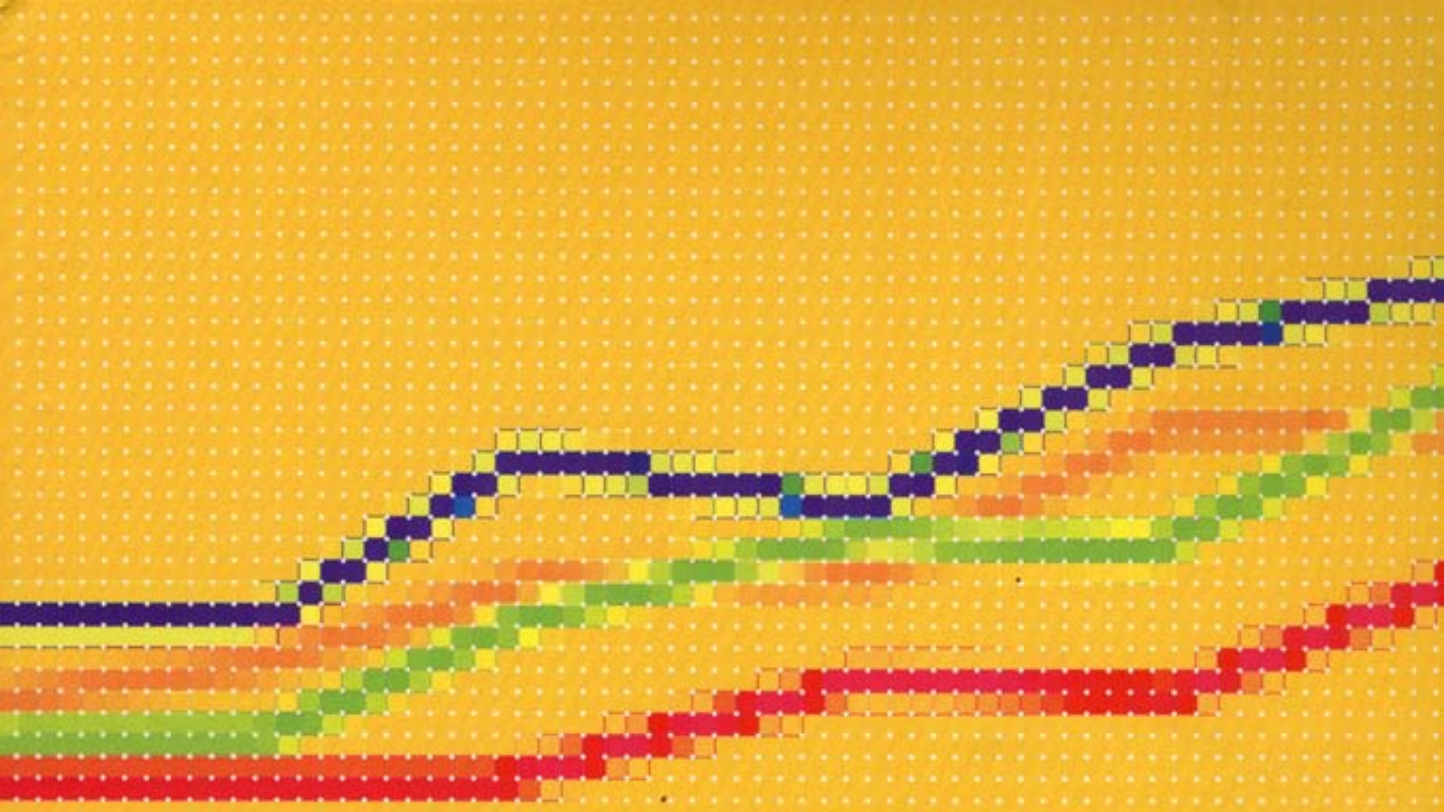


Incluye CD

6ª edición

Estadística para administración y economía

Paul Newbold
William L. Carlson
Betty Thorne

PEARSON
Prentice Hall

Estadística para Administración y Economía

Estadística para Administración y Economía

SEXTA EDICIÓN

Paul Newbold
University of Nottingham

William L. Carlson
St. Olaf College

Betty M. Thorne
Stetson University

Traducción
Esther Rabasco Espáriz

Revisión Técnica
Luis Toharia
Universidad de Alcalá de Henares

Prentice Hall, Upper Saddle River, New Jersey 07458 • Madrid

Datos de catalogación bibliográfica

Paul Newbold, William L. Carlson y Betty M. Thorne
Estadística para Administración y Economía
PEARSON EDUCACIÓN, S.A., Madrid, 2008

ISBN: 978-84-8322-403-8
Materia: 519.5 Métodos estadísticos

Formato 195 × 250 mm Páginas: 1088

Paul Newbold, William L. Carlson y Betty M. Thorne
Estadística para Administración y Economía

ISBN: 978-84-8322-403-8
Depósito legal:

PEARSON PRENTICE HALL es un sello editorial autorizado de PEARSON EDUCACIÓN, S.A.

Equipo editorial:
Editor: Alberto Cañizal
Técnico editorial: Elena Bazaco

Equipo de producción:
Director: José Antonio Clares
Técnico: José Antonio Hernán

Diseño de cubierta: Equipo de diseño de PEARSON EDUCACIÓN, S.A.

Composición: COPIBOOK, S.L.

Impreso por:

IMPRESO EN ESPAÑA - PRINTED IN SPAIN

Este libro ha sido impreso con papel y tintas ecológicos

Dedico este libro a mi mujer Charlotte, a nuestros hijos Andrea, Douglas y Larry y a nuestros nietos Ezra, Savanah, Rellana, Anna, Eva Rose y Emily

William L. Carlson

Dedico este libro a mi marido Jim y a nuestra familia Jennie, Ann, Renee, Jon, Chris, Jon, Marius, Mihaela, Cezara y Anda

Betty M. Thorne

SOBRE LOS AUTORES

Bill Carlson es profesor emérito de economía en St. Olaf College, donde lleva 31 años enseñando, ha sido varias veces director de departamento y ha desempeñado diversas funciones administrativas, entre las que se encuentra la de Director del Centro de Cálculo. También ha ocupado diversos cargos en la administración pública de Estados Unidos y en la Universidad de Minnesota, además de pronunciar conferencias en numerosas universidades. Fue elegido miembro honorario de Phi Beta Kappa. También trabajó 10 años en el sector privado y en empresas de consultoría antes de iniciar su carrera en St. Olaf. Se licenció en ingeniería en la Michigan Technological University (BS), realizó un Máster (MS) en el Illinois Institute of Technology y se doctoró (Ph.D.) en Administración Cuantitativa de Empresas en la Rackham Graduate School de la Universidad de Michigan. Entre sus investigaciones, se encuentran numerosos estudios sobre la administración de empresas, la seguridad vial y la enseñanza de la estadística. Ha publicado anteriormente dos libros de texto de estadística. Ha sido profesor encargado de numerosos grupos de estudiantes que han realizado estancias de estudio en diversos países de todo el mundo. Entre los cargos que ocupa actualmente se encuentran el de Director Ejecutivo del Cannon Valley Elder Collegium. Disfruta con sus nietos y con la ebanistería, y le encanta viajar, leer y que le encarguen trabajos en la parte septentrional del estado de Wisconsin.

Betty M. Thorne, autora, investigadora y profesora galardonada con premios a la docencia, es profesora de Ciencias de la Decisión y Directora de Estudios de Grado en la School of Business Administration de Stetson University en DeLand (Florida). Galardonada con el McEniry Award for Excellence in Teaching de la Stetson University, el máximo premio que se concede a un profesor de la Stetson University, también ha recibido el Outstanding Teacher of the Year Award y el Professor of the Year Award de la School of Business Administration en Stetson. Enseña asimismo en el programa de verano de Stetson University en Innsbruck (Austria); el College of Law

de Stetson University; el programa de MBA Ejecutivo de Stetson University, y el Executive Passport Program de Stetson University. En 2004 y 2005, fue nombrada mejor profesora del programa JD/MBA del College of Law de Stetson. Se licenció en Geneva College e hizo el Máster y el Doctorado en la Universidad de Indiana. Es miembro del comité de planificación y Secretaria/Tesorera de las jornadas tituladas *Making Statistics More Effective in Schools and Business*, en las que se reúne anualmente con estadísticos para debatir sobre cuestiones de investigación y enseñanza. También es miembro del Decision Sciences Institute, de la American Society for Quality y de la American Statistical Association. Participa en un estudio quinquenal titulado North American Fareston versus Tamoxifen Adjuvant (NAFTA) Trial sobre el cáncer de mama (http://www.naftatrial.com).

Ella y su marido, Jim, tienen cuatro hijos. Viajan mucho, asisten a clases de teología, participan en organizaciones internacionales dedicadas a ayudar a niños desfavorecidos y hacen trabajo apostólico en Rumanía.

CONTENIDO ABREVIADO

CONTENIDO

PRÓLOGO

AUDIENCIA A LA QUE VA DIRIGIDO

Estadística para los negocios y la economía (6.ª edición) se ha escrito para satisfacer la necesidad de un libro de texto que ofrezca una buena introducción a la estadística para los negocios que permita comprender los conceptos y haga hincapié en la resolución de problemas poniendo ejemplos realistas del mundo de la empresa y de la economía.

- Programas de máster o de licenciatura que enseñen estadística para los negocios.
- Programas de doctorado y de licenciatura de economía.
- Programas de MBA ejecutivo.
- Cursos de doctorado de estadística empresarial.

CONTENIDO

Hemos escrito este libro con el fin de ofrecer una buena introducción a los métodos estadísticos aplicados para que sus lectores puedan realizar un sólido análisis estadístico en muchas situaciones empresariales y económicas. Hemos hecho hincapié en la comprensión de los supuestos que son necesarios para realizar un análisis profesional. Con los ordenadores modernos, es fácil calcular a partir de los datos las salidas necesarias para muchos métodos estadísticos. Es tentador, pues, aplicar meramente sencillas «reglas» utilizando estas salidas, enfoque que se adopta en numerosos libros de texto. El nuestro es combinar los conocimientos con muchos ejemplos y ejercicios y mostrar que la comprensión de los métodos y de sus supuestos es útil para entender los problemas empresariales y económicos.

NUEVO EN ESTA EDICIÓN

Hemos actualizado y ampliado la sexta edición de este libro para satisfacer mejor las necesidades de los usuarios y ofrecer más flexibilidad. En esta edición, hemos introducido importantes cambios y novedades. Éstos son:

- Un nuevo diseño para la presentación de la estadística descriptiva.
- En cada apartado, hemos añadido ejercicios básicos antes de los ejercicios aplicados.
- Hemos introducido nuevos ejercicios aplicados que colocan a los estudiantes en situaciones empresariales reales poniendo el énfasis en las aplicaciones informáticas.

- Hemos dividido el análisis de los intervalos de confianza y del contraste de hipótesis en un capítulo dedicado a una población y otro dedicado a dos poblaciones en respuesta a las sugerencias de los usuarios y de los revisores.
- Presentaciones revisadas y más claras de los métodos de regresión simple y múltiple.
- Presentamos el análisis de cartera utilizando valores correlacionados con un extenso número de ejercicios aplicados.
- Hemos adoptado nuevos enfoques para presentar los datos utilizando imágenes gráficas.

A LOS ESTUDIANTES

El CD-ROM que acompaña a este libro contiene todos los ficheros de datos utilizados en el libro que son necesarios para hacer los problemas y los ejercicios, así como el programa TreePlan y su documentación. El PowerPoint y otros ficheros relevantes pueden encontrarse en la página web del libro (www.prenhall.com/newbold).

A LOS PROFESORES

Los ficheros de las soluciones de los capítulos y las presentaciones en PowerPoint de este libro se encuentran en formato digital descargable. Visite el Instructor Resource Center en el catálogo de Prentice Hall (www.prenhall.com). Para registrarse con el fin de utilizar los recursos del Instructor Resource Center se necesita un código de acceso como educador de Pearson.

Cada vez mejor

Una vez que se registre, no tendrá que rellenar más formularios o recordar múltiples nombres de usuario y contraseñas para acceder a nuevos títulos y/o ediciones. Como profesor registrado, puede acceder directamente a los ficheros de recursos y recibir inmediatamente el acceso y las instrucciones para instalar en el servidor de su universidad el contenido del gestor del curso.

¿Necesita ayuda?

Contamos con un entregado equipo de apoyo técnico para ayudar a los profesores a resolver cuestiones relacionadas con el material auxiliar que acompaña a este libro. Visite http://247.prenhall.com/ para las respuestas a las preguntas formuladas frecuentemente y los números de teléfono gratuitos de ayuda.

AGRADECIMIENTOS

Nos gustaría dar las gracias a las siguientes personas que han revisado el libro y han hecho perspicaces sugerencias para esta edición:

Mr. C. Patrick Kohrman-Penn State University, Berks Campus
James Thorson-Southern Connecticut State University
Mamnoon Jamil-Rutgers University, Camden
Zhimin Huang-Adelphi University

Renee Fontenot-University of Texas, Permian Basin
Allen Lynch-Mercer University
Bulent Uyar-University of Northern Iowa
David Hudgins-University of Oklahoma
Allan Lacayo-Diablo Valley College
J. Morgan Jones-University of North Carolina
Eugene Allevato-Woodbury University
Patricia Odell-Bryant University
Jay DeVore-California Polytechnic State University
Valerie Bencivenga-University of Texas
Myles J. Callan-University of Virginia
Andrew Narwold-University of San Diego
Anthony Smith-Carnegie Mellon University
Peter Baxendale-University of Southern California
Steen Anderson-Aarhus School of Business, Denmark
Eric Bentzen-Copenhagen Business School, Denmark
Hans Geilnkirchen-Erasmus University, Netherlands
Peter Reiss-Stanford University
David Hudgins-University of Oklahoma
Robert Lemke-Lake Forest College
Michael Gordinier-Washington University
Fred Wenstop-Norwegian School of Management
Sheri Aggarwal-University of Virginia
Jorgen Lauridsen-University of Southern Denmark
Robert Gillette-University of Kentucky
Peter Boatwright-Carnegie Mellon
Mark Kamstra-Simon Fraser
Albert Madansky-University of Chicago
Jeff Russell-University of Chicago
Nick Polsen-University of Chicago
Aaron Smith-University of Virginia
Yu-Chi Cheng-University of Notre Dame
Professor Mohanty-California State, Los Angeles
Ken Alexander-University of Southern California
Mendy Fygenson-University of Southern California
Matthew White-Stanford University
Stefanos Zenios-Stanford University
Lawrence Brown-Pennsylvania State University
Abba Krieger-Pennsylvania State University
Harvey Singer-George Mason
William Hausman-William and Mary University of Iowa
Jim Swanson-Central Missouri University
C. Barry Pfitzner-Randolf-Macon College

También estamos agradecidos a Annie Puciloski que ha revisado la precisión de esta edición y especialmente a Sandra Krausman, GGS Production Services, por su ayuda y pericia.

Por lo que se refiere al St. Olaf College, debemos dar las gracias a Priscilla Hall, ayudante administrativo de St. Olaf, por la labor realizada en algunas partes del libro y su

dirección del trabajo de varios estudiantes que han colaborado en el libro, entre los que se encuentran Michael Loop, Holly Malcomson, Erin McMurtry, Nelly Schwinghammer, Catharina Zuber. Este libro no habría sido posible sin su colaboración.

Por lo que se refiere a Stetson University, también damos las gracias a Jim Scheiner, Paul Dascher, Marie Gilotti, Sean A. Thomas, John Tichenor y Emma Astrom y especialmente a Jennie Bishop (*Computer Programmer Analyst II*, *State of Florida*, *Volusia County Health Department*).

Agradecemos, además, especialmente a nuestras familias su apoyo durante las numerosas horas dedicadas a este libro. Bill Carlson da las gracias especialmente a su mujer Charlotte y a sus hijos adultos Andrea, Douglas y Larry. Betty Thorne da las gracias especialmente a su marido Jim y a sus hijos adultos Jennie Bishop, Ann Thorne, Renee Payne y Jon Thorne; así como a Marius, Mihaela, Cezara y Anda Sabou.

Los autores agradecen las sólidas bases y tradición creadas por el autor original, Paul Newsbold. Paul comprendió la importancia del análisis estadístico riguroso y de sus fundamentos. Se dio cuenta de que hay algunas complejas ideas que es necesario desarrollar y se esforzó en ofrecer explicaciones claras de difíciles ideas. Además, estas ideas sólo son útiles cuando se utilizan para resolver problemas realistas. En ediciones anteriores, se incluyeron, pues, muchos ejemplos y muchos ejercicios aplicados. Nos hemos esforzado en mantener y ampliar esta tradición para hacer un libro que satisfaga las necesidades de los futuros líderes empresariales en la era de la información.

Si el lector tiene alguna sugerencia o corrección, puede ponerse en contacto con los autores a través del correo electrónico en carlsoncharbill@msn.com; bthorne@stetson.edu.

Capítulo 1

¿Por qué estudiar estadística?

Esquema del capítulo

Introducción

En nuestra era de la información, el mundo abunda en datos. En los artículos de los periódicos y en los reportajes de la televisión, se hacen afirmaciones como «El Dow Jones ha caído 6 puntos hoy» o «El índice de precios de consumo subió un 0,8 por ciento el mes pasado» o «la última encuesta indica que la tasa de aprobación del presidente es hoy de un 63 por ciento» o «El 98 por ciento de los pacientes de un estudio clínico no experimentó ningún efecto secundario significativo con un nuevo medicamento contra el cáncer de mama». Cada vez es más frecuente que para hacer una valoración inteligente de los acontecimientos actuales, necesitemos asimilar e interpretar una cantidad considerable de datos. La Administración, las empresas y los investigadores científicos gastan miles de millones de dólares en la recogida de datos. La Administración ha contribuido a ello, tanto recogiendo datos ella misma como obligando a las empresas a dar información. El sector privado también ha tenido que ver en ello. Las aireadas encuestas Gallup de las actitudes de los votantes y los índices de audiencia de Nielsen de los programas de televisión de la semana no son más que la punta de un enorme iceberg de estudios de mercado. La cantidad de datos recogidos ha aumentado a un ritmo extraordinario en los últimos años.

Debemos explicar todos los datos. La era de la informática nos ha permitido tanto procesar, resumir y analizar rápidamente los datos como producir y almacenar más datos. Los computadores ponen al alcance de la mano muchos datos, como las cotizaciones bursátiles. Debemos analizarlos e interpretarlos correctamente.

1.1. La toma de decisiones en un entorno incierto

Las decisiones a menudo se basan en información incompleta. Por ejemplo, se supone que los estudiantes universitarios de primer año, cuando son admitidos en la universidad, seleccionan una carrera. Sin embargo, muchos de estos estudiantes pueden no tener una meta profesional clara. Por poner otro ejemplo, los enfermos de cáncer pueden ser invitados a participar en un estudio clínico para probar un nuevo medicamento experimental (véase referencia bibliográfica 1) cuando aún no se dispone de información sobre los efectos secundarios, las tasas de supervivencia y las tasas de recurrencia de esta nueva medicación. Asimismo, las decisiones empresariales normalmente se toman en un entorno en el que los responsables de tomarlas no pueden estar seguros de la futura conducta de los factores que acabarán afectando al resultado de las distintas opciones consideradas.

Cuando un fabricante presenta una oferta para hacerse con un contrato, no está totalmente seguro de cuáles serán los costes totales ni de qué ofertas presentarán los competidores. A pesar de esta incertidumbre, debe hacer una oferta. Un inversor no sabe con seguridad si los mercados financieros estarán boyantes, estables o deprimidos. No obstante, debe elegir las acciones, los bonos y los instrumentos del mercado de dinero de manera que su cartera esté equilibrada sin saber cómo evolucionará el mercado en el futuro.

Consideremos las siguientes afirmaciones:

- «El precio de las acciones de IBM será más alto dentro de seis meses que ahora».
- «Si el déficit presupuestario público es tan elevado como se prevé, los tipos de interés se mantendrán altos el resto del año».
- «La renta anual de un titulado universitario será mayor que la renta anual de una persona que no tenga estudios universitarios».

Cada una de estas afirmaciones contiene un lenguaje que sugiere la existencia de una cantidad espuria de certeza. En el momento en el que se hicieron las afirmaciones, era imposible estar *seguro* de que eran ciertas. Aunque un analista crea que lo que ocurrirá en los próximos meses será tal que se prevé que el precio de las acciones de IBM subirá durante ese periodo, no estará seguro de eso. Por lo tanto, las afirmaciones deben modificarse como indican los siguientes ejemplos:

- «El precio de las acciones de IBM *probablemente* será más alto dentro de seis meses que ahora».
- «Si el déficit presupuestario público es tan elevado como se prevé, es *probable* que los tipos de interés se mantengan altos durante el resto del año».
- «La renta anual de un titulado universitario *probablemente* será mayor que la renta anual de una persona sin estudios universitarios».

Es muy importante pensar bien cómo se dicen las cosas. No es correcto sustituir las afirmaciones injustificadamente precisas por afirmaciones innecesariamente vagas. Al fin y al cabo, ¿qué significa «probablemente» o «es probable que»? Debe ponerse especial cuidado en expresar las ideas que se pretende expresar, sobre todo cuando se trata de probabilidades o cuando hay incertidumbre.

EJERCICIOS

Ejercicios básicos

1.1. Modifique las afirmaciones siguientes para que reflejen una posible incertidumbre:

a) El mejor instrumento para mejorar la cuota de mercado de este producto es una campaña publicitaria destinada al grupo de edad 18-24 años.

b) Si se presenta una oferta de esta cuantía, será más baja que las del competidor y el contrato estará asegurado.
c) El coste de la gasolina será más alto en Estados Unidos dentro de 2 meses.

1.2. Ponga un ejemplo de una decisión de comercialización que debe tomarse en condiciones de incertidumbre.

1.3. Ponga un ejemplo de una decisión financiera que debe tomarse en condiciones de incertidumbre.

1.2. El muestreo

Antes de introducir un nuevo producto en el mercado, su fabricante quiere saber cuál será el nivel probable de demanda y es posible que realice una encuesta de mercado. Lo que le interesa, en realidad, son *todos* los compradores potenciales (la población). Sin embargo, las poblaciones a menudo son tan grandes que es difícil analizarlas; sería imposible o prohibitivo recoger toda la información de una población. Incluso en las circunstancias en las que parece que se dispone de suficientes recursos, las limitaciones de tiempo obligan a examinar un subconjunto (muestra).

Población y muestra

Una **población** es el conjunto completo de todos los objetos que interesan a un investigador. El tamaño de la población, N, puede ser muy grande o incluso infinito. Una **muestra** es un subconjunto observado de valores poblacionales que tiene un tamaño muestral que viene dado por n.

Ejemplos de poblaciones son:

- Todos los votantes inscritos en un país.
- Todos los estudiantes de una universidad.
- Todas las familias que viven en una ciudad.
- Todas las acciones que cotizan en una bolsa de valores.
- Todas las reclamaciones que recibe en un año dado una compañía de seguros médicos.
- Todas las cuentas pendientes de cobro de una empresa.

Nuestro objetivo final es hacer afirmaciones basadas en datos muestrales que tengan alguna validez sobre la población en general. Necesitamos, pues, una muestra que sea representativa de la población. ¿Cómo podemos lograrlo? Uno de los principios importantes que debemos seguir en el proceso de selección de la muestra es la aleatoriedad.

Muestreo aleatorio

El **muestro aleatorio simple** es un método que se emplea para seleccionar una muestra de n objetos de una población en el que cada miembro de la población se elige estrictamente al azar, cada miembro de la población se elige con la misma probabilidad y todas las muestras posibles de un tamaño dado, n, tienen la misma probabilidad de ser seleccionadas. Este método es tan frecuente que generalmente se suprime el adjetivo *simple* y la muestra resultante se denomina **muestra aleatoria**.

El muestreo se utiliza mucho en todas las áreas de los negocios, así como en otras disciplinas. Para averiguar si un proceso de producción está funcionando correctamente, se selecciona una muestra de bienes producidos. Las auditorías de las cuentas pendientes de cobro generalmente se basan en una muestra. Durante los años de elecciones presidenciales, se hacen estimaciones de las preferencias de los votantes a partir de muestras de votantes;

también puede hacerse una encuesta a la salida de los colegios electorales para predecir qué candidato obtendrá más votos. Sin embargo, tomar una muestra es meramente un medio para llegar a un fin. Necesitamos estudiar estadística, *no para hacer afirmaciones sobre la muestra sino*, *más bien*, *para extraer conclusiones sobre la población en general*. La estadística es el estudio de cómo se toman decisiones sobre una población cuando la información procede de una muestra. Siempre quedará alguna incertidumbre.

Supongamos que queremos saber cuál es la edad media de los votantes de un país. Es evidente que el tamaño de la población es tan grande que sólo podríamos tomar una muestra aleatoria, por ejemplo, 500 votantes, y calcular su edad media. Como esta media se basa en datos muestrales, se llama *estadístico*. Si pudiéramos calcular la edad media de toda la población, la media resultante se llamaría *parámetro*. En este libro veremos cómo se toman decisiones sobre un parámetro, basándose en un estadístico. Debemos darnos cuenta de que siempre habrá una cierta incertidumbre, ya que no se conoce el valor exacto del parámetro.

Parámetro y estadístico

Un **parámetro** es una característica específica de una población. Un **estadístico** es una característica específica de una muestra.

EJERCICIOS

Ejercicios básicos

1.4. Ponga un ejemplo de un parámetro en cada una de las siguientes poblaciones:

a) Las rentas de todas las familias que viven en una ciudad.
b) Los rendimientos anuales de todas las acciones que cotizan en una bolsa de valores.
c) Los costes de todas las reclamaciones que recibe en un año dado una compañía de seguros médicos.
d) Los valores de todas las cuentas pendientes de cobro de una empresa.

1.5. Su universidad ha encuestado a sus estudiantes para averiguar el tiempo semanal medio que dedican a navegar por Internet.

a) ¿Cuál es la población?
b) ¿Cuál es la muestra?
c) ¿Cuál es el estadístico?
d) ¿Es el valor de 6,1 horas un parámetro o un estadístico?

1.6. Una compañía aérea sostiene que menos de un 1 por ciento de los vuelos programados que despegan del aeropuerto de Nueva York sale tarde. Se ha observado que el 1,5 por ciento de una muestra aleatoria de 200 vuelos salió más tarde de la hora prevista.

a) ¿Cuál es la población?
b) ¿Cuál es la muestra?
c) ¿Cuál es el estadístico?
d) ¿Es 1,5 por ciento un parámetro o un estadístico?

1.3. Estadística descriptiva e inferencial

Para pensar en términos estadísticos hay que seguir una serie de pasos que van desde la definición del problema hasta la toma de decisiones. Una vez identificado y definido el problema, se recogen datos producidos mediante diversos procesos de acuerdo con un diseño y se analizan utilizando uno o más métodos estadísticos. De este análisis se obtiene información. La información se convierte, a su vez, en conocimiento, utilizando los resultados de las experiencias específicas, la teoría y la literatura y aplicando métodos estadísticos adicionales. Para convertir los datos en un conocimiento que lleva a tomar mejores decisiones se utiliza tanto la estadística descriptiva como la inferencial.

Estadística descriptiva e inferencial

La **estadística descriptiva** está formada por los métodos gráficos y numéricos que se utilizan para resumir y procesar los datos y transformarlos en información. La **estadística inferencial** constituye la base para hacer predicciones, previsiones y estimaciones que se utilizan para transformar la información en conocimiento.

Descripción de los datos

En el ejemplo 1.1 vemos una tabla de la producción diaria de una fábrica de cereales.

EJEMPLO 1.1. Producción de cereales (estadística descriptiva)

Un jefe de producción de Cereales de Trigo formó un equipo de empleados para estudiar el proceso de producción de cereales. Durante la primera fase del estudio, se pesó una selección aleatoria de cajas y se midió la densidad del producto. A continuación, el jefe quería estudiar datos relacionados con las pautas de producción diaria. Se hallaron los niveles de producción (en miles) de un periodo de 10 días. Represente estos resultados gráficamente y comente sus observaciones:

Día	1	2	3	4	5	6	7	8	9	10
Cajas (miles)	84	81	85	82	85	84	109	110	60	63

Solución

En la Figura 1.1, el jefe de producción puede identificar los días de baja producción, así como los días de mayor producción.

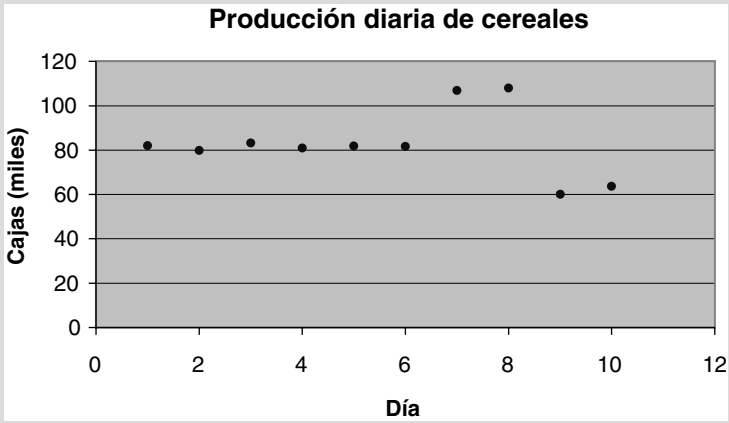


Figura 1.1. Producción diaria de cereales de trigo.

No parecía que hubiera mucha diferencia en el número de cajas producidas en los seis primeros días. Había variaciones de un día a otro, pero los seis puntos tenían valores numéricos muy parecidos. Sin embargo, en los días 7 y 8 el nivel de producción parecía que era más alto. En cambio, en los días 9 y 10 parecía que era más bajo. Basándose en estas observaciones, el equipo intentó identificar las causas por las que la productividad era más alta y más baja. Por ejemplo, tal vez en los días 9 y 10 estuvieran ausentes trabajadores clave o hubiera cambiado el método de producción o hubieran cambiado las materias primas. También se podrían identificar las causas por las que aumentó la productividad en los días 7 y 8.

Realización de inferencias

La estadística inferencial es un proceso, no un mero resultado numérico. Este proceso puede consistir en una estimación, un contraste de hipótesis, un análisis de relaciones o una predicción. En primer lugar, podemos querer *estimar un parámetro*. Supongamos que Florin's Flower Mart quiere desarrollar una nueva estrategia de comercialización. Podría ser útil la información sobre los hábitos de gasto de los clientes de la floristería. Florin puede querer:

- Estimar la edad *media* de los clientes de la tienda.
- Estimar la diferencia entre la cantidad media que los clientes cargan a una Visa y la cantidad media que cargan a American Express.
- Estimar la proporción de clientes que están insatisfechos con el sistema de reparto de la tienda.

En segundo lugar, podemos querer *contrastar una hipótesis* sobre un parámetro. Por ejemplo, Florin puede querer:

- Contrastar si los clientes tienen este año una preferencia por el color de las rosas distinta a la del año pasado.
- Contrastar si menos del 25 por ciento de los clientes de la tienda son turistas.
- Contrastar si las ventas son mayores los fines de semana que el resto de los días de la semana.
- Contrastar si la cantidad *media* que gastaron los clientes en su última compra superó los 40 $.

Las respuestas a estos tipos de preguntas pueden ayudar a Florin a lanzar una campaña publicitaria que le permita reducir los costes, incrementar los beneficios y aumentar la satisfacción de los clientes.

En tercer lugar, podemos querer *analizar las relaciones* entre dos o más variables. El director financiero de General Motors quiere tomar decisiones estratégicas que afectan a toda la compañía. En esos casos, puede utilizar series de datos macroeconómicos de los que puede disponerse en fuentes como el Bureau of Economic Analysis del Departamento de Comercio de Estados Unidos para analizar las relaciones entre variables como el producto interior bruto, el tipo de interés, la renta per cápita, la inversión total y la oferta monetaria, que indican la situación general de la economía nacional. El director financiero puede hacerse las siguientes preguntas:

- ¿Influye la tasa de crecimiento de la oferta monetaria en la tasa de inflación?
- Si General Motors sube un 5 por ciento el precio de los automóviles de tamaño intermedio, ¿cómo afectará la subida a las ventas de estos automóviles?
- ¿Afecta la legislación sobre el salario mínimo al nivel de desempleo?

¿Cómo se comienza a responder a la pregunta sobre el efecto que puede producir una subida de los precios en la demanda de automóviles? La teoría económica básica nos dice que, manteniéndose todo lo demás constante, una subida del precio va acompañada de una reducción de la cantidad demandada. Sin embargo, esa teoría es puramente cualitativa. No nos dice *cuánto* disminuye la cantidad demandada. Para avanzar más, hay que recoger información sobre cómo ha respondido la demanda a las variaciones del precio en el pasado y evaluarla. Estudiando estadística inferencial, aprenderemos a recoger información y a analizar relaciones.

En cuarto lugar, podemos necesitar **predecir**, es decir, hacer predicciones fiables. Las decisiones de inversión deben hacerse mucho antes de que pueda llevarse un nuevo

producto al mercado y, evidentemente, es deseable tener predicciones de la situación en la que se encontrará probablemente el mercado dentro de unos años. Cuando los productos están consolidados, las predicciones sobre las ventas a corto plazo son importantes para decidir los niveles de existencias y los programas de producción. Las predicciones de los futuros tipos de interés son importantes para una empresa que tiene que decidir si emite o no nueva deuda. Para formular una política económica coherente, el gobierno necesita predicciones de los resultados probables de variables como el producto interior bruto. Las predicciones de los futuros valores dependen de las regularidades descubiertas en la conducta anterior de estas variables. Por lo tanto, se recogen datos sobre la conducta anterior de la variable que va a predecirse y sobre la conducta de otras variables relacionadas con ella. Utilizaremos la estadística inferencial para analizar esta información y sugerir entonces las tendencias futuras probables.

EJERCICIOS

Ejercicios básicos

1.7. Suponga que es dueño de una tienda de alimentación.

a) Ponga un ejemplo de una pregunta que podría responderse utilizando la estadística descriptiva.

b) Ponga un ejemplo de una pregunta en la que sería útil estimar un parámetro.

c) Ponga un ejemplo de una pregunta sobre una posible relación entre dos variables que tienen interés para su tienda de alimentación.

d) Ponga un ejemplo de una cuestión en la que hay que hacer una predicción.

1.8. Averigüe si debe utilizarse la estadística descriptiva o la inferencial para obtener la siguiente información:

a) Un gráfico que muestra el número de botellas defectuosas producidas durante el turno de día a lo largo de una semana.

b) Una estimación del porcentaje de empleados que llegan tarde a trabajar.

c) Una indicación de la relación entre los años de experiencia de los empleados y la escala salarial.

RESUMEN

Las decisiones deben tomarse en condiciones de incertidumbre. Todas las áreas de negocios, así como otras disciplinas, utilizan la estadística para tomar esas decisiones. Los contables pueden necesitar seleccionar muestras para realizar auditorías. Los inversores financieros utilizan la estadística para comprender las fluctuaciones del mercado y elegir entre varias inversiones de cartera. Los directivos que quieren saber si los clientes están satisfechos con los productos o los servicios de su compañía pueden utilizar encuestas para averiguarlo. Los ejecutivos de marketing pueden querer información sobre las preferencias de los clientes, sus hábitos de compra o las características demográficas de los compradores por Internet. En cada una de estas situaciones, debemos definir meticulosamente el problema, averiguar qué datos se necesitan, recogerlos, resumirlos y hacer inferencias y tomar decisiones basadas en los datos obtenidos. La teoría estadística es esencial desde la definición inicial del problema hasta la decisión final y puede llevar a reducir los costes, a obtener más beneficios, a mejorar los procesos y a aumentar la satisfacción de los clientes.

TÉRMINOS CLAVE

- estadística descriptiva, 5
- estadística inferencial, 5
- estadístico, 4
- muestra, 3
- muestra aleatoria, 3
- muestreo aleatorio simple, 3
- parámetro, 4
- población, 3

EJERCICIO Y APLICACIÓN DEL CAPÍTULO

1.9. Se hizo a una muestra aleatoria de 100 estudiantes universitarios una serie de preguntas para obtener datos demográficos sobre su nacionalidad, la especialización cursada, el sexo, la edad, el curso en el que están y su nota media hasta ese momento. También se les hizo otras preguntas sobre su grado de satisfacción con el aparcamiento del campus universitario, las residencias del campus y los comedores del campus. Las respuestas a estas preguntas sobre su satisfacción se midieron en una escala de 1 a 5, donde 5 era el nivel de satisfacción más alto. Por último, se les preguntó si, cuando se graduaran, tenían intención de seguir estudios de postgrado en un plazo de 5 años (0: no; 1: sí). Estos datos se encuentran en el fichero de datos **Findstad and Lie Study**.

a) Ponga un ejemplo de cómo se aplica la estadística descriptiva a estos datos.

b) Ponga un ejemplo de una pregunta que conlleve una estimación a la que podría responderse por medio de la estadística inferencial.

c) Ponga un ejemplo de una relación posible entre dos variables.

Bibliografía

1. The North American Fareston versus Tamoxifen Adjuvant Trial for Breast Cancer. www.naftatrial.com.

Capítulo 2

Descripción gráfica de los datos

Esquema del capítulo

Introducción

Una vez que definimos con cuidado un problema, necesitamos recoger datos. A menudo el número de observaciones recogidas es tan grande que los resultados efectivos del estudio no están claros. Nuestro objetivo en este capítulo es resumir los datos de manera que tengamos una imagen clara y precisa. Queremos reducir lo más posible una masa de datos, evitando al mismo tiempo la posibilidad de ocultar características importantes por reducirlos excesivamente. Por desgracia, no existe una única «manera correcta» de describir los datos. La línea de ataque adecuada normalmente es específica de cada problema y depende de dos factores: el tipo de datos y el fin del estudio.

Se ha dicho que una imagen vale más que mil palabras. Asimismo, un gráfico vale más que mil cifras. En este capítulo, introducimos tablas y gráficos que nos ayudan a comprender mejor los datos y que constituyen una ayuda visual para tomar mejores decisiones. Los informes mejoran con la inclusión de tablas y gráficos adecuados, como distribuciones de frecuencia, gráficos de barras, gráficos de tarta, diagramas de Pareto, gráficos de series temporales, histogramas, diagramas de tallo y hojas u ojivas. La visualización de los datos es importante. Siempre debemos preguntarnos qué sugiere el gráfico sobre los datos, qué es lo que vemos.

La comunicación a menudo es la clave del éxito y la comunicación de datos no es una excepción. El análisis y la interpretación correctos de los datos son esenciales para comunicar los resultados de una manera que tenga sentido. Los gráficos y los diagramas pueden mejorar nuestra comunicación de los datos a los clientes, los proveedores, los consejos de administración u otros grupos. En capítulos posteriores presentaremos métodos numéricos para describir los datos.

2.1. Clasificación de las variables

Las variables pueden clasificarse de varias formas. Uno de los métodos de clasificación se refiere al tipo y la cantidad de información que contienen los datos. Los datos son categóricos o numéricos. Otro método consiste en clasificar los datos por niveles de medición, dando variables cualitativas o cuantitativas.

Categóricas o numéricas

Las **variables categóricas** producen respuestas que pertenecen a grupos o categorías. Por ejemplo, las respuestas a preguntas sí/no son categóricas. Las respuestas a «¿Tiene usted teléfono móvil?» y «¿Ha estado alguna vez en Oslo?» se limitan a un sí o un no. Una compañía de seguros médicos puede clasificar las reclamaciones incorrectas según el tipo de errores, como los errores de procedimiento y diagnóstico, los errores de información al paciente y los errores contractuales. Otros ejemplos de variables categóricas son las preguntas sobre el sexo, el estado civil y la carrera universitaria. A veces, las variables categóricas permiten elegir entre varias opciones, que pueden ir desde «totalmente en desacuerdo» hasta «totalmente de acuerdo». Consideremos, por ejemplo, una evaluación del profesorado en la que los estudiantes tienen que responder a afirmaciones como «El profesor de este curso es un buen profesor» (1: totalmente en desacuerdo; 2: un poco en desacuerdo; 3: ni de acuerdo ni en desacuerdo; 4: un poco de acuerdo; 5: totalmente de acuerdo).

Las **variables numéricas** pueden ser variables discretas o variables continuas. Una **variable numérica discreta** puede tener (pero no necesariamente) un número finito de valores. Sin embargo, el tipo más frecuente de variable numérica discreta con el que nos encontraremos produce una respuesta que proviene de un proceso de recuento. Ejemplos de variables numéricas discretas son el número de estudiantes matriculados en una clase, el número de créditos universitarios obtenidos por un estudiante al final de un cuatrimestre, el número de acciones de Microsoft que contiene la cartera de un inversor y el número de reclamaciones de indemnizaciones presentado tras un huracán.

Una **variable numérica continua** puede tomar cualquier valor de un intervalo dado de números reales y normalmente proviene de un proceso de medición (no de recuento). Ejemplos de variables numéricas continuas son la altura, el peso, el tiempo, la distancia y la temperatura. Una persona puede decir que mide 1,89 metros, pero en realidad puede tener una estatura de 1,81, 1,79 o algún otro número similar, dependiendo de la precisión del instrumento utilizado para medir la estatura. Otros ejemplos de variables numéricas continuas son el peso de las cajas de cereales, el tiempo que se hace una persona en una carrera y la distancia entre dos ciudades. En todos los casos, el valor podría desviarse dentro de un cierto margen, dependiendo de la precisión del instrumento de medición utilizado. En las conversaciones diarias tendemos a truncar las variables y a tratarlas como si fueran variables discretas sin pensarlo ni siquiera dos veces. Sin embargo, la diferencia es muy importante en estadística, ya que es uno de los factores de los que depende que un método estadístico sea mejor que otro en un determinado caso.

Niveles de medición

También podemos dividir los datos en **cualitativos** y **cuantitativos**. Con datos cualitativos, la «diferencia» entre los números no tiene ningún significado mensurable. Por ejemplo, si a un jugador de baloncesto se le asigna el número «20» y a otro el número «10», no pode-

mos extraer la conclusión de que el primero es el doble de bueno que el segundo. Sin embargo, con datos cuantitativos la diferencia entre los números tiene un significado mensurable. Cuando un estudiante obtiene una puntuación de 90 en un examen y otro obtiene una puntuación de 45, la diferencia es mensurable y tiene un significado.

Veremos que los datos cualitativos pueden ser niveles de medición nominales y ordinales. Los datos cuantitativos pueden ser niveles de medición basados en intervalos y en razones.

Los niveles de medición nominales y ordinales se refieren a los datos que se obtienen con preguntas categóricas. Las respuestas a preguntas sobre el sexo, el país de origen, la afiliación política y la propiedad de un teléfono móvil son **nominales**. Se considera que los datos nominales son el tipo de datos más bajo o más débil, ya que la identificación numérica se elige estrictamente por comodidad.

Los valores de las variables nominales son palabras que describen las categorías o clases de respuestas. Los valores de la variable sexo son hombre y mujer; los valores de «¿Ha estado alguna vez en Oslo?» son «sí» y «no». Asignamos arbitrariamente un código o un número a cada respuesta. Sin embargo, este número no se emplea más que para clasificar. Por ejemplo, podríamos codificar las respuestas sobre el sexo o las respuestas sí/no de la forma siguiente:

1 = Hombres	1 = Sí
2 = Mujeres	2 = No

Los datos **ordinales** indican el orden que ocupan los objetos y, al igual que en el caso de los datos nominales, los valores son palabras que describen las respuestas. He aquí algunos ejemplos de datos ordinales y de códigos posibles:

1. Valoración de la calidad del producto (1: malo; 2: medio; 3: bueno).
2. Valoración de la satisfacción con el servicio de comedor de la universidad (1: muy insatisfecho; 2: moderadamente insatisfecho; 3: ninguna opinión; 4: moderadamente satisfecho; 5: muy satisfecho).
3. Preferencia de los consumidores entre tres tipos de bebidas refrescantes (1: el que más se prefiere; 2: segunda opción; 3: tercera opción).

En estos ejemplos, las respuestas son ordinales, es decir, siguen un orden, pero la «diferencia» entre ellas no tiene ningún significado mensurable. Es decir, la diferencia entre la primera opción y la segunda puede no ser igual que la diferencia entre la segunda y la tercera.

Los niveles de medición basados en intervalos y en razones se refieren a los datos en una escala ordenada, en la que la *diferencia* entre las mediciones tiene un significado. Una escala de intervalos indica el orden y la distancia con respecto a un cero arbitrario medidos en intervalos unitarios. Es decir, se ofrecen datos en relación con un nivel de referencia determinado arbitrariamente. La temperatura es un ejemplo clásico de este nivel de medición; los niveles de referencia determinados arbitrariamente se basan, en general, en los grados Fahrenheit o Celsius. Supongamos que hace 80 grados Fahrenheit en Orlando (Florida) y sólo 20 en St. Paul (Minnesota). Podemos extraer la conclusión de que la diferencia de temperatura es de 60 grados, pero no podemos saber si hace el cuádruple de calor en Orlando que en St. Paul. Supongamos que cuando se estableció la temperatura Fahrenheit, el punto de congelación se fijó en 500 grados. En ese caso, en nuestro ejemplo de la temperatura de Orlando y St. Paul, ésta habría sido de 548 grados en Orlando y de 488 en St. Paul (la diferencia sigue siendo de 60 grados). El año es otro ejemplo de un nivel de medición basado en intervalos; en este caso los niveles de referencia se basan en el calendario gregoriano o en el islámico.

Los datos basados en una escala de razones sí indican tanto el orden como la distancia con respecto a un cero natural y los cocientes entre dos medidas tienen un significado. Una persona que pesa 80 kilos pesa el doble que una que pesa 40; una persona que tiene 40 años es el doble de vieja que una que tiene 20.

Después de recoger datos, primero tenemos que clasificar las respuestas en categóricas o numéricas o según la escala de medición. A continuación, asignamos un número arbitrario a cada respuesta. Algunos gráficos se utilizan generalmente para las variables categóricas y otros son adecuados para las variables numéricas.

Obsérvese que los ficheros de datos normalmente contienen «valores perdidos». Por ejemplo, los encuestados pueden decidir no responder en un cuestionario a ciertas preguntas sobre el sexo, la edad, la renta o algún otro tema delicado. Los valores perdidos requieren un código especial en la fase de introducción de los datos. Si no se resuelve correctamente la cuestión de los valores perdidos, es posible que el resultado sea erróneo. Los paquetes estadísticos resuelven la cuestión de los valores perdidos de diferentes formas.

EJERCICIOS

Ejercicios básicos

2.1. Indique si cada una de las siguientes variables es categórica o numérica. Si es categórica, indique el nivel de medición. Si es numérica, ¿es discreta o continua?

a. Número de mensajes de correo electrónico enviados diariamente por un planificador financiero.
b) Coste efectivo de los libros de texto de un estudiante para un cuatrimestre.
c) Su factura mensual de electricidad.
d) Las categorías de profesores universitarios (profesor, profesor asociado, profesor ayudante, profesor colaborador).

2.2. La oficina de relaciones públicas de un equipo de baloncesto profesional quiere información sobre los aficionados que acuden a los partidos después de la temporada. En los partidos que se celebran después de la temporada, se entrega a la entrada un cuestionario a cada aficionado. ¿Es la respuesta a cada una de las siguientes preguntas categórica o numérica? Si es categórica, indique el nivel de medición. Si es numérica, ¿es discreta o continua?

a) ¿Tiene usted una entrada de temporada?
b) ¿Vive en el condado de Orange?
c) ¿Cuánto le costó realmente la entrada para este partido de después de temporada?

2.3. En una facultad universitaria se ha repartido un cuestionario entre los estudiantes para averiguar su grado de satisfacción con diversas actividades y servicios. Por ejemplo, por lo que se refiere al «método de matriculación para las clases del próximo cuatrimestre», se pide a los estudiantes que pongan una cruz en una de las casillas siguientes:

☐ muy satisfecho
☐ moderadamente satisfecho
☐ neutral
☐ moderadamente insatisfecho
☐ muy insatisfecho

¿Es la respuesta de un estudiante a esta pregunta numérica o categórica? Si es numérica, ¿es discreta o continua? Si es categórica, indique el nivel de medición.

2.4. En una encuesta reciente se pidió al profesorado de una universidad que respondiera a una serie de preguntas. Indique el tipo de datos de cada pregunta.

a) Indique su nivel de satisfacción con la carga docente (muy satisfecho; moderadamente satisfecho; neutral; moderadamente insatisfecho; muy insatisfecho).
b) ¿Cuántos artículos ha publicado en revistas con evaluación anónima durante el último año?
c) ¿Ha asistido a la última reunión del consejo de departamento?
d) ¿Cree usted que el proceso de evaluación de la docencia debe revisarse?

2.5. Se ha formulado una serie de preguntas a una muestra de clientes de una tienda de helados. Identifique el tipo de datos que se pide en cada pregunta.

a) ¿Cuál es su sabor favorito?
b) ¿Cuántas veces al mes toma helado?
c) ¿Tiene hijos de menos de 10 años que vivan en casa?
d) ¿Ha probado el último sabor de helado?

2.6. La comunidad de propietarios de viviendas ha formulado una serie de preguntas a los residentes de

una urbanización. Identifique el tipo de datos que se pide en cada pregunta.

a) ¿Jugó al golf el mes pasado en el nuevo campo de golf de la urbanización?
b) ¿Cuántas veces ha comido en el restaurante de la urbanización en los tres últimos meses?
c) ¿Tiene usted una caravana?
d) Valore el nuevo sistema de seguridad de la urbanización (muy bueno, bueno, malo, muy malo).

Ejercicios aplicados

2.7. En una universidad se realizó una encuesta a los estudiantes para obtener información sobre varias cuestiones relacionadas con la biblioteca. Los datos se encuentran en el fichero de datos **Library**.

a) Ponga un ejemplo de una variable categórica con respuestas ordinales.
b) Ponga un ejemplo de una variable categórica con respuestas nominales.
c) Ponga un ejemplo de una variable numérica con respuestas discretas.

2.8. Un grupo de estudiantes de administración de empresas realizó una encuesta en su campus universitario para averiguar la demanda estudiantil de un producto, un suplemento proteínico para los batidos («Smoothies» en inglés). Encuestó a una muestra aleatoria de 113 estudiantes y obtuvo datos que podrían ser útiles para desarrollar su estrategia de marketing. Las respuestas a esta encuesta se encuentran en el fichero de datos **Smoothies**.

a) Ponga un ejemplo de una variable categórica con respuestas ordinales.
b) Ponga un ejemplo de una variable categórica con respuestas nominales.

2.2. Gráficos para describir variables categóricas

Las variables categóricas pueden describirse utilizando tablas de distribución de frecuencias y gráficos como gráficos de barras, gráficos de tarta y diagramas de Pareto. Estos gráficos son utilizados habitualmente por los directivos y los analistas de mercado para describir los datos procedentes de encuestas y de cuestionarios.

Distribución de frecuencias

Una **distribución de frecuencias** es una tabla utilizada para organizar datos. La columna de la izquierda (llamada clases o grupos) contiene todas las respuestas posibles sobre una variable estudiada. La columna de la derecha es una lista de las frecuencias o número de observaciones correspondientes a cada clase.

Tablas

Las clases que utilizamos para construir tablas de distribución de frecuencias de una variable categórica son sencillamente las respuestas posibles a la variable categórica.

EJEMPLO 2.1. Las principales empresas de Florida central en 2003 (gráficos de barras y de tarta)

¿Qué empresas ocuparon los primeros puestos en Florida central en 2003?

Solución

El *Orlando Sentinel* enumera anualmente las principales empresas de Florida central (véase la referencia bibliográfica 7). La Tabla 2.1 es una distribución de frecuencias de las cinco empresas que tenían el mayor número de asalariados en esta zona.

Tabla 2.1. Principales empresas de Florida central en 2003.

Empresa	Número de asalariados
Disney World	51.600
Florida Hospital	19.283
Publix Supermarkets Inc.	14.995
Wal-Mart Stores Ind.	14.995
Universal Orlando	12.000

Gráficos de barras y gráficos de tarta

Los gráficos de barras y los gráficos de tarta se utilizan normalmente para describir datos categóricos. Si nuestro objetivo es llamar la atención sobre la *frecuencia* de cada categoría, lo más probable es que tracemos un **gráfico de barras**. Si es hacer hincapié en la proporción de cada categoría, es probable que elijamos un **gráfico de tarta**. En un gráfico de barras, la altura de un rectángulo representa esta frecuencia. No es necesario que las barras se toquen. La Figura 2.1 es un gráfico de barras de los datos categóricos sobre las empresas de Florida central de la Tabla 2.1.

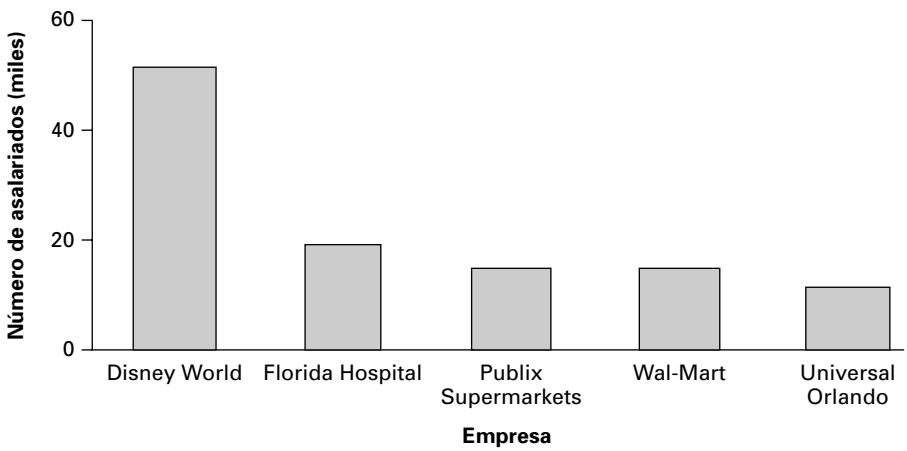


Figura 2.1. Cinco principales empresas de Florida central, 2003.

Tabla 2.2. Número de estudiantes matriculados en tres especialidades de administración de empresas, 2005 y 2005.

Especialidad	2000	2005
Finanzas	160	250
Marketing	140	200
Contabilidad	100	150

Cuando también interesan los componentes de las distintas categorías, puede utilizarse una interesante y útil extensión del gráfico de barras simple. Por ejemplo, la Tabla 2.2 muestra el número de estudiantes matriculados en tres especialidades de administración de empresas de una pequeña universidad privada en dos años distintos.

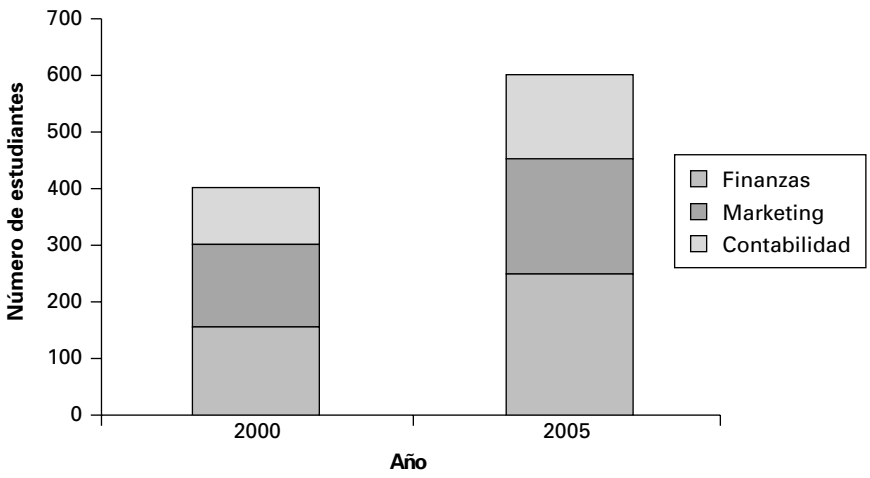


Figura 2.2A Estudiantes especializados en finanzas, marketing, 2000, 2005 (gráfico de barras por componentes).

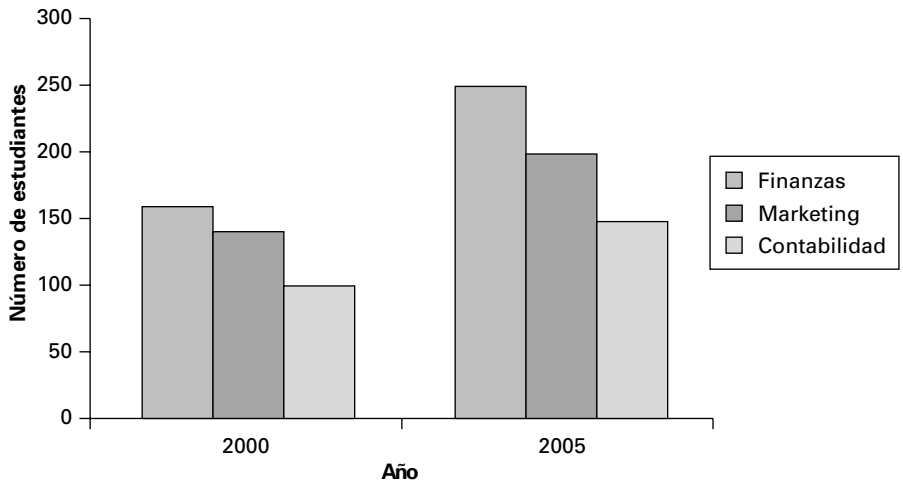


Figura 2.2B Estudiantes especializados en finanzas, marketing y contabiliad, 2000, 2005 (gráfico de barras por componentes).

Esta información puede mostrarse en un gráfico de barras desagregando el número total de estudiantes de cada año de manera que se distingan los tres componentes utilizando un sombreado diferente, como en la Figura 2.2A. Este tipo de gráfico se llama *gráfico de barras por componentes o apilado*. La Figura 2.2B muestra los mismos datos en un gráfico de barras que se denomina gráfico de barras *agrupado*. Los dos gráficos nos permiten hacer comparaciones visuales de totales y de componentes individuales. En este ejemplo, se observa que el aumento del número de matriculados que se registró entre 2000 y 2005 fue bastante uniforme en las tres especialidades.

Si queremos llamar la atención sobre la *proporción* de frecuencias en cada categoría, probablemente utilizaremos un gráfico de tarta para representar la división de un todo en sus partes integrantes. El círculo (o «tarta») representa el total y los segmentos (o «trozos de la tarta») que parten del centro representan proporciones de ese total. El gráfico de tarta se construye de tal forma que el área de cada segmento es proporcional a la frecuencia correspondiente.

EJEMPLO 2.2. Los gastos de viaje

El gerente de una universidad pidió una desagregación de los gastos de viaje de los profesores que asistían a diversas reuniones profesionales. Se observó que el 31 por ciento de los gastos estaba representado por los costes de transporte, el 25 por ciento por los costes de alojamiento, el 12 por ciento por los gastos de alimentación, el 20 por ciento por los gastos de matrícula y el resto por costes varios. Represente gráficamente estos datos.

Solución

La Figura 2.3 es un gráfico de tarta de los gastos de viaje.

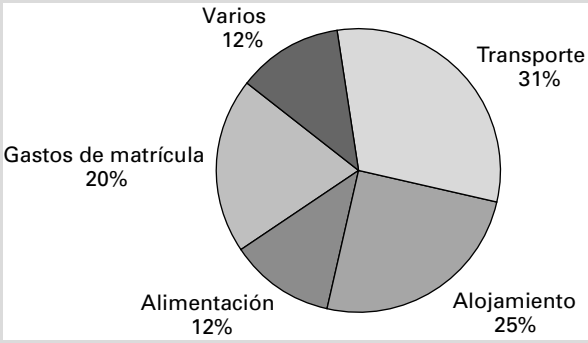


Figura 2.3. Producción diaria de cereales de trigo.

Diagramas de Pareto

Los directivos que necesitan identificar las principales causas de los problemas e intentar corregirlas rápidamente con un coste mínimo a menudo utilizan un gráfico de barras especial llamado *diagrama de Pareto*. El economista italiano Vilfredo Pareto (1848-1923) señaló que en la mayoría de los casos un pequeño número de factores es responsable de la mayoría de los problemas. Ordenamos las barras en un diagrama de Pareto de izquierda a derecha para poner énfasis en las causas más frecuentes de los defectos.

Diagrama de Pareto

Un **diagrama de Pareto** es un gráfico de barras que muestra la frecuencia de las causas de los defectos. La barra de la izquierda indica la causa más frecuente y las de la derecha indican las causas con frecuencias decrecientes. Los diagramas de Pareto se utilizan para separar lo «poco vital» de lo «mucho trivial».

El resultado de Pareto se aplica a una amplia variedad de conductas en muchos sistemas. A veces se denomina «regla del 80-20». Un fabricante de cereales puede observar que la mayoría de los errores de empaquetado se deben únicamente a unas cuantas causas. Un estudiante podría pensar que el 80 por ciento del trabajo de un proyecto de grupo ha sido realizado únicamente por el 20 por ciento de los miembros del equipo. La utilización de un

diagrama de Pareto también puede mejorar la comunicación con los empleados o con la dirección y dentro de los equipos de producción. El ejemplo 2.3 ilustra el principio de Pareto aplicado a un problema de una compañía de seguros médicos.

EJEMPLO 2.3. Errores de tramitación de las reclamaciones a un seguro (diagrama de Pareto)

Insurance

El análisis y el pago de las reclamaciones a un seguro es un complejo proceso que puede llevar a tramitar incorrectamente algunas reclamaciones. Estos errores provocan un aumento del tiempo que dedica el personal a obtener la información correcta y posiblemente a pagar indemnizaciones indebidas. El beneficiario normalmente detecta los errores cuando cobra una indemnización menor de la debida y a menudo puede pasar por alto las indemnizaciones superiores a las debidas. Estos errores pueden incrementar considerablemente los costes, además de afectar negativamente a las relaciones con los clientes. Se realizan considerables esfuerzos para analizar la actividad de presentación y de tramitación de las reclamaciones con el fin de poder desarrollar métodos para reducir lo más posible los errores. Una importante compañía de seguros médicos se fijó el objetivo de reducir un 50 por ciento los errores. Muestre cómo utilizaría el análisis de Pareto para ayudarla a averiguar los factores importantes que contribuyen a eliminar los errores. Los datos se encuentran en el fichero de datos **Insurance**.

Solución

La compañía de seguros médicos realizó una intensa investigación de todo el proceso de presentación de reclamaciones y pago de indemnizaciones. Se seleccionó un equipo de personas clave de los departamentos encargados de tramitar las reclamaciones, de relaciones con los proveedores y de marketing, de auditoría interna, de procesamiento de datos y de revisiones médicas. Basándose en su experiencia y en una revisión del proceso, los miembros del equipo llegaron finalmente a un acuerdo sobre una lista de posibles errores. Tres de ellos (códigos de procedimiento y diagnóstico, información de los proveedores e información de los pacientes) están relacionados con el proceso de presentación de reclamaciones y deben comprobarse revisando los historiales médicos de los pacientes en las clínicas y los hospitales. Tres posibles errores (tablas de precios, solicitudes de contratos y ajustes de los proveedores) están relacionados con la tramitación de las reclamaciones de indemnización dentro de la oficina de la compañía de seguros. Los errores de los programas y de los sistemas están incluidos en la categoría «Otros».

Se puso en marcha una auditoría completa de una muestra aleatoria de 1.000 reclamaciones contrastando cada reclamación con los historiales médicos de las clínicas y los hospitales hasta llegar a la fase final del pago de la indemnización. Se separaron las reclamaciones que contenían errores y se anotó el número de errores de cada tipo. Si una reclamación tenía múltiples errores, se anotaron todos. En este proceso, se tomaron muchas decisiones sobre la definición de error. Si se había dado a un niño un tratamiento que se daba normalmente a los adultos y el sistema informático de procesamiento no lo detectó, este error debía registrarse como un error 7 (errores de los programas y de los sistemas) y también como un error 3 (información de los pacientes). Si el tratamiento de un esguince estaba codificado como una fractura, debía registrarse como un error 1 (códigos de procedimientos y diagnósticos). La Tabla 2.3 es una distribución de frecuencias de las categorías y el número de errores cometidos en cada categoría.

A continuación, el equipo construyó el diagrama de Pareto de la Figura 2.4.

Tabla 2.3. Errores en la tramitación de las reclamaciones al seguro médico.

Categoría	Tipo de error	Frecuencia
1	Códigos de procedimientos y diagnósticos	40
2	Información del proveedor	9
3	Información del paciente	6
4	Tablas de precios	17
5	Solicitudes de contratos	37
6	Ajustes de los proveedores	7
7	Otros	4

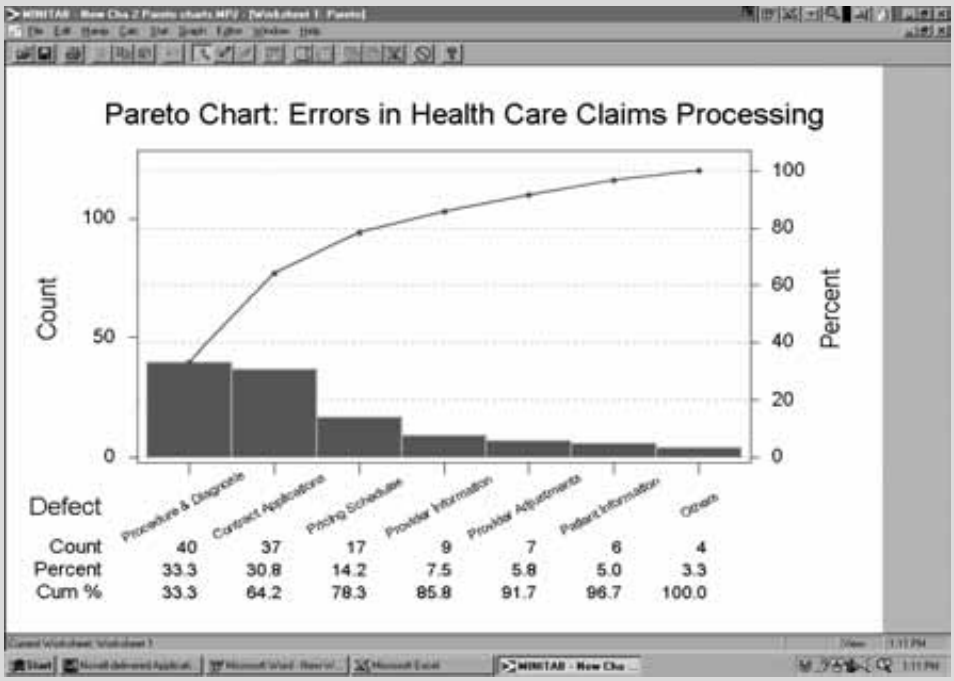


Figura 2.4. Diagrama de Pareto: errores en la tramitación de las reclamaciones al seguro médico.

Vemos en la Figura 2.4 que, cuando se van sumando los porcentajes de defectos correspondientes a los tipos de error (de izquierda a derecha), el ascenso de la línea de frecuencias acumuladas indica la mejora relativa que se obtendría corrigiendo cada uno de los problemas más frecuentes. En el diagrama de Pareto, los analistas vieron que el error 1 (códigos de procedimientos y diagnósticos) y el error 5 (solicitudes de contratos) eran las principales causas de los errores. La combinación de los errores 1, 5 y 4 (tablas de precios) provocaba casi un 80 por ciento de los errores. Examinando el diagrama de Pareto de la Figura 2.4, los analistas pueden averiguar rápidamente a qué causas debe dedicarse la mayor parte de los esfuerzos para corregir los problemas. El análisis de Pareto separó las «pocas causas vitales» de las «muchas triviales».

Pertrechado con esta información, el equipo hizo una serie de recomendaciones para reducir los errores y controlar el proceso.

1. Se harían sesiones especiales de formación para los encargados de tramitar las reclamaciones de los hospitales y las clínicas.
2. Se harían auditorías aleatorias por sorpresa para verificar los errores de codificación.
3. Se evaluaría la posibilidad de imponer sanciones monetarias a las organizaciones que cometieran excesivos errores.
4. Dos personas prepararían cada una por separado el conjunto completo de tablas de solicitud de contrato. A continuación, se compararían todas las entradas de las tablas utilizando un programa informático y se resolverían las diferencias que hubiera.
5. Se prepararían unos modelos-tipo de reclamación que se utilizarían para verificar las solicitudes correctas de contrato.

El diagrama de Pareto y las recomendaciones ayudaron a reducir los errores. Se redujeron los casos en los que se pagaban indemnizaciones de más, así como la burocracia necesaria para corregir los errores.

EJERCICIOS

Ejercicios básicos

2.9. Los gastos de viaje de una empresa son:

Concepto	Porcentaje
Compañías aéreas	41
Alojamiento	25
Comidas	12
Alquileres de automóviles	18
Otros	4

a) Construya un gráfico de tarta.
b) Construya un gráfico de barras.

2.10. Una empresa ha llegado a la conclusión de que hay siete defectos posibles en una de sus líneas de productos. Construya un diagrama de Pareto de las siguientes frecuencias de defectos:

Código de los defectos	Frecuencia
A	10
B	70
C	15
D	90
E	8
F	4
G	3

2.11. Se ha pedido a los empleados que indiquen su grado de satisfacción con el seguro médico actual. Éstas son las respuestas de una muestra aleatoria de empleados:

Muy satisfecho	29
Moderadamente satisfecho	55
Ninguna opinión	5
Moderadamente insatisfecho	20
Muy insatisfecho	9

a) Trace un gráfico de barras.
b) Trace un gráfico de tarta.

2.12. El supervisor de una planta ha obtenido una muestra aleatoria de las edades de los empleados y del tiempo que tardan en realizar una tarea (en segundos). Represente los datos con un gráfico de barras por componentes.

Edad/Tiempo	Menos de 40 segundos	Entre 40 y menos de 60 segundos	Un minuto como mínimo
Menos de 21	10	13	25
21 < 35	16	20	12
35 < 50	18	22	8
50 años o más	10	27	19

Ejercicios aplicados

2.13. Suponga que, según una estimación del gasto público, el 46 por ciento se destina a pensiones, el 18 por ciento a defensa, el 15 por ciento a regiones y municipios, el 14 por ciento a intereses de la deuda, el 6 por ciento a otros gastos de la administración central y el 1 por ciento al seguro de depósitos. Represente gráficamente esta información mediante un gráfico de tarta.

2.14. La tabla adjunta muestra una lista parcial del número de especies salvajes en peligro de extinción tanto dentro como fuera de Estados Unidos en abril de 2004 (véase la referencia bibliográfica 4):

Especie	Especies salvajes en peligro de extinción en EE.UU.	Especies salvajes en peligro de extinción en otros países
Mamíferos	69	251
Aves	77	175
Reptiles	14	64
Anfibios	12	8
Peces	71	11

FUENTE: U.S. Fish and Wildlife Service.

a) Construya un gráfico de barras del número de especies salvajes en peligro de extinción en Estados Unidos.
b) Construya un gráfico de barras del número de especies salvajes en peligro de extinción fuera de Estados Unidos.
c) Construya un gráfico de barras para comparar el número de especies salvajes en peligro de extinción en Estados Unidos y el de especies salvajes en extinción fuera de Estados Unidos.

2.15. Jon Payne, entrenador de tenis, registró del tipo de errores más grave que cometió cada uno de sus jugadores en un programa de formación de una semana. Los datos se encuentran en el fichero de datos **Tennis**.

a) Construya un diagrama de Pareto de los errores totales cometidos por todos los tenistas.
b) Construya un diagrama de Pareto de los errores totales cometidos por los tenistas masculinos.
c) Construya un diagrama de Pareto de los errores totales cometidos por los tenistas femeninos.
d) Construya un gráfico de barras por componentes que muestre el tipo de error y el sexo del tenista.

2.16. ¿A qué tipo de actividad de Internet dedica usted la mayor parte del tiempo? Las respuestas de una muestra aleatoria de 700 usuarios de Internet fueron las siguientes: realizar operaciones de banca electrónica, 40; comprar un producto, 60; obtener noticias, 150; enviar o leer correo electrónico, 200; comprar o realizar una reserva para viajar, 75; enterarse de los resultados de partidos o de información deportiva, 50; y buscar la respuesta a una pregunta, 125. Describa los datos gráficamente.

2.17. Un grupo de estudiantes de administración de empresas de una universidad decidió adquirir experiencia en la gestión de una empresa montando una para vender batidos («Smoothies») en el campus universitario. Realizaron una encuesta a una muestra aleatoria de 113 estudiantes para obtener datos que ayudaran a desarrollar su estrategia de marketing. Una de las preguntas de la encuesta les pedía que indicaran su propio nivel de concienciación sobre su estado de salud. Las respuestas a esta encuesta se encuentran en el fichero de datos **Smoothies**.

a) Trace un gráfico de barras.
b) Trace un gráfico de tarta.

2.18. Construya a partir del fichero de datos **Smoothies** gráficos de barras por componentes de las respuestas correspondientes a las siguientes variables:

a) Sexo y nivel de concienciación sobre el estado de salud.
b) Deseo de un suplemento proteínico y nivel de preocupación por el estado de salud.

2.19. El *Statistical Abstract of the United States* (véase la referencia bibliográfica 6) contiene datos sobre las exportaciones y las importaciones de Estados Unidos y sobre su balanza comercial de mercancías por países.

a) Represente gráficamente los 10 principales compradores de exportaciones de Estados Unidos en el año más reciente del que se dispone.
b) Represente gráficamente los 10 principales proveedores de importaciones de Estados Unidos en el año más reciente del que se dispone.

2.3. Gráficos para describir datos de series temporales

Supongamos que tomamos una muestra aleatoria de 100 cajas de una nueva variedad de cereales. Si recogemos nuestra muestra en un momento del tiempo y ponderamos cada caja, las mediciones obtenidas se conocen con el nombre de datos de *corte transversal*. Sin embargo, podríamos recoger y medir una muestra aleatoria de 5 cajas cada 15 minutos o de 10 cajas cada 20 minutos. Los datos medidos en sucesivos momentos del tiempo se denominan datos de *series temporales*. En el Capítulo 19 estudiaremos en mayor profundidad este tipo de datos. Pero de momento examinaremos un gráfico de datos de series temporales llamado *gráfico de series temporales*.

Gráfico de series temporales

Un **gráfico de series temporales** representa una serie de datos en varios intervalos de tiempo. Midiendo el tiempo en el eje de abscisas y la cantidad numérica que interesa en el de ordenadas se obtiene un punto en el gráfico por cada observación. Uniendo los puntos contiguos en el tiempo por medio de líneas rectas se obtiene un gráfico de series temporales.

La tecnología del siglo XXI permite acceder rápidamente a datos que pueden ayudar a tomar decisiones y muchos de estos datos son de series temporales. El comercio electrónico es importante para todos nosotros. Se puede comprar casi todo: billetes de avión, automóviles, electrónica, libros, flores, acciones, etc. Los minoristas del país notifican a las autoridades cuánto negocio hacen en línea y esta información se utiliza en los informes oficiales mensuales sobre la situación de la economía. Estos datos se recogen a intervalos sucesivos de tiempo.

Numerosas empresas analizan y venden encuestas y datos estadísticos por Internet. Para desarrollar planes de marketing, muchas empresas necesitan las características demográficas de los compradores por Internet, así como del resto de los compradores. Muchas veces las observaciones se miden a sucesivos intervalos de tiempo (anual, mensual o semanalmente, por horas, etc.). Las universidades estudian la evolución de las cifras de matriculados para comprender mejor sus tendencias. Los médicos controlan semanal o mensualmente los análisis de sangre de los pacientes de cáncer. Para describir gráficamente todos estos ejemplos se utiliza un gráfico de series temporales.

EJEMPLO 2.4. Tendencias del número de matriculados en una universidad (gráfico de series temporales)

University Enrollments, 1995-2005

El rector de una pequeña universidad privada solicitó datos sobre el número de estudiantes de primer año y sobre el número de estudiantes procedentes de otras universidades que entraron en la universidad entre 1995 y 2005. Los datos se encuentran en el fichero de datos **University Enrollments, 1995-2005**.

Solución

En la Figura 2.5 podemos ver que el número de matriculados de primer año ha aumentado desde 2000 y que el máximo que alcanzó el número de estudiantes procedentes de otras universidades en 2002 fue seguido de un continuo descenso. El personal de admisiones debe averiguar cuáles son los factores que explican ambas tendencias.

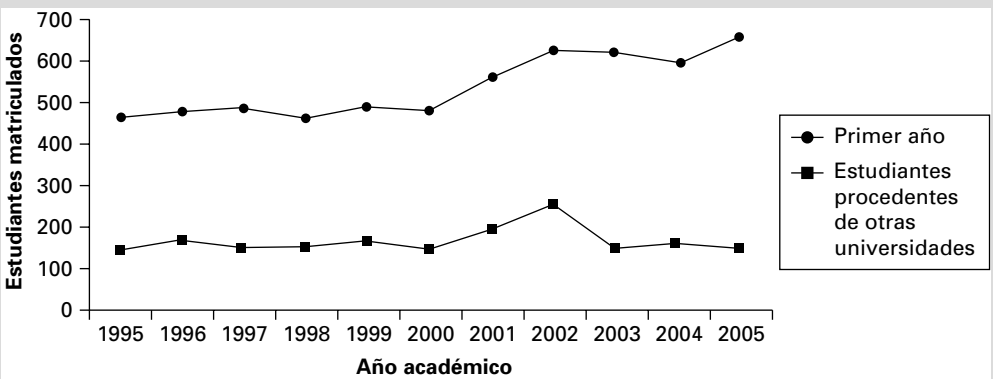


Figura 2.5. Estudiantes matriculados por primera vez, 1995-2005.

EJEMPLO 2.5. Ventas trimestrales de una empresa durante seis años (gráfico de series temporales)

Quarterly Sales 2001-2006

La Tabla 2.4 muestra las ventas trimestrales realizadas por una empresa durante un periodo de 6 años que se encuentran en el fichero de datos **Quarterly Sales 2001-2006**. Describa los datos gráficamente.

Solución

La Figura 2.6 es un gráfico de series temporales de los 24 intervalos de tiempo. Observamos que las ventas del primer trimestre van seguidas sistemáticamente de una disminución de las ventas en el segundo. Tal vez la estación del año sea una explicación. En el Capítulo 19 presentaremos métodos para ajustar los datos de series temporales con el fin de tener en cuenta la estacionalidad, las tendencias, la conducta cíclica o algún otro componente irregular.

Tabla 2.4 Ventas trimestrales, 2001-2006 (miles).

	Trimestre			
Año	1	2	3	4
2001	271	199	240	255
2002	341	246	245	275
2003	351	283	353	292
2004	401	282	306	291
2005	370	242	281	274
2006	356	245	304	279

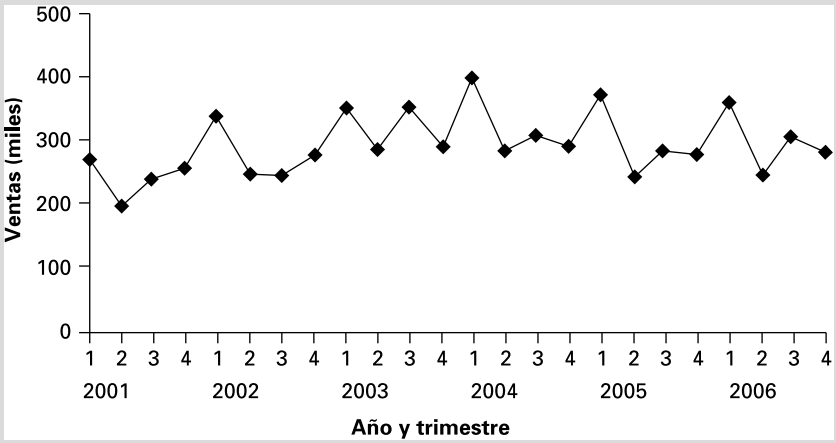


Figura 2.6. Ventas trimestrales, 2001-2006 (gráfico de series temporales).

Si sólo nos interesa comparar las ventas del primer trimestre con las del segundo, puede ser interesante un gráfico de series temporales como el de la Figura 2.7.

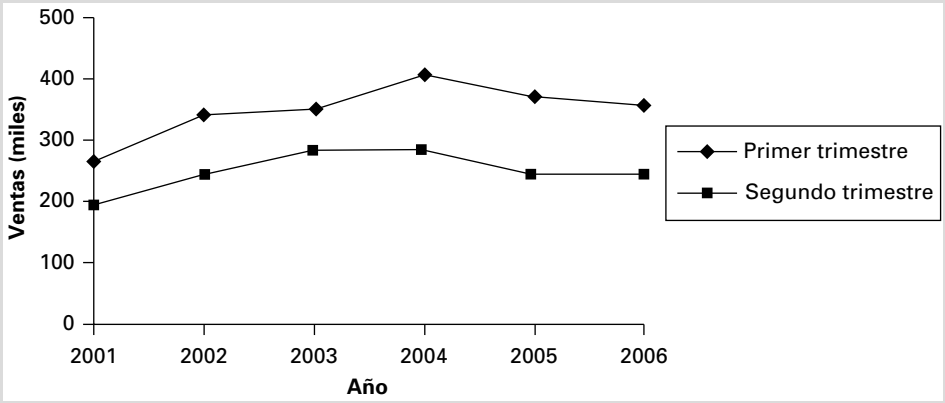


Figura 2.7. Ventas del primero y del segundo trimestre, 2001-2006 (gráfico de series temporales).

EJERCICIOS

Ejercicios básicos

2.20. Construya un gráfico de series temporales de los siguientes datos sobre el uso del teléfono móvil durante el fin de semana o por la noche (en minutos):

Mes	Fin de semana o noche
Enero	575
Febrero	603
Marzo	469
Abril	500
Mayo	586
Junio	540

2.21. ¿Qué porcentaje de antiguos alumnos hizo donaciones a su universidad? La tabla adjunta muestra los porcentajes que obtuvo una universidad en el periodo 2001-2005. Trace un gráfico de series temporales de los datos. ¿Qué medida podría tomar la universidad?

Año	Porcentaje
2001	26,72
2002	27,48
2003	24,89
2004	25,83
2005	30,22

Ejercicios aplicados

2.22. El fichero de datos **Degrees 1998-2005** contiene los títulos concedidos entre 1998 y 2005 por tipo de título en una universidad privada.

a) Represente los datos en un gráfico de series temporales.

b) ¿Qué conclusiones o qué medidas podría considera la universidad?

2.23. El *Statistical Abstract of the United States* (Section 21: Manufactures) (véase la referencia bibliográfica 5) contiene información sobre el producto interior bruto de la industria manufacturera. El U.S. Census Bureau, el *Annual Survey of Manufacturers* y el *Current Industrials Report* son las principales fuentes de esta información.

a) Utilice un gráfico de series temporales para representar el producto interior bruto de la industria manufacturera en dólares corrientes por industrias de bienes duraderos (como productos de madera, muebles y productos afines, vehículos de motor y equipo) entre 1998 y 2003.

b) Utilice un gráfico de series temporales para representar el producto interior bruto de la industria manufacturera en dólares encadenados (2000) por industrias de bienes no duraderos (como alimentación, ropa y productos de cuero) entre 1998 y 2003.

2.24. El fichero de datos **Exchange Rate** muestra un índice del valor del dólar frente a las monedas

de sus socios comerciales durante 12 meses consecutivos. Represente los datos mediante un gráfico de series temporales.

2.25. El fichero de datos **Inventory Sales** contiene el cociente entre las existencias y las ventas de la industria manufacturera y el comercio de Estados Unidos en un periodo de 12 años. Represente los datos con un gráfico de series temporales.

2.26. Seleccione en Internet los rendimientos anuales de un índice de cotizaciones bursátiles durante 14 años. Represente los datos con un gráfico de series temporales.

2.27. El fichero de datos **Gold Price** muestra el precio del oro (en dólares) a final de año durante 14 años consecutivos. Represente los datos con un gráfico de series temporales.

2.28. El fichero de datos **Housing Starts** muestra las viviendas privadas iniciadas por mil habitantes de la población de Estados Unidos en un periodo de 24 años. Describa los datos con un gráfico.

2.29. El fichero de datos **Earnings per share** contiene los beneficios por acción de una empresa en un periodo de 28 años. Represente gráficamente la serie y coméntela en el gráfico.

2.4. Gráficos para describir variables numéricas

En este apartado presentamos brevemente histogramas, ojivas y diagramas de tallo y hojas que resumen y describen datos numéricos. Primero examinamos una distribución de frecuencias de datos numéricos.

Distribuciones de frecuencias

Una distribución de frecuencias de datos numéricos es, al igual que una distribución de frecuencias de datos categóricos (apartado 2.2), una tabla que resume datos enumerando las clases en la columna de la izquierda y el número de observaciones de cada clase en la columna de la derecha. Sin embargo, en una distribución de frecuencias de datos numéricos las clases o intervalos no son fácilmente identificables.

Para decidir los intervalos de una distribución de frecuencias de datos numéricos es necesario responder a ciertas preguntas: ¿cuántos intervalos deben utilizarse? ¿De qué amplitud debe ser cada intervalo? Hay algunas reglas generales (como las ecuaciones 2.1 y 2.2) para preparar distribuciones de frecuencias que nos permitan responder más fácilmente a este tipo de cuestiones, para resumir datos y para comunicar los resultados.

Construcción de una distribución de frecuencias

Regla 1: Decidir k, el número de intervalos (clases).

Regla 2: Los intervalos (clases) deben ser de la misma amplitud, w; la amplitud viene determinada por lo siguiente:

$$w = \text{Amplitud de los intervalos} = \frac{(\text{Número mayor} - \text{Número menor})}{\text{Número de intervalos}} \qquad \textbf{(2.1)}$$

Tanto k como w deben redondearse al alza, posiblemente al siguiente número entero mayor.

Regla 3: Los intervalos (clases) deben ser inclusivos y no solaparse.

Regla 1. Número de intervalos

El número de intervalos (clases) utilizados en una distribución de frecuencias se decide de una manera algo arbitraria.

Guía rápida para decidir un número aproximado de intervalos de una distribución de frecuencias

Tamaño de la muestra	Número de intervalos
Menos de 50	5-7
De 50 a 100	7-8
De 101 a 500	8-10
De 501 a 1.000	10-11
De 1.001 a 5.000	11-14
Más de 5.000	14-20

(2.2)

La práctica y la experiencia son la mejor guía. Los conjuntos de datos mayores requieren más intervalos; los conjuntos de datos menores requieren menos intervalos. Si seleccionamos excesivamente pocas clases, las pautas y algunas características de los datos pueden quedar ocultas. Si seleccionamos demasiadas clases, descubriremos que algunos intervalos no contienen ninguna observación o tienen una frecuencia muy pequeña.

Regla 2. Amplitud de los intervalos

Después de elegir el número de intervalos, el paso siguiente es elegir la amplitud de los intervalos:

$$w = \text{Amplitud de los intervalos} = \frac{(\text{Número mayor} - \text{Número menor})}{\text{Número de intervalos}}$$

La amplitud de los intervalos a menudo se redondea a un número entero para facilitar la interpretación.

Regla 3. Intervalos inclusivos y que no se solapen

Los intervalos deben ser inclusivos y no solaparse. Cada observación debe pertenecer a uno y sólo un intervalo. Consideremos una distribución de frecuencias de las edades (redondeadas al año más próximo) de un grupo de personas. Si la distribución de frecuencias contiene los intervalos «20-30 años» y «30-40 años», ¿a cuál de estas dos clases pertenecería una persona de 30 años?

Los *límites* o extremos de cada clase deben estar claramente definidos. Para evitar solapamientos, los intervalos de edades podrían definirse de la forma siguiente: «20 años *pero menos de* 30», seguido de «30 años *pero menos de* 40», y así sucesivamente. Otra posibilidad es definir los intervalos de edad del modo siguiente: «20-29», «30-39», etc. Dado que la edad es un número entero, no hay ningún solapamiento. La selección de los límites es subjetiva. Hay que asegurarse simplemente de definir unos límites que permitan comprender e interpretar claramente los datos.

No debemos hacer excesivo hincapié en las reglas para determinar el número de intervalos y su amplitud o hacer demasiado poco hincapié en la selección del número de clases que muestren las pautas de los datos más claras.

Dos distribuciones de frecuencias especiales son la *distribución de frecuencias acumuladas* y la *distribución de frecuencias relativas acumuladas*.

Distribuciones de frecuencias relativas, acumuladas y relativas acumuladas

Se obtiene una **distribución de frecuencias relativas** dividiendo cada frecuencia por el número de observaciones y multiplicando la proporción resultante por 100 por ciento. Una **distribución de frecuencias acumuladas** contiene el número total de observaciones cuyos valores son menores que el límite superior de cada intervalo. Se construye sumando las frecuencias de todos los intervalos de la distribución de frecuencias e incluyendo el presente intervalo. En una **distribución de frecuencias relativas acumuladas,** las frecuencias acumuladas pueden expresarse en proporciones o porcentajes acumulados.

EJEMPLO 2.6. El uso del teléfono móvil (pensar en términos estadísticos)

Jennie Bishop, directora de marketing de una importante compañía de telefonía móvil, obtuvo los registros de los minutos consumidos por una muestra aleatoria de 110 abonados al plan más barato de la empresa (250 minutos mensuales como máximo en hora punta). La Tabla 2.5 contiene una lista de los minutos consumidos por cada abonado de la muestra durante un mes. Los datos se encuentran en el fichero de datos **Mobile Usage**. ¿Qué indican los datos?

271	236	294	252	254	263	266	222	262	278	288
262	237	247	282	224	263	267	254	271	278	263
262	288	247	252	264	263	247	225	281	279	238
252	242	248	263	255	294	268	255	272	271	291
263	242	288	252	226	263	269	227	273	281	267
263	244	249	252	256	263	252	261	245	252	294
288	245	251	269	256	264	252	232	275	284	252
263	274	252	252	256	254	269	234	285	275	263
263	246	294	252	231	265	269	235	275	288	294
263	247	252	269	261	266	269	236	276	248	298

Solución

La Tabla 2.5 en sí misma no sirve de mucho a la directora de marketing para desarrollar una estrategia de marketing. Podemos encontrar alguna información en esa tabla: la cantidad mínima de minutos consumidos en hora punta fue de 222 y el tiempo máximo consumido fue de 298. Sin embargo, necesitamos más información que ésta antes de presentar un informe a los altos ejecutivos. Para comprender mejor lo que indican los datos de la Tabla 2.5, primero desarrollamos una distribución de frecuencias.

Basándonos en la guía rápida, desarrollamos una distribución de frecuencias con ocho clases para los datos de la Tabla 2.5. Según la ecuación 2.1, la amplitud de cada clase es

$$w = \frac{299 - 222}{8} = 10 \text{ (redondeando)}$$

Dado que el valor más bajo es 222, el primer intervalo podría ser «220 pero menos que 230». A continuación, se van añadiendo intervalos de igual amplitud a la distribución de frecuencias, así como el número de minutos que pertenecen a cada clase. La Tabla 2.6 es una distribución de frecuencias correspondiente a los datos de la Tabla 2.5 sobre el uso de los teléfonos móviles.

Tabla 2.6. Distribuciones de frecuencia y de frecuencias relativas del uso del teléfono móvil

Uso del teléfono móvil (en minutos)	Frecuencia	Porcentaje
220 menos de 230	5	4,5
230 menos de 240	8	7,3
240 menos de 250	13	11,8
250 menos de 260	22	20,0
260 menos de 270	32	29,1
270 menos de 280	13	11,8
280 menos de 290	10	9,1
290 menos de 300	7	6,4

El director puede querer saber cuál es el uso del teléfono móvil por debajo (o por encima) de una cierta cantidad de tiempo. La Tabla 2.7 contiene una distribución de frecuencias acumuladas y una distribución de porcentajes acumulados.

Las distribuciones de frecuencias de las Tablas 2.6 y 2.7 son una mejora con respecto a la lista inicial de datos de la 2.5. Hemos resumido al menos 110 observaciones en 8 categorías y podemos decir a Jennie que durante el mes estudiado menos de una cuarta parte (el 23,6 por ciento) de los abonados de la muestra utilizó el teléfono móvil respetando los límites de sus planes. La directora de marketing podría sugerir que se pusiera en marcha una campaña publicitaria para promover un plan que conllevara un aumento de los minutos en hora punta.

Tabla 2.7. Distribuciones de frecuencias acumuladas y de frecuencias relativas acumuladas del uso del teléfono móvil

Uso del teléfono móvil (en minutos)	Frecuencia	Porcentaje
Menos de 230	5	4,5
Menos de 240	13	11,8
Menos de 250	26	23,6
Menos de 260	48	43,6
Menos de 270	80	72,7
Menos de 280	93	84,5
Menos de 290	103	93,6
Menos de 300	110	100,0

Histogramas y ojivas

Una vez desarrolladas las distribuciones de frecuencias, podemos representar gráficamente esta información. Analizaremos brevemente los *histogramas* y las *ojivas*.

Histograma

Un **histograma** es un gráfico formado por barras verticales construidas sobre una línea recta horizontal delimitada por los intervalos de la variable mostrada. Los intervalos corresponden a los de una tabla de distribución de frecuencias. La altura de cada barra es proporcional al número de observaciones que hay en ese intervalo. El número de observaciones puede indicarse encima de las barras.

Ojiva

Una **ojiva**, llamada a veces *gráfico de frecuencias acumuladas*, es una línea que conecta puntos que son el porcentaje acumulado de observaciones situadas por debajo del límite superior de cada intervalo en una distribución de frecuencias acumuladas.

La Figura 2.8 es un histograma de las frecuencias de uso del teléfono móvil de la Tabla 2.6. La 2.9 es una ojiva que describe las frecuencias relativas acumuladas de la Tabla 2.7.

La forma de un histograma revela si los datos están repartidos de una manera uniforme a un lado y a otro del punto medio del gráfico. Es decir, en algunos histogramas veremos que la mitad o el centro del gráfico los divide en dos «imágenes gemelas», de manera que la parte de uno de los lados es casi idéntica a la del otro. Los histogramas que tienen esta forma son *simétricos*; los que no la tienen son asimétricos o *sesgados*.

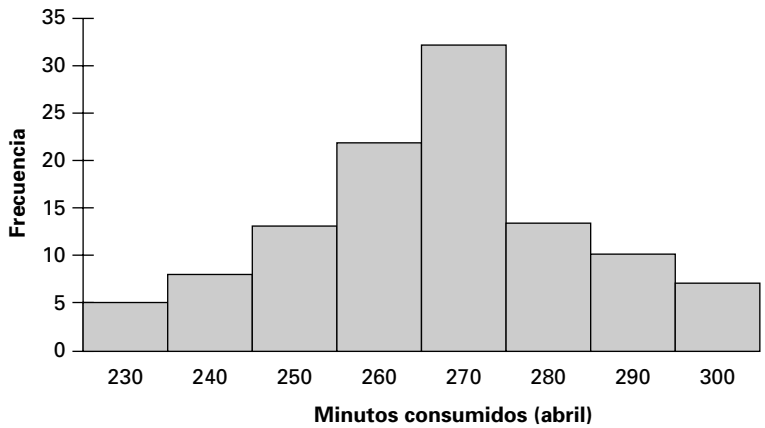


Figura 2.8.
Uso del teléfono móvil (histograma).

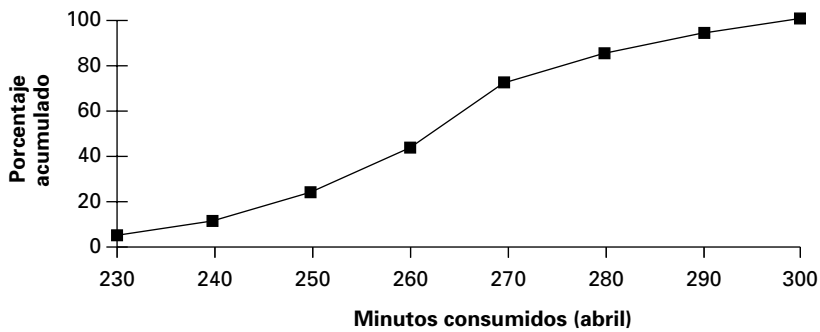


Figura 2.9.
Uso del teléfono móvil (ojiva).

Simetría

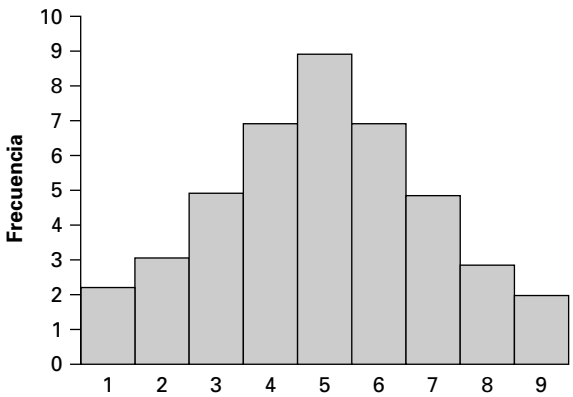
Se dice que la forma de un histograma es **simétrica** si las observaciones están equilibradas, es decir, distribuidas de una manera uniforme a un lado y a otro del punto medio del histograma.

Sesgo

Una distribución está **sesgada** o es asimétrica si las observaciones no están distribuidas simétricamente en ninguno de los lados de la mitad. Una distribución *sesgada positivamente* tiene una cola que se extiende hacia la derecha, en la dirección de los valores positivos. Una distribución *sesgada negativamente* tiene una cola que se extiende hacia la izquierda, en la dirección de los valores negativos.

La Figura 2.10A representa un histograma simétrico. En cambio, el de la 2.10B tiene una larga cola hacia la derecha, con una disminución mucho más brusca hacia la izquierda. Esta distribución está *sesgada hacia la derecha*, es decir, está *sesgada positivamente*. La distribución de la Figura 2.10C está sesgada negativamente: las observaciones más bajas están más extendidas hacia la izquierda. En el Capítulo 3 aprenderemos más sobre los factores que contribuyen al sesgo y veremos cómo se mide éste numéricamente.

Figura 2.10A Distribución simétrica.



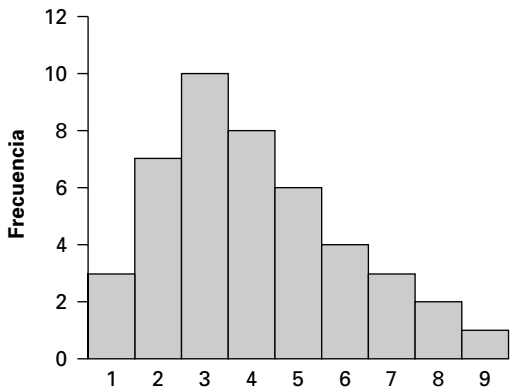


Figura 2.10B. Distribución sesgada positivamente.

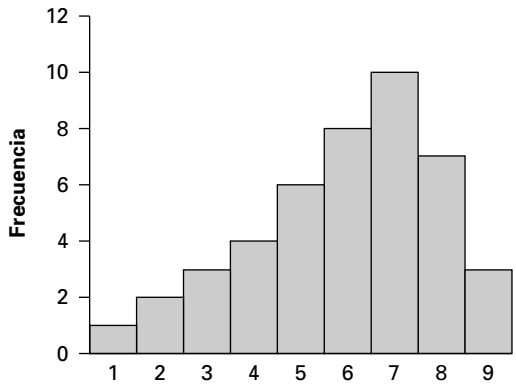


Figura 2.10C. Distribución sesgada negativamente.

Aunque los histogramas pueden permitirnos conocer la forma de la distribución, es importante recordar que pueden no ser «matemáticamente correctos», ya que a menudo su escala vertical no puede ajustarse. En el apartado 2.6 hacemos algunas advertencias sobre los histogramas que distorsionan la verdad.

Diagramas de tallo y hojas

El análisis exploratorio de datos (AED) consiste en los métodos utilizados para describir los datos en sencillos términos aritméticos con imágenes fáciles de trazar con lápiz y papel (véase la referencia bibliográfica 11). Uno de esos métodos es el *diagrama de tallo y hojas*. Antes de que existieran los computadores, este método permitía identificar rápidamente las pautas posibles en pequeños conjuntos de datos. Aquí sólo lo analizamos brevemente.

Diagrama de tallo y hojas

Un **diagrama de tallo y hojas** es un gráfico AED que es una alternativa al histograma. Los datos se agrupan de acuerdo con sus primeros dígitos (llamados tallo) y se hace un listado de los últimos dígitos (llamados hojas) de cada miembro de una clase. Las hojas se muestran individualmente en orden ascendente después de cada uno de los tallos.

El número de dígitos de cada clase indica la frecuencia de clase. Los dígitos individuales indican la pauta de valores dentro de cada clase. Salvo los casos *atípicos* extremos (los valores de datos que son mucho mayores o menores que otros valores del conjunto de datos), se incluyen todos los tallos aunque no haya observaciones en el subconjunto correspondiente. El número de dígitos del tallo depende del conjunto de datos.

Accounting GPAs

EJEMPLO 2.7. Calificaciones medias (diagrama de tallo y hojas)

Se han obtenido las calificaciones medias (GPA, por sus iniciales en inglés) en la especialización de contabilidad de una muestra aleatoria de estudiantes que acaban de terminar los estudios. ¿Qué información suministra el diagrama de tallo y hojas de la Figura 2.11? Los datos se encuentran en el fichero de datos **Accounting GPAs**.

Solución

La calificación media de cada estudiante se ha redondeado a su valor entero más próximo. La Figura 2.11 muestra la salida Minitab (con este programa pueden obtenerse

Frecuencia acumulada	Tallo	Hoja
1	21	2
3	22	2 9
7	23	3 4 5 9
13	24	0 1 3 4 7 9
19	25	1 2 3 5 5 7
24	26	1 1 1 2 6
30	27	1 2 3 5 6 8
40	28	0 2 3 4 4 4 5 6 9 9
51	29	0 1 2 2 4 4 4 5 7 7 7
(10)	30	1 1 1 2 6 7 8 8 8 9
51	31	0 1 1 1 2 4 5 6 8
42	32	1 1 4 5 6 8 9
35	33	1 2 3 5 7 8 8 9
27	34	0 0 1 1 1 3 3 3 4 6
17	35	1 6 7 7
13	36	0 1 2 5 5 6 6 8 8
4	37	2 3
2	38	0 7

Figura 2.11. Diagrama de tallo y hojas de las calificaciones medias.

distintas versiones del diagrama de tallo y hojas). Podemos hacer varias observaciones a partir de la Figura 2.11. Por ejemplo, vemos que una calificación media de 3,25 se registra como un tallo de «32» y una hoja de «5». La más baja es 2,12 y la más alta 3,87. La columna situada más a la izquierda de la salida Minitab contiene las frecuencias acumuladas, separadas por un número entre paréntesis. En la Figura 2.11, el número 10 (entre paréntesis) nos dice que los datos están centrados en las calificaciones medias comprendidas entre 3,00 y 3,09. El número 40 de la columna situada más a la izquierda indica que 40 estudiantes obtuvieron una calificación media de menos de 2,90. El número 27 de la columna situada más a la izquierda nos dice que 27 estudiantes obtuvieron una calificación media de al menos 3,40.

EJERCICIOS

Ejercicios básicos

2.30. Utilice la guía rápida para hallar un número aproximado de clases de una distribución de frecuencias suponiendo que el tamaño de la muestra es:

a) $n = 47$
b) $n = 80$
c) $n = 150$
d) $n = 400$
e) $n = 650$

2.31. Halle la amplitud que deben tener los intervalos para una muestra aleatoria de 110 observaciones que se encuentran

a) entre 20 y 85 (inclusive)
b) entre 30 y 190 (inclusive)
c) entre 40 y 230 (inclusive)
d) entre 140 y 500 (inclusive)

2.32. Considere los datos siguientes:

17	62	15	65
28	51	24	65
39	41	35	15
39	32	36	37
40	21	44	37
59	13	44	56
12	54	64	59

a) Construya una distribución de frecuencias.
b) Trace un histograma.
c) Trace una ojiva.
d) Trace un diagrama de tallo y hojas.

2.33. Construya un diagrama de tallo y hojas de las horas que dedican 20 estudiantes a estudiar para un examen de marketing.

3,5 2,8 4,5 62, 4,8 2,3 2,6 3,9 4,4 5,5
5,2 6,7 3,0 2,4 5,0 3,6 2,9 1,0 2,8 3,6

2.34. Considere la siguiente distribución de frecuencias

Clase	Frecuencia
$0 < 10$	8
$10 < 20$	10
$20 < 30$	13
$30 < 40$	12
$40 < 50$	6

a) Construya una distribución de frecuencias relativas.
b) Construya una distribución de frecuencias acumuladas.
c) Construya una distribución de frecuencias relativas acumuladas.

Ejercicios aplicados

2.35. La tabla siguiente muestra la distribución por edades de los visitantes de páginas web de agencias de viajes durante diciembre de 2003 (véase la referencia bibliográfica 12):

Edad	Porcentaje
18-24	11,30
25-34	19,11
35-44	23,64
45-54	23,48
55+	22,48

a) Construya una distribución de frecuencias relativas acumuladas.
b) ¿Qué porcentaje de visitantes de Internet tenía menos de 45 años?
c) ¿Qué porcentaje de visitantes de Internet tenía al menos 35 años?

2.36. La demanda de agua embotellada aumenta durante la temporada de huracanes en Florida. El director de operaciones de una planta que embotella agua quiere estar seguro de que el proceso de embotellado de botellas de 1 galón está funcionando correctamente. Actualmente, la compañía está comprobando el volumen de las botellas de 1 galón. Se comprueba una muestra aleatoria de 75 botellas. Estudie el proceso de embotellado de este producto y presente un informe de sus resultados al director de operaciones. Construya una distribución de frecuencias, una distribución de frecuencias acumuladas, un histograma, una ojiva y un diagrama de tallo y hojas. Incorpore estos gráficos a un resumen bien redactado. ¿Cómo podríamos pensar en términos estadísticos en esta situación? Los datos se encuentran en el fichero de datos **Water**.

2.37. El fichero de datos llamado **Scores** contiene las puntuaciones obtenidas por 40 estudiantes en un test.

a) Construya una distribución de frecuencias de los datos.
b) Construya una distribución de frecuencias acumuladas de los datos.
c) Basándose en su respuesta al apartado a), construya un histograma adecuado de los datos.
d) Construya un diagrama de tallo y hojas de los datos.

2.38. El fichero de datos **Returns** contiene los rendimientos porcentuales obtenidos en un día específico por los fondos de inversión en acciones ordinarias de las 25 mayores empresas de Estados Unidos.
a) Construya un histograma para describir los datos.
b) Trace un diagrama de tallo y hojas para describir los datos.
c) Construya una ojiva para describir los datos.

2.39. Ann Thorne, la directora de operaciones de una fábrica de cremas bronceadoras, quiere asegurarse de que el proceso que se emplea para llenar los botes de 8 onzas (237 ml) de SunProtector está funcionando correctamente. Suponga que se selecciona una muestra aleatoria de 100 botes de esta crema, se miden los contenidos y se almacenan los volúmenes (en ml) en el fichero de datos **Sun**. Describa los datos gráficamente.

2.5. Tablas y gráficos para describir relaciones entre variables

En los apartados anteriores hemos desarrollado gráficos para describir una única variable. Estas «imágenes» nos han ayudado a analizar mejor la información que contenía un gran conjunto de datos. En este apartado, ampliamos las medidas gráficas para describir las relaciones entre dos variables. En primer lugar, presentamos un *diagrama de puntos dispersos* para estudiar las posibles relaciones entre dos variables cuantitativas. A continuación, analizamos *tablas cruzadas* de dos variables para examinar posibles relaciones entre variables cualitativas.

Los análisis empresariales y económicos a menudo se refieren a relaciones entre variables. ¿Obtienen mejores calificaciones medias en la universidad los alumnos que tienen mejores notas en el examen de selectividad? ¿Cuánto varía la cantidad vendida cuando varía el precio? ¿Cómo influye en las ventas totales la renta total disponible en una región geográfica? ¿Aumenta la publicidad las ventas? ¿Cómo varía la mortalidad infantil en los países en vías de desarrollo cuando aumenta la renta per cápita?

En estos ejemplos, observamos que una variable puede depender en alguna medida de la otra. Por ejemplo, la calificación media de un estudiante universitario puede depender de la nota que obtuvo en la prueba de matemáticas de la selectividad. En ese caso, llamamos a la calificación media *variable dependiente* y la representamos por medio de Y y a la

puntuación obtenida en la prueba de matemáticas de la selectividad *variable independiente* y la representamos por medio de X. Asimismo, llamaríamos Y a la cantidad vendida y X al precio de la mercancía.

Para responder a estas preguntas, reunimos y analizamos muestras aleatorias de datos recogidos en poblaciones relevantes. Nuestro análisis comienza con la construcción de un gráfico llamado diagrama de puntos dispersos.

Diagramas de puntos dispersos

Una imagen a menudo muestra la relación que puede existir entre dos variables.

Diagrama de puntos dispersos

Podemos trazar un **diagrama de puntos dispersos** localizando un punto por cada par de dos variables que representan una observación del conjunto de datos. El diagrama de puntos dispersos es una representación de los datos, que comprende lo siguiente:

1) El rango de cada variable.
2) La pauta de valores existente dentro del rango.
3) Una sugerencia sobre la posible relación entre las dos variables.
4) Una indicación de los casos atípicos (puntos extremos).

Podríamos trazar diagramas de puntos dispersos representando puntos en un papel milimetrado. Sin embargo, todos los paquetes estadísticos modernos contienen rutinas para realizar directamente diagramas de puntos dispersos a partir de un fichero de datos electrónico. Como se muestra en el ejemplo 2.8, la realización de un diagrama de ese tipo es una tarea habitual en cualquier análisis inicial de datos que se realiza al principio de un estudio económico o empresarial. En el ejemplo citado mostramos un diagrama de puntos dispersos de dos variables cuantitativas.

EJEMPLO 2.8. Las notas de los exámenes de admisión en las universidades en Estados Unidos y las calificaciones medias de los estudios universitarios (diagramas de puntos dispersos)

¿Son las notas obtenidas en la prueba de matemáticas del SAT para acceder a la universidad un buen indicador de éxito en la universidad? En Estados Unidos, todos los estudiantes realizan uno o más tests de aptitud para entrar en una universidad. El personal de admisiones de las universidades utiliza los resultados para admitir o no a los estudiantes. La Tabla 2.8 muestra las notas obtenidas en la prueba de matemáticas realizada antes de ser admitido en la universidad por una muestra aleatoria de 11 estudiantes de

Tabla 2.8. Relación entre la nota de la prueba de matemáticas del SAT y la calificación media de los estudios universitarios.

Nota de matemáticas en el SAT	450	480	500	520	560	580	590	600	620	650	700
Calificación media en los estudios universitarios	3,25	2,60	2,88	2,85	3,30	3,10	3,35	3,20	3,50	3,59	3,95

una pequeña universidad del Medio Oeste y la calificación media obtenida al terminar los estudios universitarios. Trace un diagrama de puntos dispersos e indique qué información suministra.

Solución

Utilizando el programa Excel, obtenemos la Figura 2.12, que es un diagrama de puntos dispersos de la variable dependiente, la calificación media, y la variable independiente, la nota obtenida en la prueba de matemáticas.

Podemos hacer varias observaciones examinando el diagrama de puntos dispersos de la Figura 2.12. Las calificaciones medias van desde alrededor de 2,5 hasta 4 y las notas obtenidas en la prueba de matemáticas van desde 450 hasta 700. Una interesante pauta es la tendencia ascendente positiva: las calificaciones medias tienden a aumentar directamente con los aumentos de las notas obtenidas en la prueba de matemáticas. Obsérvese también que la relación no suministra una predicción exacta. Algunos estudiantes que obtienen una baja nota en la prueba de matemáticas tienen una calificación media más alta que los estudiantes que obtienen una nota mejor en la prueba de matemáticas. Vemos que la pauta básica indica que las notas más altas obtenidas en los exámenes de admisión predicen mayores calificaciones medias, pero los resultados no son perfectos.

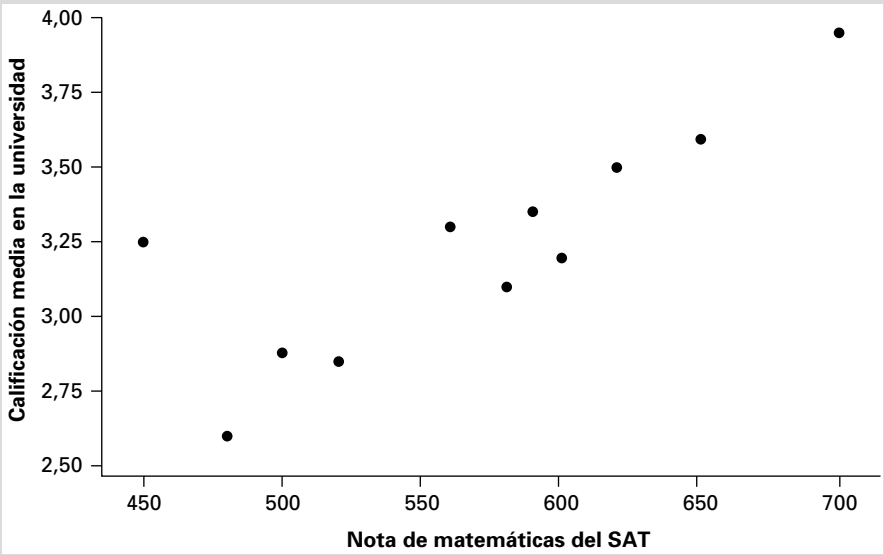


Figura 2.12. Relación entre la calificación media de los estudios universitarios y la nota de la prueba de matemáticas del SAT.

Tablas cruzadas

Hay situaciones en las que necesitamos describir relaciones entre variables categóricas u ordinales. Las empresas de estudios de mercado describen las actitudes hacia los productos, medidas en una escala ordinal, en función de los niveles de estudios, de medidas del estatus social, de las zonas geográficas y de otras variables ordinales o categóricas. Los departamentos de personal estudian los niveles de evaluación de los empleados en relación con las clasificaciones de los puestos, los niveles de estudios y otras variables de los empleados. Los analistas de producción estudian las relaciones entre los departamentos

o líneas de producción y las medidas del rendimiento para averiguar las causas de los cambios de los productos, las causas de la interrupción de la producción y la calidad del producto. Estas situaciones normalmente se describen por medio de tablas cruzadas y se representan mediante gráficos de barras.

Tablas cruzadas

Una **tabla cruzada**, llamada a veces tabla de contingencia, enumera el número de observaciones correspondiente a cada combinación de valores de dos variables categóricas u ordinales. La combinación de todos los intervalos posibles de las dos variables define las casillas en una tabla. Una tabla cruzada de r filas y c columnas se denomina tabla cruzada de dimensión $r \times c$.

EJEMPLO 2.9. La demanda de un producto por zonas residenciales (tabla cruzada)

Un minorista de materiales de construcción ha estado estudiando un plan para abrir tiendas en nuevos lugares dentro de su programa de expansión regional. En una ciudad propuesta para la expansión hay tres lugares posibles: norte, este y oeste. El minorista sabe por experiencia que los tres mayores centros de beneficios de sus tiendas son los de herramientas, madera y pintura. Para seleccionar un lugar, son importantes las pautas de demanda de las diferentes partes de la ciudad. Ha pedido, pues, ayuda al departamento de estudios de mercado para obtener y analizar los datos relevantes. Este minorista cree que tiene una ventaja comparativa en la venta de herramientas.

Solución

La Tabla 2.9 es una tabla de contingencia de 3×4 de las variables «lugar residencial» y «producto comprado». Ha sido realizada por el personal del departamento de estudios de mercado utilizando datos procedentes de una muestra aleatoria de hogares de tres grandes zonas residenciales de la ciudad. Cada zona residencial tenía un prefijo telefónico distinto y se eligieron los cuatro últimos dígitos utilizando un generador de números aleatorios por computador. Si el número no correspondía a una residencia, se generó aleatoriamente otro número telefónico. Si no contestaba nadie a un número, se llamó hasta un máximo de cinco veces para garantizar una elevada tasa de participación.

En cada zona residencial, se contactó con 250 hogares por teléfono y se les pidió que indicaran cuál de tres categorías de productos habían comprado la última vez que habían ido a una tienda de materiales de construcción. La encuesta se realizó para determinar la demanda de herramientas, madera y pintura. Las tres zonas residenciales contienen el mismo número de hogares y, por lo tanto, la muestra aleatoria de 750 representa la población de hogares de toda la ciudad.

Tabla 2.9. Tabla cruzada de la demanda de productos por parte de los hogares por zonas residenciales.

Zona	Herramientas	Madera	Pintura	Ninguna	Total
Este	100	50	50	50	250
Norte	50	95	45	60	250
Oeste	65	70	75	40	250
Total	215	215	170	150	750

Cada casilla de la Tabla 2.9 muestra el número de hogares encuestados en cada una de las zonas residenciales que habían comprado herramientas, madera o pintura el mes anterior. Si habían comprado artículos de más de una categoría, indicaban la categoría en la que más habían gastado. Por ejemplo, 100 hogares encuestados en la zona este habían comprado herramientas y 75 encuestados en la zona oeste habían comprado pintura. En el lado derecho de cada fila observamos el número total de hogares encuestados (250) en esa fila. Asimismo, en la parte inferior de cada columna mostramos el número de hogares encuestados que habían comprado en cada categoría de productos. Los números situados en el lado derecho de las filas y en la parte inferior de las columnas se denominan distribuciones marginales. Estos números son las distribuciones de frecuencias de cada una de las dos variables presentadas en la tabla cruzada.

La Tabla 2.9 contiene un resumen de las pautas de compra de los hogares de los tres barrios. La Figura 2.13 es un gráfico de barras agrupado de la citada tabla. Si la región geográfica y los productos comprados no estuvieran relacionados, sería de esperar que hubiera similitudes en los gráficos de barras.

Sin embargo, observamos que los gráficos de barras sí son diferentes, lo cual induce a pensar que existe una relación entre estas dos variables. Basándose en esta investigación, el personal de marketing ahora sabe que la gente de la zona este compra más a menudo herramientas, mientras que los hogares del norte compran más madera. La demanda de pintura es mayor en el oeste. Basándose en estas pautas, el minorista decide instalar tiendas en el este, debido a que es mayor el potencial de ventas de herramientas.

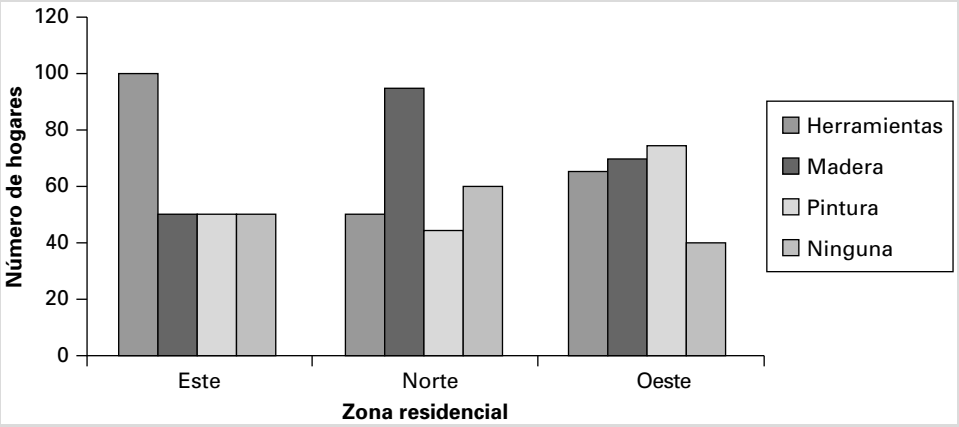


Figura 2.13. Demanda de productos por parte de los hogares por zonas residenciales.

EJEMPLO 2.10. Fuentes del consumo de alcohol de los conductores de vehículos de motor (tabla cruzada)

Un equipo de investigación recibió la tarea de averiguar las fuentes del consumo de alcohol de los conductores de vehículos de motor que tenían diversos niveles de alcohol en la sangre.

Solución

Se obtuvo una muestra aleatoria de automovilistas y se utilizaron los datos resultantes para preparar la Tabla 2.10. Esta tabla muestra la relación entre la concentración de

Tabla 2.10. Tabla cruzada de la CAS de los conductores según el lugar en el que consumieron la primera bebida alcohólica.

	CAS				
Lugar	$\leqslant$0,02%	0,03-0,04%	0,05-0,09%	$\geqslant$0,10%	**Total**
Bar					
Número	22	25	17	14	78
Porcentaje	28,2	32,1	21,8	17,9	100,0
Restaurante					
Número	11	3	9	1	24
Porcentaje	45,8	12,5	37,5	4,2	100,0
En su casa					
Número	45	16	11	10	82
Porcentaje	54,9	19,5	13,4	12,2	100,0
En otra casa					
Número	42	10	6	0	58
Porcentaje	72,5	17,2	10,3	0	100,0
Total					
Número	120	54	43	25	242
Porcentaje	49,6	22,3	17,8	10,3	100,0

alcohol en la sangre y el lugar en el que habían consumido la primera bebida alcohólica las personas que iban conduciendo por la noche y que habían estado bebiendo. Los datos de esta tabla proceden de una muestra aleatoria de personas que conducían un automóvil en el condado Washtenaw (Michigan) entre las 7 de la tarde y las 3 de la madrugada. Las columnas indican la concentración de alcohol en la sangre (CAS) del conductor y se obtuvieron por medio de un alcoholímetro. Normalmente, se considera que cuando estas concentraciones son de $\leqslant$0,02 por ciento, no hay casi ningún alcohol en la sangre y ninguna merma de la capacidad para conducir; cuando están comprendidas entre 0,03 y 0,04 por ciento, hay alcohol en la sangre sin pérdida de capacidad para conducir en el caso de la mayoría de los conductores; cuando están comprendidas entre 0,05 y 0,09 por ciento, casi todos los conductores sufren una pérdida visible de capacidad para conducir y pueden ser condenados por un tribunal; cuando son de $\geqslant$0,10 por ciento, todos están seriamente afectados y representan una amenaza para otros vehículos y peatones. La Tabla 2.10 también indica el porcentaje de conductores que hay en cada categoría de intoxicación dentro de cada fila. Eso permite comparar fácilmente las distintas fuentes del consumo de alcohol de los conductores, a pesar de que el número de conductores de cada fuente es diferente.

La citada tabla permitió obtener algunas indicaciones importantes sobre el consumo de alcohol y la conducta al volante. La muestra sólo contenía conductores que habían consumido al menos una bebida alcohólica durante el día. Comenzando por la fila inferior, que resume toda la muestra, más del 70 por ciento no tenía una CAS que redujera seriamente su capacidad para conducir (por ejemplo, $\leqslant$0,02 por ciento y entre 0,03 y 0,04 por ciento). La fuente más probable de los conductores seriamente afectados eran los bares. En el caso de las 78 personas que consumieron la primera bebida alcohólica en un bar, el 17,9 por ciento tenía una CAS de 0,10 por ciento o más. En el caso de los 82 conductores que comenzaron bebiendo en casa, el 12,2 por ciento tenía el nivel de CAS más alto. Sin embargo, en este grupo de bebedores en casa casi el 75 por ciento

se encontraba en las dos categorías más bajas de CAS y, por lo tanto, no estaban seriamente afectados. Las personas que habían consumido la primera bebida alcohólica en otra casa eran las que menos probabilidades tenían de presentar un alto nivel de CAS. Un importante resultado de este análisis es que en los intentos de reducir el número de conductores seriamente afectados deberían considerarse los bares como una importante fuente (véase la nota bibliográfica 2).

Los gráficos producen un efecto visual mayor que las tablas cruzadas. El gráfico de barras por componentes de la Figura 2.14 es, desde luego, una presentación visual más fuerte del contenido de alcohol en la sangre que la tabla cruzada de la Tabla 2.10.

Los programas informáticos estadísticos pueden realizar la mayoría de estas tablas. En el Capítulo 16 presentamos métodos estadísticos más poderosos para analizar las tablas cruzadas.

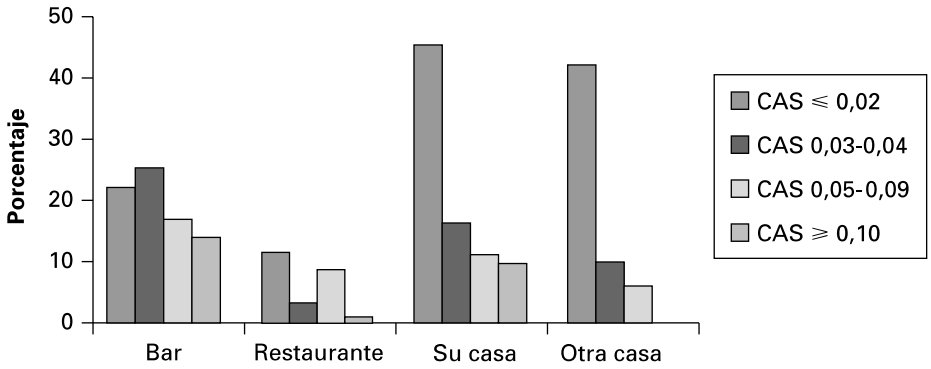


Figura 2.14. CAS de los conductores según el lugar en el que consumieron la primera bebida alcohólica.

EJERCICIOS

Ejercicios básicos

2.40. Realice un diagrama de puntos dispersos con los datos siguientes:

(5, 53)	(21, 65)	(14, 48)	(11, 66)	(9, 46)
(4, 56)	(7, 53)	(21, 57)	(17, 49)	(14, 66)
(9, 54)	(7, 56)	(9, 53)	(21, 52)	(13, 49)
(14, 56)	(9, 59)	(4, 56)		

2.41. Volviendo al ejemplo 2.9, suponga que los datos de la encuesta de mercado no fueran los de la Tabla 2.9 sino los de la tabla adjunta. Explique las conclusiones de esta encuesta desde el punto de vista de la estrategia de producción.

Tabla cruzada revisada de la demanda de productos por parte de los hogares por zonas residenciales

Zona	Herramientas	Madera	Pintura	Ninguno	Total
Este	100	40	60	50	250
Norte	70	45	95	40	250
Oeste	75	70	65	40	250
Total	245	155	220	130	750

2.42. Tres subcontratistas, A, B y C, suministraron 58, 70 y 72 piezas, respectivamente, a una planta la semana pasada. De las piezas suministradas por el subcontratista A, sólo 4 estaban defectuosas. De las piezas suministradas por el B, 60 estaban bien; de las piezas suministradas por el C, sólo 6 estaban defectuosas.

a) Realice una tabla cruzada con los datos.
b) Trace un gráfico de barras.

Ejercicios aplicados

2.43. El supermercado Bishop's registra el precio efectivo de los productos de alimentación y las cantidades vendidas semanalmente. Utilice el fichero de datos **Bishop** para obtener el diagrama de puntos dispersos del precio efectivo de un galón de zumo de naranja y todas las cantidades semanales vendidas a ese precio. ¿Sigue el diagrama de puntos dispersos la pauta que indica la teoría económica?

2.44. Acme Delivery ofrece tres tarifas distintas de envío de paquetes de menos de 5 libras de Maine a

la costa oeste: ordinario, 3 $; urgente, 5 $; y superurgente, 10 $. Para comprobar la calidad de estos servicios, un importante minorista de venta por correo envió 15 paquetes de Maine a Tacoma (Washington) en momentos elegidos aleatoriamente. Los paquetes fueron enviados en grupos de tres por los tres servicios al mismo tiempo para reducir las diferencias resultantes del día del envío. Los datos siguientes muestran el coste de envío, x, y el número de días, y, en pares (x, y):

(3, 7) (5, 5) (10, 2) (3, 9) (5, 6) (10, 5)
(3, 6) (5, 6) (10, 1) (3, 10) (5, 7) (10, 4)
(3, 5) (5, 6) (10, 4)

Trace un diagrama de puntos dispersos de los puntos y comente la relación entre el coste de envío y el momento observado de entrega.

2.45. El fichero de datos **Stordata** contiene los ingresos totales por ventas (en dólares) según el día de la semana. Realice una tabla cruzada en la que aparezcan los días de la semana en las filas y los cuatro intervalos cuartílicos en las columnas.

a) Calcule los porcentajes por filas.

b) ¿Cuáles son las principales diferencias entre los niveles de ventas de los distintos días de la semana según los porcentajes por filas?

c) Describa las pautas esperadas del volumen de ventas a lo largo de la semana basándose en esta tabla.

2.46. Muchas ciudades pequeñas hacen muchos esfuerzos para atraer actividades comerciales, como centros comerciales y grandes almacenes. Uno de los argumentos es que estas instalaciones aumentan las propiedades que puede gravarse y, por lo tanto, generan más fondos para satisfacer las necesidades de las administraciones locales. Los datos del fichero de datos **Citydat** proceden de un estudio de la capacidad municipal de generación de ingresos. Realice un diagrama de puntos dispersos de la variable «taxbase», o sea, de la base imponible, es decir, del valor catastral de todas las propiedades municipales en millones de dólares, en relación con la variable «comper», que es el porcentaje del valor catastral de las propiedades que son propiedades comerciales. ¿Qué información suministra este diagrama de puntos dispersos sobre la base imponible y el porcentaje de propiedades comerciales que hay en la ciudad?

2.6. Errores en la presentación de datos

Los gráficos mal realizados pueden distorsionar fácilmente la verdad. Hemos examinado varios gráficos que resumen y presentan datos. Si se emplean de una manera sensata y prudente, pueden ser excelentes instrumentos para extraer la información esencial de lo que, de lo contrario, sería una mera masa de números. Desgraciadamente, no siempre se intenta resumir los datos de una manera sensata o prudente. En esas circunstancias, es fácil que la manera en que se presenta el resumen induzca a error. Debemos extraer de los datos la imagen más clara y precisa posible. Los gráficos incorrectos pueden ofrecer una imagen distorsionada y dar una falsa impresión. Es posible transmitir un mensaje erróneo sin ser deliberadamente deshonesto.

En este apartado presentamos algunos ejemplos de gráficos engañosos, no con el fin de animar a no utilizarlos sino con el fin de advertir de sus riesgos. El ejemplo 2.11 muestra que las distorsiones en los histogramas pueden llevar a extraer conclusiones incorrectas. El 2.12 muestra que la elección de una u otra opción para el eje de ordenadas en los gráficos de series temporales puede llevar a extraer conclusiones diferentes. Existen otras muchas posibilidades de que los gráficos sean engañosos y para profundizar recomendamos la lectura de Edward Tufte (véase la referencia bibliográfica 10) y de Howard Wainer (véase la referencia bibliográfica 13), que son líderes en el campo de la presentación de datos. Han estudiado el diseño adecuado de los gráficos, así como las causas y los riesgos de hacer deducciones de gráficos mal trazados.

Histogramas engañosos

Sabemos que la amplitud de todos los intervalos debe ser la misma. Supongamos que un conjunto de datos contiene muchas observaciones que se encuentran dentro de una parte relativamente reducida del rango, mientras que otras están muy dispersas. Podríamos tener la tentación de construir una distribución de frecuencias con intervalos reducidos en los que se encontrara la mayoría de las observaciones e intervalos más amplios en otra parte. Aunque recordemos que son las *áreas*, no las alturas, de los rectángulos del histograma las que deben ser proporcionales a las frecuencias, nunca es una opción deseable construir un histograma con diferentes anchos de columnas, ya que puede engañar o distorsionar los resultados. Incluimos este apartado simplemente para señalar los errores que podemos encontrarnos en los histogramas. En el ejemplo 2.11 mostramos cómo se construye un histograma cuando los intervalos no tienen todos ellos la misma amplitud.

EJEMPLO 2.11. Recibos de una tienda de alimentación (intervalos de distinta amplitud)

La Tabla 2.11 muestra las cantidades en dólares de una muestra aleatoria de 692 recibos de una tienda de alimentación.

Tabla 2.11. Recibos de una tienda de alimentación (cantidades en dólares).

Cantidad de dólares	Número de recibos	Proporciones
$0 < 10$	84	84/692
$10 < 20$	113	113/692
$20 < 30$	112	112/692
$30 < 40$	85	85/692
$40 < 50$	77	77/692
$50 < 60$	58	58/692
$60 < 80$	75	75/692
$80 < 100$	48	48/692
$100 < 200$	40	40/92

Uno de los errores que pueden cometerse cuando se realiza un histograma es hacer que sean proporcionales a las frecuencias las *alturas* de los rectángulos en lugar de sus *áreas*. Vemos este histograma engañoso en la Figura 2.15. La observación de este histograma incorrecto nos da la falsa impresión de que hay una elevada proporción de observaciones en la clase más alta. *Bajo ninguna circunstancia debemos construir nunca un histograma con este error. Lo ilustramos únicamente como advertencia contra los gráficos engañosos.*

Las continuas mejoras introducidas en los paquetes informáticos han ido acompañadas de un aumento del uso y el abuso de los gráficos generados por computador. La Figura 2.16 muestra un histograma generado por computador, en el que los intervalos tienen la misma amplitud, a pesar de que tres de las clases tienen diferente amplitud. De nuevo, *bajo ninguna circunstancia debemos construir nunca un histograma con este error. Lo ilustramos únicamente como advertencia contra los gráficos engañosos.*

Para construir un histograma, debemos observar que las cantidades de la Tabla 2.11 se interpretan de la manera habitual. Así, de todos estos recibos, 113/692, o sea, el 16,3 por ciento, se encontraba en el intervalo comprendido entre 10 $ y menos de 20 $.

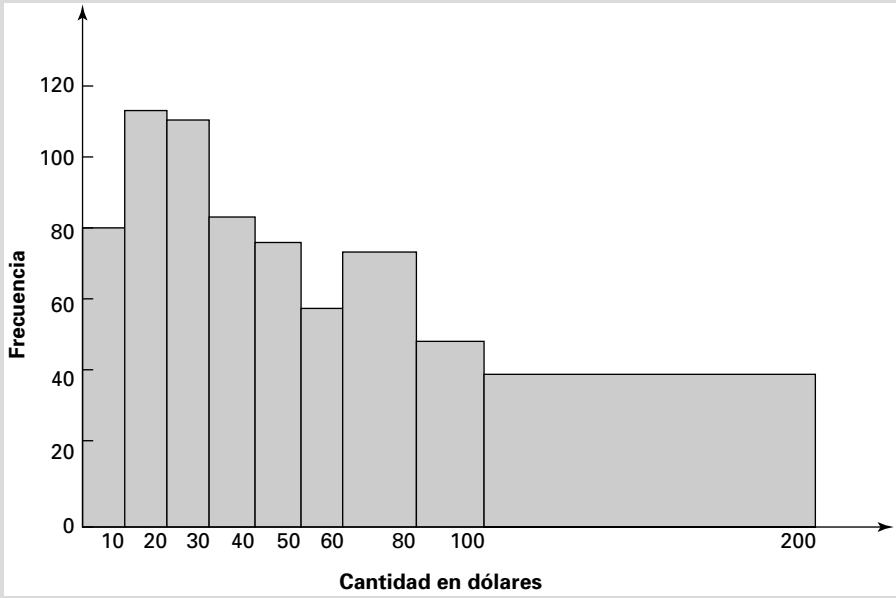


Figura 2.15. Histograma engañoso de los recibos de una tienda de alimentación (error: alturas proporcionales a las frecuencias).

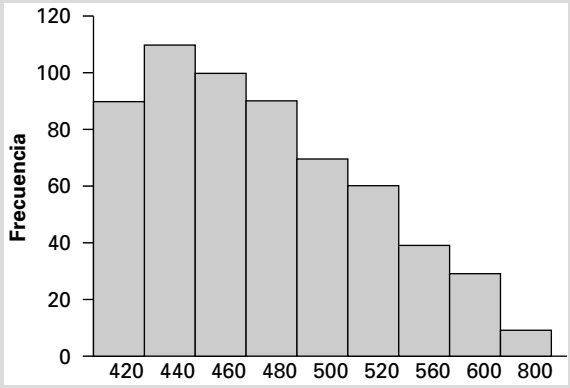


Figura 2.16. Histograma engañoso de los recibos de una tienda de alimentación (error: amplitud desigual de los intervalos).

Tenemos que representar un histograma de manera que las *áreas* de los rectángulos situados sobre los intervalos sean proporcionales a sus frecuencias. Como cada uno de los seis primeros intervalos tiene una amplitud de 10, podemos trazar rectángulos de alturas 84, 113, 112, 85, 77 y 58 sobre estos intervalos. Los dos siguientes intervalos tienen una amplitud de 20, es decir, el doble de la amplitud de cada uno de los seis primeros. Por lo tanto, para que sus áreas sean proporcionales a las frecuencias, los rectángulos representados sobre estos intervalos deben tener alturas que sean la mitad de las frecuencias correspondientes, es decir, 37,5 y 24.

Finalmente, el último intervalo tiene una amplitud de 100, diez veces la amplitud de cada uno de los seis primeros. Por lo tanto, la altura del rectángulo trazado sobre este último intervalo debe ser un décimo de la frecuencia. Es decir, la altura del último

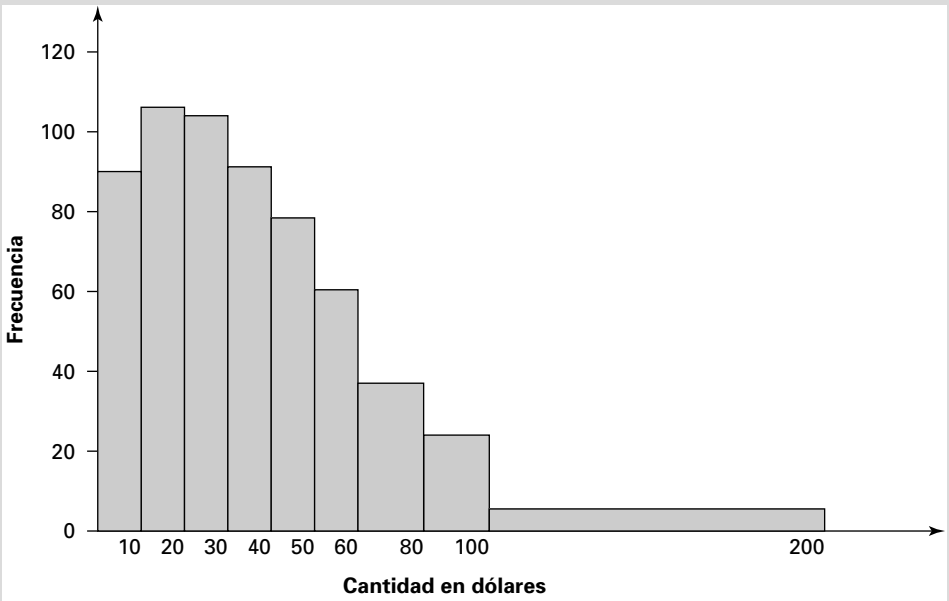


Figura 2.17. Recibos de una tienda de alimentación (histograma).

rectángulo debe ser 4. La razón por la que hacemos que las áreas de estos rectángulos sean proporcionales a las frecuencias se halla en que visualmente asociamos área con tamaño. En la Figura 2.17 vemos un histograma que evita los errores ilustrados en las Figuras 2.15 y 2.16.

Gráficos de series temporales engañosos

Seleccionando una determinada escala de medición, podemos dar la impresión en un gráfico de series temporales de que hay una relativa estabilidad o considerables fluctuaciones a lo largo del tiempo.

EJEMPLO 2.12. Notas obtenidas en la prueba de matemáticas del SAT de 1986-2006 (elección de la escala para realizar gráficos de series temporales)

SAT Math 1986-2006

El fichero de datos **SAT Math 1986-2006** contiene las notas medias de la prueba de matemáticas del SAT obtenidas por los estudiantes de primer año de una universidad entre 1986 y 2006. Represente estos datos con un gráfico de series temporales.

Solución

Mostramos aquí dos gráficos posibles de series temporales para representar las notas de la prueba de matemáticas del SAT que contiene el fichero de datos **SAT Math**. La Figura 2.18 sugiere que las notas medias experimentan grandes fluctuaciones. Esta misma

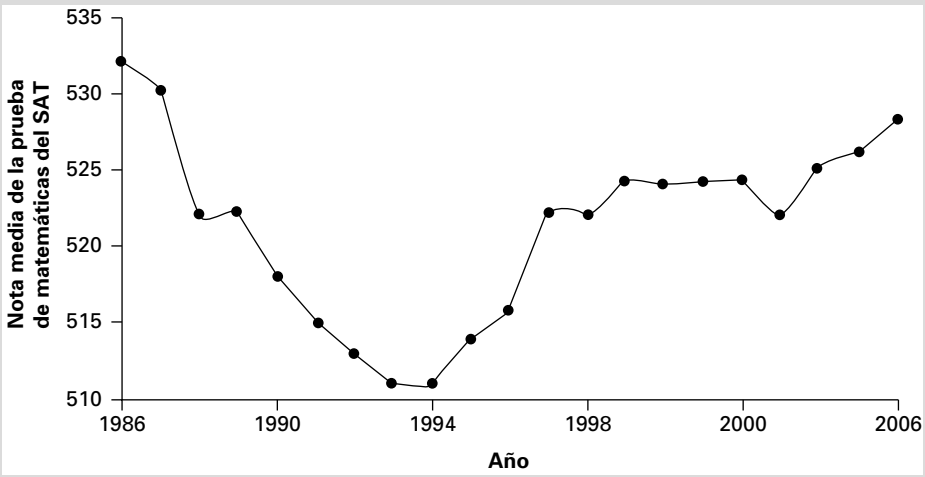


Figura 2.18. Notas de la prueba de matemáticas del SAT: estudiantes de primer año, 1986-2006.

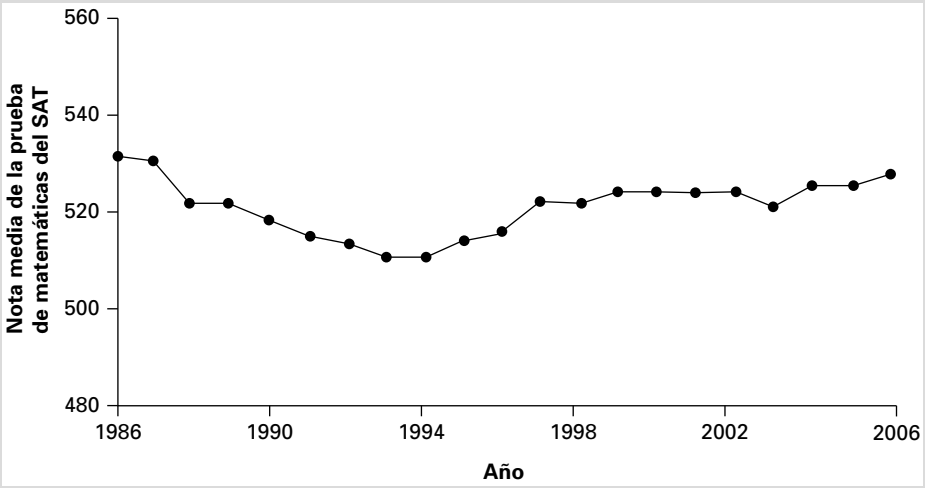


Figura 2.19. Gráfico de series temporales revisado de las notas de la prueba de matemáticas del SAT: estudiantes de primer año, 1986-2006.

información se representa en la Figura 2.19, pero ahora con una escala mucho más amplia en el eje de ordenadas. La imagen resultante es mucho más plana y sugiere que las notas medias han variado mucho menos a lo largo del tiempo.

No existe una elección «correcta» de la escala para ningún gráfico de series temporales. El ejemplo 2.12 lleva a la conclusión de que observar meramente la forma del gráfico es insuficiente para obtener una imagen clara de los datos. También es necesario tener presente la escala en la que se hacen las mediciones.

EJERCICIOS

Ejercicios básicos

2.47. Un supervisor de una planta llevaba un registro del tiempo (en segundos) que necesitaban los empleados para realizar una determinada tarea. La tabla adjunta resume los datos:

Tiempo	30 < 40	40 < 50	50 < 60	60 < 80	80 < 100	100 < 150
Número	10	15	20	30	24	20

a) Represente gráficamente los datos con un histograma.
b) Analice los posibles errores.

2.48. La tabla adjunta enumera el número de visitas diarias realizadas a la nueva página web de una empresa durante 2006.

Mes	**Número**	**Mes**	**Número**
En-06	5,400	Jul-06	5,600
Feb-06	5,372	Ag-06	5,520
Mar-06	5,265	Sep-06	5,280
Abr-06	5,250	Oct-06	5,400
Mayo-06	5,289	Nov-06	5,448
Jun-06	5,350	Dic-06	5,500

a) Represente los datos con un gráfico de series temporales utilizando una escala vertical de 5.000 a 5.700.
b) Represente los datos con un gráfico de series temporales utilizando una escala vertical de 4.000 a 7.000.
c) Comente la diferencia entre estos dos gráficos de series temporales.

Ejercicios aplicados

2.49. El fichero de datos **Exchange Rate** muestra un índice del valor del dólar estadounidense frente a las monedas de sus socios comerciales durante 12 meses consecutivos.
a) Represente estos datos en un gráfico de series temporales utilizando un eje de ordenadas que vaya de 92 a 106.
b) Represente estos datos en un gráfico de series temporales utilizando un eje de ordenadas que vaya de 75 a 120.
c) Comente estos dos gráficos de series temporales.

2.50. El fichero de datos **Inventory Sales** muestra el cociente entre las existencias y las ventas de la industria manufacturera y el comercio de Estados Unidos en un periodo de 12 años. Represente dos gráficos de series temporales de esta serie con diferentes escalas verticales. Comente los resultados.

RESUMEN

En este capítulo hemos mostrado cómo se describen datos por medio de gráficos. En primer lugar, hemos estudiado gráficos como el histograma para resumir una variable numérica. A continuación, hemos utilizado gráficos de barras, gráficos de tarta y diagramas de Pareto para describir una variable categórica. Después hemos examinado la descripción de las relaciones entre (1) dos variables cuantitativas, (2) una variable cuantitativa y una variable ordinal y (3) dos variables categóricas. Hemos mostrado que los diagramas de puntos dispersos pueden suministrar valiosa información al comienzo de un estudio sobre la pauta posible de los puntos de datos.

TÉRMINOS CLAVE

cualitativos (datos), 10
cuantitativos (datos), 10
diagrama de Pareto, 16
diagrama de puntos dispersos, 33
diagrama de tallo y hojas, 30
distribución de frecuencias, 24
distribución de frecuencias acumuladas, 26
distribución de frecuencias relativas, 26
distribución de frecuencias relativas acumuladas, 26
gráfico de barras, 14
gráfico de series temporales, 21
gráfico de tarta, 14
histograma, 27
niveles de medición, 10
nominal, 11
ojiva, 28
ordinal, 11
sesgo, 29
simetría, 28
tabla cruzada, 35
variable categórica, 10
variable numérica, 10
variable numérica continua, 10
variable numérica discreta, 10

2.51. Describa gráficamente el tiempo (en horas) que dedicaron 20 estudiantes a estudiar para un examen de estadística.

6,5	5,8	4,5	6,2	4,8	7,3	4,6	3,9	4,4	5,5
5,2	6,7	3,0	2,4	5,0	3,6	2,9	4,0	2,8	3,6

2.52. Una muestra de 20 analistas financieros ha recibido el encargo de predecir los beneficios por acción que obtendrá una empresa el próximo año. La tabla adjunta resume los resultados.

Predicción	9,95	10,45	10,95	11,45	11,95
($ por acción)	<10,45	<10,95	<11,45	<11,95	<12,45
Número	2	8	6	3	1

a) Trace el histograma.
b) Halle las frecuencias relativas.
c) Halle las frecuencias acumuladas.
d) Halle e interprete las frecuencias relativas acumuladas.

2.53. En una región se observó que utilizaba Internet el 28 por ciento de las personas que tenían una renta de menos de 50.000 $, el 48 por ciento de las que tenían una renta de entre 50.000 $ y 74.999 $ y el 70 por ciento de las que tenían una renta de 75.000 $ como mínimo. Utilice un gráfico de tarta o un gráfico de barras para representar estos datos.

2.54. El Dr. James Mallet, profesor y director del Roland George Investment Institute de la Stetson University, declaró en *USA Today* (véase la referencia bibliográfica 3) que los fondos gestionados por los estudiantes muestran una tendencia ascendente. Utilice un gráfico de series temporales para describir los rendimientos trimestrales de un fondo de inversión gestionado por los estudiantes del máster de administración de empresas de una universidad en relación con los del S&P 500:

	Nov. 1998	**Feb. 1999**	**Mayo 1999**	**Ag. 1999**	**Nov. 1999**
Fondo de inversión de estudiantes de administración de empresas	16,1%	12,5%	2,5%	3,6%	7,0%
S&P 500	21,6%	6,4%	5,1%	1,4%	5,2%

2.55. ¿Están familiarizados los estadounidenses con la nueva legislación tributaria? Según una encuesta (véase la referencia bibliográfica 1), los porcentajes de encuestados que estaban familiarizados con los cambios de la legislación tributaria eran los siguientes: el 70 por ciento conocía la deducción fiscal por hijos, el 52 por ciento la penalización por matrimonio, el 51 por ciento las ganancias de capital, el 44 por ciento los dividendos y el 41 por ciento los tipos impositivos marginales; el 25 por ciento desconocía los cambios. Represente los datos gráficamente.

2.56. Un equipo de estudiantes de administración de empresas recibió el encargo de recomendar cambios que mejoraran el proceso de introducción de datos en la oficina del catastro provincial. El equipo identificó varios tipos de errores, como escribir mal el nombre del titular o el número de la finca. Se pidió a los tasadores que llevaran un registro de los errores que contuvieran los datos que les enviaban. La tabla siguiente es una distribución de frecuencias de los errores:

Error	**Total**
Escribir mal el nombre del titular	23
Escribir mal el número de la finca	21
Propiedad vendida después de que se enviara por correo la notificación del impuesto	5
Finca situada fuera de los límites de la provincia	18
Descripción legal errónea o incompleta	4
Escrituras recibidas después de imprimir la notificación del impuesto	6
Errores de correspondencia	2
Errores varios	1

a) Construya un diagrama de Pareto de estos defectos en la entrada de datos.
b) ¿Qué recomendaciones sugeriría a la oficina del catastro provincial?

2.57. ¿Cuáles son las principales páginas de Internet (medidas por el número total de usuarios que las visitan realmente durante un mes dado)? La tabla adjunta indica las seis páginas principales en diciembre de 2003 (véase la referencia bibliográfica 9). Represente gráficamente los datos.

Página	**Número de visitantes diferentes (miles)**
Páginas de Yahoo!	111.271
Time Warner Network	110.471
Páginas de MSN-Microsoft	110.021
eBay	69.169
Páginas de Google	61.501

2.58. La tabla adjunta basada en Nielsen/Net Ratings de enero de 2004 (véase la referencia bibliográfica 8) muestra el aumento del tráfico semanal de las cinco principales páginas de Internet dedicadas a la salud, el estado físico y la nutrición. Represente gráficamente y analice los factores que pueden haber contribuido a este crecimiento.

Página	Número de visitantes diferentes 4/1/2004	Número de visitantes diferentes 28/12/2003
eDiets	1.036.000	472.000
Weight Watchers	876.000	445.000
WebMD	853.000	524.000
AOL Health	713,000	448.000
Yahoo! Health	590.000	396.000

2.59. ¿Qué relación existe entre el precio de una pintura y su demanda? Se ha obtenido una muestra aleatoria de datos (precio, cantidad) de siete días de funcionamiento. Trace un gráfico y describa la relación entre la cantidad y el precio poniendo énfasis en las observaciones atípicas.

(110, 100) (8, 120) (5, 200) (4, 200)
(10, 90) (7, 110) (6, 150)

2.60. Una empresa de bienes de consumo ha estado estudiando la influencia de la publicidad en los beneficios totales. Se han recogido como parte del estudio datos sobre los gastos publicitarios (miles) y las ventas totales (miles) de un periodo de cinco meses y son los siguientes:

(10, 100) (15, 200) (7, 80) (12, 120) (14, 150)

La primera cifra son los gastos publicitarios y la segunda son las ventas totales. Represente gráficamente los datos.

2.61. El presidente de Pavimentos S.A. quiere información sobre la relación entre la experiencia en la venta al por menor (años) y las ventas semanales (en cientos de dólares). Ha obtenido la siguiente muestra aleatoria sobre la experiencia y las ventas semanales:

(2, 5) (4, 10) (3, 8) (6, 18) (3, 6)
(5, 15) (6, 20) (2, 4)

La primera cifra de cada observación son los años de experiencia y la segunda son las ventas semanales. Represente gráficamente los datos.

2.62. Una muestra aleatoria de 12 jugadores de béisbol universitarios participó en un programa especial de entrenamiento de fuerza en un intento de mejorar sus medias de bateo. El programa duró 20 semanas y se realizó inmediatamente antes del comienzo de la temporada de béisbol. El número medio de horas semanales y la variación de las medias de bateo con respecto a la temporada anterior son los siguientes:

(8,0, 10) (20,0, 100) (5,4, −10) (12,4, 79)
(9,2, 50) (15,0, 89) (6,0, 34) (8,0, 30)
(18,0, 68) (25,0, 110) (10,0, 34) (5,0, 10)

Represente gráficamente los datos. ¿Le parece que tuvo éxito el programa de entrenamiento?

2.63. Un banco ofrece cuatro tipos de cuentas corrientes. Suponga que hace poco se hizo una encuesta a una muestra aleatoria de 300 clientes del banco y se les formularon varias preguntas. Se observó que el 60 por ciento de los encuestados prefería la Cuenta Fácil, el 12 por ciento prefería la Cuenta Inteligente, el 18 por ciento prefería la Supercuenta y el resto la Cuenta Moderna. De los que seleccionaron la Cuenta Fácil, 100 eran mujeres; un tercio de los que seleccionaron la Cuenta Inteligente eran hombres; la mitad de los que seleccionaron la Supercuenta eran hombres; el 80 por ciento de los que seleccionaron la Cuenta Moderna eran hombres.

a) Describa los datos con una tabla cruzada.
b) Describa gráficamente los datos.

2.64. ¿Cómo se entera la gente por primera vez de la existencia de un nuevo producto? Una tienda preguntó a una muestra aleatoria de 200 clientes su edad y si se habían enterado de la existencia del producto por un amigo o por la publicidad de la prensa local. Los resultados indicaron que 50 encuestados tenían menos de 21 años, 90 tenían entre 21 y 35 años y 60 tenían más de 35 años. De los que tenían menos de 21 años, 30 se enteraron de la existencia del producto por un amigo y el resto por la publicidad de la prensa local. Un tercio de las personas del grupo de edad 21-35 años se enteró por primera vez de la existencia del producto por la misma publicidad; el resto por un amigo. Un amigo habló del producto por primera vez al 30 por ciento de la gente de más de 35 años; el resto se enteró por la publicidad de la prensa local.

a) Describa los datos con una tabla cruzada.
b) Describa gráficamente los datos.

2.65. En una encuesta se pidió a una muestra aleatoria de clientes que seleccionara su bebida refrescante favorita de una lista de cinco marcas. Los resultados mostraron que 30 preferían la

marca A, 50 preferían la B, 46 preferían la C, 100 preferían la D y 14 preferían la E.

a) Construya un gráfico de tarta.
b) Construya un gráfico de barras.

2.66. Partiendo del fichero de datos **Smoothies**, construya tablas cruzadas de estas variables:

a) Sexo y nivel de preocupación por la salud.
b) Deseo de suplementos proteínicos y nivel de preocupación de la salud.

2.67. Construya un gráfico de series temporales del crecimiento de la población en el estado de Nueva York desde 1997 hasta la actualidad (*pista*: consulte las páginas www.census.gov o www.bea.doc.gov).

2.68. Partiendo del fichero de datos **Florin**, construya lo siguiente:

a) Una tabla cruzada de las variables «método de pago» y «día de compra».
b) Un gráfico de tarta de la preferencia por el color «Rosa».

2.69. Un promotor de supermercados ha realizado un gran estudio para averiguar las preferencias por las bebidas alcohólicas basándose en el tipo de vehículo utilizado normalmente para ir a un centro comercial. Se entrevistó a una muestra aleatoria de 100 clientes que conducían un automóvil y a una segunda muestra aleatoria de 100 clientes que conducían una camioneta y se les pidió que indicaran sus preferencias por la cerveza o el vino. Los resultados indicaron que el 68 por ciento de los que conducían un automóvil prefería el vino, mientras que el 71 por ciento de los que llevaban una camioneta prefería la cerveza. Construya una tabla cruzada y un gráfico de barras con esta información.

Bibliografía

1. Block, Sandra. Fuente: H&R Block November 2003 survey. Reimpreso en «The Trouble with Taxes: They're Too Hard, They Don't Make Sense, and There's No Easy Fix», *USA Today*, 9 de abril de 2004, pág. B1.
2. Carlson, William L., «Alcohol Usage of the Nighttime Driver», *Journal of Safety Research* 4, marzo, 1972, pág. 12.
3. Fogarty, Thomas A., «Student-Run Funds Teach Real Skills with Real Cash», *USA Today*, 13 de diciembre de 1999, pág. 12B.
4. «N.º 373. Threatened and Endangered Wildlife and Plant Species Number: 2004». Fuente: U.S. Fish and Wildlife Service, *Endangered Species Bulletin*. Reimpreso en *Statistical Abstract of the United States*, Sección 6, Geography and Environment, pág. 227. Véase http://www.census.gov/prd/2004pubs/04statab/geo.pdf. Para información del año en curso, véase http://www.census.gov/statab/www/.
5. «N.º 972. Gross Domestic Product in Manufacturing in Current and Real (2000) Dollars by Industry: 1998-2003». Fuente: U.S. Bureau of Economic Analysis, *Survey of Current Business*, julio, 2004. Reimpreso en *Statistical Abstract of the United States*, sección 21, Manufacturers, pág. 628. Véase http://www.census.gov/prod/2004pubs/04statab/manufact.pdf. Para información sobre el año en curso, véase http://www.census.gov/statb/www/.
6. «N.º 1298. U.S. Exports, Imports, and Merchandise Trade Balance by Country: 1999-2003». Fuente: U.S. Census Bureau. Reimpreso en *Statistical Abstract of the United States*, sección 28, Foreign Commerce and Aid, págs. 814-817. Véase http://www.census.gov/prod/2004pubs/04statab/foreign.pdf. Para información sobre el año en curso, véase http://www.census.gov/statab/www/.
7. «Top Employers by Industry: Top-Ranked Companies Among the 100 Biggest Employers in Central Florida». Fuente: respectivas empresas; investigación de *Sentinel*. *Orlando Sentinel*, 8 de diciembre de 2003. Véase http://www.orlandosentinel.com/business.
8. «Top Health, Fitness & Nutrition Sites, Week Ending January 4 (U.S., Home)». Fuente: Nielsen/NetRatings, enero, 2004. Reimpreso por Janis Mara en «Users Shrink, Sites Expand», ClickZ Stats, 19 de enero de 2004. Véase www.clickz.com/stats/markets/healthcare/article.php/10101_3298631.

9. «Top Properties of December 2003 U.S., Home, Work and University». Fuente: conScore Media Metrix. Reimpreso en «U.S. Web Usage and Traffic, December 2003», ClickZ Stats, 27 de enero de 2004. Véase www.clizkz.com/stats/big_picture/traffic_patterns/article.php/5931_3301321.
10. Tufte, E. R., *The Visual Display of Quantitative Information*, Cheshire, CT, Graphics Press, 1983.
11. Turkey, J., *Exploratory Data Analysis*, Reading, MA, Addison-Wesley, 1977.
12. «Visitors to Travel Agency Sites by Age, U.S. December 2003». Fuente: Hitwise. Reimpreso por Robyn Greenspan en «Internet High on Travel Destinations», ClickZ Stats, 28 de enero de 2004. Véase www.clickz.com/stats/markets/travel/article.php/6071_3304691.
13. Wainer, H., *Visual Revelations*: *Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*, Nueva York, Copernicus/Springer-Verlag, 1997.

Capítulo 3

Descripción numérica de los datos

Esquema del capítulo

Introducción

En el Capítulo 2 hemos descrito los datos gráficamente. En éste, los describimos numéricamente con medidas de la tendencia central, medidas de la variabilidad, medidas de datos agrupados y medidas del sentido y del grado de relación entre dos variables.

3.1. Medidas de la tendencia central

A menudo podemos averiguar si los datos tienden a estar centrados o a agruparse en torno a algún valor construyendo un histograma. Las medidas de la tendencia central suministran información numérica sobre una observación «típica» de los datos. En este apartado analizamos la media, la mediana, la moda y la simetría de los datos (para la media geométrica, véase el apéndice de este capítulo).

Media, mediana, moda

En el Capítulo 1 presentamos los términos *parámetro* y *estadístico*. Un parámetro se refiere a una característica poblacional específica; un estadístico se refiere a una característica muestral específica. Las medidas de la tendencia central normalmente se calculan a partir de datos muestrales más que a partir de datos poblacionales. Una de las medidas de la tendencia central que nos viene rápidamente a la mente es la *media*.

Media aritmética

La **media aritmética** (o media simplemente) de un conjunto de datos es la suma de los valores de los datos dividida por el número de observaciones. Si el conjunto de datos es toda la población de datos, la *media poblacional*, μ, es un *parámetro* que viene dado por

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{3.1}$$

donde N = tamaño de la población y Σ significa «la suma de».

Si el conjunto de datos procede de una muestra, entonces la *media muestral*, $\bar{x}$, es un *estadístico* que viene dado por

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.2}$$

donde n = tamaño de la muestra y Σ significa «la suma de».

Para localizar la *mediana*, debemos reordenar los datos en sentido ascendente o descendente.

Mediana

La **mediana** es la observación que ocupa el lugar central de un conjunto de observaciones ordenadas en sentido ascendente (o descendente). Si el tamaño de la muestra, n, es un número impar, la mediana es la observación que se encuentra en el medio. Si el tamaño de la muestra, n, es un número par, la mediana es la media de las dos observaciones que se encuentran en el medio. La mediana se encontrará en la

$$0{,}50(n + 1) \text{ primera posición ordenada} \tag{3.3}$$

Moda

La **moda**, si existe, es el valor que aparece con más frecuencia.

EJEMPLO 3.1. Ejemplo 3.1 Tiempos realizados en una carrera de 5.000 metros (medidas de la tendencia central)

La Komen Race for the Cure® Series es la serie de carreras de 5.000 metros más multitudinaria del mundo. La Susan G. Komen Breast Cancer Foundation recauda fondos para financiar la lucha contra el cáncer de mama y para darla a conocer; apoya los proyectos de educación, selección y tratamiento en comunidades de todo el mundo; alaba a las mujeres que han sobrevivido y honra a las que han perdido la batalla contra la enfermedad (véase la referencia bibliográfica 3). Halle las medidas de la tendencia central de una muestra de cinco tiempos (en minutos) que hicieron los participantes en una reciente Race for the Cure®:

45 53 45 50 48

Solución

El tiempo medio muestral es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45 + 53 + 45 + 50 + 48}{5} = 48{,}2$$

Cuando ordenamos los datos en sentido ascendente:

45 45 48 50 53

la mediana es 48; dos números son inferiores a 48 minutos y dos son superiores a 48 minutos. Como la media y la mediana son muy parecidas, no importa mucho el valor que utilicemos para describir el centro de los datos. La moda es 45 minutos, ya que aparece dos veces y todos los demás tiempos sólo aparecen una vez. Sin embargo, en este caso, la moda es el valor más bajo y no es el mejor indicador de la tendencia central. Si la muestra incluyera el tiempo de 53 minutos del sexto participante,

45 45 48 50 53 53

la mediana se encontraría situada en la $0{,}5(n + 1)$ primera posición, o sea, la 3,5.ª observación ordenada, que sería 49 minutos. Ahora vemos que los datos son bimodales y que las modas son 45 y 53.

¿Cuál es la mejor medida para describir la tendencia central de los datos: la media, la mediana o la moda? Depende del contexto. Uno de los factores que influyen en la decisión es el tipo de datos, categóricos o numéricos, definidos en el Capítulo 2. La media generalmente es la medida preferida para describir datos numéricos, pero no datos categóricos. Si una persona está totalmente de acuerdo con una afirmación (código 5) y otra está totalmente en desacuerdo (código 1), ¿es la media «ninguna opinión»? Por poner otro ejemplo, supongamos que un comité está formado por dos hombres (cada uno responde 1) y tres mujeres (cada una responde 2). La media aritmética $[(1 + 1 + 2 + 2 + 2)/5 = 1{,}6]$ no tiene sentido. Pero la moda de 2 indica que hay más mujeres que hombres en este comité. Es evidente que los datos categóricos se describen mejor por medio de la moda o de la mediana. Quizá el uso más obvio de la mediana y la moda sea el de los fabricantes que producen bienes, como prendas de vestir, de varias tallas. La talla de los artículos que se venden más a menudo, la moda, es, pues, la más demandada. Saber que la talla media de

las camisas de los hombres europeos es 41,13 o que el número medio del calzado de las mujeres estadounidenses es 8,24 no sirve de nada, pero saber que la talla modal de las camisas es 40 o que el número modal del calzado es 7 es valioso para tomar decisiones sobre las existencias. Sin embargo, la moda puede no representar el verdadero centro de los datos numéricos. Por este motivo, se utiliza menos que la media o la mediana en las aplicaciones empresariales.

EJEMPLO 3.2. Variación porcentual de los beneficios por acción (medidas de la tendencia central)

En una muestra aleatoria de ocho empresas estadounidenses, los beneficios por acción han experimentado este año las siguientes variaciones porcentuales en comparación con el año pasado:

0% 0% 8,1% 13,6% 19,4% 20,7% 10,0% 14,2%

Solución

La variación porcentual media de los beneficios por acción de esta muestra es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0 + 0 + 8{,}1 + 13{,}6 + \cdots + 14{,}2}{8} = 10{,}75, \quad \text{o sea, } 10{,}75\%$$

y la variación porcentual mediana de los beneficios por acción es 11,8 por ciento. La moda es 0 por ciento, ya que aparece dos veces y los otros porcentajes sólo aparecen una vez. Pero esta tasa porcentual modal no representa el centro de estos datos muestrales.

Otro factor que debe considerarse es la presencia de casos atípicos. Siempre que hay casos atípicos en los datos, hay que buscar las causas posibles. En primer lugar, hay que averiguar si se ha cometido un error en la introducción de los datos. La media será mayor si hay grandes casos atípicos y será menor cuando los datos contienen pequeños casos atípicos. La mediana es la medida preferida para describir la distribución de la renta en una ciudad, una región o un país. Como la renta normalmente contiene una pequeña proporción de valores muy altos, la renta media será más alta. Sin embargo, la renta mediana es el nivel de renta o de riqueza por encima del cual se encuentra la mitad de los hogares de la población. Aunque los casos atípicos influyan en la media, en el Capítulo 8 veremos que en muchas situaciones la media tiene algunas propiedades por las que es más atractiva que la mediana.

La mediana, a pesar de su ventaja para descontar las observaciones extremas, se utiliza menos a menudo que la media. La razón se halla en que el desarrollo teórico de los métodos de inferencia basados en la media y las medidas relacionadas con ella es considerablemente más sencillo que el desarrollo de métodos basados en la mediana.

Forma de la distribución

En la Figura 2.10 del Capítulo 2 presentamos histogramas que eran **simétricos**, estaban **sesgados** positivamente y sesgados negativamente. La media y la mediana de una distribución simétrica son iguales, ya que las observaciones están equilibradas, o sea, están distribuidas uniformemente en torno al centro. La media de una distribución sesgada positiva-

mente es mayor que su mediana. La media de una distribución sesgada negativamente es menor que su mediana.

Las distribuciones de la renta o de la riqueza de los hogares de una ciudad, una región o un país tienden a contener una proporción relativamente pequeña de valores altos. Una elevada proporción de la población tiene una renta relativamente modesta, pero, por ejemplo, la renta del 10 por ciento superior de todos los perceptores de renta se extiende a lo largo de un considerable intervalo de valores. Como consecuencia, la media de esas distribuciones normalmente es mucho más alta que la mediana. La media, que es inflada por las personas muy ricas, da una visión demasiado optimista del bienestar económico de la comunidad. Se prefiere, pues, la mediana a la media.

Una de las causas posibles del sesgo es la presencia de casos atípicos. Las observaciones excepcionalmente grandes tienden a aumentar la media, lo que provoca posiblemente un sesgo positivo. Asimismo, si hay observaciones excepcionalmente pequeñas en los datos, el valor de la media disminuye, lo que provoca posiblemente un sesgo negativo. A veces el sesgo es simplemente inherente a la distribución. Si es cero o cercano a cero, la distribución es simétrica o aproximadamente simétrica. Si el valor del sesgo es negativo, la distribución está sesgada hacia la izquierda y, si es positivo, la distribución está sesgada hacia la derecha.

El cálculo manual del sesgo requiere medidas descriptivas que se presentan más adelante en este capítulo. En el ejemplo 3.3 haremos uso de la tecnología, dejando el análisis más extenso del sesgo para el apéndice de este capítulo.

EJEMPLO 3.3. El sueldo anual (sesgo)

Los sueldos anuales de una muestra de cinco empleados son

39.000 $ 37.500 $ 35.200 $ 40.400 $ 100.000 $

Describa la tendencia central y la simetría de los datos.

Solución

En primer lugar, verificamos la exactitud de los datos. No encontrando ningún error (el caso atípico de 100.000 es un sueldo correcto), calculamos el sueldo anual medio, 50.420, un valor que no parece que sea un sueldo «representativo». El sueldo mediano de 39.000 es la medida preferida de la tendencia central. Estos datos no tienen moda. Como la media es mucho mayor que la mediana, es de suponer que los datos están sesgados positivamente, lo que se confirma en la Figura 3.1, en la que el sesgo es aproximadamente igual a 2,21.

Sueldos anuales	
Media	**50.420**
Error típico	12.424,91
Mediana	**39.000**
Moda	#N/A
Desviación típica	27.782,94
Varianza muestral	7,72E+08
Curtosis	4,905059
Sesgo	**2,209069**

Figura 3.1. Sesgo de los sueldos anuales (salida Excel).

Queremos insistir en que la elección de la medida de la tendencia central depende del contexto o del problema. Con eso *no* queremos decir que *siempre* deba preferirse la mediana a la media cuando la población o la muestra está sesgada. Hay veces en las que la media seguiría siendo la medida preferida aunque la distribución estuviera sesgada. Consideremos el caso de una compañía de seguros que es muy probable que se enfrente a una distribución de las reclamaciones sesgada hacia la derecha. Si quiere saber cuál es la cuantía de las reclamaciones más representativa, se prefiere la mediana. Pero supongamos que quiere saber cuánto dinero necesita presupuestar para cubrir las reclamaciones. En ese caso, se prefiere la media.

EJERCICIOS

Ejercicios básicos

3.1. En una muestra aleatoria de 5 semanas se observó que una agencia de cruceros recibía el siguiente número de programas semanales especiales de cruceros al Caribe:

20 73 75 80 82

a) Calcule la media, la mediana y la moda.
b) ¿Qué medida de la tendencia central describe mejor los datos?

3.2. El director de unos grandes almacenes tiene interés en saber cuántas reclamaciones recibe el departamento de atención al cliente sobre la calidad de los aparatos eléctricos que venden los almacenes. Los registros de un periodo de 5 semanas muestran el siguiente número de reclamaciones semanales:

13 15 8 16 8

a) Calcule el número medio de reclamaciones semanales.
b) Calcule el número mediano de reclamaciones semanales.
c) Halle la moda.

3.3. Diez economistas recibieron el encargo de predecir el crecimiento porcentual que experimentará el índice de precios de consumo el próximo año. Sus predicciones fueron

3,6 3,1 3,9 3,7 3,5
3,7 3,4 3,0 3,7 3,4

a) Calcule la media muestral.
b) Calcule la mediana muestral.
c) ¿Cuál es la moda?

3.4. Una cadena de grandes almacenes eligió aleatoriamente 10 establecimientos situados en una región. Tras examinar los datos de ventas, observó que ese año se habían conseguido en las Navidades los siguientes aumentos porcentuales de las ventas en dólares con respecto al año anterior:

10,2 3,1 5,9 7,0 3,7
2,9 6,8 7,3 8,2 4,3

a) Calcule el aumento porcentual medio de las ventas en dólares.
b) Calcule la mediana.
c) Comente la simetría.

3.5. Los porcentajes de la remuneración total correspondientes al pago de pluses de una muestra de 12 altos ejecutivos son los siguientes:

15,8 17,3 28,4 18,2 15,0 24,7
13,1 10,2 29,3 34,7 16,9 25,3

a) Calcule la mediana muestral.
b) Calcule la media muestral.

3.6. La demanda de agua embotellada aumenta durante la temporada de huracanes en Florida. En una muestra aleatoria de 7 horas, se observó que en una tienda se vendió el siguiente número de botellas de 1 galón:

40 55 62 43 50 60 65

a) Describa la tendencia central de los datos.
b) Comente la simetría o el sesgo.

3.7. Un fabricante de radios portátiles obtuvo una muestra de 50 radios de la producción de una semana. Las radios se examinaron minuciosamente y el número de defectos encontrados fue el siguiente:

Número de defectos	0	1	2	3
Número de radios	12	15	17	6

Halle las medidas de la tendencia central.

3.8. Las edades de una muestra de 12 estudiantes matriculados en un curso de macroeconomía en línea son

21 22 27 36 18 19
22 23 22 28 36 33

a) ¿Cuál es la edad media de esta muestra?
b) Halle la edad mediana.
c) ¿Cuál es la edad modal?

Ejercicios aplicados

3.9. El fichero de datos **Rates** contiene las tasaciones (en porcentaje) que se hicieron en 2005 de una muestra aleatoria de 40 solares de una zona comercial.
a) Calcule la tasa porcentual de tasación media, la mediana y la modal.
b) Describa la asimetría o el sesgo de los datos.

3.10. Una muestra de 33 estudiantes de contabilidad anotó el número de horas dedicadas a estudiar la materia de la asignatura durante la semana anterior al examen final. Los datos se encuentran en el fichero de datos **Study**.
a) Calcule la media muestral.
b) Calcule la mediana muestral.
c) Comente la simetría o el sesgo.

3.11. El fichero de datos **Sun** contiene los volúmenes de una muestra aleatoria de 100 envases (de 237 ml) de una nueva crema bronceadora.
a) Halle e interprete el volumen medio.
b) Halle el volumen mediano.
c) ¿Son simétricos los datos o están sesgados? Explique su respuesta.

3.2. Medidas de la variabilidad

La media no es por sí sola una descripción completa o suficiente de los datos. En este apartado presentamos números descriptivos que miden la variabilidad o dispersión de las observaciones con respecto a la media. En concreto, incluimos el rango, el rango intercuartílico, la varianza, la desviación típica y el coeficiente de variación. También describimos los datos numéricamente por medio del resumen de cinco números, con un breve análisis de las reglas básicas para ayudarnos a hallar el porcentaje de observaciones que se encuentran a diversas distancias de la media.

No existen dos cosas exactamente iguales. Éste es uno de los principios básicos del control de calidad estadístico. En todas las áreas hay variaciones. En los deportes, el jugador estrella de baloncesto puede anotar cinco canastas de 3 puntos en un partido y ninguna en el siguiente o puede jugar 40 minutos en un partido y sólo 24 en el siguiente. La variación es obvia en el sector de la música; el tiempo meteorológico varía mucho de un día a otro e incluso de una hora a otra; las calificaciones de un examen varían de unos alumnos a otros dentro de un mismo curso con un mismo profesor; la presión sanguínea, el pulso, el nivel de colesterol y la ingesta de calorías de una persona varían diariamente.

Aunque dos conjuntos de datos tuvieran la misma media, las observaciones individuales de uno de ellos podrían variar con respecto a la media más que las del segundo. Consideremos los dos conjuntos siguientes de datos muestrales:

Muestra A	1	2	1	36
Muestra B	8	9	10	13

Aunque la media es 10 en ambas muestras, es evidente que los datos de la muestra A están más alejados de 10 que los de la muestra B. Necesitamos números descriptivos para medir esta dispersión.

Rango y rango intercuartílico

Rango

Rango es la diferencia entre la observación mayor y la menor.

Cuanto mayor es la dispersión de los datos con respecto al centro de la distribución, mayor es el rango. Como el rango sólo tiene en cuenta la observación mayor y la menor, puede estar muy distorsionado si hay una observación excepcionalmente extrema. Aunque el rango mide la dispersión *total* de los datos, puede ser una medida insatisfactoria de la variabilidad (dispersión) debido a que los casos atípicos, o bien muy altos o bien muy bajos, influyen en él. Una manera de evitar esta dificultad es ordenar los datos en sentido ascendente o descendente, descartar algunos de los números más altos y algunos de los más bajos y hallar el rango del resto. El *rango intercuartílico* mide la dispersión del 50 por ciento intermedio de los datos.

Rango intercuartílico

El **rango intercuartílico (*RIC*)** mide la dispersión que hay en el *50 por ciento central* de los datos; es la diferencia entre la observación de Q_3, el **tercer cuartil** (o sea, el 75.º **percentil**) y la observación de Q_1, el **primer cuartil** (o sea, el 2.º *percentil*). Por lo tanto,

$$RIC = Q_3 - Q_1 \qquad \textbf{(3.4)}$$

donde Q_3 se encuentra situado en la $0{,}75(n + 1)$ primera posición cuando los datos están ordenados en sentido ascendente y Q_1 está situado en la $0{,}25(n + 1)$ primera posición cuando los datos están ordenados en sentido ascendente.

En la ecuación 3.3 ya hemos visto que la mediana es el 50.º percentil, o sea el segundo cuartil (Q_2), y se encuentra situada en la $0{,}50(n + 1)$ primera posición ordenada.

Resumen de cinco números

El **resumen de cinco números** se refiere a las cinco medidas descriptivas: mínimo, primer cuartil, mediana, tercer cuartil y máximo. Es evidente que

$$\text{Mínimo} < Q_1 < \text{Mediana} < Q_3 < \text{Máximo}$$

EJEMPLO 3.4. Tiempos de espera en Comestibles Gilera (resumen de cinco números)

Comestibles Gilera anuncia que los clientes tienen que esperar menos de 1 minuto para pagar si utilizan la Caja rápida. La Figura 3.2 es un diagrama de tallo y hojas de una muestra de 25 tiempos de espera (en segundos). Calcule el resumen de cinco números.

```
Stem-and-leaf
Minutes   N  = 25
Leaf Unit = 1.0

 9    1 1 2 4 6 7 8 8 9 9
(9)   2 1 2 2 2 4 6 8 9 9
 7    3 0 1 2 3 4
 2    4 0 2
```

Figura 3.2. Tiempos de espera en Comestibles Gilera.

Solución

En el diagrama de tallo y hojas vemos que el tiempo mínimo es de 11 segundos y el máximo es de 42. El primer cuartil, Q_1, se encuentra en la 0,25(25 + 1) primera posición ordenada = 6,5 primera posición ordenada. El valor es de 18 segundos. El tercer cuartil, Q_3, se encuentra en la 0,75(25 + 1) primera posición ordenada = 19,5 primera posición ordenada. El valor es de 30,5 segundos. El tiempo mediano es de 22 segundos. El rango es 42 − 11 = 31 segundos; el rango intercuartílico es 30,5 − 18 = 12,5 segundos; es decir, el *50 por ciento central* de los datos tiene una dispersión de 12,5 segundos solamente.

Varianza y desviación típica

Aunque el rango y el rango intercuartílico miden la dispersión de los datos, ambas medidas sólo tienen en cuenta dos de los valores de los datos. Necesitamos una medida que considere cada uno de los valores de los datos. Esa medida *promediaría* la distancia total (Σ) entre cada observación y la media. Esta distancia sería negativa en el caso de los valores menores que la media (y la distancia no es negativa). Si se eleva al cuadrado cada una de estas diferencias, $(x_i - \bar{x})^2$, cada observación (tanto por encima como por debajo de la media) contribuye a la suma de los términos al cuadrado. La media de la suma de los términos al cuadrado se llama *varianza*.

Varianza

Con respecto a la **varianza**, la *varianza poblacional*, σ^2, es la suma de los cuadrados de las diferencias entre cada observación y la media poblacional dividida por el tamaño de la población, N:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{3.5}$$

La *varianza muestral*, s^2, es la suma de los cuadrados de las diferencias entre cada observación y la media muestral dividida por el tamaño de la muestra, n, menos 1.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \tag{3.6}$$

Obsérvese que, en el caso de los datos muestrales, en la ecuación 3.6 la varianza se halla dividiendo el numerador por $(n - 1)$, y no por n. Como nuestro objetivo es hallar una media de los cuadrados de las desviaciones en torno a la media, sería de esperar que hubiera que dividir por n. ¿Por qué se calcula entonces la varianza muestral dividiendo por $(n - 1)$? Si tomáramos un número muy grande de muestras, cada una del tamaño n, de la población y calculáramos la varianza muestral, como se hace en la ecuación 3.6 para cada una de estas muestras, la media de todas estas varianzas muestrales sería la varianza poblacional, σ^2. En el Capítulo 8 veremos que esta propiedad indica que la varianza muestral es un «estimador insesgado» de la varianza poblacional, σ^2. De momento, nos basamos en los estadísticos matemáticos que han demostrado que, si no se conoce la varianza poblacional,

una varianza muestral es un estimador mejor de la varianza poblacional si el denominador de la varianza muestral es $(n - 1)$, en lugar de n.

Para calcular la varianza hay que elevar al cuadrado las distancias, lo que altera la unidad de medición, que ahora son unidades al cuadrado. La *desviación típica*, que es la raíz cuadrada de la varianza, hace que los datos vuelvan a su unidad original de medición. Si las mediciones originales estuvieran en pies, la varianza estaría en pies cuadrados, pero la desviación típica estaría en pies. La desviación típica mide la dispersión *media* en torno a la media.

Desviación típica

Con respecto a la **desviación típica**, la *desviación típica* poblacional, σ, es la raíz cuadrada (positiva) de la varianza poblacional y se define de la forma siguiente:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \tag{3.7}$$

La *desviación típica muestral*, s, es

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \tag{3.8}$$

EJEMPLO 3.5. Calificaciones de los exámenes de un clase de introducción al marketing (medidas de la variabilidad)

Un profesor enseña a dos grandes grupos de introducción al marketing y selecciona aleatoriamente una muestra de calificaciones de los exámenes realizados por los dos grupos. Halle el rango y la desviación típica de cada muestra:

Grupo 1	50	60	70	80	90
Grupo 2	72	68	70	74	66

Solución

Aunque la calificación media de los dos grupos es 70, observamos que las calificaciones del grupo 2 son más cercanas a la media, 70, que las del grupo 1. Y, como cabría esperar, el rango del grupo 1, 40, es mayor que el del grupo 2, que es 8.

Asimismo, sería de esperar que la desviación típica del grupo 1 fuera mayor que la del grupo 2.

$$s_1 = \sqrt{s_1^2} = \sqrt{\frac{(50 - 70)^2 + (60 - 70)^2 + (70 - 70)^2 + (80 - 70)^2 + (90 - 70)^2}{4}} =$$

$$= \sqrt{250} = 15{,}8$$

$$s_2 = \sqrt{s_2^2} = \sqrt{\frac{(72 - 70)^2 + (68 - 70)^2 + (70 - 70)^2 + (74 - 70)^2 + (66 - 70)^2}{4}} =$$

$$= \sqrt{10} = 3{,}16$$

El ejemplo 3.6 ilustra una aplicación de la desviación típica en el área de las finanzas.

EJEMPLO 3.6. El riesgo de un activo (desviación típica)

Vanesa y Jimena Mora, dueñas de una tienda de fotografía, están considerando la posibilidad de invertir en el activo A o en el B. No saben cuál de los dos es mejor y le piden consejo a Sara Nieves, planificadora financiera.

Solución

Sara sabe que la desviación típica, s, es el indicador más frecuente del riesgo o variabilidad de un activo. En las situaciones financieras, la fluctuación en torno a la tasa efectiva de rendimiento de las acciones de una empresa y su tasa esperada de rendimiento se denomina *riesgo* de las acciones. La desviación típica mide la variación de los rendimientos en torno a la media de un activo. Sara obtiene las tasas de rendimiento de cada activo de los cinco últimos años y calcula las medias y las desviaciones típicas de cada uno. La Tabla 3.1 muestra los resultados. Obsérvese que los dos activos tienen la misma tasa media de rendimiento del 12,2 por ciento. Sin embargo, cuando Sara halla las desviaciones típicas, es evidente que el activo B es una inversión más arriesgada.

Tabla 3.1. Tasas de rendimiento: activos A y B.

Años	Tasas de rendimiento	
	Activo A	**Activo B**
Hace 5 años	11,3%	9,4%
Hace 4 años	12,5	17,1
Hace 3 años	13,0	13,3
Hace 2 años	12,0	10,0
Hace 1 años	12,2	11,2
Total	61,0	61,0
Tasa media de rendimiento	12,2%	12,2%
Desviación típica	**0,63**	**3,12**

Teorema de Chebychev y regla empírica

Un matemático ruso, Pafnuty Lvovich Chebychev (1821-1894), estableció los intervalos de datos de cualquier conjunto de datos, *independientemente* de la forma de la distribución.

Teorema de Chebychev

Para cualquier población de media σ, desviación típica σ y $k > 1$, el porcentaje de observaciones que se encuentran dentro del intervalo $[\mu + k\sigma]$ es

$$\text{al menos } 100[1 - (1/k^2)]\% \qquad \textbf{(3.9)}$$

donde k es el número de desviaciones típicas.

Para ver cómo funciona en la práctica el teorema de Chebychev, construimos la Tabla 3.2 para algunos valores de k. Supongamos que la calificación media de un examen es 72 y la desviación típica es 4. Según el teorema de Chebychev, al menos el 75 por ciento de las calificaciones se encuentra en el intervalo comprendido entre 64 y 80 y al menos

Tabla 3.2. Teorema de Chebychev para algunos valores de *k*.

Algunos valores de *k*	1,5	2	2,5	3
$[1 - (1/k^2)]\%$	55,6%	75%	84%	88,9%

el 88,9 por ciento se encuentra en el intervalo comprendido entre 60 y 84. O supongamos que el salario medio de una muestra de trabajadores es de 33.500 $ y la desviación típica es de 1.554 $. Por el teorema de Chebychev, al menos el 55,6 por ciento de los salarios debe encontrarse dentro de (1,5) (1.554 $) = 2.331 $ en torno a la media, es decir, dentro del intervalo comprendido entre 31.169 $ y 35.831 $. Asimismo, al menos el 75 por ciento de los salarios de esta población debe encontrarse dentro de 3.108 $ en torno a la media, es decir, dentro del intervalo comprendido entre 30.392 $ y 36.608 $.

La ventaja del teorema de Chebychev es que puede aplicarse a cualquier población. Sin embargo, en esa garantía se encuentra su principal inconveniente. En el caso de muchas poblaciones, el porcentaje de valores que se encuentran dentro de un intervalo determinado es mucho mayor que el *mínimo* asegurado por el teorema de Chebychev. En el mundo real, muchas grandes poblaciones proporcionan datos en forma de campana que son simétricos, al menos aproximadamente, y muchos de los puntos de datos están agrupados en torno a la media. En el Capítulo 6, analizaremos una fórmula más exacta, pero de momento sólo introduciremos una regla que se aplica a muchas distribuciones en forma de campana.

Regla empírica (68 por ciento, 95 por ciento o casi todo)

En el caso de muchas grandes poblaciones, la **regla empírica** da una estimación del porcentaje aproximado de observaciones que están contenidas en una, dos o tres desviaciones típicas de la media:

- Alrededor del **68 por ciento** de las observaciones se encuentra en el intervalo $\mu \pm 1\sigma$.
- Alrededor del **95 por ciento** de las observaciones se encuentra en el intervalo $\mu \pm 2\sigma$.
- Casi todas las observaciones se encuentran en el intervalo $\mu \pm 3\sigma$.

Supongamos que tenemos una gran población de salarios que tiene una media de 33.500 $ y una desviación típica de 1.554 $. Aplicando la regla empírica, estimamos que alrededor del 68 por ciento de los salarios se encuentra comprendido entre 31.946 $ y 35.054 $ y que alrededor del 95 por ciento se encuentra comprendido entre 30.392 $ y 36.608 $. Sólo hay una probabilidad relativamente pequeña de que una observación se aleje de la media más de $\pm 2\sigma$; cualquier observación que se aleja de la media más de $\pm 3\sigma$ es un caso atípico.

EJEMPLO 3.7. Tiempo que tarda un paquete en llegar a su destino (teorema de Chebychev y regla empírica)

Un grupo de 13 estudiantes está estudiando en Estambul (Turquía) durante cinco semanas. Como parte de su estudio de la economía local, cada uno ha comprado una alfombra oriental y ha hecho las gestiones oportunas para que se la enviaran a Estados Unidos. El tiempo que tardaba en llegar cada alfombra era, en días,

31	31	42	39	42	43	34
30	28	36	37	35	40	

Estime el porcentaje de días que se encuentran dentro de dos desviaciones típicas de la media. ¿Es probable que se tarde 2 meses en enviar la alfombra?

Solución

La media es de 36 días y la desviación típica es de alrededor de 5 días. Según el teorema de Chebychev, al menos el 75 por ciento de los tiempos de envío estaría comprendido entre 26 y 46 días. Observamos que la mediana también es 36. Se prefiere la regla empírica, según la cual alrededor del 95 por ciento de las veces se tardará entre 26 y 46 días en enviar la alfombra. Es improbable que se tarde 2 meses, ya que 60 días es un caso atípico.

Coeficiente de variación

El *coeficiente de variación* expresa la desviación típica en porcentaje de la media.

Coeficiente de variación

El **coeficiente de variación, *CV***, es una medida de la dispersión relativa que expresa la desviación típica en porcentaje de la media (siempre que la media sea positiva).

El *coeficiente de variación poblacional* es

$$CV = \frac{\sigma}{\mu} \times 100\% \qquad \text{si } \mu > 0 \tag{3.10}$$

El *coeficiente de variación muestral* es

$$CV = \frac{s}{\bar{x}} \times 100\% \quad \text{si } \bar{x} > 0 \tag{3.11}$$

Si se comparan las desviaciones típicas de las ventas de los grandes y los pequeños almacenes que venden bienes similares, la desviación típica de los grandes almacenes casi siempre será mayor. Una sencilla explicación es que los grandes almacenes pueden concebirse como un conjunto de pequeños almacenes. La comparación de la variación utilizando la desviación típica sería engañosa. El coeficiente de variación resuelve este problema teniendo en cuenta la escala en la que se miden las unidades poblacionales.

EJEMPLO 3.8. Comparación de acciones (coeficiente de variación)

En el ejemplo 3.6, hemos examinado dos inversiones que tenían la misma tasa media de rendimiento. Ahora los propietarios están considerando la posibilidad de comprar acciones de la empresa A o de la empresa B que cotizan en bolsa. Basándose en los precios de cierre de las acciones de las dos empresas de los últimos meses, se observó que las desviaciones típicas eran muy diferentes: $s_A = 2{,}00$ \$ y $s_B = 8{,}00$ \$. ¿Deben comprarse las acciones de la empresa A, dado que la desviación típica de las acciones de la B es mayor?

Solución

Podríamos creer que las acciones de la empresa B son más volátiles que las de la A. El precio medio de cierre de las acciones de las dos empresas es $\bar{x}_A = 4{,}00$ \$ y $\bar{x}_B = 80{,}00$ \$. A continuación, se calculan los coeficientes de variación para medir y comparar el riesgo de estas oportunidades de inversión:

$$CV_A = \frac{2{,}00\ \$}{4{,}00\ \$} \times 100\% = 50\% \quad \text{y} \quad CV_B = \frac{8{,}00\ \$}{80{,}00\ \$} \times 100\% = 10\%$$

Obsérvese que el valor de mercado de las acciones de A fluctúa más de un periodo a otro que el de las acciones de B.

Cuando se trata de grandes conjuntos de datos, recomendamos que se utilice el computador para obtener las medidas numéricas analizadas en este capítulo. Concluimos este apartado examinando de nuevo el uso del teléfono móvil (véase el ejemplo 2.6) y los datos que se encuentran en el fichero de datos **Mobile Usage**.

Mobile Usage

EJEMPLO 3.9. El uso del teléfono móvil

Los registros de los minutos consumidos por una muestra de 110 abonados al plan más barato de de una compañía de telefonía móvil (250 mensuales como máximo en hora punta) se encuentran en el fichero de datos **Mobile Usage** (véase el ejemplo 2.6). Describa los datos numéricamente.

Solución

Para describir los datos numéricamente, calculamos la media, la mediana, la moda, el rango, la varianza, la desviación típica, el sesgo, el coeficiente de variación y el resumen de cinco números. La media de 261 minutos es algo menor que la mediana de 263 minutos y, según la Figura 3.3, el sesgo es cercano a 0. El tiempo modal es 252 minutos y los datos van desde un máximo de 299 minutos hasta un mínimo de 222. La desviación típica es de 17,5 minutos. La Figura 3.4 incluye el coeficiente de variación, el resumen de cinco números y el rango intercuartílico.

Minutos consumidos	
Media	**261,0636**
Error típico	1,669741
Mediana	263
Moda	252
Desviación típica	**17,5124**
Varianza muestral	**306.684**
Curtosis	−0,33805
Sesgo	**0,001613**
Rango	77
Mínimo	222
Máximo	299
Suma	28.717
Número de casos	110

Figura 3.3. El uso del teléfono móvil (salida Excel).

Descriptive Statistics: Minutes/April

Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum
Minutes	110	0	261.06	1.67	17.51	306.68	6.71	222.00

Variable	Q1	Median	Q3	Maximum	Range	IQR	Skewness
Minutes	251.75	263.00	271.25	299.00	77.00	19.50	0.00

Figura 3.4. Uso del teléfono móvil (salida Minitab).

EJERCICIOS

Ejercicios básicos

3.12. Calcule la varianza y la desviación típica de los siguientes datos muestrales:

6 8 7 10 3 5 9 8

3.13. Calcule la varianza y la desviación típica de los siguientes datos muestrales:

3 0 −2 −1 5 10

3.14. Calcule el coeficiente de variación de los siguientes datos muestrales:

10 8 11 7 9

3.15. El tiempo (en segundos) que tardó una muestra aleatoria de empleados en realizar una tarea es

23	35	14	37	28	45
12	40	27	13	26	25
37	20	29	49	40	13
27	16	40	20	13	66

a) Halle el tiempo medio.
b) Halle la desviación típica.
c) Halle el resumen de cinco números
d) Halle el coeficiente de variación.

3.16. El siguiente diagrama de tallo y hojas contiene los siguientes datos muestrales:

Unidad de tallo

3	**0 1**
4	**5 8 8**
5	**0 3 4 5 7 8 9**
6	**1 4 7 9**
7	**3 6 9**
8	**0 3 7**

a) Calcule el *RIC*.
b) Halle el 8.º decil.
c) Halle el 92.º decil.

3.17. Una muestra aleatoria de datos tiene una media de 75 y una varianza de 25.

a) Utilice el teorema de Chebychev para hallar el porcentaje de observaciones comprendidas entre 65 y 85.

b) Si los datos tienen forma de campana, utilice la regla empírica para hallar el porcentaje aproximado de observaciones comprendidas entre 65 y 85.

3.18. Utilice el teorema de Chebychev para calcular aproximadamente cada una de las siguientes observaciones suponiendo que la media es 250 y la desviación típica es 20. ¿Qué proporción aproximadamente de las observaciones se encuentra

a) Entre 190 y 310?
b) Entre 210 y 290?
c) Entre 230 y 270?

3.19. Un conjunto de datos tiene forma de campana y tiene una media de 450 y una varianza de 625. Indique qué proporción aproximadamente de las observaciones es

a) Superior a 425.
b) Inferior a 500.
c) Superior a 525.

Ejercicios aplicados

3.20. Los rendimientos porcentuales anuales de las acciones ordinarias fueron los siguientes en un periodo de 7 años:

4,0% 14,3% 19,0% −14,7% −26,5% 37,2% 23,8%

Durante ese mismo periodo, los rendimientos porcentuales anuales de las letras del Tesoro de Estados Unidos fueron los siguientes:

6,5% 4,4% 3,8% 6,9% 8,0% 5,8% 5,1%

a) Compare las medias de estas dos distribuciones poblacionales.
b) Compare las desviaciones típicas de estas dos distribuciones poblacionales.

3.21. Los beneficios por acción de una muestra de ocho empresas estadounidenses experimentaron

las siguientes variaciones porcentuales este año en comparación con el anterior:

13,6%	25,5%	43,6%	−19,8%
12,0%	36,3%	14,3%	−13,8%

Halle la variación porcentual media muestral de los beneficios por acción.

3.22. El director de operaciones de una planta embotelladora de agua mineral quiere estar seguro de que el proceso de embotellado de botellas de 1 galón está funcionando correctamente. Se selecciona una muestra aleatoria de 75 botellas y se mide el contenido. El volumen de cada botella se encuentra en el fichero de datos **Water**.

a) Halle el rango, la varianza y la desviación típica de los volúmenes.
b) Halle el resumen de cinco números de los volúmenes.
c) Halle e interprete el rango intercuartílico de los datos.
d) Halle el valor del coeficiente de variación.

3.23. El fichero de datos **Stores** contiene las calificaciones obtenidas por 40 estudiantes en un examen.

a) Halle la calificación media obtenida en este examen.
b) Halle la desviación típica de las calificaciones del examen.
c) Halle el coeficiente de variación.
d) Halle e interprete el rango intercuartílico.

3.24. El fichero de datos **Rates** contiene las tasaciones (en porcentaje) que se hicieron en 2005 de una muestra aleatoria de 40 solares de uso comercial.

a) ¿Cuál es la desviación típica de las tasaciones?
b) ¿Qué proporción aproximadamente de las tasaciones se encontrará dentro de un intervalo de ± 2 desviaciones típicas con respecto a la media?

3.25. Calcule la cantidad media en dólares y la desviación típica de las cantidades en dólares cargadas a una cuenta Visa de Florin's Flower Shop. Los datos se encuentran en la base de datos **Florin**.

3.3. Media ponderada y medidas de datos agrupados

Algunas situaciones requieren un tipo especial de media llamado *media ponderada*.

Media ponderada

La **media ponderada** de un conjunto de datos es

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{\sum w_i} \qquad \textbf{(3.12)}$$

donde w_i = ponderación de la i-ésima observación.

Una importante situación que requiere el uso de una media ponderada es el cálculo de la calificación media.

EJEMPLO 3.10. Calificación media (media ponderada)

Suponga que un estudiante que ha realizado 15 créditos en una universidad durante el primer cuatrimestre ha obtenido una A, una B, una C y una D. Suponga que se asigna un valor de 4 a A, un valor de 3 a A, un valor de 2 a C, un valor de 1 a D y un valor de 0 a F. Calcule la calificación cuatrimestral media del estudiante.

Solución

La calificación media calculada por medio de la media simple es

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{4 + 3 + 2 + 1}{4} = 2{,}5$$

Pero ésta no es la calificación media correcta. Para calcular la media simple, suponemos que cada asignatura tiene la misma importancia o «ponderación», pero este supuesto no tiene en cuenta el hecho de que todas las asignaturas no tienen el mismo número de créditos. Es decir, la calificación A se obtuvo en un curso de inglés de *tres* créditos y la B en un curso de matemáticas de *tres* créditos, pero la C se obtuvo en un laboratorio de biología de *cuatro* créditos y la D, desgraciadamente, en un curso de español de *cinco* créditos. Esta información se resume en la Tabla 3.3.

Utilizando los créditos como ponderaciones, es decir, w_i = número de créditos, y $\Sigma w_i = 15$, la calificación media correcta es 2,267 y no 2,5.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{\sum w_i} = \frac{3(4) + 3(3) + 4(2) + 5(1)}{15} = \frac{34}{15} = 2{,}267$$

Tabla 3.3. Expediente académico cuatrimestral.

Asignatura	Calificación	Valor	Créditos	(valor) × Créditos
Inglés	A	4	3	12
Matemáticas	B	3	3	9
Lab biología	C	2	4	8
Español	D	1	5	5
Total			15	34

La renta personal per cápita es la renta personal total dividida por la población total a mediados del año. Los economistas utilizan una media ponderada para calcular la renta personal per cápita *media* de un año dado. En Estados Unidos, pueden obtenerse los datos sobre la renta personal, la renta y el empleo y los perfiles económicos de cada estado a través del Regional Economic Information System del Bureau of Economic Analysis (www.bea.doc.gov). Las estimaciones de la población a mediados de año se basan en datos suministrados por el Bureau of the Census.

EJEMPLO 3.11. Renta personal per cápita en 2 (media ponderada)

La Tabla 3.4 contiene el tamaño de la población y la renta personal per cápita de una muestra aleatoria de cinco estados de Estados Unidos. Calcule la renta personal per cápita media de 2002 (véanse las referencias bibliográficas 1 y 2).

Solución

Dado que el tamaño de la población varía de unos estados a otros, la renta personal media per cápita de 2002 se calcula por medio de una media ponderada, utilizando las poblaciones de los estados como ponderaciones.

$$\text{Media ponderada: } \frac{\sum_{i=1}^{n} w_i x_i}{\Sigma w_i} = \frac{35.001.986(32.989\ \$) + \cdots + 616.408(29.764\ \$)}{57.968.797} = 31.986{,}12\ \$$$

Tabla 3.4. Población y renta persona per cápita, 2002.

Población	Población	Renta personal per cápita
California	35.001,986	32.989 $
Florida	16.691.701	29.758
Minnesota	5.024.791	33.322
Dakota del Norte	633.911	26.852
Vermont	616.408	29.764
Total	57.968.797	152.685 $

Por lo tanto, la renta personal per cápita media de 2002 es 31.986,12 $ y no 30.537 como sería si se calculara la media aritmética simple.

Una encuesta puede pedir a los encuestados que seleccionen una categoría de edad como «18-25» en lugar de indicar su edad específica. En ese caso, no es posible hallar los valores *exactos* de la media y la varianza. Sin embargo, es posible calcularlas aproximadamente.

Media y varianza aproximadas de datos agrupados

Supongamos que los datos se agrupan en K clases y que las frecuencias son $f_1, f_2, \ldots, f_K$. Si los puntos medios de estas clases son $m_1, m_2, \ldots, m_K$, la media poblacional y la varianza poblacional de los datos agrupados se estiman de la siguiente manera:

a) Para una *población* de N observaciones, tal que

$$N = \sum_{i=1}^{k} f_i$$

la media es

$$\mu = \frac{\sum_{i=1}^{K} f_i m_i}{N} \tag{3.13}$$

y la varianza es

$$\sigma^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \mu)^2}{N} \tag{3.14}$$

Población y muestra

b) Para una *muestra* de *n* observaciones, tal que

$$n = \sum_{i=1}^{K} f_i$$

la media es

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n} \tag{3.15}$$

y la varianza es

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n - 1} \tag{3.16}$$

EJEMPLO 3.12. Análisis de un producto químico para hallar la concentración de impurezas (media y varianza de valores agrupados)

Se ha analizado una muestra de 20 lotes de un producto químico para hallar la concentración de impurezas. Los resultados obtenidos son

Porcentaje de impurezas	$0 < 2$	$2 < 4$	$4 < 6$	$6 < 8$	$8 < 10$
Lotes	2	3	6	5	4

Halle la media y la desviación típica muestrales de estos niveles porcentuales de impurezas.

Solución

Los cálculos se muestran en la Tabla 3.5.

Tabla 3.5. Lotes de un producto químico (cálculo de datos agrupados).

Clases	m_i	f_i	$m_i f_i$	$(m_i - \bar{x})$	$(m_i - \bar{x})^2$	$f_i(m_i - \bar{x})^2$
$0 < 2$	1	2	2	−4,6	21,16	42,32
$2 < 4$	3	3	9	−2,6	6,76	20,28
$4 < 6$	5	6	30	−0,6	0,36	2,16
$6 < 8$	7	5	35	1,4	1,96	9,8
$8 < 10$	9	4	36	3,4	11,56	46,24
	Suma	20	112			120,8

En esta tabla vemos que

$$\sum_{i=1}^{K} f_i = n = 20 \qquad \sum_{i=1}^{K} f_i m_i = 112$$

La media muestral se estima de la siguiente manera:

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n} = \frac{112}{20} = 5{,}6$$

Dado que son datos muestrales, la varianza se estima de la siguiente manera:

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n - 1} = \frac{120{,}8}{19} = 6{,}3579$$

Por lo tanto, la desviación típica muestral se estima de la siguiente manera:

$$s = \sqrt{s^2} = \sqrt{6{,}3579} = 2{,}52$$

Se estima, pues, que en esta muestra la concentración media de impurezas es del 5,6 por ciento y la desviación típica muestral es del 2,52 por ciento.

EJERCICIOS

Ejercicios básicos

3.26. Considere la siguiente muestra de cinco valores y las ponderaciones correspondientes:

x_i	w_i
4,6	8
3,2	3
5,4	6
2,6	2
5,2	5

a) Calcule la media aritmética de los x_i valores sin ponderaciones.
b) Calcule la media ponderada de los x_i valores.

3.27. Considere la siguiente distribución de frecuencias de una muestra de 40 observaciones:

Clase	Frecuencia
0-4	5
5-9	8
10-14	11
15-19	9
20-24	7

a) Calcule la media muestral.
b) Calcule la varianza muestral y la desviación típica muestral.

Ejercicios aplicados

3.28. Halle la renta personal media per cápita ponderada de la siguiente muestra aleatoria de siete estados de Estados Unidos de 2003 (véanse las referencias bibliográficas 1 y 2):

Estado	Población	Per cápita Renta personal
Alabama	4.500.752	26.338
Georgia	8.684.715	29.442
Illinois	12.653.544	33.690
Indiana	6.195.643	28.783
Nueva York	19.190.115	36.574
Pensilvania	12.365.455	31.998
Tennessee	5.841.748	28.455

3.29. Un fabricante de radios portátiles obtuvo una muestra de 50 radios de la producción de una semana. Las radios se comprobaron minuciosamente y el número de defectos encontrados fue el siguiente:

Número de defectos	0	1	2	3
Número de radios	12	15	17	6

Calcule la desviación típica.

3.30. En una muestra aleatoria de 50 pólizas de seguro de propiedades personales se encontró el siguiente número de reclamaciones en los dos últimos años.

Número de reclamaciones	0	1	2	3	4	5	6
Número de pólizas	21	13	5	4	2	3	2

a) Halle el número medio de reclamaciones al día.
b) Halle la varianza y la desviación típica muestrales.

3.31. La tabla adjunta muestra la cantidad de tiempo (en horas) dedicada a estudiar para un examen por una muestra aleatoria de 25 estudiantes de una clase numerosa.

Número de estudio	$0<4$	$4<8$	$8<12$	$12<16$	$16<20$
Número de estudiantes	3	7	8	5	2

a) Estime la media muestral del tiempo de estudio.
b) Estime la desviación típica muestral.

3.32. Se ha pedido a una muestra de 20 analistas financieros que hagan una predicción de los beneficios por acción que obtendrá una empresa el próximo año. La tabla adjunta resume los resultados:

Predicción ($ por acción)	9,95 $<10,45$	10,45 $<10,95$	10,95 $<11,45$	11,45 $<11,95$	11,95 $<12,45$
Número de analistas	2	8	6	3	1

a) Estime la predicción media muestral.
b) Estime la desviación típica muestral.

3.33. Un editor recibe de una imprenta un ejemplar de un libro de texto de 500 páginas. Las pruebas se leen minuciosamente, se anota el número de erratas que hay en cada página y se obtienen los datos de la tabla siguiente:

Número de erratas	0	1	2	3	4	5
Número de páginas	102	138	140	79	33	8

Halle la media y la desviación típica del número de erratas por página.

3.34. En el ejemplo 3.9 se han calculado la media y la desviación típica de los minutos utilizados por una muestra aleatoria de clientes de teléfonos móviles. Ahora calcule y compare la media y la desviación típica basándose solamente en la distribución de frecuencias de la Tabla 2.6.

3.4. Medidas de las relaciones entre variables

En el Capítulo 2 presentamos los diagramas de puntos dispersos para describir gráficamente una relación entre dos variables. En este apartado introducimos la *covarianza* y la *correlación*, que permiten describir numéricamente una relación lineal y a las que prestamos más atención en los Capítulos 12 a 14. La covarianza es una media del *sentido* de una relación lineal entre dos variables.

Covarianza

La **covarianza (Cov)** es una medida de la relación lineal entre dos variables. Un valor positivo indica una relación lineal directa o creciente y un valor negativo indica una relación lineal decreciente.

Una *covarianza poblacional* es

$$\text{Cov}(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N} \qquad \textbf{(3.17)}$$

donde x_i e y_i son los valores observados, μ_x y μ_y son las medias poblacionales y N es el tamaño de la población.

Una *covarianza muestral* es

$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{3.18}$$

donde x_i e y_i son los valores observados, $\bar{x}$ e $\bar{y}$ son las medias muestrales y n es el tamaño de la muestra.

El coeficiente de correlación muestral nos da una medida estandarizada de la relación lineal entre dos variables. Generalmente es una medida más útil, ya que indica tanto el *sentido* como el *grado* de relación. La covarianza y el coeficiente de correlación correspondiente tienen el mismo signo (ambos son positivos o ambos son negativos).

Coeficiente de correlación

El **coeficiente de correlación** se calcula dividiendo la covarianza por el producto de las desviaciones típicas de las dos variables.

Un *coeficiente de correlación poblacional*, ρ, es

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} \tag{3.19}$$

donde σ_x y σ_y son las desviaciones típicas poblacionales de las dos variables.

Un *coeficiente de correlación muestral*, r, es

$$r = \frac{\text{Cov}(x, y)}{s_x s_y} \tag{3.20}$$

donde s_x y s_y son las desviaciones típicas muestrales de las dos variables. Una útil regla práctica es que existe una relación si

$$|r| \geqslant \frac{2}{\sqrt{n}} \tag{3.21}$$

El coeficiente de correlación va de -1 a $+1$. Cuanto más cerca se encuentra r de $+1$, más cerca se encuentran los datos de puntos de una línea recta ascendente que indica una relación lineal *positiva*. Cuanto más cerca se encuentra r de -1, más cerca se encuentran los datos de puntos de una línea recta descendente que indica una relación lineal *negativa*. Cuando $r = 0$, no existe ninguna relación *lineal* entre x e y, pero eso no quiere decir necesariamente que no exista ninguna relación. En el Capítulo 2 presentamos los diagramas de puntos dispersos, que eran una medida gráfica para determinar la relación. La Figura 3.5 muestra algunos ejemplos de diagramas de puntos dispersos y sus correspondientes coeficientes de correlación. La Figura 3.6 es un diagrama de las ventas trimestrales de una gran empresa minorista. Obsérvese que las ventas varían según el trimestre del año, reflejando las pautas de compra de los consumidores. El coeficiente de correlación entre la variable tiempo y las ventas trimestrales es cero. Vemos la existencia de una relación estacional muy clara, pero no es una relación lineal.

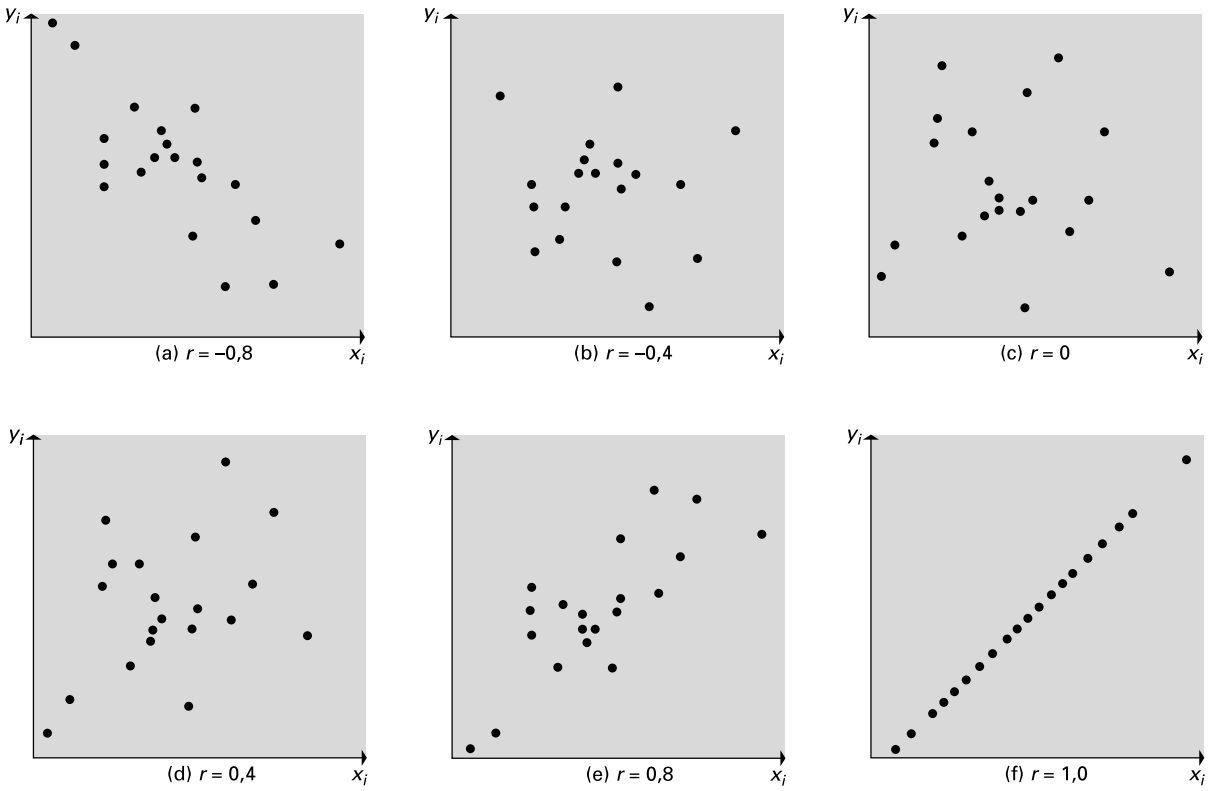
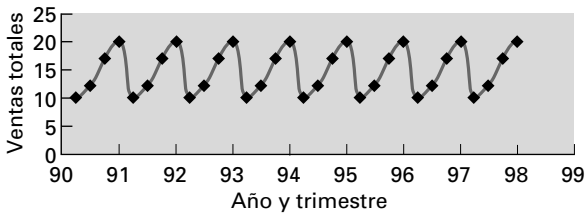


Figura 3.5. Diagramas de puntos dispersos y correlación.

Figura 3.6.
Ventas al por menor por trimestre.



EJEMPLO 3.13. Planta manufacturera (covarianza y coeficiente de correlación)

Rising Hills Manufacturing Inc. desea estudiar la relación entre el número de trabajadores, X, y el número de mesas, Y, producidas en su planta de Redwood Falls. Ha tomado una muestra aleatoria de 10 horas de producción. Se han obtenido las siguientes combinaciones (x, y) de puntos:

(12, 20)	(30, 60)	(15, 27)	(24, 50)	(14, 21)
(18, 30)	(28, 61)	(26, 54)	(19, 32)	(27, 57)

Calcule la covarianza y el coeficiente de correlación. Analice brevemente la relación entre el número de trabajadores y el número de mesas producidas por hora. Los datos se encuentran en el fichero de datos **Rising Hills**.

Rising Hills

Solución

Los cálculos se indican en la Tabla 3.6.

Tabla 3.6. Cálculos de la covarianza y la correlación.

x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
12	20	−9,3	86,49	−21,2	449,44	197,16
30	60	8,7	75,69	18,8	353,44	163,56
15	27	−6,3	39,69	−14,2	201,64	89,46
24	50	2,7	7,29	8,8	77,44	23,76
14	21	−7,3	53,29	−20,2	408,04	147,46
18	30	−3,3	10,89	−11,2	125,44	36,96
28	61	6,7	44,89	19,8	392,04	132,66
26	54	4,7	22,09	12,8	163,84	60,16
19	32	−2,3	5,29	−9,2	84,64	21,16
27	57	5,7	32,49	15,8	249,64	90,06
Σ = 213	**Σ = 412**		**Σ = 378,1**		**Σ = 2.505,6**	**Σ = 962,4**

Aplicando la ecuación 3.18, tenemos que

$$\text{Cov}\,(x,\, y) = s_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{962,4}{9} = 106,93$$

Aplicando la ecuación 3.20, tenemos que

$$r = \frac{\text{Cov}\,(x,\, y)}{s_x s_y} = \frac{106,93}{\sqrt{42,01}\,\sqrt{278,4}} = 0,989$$

Aplicando la ecuación 3.21, tenemos que

$$|0,989| \geqslant \frac{2}{\sqrt{10}} \cong 0,64$$

Llegamos a la conclusión de que existe una estrecha relación positiva entre el número de trabajadores y el número de mesas producidas por hora.

EJEMPLO 3.14. Análisis de carteras de acciones (análisis de los coeficientes de correlación)

Alicia Viera, analista financiera de Títulos Integrados, está examinando diferentes acciones para un nuevo fondo de inversión que está desarrollando. Una de sus preguntas se refiere a los coeficientes de correlación entre los precios de las diferentes acciones. Para hallar las pautas de los precios de las acciones, ha elaborado una serie de diagramas de puntos dispersos y ha calculado el coeficiente de correlación muestral de cada diagrama. ¿Qué información suministra la Figura 3.7 a Alicia?

Solución

Alicia ve que es posible controlar la variación del precio medio del fondo de inversión combinando diferentes acciones en una cartera. La variación de la cartera aumenta si se incluyen acciones que tienen coeficientes de correlación positivos, ya que los precios tienden a subir juntos. En cambio, la variación de la cartera es menor si se incluyen acciones que tienen coeficientes de correlación negativos. Cuando sube el precio de las acciones de una de las empresas, baja el precio de las de otra y el precio conjunto es más estable. Los observadores de los precios de las acciones que tienen experiencia podrían cuestionar la posibilidad de que existan coeficientes de correlación negativos muy altos. Nuestro objetivo aquí es ilustrar gráficamente los coeficientes de correlación de ciertas pautas de datos observados y no describir exactamente un mercado concreto. Después de examinar estos coeficientes de correlación, Alicia está preparada para comenzar a construir su cartera. En el Capítulo 6 mostramos cómo afectan exactamente los coeficientes de correlación entre los precios de las acciones a la variación de toda la cartera.

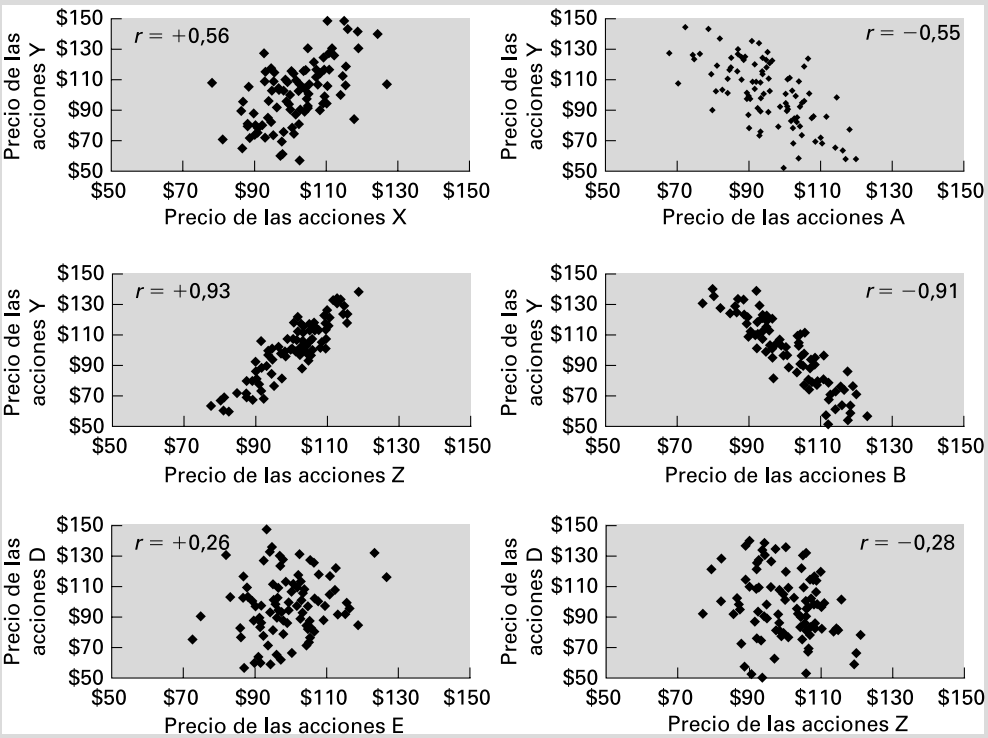


Figura 3.7. Relaciones entre los precios de varias acciones.

Para calcular medidas descriptivas como la covarianza muestral y el coeficiente de correlación muestral puede utilizarse el programa Minitab, el Excel, el SPSS, el SAS y otros muchos paquetes estadísticos. La Figura 3.8 muestra la salida Minitab correspondiente a la covarianza y la correlación.

Si se utiliza el programa Excel para calcular la covarianza, hay que tener especial cuidado. Obsérvese que el valor que figura en la salida Excel de la Figura 3.9 da una covarianza de 96,24; sin embargo, sabemos que la covarianza muestral es de 106,93 en el caso

de estos datos. Excel (XP o 2000) calcula automáticamente la covarianza poblacional como se indica en la ecuación 3.17. Obtenemos la covarianza muestral multiplicando la covarianza poblacional de 96,24 por $n/(n-1)$.

$$(96{,}24)\,\frac{n}{n-1} = (96{,}24)\,\frac{10}{9} = 106{,}93$$

Covarianzas: trabajadores, mesas

	Trabajadores	Número de mesas
X, trabajadores	42,0111	
Y, mesas	106,9333	278,4000

Correlaciones: trabajadores, mesas

Correlación de x e y en personas = 0,989
Valor P = 0,000

Figura 3.8. Covarianza y correlación: trabajadores, mesas (salida Minitab).

Covarianza: trabajadores, mesas		
	Trabajadores	**Mesas**
Trabajadores	37,81	
Mesas	**96,24**	250,56

Correlación: trabajadores, mesas		
	Trabajadores	**Mesas**
Trabajadores	1	
Mesas	**0,988773**	1

Figura 3.9. Covarianza y correlación: trabajadores, mesas (salida Excel).

EJERCICIOS

Ejercicios básicos

3.35. A continuación, se presenta una muestra aleatoria de siete pares (x, y) de puntos de datos:

(1, 5) (3, 7) (4, 6) (5, 8) (7, 9)
(3, 6) (5, 7)

a) Calcule la covarianza.
b) Calcule el coeficiente de correlación.

3.36. A continuación, se presenta una muestra aleatoria de cinco pares (x, y) de puntos de datos:

(12, 200) (30, 600) (15, 270) (24, 500)
(14, 210)

a) Calcule la covarianza.
b) Calcule el coeficiente de correlación.

3.37. A continuación, se presenta una muestra aleatoria del precio por tabla de contrachapado, X, y la cantidad vendida, Y (en miles):

Precio por trozo (X)	Miles de trozos vendidos (Y)
6 $	80
7	60
8	70
9	40
10	0

a) Calcule la covarianza.
b) Calcule el coeficiente de correlación.

Ejercicios aplicados

3.38. Un hospital tiene interés en averiguar la eficacia de un nuevo medicamento para reducir el tiempo necesario para recuperarse totalmente de una operación de rodilla. La recuperación total se mide por medio de una serie de tests de fuerza que comparan la rodilla operada con la rodilla sin operar. El medicamento se administró en dosis diferentes a 18 pacientes durante un periodo de 6 meses. Los datos (x, y) siguientes indican el número de unidades de medicamento, X, y los días necesarios para la recuperación total Y de cada paciente:

(5, 53) (21, 65) (14, 48) (11, 66) (9, 46)
(4, 56) (7, 53) (21, 57) (17, 49) (14, 66)
(9, 54) (7, 56) (9, 53) (21, 52) (13, 49)
(14, 56) (9, 59) (4, 56)

a) Calcule la covarianza.
b) Calcule el coeficiente de correlación.
c) Analice brevemente la relación entre el número de unidades de medicamento y el tiempo de recuperación. ¿Qué dosis deberíamos recomendar basándonos en este análisis inicial?

3.39. Acme Delivery ofrece tres tarifas distintas de envío de paquetes de menos de 5 libras de Maine a la costa oeste: ordinario, 3 $; urgente, 5 $, y superurgente, 10 $. Para comprobar la calidad de estos servicios, un importante minorista de venta por correo envió 15 paquetes de Maine a Tacoma (Washington) en momentos elegidos aleatoriamente. Los paquetes fueron enviados en grupos de tres por los tres servicios al mismo tiempo para reducir las diferencias resultantes del día del envío. Los datos siguientes muestran el coste de envío, X, y el número de días, Y, en pares (x, y):

(3, 7) (5, 5) (10, 2) (3, 9) (5, 6) (10, 5)
(3, 6) (5, 6) (10, 1) (3, 10) (5, 7) (10, 4)
(3, 5) (5, 6) (10, 4)

a) Describa los datos numéricamente (covarianza y correlación).
b) Analice el valor de los servicios de precio más alto desde el punto de vista del envío más rápido.

3.5. Obtención de relaciones lineales

Hemos visto cómo puede describirse la relación entre dos variables utilizando datos muestrales. Los diagramas de puntos dispersos representan la relación y los coeficientes de correlación son una medida numérica. En muchos problemas económicos y empresariales se desea una relación funcional específica.

- ¿Qué nivel medio de ventas cabe esperar si el precio se fija en 10 $ por unidad?
- Si se emplean 250 trabajadores, ¿cuántas unidades cabe esperar?
- Si un país en vías de desarrollo aumenta su producción de fertilizantes en 1 millón de toneladas, ¿cuánto cabe esperar que aumente la producción de cereales?

Los modelos económicos utilizan relaciones funcionales específicas para indicar el efecto que producen en una variable dependiente, Y, algunas variaciones de la variable independiente, X. En muchos casos, podemos calcular aproximadamente las relaciones funcionales deseadas mediante una ecuación lineal:

$$Y = \beta_0 + \beta_1 X$$

donde Y es la variable dependiente, X es la variable independiente, β_0 es la ordenada en el origen y β_1 es la pendiente de la recta, o sea, la variación que experimenta Y por cada variación unitaria de X. En nuestras aplicaciones, partimos del supuesto nominal de que podemos fijar X en diferentes valores y a cada uno le corresponderá un valor medio de Y debido a la relación lineal subyacente en el proceso estudiado. El modelo de la ecuación lineal calcula la media de Y para cada valor de X. Esta idea es la base para obtener muchas relaciones económicas y empresariales, entre las que se encuentran las funciones de demanda, las funciones de producción, las funciones de consumo y las predicciones sobre las ventas.

Utilizamos regresiones para averiguar cuál es la mejor relación entre Y y X para una aplicación específica. Para eso es necesario hallar los mejores valores de los coeficientes β_0 y β_1. Generalmente, utilizamos los datos del proceso para calcular «estimaciones» o valores numéricos de los coeficientes β_0 y β_1. Estas estimaciones —b_0 y b_1— generalmente se calculan utilizando una *regresión por mínimos cuadrados*, técnica que se aplica mucho en paquetes estadísticos como Minitab y en hojas de cálculo como Excel. El método de mínimos cuadrados selecciona la recta que mejor se ajusta, dado un conjunto de puntos de datos. Consideremos una representación característica de puntos de un proceso que tiene una relación lineal mostrada en la Figura 3.10.

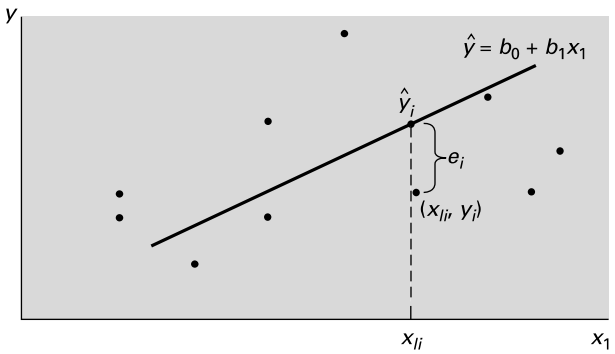


Figura 3.10. Función lineal y puntos de datos.

La ecuación lineal representada por la recta es la ecuación lineal que mejor se ajusta. Vemos que los puntos de datos individuales se encuentran por encima y por debajo de la recta y que ésta tiene puntos con desviaciones tanto positivas como negativas. La distancia de cada punto (x_i, y_i) con respecto a la ecuación lineal es el residuo, e_i. Nos gustaría elegir la ecuación de manera que alguna función de los residuos positivos y negativos fuera lo más pequeña posible. Eso significa estimar los coeficientes β_0 y β_1.

Los primeros matemáticos trataron denodadamente de desarrollar un método para estimar los coeficientes de la ecuación lineal. No era útil minimizar simplemente las desviaciones, ya que las desviaciones tienen tanto signo positivo como negativo. También se han desarrollado algunos métodos que utilizan valores absolutos, pero ninguno ha resultado tan útil o tan popular como la regresión por mínimos cuadrados. Más adelante veremos que los coeficientes desarrollados utilizando este método tienen propiedades estadísticas muy útiles. Una importante cautela en el caso de los mínimos cuadrados es que los puntos atípicos extremos pueden tener tal influencia en la recta de regresión que toda la recta se dirija hacia esos puntos. Por lo tanto, siempre debemos examinar los diagramas de puntos dispersos para asegurarnos de que la relación de regresión no se basa solamente en unos cuantos puntos extremos.

Desarrollamos ecuaciones para calcular estas estimaciones utilizando el método de regresión por mínimos cuadrados que presentaremos con mayor profundidad en el Capítulo 12. La regresión por mínimos cuadrados elige los valores de b_0 y b_1 con los que se minimiza la suma de los cuadrados de los residuos.

Regresión por mínimos cuadrados

La recta de regresión por mínimos cuadrados basada en datos muestrales es

$$\hat{y} = b_0 + b_1 x \tag{3.22}$$

b_1 es la pendiente de la recta, o sea la variación de y por cada variación unitaria de x, y se calcula de la forma siguiente:

$$b_1 = \frac{\text{Cov}(x, y)}{s_x^2} = \frac{s_y}{s_x} \tag{3.23}$$

donde b_0 es la ordenada en el origen y se calcula de la forma siguiente:

$$b_0 = \bar{y} - b_1 \bar{x} \tag{3.24}$$

Rising Hills

EJEMPLO 3.15. Planta manufacturera (recta de regresión)

En el ejemplo 3.13, presentamos el número de trabajadores, X, y el número de mesas producidas por hora, Y, por una muestra de 10 trabajadores. Si la dirección decide emplear 25 trabajadores, estime el número esperado de meses que es probable que se produzcan. Los datos se encuentran en el fichero de datos **Rising Hills**.

Solución

En el ejemplo 3.13 hemos calculado la covarianza y la correlación de estos datos muestrales:

$$\text{Cov}(x, y) = 106{,}93$$

$$r = 0{,}989$$

La covarianza muestra que el sentido de la relación es *positivo*; la elevada correlación de 0,989 también indica que los puntos de datos muestrales están muy cerca de una recta ascendente, como se observa en la Figura 3.11.

Con los datos de la Tabla 3.6, calculamos los coeficientes de regresión muestrales:

$$b_1 = \frac{\text{Cov}(x, y)}{s_x^2} = \frac{106{,}93}{42{,}01} = 2{,}545$$

$$b_0 = \bar{y} - b_1\bar{x} = 41{,}21 - 2{,}545(21{,}3) = -13{,}02$$

De esta expresión se desprende que la recta de regresión muestral es

$$\hat{y} = b_0 + b_1 x = -13{,}02 + 2{,}545x$$

Con 25 trabajadores, es de esperar que se produzcan

$$\hat{y} = -13{,}02 + 2{,}545(25) = 50{,}62$$

o sea, alrededor de 51 mesas.

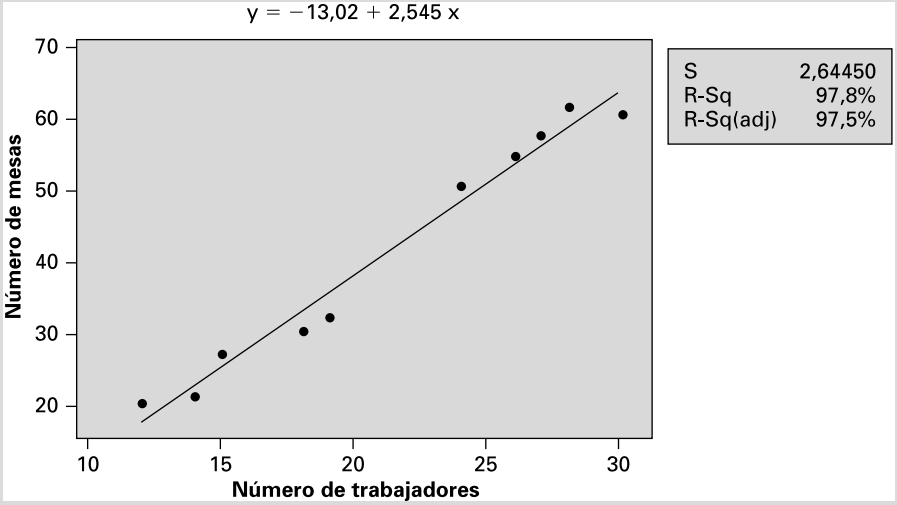


Figura 3.11. Recta de regresión: trabajadores, mesas (salida Minitab).

También podemos utilizar un paquete estadístico como Minitab o una hoja de cálculo como Excel para hallar los mismos coeficientes de regresión y la misma recta de regresión. La Figura 3.12 muestra la salida Minitab obtenida con estos datos.

The regression equation is

Number of Tables = −13.0 + 2.55 (Number of Workers)

Predictor	Coef	SE Coef	T	P
Constant	−13.016	3.015	−4.32	0.003
Number of Workers	2.5454	0.1360	18.72	0.000

S = 2.64450 R-Sq = 97.8% R-Sq(adj) = 97.5%

Figura 3.12. Análisis de regresión: número de mesas en relación con el número de trabajadores (salida Minitab).

No pretendemos sugerir que *siempre* podemos introducir *cualquier* valor de x en una recta mínimo-cuadrática y tomar una decisión razonable. A veces la relación es meramente espuria, es decir, el valor de x puede encontrarse fuera de un intervalo aceptable de valores. Por ejemplo, como el número de trabajadores de la planta manufacturera Rising Hills estaba comprendido entre 12 y 30, no podemos predecir el número de mesas que se producirían por hora si se emplearan 100 trabajadores.

Al igual que ocurre en todo el capítulo, nuestro objetivo es aprender a describir datos numéricamente y no a realizar un sofisticado análisis estadístico de modelos de regresión lineal. Eso ya llegará en el Capítulo 12 y posteriores. Utilizaremos el computador para calcular coeficientes de regresión de datos más realistas, ya que el tamaño de la muestra generalmente hace que los cálculos sean tediosos.

EJERCICIOS

Ejercicios básicos

3.40. Dados estos pares (x, y) de puntos de datos:

(1, 5) (3, 7) (4, 6) (5, 8) (7, 9)

a) Calcule b_1.
b) Calcule b_0.
c) ¿Cuál es la ecuación de la recta de regresión?

3.41. Los datos siguientes muestran X, el precio cobrado por tabla de contrachapado, e Y, la cantidad vendida (en miles):

Precio por tabla (X)	Miles de tablas vendidos (Y)
6 $	80
7	60
8	70
9	40
10	0

a) Calcule la covarianza.
b) ¿Qué información suministra el coeficiente de correlación?
c) Calcule e interprete b_1.
d) Calcule b_0.
e) ¿Qué cantidad de tablas es de esperar que vendamos si el precio es de 7 $ por tabla?

3.42. Una muestra aleatoria de 7 días de operaciones produjo los siguientes valores de los datos (precio, cantidad):

Precio por litro de pintura (X)	Cantidad vendida (Y)
10	100
8	120
5	200
4	200
10	90
7	110
6	150

a) Describa los datos numéricamente (calcule la covarianza y la correlación).
b) Calcule e interprete b_1.
c) Calcule e interprete b_0.
d) ¿Cuántos litros de pintura es de esperar que vendamos si el precio es de 7 $ el litro?

Ejercicios aplicados

3.43. Una empresa de bienes de consumo ha estado estudiando la influencia de la publicidad en los beneficios totales. En este estudio, se han recogido los siguientes datos sobre los gastos publicitarios (en miles) y las ventas totales (en miles) de un periodo de cinco meses:

(10, 100) (15, 200) (7, 80) (12, 120) (14, 150)

El primer número son los gastos publicitarios y el segundo son las ventas totales.

a) Represente gráficamente los datos y calcule el coeficiente de correlación.
b) ¿Demuestran estos resultados que la publicidad influye positivamente en las ventas?
c) Calcule los coeficientes de regresión, b_0 y b_1.

3.44. El presidente de Pavimentos S.A. quiere información sobre la relación entre la experiencia en la venta al por menor (años) y las ventas semanales (en cientos de dólares). Ha obtenido la siguiente muestra aleatoria sobre la experiencia y las ventas semanales:

(2, 5) (4, 10) (3, 8) (6, 18) (3, 6)
(5, 15) (6, 20) (2, 4)

La primera cifra de cada observación son los años de experiencia y la segunda son las ventas semanales.

a) Calcule la covarianza y la correlación.
b) Calcule los coeficientes de regresión, b_0 y b_1.
c) Explique brevemente la ecuación de regresión que podría utilizarse para predecir las ventas. Incluya una indicación del rango al que podría aplicarse la ecuación.

3.45. Una muestra aleatoria de 12 jugadores de béisbol universitarios participó en un programa especial de entrenamiento de fuerza en un intento de mejorar sus medias de bateo. El programa duró 20 semanas y se realizó inmediatamente antes del comienzo de la temporada de béisbol. El número medio de horas semanales y la variación de las medias de bateo con respecto a la temporada anterior son los siguientes:

(8,0, 10)	(20,0, 100)	(5,4, −10)	(12,4, 79)
(9,2, 50)	(15,0, 89)	(6,0, 34)	(8,0, 30)
(18,0, 68)	(25,0, 110)	(10,0, 34)	(5,0, 10)

a) Represente gráficamente los datos. ¿Le parece que tuvo éxito el programa de entrenamiento?
b) Estime la ecuación de regresión.

RESUMEN

El tema de este capítulo son las medidas numéricas que se emplean para describir datos. Hemos descrito la tendencia central por medio de la media, la mediana y la moda y la variabilidad por medio del rango, el rango intercuartílico, la varianza, la desviación típica y el coeficiente de variación. Hemos presentado el teorema de Chebychev, la regla empírica, así como métodos para calcular una proporción aproximada de los datos dentro de un cierto intervalo en torno a la media. Hemos analizado aproximaciones de la media y la varianza de datos agrupados. Por último, hemos introducido brevemente dos números, la covarianza y el coeficiente de correlación, como medidas numéricas de las relaciones entre variables. También hemos analizado el método de regresión por mínimos cuadrados. En el Capítulo 2 presentamos métodos gráficos para describir los datos. En el 3 presentamos métodos numéricos para describirlos.

TÉRMINOS CLAVE

coeficiente de correlación, 70
coeficiente de variación, 61
covarianza, 69
desviación típica, 58
media aritmética, 50
media ponderada, 64
mediana, 50
moda, 50
primer cuartil, 56
rango, 55
rango intercuartílico (*RIC*), 56
regla empírica, 60
resumen de cinco números, 56
sesgado, 52
simetría, 52
tercer cuartil, 56
varianza, 57

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

3.46. Un importante aeropuerto ha contratado recientemente al consultor Juan Cadaqués para estudiar el problema de los retrasos en el tráfico aéreo. Éste ha anotado el número de minutos de retraso de una muestra de vuelos en la siguiente tabla:

Minutos de retaso	0<10	10<20	20<30	30<40	40<50	50<60
Número de vuelos	30	25	13	6	5	4

a) Estime el número medio de minutos de retraso.
b) Estime la varianza y la desviación típica muestrales.

3.47. Snappy Lawn Inc. lleva un registro de lo que cobra por los servicios profesionales de jardinería. El fichero de datos **Snappy** contiene una muestra aleatoria de lo que cobra. Describa los datos numéricamente.

3.48. El fichero de datos **Cotton** contiene información de la producción de fibra de algodón.

a) Represente gráficamente la producción de algodón en relación con el precio al por mayor. Represente una relación lineal aproximada.
b) Halle la constante y la pendiente de la ecuación de regresión. ¿Qué efecto marginal produce cada variación unitaria del precio en la cantidad producida?
c) Estime la relación entre el tejido de algodón exportado y la producción de fibra de algodón.

3.49. Basándose en el fichero de datos **Cotton**,

a) Represente gráficamente la relación entre la producción de algodón y la cantidad exportada de tejido de algodón. Represente una relación lineal aproximada.
b) Calcule la constante y la pendiente de la ecuación de regresión. ¿Qué efecto marginal produce cada variación unitaria de la cantidad de tejido exportado en la cantidad producida?

3.50. ¿Son las notas obtenidas en la prueba de matemáticas del SAT un buen indicador de éxito en la universidad? En el ejemplo 2.8 describimos gráficamente (diagrama de puntos dispersos) las variables de las notas obtenidas en la prueba de matemáticas del SAT y la calificación media obtenida en los estudios universitarios por una muestra aleatoria de 11 estudiantes que termina los estudios universitarios. La tabla siguiente muestra los datos:

Matemáticas SAT	**GPA**
450	3,25
480	2,60
500	2,88
520	2,85
560	3,30
580	3,10
590	3,35
600	3,20
620	3,50
650	3,59
700	3,95

a) Describa el sentido y el grado de relación entre estas dos variables.
b) Calcule e interprete b_1.
c) Calcule b_0.
d) Si la nota obtenida por un estudiante es 530, prediga la calificación media que obtendrá cuando termine los estudios.
e) Basándonos en los datos, ¿podemos predecir la calificación media de un estudiante que obtuvo 375 en la prueba de matemáticas?

3.51. Describa numéricamente los datos siguientes:

(5, 53) (21, 65) (14, 48) (11, 66) (9, 46) (4, 56) (7, 53) (21, 57) (17, 49) (14, 66) (9, 54) (7, 56) (9, 53) (21, 52) (13, 49) (14, 56) (9, 59) (4, 56)

3.52. El fichero de datos **Student GPA** contiene la calificación media obtenida en los estudios universitarios en relación con la nota obtenida en la prueba de lengua del SAT por una muestra aleatoria de 67 estudiantes.

a) Describa gráficamente los datos.
b) Describa numéricamente los datos.
c) Estime la calificación media de un estudiante que obtuvo una nota de 520 en la prueba de lengua.

3.53. Considere las cuatro poblaciones siguientes:

- 1, 2, 3, 4, 5, 6, 7, 8
- 1, 1, 1, 1, 8, 8, 8, 8
- 1, 1, 4, 4, 5, 5, 8, 8
- −6, −3, 0, 3, 6, 9, 12, 15

Todas estas poblaciones tienen la misma media. *Sin hacer los cálculos*, ordene las poblaciones en

función de las magnitudes de sus varianzas, de menor a mayor. A continuación, calcule manualmente cada una de las varianzas.

3.54. Un auditor observa que los valores de las cuentas pendientes de cobro de una empresa tienen una media de 295 $ y una desviación típica de 63 $.

a) Halle un intervalo en el que pueda garantizarse que se encuentra el 60 por ciento de estos valores.

b) Halle un intervalo en el que pueda garantizarse que se encuentra el 84 por ciento de estos valores.

3.55. En un año, el crecimiento de los beneficios de las 500 mayores empresas de Estados Unidos fue, en promedio, de un 9,2 por ciento; la desviación típica fue de 3,5 por ciento.

a) Halle un intervalo en el que pueda garantizarse que se encuentra el 84 por ciento de las cifras de crecimiento de los beneficios.

b) Utilizando la regla empírica, halle un intervalo en el que pueda estimarse que se encuentra aproximadamente el 68 por ciento de estas cifras de crecimiento de los beneficios.

3.56. Los neumáticos de una determinada marca tienen una duración media de 29.000 kilómetros y una desviación típica de 3.000 kilómetros.

a) Halle un intervalo en el que pueda garantizarse que se encuentra el 75 por ciento de las duraciones de los neumáticos de esta marca.

b) Utilizando la regla empírica, halle un intervalo en el que pueda estimarse que se encuentra aproximadamente el 95 por ciento de las duraciones de los neumáticos de esta marca.

Apéndice

1. Media geométrica

Otra medida de la tendencia central que es importante en las empresas y en economía, pero que a menudo se pasa por alto, es la *media geométrica*. Los analistas de empresas y los economistas que tienen interés en saber cuál es el crecimiento en una serie de periodos de tiempo utilizan la media geométrica. Entre las aplicaciones de la media geométrica en las finanzas se encuentran el interés compuesto a lo largo de varios años, el crecimiento de las ventas totales y el crecimiento de la población. Una importante cuestión es el crecimiento anual medio que provoca un cierto crecimiento total en varios años.

Media geométrica

La **media geométrica**, $\bar{x}_g$, es la n-ésima raíz del producto de n números:

$$\bar{x}_g = \sqrt[n]{(x_1 \cdot x_2 \cdot \cdots \cdot x_n)} = (x_1 \cdot x_2 \cdot \cdots \cdot x_n)^{1/n} \qquad \textbf{(3.25)}$$

La media geométrica se utiliza para hallar el crecimiento medio de varios periodos, dado el crecimiento compuesto de cada periodo.

Por ejemplo, la media geométrica de

$$1{,}05 \qquad 1{,}02 \qquad 1{,}10 \qquad 1{,}06$$

es

$$\bar{x}_g = [(1{,}05)(1{,}02)(1{,}10)(1{,}06)]^{1/4} = 1{,}0571$$

EJEMPLO 3.16. Tasa anual de crecimiento (media geométrica)

Halle la tasa anual de crecimiento suponiendo que las ventas han crecido un 25 por ciento en 5 años.

Solución

La tentación intuitiva, pero ingenua, es dividir simplemente el crecimiento total, 25 por ciento, por el número de periodos, 5, y concluir que la tasa anual media de crecimiento es del 5 por ciento. Este resultado es incorrecto porque no tiene en cuenta el efecto compuesto del crecimiento.

Supongamos que la tasa anual de crecimiento es realmente del 5 por ciento; en ese caso, el crecimiento total en 5 años será

$$(1{,}05)(1{,}05)(1{,}05)(1{,}05)(1{,}05) = 1{,}2763$$

o sea, 27,63 por ciento. Sin embargo, la tasa anual de crecimiento, r, que daría un 25 por ciento en 5 años debe satisfacer esta ecuación:

$$(1 + r)^5 = 1{,}25$$

Primero, hallamos la media geométrica:

$$\bar{x}_g = 1 + r = (1{,}25)^{1/5} = 1{,}046$$

La tasa de crecimiento es $r = 0{,}046$, o sea, 4,6 por ciento.

2. Sesgo

Sesgo

El **sesgo** es

$$\text{Sesgo} = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \tag{3.26}$$

La parte importante de esta expresión es el numerador; el denominador tiene por objeto la estandarización, que hace que las unidades de medición sean irrelevantes. El sesgo es positivo si una distribución está sesgada hacia la derecha, ya que las discrepancias medias al cubo en torno a la media son positivas. El sesgo es negativo en el caso de las distribuciones sesgadas hacia la izquierda y 0 en el de las distribuciones, como la distribución normal, que son simétricas en torno a la media.

Bibliografía

1. Bureau of Economic Analysis, http://www.bea.doc.gov/bea/regional/spi/default.cfm, Table SA1-3-Per Capita Personal Income, 28 de mayo de 2004.
2. Bureau of Economic Analysis, http://www.bea.doc.gov/bea/regional/spi/default.cfm, Table SA1-3-Population, 28 de mayo de 2004.
3. Susan G. Komen Breast Cancer Foundation, About Komen, http://www.komen.org, 19 de mayo de 2004.

Capítulo 4

Probabilidad

Esquema del capítulo

Introducción

En este capítulo desarrollamos modelos de probabilidad que pueden utilizarse para estudiar problemas empresariales y económicos cuyos futuros resultados se desconocen.

Consideremos el problema al que se enfrenta Jorge Sánchez, presidente de Desarrollo de Sistemas Avanzados, S.A. (DSA). La empresa ha presentado cinco propuestas de proyectos distintos para el próximo año. Jorge sabe que la empresa tendrá que realizar hasta cinco proyectos el próximo año. Actualmente, el personal de la empresa puede realizar hasta dos y se podría contratar personal para realizar un tercer proyecto. Pero si se adjudican cuatro o cinco proyectos a DSA, tendrá que subcontratar o ampliar significativamente la plantilla. En este capítulo desarrollamos conceptos de probabilidad que puede utilizar Jorge para hallar la ocurrencia probable de los sucesos posibles: la adjudicación de 0, 1, 2, 3, 4 o 5 proyectos. La probabilidad de que ocurra cada suceso es un número comprendido entre 0 y 1, de tal manera que las probabilidades de los seis sucesos suman exactamente 1,0. Cuanto mayor es la probabilidad de que ocurra un suceso, más probable es que ocurra, en comparación con los demás. Si la probabilidad de que se adjudiquen exactamente dos contratos es de 0,80, Jorge estará más seguro de que se producirá ese suceso en comparación con el caso en el que la probabilidad es de 0,20. Pero, en cualquiera de los dos casos, Jorge no puede estar seguro de que ocurrirá el suceso.

Un hospital sabe por experiencia que los sábados por la tarde se registra una media de 1,0 ingresos por hora en la sala de urgencias. La sala de urgencias tiene tres salas de cuidados intensivos. Si se mantiene esta pauta en el futuro, al hospital le gustaría saber cuál es la probabilidad de que sean ingresadas más de tres personas en la sala de urgencias en cualquier hora. Si la probabilidad de que ocurra ese suceso es alta, el hospital necesitará abrir más salas de cuidados intensivos para satisfacer la demanda de los pacientes. Pero si la probabilidad de que haya más de tres ingresos es baja, las caras instalaciones de cuidados intensivos estarán vacías la mayor parte del tiempo, por lo que sería mejor utilizar los recursos para otros fines médicos. Las probabilidades de que ocurran estos sucesos son, pues, muy importantes para decidir el número de salas que deben crearse.

Mostraremos cómo se utilizan modelos de probabilidad para estudiar la variación de los datos observados de manera que puedan hacerse inferencias sobre el proceso subyacente. Nuestro objetivo, tanto en este capítulo como en los dos siguientes, es comprender las probabilidades y cómo pueden hallarse.

4.1. Experimento aleatorio, resultados, sucesos

Para el directivo, la probabilidad de que ocurra un suceso en el futuro presenta un nivel de conocimiento. El directivo podría saber con certeza que el suceso ocurrirá; por ejemplo, habrá un contrato legal. O podría no saber si ocurrirá; por ejemplo, el suceso podría ocurrir o no como parte de una nueva oportunidad empresarial. En la mayoría de las situaciones empresariales, no podemos estar seguros de que ocurrirá un suceso en el futuro, pero si se conoce la probabilidad de que ocurra, tenemos más probabilidades de tomar la mejor decisión posible, en comparación con la situación en la que no conocemos la ocurrencia probable del suceso. Las decisiones y las políticas empresariales a menudo se basan en un conjunto implícito o supuesto de probabilidades.

Para hacer afirmaciones sobre las probabilidades en un entorno incierto, necesitamos desarrollar definiciones y conceptos, como espacio muestral, resultados y sucesos. Éstos son los elementos básicos para definir y calcular probabilidades.

Para nuestro estudio de la probabilidad examinaremos procesos que pueden tener dos resultados o más y existe incertidumbre sobre el resultado que se obtendrá.

Experimento aleatorio

Un **experimento aleatorio** es un proceso que tiene dos o más resultados posibles y existe incertidumbre sobre el resultado que se obtendrá.

Ejemplos de experimentos aleatorios:

1. Se lanza una moneda al aire y el resultado puede ser cara o cruz.
2. En el ejemplo de DSA, la empresa tiene la posibilidad de que le adjudiquen entre 0 y 5 contratos.
3. En una hora se ingresa en la sala de urgencias de un hospital un cierto número de personas.
4. Un cliente entra en una tienda y compra una camisa o no la compra.
5. Se observa la evolución diaria de un índice bursátil.
6. Se selecciona una caja de cereales de una cadena de empaquetado y se pesa para averiguar si el peso es superior o inferior al que viene indicado en la caja.
7. Se lanza al aire un dado de seis lados.

En cada uno de los experimentos aleatorios citados podemos especificar los resultados posibles, que denominamos *resultados básicos*. Por ejemplo, un cliente compra o no una camisa.

Espacio muestral

Los resultados posibles de un experimento aleatorio se llaman **resultados básicos** y el conjunto de todos los resultados básicos se llama **espacio muestral** y se representa por medio del símbolo *S*.

Los resultados básicos deben definirse de tal forma que no puedan ocurrir simultáneamente dos resultados. Además, el experimento aleatorio debe llevar necesariamente a la ocurrencia de uno de los resultados básicos.

EJEMPLO 4.1. Lanzamiento de un dado al aire (espacio muestral)

¿Cuál es el espacio muestral del lanzamiento al aire de un dado de seis caras?

Solución

Los resultados básicos son los seis números posibles y el espacio muestral es

$$S = [1, 2, 3, 4, 5, 6]$$

El espacio muestral contiene seis resultados básicos. No pueden ocurrir dos resultados simultáneamente y debe ocurrir uno de los seis.

EJEMPLO 4.2. Resultados de una inversión (espacio muestral)

Un inversor sigue el índice bursátil Dow-Jones. ¿Cuáles son los resultados básicos posibles al cierre de la sesión?

Solución

El espacio muestral de este experimento es

$$S = [\{1.\text{ El índice será más alto que al cierre de ayer}\},\ \{2.\text{ El índice no será más alto que al cierre de ayer}\}]$$

Debe ocurrir uno de estos dos resultados. No pueden ocurrir simultáneamente. Por lo tanto, los dos resultados constituyen un espacio muestral.

En muchos casos, nos interesa un subconjunto de los resultados básicos y no los resultados por separado. Por ejemplo, en el caso del lanzamiento de un dado al aire, podría interesarnos saber si el resultado es par, es decir, 2, 4 o 6.

Suceso

Un **suceso**, *E*, es cualquier subconjunto de resultados básicos del espacio muestral. Un suceso ocurre si el experimento aleatorio genera uno de los resultados básicos que lo constituyen. El suceso nulo representa la ausencia de un resultado básico y se representa por medio de $\varnothing$.

En algunas aplicaciones, nos interesa la ocurrencia simultánea de dos o más sucesos. Por ejemplo, si se lanza un dado al aire, dos sucesos que podrían considerarse son «el número resultante es par» y «el número resultante es como mínimo un 4». Una posibilidad es que ocurran todos los sucesos de interés. Ocurrirán si el resultado básico del experimento aleatorio pertenece a todos estos sucesos. El conjunto de resultados básicos que pertenecen a todos los sucesos de un grupo de sucesos se denomina *intersección* de estos sucesos. La intersección de los sucesos «el número resultante es par» y «el número resultante es como mínimo un 4» sería que las caras del dado sean iguales a 4 o a 6.

Intersección de sucesos

Sean A y B dos sucesos contenidos en el espacio muestral S. Su **intersección**, representada por $A \cap B$, es el conjunto de todos los resultados básicos en S que pertenecen tanto a A como a B. Por lo tanto, la intersección $A \cap B$ ocurre si y sólo si ocurren tanto A como B. Utilizaremos la expresión **probabilidad conjunta** de A y B para representar la probabilidad de la intersección de A y B.

En términos más generales, dados K sucesos E_1, E_2, ..., E_K, su intersección, $E_1 \cap E_2 \cap \cdots \cap E_K$ es el conjunto de todos los resultados básicos que pertenecen a todos los E_i ($i = 1, 2, ..., K$).

Es posible que la intersección de dos sucesos sea el conjunto vacío.

Mutuamente excluyentes

Si los sucesos A y B no tienen ningún resultado básico común, se llaman **mutuamente excluyentes** y se dice que su intersección, $A \cap B$, es el conjunto vacío que indica que $A \cap B$ no puede ocurrir.

En términos más generales, se dice que los K sucesos E_1, E_2, ..., E_K son mutuamente excluyentes si todo par (E_i, E_j) es un par de sucesos mutuamente excluyentes.

La Figura 4.1 ilustra las intersecciones utilizando un diagrama de Venn. En la parte (a) de la figura, el rectángulo S representa el espacio muestral y los dos círculos representan los sucesos A y B. Los resultados básicos pertenecientes a A están dentro del círculo A y los resultados básicos pertenecientes a B están en el círculo B correspondiente. La intersección de A y B, $A \cap B$, se indica por medio del área sombreada en la que se cortan los círculos. Vemos que un resultado básico pertenece a $A \cap B$ si y sólo si pertenece tanto a A como a B. Así, por ejemplo, cuando se lanza un dado al aire, los resultados 4 y 6 pertenecen ambos a los dos sucesos «sale un número par» y «sale como mínimo un 4». En la Figura 4.1(b), los círculos no se cortan, lo que indica que los sucesos A y B son mutuamente excluyentes. Por ejemplo, si se audita un conjunto de cuentas, los sucesos «menos del 5 por ciento contiene errores importantes» y «más del 10 por ciento contiene errores importantes» son mutuamente excluyentes.

Cuando consideramos conjuntamente varios sucesos, otra posibilidad interesante es que ocurra al menos uno de ellos. Eso sucederá si el resultado básico del experimento aleatorio pertenece al menos a uno de los sucesos. El conjunto de resultados básicos pertenecientes al menos a uno de los sucesos se llama *unión*. Por ejemplo, cuando se lanza un dado al aire, los resultados básicos 2, 4, 5 y 6 pertenecen todos ellos al menos a uno de los sucesos «sale un número par» o «sale un número impar».

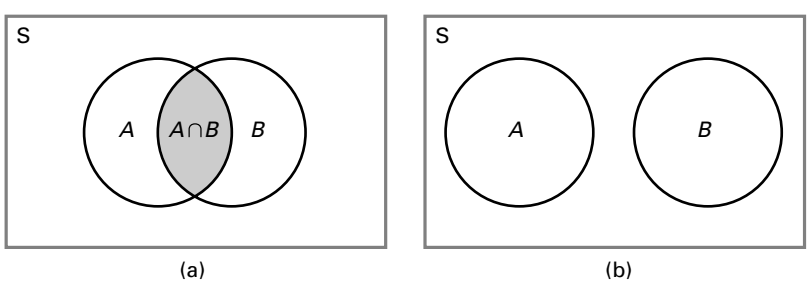


Figura 4.1. Diagramas de Venn de la intersección de los sucesos A y B: (a) $a \cap B$ es el área sombreada; (b) A y B son mutuamente excluyentes.

Unión

Sean A y B dos sucesos contenidos en el espacio muestral, S. Su **unión**, representada por $A \cup B$, es el conjunto de todos los resultados básicos contenidos en S que pertenecen al menos a uno de estos dos sucesos. Por lo tanto, la unión $A \cup B$ ocurre si y sólo si ocurre A o B o ambos.

En términos más generales, dados K sucesos E_1, E_2, ..., E_K, su unión, $E_1 \cup E_2 \cdots \cup E_K$, es el conjunto de todos los resultados básicos pertenecientes al menos a uno de estos K sucesos.

El diagrama de Venn de la Figura 4.2 muestra la unión; se observa claramente que un resultado básico estará en $A \cup B$ si y sólo si está en A o en B o en ambos.

Si la unión de varios sucesos cubre todo el espacio muestral, S, decimos que estos sucesos son *colectivamente exhaustivos*. Dado que todos los resultados básicos están en S, se deduce que todo resultado del experimento aleatorio estará al menos en uno de estos sucesos. Por ejemplo, si se lanza un dado al aire, los sucesos «el resultado es como mínimo un 3» y «el resultado es como máximo un 5» son colectivamente exhaustivos.

Colectivamente exhaustivo

Dados K sucesos E_1, E_2, ..., E_K contenidos en el espacio muestral, S, si $E_1 \cup E_2, \cup \cdots \cup E_K = S$, se dice que estos K sucesos son **colectivamente exhaustivos**.

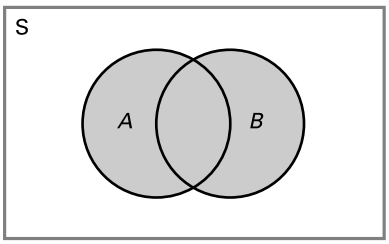


Figura 4.2. Diagrama de Venn de la unión de los sucesos A y B.

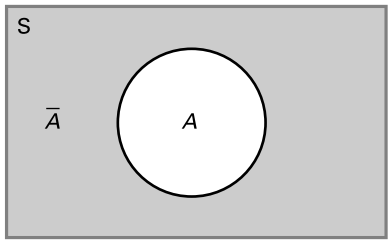


Figura 4.3. Diagrama de Venn del complementario del suceso A.

Podemos ver que el conjunto de todos los resultados básicos contenidos en un espacio muestral es tanto mutuamente excluyente como colectivamente exhaustivo. Ya hemos señalado que estos resultados son tales que debe ocurrir uno, pero no puede ocurrir simultáneamente más de uno.

A continuación, sea A un suceso. Supongamos que nos interesan todos los resultados básicos no incluidos en A.

Complementario

Sea A un suceso contenido en el espacio muestral, S. El conjunto de resultados básicos de un experimento aleatorio perteneciente a S pero no a A se llama **complementario** de A y se representa por medio de $\overline{A}$.

Es evidente que los sucesos A y $\overline{A}$ son mutuamente excluyentes, es decir, ningún resultado básico puede pertenecer a ambos, y colectivamente exhaustivos, es decir, todos los resultados básicos deben pertenecer a uno o al otro. La Figura 4.3 muestra el complementario de A utilizando un diagrama de Venn.

Ya hemos definido tres conceptos importantes —la intersección, la unión y el complementario— que serán importantes en nuestro desarrollo de la probabilidad. Los siguientes ejemplos ayudan a ilustrar estos conceptos.

EJEMPLO 4.3. El lanzamiento de un dado al aire (uniones, intersecciones y complementarios)

Se lanza un dado al aire. Sea A el suceso «el número resultante es par» y B el suceso «el número resultante es como mínimo un 4». En ese caso,

$$A = [2, 4, 6] \quad \text{y} \quad B = [4, 5, 6]$$

Halle el complementario de cada suceso, la intersección y la unión de A y B y la intersección de $\bar{A}$ y B.

Solución

Los complementarios de estos sucesos son, respectivamente,

$$\bar{A} = [1, 3, 5] \quad \text{y} \quad \bar{B} = [1, 2, 3]$$

La intersección de A y B es el suceso «el número resultante es par y como mínimo un 4», por lo que

$$A \cap B = [4, 6]$$

La unión de A y B es el suceso «el número resultante es par o como mínimo un 4 o ambas cosas a la vez» y, por lo tanto,

$$A \cup B = [2, 4, 5, 6]$$

Obsérvese también que los sucesos A y $\bar{A}$ son mutuamente excluyentes, ya que su intersección es el conjunto vacío, y colectivamente exhaustivos, ya que su unión es el espacio muestral S; es decir,

$$A \cup \bar{A} = [1, 2, 3, 4, 5, 6] = S$$

Puede decirse lo mismo de los sucesos B y $\bar{B}$.

Consideremos otra intersección de los sucesos $\bar{A}$ y B. Dado que el único resultado que es «no par» y «como mínimo un 4» es 5, se deduce que $\bar{A} \cap B = [5]$.

EJEMPLO 4.4. Índice bursátil Dow-Jones (uniones, intersecciones y complementarios)

Éstos son cuatro resultados básicos del índice bursátil en 2 días consecutivos:

O_1: El índice sube los dos días.
O_2: El índice sube el primer día, pero no sube el segundo.
O_3: El índice no sube el primer día, pero sube el segundo.
O_4: el índice no sube ninguno de los dos días.

Es evidente que debe ocurrir uno de estos resultados, pero no puede ocurrir más de uno al mismo tiempo. Por lo tanto, podemos representar el espacio muestral de la forma siguiente: $S = [O_1, O_2, O_3, O_4]$. Consideraremos ahora estos dos sucesos:

A: El índice sube el primer día.
B: El índice sube el segundo día.

Halle la intersección, la unión y el complementario de A y B.

Solución

Vemos que A ocurre si ocurre O_1 u O_2 y, por lo tanto,

$$A = [O_1, O_2] \quad \text{y} \quad B = [O_1, O_3]$$

La intersección de A y B es el suceso «el índice sube el primer día y sube el segundo». Éste es el conjunto de todos los resultados básicos pertenecientes tanto a A como a B, $A \cap B = [O_1]$.

La unión de A y B es el suceso «el índice sube como mínimo uno de los días». Éste es el conjunto de todos los resultados pertenecientes a A o a B o a ambos. Por lo tanto,

$$A \cup B = [O_1, O_2, O_3]$$

Por último, el complementario de A es el suceso «el índice no sube el primer día». Éste es el conjunto de todos los resultados básicos contenidos en el espacio muestral, S, que no pertenecen a A. Por lo tanto,

$$\bar{A} = [O_3, O_4] \quad \text{y, asimismo,} \quad \bar{B} = [O_2, O_4]$$

La Figura 4.4 muestra la intersección de los sucesos $\bar{A}$ y B. Esta intersección contiene todos los resultados que pertenecen tanto a $\bar{A}$ como a B. Claramente, $\bar{A} \cap B = [O_3]$.

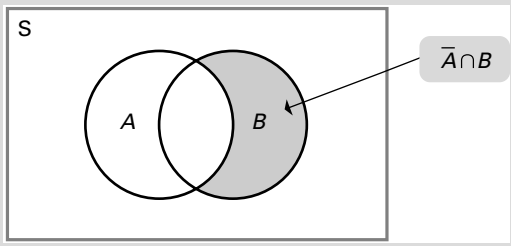


Figura 4.4. Diagrama de Venn de la intersección de $\bar{A}$ y B.

Los diagramas de Venn de las Figuras 4.5, 4.6 y 4.7 muestran tres resultados que implican uniones e intersecciones de sucesos.

Reultado 1

Sean A y B dos sucesos. Los sucesos $A \cap B$ y $\bar{A} \cap B$ son mutuamente excluyentes y su unión es B, como muestra el diagrama de Venn de la Figura 4.5. Claramente,

$$(A \cap B) \cup (\bar{A} \cap B) = B \tag{4.1}$$

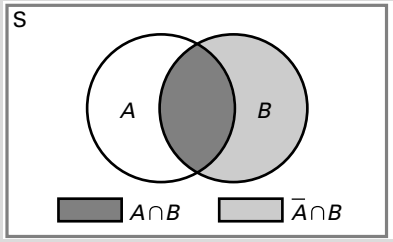


Figura 4.5. Diagrama de Venn del resultado 1: $(A \cap B) \cup (\bar{A} \cap B) = B$.

Resultado 2

Sean A y B dos sucesos. Los sucesos A y $\bar{A} \cap B$ son mutuamente excluyentes y su unión es $A \cap B$, como muestra el diagrama de Venn de la Figura 4.6. Es decir,

$$A \cup (\bar{A} \cap B) = A \cup B \tag{4.2}$$

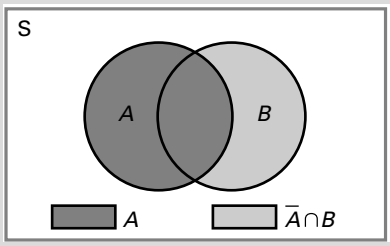


Figura 4.6. Diagrama de Venn del resultado 2: $A \cup (\bar{A} \cap B) = A \cup B$.

Resultado 3

Sean E_1, E_2, ..., E_K K sucesos mutuamente excluyentes y colectivamente exhaustivos y A algún otro suceso. Entonces, los K sucesos $E_1 \cap A$, $E_2 \cap A$, ..., $E_K \cap A$ son mutuamente excluyentes y su unión es A. Es decir,

$$(E_1 \cap A) \cup (E_2 \cap A) \cup \cdots \cup (E_K \cap A) = A \tag{4.3}$$

Podemos comprender mejor la tercera afirmación examinando el diagrama de Venn de la Figura 4.7. El rectángulo grande representa todo el espacio muestral y está dividido en rectángulos más pequeños que representan K sucesos mutuamente excluyentes y colectivamente exhaustivos, E_1, E_2, ..., E_K. El suceso A está representado por la primera fila. Vemos que los sucesos formados por la intersección de A con cada uno de los E sucesos son, de hecho, excluyentes y que su unión es simplemente el suceso A. Por lo tanto, tenemos que

$$(E_1 \cap A) \cup (E_2 \cap A) \cup \cdots \cup (E_K \cap A) = A$$

Figura 4.7. Diagrama de Venn del resultado 3: $(E_1 \cap A) \cup (E_2 \cap A) \cup \cdots \cup (E_K \cap A) = A$.

	E_1	E_2	E_3	E_4	E_5	.....	E_K
A	$E_1 \cap A$	$E_2 \cap A$	$E_3 \cap A$	$E_4 \cap A$	$E_5 \cap A$	.....	$E_K \cap A$
$\bar{A}$							

EJEMPLO 4.5. Lanzamiento de un dado al aire (resultados 1 y 2)

Considere el experimento del lanzamiento de un dado al aire del ejemplo 4.3, donde $A = [2, 4, 6]$ y $B = [4, 5, 6]$. Demuestre lo siguiente:

a) $(A \cap B) \cup (\bar{A} \cap B) = B$

b) $A \cup (\bar{A} \cap B) = A \cup B$

Solución

Sabemos que

$$\bar{A} = [1, 3, 5]$$

Se deduce que

$$A \cap B = [4, 6] \quad \text{y} \quad \bar{A} \cap B = [5]$$

Entonces, $A \cap B$ y $\bar{A} \cap B$ son mutuamente excluyentes y su unión es $B = [4, 5, 6]$; es decir,

$$(A \cap B) \cup (\bar{A} \cap B) = [4, 5, 6] = B \qquad \textbf{(resultado 1)}$$

También, A y $\bar{A} \cap B$ son mutuamente excluyentes y su unión es

$$A \cup (\bar{A} \cap B) = [2, 4, 5, 6] = A \cup B \qquad \textbf{(resultado 2)}$$

EJEMPLO 4.6. Lanzamiento de un dado al aire (resultado 3)

Considere el experimento del lanzamiento de un dado al aire en el que los sucesos A, E_1, E_2 y E_3 vienen dados por

$$A = [2, 4, 6] \quad E_1 = [1, 2] \quad E_2 = [3, 4] \quad E_3 = [5, 6]$$

Demuestre que $E_1 \cap A$, $E_2 \cap A$ y $E_3 \cap A$ son mutuamente excluyentes y que su unión es A.

Solución

En primer lugar, observamos que E_1, E_2 y E_3 son mutuamente excluyentes y colectivamente exhaustivos. Entonces,

$$E_1 \cap A = [2] \quad E_2 \cap A = [4] \quad E_3 \cap A = [6]$$

Claramente, estos tres sucesos son mutuamente excluyentes y su unión es

$$(E_1 \cap A) \cup (E_2 \cap A) \cup (E_3 \cap A) = [2, 4, 6] = A$$

EJERCICIOS

Ejercicios básicos

Para los ejercicios 4.1-4.4 utilice el espacio muestral S definido de la forma siguiente:

$$S = [E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9, E_{10}]$$

4.1. Dado $A = [E_1, E_3, E_6, E_9]$, defina $\bar{A}$.

4.2. Dados $A = [E_1, E_3, E_7, E_9]$ y $B = [E_2, E_3, E_8, E_9]$,

a) ¿Cuál es la intersección de A y B?
b) ¿Cuál es la unión de A y B?
c) ¿Es la unión de A y B colectivamente exhaustiva?

4.3. Dados $\bar{A} = [E_1, E_3, E_7, E_9]$ y $\bar{B} = [E_2, E_3, E_8, E_9]$,

a) ¿Cuál es la intersección de A y B?
b) ¿Cuál es la unión de A y B?
c) ¿Es la unión de A y B colectivamente exhaustiva?

4.4. Dados $A = [E_3, E_5, E_6, E_{10}]$ y $B = [E_3, E_4, E_6, E_9]$,

a) ¿Cuál es la intersección de A y B?
b) ¿Cuál es la unión de A y B?
c) ¿Es la unión de A y B colectivamente exhaustiva?

Ejercicios aplicados

4.5. Una empresa adquiere una nueva máquina que debe instalarse y probarse antes de que esté lista para su uso. La empresa está segura de que no tardará más de 7 días en instalarla y probarla. Sea A el suceso «se necesitarán más de 4 días para que la máquina esté lista» y B el suceso «se necesitarán menos de 6 días para que la máquina esté lista».

a) Describa el suceso que es complementario del suceso A.
b) Describa el suceso que es la intersección de los sucesos A y B.

c) Describa el suceso que es la unión de los sucesos A y B.
d) ¿Son los sucesos A y B mutuamente excluyentes?
e) ¿Son los sucesos A y B colectivamente exhaustivos?
f) Demuestre que $(A \cap B) \cup (\bar{A} \cap B) = B$.
g) Demuestre que $A \cup (\bar{A} \cap B) = A \cup B$.

4.6. Considere el ejemplo 4.4, en el que éstos son cuatro resultados básicos del índice bursátil en 2 días consecutivos:

O_1: El índice sube los dos días.
O_2: El índice sube el primer día, pero no sube el segundo.
O_3: El índice no sube el primer día, pero sube el segundo.
O_4: El índice no sube ninguno de los dos días.

Sean los sucesos A y B los siguientes:

A: El índice sube el primer día.
B: El índice sube el segundo día.

a) Demuestre que $(A \cap B) \cup (\bar{A} \cap B) = B$.
b) Demuestre que $A \cup (\bar{A} \cap B) = A \cup B$.

4.7. Florencio Frentes tiene una pequeña tienda de automóviles usados en la que tiene tres Mercedes (M_1, M_2, M_3) y dos Toyotas (T_1, T_2). Dos clientes, César y Andrés, entran en la tienda y selecciona cada uno un automóvil. Los clientes no se conocen y no hay comunicación entre ellos. Sean A y B los sucesos siguientes:

A: Los clientes seleccionan como mínimo un Toyota.
B: Los clientes seleccionan dos automóviles del mismo modelo.

a) Identifique los pares de automóviles en el espacio muestral.
b) Describa el suceso A.
c) Describa el suceso B.
d) Describa el complementario de A.
e) Demuestre que $(A \cap B) \cup (\bar{A} \cap B) = B$.
f) Demuestre que $A \cup (\bar{A} \cap B) = A \cup B$.

4.2. La probabilidad y sus postulados

Estamos ya en condiciones de utilizar el lenguaje y los conceptos desarrollados en el apartado anterior para averiguar cómo se halla una probabilidad efectiva de que ocurra un proceso. Supongamos que se realiza un experimento aleatorio y que queremos averiguar la probabilidad de que ocurra un determinado suceso. La probabilidad se mide en una escala de 0 a 1. Una probabilidad de 0 indica que el suceso no ocurrirá y una probabilidad de 1 indica que el suceso es seguro que ocurra. Ninguno de estos dos extremos es habitual en los problemas aplicados. Por lo tanto, nos interesa asignar probabilidades comprendidas entre 0 y 1 a los sucesos inciertos. Para ello, es necesario utilizar toda la información de que podamos disponer. Por ejemplo, si las rentas son altas, será más frecuente que se vendan automóviles de lujo. Un director de ventas con experiencia puede ser capaz de saber qué probabilidad tienen las ventas de ser superiores al nivel de rentabilidad que se ha fijado la empresa como objetivo. En este apartado examinamos tres definiciones de probabilidad:

1. Probabilidad clásica.
2. Frecuencia relativa.
3. Probabilidad subjetiva.

Probabilidad clásica

Probabilidad clásica

La **probabilidad clásica** es la proporción de veces que ocurrirá un suceso, suponiendo que todos los resultados contenidos en un espacio muestral tienen la misma probabilidad de ocurrir. La división del número de resultados contenidos en el espacio muestral que satisface el suceso

por el número total de resultados contenidos en el espacio muestral se obtiene la probabilidad de un suceso. La probabilidad de un suceso *A* es

$$P(A) = \frac{N_A}{N} \tag{4.4}$$

donde N_A es el número de resultados que satisfacen la condición del suceso *A* y *N* es el número total de resultados contenidos en el espacio muestral. La idea importante aquí es que se puede hallar una probabilidad a partir de un razonamiento fundamental sobre el proceso.

En el método de la probabilidad clásica, hay que contar los resultados contenidos en el espacio muestral. A continuación, se utiliza el recuento para hallar la probabilidad. El siguiente ejemplo indica cómo puede utilizarse la probabilidad clásica en un problema relativamente sencillo.

EJEMPLO 4.7. Selección de un computador (probabilidad clásica)

Carla Alcántara tiene una pequeña tienda de computadores. Un día tiene tres Gateway y dos Compaq en existencias. Supongamos que entra en la tienda Susana Eslava a comprar dos computadores. A Susana le da igual la marca —todos los computadores tienen las mismas especificaciones técnicas—, por lo que selecciona los computadores puramente al azar: cualquiera de los computadores del estante tiene la misma probabilidad de ser elegido. ¿Cuál es la probabilidad de que Susana compre un Gateway y un Compaq?

Solución

La respuesta puede hallarse utilizando la probabilidad clásica. Primero se define el espacio muestral, que son todos los pares posibles de dos computadores que pueden seleccionarse en la tienda. A continuación, se cuenta el número de pares, que es el número de resultados que satisfacen la condición: un Gateway y un Compaq. Representemos los tres computadores Gateway por medio de G_1, G_2 y G_3 y los dos Compaq por medio de C_1 y C_2. El espacio muestral, *S*, contiene los siguientes pares de computadores:

$$S = \{G_1C_1, G_1C_2, G_2C_1, G_2C_2, G_3C_1, G_3C_2, G_1G_2, G_1G_3, G_2G_3, C_1C_2\}$$

El número de resultados contenidos en el espacio muestral es 10. Si *A* es el suceso «se elige un Gateway y un Compaq», el número, N_A, de resultados que tienen un Gateway y un Compaq es 6. Por lo tanto, la probabilidad de que ocurra el suceso *A* —un Gateway y un Compaq— es

$$P(A) = \frac{N_A}{N} = \frac{6}{10} = 0{,}6$$

El recuento de todos los resultados llevaría mucho tiempo si tuviéramos que identificar primero todos los resultados posibles. Sin embargo, muchos de los lectores habrán aprendido en cursos anteriores la fórmula básica para calcular *el número de combinaciones* de *n* objetos que se toman *k* de cada vez.

Fórmula para hallar el número de combinaciones

El proceso de recuento puede generalizarse utilizando la siguiente ecuación para calcular el **número de combinaciones** de *n* objetos que se toman *k* de cada vez:

$$C_k^n = \frac{n!}{k!(n-k)!} \qquad 0! = 1 \tag{4.5}$$

En el apéndice que se encuentra al final de este capítulo se desarrollan combinaciones; el lector debe estudiarlo si necesita aprender o repasar las combinaciones.

Ilustramos la ecuación de combinación, la ecuación 4.5, señalando que en el ejemplo 4.7 el número de combinaciones de los cinco computadores que se toman dos de cada vez es el número de elementos contenidos en el espacio muestral:

$$C_2^5 = \frac{5!}{2!(5-2)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1(3 \cdot 2 \cdot 1)} = 10$$

En el ejemplo 4.8, aplicamos la probabilidad clásica a un problema más difícil.

EJEMPLO 4.8. Reconsideración de la selección de un computador (probabilidad clásica)

Supongamos que ahora en la tienda de Carla hay 10 computadores Gateway, 5 Compaq y 5 Acer. Susana entra en la tienda y quiere comprar 3. Los selecciona puramente al azar. ¿Cuál es ahora la probabilidad de que seleccione 2 Gateway y 1 Compaq?

Solución

Utilizaremos la definición clásica de probabilidad. Pero en este ejemplo utilizaremos la fórmula de las combinaciones para hallar el número de resultados contenidos en el espacio muestral y el número de resultados que satisfacen la condición *A*: [2 Gateways y 1 Compaq].

El número total de resultados contenidos en el espacio muestral es

$$N = C_3^{20} = \frac{20!}{3!(20-3)!} = 1,\,140$$

El número de formas en que podemos seleccionar 2 computadores Gateway de los 10 que hay se calcula de la forma siguiente:

$$C_2^{10} = \frac{10!}{2!(10-2)!} = 45$$

Asimismo, el número de formas en que podemos seleccionar 1 computador Compaq de los 5 que hay se calcula de la forma siguiente:

$$C_1^5 = \frac{5!}{1!(5-1)!} = 5$$

Por lo tanto, el número de resultados que satisfacen el suceso A es

$$N_A = C_2^{10} \times C_1^5 = 45 \times 5 = 225$$

Por último, la probabilidad de A = [2 Gateways y 1 Compaq] es

$$P_A = \frac{N_A}{N} = \frac{C_2^{10} \times C_1^5}{C_3^{20}} = \frac{45 \times 5}{1,\ 140} = 0{,}197$$

Frecuencia relativa

A menudo utilizamos la frecuencia relativa para hallar las probabilidades de una determinada población. La *frecuencia relativa* es el número de sucesos contenidos en la población que satisfacen la condición dividido por el número total de sucesos. Estas probabilidades indican la frecuencia con que ocurrirá un suceso en comparación con otros. Por ejemplo, si el suceso A tiene una probabilidad de 0,40, sabemos que ocurrirá el 40 por ciento de las veces. Es más frecuente que el suceso B si el suceso B sólo tiene una probabilidad de 0,30 de ocurrir. Pero no sabemos qué suceso, el A o el B, ocurrirá a continuación.

Frecuencia relativa

La **frecuencia relativa** es el límite de la proporción de veces que ocurre el suceso A en un gran número de pruebas, n:

$$P(A) = \frac{n_A}{n} \qquad \textbf{(4.6)}$$

donde n_A es el número de veces que se obtiene A y n es el número total de pruebas o resultados. La probabilidad es el límite a medida que n se hace más grande (o tiende a infinito).

EJEMPLO 4.9. Probabilidad de que las rentas sean de más de 50.000 $ (probabilidad relativa)

Sara Olmedo está considerando la posibilidad de abrir un nuevo concesionario de automóviles en una ciudad que tiene una población de 150.000 habitantes. La experiencia de otros muchos concesionarios indica que en ciudades parecidas un concesionario tiene éxito si al menos el 40 por ciento de los hogares tiene una renta anual de más de 50.000 $. Ha pedido a Pablo Sánchez, consultor de marketing, que estime la proporción de rentas familiares de más de 50.000 $, o sea, la probabilidad de esas rentas.

Solución

Después de examinar el problema, Pablo llega a la conclusión de que la probabilidad debe basarse en la frecuencia relativa. Primero examina los datos censales más recientes y observa que en la ciudad había 54.345 hogares y que 31.496 tenían una renta de más de 50.000 $. Pablo calcula la probabilidad del suceso A, «renta familiar de más de 50.000 $», de la forma siguiente:

$$P(A) = \frac{n_A}{n} = \frac{31.496}{54.345} = 0{,}580$$

Como Pablo sabe que hay varios errores en los datos censales, también consulta datos similares publicados en una revista del sector. Basándose en esta fuente, obtiene 55.100 hogares, de los que 32.047 tienen una renta de más de 50.000 $. Pablo calcula la probabilidad del suceso A a partir de esta fuente de la forma siguiente:

$$P(A) = \frac{n_A}{n} = \frac{32.047}{55.100} = 0{,}582$$

Como estas cifras son parecidas, podría dar las dos. Pablo decide dar una probabilidad de 0,58.

Este ejemplo muestra que las probabilidades basadas en el enfoque de la frecuencia relativa a menudo pueden obtenerse utilizando las fuentes de datos existentes. También indica que pueden ocurrir y ocurren diferentes resultados y que los analistas y los directivos con experiencia tratarán de verificar sus resultados utilizando más de una fuente. Se necesita experiencia y mucho criterio para decidir si los diferentes datos son suficientemente parecidos.

Probabilidad subjetiva

Probabilidad subjetiva

La **probabilidad subjetiva** expresa el grado en que una persona cree que ocurrirá un suceso. Estas probabilidades subjetivas se utilizan en algunos procedimientos empresariales de toma de decisiones.

Podemos comprender el concepto de probabilidad subjetiva utilizando el concepto de apuestas justas. Por ejemplo, si afirmo que la probabilidad de que suba el precio de las acciones de una empresa la próxima semana es 0,5, creo que el precio de las acciones tiene tantas probabilidades de subir como de bajar. Cuando expreso esta probabilidad subjetiva, no estoy pensando necesariamente en un experimento repetido sino en el precio que tendrán las acciones la próxima semana. La probabilidad subjetiva que expreso implica que consideraría justa una apuesta en la que hay que pagar 1 $ si el precio baja y se recibe 1 $ si el precio sube. Si recibiera más de 1 $ por una subida del precio, consideraría que la apuesta me favorece. Asimismo, si creo que la probabilidad de que un caballo gane una carrera es 0,4, estoy expresando mi opinión personal de que hay una posibilidad del 40 por ciento de que gane. Dada esta creencia, consideraría justa una apuesta en la que recibiera 3 $ si el caballo ganara y perdiera 2 $ si el caballo perdiera.

Queremos hacer hincapié en que las probabilidades subjetivas son personales. No es necesario que todo el mundo piense que un suceso tiene las mismas probabilidades. En el ejemplo del precio de las acciones, la mayoría de la gente llegaría a la conclusión de que la probabilidad correcta de que suban las acciones es 0,50. Sin embargo, una persona que tenga más información sobre las acciones podría creer otra cosa. En el ejemplo de la carrera de caballos, es probable que dos apostantes tengan probabilidades subjetivas diferentes. Pueden no tener la misma información y, aunque la tengan, pueden interpretarla de manera distinta. Sabemos que los inversores no tienen todos ellos las mismas opiniones sobre la futura conducta del mercado de valores. Cabría pensar que sus probabilidades subjetivas

dependen de la información que tienen y del modo en que la interpretan. Los directivos de diferentes empresas tienen probabilidades subjetivas diferentes sobre las oportunidades de ventas en un mercado regional y, por lo tanto, toman decisiones diferentes.

Postulados probabilísticos

Necesitamos desarrollar un marco para evaluar y manipular las probabilidades. Para ello, primero formularemos tres reglas (o postulados) que deben cumplir las probabilidades y demostraremos que estos requisitos son «razonables».

Postulados probabilísticos

Sea S el espacio muestral de un experimento aleatorio, O_i los resultados básicos y A un suceso. Para cada suceso A del espacio muestral, S, suponemos que se define $P(A)$ y tenemos los siguientes **postulados probabilísticos**:

1. Si A es cualquier suceso del espacio muestral, S,
$$0 \leqslant P(A) \leqslant 1$$
2. Sea A un suceso de S y sea O_i los resultados básicos. Entonces,
$$P(A) = \sum_{A} P(O_i)$$
donde la notación implica que el sumatorio abarca todos los resultados básicos contenidos en A.
3. $P(S) = 1$.

El primer postulado requiere que la probabilidad se encuentre entre 0 y 1. El segundo puede comprenderse por medio de las frecuencias relativas. Supongamos que un experimento aleatorio se repite N veces. Sea N_i el número de veces que ocurre el resultado básico O_i y N_A el número de veces que ocurre el suceso A. Entonces, dado que los resultados básicos son mutuamente excluyentes, N_A es simplemente la suma de N_i correspondiente a todos los resultados básicos contenidos en A; es decir,

$$N_A = \sum_{A} N_i$$

y dividiendo por el número de pruebas, N, obtenemos

$$\frac{N_A}{N} = \sum_{A} \frac{N_i}{N}$$

Pero según el concepto de frecuenta relativa, N_A/N tiende a $P(A)$ y cada N_i/N tiende a $P(O_i)$ a medida que N se hace infinitamente grande. Por lo tanto, el segundo postulado puede considerarse un requisito lógico cuando la probabilidad se ve de esta forma.

El tercer postulado puede parafrasearse de la siguiente manera: «cuando se realiza un experimento aleatorio, algo tiene que ocurrir». Sustituyendo A por el espacio muestral, S, en el segundo postulado, tenemos que

$$P(S) = \sum_{S} P(O_i)$$

donde el sumatorio abarca todos los resultados básicos del espacio muestral. Pero como $P(S) = 1$ según el tercer postulado, se deduce que

$$\sum_{S} P(O_i) = 1$$

Es decir, la suma de las probabilidades de todos los resultados básicos del espacio muestral es 1.

Consecuencias de los postulados

A continuación, enumeramos e ilustramos algunas consecuencias inmediatas de los tres postulados.

1. Si el espacio muestral, S, está formado por n resultados básicos igualmente probables, E_1, E_2, ..., E_n, entonces

$$P(E_i) = \frac{1}{n} \qquad i = 1, 2, ..., n$$

ya que los n resultados cubren el espacio muestral y son igualmente probables. Por ejemplo, si se lanza al aire un dado equilibrado, la probabilidad de que salga cada uno de los seis resultados básicos es 1/6.

2. Si el espacio muestral, S, está formado por n resultados básicos igualmente probables y el suceso A está formado por n_A de estos resultados, entonces

$$P(A) = \frac{n_A}{n}$$

Este resultado se deduce de la consecuencia 1 y el postulado 2. Todo resultado básico tiene la probabilidad $1/n$ y, por el postulado 2, $P(A)$ es simplemente la suma de las probabilidades de los n_A resultados básicos de A. Por ejemplo, si se lanza al aire un dado equilibrado y A es el suceso «sale un número par», hay $n = 6$ resultados básicos y $n_A = 3$ de ellos se encuentran en A. Por lo tanto, $P(A) = 3/6 = 1/2$.

3. Sean A y B sucesos mutuamente excluyentes. En ese caso, la probabilidad de su unión es la suma de sus probabilidades individuales; es decir,

$$P(A \cup B) = P(A) + P(B)$$

En general, si E_1, E_2, ..., E_K son sucesos mutuamente excluyentes,

$$P(E_1 \cup E_2 \cup \cdots \cup E_K) = P(E_1) + P(E_2) + \cdots + P(E_K)$$

Este resultado es una consecuencia del postulado 2. La probabilidad de la unión de A y B es

$$P(A \cup B) = \sum_{A \cup B} P(O_i)$$

donde el sumatorio abarca todos los resultados básicos de $A \cup B$. Pero, dado que A y B son mutuamente excluyentes, ningún resultado básico pertenece a ambos, por lo que

$$\sum_{A \cup B} P(O_i) = \sum_{A} P(O_i) + \sum_{B} P(O_i) = P(A) + P(B)$$

4. Si E_1, E_2, ..., E_K son sucesos colectivamente exhaustivos, la probabilidad de su unión es

$$P(E_1 \cup E_2 \cup \cdots \cup E_K) = 1$$

Dado que los sucesos son colectivamente exhaustivos, su unión es todo el espacio muestral, S, y el resultado se deduce del postulado 3.

EJEMPLO 4.10. Lotería (probabilidad)

Una organización benéfica vende 1.000 billetes de lotería. Hay 10 premios grandes y 100 premios pequeños y todos deben repartirse. El proceso de selección de los ganadores es tal que al principio todos los billetes tienen las mismas probabilidades de ganar un premio grande y todos tienen las mismas probabilidades de ganar un premio pequeño. Ninguno puede ganar más de un premio. ¿Cuál es la probabilidad de ganar un premio grande con un único billete? ¿Cuál es la probabilidad de ganar un premio pequeño? ¿Cuál es la probabilidad de ganar algún premio?

Solución

De los 1.000 billetes, 10 ganarán premios grandes 100 ganarán premios pequeños y 890 no ganarán ningún premio. Nuestro único billete es seleccionado de entre 1.000. Sea A el suceso «el billete seleccionado gana un premio grande» y B el suceso «el billete seleccionado gana un premio pequeño». Las probabilidades son

$$P(A) = \frac{10}{1.000} = 0{,}01$$

$$P(B) = \frac{100}{1.000} = 0{,}10$$

El suceso «el billete gana algún premio» es la unión de los sucesos A y B. Como sólo se permite un premio, estos sucesos son mutuamente excluyentes y

$$P(A \cup B) = P(A) + P(B) = 0{,}01 + 0{,}10 = 0{,}11$$

EJEMPLO 4.11. Reconsideración del índice bursátil Dow-Jones (probabilidad)

En el ejemplo 4.4, hemos examinado la evolución del índice bursátil Dow-Jones en 2 días y hemos definido cuatro resultados básicos:

O_1: El índice sube los dos días.
O_2: El índice sube el primer día, pero no sube el segundo.
O_3: El índice no sube el primer día, pero sube el segundo.
O_4: El índice no sube ninguno de los dos días.

Suponga que estos cuatro resultados básicos son igual de probables. En ese caso, ¿cuál es la probabilidad de que el mercado suba como mínimo 1 de los 2 días?

Solución

El suceso que nos interesa, «el mercado sube como mínimo 1 de los 2 días», contiene tres de los cuatro resultados básicos, O_1, O_2 y O_3. Como los resultados básicos son todos igual de probables, se deduce que la probabilidad de este suceso es 3/4, o sea, 0,75.

EJEMPLO 4.12. Prospecciones petroleras (probabilidad)

En las primeras fases del desarrollo de una plataforma petrolera en el océano Atlántico, una empresa petrolera estimó que había una probabilidad de 0,1 de que las reservas económicamente recuperables superaran los 2.000 millones de barriles. La probabilidad de que superaran los 1.000 millones se estimó en 0,5. Dada esta información, ¿cuál es la probabilidad estimada de que las reservas se encuentren entre 1.000 y 2.000 millones de barriles?

Solución

Sea A el suceso «las reservas superan los 2.000 millones de barriles» y B el suceso «las reservas se encuentran entre 1.000 y 2.000 millones de barriles». Éstos son mutuamente excluyentes y su unión, $A \cup B$, es el suceso «las reservas superan los 1.000 millones de barriles». Por lo tanto, tenemos que

$$P(A) = 0{,}1 \qquad P(A \cup B) = 0{,}5$$

Entonces, dado que A y B son mutuamente excluyentes,

$$P(B) = P(A \cup B) - P(A) = 0{,}5 - 0{,}1 = 0{,}4$$

EJERCICIOS

Ejercicios básicos

4.8. El espacio muestral contiene 5 A y 7 B. ¿Cuál es la probabilidad de que un conjunto de 2 seleccionado aleatoriamente contenga 1 A y 1 B?

4.9. El espacio muestral contiene 6 A y 4 B. ¿Cuál es la probabilidad de que un conjunto de 3 seleccionado aleatoriamente contenga 1 A y 2 B?

4.10. El espacio muestral contiene 10 A y 6 B. ¿Cuál es la probabilidad de que un conjunto de 4 seleccionado aleatoriamente contenga 2 A y 2 B?

4.11. En una ciudad de 120.000 personas hay 20.000 noruegos. ¿Cuál es la probabilidad de que una persona de la ciudad seleccionada aleatoriamente sea noruega?

4.12. En una ciudad de 180.000 personas hay 20.000 noruegos. ¿Cuál es la probabilidad de que una muestra aleatoria de 2 personas de la ciudad contenga 2 noruegos?

Ejercicios aplicados

4.13. Recuerde la empresa del ejercicio 4.5. Su nueva máquina debe instalarse y probarse antes de que esté lista para funcionar. La tabla adjunta muestra la valoración del directivo de la probabilidad del número de días necesarios para que la máquina esté lista para usarla.

Número de días	3	4	5	6	7
Probabilidad	0,08	0,24	0,41	0,20	0,07

Sea A el suceso «se necesitarán más de 4 días para que la máquina esté lista para funcionar» y sea B el suceso «se necesitarán menos de 6 días para que la máquina esté lista para funcionar».

a) Halle la probabilidad del suceso A.

b) Halle la probabilidad del suceso B.

c) Halle la probabilidad del complementario del suceso *A*.
d) Halle la probabilidad de la intersección de los sucesos *A* y *B*.
e) Halle la probabilidad de la unión de los sucesos *A* y *B*.

4.14. El gestor de un fondo está considerando la posibilidad de invertir en las acciones de una compañía de asistencia sanitaria. La tabla adjunta resume su valoración de las probabilidades de las tasas de rendimiento de estas acciones durante el próximo año. Sea *A* el suceso «la tasa de rendimiento será de más del 10 por ciento» y *B* el suceso «la tasa de rendimiento será negativa».

Tasa de rendimiento	Menos de 10%	Entre −10% y 0%	Entre 0% y 10%	Entre 10% y 20%	Más de 20%
Probabilidad	0,04	0,14	0,28	0,33	0,21

a) Halle la probabilidad del suceso *A*.
b) Halle la probabilidad del suceso *B*.
c) Describa el suceso que es el complementario de *A*.
d) Halle la probabilidad del complementario de *A*.
e) Describa el suceso que es la intersección de *A* y *B*.
f) Halle la probabilidad de la intersección de *A* y *B*.
g) Describa el suceso que es la unión de *A* y *B*.
h) Halle la probabilidad de la unión de *A* y *B*.
i) ¿Son *A* y *B* mutuamente excluyentes?
j) ¿Son *A* y *B* colectivamente exhaustivos?

4.15. Un directivo tiene ocho empleados que podría asignar a la tarea de supervisar un proyecto. Cuatro son mujeres y cuatro son hombres. Dos de los hombres son hermanos. El directivo va a asignar la tarea aleatoriamente, por lo que los ocho empleados tienen las mismas probabilidades de ser elegidos. Sea *A* el suceso «el empleado elegido es un hombre» y *B* el suceso «el empleado elegido es uno de los hermanos».

a) Halle la probabilidad del suceso *A*.
b) Halle la probabilidad del suceso *B*.
c) Halle la probabilidad de la intersección de *A* y *B*.

4.16. Si dos sucesos son mutuamente excluyentes, sabemos que la probabilidad de su unión es la suma de sus probabilidades individuales. Sin embargo, *no* es así en el caso de los sucesos que no son mutuamente excluyentes. Verifique esta afirmación considerando los sucesos *A* y *B* del ejercicio 4.2.

4.17. El director de unos grandes almacenes ha examinado el número de reclamaciones que se reciben semanalmente por la mala calidad del servicio. La tabla adjunta muestra las probabilidades de los números de quejas semanales obtenidas en este examen. Sea *A* el suceso «habrá como mínimo una reclamación a la semana» y *B* el suceso «habrá menos de 10 reclamaciones a la semana».

Número de reclamaciones	0	Entre 1 y 3	Entre 4 y 6	Entre 7 y 9	Entre 10 y 12	Más de 12
Probabilidad	0,14	0,39	0,23	0,15	0,06	0,03

a) Halle la probabilidad de *A*.
b) Halle la probabilidad de *B*.
c) Halle la probabilidad del complementario de *A*.
d) Halle la probabilidad de la unión de *A* y *B*.
e) Halle la probabilidad de la intersección de *A* y *B*.
f) ¿Son *A* y *B* mutuamente excluyentes?
g) ¿Son *A* y *B* colectivamente exhaustivos?

4.18. Una empresa recibe una pieza en envíos de 100. Según un estudio, las probabilidades del número de piezas defectuosas que hay en un envío son las que se muestran en la tabla adjunta.

Número defectuoso	0	1	2	3	Más de 3
Probabilidad	0,29	0,36	0,22	0,10	0,03

a) ¿Cuál es la probabilidad de que haya menos de 3 piezas defectuosas en un envío?
b) ¿Cuál es la probabilidad de que haya más de 1 pieza defectuosa en un envío?
c) Las cinco probabilidades de la tabla suman 1. ¿Por qué debe ser así?

4.3. Reglas de la probabilidad

A continuación presentamos algunas reglas importantes para calcular las probabilidades de sucesos compuestos. Comenzamos definiendo A como un suceso contenido en el espacio muestral, S, por lo que A y su complementario, $\bar{A}$, son mutuamente excluyentes y colectivamente exhaustivos.

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1$$

Ésta es la *regla del complementario*.

Regla del complementario

Sea A un suceso y $\bar{A}$ su complementario. La **regla del complementario** es

$$P(\bar{A}) = 1 - P(A) \qquad \textbf{(4.7)}$$

Por ejemplo, cuando se lanza un dado al aire, la probabilidad de que salga un 1 es 1/6 y, por lo tanto, según la regla del complementario, la probabilidad de no salga un 1 es 5/6. Este resultado es importante porque en algunos problemas puede ser más fácil hallar $P(\bar{A})$ y hallar después $P(A)$, como se observa en el ejemplo 4.13.

EJEMPLO 4.13. Selección de personal (regla del complementario)

Una empresa está contratando directivos para cubrir cuatro puestos clave. Los candidatos son cinco hombres y tres mujeres. Suponiendo que todas las combinaciones de hombres y mujeres tienen las mismas probabilidades de ser seleccionadas, ¿cuál es la probabilidad de que se seleccione como mínimo una mujer?

Solución

Resolveremos este problema calculando primero la probabilidad del complementario de A, «no se selecciona ninguna mujer» y utilizando a continuación la regla del complementario para calcular la probabilidad de A, «se selecciona como mínimo una mujer». Esta probabilidad es más fácil de calcular que las probabilidades de que se seleccione entre una y tres mujeres. Utilizando el método de la probabilidad clásica,

$$P(\bar{A}) = \frac{C_4^5}{C_4^8} = \frac{1}{14}$$

y, por lo tanto, la probabilidad es

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{1}{14} = \frac{13}{14}$$

Antes hemos demostrado que, si dos sucesos son mutuamente excluyentes, la probabilidad de su unión es la suma de las probabilidades de cada suceso:

$$P(A \cup B) = P(A) + P(B)$$

A continuación, queremos averiguar el resultado cuando los sucesos A y B no son mutuamente excluyentes. En el apartado 4.1 hemos señalado que los sucesos A y $\bar{A} \cap B$ son mutuamente excluyentes —repase el lector el resultado 2 y la Figura 4.6— y, por lo tanto,

$$P(A \cup B) = P(A) + P(\bar{A} \cap B)$$

Además, los sucesos $A \cap B$ y $\bar{A} \cap B$ son mutuamente excluyentes y su unión es B (repase el lector el resultado 1 y la Figura 4.5):

$$P(B) = P(A \cap B) \cup P(\bar{A} \cap B)$$

A partir de esta expresión, obtenemos el resultado

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

Combinando estos dos resultados, tenemos la *regla de la suma de probabilidades*.

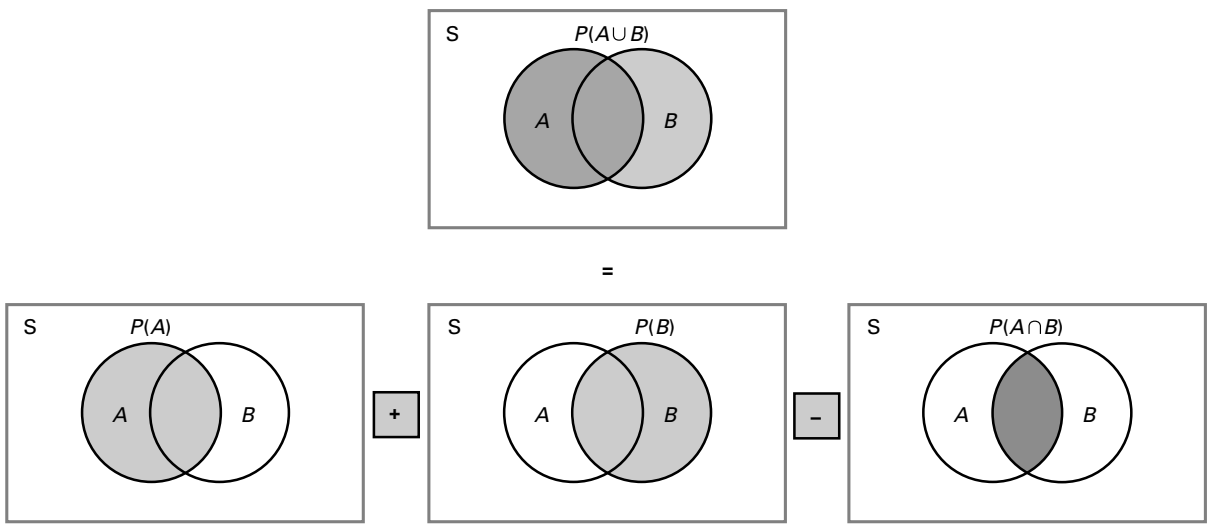


Figura 4.8. Diagrama de Venn de la regla de la suma: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

La regla de la suma de probabilidades

Sean A y B dos sucesos. Utilizando la **regla de la suma de probabilidades**, la probabilidad de su unión es

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \textbf{(4.8)}$$

El diagrama de Venn de la Figura 4.8 permite comprender intuitivamente la regla de la suma. El rectángulo mayor, S, representa todo el espacio muestral. Los círculos más pequeños, A y B, representan los sucesos A y B. Podemos ver que el área en la que A y B se solapan representan la intersección de las dos probabilidades, $P(A \cap B)$. Para calcular la probabilidad de la unión de los sucesos A y B, primero sumamos las probabilidades de los sucesos, $P(A) + P(B)$. Obsérvese, sin embargo, que la probabilidad de la intersección, $P(A \cap B)$, se contabiliza dos veces y, por lo tanto, debe restarse una vez.

EJEMPLO 4.14. Selección de productos (regla de la suma)

Una cadena de hamburgueserías observó que el 75 por ciento de todos los clientes consume mostaza, el 80 por ciento consume ketchup y el 65 por ciento consume los dos. ¿Cuál es la probabilidad de que un cliente consuma al menos uno de los dos?

Solución

Sea A el suceso «el cliente consume mostaza» y B el suceso «el cliente consume ketchup». Por lo tanto, tenemos que

$$P(A) = 0{,}75 \qquad P(B) = 0{,}80 \qquad \text{y} \qquad P(A \cap B) = 0{,}65$$

La probabilidad es

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0{,}75 + 0{,}80 - 0{,}65 = 0{,}90 \end{aligned}$$

Probabilidad condicionada

Consideremos un par de sucesos, A y B. Supongamos que nos interesa saber cuál es la probabilidad de A, dado que ha ocurrido B. Este problema puede analizarse por medio del concepto de *probabilidad condicionada*. La idea básica es que la probabilidad de que ocurra cualquier suceso a menudo depende de que hayan ocurrido o no otros sucesos. Por ejemplo, un fabricante que está considerando la posibilidad de introducir una nueva marca puede hacer una prueba ofreciendo el producto en unas cuantas tiendas. Este fabricante estará mucho más seguro del éxito de la marca en el mercado en general si tiene una buena acogida en esas cuantas tiendas que en caso contrario. La valoración de la empresa de la probabilidad de que las ventas sean altas dependerá, pues, del resultado obtenido en esas cuantas tiendas.

Si supiéramos que los tipos de interés van a bajar el año que viene, seríamos más optimistas sobre la bolsa de valores que si creyéramos que van a subir. Lo que sabemos o creemos sobre los tipos de interés condiciona nuestra valoración de la probabilidad de la evolución de los precios de las acciones. A continuación, formulamos en términos formales la probabilidad condicionada que puede utilizarse para averiguar cómo afectan los resultados anteriores a la probabilidad.

Probabilidad condicionada

Sean A y B dos sucesos. La **probabilidad condicionada** del suceso A, dado que ha ocurrido el suceso B, se representa por medio del símbolo $P(A|B)$ y es

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{siempre que } P(B) > 0 \qquad \textbf{(4.9)}$$

Asimismo,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{siempre que } P(A) > 0$$

Podemos comprender mejor estos resultados y los siguientes examinando la Tabla 4.1. La probabilidad condicionada, $P(A|B)$, es el cociente entre la probabilidad conjunta, $P(A \cap B)$ y la probabilidad de la variable condicionada, $P(B)$. Podemos imaginar que esta probabilidad

Tabla 4.1. Probabilidad conjunta de A y B.

	A	$\bar{A}$	
B	$P(A \cap B)$	$P(\bar{A} \cap B)$	$P(B)$
$\bar{B}$	$P(A \cap \bar{B})$	$P(\bar{A} \cap \bar{B})$	$P(\bar{B})$
	$P(A)$	$P(\bar{A})$	1,0

condicionada equivale a utilizar solamente la primera fila de la tabla que se refiere a la condición B. Podría realizarse un análisis similar con la probabilidad condicionada $P(B|A)$.

Las frecuencias relativas también pueden ayudarnos a comprender la probabilidad condicionada. Supongamos que repetimos un experimento aleatorio n veces y que hay n_B ocurrencias del suceso B y $n_{A\cap B}$ ocurrencias de A y B juntos. En ese caso, la proporción de veces que ocurre A, cuando ha ocurrido B, es $n_{A\cap B}/n_B$, y se puede concebir la probabilidad condicionada de A, dado B, como el límite de esta proporción cuando el número de repeticiones del experimento se vuelve infinitamente grande:

$$\frac{n_{A\cap B}}{n_B} = \frac{n_{A\cap B}/n}{n_B/n}$$

y entonces, a medida que n se hace grande, el numerador y el denominador del segundo miembro de esta expresión tienden a $P(A \cap B)$ y $P(B)$, respectivamente.

EJEMPLO 4.15. Elección de productos: ketchup y mostaza (probabilidad condicionada)

En el ejemplo 4.14 hemos señalado que el 75 por ciento de los clientes de la cadena consume mostaza, el 80 por ciento consume ketchup y el 65 por ciento consume los dos. ¿Cuáles son las probabilidades de que un consumidor de ketchup utilice mostaza y de que un consumidor de mostaza utilice ketchup?

Solución

En el ejemplo 4.14 hemos visto que $P(A) = 0{,}75$, $P(B) = 0{,}80$ y $P(A \cap B) = 0{,}65$. La probabilidad de que un consumidor de ketchup utilice mostaza es la probabilidad condicionada del suceso A, dado el suceso B.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}65}{0{,}80} = 0{,}8125$$

De la misma forma, la probabilidad de que un consumidor de mostaza utilice ketchup es

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0{,}65}{0{,}75} = 0{,}8667$$

Estos cálculos también pueden realizarse utilizando la Tabla 4.2, que tiene un formato parecido al de la 4.1. Obsérvese que la probabilidad condicionada de que un consumidor de ketchup también utilice mostaza es la probabilidad conjunta, 0,65, dividida por la probabilidad de un consumidor de ketchup, 0,80. La otra probabilidad condicionada puede hallarse realizando un cálculo similar. Hemos observado que muchas

Tabla 4.2. Probabilidad conjunta de la mostaza y el ketchup del ejemplo 4.15.

	Mostaza	No mostaza	
Ketchup	0,65	0,15	0,80
No ketchup	0,10	0,10	0,20
	0,75	0,25	1,0

personas piensan que la utilización de una tabla como la 4.2 las anima más y les permite resolver mejor la probabilidad condicionada y los problemas parecidos siguientes. Utilizando correctamente la tabla, se obtienen exactamente los mismos resultados que utilizando ecuaciones. El lector puede sentirse absolutamente cómodo utilizando tablas para resolver los problemas.

Una consecuencia inmediata de la probabilidad condicionada es la *regla del producto de probabilidades*, que expresa la probabilidad de una intersección por medio de las probabilidades de sucesos individuales y las probabilidades condicionadas.

La regla del producto de probabilidades

Sean A y B dos sucesos. Utilizando la **regla del producto de probabilidades**, la probabilidad de su intersección puede deducirse de la probabilidad condicionada de la forma siguiente:

$$PA \cap B) = P(A|B)P(B) \qquad \textbf{(4.10)}$$

También,

$$P(A \cap B) = P(B|A)P(A)$$

EJEMPLO 4.16. Elección de productos: ketchup y mostaza II (regla del producto)

Cuando la probabilidad condicionada del consumo de mostaza, dado el consumo de ketchup,

$$P(A|B) = \frac{0{,}65}{0{,}80} = 0{,}8125$$

se multiplica por la probabilidad del consumo de ketchup, tenemos la probabilidad conjunta tanto del consumo de mostaza como del consumo de ketchup:

$$P(A \cap B) = (0{,}8125)(0{,}80) = 0{,}65$$

En el ejemplo siguiente vemos una interesante aplicación de la regla del producto de probabilidades. También reunimos algunas ideas presentadas anteriormente.

EJEMPLO 4.17. Preguntas delicadas (regla del producto)

Suponga que en una ciudad se realizó una encuesta y que a cada encuestado se le hicieron las dos preguntas siguientes:

a) ¿Es el último dígito del número de su documento nacional de identidad un número impar?
b) ¿Ha mentido alguna vez en una solicitud de empleo?

La segunda pregunta es, por supuesto, muy delicada y es de suponer que algunas personas no dirán la verdad por diversas razones, sobre todo si su respuesta es sí. Para eliminar este posible sesgo, se pidió a los encuestados que lanzaran una moneda al aire y respondieran a la pregunta (a) si el resultado era «cara» y a la (b) en caso contrario. El 37 por ciento de los encuestados respondió «sí». ¿Cuál es la probabilidad de que un encuestado que estaba respondiendo a la pregunta delicada (b), respondiera afirmativamente?

Solución

Definimos los siguientes sucesos:

A: El encuestado responde afirmativamente.
E_1: El encuestado responde a la pregunta (a).
E_2: El encuestado responde a la pregunta (b).

Por el análisis del problema sabemos que $P(A) = 0{,}37$. También sabemos que la elección de la pregunta se hace lanzando una moneda al aire, por lo que $P(E_1) = 0{,}50$ y $P(E_2) = 0{,}50$. Sabemos, además, cuáles son las respuestas a la pregunta (a). Como el último dígito de la mitad de todos los números del documento nacional de identidad es impar, la probabilidad de que la respuesta sea afirmativa, dado que se ha respondido a la pregunta (a), debe ser 0,50, es decir, $P(A|E_1) = 0{,}50$.

Sin embargo, necesitamos $P(A\,|\,E_2)$, que es la probabilidad condicionada de que la respuesta sea afirmativa, dado que se respondió a la pregunta (b). Podemos hallar esta probabilidad utilizando dos resultados de los apartados anteriores. Sabemos que E_1 y E_2 son mutuamente excluyentes y colectivamente exhaustivos. También sabemos que las intersecciones $E_1 \cap A$ y $E_2 \cap A$ son mutuamente excluyentes y que su unión es A. Por lo tanto, la suma de las probabilidades de estas dos intersecciones es la probabilidad de A, por lo que

$$P(A) = P(E_1 \cap A) + P(E_2 \cap A)$$

A continuación, utilizando la regla del producto, tenemos que

$$P(E_1 \cap A) = P(A\,|\,E_1)P(E_1) = (0{,}50)(0{,}50) = 0{,}25$$

Y

$$P(E_2 \cap A) = P(A) - P(E_1 \cap A) = 0{,}37 - 0{,}25 = 0{,}12$$

A continuación, podemos hallar la probabilidad condicionada:

$$P(A\,|\,E_2) = \frac{P(E_2 \cap A)}{P(E_2)} = \frac{0{,}12}{0{,}50} = 0{,}24$$

Partiendo de este resultado, estimamos que el 24 por ciento de la población encuestada ha mentido en alguna solicitud de empleo.

Independencia estadística

La *independencia estadística* es un caso especial en el que la probabilidad condicionada de A, dado B, es igual que la probabilidad incondicionada de A. Es decir, $P(A|B) = P(A)$. En general, este resultado no es cierto, pero cuando lo es, vemos que el hecho de saber que el suceso B no ha ocurrido no altera la probabilidad del suceso A.

Independencia estadística

Sean A y B dos sucesos. Se dice que estos sucesos son **estadísticamente independientes** si y sólo si

$$P(A \cap B) = P(A)P(B)$$

También se deduce de la regla del producto que

$$P(A|B) = P(A) \qquad (\text{si } P(B) > 0)$$

$$P(B|A) = P(B) \qquad (\text{si } P(A) > 0)$$

En términos más generales, los sucesos E_1, E_2, ..., E_K son independientes estadísticamente si y sólo si

$$P(E_1 \cap E_2 \cap \cdots \cap E_K) = P(E_1)P(E_2) \cdots P(E_K)$$

Como mejor se ve la base lógica de la definición de independencia estadística es por medio de las probabilidades condicionadas y como más atractiva resulta es por medio de la probabilidad subjetiva. Supongamos que creemos que la probabilidad de que ocurra el suceso A es $P(A)$. Ahora se nos da la información de que ha ocurrido el suceso B. Si esta nueva información no cambia mi valoración de la probabilidad de A, entonces $P(A) = P(A|B)$ y la información sobre la ocurrencia de B no tiene ningún valor en la determinación de $P(A)$. Esta definición de independencia estadística coincide con el concepto de sentido común de «independencia». Para ayudar a comprender la independencia, presentamos en la Tabla 4.3 una versión revisada de nuestro problema de la mostaza y el ketchup. En este caso, las probabilidades marginales del ketchup y la mostaza son iguales, pero su consumo es independiente. Obsérvese que las definiciones anteriores de independencia llevan a una conclusión de independencia en el caso de la Tabla 4.3, pero no en el de la 4.2.

En nuestros análisis siguientes llamaremos «independientes» a los sucesos. Por ejemplo, los sucesos «el índice Dow-Jones subirá» y «las corbatas son más anchas» son independientes. Lo que creamos sobre la probabilidad del segundo no influirá en las posibilidades de que ocurra el primero. El ejemplo 4.18 muestra cómo se sabe si dos sucesos son independientes.

Tabla 4.3. Probabilidad conjunta de la mostaza y el ketchup cuando son independientes.

	Mostaza	**No mostaza**	
Ketchup	0,60	0,20	0,80
No ketchup	0,15	0,05	0,20
	0,75	0,25	1,0

EJEMPLO 4.18. Probabilidad de los títulos universitarios (independencia estadística)

Supongamos que las mujeres obtienen el 48 por ciento de todos los títulos de licenciatura en un país y que el 17,5 por ciento de todos los títulos de licenciatura son de administración de empresas. Además, el 6 por ciento de todos los títulos de licenciatura va a parar a mujeres que se licencian en administración de empresas. ¿Son los sucesos «el licenciado es una mujer» y «la licenciatura es de administración de empresas» estadísticamente independientes?

Solución

Sea A el suceso «el licenciado es una mujer» y B «la licenciatura es de administración de empresas». Tenemos que

$$P(A) = 0{,}48 \qquad P(B) = 0{,}175 \qquad P(A \cap B) = 0{,}06$$

Dado que

$$P(A)P(B) = (0{,}48)(0{,}175) = 0{,}084 \neq 0{,}06 = P(A \cap B)$$

estos sucesos no son independientes. La dependencia puede comprobarse por medio de la probabilidad condicionada:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0{,}06}{0{,}175} = 0{,}343 \neq 0{,}48 = P(A)$$

Por lo tanto, en el país examinado sólo el 34,3 por ciento de las licenciaturas va a parar a mujeres, mientras que las mujeres constituyen el 48 por ciento de todos los licenciados.

También es importante distinguir entre los términos *mutuamente excluyente* e *independiente*. Dos sucesos son mutuamente excluyentes si no pueden ocurrir conjuntamente; es decir, la probabilidad de su intersección es 0. Cuando los sucesos son independientes, la probabilidad de su intersección es el producto de sus probabilidades individuales y, en general, esa probabilidad no es 0 (a menos que la probabilidad de uno de los sucesos sea 0, y ese resultado no es muy interesante). También debe señalarse que si sabemos que dos sucesos son mutuamente excluyentes, entonces si ocurre uno, el otro no puede ocurrir, y los sucesos no son independientes.

En algunas circunstancias, la independencia puede deducirse, o al menos inferirse razonablemente, de la naturaleza de un experimento aleatorio. Por ejemplo, si lanzamos al aire dos veces o más una moneda equilibrada, la probabilidad de que salga «cara» es la misma en todos los lanzamientos y en ella no influye el resultado de los lanzamientos anteriores. En ese caso, la probabilidad de la intersección puede calcularse multiplicando las probabilidades. Este resultado es especialmente útil en el caso de los experimentos repetidos que son lógicamente independientes.

EJEMPLO 4.19. Reparación de computadores (independencia)

La experiencia dice que el 90 por ciento de los computadores de un determinado modelo funcionan como mínimo 1 año antes de que haya que efectuar alguna reparación. Un directivo compra tres computadores de este modelo. ¿Cuál es la probabilidad de que los tres funcionen 1 año sin necesidad de reparación alguna?

Solución

En este caso, es razonable suponer que las averías de los tres computadores son independientes. Los tres se fabricaron en la misma cadena de montaje y su uso en la empresa probablemente es similar. Dado el supuesto de la independencia, sea E_i «el i-ésimo computador funciona 1 año sin necesidad de ninguna reparación». El supuesto de la independencia lleva entonces a

$$P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3) = 0{,}90^3 = 0{,}729$$

Debemos hacer hincapié en que los sucesos no siempre son independientes. En el ejemplo 4.19 los computadores podrían recibir la corriente eléctrica del mismo circuito y ese circuito puede no estar protegido de las subidas de la tensión. En ese caso, una subida de la tensión que aumentara la probabilidad de que se averíe un computador provocaría un aumento de la probabilidad de que se averiaran todos los computadores. Por lo tanto, los sucesos no son independientes. La condición de la independencia de los sucesos es un supuesto y sólo debe utilizarse tras un detenido análisis del proceso examinado.

Los dos ejemplos siguientes muestran cómo podemos simplificar a menudo el cálculo de la probabilidad de un suceso calculando primero la probabilidad del complementario y utilizándola después para hallar la probabilidad del suceso que nos interesa.

EJEMPLO 4.20. El problema del día de nacimiento (regla del complementario)

Una gran pregunta en una fiesta es «¿qué probabilidades hay de que al menos dos personas de las que se encuentran en esta habitación hayan nacido el mismo día?». Desgraciadamente, será difícil para el lector compartir con los asistentes a la fiesta el método para hallar la solución.

Para que el problema sea manejable, asignamos todos los nacidos el 29 de febrero al 1 de marzo y suponemos que los 365 días del año son igual de probables en el conjunto de la población. También suponemos que las personas que hay en la habitación son una muestra aleatoria, con respecto a las fechas de nacimiento, de la población en general (estas simplificaciones apenas afectan a los resultados numéricos).

Solución

Sea M el número de personas que hay en el grupo y A el suceso «al menos un par nacieron el mismo día». Ahora bien, sería muy tedioso hallar la probabilidad de A directamente, ya que tendríamos que tener en cuenta la posibilidad de que hubiera más de un par de personas cuya fecha de nacimiento coincidiera. Es más fácil hallar la probabilidad de que «todas las M personas nacieran en días diferentes», es decir, $\bar{A}$.

Como hay 365 fechas de nacimiento posibles para cada persona y cada una puede relacionarse con todas las fechas de nacimiento posibles de otras personas, el número total de ordenaciones igualmente probables de M personas es 365^M. A continuación, nos preguntamos cuántos de estos resultados están contenidos en el suceso $\bar{A}$, es decir, cuántos pares que implican a los M individuos tienen fechas de nacimiento diferentes. Eso es exactamente lo mismo que preguntar de cuántas formas pueden seleccionarse M fechas de nacimiento de 365 fechas de nacimiento posibles y ordenarlas. La fecha de nacimiento de la primera persona puede ocurrir en cualquiera de 365 días, la segunda en

cualquiera de 364 días, la tercera en cualquiera de 363 días, y así sucesivamente. Por lo tanto, en el caso de M personas el número de fechas de nacimiento diferentes es

$$(365)\,(364)\,(363)\cdots(365 - M + 1)$$

El número de fechas de nacimiento posibles de M personas es 365^M. Por lo tanto, la probabilidad de que las M fechas de nacimiento sean diferentes es

$$P(\bar{A}) = \frac{(365)(364)\cdots(365 - M + 1)}{365^M}$$

La probabilidad de que haya al menos dos personas es el complementario

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{(365)(364)\cdots(365 - M + 1)}{365^M}$$

Las probabilidades de algunos números de personas, M, son

M	10	20	22	23	30	40	60
$P(A)$	0,117	0,411	0,476	0,507	0,706	0,891	0,994

Si hay al menos 23 personas en el grupo, la probabilidad de que al menos un par naciera el mismo día es de más de 0,50. Esta probabilidad aumenta vertiginosamente a medida que es mayor el grupo hasta que, cuando está formado por 60 personas, es casi seguro que encontraremos al menos un par. Este resultado es sorprendente para la mayoría de la gente. La probabilidad de que cualquier par dado de personas haya nacido el mismo día es de 1/365. Pero a medida que aumenta el grupo, el número de posibles coincidencias aumenta hasta que la probabilidad de que haya al menos una coincidencia es bastante grande. Aquí tenemos un caso de unión de sucesos que son individualmente improbables, pero que, cuando se consideran conjuntamente, la probabilidad es bastante grande. La utilización de reglas de probabilidad bastante sencillas a veces da sorprendentes resultados.

EJEMPLO 4.21. Viajes en avión gratuitos (regla del complementario)

En una promoción de una compañía aérea, los clientes y los posibles clientes recibieron vales. Uno de cada 325 de estos vales contenía un regalo de un billete de ida y vuelta para viajar a cualquier lugar al que volase la compañía. ¿Cuántos vales necesitaría una persona para tener un 50 por ciento de probabilidades de conseguir al menos un viaje gratuito?

Solución

El suceso que nos interesa, A, es «con M vales se consigue al menos un viaje gratuito». De nuevo, es más fácil hallar primero la probabilidad del complementario, $\bar{A}$, donde $\bar{A}$ es el suceso «con M vales no se consigue ningún viaje gratuito». La probabilidad de conseguir un viaje con un vale es 1/325 y, por lo tanto, la probabilidad de no ganar es 324/325. Si el individuo tiene M vales, el suceso de que no se consigue con ninguno de

ellos es justamente la intersección de los sucesos «No ha conseguido un viaje» para cada uno de los vales. Por otra parte, estos sucesos son independientes y, por lo tanto,

$$P(\bar{A}) = \left(\frac{324}{325}\right)^M$$

y la probabilidad de conseguir al menos un viaje es

$$P(A) = 1 - P(\bar{A}) = 1 - \left(\frac{324}{325}\right)^M$$

Para que $P(A)$ sea al menos 0,5, el individuo necesita como mínimo $M = 225$ vales.

De nuevo, este resultado es sorprendente. Cabría imaginar que si la probabilidad de conseguir un viaje con un único vale es 1/325, bastarían 163 vales para tener un 50 por ciento de probabilidades de ganar. Sin embargo, en ese caso estaríamos suponiendo implícitamente que la probabilidad de una unión es la suma de las probabilidades individuales y no tendríamos en cuenta que hay que restar las probabilidades correspondientes a las intersecciones que se han contado dos veces (lo que en este caso implicaría que en M vales hay más de uno que regala un viaje).

EJERCICIOS

Ejercicios básicos

4.19. La probabilidad de A es 0,60 y la de B es 0,45 y la de cualquiera de los dos es 0,80. ¿Cuál es la probabilidad tanto de A como de B?

4.20. La probabilidad de A es 0,40 y la de B es 0,45 y la de cualquiera de los dos es 0,85. ¿Cuál es la probabilidad tanto de A como de B?

4.21. La probabilidad de A es 0,60 y la de B es 0,40 y la de cualquiera de los dos es 0,76. ¿Cuál es la probabilidad tanto de A como de B?

4.22. La probabilidad de A es 0,60 y la de B es 0,45 y la de cualquiera de los dos es 0,30. ¿Cuál es la probabilidad tanto de A como de B?

4.23. La probabilidad de A es 0,60 y la de B es 0,45 y la de cualquiera de los dos es 0,30. ¿Cuál es la probabilidad condicionada de A, dado B? ¿Son A y B independientes en el sentido probabilístico?

4.24. La probabilidad de A es 0,80 y la de B es 0,10 y la de cualquiera de los dos es 0,08. ¿Cuál es la probabilidad condicionada de A, dado B? ¿Son A y B independientes en el sentido probabilístico?

4.25. La probabilidad de A es 0,30 y la de B es 0,40 y la de cualquiera de los dos es 0,30. ¿Cuál es la probabilidad condicionada de A, dado B? ¿Son A y B independientes en el sentido probabilístico?

4.26. La probabilidad de A es 0,70 y la de B es 0,80 y la de cualquiera de los dos es 0,50. ¿Cuál es la probabilidad condicionada de A, dado B? ¿Son A y B independientes en el sentido probabilístico?

Ejercicios aplicados

4.27. Una empresa sabe que una competidora está a punto de introducir en el mercado un producto rival. Cree que esta empresa tiene en mente tres planes posibles de empaquetado (superior, normal y barato) y que todos son igual de probables. Además, hay tres estrategias de marketing igual de probables (publicidad intensa en los medios de comunicación, descuentos de precios y utilización de un cupón para reducir el precio de futuras compras). ¿Cuál es la probabilidad de que la empresa competidora emplee un empaquetado superior junto con una intensa campaña publicitaria en los medios de comunicación? Suponga que los planes de empaquetado y las estrategias de marketing se deciden independientemente.

4.28. Un analista financiero recibió el encargo de evaluar las perspectivas de beneficios de siete empresas para el próximo año y de ordenarlas en función de las tasas previstas de crecimiento de los beneficios.

a) ¿Cuántas ordenaciones son posibles?

b) Si una ordenación es, de hecho, el resultado de una conjetura, ¿cuál es la probabilidad de que esta conjetura resulte correcta?

4.29. Una empresa tiene 50 representantes de ventas. Decide que el que tuvo más éxito el año pasado será premiado con unas vacaciones en Hawai en enero, mientras que el segundo será premiado con unas vacaciones en Las Vegas. Los demás representantes deberán asistir a una conferencia sobre los métodos modernos de ventas que se celebrará en Buffalo. ¿Cuántos resultados son posibles?

4.30. Un analista de títulos sostiene que, dada una lista específica de acciones ordinarias de seis empresas, es posible predecir en el orden correcto las tres que obtendrán mejores resultados el próximo año. ¿Qué probabilidades hay de que se haga la selección correcta por casualidad?

4.31. Un comité de estudiantes tiene seis miembros: cuatro estudiantes de licenciatura y dos de doctorado. Hay que elegir aleatoriamente a un subcomité de tres miembros de manera que todas las combinaciones posibles de tres de los seis estudiantes tengan las mismas probabilidades de salir elegidas. ¿Cuál es la probabilidad de que no haya estudiantes de doctorado en el subcomité?

4.32. En un torneo de baloncesto que se celebra en una ciudad participan cinco equipos. Hay que predecir por orden cuáles serán los tres mejores al final de la temporada. Dejando a un lado la posibilidad de que haya empates, calcule el número de predicciones que pueden hacerse. ¿Cuál es la probabilidad de que se haga la predicción correcta por casualidad?

4.33. Un directivo tiene cuatro ayudantes —Juan, Jorge, María y Javier— para asignar a cuatro tareas. Cada ayudante es asignado a una de las tareas y hay un ayudante para cada tarea.

a) ¿Cuántas asignaciones diferentes son posibles?
b) Si las asignaciones se realizan aleatoriamente, ¿qué probabilidades hay de que María sea asignada a una tarea específica.

4.34. La dirección de una empresa ha decidido que en el futuro repartirá su presupuesto publicitario entre dos agencias. Actualmente, está considerando ocho agencias para hacer ese trabajo. ¿Cuántas elecciones de dos agencias son posibles?

4.35. Suponga que es una de las siete candidatas que se presentan a una prueba para representar dos papeles —la heroína y su mejor amiga— en una obra. Antes de la prueba, no sabe nada de las demás candidatas y supone que todas tienen las mismas probabilidades de representar los papeles.

a) ¿Cuántas elecciones son posibles para representar los dos papeles?
b) ¿En cuántas de las posibilidades del apartado (a) sería elegida para representar la heroína?
c) ¿En cuántas de las posibilidades del apartado (a) sería elegida para representar a la mejor amiga?
d) Utilice los resultados de los apartados (a) y (b) para hallar la probabilidad de que sea elegida para representar a la heroína. Indique una forma más directa de hallar esta probabilidad.
e) Utilice los resultados de las preguntas (a), (b) y (c) para hallar la probabilidad de que sea elegida para representar uno de los dos papeles. Indique una forma más directa de hallar esta probabilidad.

4.36. Para realizar un proyecto de construcción hay que formar una cuadrilla en la que debe haber dos oficiales y cuatro peones seleccionados de un total de cinco oficiales y seis peones.

a) ¿Cuántas combinaciones son posibles?
b) El hermano de uno de los oficiales es peón. Si la cuadrilla se forma aleatoriamente, ¿cuál es la probabilidad de que sean seleccionados los dos hermanos?
c) ¿Cuál es la probabilidad de que no sea seleccionado ninguno de los hermanos?

4.37. Un fondo de inversión tiene seis fondos que invierten en el mercado de Estados Unidos y cuatro que invierten en mercados internacionales. Un cliente quiere invertir en dos fondos estadounidenses y dos fondos internacionales.

a) ¿Cuántos conjuntos de fondos de esta empresa podría elegir el inversor?
b) Uno de los fondos estadounidenses y uno de los fondos internacionales obtendrá muy malos resultados el próximo año, pero el inversor no lo sabe. Si el inversor selecciona fondos para comprar aleatoriamente, ¿cuál es la probabilidad de que al menos uno de los fondos elegidos obtenga muy malos resultados el año que viene?

4.38. Se ha estimado que el 30 por ciento de todos los estudiantes de último curso que hay en una universidad está realmente preocupado por sus perspectivas de empleo, el 25 por ciento está muy preocupado por las calificaciones y el 20 por ciento está muy preocupado por ambas cosas. ¿Cuál es la probabilidad de que un estudiante de

esta universidad elegido aleatoriamente esté muy preocupado al menos por una de estas dos cosas?

4.39. El dueño de una tienda de música observa que el 30 por ciento de los clientes que entran en la tienda pide ayuda a un dependiente y que el 20 por ciento compra antes de irse. También observa que el 15 por ciento de todos los clientes pide ayuda y compra algo. ¿Cuál es la probabilidad de que un cliente haga al menos una de estas dos cosas?

4.40. Volviendo a la información del ejercicio 4.39, considere dos sucesos: «el cliente pide ayuda» y «el cliente compra algo». Responda a las siguientes preguntas justificando su respuesta por medio de las probabilidades de los sucesos relevantes.

a) ¿Son los dos sucesos mutuamente excluyentes?
b) ¿Son los dos sucesos colectivamente exhaustivos?
c) ¿Son los dos sucesos estadísticamente independientes?

4.41. Una organización local solicita donaciones por teléfono. Se ha estimado que la probabilidad de que cualquier individuo haga inmediatamente una donación mediante tarjeta de crédito para una determinada lista de proyectos es de 0,05, la probabilidad de que no haga una donación inmediatamente pero solicite más información por correo es de 0,25 y la probabilidad de que no muestre ningún interés es de 0,7. Se envía información por correo a todas las personas que la solicitan y se estima que el 20 por ciento de estas personas acabará haciendo una donación. Un operador hace una serie de llamadas, cuyos resultados puede suponerse que son independientes.

a) ¿Cuál es la probabilidad de que no se reciba inmediatamente ninguna donación mediante tarjeta de crédito hasta que se hagan al menos cuatro llamadas sin éxito?
b) ¿Cuál es la probabilidad de que la primera llamada que consigue una donación (inmediatamente o finalmente por correo) vaya precedida como mínimo de cuatro llamadas sin éxito?

4.42. Una empresa de venta por correo considera tres sucesos posibles al enviar un pedido:

A: Se envía un artículo que no es el solicitado.
B: El artículo se pierde en el camino.
C: El artículo sufre daños en el camino.

Suponga que *A* es independiente tanto de *B* como de *C* y que *B* y *C* son mutuamente excluyentes. Las probabilidades de los sucesos individuales son $P(A) = 0{,}02$ y $P(B) = 0{,}01$ y $P(C) = 0{,}04$. Halle la probabilidad de que ocurra al menos uno de estos desastres en el caso de un pedido elegido aleatoriamente.

4.43. Un entrenador selecciona para un equipo universitario a un jugador estrella que está actualmente en el último curso de secundaria. Para poder jugar el próximo año este jugador debe haber terminado los estudios secundarios con buenas notas y haber aprobado un examen de acceso a la universidad. El entrenador estima que la probabilidad de que el deportista no obtenga buenas notas en secundaria es 0,02, que la probabilidad de que no apruebe el examen de acceso a la universidad es 0,15 y que estos sucesos son independientes. Según estas estimaciones, ¿cuál es la probabilidad de que este estudiante reúna las condiciones para poder jugar el año que viene en la universidad?

4.44. Según un estudio de mercado realizado en una ciudad, en una semana el 18 por ciento de todos los adultos ve un programa de televisión sobre temas empresariales y financieros, el 12 por ciento lee una publicación dedicada a estos temas y el 10 por ciento hace las dos cosas.

a) ¿Qué probabilidad hay de que un adulto de esta ciudad que vea un programa de televisión sobre temas empresariales y financieros lea una publicación dedicada a estos temas?
b) ¿Qué probabilidad hay de que un adulto de esta ciudad que lea una publicación dedicada a temas empresariales y financieros vea un programa de televisión sobre estos temas?

4.45. Un inspector examina artículos que salen de una cadena de montaje. Sus anotaciones revelan que sólo acepta el 8 por ciento de todos los artículos defectuosos. También se ha observado que el 1 por ciento de todos los artículos que salen de la cadena de montaje son defectuosos y son aceptados por el inspector. ¿Cuál es la probabilidad de que un artículo de esta cadena de montaje elegido aleatoriamente sea defectuoso?

4.46. Un analista recibe listas de cuatro acciones y cinco bonos. Recibe el encargo de predecir por orden qué dos acciones obtendrán el mayor rendimiento el próximo año y qué dos bonos obtendrán el mayor rendimiento el próximo año. Suponga que estas predicciones se hacen aleatoriamente e independientemente la una de la

otra. ¿Qué probabilidades hay de que el analista tenga éxito al menos en una de las dos tareas?

4.47. Un banco clasifica a los prestatarios en dos grupos: de alto riesgo y de bajo riesgo. Sólo concede el 15 por ciento de sus préstamos a prestatarios de alto riesgo. El 5 por ciento de todos sus préstamos no se devuelve y el 40 por ciento de los que no se devuelven se concedió a prestatarios de alto riesgo. ¿Cuál es la probabilidad de que un prestatario de alto riesgo no devuelva su préstamo?

4.48. Una conferencia empezó al mediodía con dos sesiones paralelas. A la sesión sobre gestión de carteras asistió el 40 por ciento de los delegados, mientras que a la sesión sobre «chartismo» asistió el 50 por ciento. La sesión de la tarde era una charla titulada «¿Ha muerto el paseo aleatorio?». A ella asistió el 80 por ciento de todos los delegados.

a) Si la asistencia a la sesión sobre gestión de carteras y la asistencia a la sesión sobre «chartismo» son mutuamente excluyentes, ¿cuál es la probabilidad de que un delegado seleccionado aleatoriamente asistiera al menos a una de estas sesiones?

b) Si la asistencia a la sesión sobre gestión de carteras y la asistencia a la sesión de la tarde son estadísticamente independientes, ¿cuál es la probabilidad de que un delegado seleccionado aleatoriamente asistiera al menos a una de estas sesiones?

c) El 75 por ciento de los que asistieron a la sesión sobre «chartismo» también asistió a la sesión de la tarde. ¿Cuál es la probabilidad de que un delegado seleccionado aleatoriamente asistiera al menos a una de estas dos sesiones?

4.49. Un analista de bolsa sostiene que es experto en la selección de acciones que obtendrán resultados mejores de lo normal. Este analista recibe una lista de acciones de cinco empresas de alta tecnología y una lista de acciones de cinco compañías aéreas y debe indicar por orden cuáles son las acciones de tres empresas que obtendrán mejores resultados en cada una de estas dos listas el año que viene. El analista sostiene que acertar en una de estas dos tareas ya sería un gran éxito. Si elige de hecho aleatoria e independientemente, ¿cuál es la probabilidad de que tenga éxito al menos en una de las dos tareas meramente por causalidad? Dado este resultado, ¿qué piensa de la afirmación del analista?

4.50. Un director de control de calidad observó que el 30 por ciento de los problemas relacionados con el trabajo ocurría los lunes y que el 20 por ciento ocurría en la última hora del turno de día. También observó que el 4 por ciento de los problemas relacionados con los trabajadores ocurría en la última hora del turno del lunes.

a) ¿Qué probabilidades hay de que un problema relacionado con los trabajadores que ocurre en lunes no ocurra en la última hora del turno de día?

b) ¿Son estadísticamente independientes los sucesos «el problema ocurre el lunes» y «el problema ocurre en la última hora del turno de día»?

4.51. A una empresa le preocupaba el nivel de estudios básicos de sus trabajadores y decidió ofrecer a un grupo seleccionado clases de lectura y de matemáticas. El 40 por ciento de estos trabajadores se apuntó a las clases de lectura y el 50 por ciento a las de matemáticas. El 30 por ciento de los que se apuntaron a las clases de lectura se apuntó a las clases de matemáticas.

a) ¿Cuál es la probabilidad de que un trabajador seleccionado aleatoriamente se apuntara a las dos clases?

b) ¿Cuál es la probabilidad de que un trabajador seleccionado aleatoriamente que se apuntara a las clases de matemáticas se apuntara también a las de lectura?

c) ¿Cuál es la probabilidad de que un trabajador seleccionado aleatoriamente se apuntara al menos a una de estas dos clases?

d) ¿Son estadísticamente independientes los sucesos «se apunta a las clases de lectura» y «se apunta a las clases de matemáticas»?

4.52. Una empresa de trabajos de jardinería ha realizado llamadas telefónicas para captar clientes para la próxima temporada. Según sus datos, en el 15 por ciento de estas llamadas consiguió nuevos clientes y el 80 por ciento de estos nuevos clientes había utilizado los servicios de alguna empresa de la competencia el año anterior. También se estima que el 60 por ciento de todas las personas a las que llamó habían utilizado los servicios de una empresa rival el año anterior. ¿Qué probabilidades hay de que una llamada a una persona que utilizó los servicios de una empresa rival el año pasado consiga un nuevo cliente?

4.53. Una editorial puede utilizar todas las estrategias posibles para mejorar las ventas de un libro, algunas o ninguna:

a) Una cara promoción antes de la publicación.
b) Un caro diseño de cubierta.
c) Una prima a los representantes de ventas que vendan un número de libros determinado de antemano.

Hasta ahora estas tres estrategias se han aplicado simultáneamente sólo al 2 por ciento de los libros de la editorial. El 20 por ciento de los libros tenía un caro diseño de cubierta, de los cuales el 80 por ciento había tenido una cara promoción antes de su publicación. Una editorial de la competencia se entera de que un nuevo libro va a tener tanto una cara promoción antes de la publicación como un caro diseño de cubierta y ahora quiere saber qué probabilidades hay de que se introduzca un sistema de primas para los representantes de ventas. Calcule la probabilidad que le interesa a la editorial rival.

4.4. Probabilidades bivariantes

En este apartado introducimos una clase de problemas en los que hay dos conjuntos distintos de sucesos, que llamamos $A_1, A_2, ..., A_h$ y $B_1, B_2, ..., B_k$. Estos problemas tienen muchas aplicaciones en el mundo de la empresa y en economía. Pueden estudiarse construyendo tablas de doble entrada que permiten solucionar intuitivamente los problemas. Los sucesos A_i y B_j son mutuamente excluyentes y colectivamente exhaustivos dentro de sus conjuntos, pero puede haber intersecciones $(A_i \cap B_j)$ entre todos los sucesos de los dos conjuntos. Estas intersecciones pueden considerarse resultados básicos de un experimento aleatorio. Dos conjuntos de sucesos, considerados conjuntamente de esta forma, se llaman *bivariantes* y las probabilidades se denominan *probabilidades bivariantes*.

También examinamos situaciones en las que es difícil hallar las probabilidades condicionadas deseadas, pero en las que se dispone de probabilidades condicionadas alternativas. Puede ser difícil hallar las probabilidades porque los costes de enumeración son altos o porque alguna restricción crítica, ética o legal impide obtener directamente las probabilidades.

La Tabla 4.4 ilustra los resultados de sucesos bivariantes llamados $A_1, A_2, ..., A_h$ y $B_1, B_2, ..., B_k$. Si pueden asignarse probabilidades a todas las intersecciones $(A_i \cap B_j)$, entonces se conoce toda la estructura de probabilidades del experimento aleatorio, por lo que se pueden calcular otras probabilidades de interés.

Tabla 4.4. Resultados correspondientes a sucesos bivariantes.

	B_1	B_2	$\cdots$	B_K
A_1	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$	$\cdots$	$P(A_1 \cap B_k)$
A_2	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	$\cdots$	$P(A_2 \cap B_k)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_h	$P(A_h \cap B_1)$	$P(A_h \cap B_2)$	$\cdots$	$P(A_h \cap B_k)$

Consideremos a modo de ilustración el caso de un publicista que quiere saber cuáles son las características de renta y otras características relevantes de la audiencia de un determinado programa de televisión. Las familias pueden clasificarse en tres categorías —ven habitualmente el programa de televisión, lo ven de vez en cuando y no lo ven nunca— utilizando A_i. También pueden clasificarse en tres subgrupos —renta baja, renta media y renta alta— utilizando B_j. A continuación, pueden mostrarse las clasificaciones cruzadas posibles por medio de una tabla como la 4.4, en la que $h = 3$ y $k = 3$. También pueden representarse en un diagrama de árbol como el de la Figura 4.9. Lo primero que tenemos a la izquierda es toda la población de familias. Esta población se divide en tres ramas, que

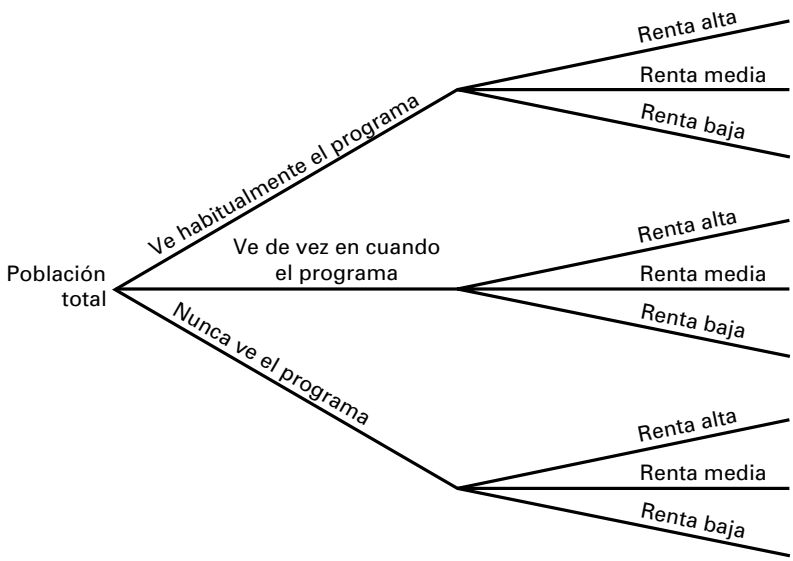


Figura 4.9. Tres diagramas del ejemplo de ver el programa de televisión y la renta.

dependen de la frecuencia con que ven el programa de televisión. Cada una de estas ramas se divide a su vez en tres subramas en función del nivel de renta familiar. Hay, pues, nueve subramas que corresponden a todas las combinaciones de frecuencia con que se ve el programa de televisión y nivel de renta.

Ahora tenemos que hallar las probabilidades de cada una de las intersecciones de sucesos. Estas probabilidades, obtenidas por medio de encuestas a los espectadores, se presentan en la Tabla 4.5. Por ejemplo, el 10 por ciento de las familias es de renta alta y ve de vez en cuando el programa de televisión. Estas probabilidades se hallan utilizando el concepto de frecuencia relativa, suponiendo que la encuesta es lo suficientemente grande para que sea posible considerar aproximadamente las proporciones como probabilidades. Basándose en esta información, la probabilidad de que una familia elegida aleatoriamente en la población tenga una renta alta y vea de vez en cuando el programa es 0,10.

Tabla 4.5. Probabilidades del ejemplo de ver el programa y la renta.

Frecuencia con que se ve el programa	Renta alta	Renta media	Renta baja	Total
Habitualmente	0,04	0,13	0,04	0,21
De vez en cuando	0,10	0,11	0,06	0,27
Nunca	0,13	0,17	0,22	0,52
Totales	0,27	0,41	0,32	1,00

Probabilidades conjuntas y marginales

En el contexto de las probabilidades bivariantes, las probabilidades de la intersección, $P(A_i \cap B_j)$, se llaman probabilidades conjuntas. Las probabilidades de sucesos individuales, $P(A_i)$ o $P(B_j)$, se denominan **probabilidades marginales**. Las probabilidades marginales se encuentran en el margen de una tabla como la 4.5 y pueden calcularse sumando la fila o la columna correspondiente.

Para hallar las probabilidades marginales de un suceso, sumamos simplemente las correspondientes probabilidades conjuntas mutuamente excluyentes:

$$P(A_i) = P(A_i \cap B_1) + P(A_i \cap B_2) + \cdots + P(A_i \cap B_k)$$

Obsérvese que eso equivaldría a sumar las probabilidades de una fila de la Tabla 4.5. Siguiendo el mismo razonamiento, las probabilidades de B_j son los totales de cada columna.

Continuando con el ejemplo, definamos los subgrupos que ven el programa de televisión: A_1, «habitualmente»; A_2, «de vez en cuando», y A_3, «nunca». Definamos también los subgrupos de renta: B_1, «alta»; B_2, «media», y B_3, «baja». La probabilidad de que una familia vea de vez en cuando el programa es

$$\begin{aligned} P(A_2) &= P(A_2 \cap B_1) + P(A_2 \cap B_2) + P(A_2 \cap B_3) \\ &= 0{,}10 + 0{,}11 + 0{,}06 = 0{,}27 \end{aligned}$$

Asimismo, sumando las otras filas de la Tabla 4.5, tenemos que $P(A_1) = 0{,}21$ y $P(A_3) = 0{,}52$. También podemos sumar las columnas de la Tabla 4.5 y obtener

$$P(B_1) = 0{,}27 \qquad P(B_2) = 0{,}41 \qquad \text{y} \qquad P(B_3) = 0{,}32$$

También pueden hallarse las probabilidades marginales a partir de diagramas de árbol como la Figura 4.10, que tiene las mismas ramas que la 4.9. La parte de la derecha contiene todas las probabilidades conjuntas; las probabilidades marginales de los tres sucesos de frecuencia se introducen en las ramas principales sumando las probabilidades de las subramas correspondientes. El modelo de las ramas de árbol es especialmente útil cuando hay más de dos sucesos de interés. En este caso, por ejemplo, al publicista también podría interesarle la edad del cabeza de familia o el número de hijos. Las probabilidades marginales de los distintos sucesos suman 1 porque esos sucesos son mutuamente excluyentes y mutuamente exhaustivos.

En muchas aplicaciones, observamos que las probabilidades condicionadas tienen más interés que las probabilidades marginales. A un publicista puede interesarle más la probabilidad de que una familia de renta alta vea la televisión que la probabilidad de que la vea cualquier familia. La probabilidad condicionada puede obtenerse fácilmente a partir de la

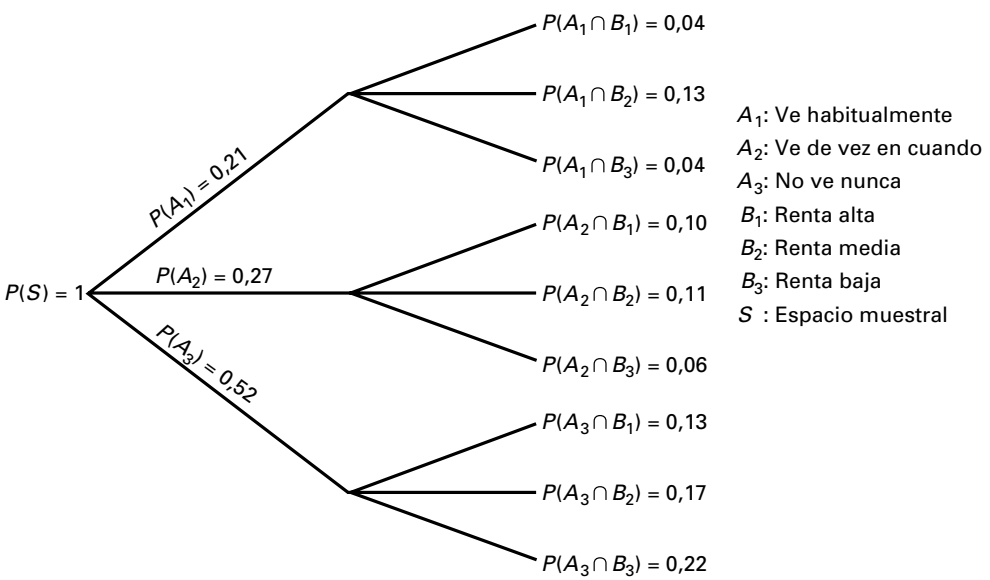


Figura 4.10. Diagrama de árbol del ejemplo de ver el programa y la renta, que muestra las probabilidades conjuntas y marginales.

Tabla 4.6. Probabilidades condicionadas de las frecuencias de ver el programa y la renta, que muestra las probabilidades conjuntas y marginales.

Frecuencia con que se ve el programa	Renta alta	Renta media	Renta baja
Habitualmente	0,15	0,32	0,12
De vez en cuando	0,37	0,27	0,19
Nunca	0,48	0,41	0,69

tabla porque tenemos todas las probabilidades conjuntas y las probabilidades marginales. Por ejemplo, la probabilidad de que una familia de renta alta vea habitualmente el programa es

$$P(A_1 | B_1) = \frac{P(A_1 \cap B_1)}{P(B_1)} = \frac{0{,}04}{0{,}27} = 0{,}15$$

La Tabla 4.6 muestra la probabilidad de los grupos de espectadores condicionada a los niveles de renta. Obsérvese que las probabilidades condicionadas con respecto a un determinado grupo de renta siempre suman 1, como se observa en las tres columnas de la citada tabla. Eso siempre es así, como se observa en la siguiente expresión:

$$\sum_{i=1}^{h} P(A_i | B_j) = \sum_{i=1}^{h} \frac{P(A_i \cap B_j)}{P(B_j)} = \frac{P(B_j)}{P(B_j)} = 1$$

También pueden calcularse, como muestra la Tabla 4.7, las probabilidades condicionadas de los grupos de renta, dadas las frecuencias con que se ve el programa, utilizando la definición de probabilidad condicionada y las probabilidades conjuntas y marginales.

Para hallar las probabilidades condicionadas con respecto a los grupos de renta de la Tabla 4.5 dividimos cada una de las probabilidades conjuntas de una fila por la probabilidad marginal de la columna de la derecha. Por ejemplo,

$$P(\text{Renta baja} \,|\, \text{De vez en cuando}) = \frac{0{,}06}{0{,}27} = 0{,}22$$

También podemos comprobar utilizando una tabla de doble entrada si los sucesos por pares son estadísticamente independientes. Recuérdese que los sucesos A_i y B_j son independientes si y sólo si su probabilidad conjunta es el producto de sus probabilidades marginales, es decir, si

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

Tabla 4.7. Probabilidades condicionadas de los niveles de renta, dadas las frecuencias de ver el programa.

Frecuencia con que se ve el programa	Renta alta	Renta media	Renta baja
Habitualmente	0,19	0,62	0,19
De vez en cuando	0,37	0,41	0,22
Nunca	0,25	0,33	0,42

En la Tabla 4.5, los sucesos conjuntos A_2 («de vez en cuando») y B_1 («renta alta») tienen una probabilidad

$$P(A_2 \cap B_1) = 0{,}10$$

y

$$P(A_2) = 0{,}27 \qquad P(B_1) = 0{,}27$$

El producto de estas probabilidades marginales es 0,0729 y, por lo tanto, no es igual a la probabilidad conjunta de 0,10; de ahí que los sucesos A_i y B_j no sean estadísticamente independientes.

Sucesos independientes

Sean A y B un par de sucesos, cada uno dividido en categorías de sucesos mutuamente excluyentes y colectivamente exhaustivos representados por A_1, A_2, ..., A_h y B_1, B_2, ..., B_K. Si todo suceso A_i es estadísticamente independiente de todo suceso B_j, entonces A y B son **sucesos independientes**.

Dado que A_2 y B_1 no son estadísticamente independientes, se deduce que los sucesos «frecuencia de ver» y «renta» no son independientes.

En muchas aplicaciones prácticas, no se conocen con precisión las probabilidades conjuntas. Se obtiene una muestra de una población y se estiman las probabilidades conjuntas a partir de los datos muestrales. Queremos saber, basándonos en esta evidencia muestral, si estos sucesos son independientes unos de otros. Más adelante en este libro presentamos un método para realizar un contraste de ese tipo.

Ventaja (odds)

La ventaja se emplea para transmitir información sobre las probabilidades en algunas situaciones. Por ejemplo, un comentarista deportivo podría afirmar que la ventaja a favor del equipo A frente al equipo B es de 2 a 1. La ventaja puede convertirse directamente en probabilidades y las probabilidades pueden convertirse en ventaja utilizando la siguiente ecuación.

Ventaja

La **ventaja** de un suceso es el cociente entre la probabilidad del suceso dividida por la probabilidad de su complementario. La ventaja a favor de A es

$$\text{Ventaja} = \frac{P(A)}{1 - P(A)} = \frac{P(A)}{P(\bar{A})} \qquad \textbf{(4.11)}$$

Por lo tanto, la ventaja de 2 a 1 puede convertirse en la probabilidad de que gane A:

$$\frac{2}{1} = \frac{P(A)}{1 - P(A)}$$

y aplicando el álgebra básica

$$2 \times (1 - P(A)) = P(A)$$

de donde

$$P(A) = 0{,}67$$

Asimismo, si la ventaja a favor de ganar es de 3 a 2, la probabilidad de ganar es 0,60. Obsérvese que 0,60/0,40 es igual a 3/2.

Cociente de «sobreparticipación»

Hay algunas situaciones en las que es difícil hallar las probabilidades condicionadas deseadas, pero se dispone de probabilidades condicionadas alternativas. Puede ser difícil hallar las probabilidades porque los costes de enumeración son altos o porque alguna restricción crítica, ética o legal impide hallar directamente las probabilidades. En algunos de esos casos, es posible utilizar relaciones probabilísticas básicas para hallar las probabilidades deseadas a partir de las probabilidades de las que se dispone. En este apartado presentamos uno de esos métodos basado en el uso de cocientes de «sobreparticipación» (véase la referencia bibliográfica 3).

Comenzamos examinando un sencillo ejemplo. Supongamos que sabemos que el 60 por ciento de los que compran nuestro producto ha visto nuestro anuncio, pero sólo lo ha visto el 30 por ciento de los que no lo compran. El cociente entre 60 y 30 por ciento es el cociente de «sobreparticipación» del suceso «ha visto nuestro anuncio» en el grupo de los que compran el producto, en comparación con el grupo que no lo compra. En el análisis siguiente mostramos que un cociente de «sobreparticipación» mayor que 1,0 es una prueba, por ejemplo, de que la publicidad influye en la conducta de compra.

El cociente de «sobreparticipación», presentado en la ecuación 4.12, es el cociente de la probabilidad de un suceso —como ver un anuncio— en el que sólo hay dos resultados mutuamente excluyentes y complementarios, como la venta de un producto o la no venta de un producto. Si el cociente de las probabilidades condicionadas no es igual a 1,0, entonces el suceso influye en los resultados. Estos cocientes tienen aplicaciones en algunas situaciones empresariales, entre las que se encuentran el marketing, la producción y la contabilidad. En este apartado desarrollamos la teoría y la aplicación de los *cocientes de «sobreparticipación»*.

Cocientes de «sobreparticipación»

La probabilidad del suceso A_1, condicionada al suceso B_1, dividida por la probabilidad de A_1, condicionada al suceso B_2, es el **cociente de «sobreparticipación»**:

$$\frac{P(A_1|B_1)}{P(A_1|B_2)} \tag{4.12}$$

Un cociente de «sobreparticipación» mayor que 1:

$$\frac{P(A_1|B_1)}{P(A_1|B_2)} > 1{,}0$$

implica que el suceso A_1 aumenta la ventaja a favor de B_1:

$$\frac{P(B_1|A_1)}{P(B_2|A_1)} > \frac{P(B_1)}{P(B_2)}$$

Consideremos una empresa que desea averiguar la eficacia de un nuevo anuncio. Se realiza un experimento en el que se muestra el anuncio a un grupo de clientes y no a otro y se observa la conducta de compra de los dos grupos. Este tipo de estudios tiene una alta probabilidad de error; puede estar sesgado porque la gente a menudo se comporta de forma distinta cuando se la observa de cerca y cuando no se la observa. Sin embargo, es posible calcular el porcentaje de compradores que han visto un anuncio y el porcentaje de no compradores que lo han visto. Veamos cómo pueden analizarse esos datos para hallar la eficacia del nuevo anuncio.

La eficacia de la publicidad se averigua realizando el siguiente análisis. La población se divide en

B_1: Compradores.
B_2: No compradores.

y en

A_1: Los que han visto el anuncio.
A_2: Los que no han visto el anuncio.

La ventaja a favor del comprador en este problema es

$$\frac{P(B_1)}{P(B_2)}$$

Asimismo, podemos definir la ventaja condicionada, en la que utilizamos el cociente de las probabilidades que son ambas condicionadas al mismo suceso. En este problema, la ventaja de un comprador condicionada a «haber visto un anuncio» es

$$\frac{P(B_1|A_1)}{P(B_2|A_1)}$$

Si la ventaja condicionada es mayor que la incondicionada, se dice que el suceso condicionante influye en el suceso que nos interesa. Por lo tanto, la publicidad se consideraría eficaz si

$$\frac{P(B_1|A_1)}{P(B_2|A_1)} > \frac{P(B_1)}{P(B_2)}$$

Los términos del primer miembro son iguales a

$$P(B_1|A_1) = \frac{P(A_1|B_1)P(B_1)}{P(A_1)}$$

$$P(B_2|A_1) = \frac{P(A_1|B_2)P(B_2)}{P(A_1)}$$

Introduciendo estos últimos términos en la primera ecuación, tenemos que

$$\frac{P(A_1|B_1)P(B_1)}{P(A_1|B_2)P(B_2)} > \frac{P(B_1)}{P(B_2)}$$

Dividiendo los dos miembros por el cociente de la derecha, tenemos que

$$\frac{P(A_1|B_1)}{P(A_1|B_2)} > 1{,}0$$

Este resultado muestra que si el porcentaje de compradores que han visto el anuncio es mayor que el porcentaje de no compradores que lo han visto, la ventaja a favor de comprar condicionada a haber visto el anuncio es mayor que la ventaja incondicionada. Por lo tanto, tenemos pruebas de que la publicidad va acompañada de un aumento de la probabilidad de compra.

En el problema inicial, el 60 por ciento de los compradores y el 30 por ciento de los no compradores habían visto el anuncio. El cociente de «sobreparticipación» es 2,0 (60/30) y, por lo tanto, llegamos a la conclusión de que la publicidad aumenta la probabilidad de compra. Los analistas de mercado utilizan este resultado para evaluar la eficacia de la publicidad y de otras actividades de promoción de las ventas. Se pregunta a los compradores de productos si han visto determinados anuncios y se realizan encuestas a hogares basadas en un muestreo aleatorio a partir de las cuales se halla el porcentaje de personas que han visto el anuncio y no han comprado el producto.

Consideremos otra situación en la que es difícil, ilegal o poco ético hallar las probabilidades.

EJEMPLO 4.22. El alcohol y los accidentes de tráfico (cocientes de «sobreparticipación»)

Los investigadores de la National Highway Traffic Safety Administration del Departamento de Transporte de Estados Unidos querían averiguar la influencia del alcohol en los accidentes de tráfico. Es evidente que no sería ético dar alcohol a un grupo de automovilistas y comparar su participación en accidentes con la de un grupo que no hubiera tomado alcohol. Sin embargo, los investigadores sí observaron que había bebido el 10,3 por ciento de los automovilistas que iban conduciendo de noche por un determinado condado y el 32,4 por ciento de los automovilistas involucrados en un accidente de un solo vehículo que iban conduciendo en ese mismo momento por ese mismo condado. Los accidentes en los que había un solo vehículo involucrado se eligieron para garantizar que el error de un automovilista podía atribuirse solamente a un automovilista, cuyo consumo de alcohol se había medido. Basándose en estos resultados, querían saber si había pruebas para concluir que los accidentes aumentaban por la noche cuando los automovilistas habían bebido. Utilice los datos para averiguar si el consumo de alcohol aumenta la probabilidad de que haya accidentes (véase la referencia bibliográfica 2).

Solución

Este problema puede resolverse utilizando los cocientes de «sobreparticipación». En primer lugar, hay que definir los sucesos del espacio muestral:

A_1: El automovilista había bebido.

A_2: El automovilista no había bebido.

C_1: El automovilista se vio involucrado en un accidente.

C_2: El automovilista no se vio involucrado en un accidente.

Sabemos que el alcohol, A_1, aumenta la probabilidad de que haya un accidente si

$$\frac{P(A_1 \mid C_1)}{P(A_1 \mid C_2)} > 1{,}0$$

La investigación permite saber que las probabilidades condicionadas son

$$P(A_1|C_1) = 0{,}324$$
$$P(A_1|C_2) = 0{,}103$$

Utilizando estos resultados, el cociente de «sobreparticipación» es

$$\frac{P(A_1|C_1)}{P(A_1|C_2)} = \frac{0{,}324}{0{,}103} = 3{,}15$$

Basándose en este análisis, hay pruebas para concluir que el alcohol aumenta la probabilidad de que haya accidentes de tráfico.

El cociente de «sobreparticipación» es un buen ejemplo de cómo pueden utilizarse las manipulaciones matemáticas de las probabilidades para obtener resultados útiles para tomar decisiones empresariales. La frecuente utilización de métodos automatizados de recogida de datos, incluidos los escáneres de códigos de barras, la segmentación de la audiencia y los datos censales en cintas y discos, permite calcular muchas probabilidades diferentes, probabilidades condicionadas y cocientes de «sobreparticipación». Como consecuencia, los análisis parecidos a los que presentamos en este capítulo han pasado a formar parte de la rutina diaria de los analistas de marketing y de los directores de productos.

EJERCICIOS

Ejercicios básicos

Los ejercicios básicos 4.54 a 4.60 se refieren a la Tabla 4.8.

4.54. ¿Cuál es la probabilidad conjunta de «renta alta» y «nunca»?

4.55. ¿Cuál es la probabilidad conjunta de «renta baja» y «habitualmente»?

4.56. ¿Cuál es la probabilidad conjunta de «renta media» y «nunca»?

4.57. ¿Cuál es la probabilidad conjunta de «renta media» y «de vez en cuando»?

4.58. ¿Cuál es la probabilidad condicionada de «renta alta», dado «nunca»?

4.59. ¿Cuál es la probabilidad condicionada de «renta baja», dado «de vez en cuando»?

4.60. ¿Cuál es la probabilidad condicionada de «habitualmente», dado «renta alta»?

4.61. La probabilidad de una venta es de 0,80. ¿Cuál es la ventaja a favor de una venta?

4.62. La probabilidad de una venta es de 0,50. ¿Cuál es la ventaja a favor de una venta?

4.63. Considere dos grupos de estudiantes: B_1, estudiantes que recibieron una buena nota en los exámenes, y B_2, estudiantes que recibieron una mala nota en los exámenes. En el grupo B_1, el 80 por ciento estudia más de 25 horas a la semana y en el B_2 el 40 por ciento estudia más de 25 horas a la semana. ¿Cuál es el cociente de «sobreparticipación» de los elevados niveles de estudio en las buenas notas con respecto a las malas notas?

Tabla 4.8. Probabilidades del ejemplo de ver el programa y la renta.

Frecuencia con que se ve el programa	Renta alta	Renta media	Renta baja	Total
Habitualmente	0,10	0,15	0,05	0,30
De vez en cuando	0,10	0,20	0,10	0,40
Nunca	0,05	0,05	0,20	0,30
Totales	0,25	0,40	0,35	1,00

4.64. Considere dos grupos de estudiantes: B_1, estudiantes que recibieron una buena nota en los exámenes, y B_2, estudiantes que recibieron una mala nota en los exámenes. En el grupo B_1, el 40 por ciento estudia más de 25 horas a la semana y en el B_2 el 20 por ciento estudia más de 25 horas a la semana. ¿Cuál es el cociente de «sobreparticipación» de los elevados niveles de estudio en las buenas notas con respecto a las malas notas?

4.65. Considere dos grupos de estudiantes: B_1, estudiantes que recibieron una buena nota en los exámenes, y B_2, estudiantes que recibieron una mala nota en los exámenes. En el grupo B_1, el 20 por ciento estudia más de 25 horas a la semana y en el B_2 el 40 por ciento estudia más de 25 horas a la semana. ¿Cuál es el cociente de «sobreparticipación» de los elevados niveles de estudio en las buenas notas con respecto a las malas notas?

Ejercicios aplicados

4.66. En una encuesta realizada para un supermercado, se ha clasificado a los clientes en los que van frecuentemente o infrecuentemente a la tienda y los que compran productos genéricos a menudo, a veces o nunca. La tabla adjunta muestra las proporciones de personas encuestadas en cada una de las seis clasificaciones conjuntas.

Frecuencia de las visitas	Compra de productos genéricos		
	A menudo	A veces	Nunca
Frecuente	0,12	0,48	0,19
Infrecuente	0,07	0,06	0,08

a) ¿Cuál es la probabilidad de que un cliente sea un comprador frecuente y compre a menudo productos genéricos?
b) ¿Cuál es la probabilidad de que un cliente que nunca compra productos genéricos vaya a la tienda frecuentemente?
c) ¿Son independientes los sucesos «nunca compra productos genéricos» y «va a la tienda frecuentemente»?
d) ¿Cuál es la probabilidad de que un cliente que va infrecuentemente a la tienda compre a menudo productos genéricos?
e) ¿Son independientes los sucesos «compra a menudo productos genéricos» y «va infrecuentemente a la tienda»?
f) ¿Cuál es la probabilidad de que un cliente vaya frecuentemente a la tienda?
g) ¿Cuál es la probabilidad de que un cliente no compre nunca productos genéricos?
h) ¿Cuál es la probabilidad de que un cliente vaya frecuentemente a la tienda o no compre nunca productos genéricos o ambas cosas?

4.67. Una consultora predice si el próximo año los beneficios de las empresas serán excepcionalmente bajos, excepcionalmente altos o normales. Antes de decidir si continúa comprando estas predicciones, un corredor de bolsa compara las predicciones pasadas con los resultados efectivos. La tabla adjunta muestra las proporciones en las nueve clasificaciones conjuntas.

Frecuentemente	Predicción		
	Excepcionalmente altos	Normales	Excepcionalmente bajos
Excepcionalmente altos	0,23	0,12	0,03
Normales	0,06	0,22	0,08
Excepcionalmente bajos	0,01	0,06	0,19

a) ¿En qué proporción se predice que los beneficios serán excepcionalmente altos?
b) ¿En qué proporción han sido los beneficios excepcionalmente altos?
c) Si una empresa tuviera unos beneficios excepcionalmente altos, ¿cuál es la probabilidad de que la consultora predijera correctamente este suceso?
d) Si la consultora predijera que una empresa va a tener unos beneficios excepcionalmente altos, ¿cuál es la probabilidad de que se materializaran?
e) ¿Cuál es la probabilidad de que una empresa de la que se hubiera predicho que iba a tener unos beneficios excepcionalmente altos tenga unos beneficios excepcionalmente bajos?

4.68. A los suscriptores de un periódico local se les preguntó si leían frecuentemente, de vez en cuando o nunca la sección económica y si tenían acciones ordinarias cotizadas en bolsa (o participaciones en un fondo de inversión) el año pasado. La tabla adjunta muestra las proporciones de suscriptores en las seis clasificaciones conjuntas.

	Leer la sección económica		
Acciones cotizadas	**Frecuentemente**	**De vez en cuando**	**Nunca**
Sí	0,18	0,10	0,04
No	0,16	0,31	0,21

a) ¿Cuál es la probabilidad de que un suscriptor seleccionado aleatoriamente no lea nunca la sección económica?
b) ¿Cuál es la probabilidad de que un suscriptor seleccionado aleatoriamente tuviera acciones cotizadas el año pasado?
c) ¿Cuál es la probabilidad de que un suscriptor que nunca lee la sección económica tuviera acciones cotizadas el año pasado?
d) Cuál es la probabilidad de que un suscriptor que tuviera acciones cotizadas el año pasado nunca lea la sección económica?
e) Cuál es la probabilidad de que un suscriptor que no lee habitualmente la sección económica tuviera acciones cotizadas el año pasado?

4.69. Una empresa recibe habitualmente una pieza delicada de tres subcontratistas. Observa que la proporción de piezas que son buenas o defectuosas del total recibido es la que muestra la tabla adjunta:

	Subcontratista		
Pieza	**A**	**B**	**C**
Buena	0,27	0,30	0,33
Defectuosa	0,02	0,05	0,03

a) Si se selecciona aleatoriamente una pieza de todas las piezas recibidas, ¿cuál es la probabilidad de que sea defectuosa?
b) Si se selecciona aleatoriamente una pieza de todas las piezas recibidas, ¿cuál es la probabilidad de que proceda del subcontratista B?
c) ¿Cuál es la probabilidad de que una pieza procedente del subcontratista B sea defectuosa?
d) ¿Cuál es la probabilidad de que una pieza defectuosa seleccionada aleatoriamente proceda del subcontratista B?
e) ¿Es la calidad de una pieza independiente de la fuente de suministro?
f) Desde el punto de vista de la calidad, ¿cuál de los tres subcontratistas es más fiable?

4.70. A los estudiantes de una clase de estadística para los negocios se les preguntó qué nota esperaban sacar en el curso y si hacían más problemas de los que ponía el profesor. La tabla adjunta muestra las proporciones de estudiantes en cada una de las ocho clasificaciones conjuntas.

	Nota esperada			
Problemas realizados	**A**	**B**	**C**	**Menos de C**
Sí	0,12	0,06	0,12	0,02
No	0,13	0,21	0,26	0,08

a) Halle la probabilidad de que un estudiante seleccionado aleatoriamente en esta clase hiciera más problemas.
b) Halle la probabilidad de que un estudiante seleccionado aleatoriamente en esta clase espere una A.
c) Halle la probabilidad de que un estudiante seleccionado aleatoriamente que hiciera más problemas espere una A.
d) Halle la probabilidad de que un estudiante seleccionado aleatoriamente que espere una A hiciera más problemas.
e) Halle la probabilidad de que un estudiante seleccionado aleatoriamente que hiciera más problemas espere una calificación de menos de B.
f) ¿Son independientes «resolución de más problemas» y «nota esperada»?

4.71. La tabla adjunta muestra las proporciones de vendedores de computadores clasificados según su estado civil y según que abandonaran el empleo o permanecieran en él 1 año.

Estado civil	**Permaneció 1 año**	**Se fue**
Casado	0,64	0,13
Soltero	0,17	0,06

a) ¿Cuál es la probabilidad de que un vendedor seleccionado aleatoriamente estuviera casado?
b) ¿Cuál es la probabilidad de que un vendedor seleccionado aleatoriamente dejara el empleo antes de un año?
c) ¿Cuál es la probabilidad de que un vendedor soltero seleccionado aleatoriamente dejara el empleo antes de un año?
d) ¿Cuál es la probabilidad de que un vendedor seleccionado aleatoriamente que permaneció un año estuviera casado?

4.72. La tabla adjunta muestra las proporciones de adultos que hay en zonas no metropolitanas, cla-

sificados según que lean o no periódicos y que votaran o no en las últimas elecciones.

Votaron	Lectores	No lectores
Casado	0,63	0,13
Soltero	0,14	0,10

a) ¿Cuál es la probabilidad de que un adulto de esta población seleccionado aleatoriamente votara?
b) ¿Cuál es la probabilidad de que un adulto de esta población seleccionado aleatoriamente lea periódicos?
c) ¿Cuál es la probabilidad de que un adulto de esta población seleccionado aleatoriamente que no lea periódicos no votara?

4.73. Un club de estudiantes universitarios distribuyó información sobre las condiciones para hacerse socio entre los nuevos estudiantes que asistieron a una reunión informativa. El 40 por ciento de los que recibieron esta información eran hombres y el 60 por ciento eran mujeres. Posteriormente, se observó que el 7 por ciento de los hombres y el 9 por ciento de las mujeres que recibieron esta información entraron en el club.

a) Halle la probabilidad de que entre en el club un nuevo estudiante seleccionado aleatoriamente que recibe información.
b) Halle la probabilidad de que un nuevo estudiante seleccionado aleatoriamente que entra en el club después de recibir información sea una mujer.

4.74. Un analista que está intentando predecir los beneficios que obtendrá una empresa el próximo año cree que el negocio de esa empresa es muy sensible al nivel de los tipos de interés. Cree que si el año que viene los tipos medios son más de un 1 por ciento más altos que este año, la probabilidad de que los beneficios crezcan significativamente es 0,1. Si el próximo año los tipos medios son más de un 1 por ciento más bajos que este año, se estima que la probabilidad de que los beneficios crezcan significativamente es 0,8. Por último, si el próximo año los tipos de interés medios se encuentran a una distancia máxima de un 1 por ciento de los tipos de este año, la probabilidad de que los beneficios crezcan significativamente es 0,5. El analista estima que la probabilidad de que los tipos sean el próximo año más de un 1 por ciento más altos es 0,25 y que la probabilidad de que sean más de un 1 por ciento más bajos que este año es 0,15.

a) ¿Cuál es la probabilidad estimada tanto de que los tipos de interés sean un 1 por ciento más altos como de que crezcan significativamente?
b) ¿Cuál es la probabilidad de que los beneficios de esta empresa crezcan significativamente?
c) Si los beneficios de esta empresa crecen significativamente, ¿cuál es la probabilidad de que los tipos de interés hayan sido más de un 1 por ciento más bajos que este año?

4.75. El 42 por ciento de los obreros de una empresa está a favor de un plan médico modificado y el 22 por ciento de sus obreros está a favor de una propuesta para cambiar el horario de trabajo. El 34 por ciento de los partidarios de la modificación del plan médico es partidario de que se cambie el horario de trabajo.

a) ¿Cuál es la probabilidad de que un obrero seleccionado aleatoriamente esté a favor tanto del plan médico modificado como del cambio del horario de trabajo?
b) ¿Cuál es la probabilidad de que un obrero seleccionado aleatoriamente esté a favor al menos de uno de los dos cambios?
c) ¿Cuál es la probabilidad de que un obrero seleccionado aleatoriamente que esté a favor del cambio del horario de trabajo también sea partidario del plan médico modificado?

4.76. Se han analizado las calificaciones de una clase de estudiantes universitarios de primer curso. El 70 por ciento de los estudiantes del cuarto superior de la clase universitaria había terminado la enseñanza secundaria en el 10 por ciento superior de su clase, al igual que el 50 por ciento de los estudiantes de la mitad central de la clase universitaria y el 20 por ciento de los estudiantes del cuarto inferior de la clase universitaria.

a) ¿Cuál es la probabilidad de que un alumno de primer año seleccionado aleatoriamente estuviera en el 10 por ciento superior de su clase de secundaria?
b) ¿Cuál es la probabilidad de que un alumno de primer año seleccionado aleatoriamente que estuviera en el 10 por ciento superior de su clase de secundaria esté en el cuarto superior de la clase universitaria?
c) ¿Cuál es la probabilidad de que un alumno de primer año seleccionado aleatoriamente que no estuviera en el 10 por ciento superior de su clase de secundaria no esté en el cuarto superior de la clase universitaria?

4.77. Antes de que se comercialicen los libros destinados a los niños de preescolar, se observan las reacciones de un grupo de niños de preescolar. Estas reacciones se dividen en «favorables», «neutrales» o «desfavorables». A continuación, se dividen las ventas de los libros en «altas», «moderadas» o «bajas», según las normas de este mercado. En el pasado se han evaluado 1.000 libros siguiendo este procedimiento. La tabla adjunta muestra sus reacciones y los resultados de los libros en el mercado.

	Reacción del grupo		
Ventas	**Favorable**	**Neutral**	**Desfavorable**
Altas	173	101	61
Moderadas	88	211	70
Bajas	42	113	141

a) Si la reacción del grupo es favorable, ¿cuál es la probabilidad de que las ventas sean altas?
b) Si la reacción del grupo es desfavorable, ¿cuál es la probabilidad de que las ventas sean bajas?
c) Si la reacción del grupo es neutral o mejor, ¿cuál es la probabilidad de que las ventas sean bajas?
d) Si las ventas son bajas, ¿cuál es la probabilidad de que la reacción del grupo fuera neutral o mejor?

4.78. Un fabricante produce cajas de caramelos, cada una de las cuales contiene 10 caramelos. Se utilizan dos máquinas para empaquetarlas. Después de producir un gran lote, se descubre que una de las máquinas, que produce el 40 por ciento de la producción total, tiene un defecto por el que el 10 por ciento de los caramelos que produce tiene una impureza. Se selecciona aleatoriamente un caramelo de una caja y se prueba. Si ese caramelo no contiene ninguna impureza, ¿cuál es la probabilidad de que la máquina defectuosa produjera la caja de la que procede?

4.79. Un estudiante piensa que el 70 por ciento de las asignaturas universitarias ha sido ameno y el resto ha sido aburrido. Este estudiante tiene acceso a las evaluaciones de los profesores realizadas por los estudiantes y observa que los profesores que han recibido anteriormente evaluaciones muy positivas de sus estudiantes han enseñado el 60 por ciento de sus asignaturas amenas y el 25 por ciento de sus asignaturas aburridas. El próximo cuatrimestre el estudiante decide hacer tres asignaturas impartidas todas ellas por profesores que han recibido evaluaciones muy positivas. Suponga que las reacciones del estudiante a las tres asignaturas son independientes unas de otras.
a) ¿Cuál es la probabilidad de que este estudiante piense que las tres asignaturas son amenas?
b) ¿Cuál es la probabilidad de que este estudiante piense que al menos una de las tres asignaturas es amena?

4.5. El teorema de Bayes

En este apartado introducimos un importante resultado que tiene muchas aplicaciones en la toma de decisiones empresariales. El teorema de Bayes permite reconsiderar las probabilidades condicionadas utilizando la información de que se dispone. También permite saber cómo deben ajustarse las estimaciones de la probabilidad, dada la información adicional.

El reverendo Thomas Bayes (1702-1761) desarrolló el teorema de Bayes, publicado inicialmente en 1763 después de su muerte y de nuevo en 1958 (véase la referencia bibliográfica 1). Como los juegos de azar y, por lo tanto, la probabilidad se consideraban obras del demonio, los resultados no fueron muy divulgados. Desde la Segunda Guerra Mundial, se ha desarrollado un importante campo de la estadística y un importante campo de la teoría de las decisiones empresariales, basados en las obras originales de Thomas Bayes. Comenzamos nuestra exposición con un ejemplo seguido de un desarrollo más formal.

EJEMPLO 4.23. Pruebas médicas para detectar el consumo de drogas (teorema de Bayes)

Algunas empresas realizan habitualmente pruebas para detectar si los demandantes de empleo consumen drogas o tienen algunas enfermedades o ambas cosas. Juana Sánchez, presidenta de Buen Tiempo, S.A., ha solicitado un análisis para averiguar si se pueden realizar pruebas a los demandantes de empleo para averiguar si son seropositivos. Los futuros costes médicos de esas personas pueden aumentar espectacularmente el coste del seguro médico de los empleados de la empresa y a Juana le gustaría minimizar las probabilidades de tener que incurrir en esos costes. Supongamos que el 10 por ciento de los demandantes de empleo es seropositivo. Existe, además, una prueba que identifica correctamente el estado de una persona el 90 por ciento de las veces. Si una persona es seropositiva, hay una probabilidad de 0,90 de que la prueba la identifique correctamente. Asimismo, si la persona no es seropositiva, hay una probabilidad de 0,90 de que la prueba identifique correctamente a la persona que no es seropositiva.

Debemos señalar que la negativa a dar empleo basándose en razones de salud puede plantear cuestiones éticas y legales. Naturalmente, esas cuestiones constituyen una parte muy importante de la decisión de hacer la prueba. En este caso, nos interesa la posibilidad de hacer esa prueba si se ha llegado a la conclusión de que es correcto hacerla, dados el sistema jurídico y el sistema de valores.

Solución

El primer paso del análisis es identificar los sucesos contenidos en el espacio muestral:

H_1: La persona es seropositiva.
H_2: La persona no es seropositiva.

La prueba propuesta da resultados positivos o negativos:

T_1: La prueba dice que la persona es seropositiva.
T_2: La prueba dice que la persona no es seropositiva.

Basándose en la información suministrada, pueden definirse las siguientes probabilidades:

$$P(H_1) = 0{,}10 \qquad P(H_2) = 0{,}90$$
$$P(T_1|H_1) = 0{,}90 \qquad P(T_2|H_1) = 0{,}10$$
$$P(T_1|H_2) = 0{,}10 \qquad P(T_2|H_2) = 0{,}90$$

Utilizando estas probabilidades, es posible hacer una tabla de doble entrada que contenga las probabilidades conjuntas:

$$P(H_1 \cap T_1) = P(T_1|H_1)P(H_1) = 0{,}90 \times 0{,}10 = 0{,}09$$
$$P(H_1 \cap T_2) = P(T_2|H_1)P(H_1) = 0{,}10 \times 0{,}10 = 0{,}01$$
$$P(H_2 \cap T_1) = P(T_1|H_2)P(H_2) = 0{,}10 \times 0{,}90 = 0{,}09$$
$$P(H_2 \cap T_2) = P(T_2|H_2)P(H_2) = 0{,}90 \times 0{,}90 = 0{,}81$$

Basándose en la Tabla 4.9, es posible averiguar fácilmente la probabilidad condicionada de ser seropositivo, dado que la prueba dice si una persona es seropositiva, dividiendo la probabilidad conjunta de H_1 y T_1 (0,09) por la probabilidad marginal de T_1 (0,18):

$$P(H_1|T_1) = \frac{P(H_1 \cap T_1)}{P(T_1)} = \frac{0{,}09}{0{,}18} = 0{,}50$$

Tabla 4.9. Subgrupos utilizados para probar el medicamento.

	T_1 (prueba dice seropositivo)	T_2 (prueba dice no seropositivo)	Total
H_1 (seropositivo)	0,09	0,01	0,10
H_2 (no seropositivo)	0,09	0,81	0,90
Total	0,18	0,82	1,0

Asimismo, la probabilidad de que una persona no sea seropositiva, dado que la prueba dice si una persona no es seropositiva, puede hallarse a partir de la segunda columna de la Tabla 4.9:

$$P(H_2|T_2) = \frac{P(H_2 \cap T_2)}{P(T_2)} = \frac{0{,}81}{0{,}82} = 0{,}988$$

Estos resultados nos permiten ver que, si la prueba dice que una persona no es seropositiva, hay una probabilidad muy alta de que el resultado de la prueba sea correcto. Sin embargo, si la prueba dice que la persona es seropositiva, sólo hay una probabilidad del 0,50 de que lo sea. Es un gran aumento con respecto a la probabilidad del 0,10 de una persona seleccionada aleatoriamente. Sin embargo, está claro que la empresa no querría rechazar a los demandantes de empleo basándose simplemente en los resultados de esta prueba. Las posibilidades de que se utilizaran métodos de contratación poco éticos y de que se emprendieran serias acciones legales serían demasiado grandes. La mejor estrategia sería hacer una segunda prueba independiente para seleccionar mejor a las personas que, según la primera, son seropositivas. Hacemos de nuevo hincapié en que la denegación de empleo a una persona por ser seropositiva plantea serias cuestiones éticas y médicas.

Con esta información, formulamos a continuación en términos más formales el teorema de Bayes. En primer lugar, repasamos la regla del producto, la ecuación 4.10:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

El *teorema de Bayes* se deduce de esta regla.

Teorema de Bayes

Sean A y B dos sucesos. El **teorema de Bayes** establece que

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{4.13}$$

y

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Se ha realizado una interesante interpretación del teorema de Bayes en el contexto de las probabilidades subjetivas. Supongamos que una persona está interesada en el suceso B y tiene una opinión subjetiva sobre la probabilidad de que ocurra; en este contexto, la probabilidad $P(B)$ se llama probabilidad *a priori*. Si obtiene entonces más información —a saber, que ha ocurrido el suceso A—, eso puede cambiar su opinión personal sobre la

probabilidad de que ocurra B. Como se sabe que A ha ocurrido, la probabilidad relevante de B ahora es la probabilidad condicionada de B, dado A, y se denomina probabilidad *a posteriori*. Podemos considerar que el teorema de Bayes, visto de esta forma, es un mecanismo para actualizar una probabilidad *a priori* y convertirla en una probabilidad *a posteriori* cuando se dispone de la información de que ha ocurrido A. El teorema establece que la actualización se logra multiplicando la probabilidad *a priori* por $P(A|B)/P(A)$.

Sabemos que la gente normalmente hace valoraciones sobre la probabilidad subjetiva y luego las modifica. Por ejemplo, una parte importante de la labor de un auditor es averiguar si la contabilidad es correcta. Antes de examinar una determinada cuenta, el auditor se habrá formado una opinión, basada en auditorías anteriores, de la probabilidad de que haya un error. Sin embargo, si observa que el saldo es muy diferente de lo que cabría esperar, dadas las cifras de los últimos años, el auditor creerá que la probabilidad de que haya un error es mayor y, por lo tanto, prestará especial atención a esa cuenta. En este caso, la probabilidad *a priori* se ha actualizado a la luz de la información adicional.

EJEMPLO 4.24. Auditoría de las cuentas de una empresa (teorema de Bayes)

Basándose en el examen de la contabilidad anterior de una empresa, un auditor observa que el 15 por ciento contenía errores. Considera que en el 60 por ciento de los saldos contables que contienen errores, los valores son inusuales a juzgar por las cifras anteriores. El 20 por ciento de todos los saldos contables son valores inusuales. Si parece que la cifra de un saldo contable concreto es inusual según este criterio, ¿cuál es la probabilidad de que sea errónea?

Solución

Sea A «error en el saldo contable» y B «valor inusual a juzgar por las cifras anteriores». De la información de la que se dispone se deduce que

$$P(A) = 0{,}15 \qquad P(B) = 0{,}20 \qquad P(B|A) = 0{,}60$$

Utilizando el teorema de Bayes,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{(0{,}60)(0{,}15)}{0{,}20} = 0{,}45$$

Por lo tanto, dada la información de que el saldo contable parece inusual, la probabilidad de que sea erróneo se modifica y pasa de una probabilidad *a priori* de 0,15 a una probabilidad *a posteriori* de 0,45.

El teorema de Bayes se expresa a menudo de una forma diferente, pero equivalente, que utiliza información más detallada. Sean E_1, E_2, ..., E_K K sucesos mutuamente excluyentes y colectivamente exhaustivos y sea A algún otro suceso. Podemos hallar la probabilidad de E_i, dado A, utilizando el teorema de Bayes:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)}$$

El denominador puede expresarse por medio de las probabilidades de A, dados los diversos E_i, utilizando las intersecciones y la regla del producto:

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \cdots + P(A \cap E_K) =$$
$$= P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots + P(A|E_K)P(E_K)$$

Estos resultados pueden combinarse para obtener una segunda formulación del teorema de Bayes.

Teorema de Bayes (formulación alternativa)

Sean E_1, E_2, ..., E_K K sucesos mutuamente excluyentes y colectivamente exhaustivos y sea A algún otro suceso. La probabilidad condicionada de E_i, dado A, puede expresarse como el teorema de Bayes:

$$P(E_i|A) = \frac{P(A|E_i)P(E_i)}{P(A)} =$$
$$= \frac{P(A|E_i)P(E_i)}{P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots + P(A|E_k)P(E_k)} \qquad \textbf{(4.14)}$$

donde

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \cdots + P(A \cap E_k) =$$
$$= P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots + P(A|E_k)P(E_k)$$

La ventaja de esta reformulación del teorema se halla en que las probabilidades que implica a menudo son precisamente las probabilidades de las que se dispone directamente.

Este proceso para hallar la probabilidad condicionada y resolver problemas de Bayes puede resumirse de la forma siguiente.

Pasos para calcular la probabilidad por medio del teorema de Bayes

1. Se definen los sucesos de los subconjuntos, dado el problema.
2. Se definen las probabilidades de los sucesos definidos en el paso 1.
3. Se calculan los complementarios de las probabilidades.
4. Se aplica el teorema de Bayes para calcular la probabilidad que es la solución del problema.

Aquí seguimos estos pasos para resolver un problema que requiere un detenido análisis. Consideramos de nuevo el ejemplo 4.23. La primera tarea es identificar los sucesos en el espacio muestral. En el ejemplo citado, el espacio muestral son los demandantes de empleo divididos en H_1, seropositivos, y H_2, no seropositivos. Para eso fue necesario recurrir a un estudio independiente que determinara qué personas eran realmente seropositivas y cuáles no. Estos sucesos abarcan el espacio muestral. Los sucesos también se identificaron por medio de su clasificación en una prueba. Los sucesos son T_1, la prueba indica que el individuo es seropositivo, y T_2, la prueba indica que el individuo no lo es. Estos sucesos también abarcan el espacio muestral. Obsérvese que un resultado T_1, que indica que el individuo es seropositivo, no garantiza que lo sea, H_1.

Una vez definidos los sucesos, hay que examinar la capacidad del método para hacer predicciones, utilizando los datos. Así, en el ejemplo 4.23 se hizo la prueba a un grupo de personas que se sabía que eran seropositivas y a otro que se sabía que no lo eran. Estos

resultados de la prueba dieron las probabilidades condicionadas de los resultados de la prueba, dado un resultado positivo o no en la prueba. Los datos se convirtieron en información sobre la calidad de las predicciones de la prueba utilizando el teorema de Bayes. La tarea final es expresar una o más cuestiones en forma del teorema de Bayes. En el ejemplo 4.23 nos interesaba saber cuál era la probabilidad de que un demandante de empleo fuera seropositivo, dado que la persona dio un resultado positivo en la prueba. También nos dimos cuenta de que era importante saber cuál era la probabilidad de que una persona no fuera seropositiva, dado que dio un resultado positivo en la prueba.

EJEMPLO 4.25. Incentivos en la venta de automóviles (teorema de Bayes)

Un concesionario de automóviles sabe por experiencia que el 10 por ciento de las personas que entran en la tienda y hablan con un vendedor acaba comprando un automóvil. Para aumentar las posibilidades de éxito, propusimos ofrecer una cena gratis con un vendedor a todas las personas que estuvieran dispuestas a escuchar la presentación completa del vendedor. Sabíamos que algunas personas hacen cualquier cosa por cenar gratis aunque no tengan intención de comprar un automóvil. Sin embargo, algunas prefieren no cenar con un vendedor de automóviles. Por lo tanto, queríamos comprobar la eficacia de este incentivo. El proyecto se realizó durante seis meses y el 40 por ciento de las personas que compraron un automóvil cenó gratis. También cenó gratis el 10 por ciento de las personas que no compraron un automóvil.

Las preguntas para las que queremos encontrar una respuesta son las siguientes:

a) ¿Tienen las personas que aceptan la cena una probabilidad mayor de comprar un automóvil?

b) ¿Qué probabilidad hay de que una persona que no acepta una cena gratis compre un automóvil?

Solución

Paso 1. Definimos los sucesos de los subconjuntos, dado el problema:

D_1: El cliente cena con el vendedor.
D_2: El cliente no cena con el vendedor.
P_1: El cliente compra un automóvil.
P_2: El cliente no compra un automóvil.

Paso 2. Definimos las probabilidades de los sucesos definidos en el paso 1:

$$P(P_1) = 0{,}10 \qquad P(D_1|P_1) = 0{,}40 \qquad P(D_1|P_2) = 0{,}10$$

Paso 3. Calculamos los complementarios de las probabilidades:

$$P(P_2) = 0{,}90 \qquad P(D_2|P_1) = 0{,}60 \qquad P(D_2|P_2) = 0{,}90$$

Paso 4. Aplicamos el teorema de Bayes para calcular la probabilidad que es la solución del problema.

a) Sabemos que el plan de promoción de las ventas ha aumentado la probabilidad de que se compre un automóvil si más del 10 por ciento de los que cenaron compró un automóvil.

Concretamente, preguntamos si

$$P(P_1 \mid D_1) > P(P_1)$$

$$P(P_1 \mid D_1) > 0{,}10$$

Utilizando el teorema de Bayes, observamos que

$$P(P_1 \mid D_1) = \frac{P(D_1 \mid P_1)P(P_1)}{P(D_1 \mid P_1)P(P_1) + P(D_1 \mid P_2)P(P_2)}$$

$$= \frac{0{,}40 \times 0{,}10}{0{,}40 \times 0{,}10 + 0{,}10 \times 0{,}90}$$

$$= 0{,}308$$

Por lo tanto, la probabilidad de que se compre un automóvil es mayor, dada la cena con el vendedor.

b) Se pide que calculemos la probabilidad de que se compre un automóvil, P_1, dado que el cliente no cena con el vendedor, D_2. Aplicando de nuevo el teorema de Bayes, tenemos que

$$P(P_1 \mid P_2) = \frac{P(D_2 \mid P_1)P(P_1)}{P(D_2 \mid P_1)P(P_1) + P(D_2 \mid P_2)P(P_2)}$$

$$= \frac{0{,}60 \times 0{,}10}{0{,}60 \times 0{,}10 + 0{,}90 \times 0{,}90}$$

$$= 0{,}069$$

Vemos que los que rechazan la cena tienen menos probabilidades de comprar un automóvil. Para realizar una evaluación adicional del programa de ventas, también podríamos comparar las ventas realizadas durante 6 meses con las de otros concesionarios y con la de otros programas anteriores, dadas unas condiciones económicas similares.

Hemos presentado paso a paso un método lógico o lineal para resolver problemas de Bayes. Este método funciona muy bien en el caso de las personas que tienen experiencia en la resolución de este tipo de problema. También puede ayudar a organizar los problemas de Bayes. Sin embargo, los problemas reales en situaciones nuevas no se resuelven casi ninguno siguiendo un método paso a paso o lineal. Es probable, pues, que el lector tenga que volver a los pasos anteriores y revisar las definiciones iniciales. En algunos casos, puede resultar útil formular el teorema de Bayes antes de definir las probabilidades. La forma matemática define las probabilidades que deben obtenerse de la descripción del problema. También se puede construir una tabla de doble sentido, como la del ejemplo 4.23. Cuando el lector se disponga a resolver estos problemas, utilice la estructura, pero aprenda a ser creativo y a estar dispuesto a volver a repetir los pasos anteriores.

EJERCICIOS

Ejercicios básicos

Los siguientes ejercicios básicos utilizan un espacio muestral definido por los sucesos A_1, A_2, B_1 y B_2.

4.80. Dados $P(A_1) = 0{,}40$, $P(B_1 \mid A_1) = 0{,}60$ y $P(B_1 \mid A_2) = 0{,}70$, ¿cuál es la probabilidad de $P(A_1 \mid B_1)$?

4.81. Dados $P(A_1) = 0{,}80$, $P(B_1 \mid A_1) = 0{,}60$ y $P(B_1 \mid A_2) = 0{,}20$, ¿cuál es la probabilidad de $P(A_1 \mid B_1)$?

4.82. Dados $P(A_1) = 0{,}50$, $P(B_1 \mid A_1) = 0{,}40$ y $P(B_1 \mid A_2) = 0{,}70$, ¿cuál es la probabilidad de $P(A_1 \mid B_2)$?

4.83. Dados $P(A_1) = 0{,}40$, $P(B_1 \mid A_1) = 0{,}60$ y $P(B_1 \mid A_2) = 0{,}70$, ¿cuál es la probabilidad de $P(A_2 \mid B_2)$?

4.84. Dados $P(A_1) = 0{,}60$, $P(B_1 \mid A_1) = 0{,}60$ y $P(B_1 \mid A_2) = 0{,}40$, ¿cuál es la probabilidad de $P(A_1 \mid B_1)$?

Ejercicios aplicados

4.85. Una editorial envía publicidad de un libro de texto de contabilidad al 80 por ciento de todos los profesores que imparten la asignatura de contabilidad. El 30 por ciento de los profesores que recibe esta publicidad adopta el libro, al igual que el 10 por ciento de los que no la reciben. ¿Cuál es la probabilidad de que un profesor que adopta el libro haya recibido la publicidad?

4.86. Un analista bursátil examinó las perspectivas de las acciones de un gran número de empresas. Cuando analizó los resultados de estas acciones un año más tarde, resultó que el 25 por ciento obtuvo unos resultados mucho mejores que la media, el 25 por ciento obtuvo unos resultados mucho peores y el 50 por ciento restante obtuvo unos resultados parecidos a la media. El 40 por ciento de las acciones que obtuvieron unos resultados mucho mejores que la media fueron calificados de «buenas compras» por el analista, al igual que el 20 por ciento de los que obtuvieron unos resultados parecidos a la media y el 10 por ciento de los que obtuvieron unos resultados mucho peores que la media. ¿Cuál es la probabilidad de que una acción calificada de «buena compra» por el analista obtuviera unos resultados mucho mejores que la media?

RESUMEN

En este capítulo hemos introducido las ideas básicas de la probabilidad. Un riguroso conjunto de definiciones y reglas permite desarrollar métodos para resolver el núcleo de problemas de probabilidad que se plantean en el mundo de la empresa y en economía. Hemos desarrollado estos métodos para resolver problemas utilizando las probabilidades conjuntas, las probabilidades marginales, la independencia, las probabilidades condicionadas, los cocientes de «sobreparticipación» y el teorema de Bayes. Los métodos para resolver problemas son las ecuaciones, los diagramas de Venn y las tablas de doble entrada.

TÉRMINOS CLAVE

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

4.87. Suponga que tiene un amigo inteligente que no ha estudiado probabilidad. ¿Cómo le explicaría la distinción entre sucesos mutuamente excluyentes y sucesos independientes? Ilustre su respuesta con ejemplos adecuados.

4.88. Indique si cada una de las afirmaciones siguientes es verdadera o falsa y arguméntelo.

a) El complementario de la unión de dos sucesos es la intersección de sus complementarios.
b) La suma de las probabilidades de sucesos colectivamente exhaustivos debe ser igual a 1.
c) El número de combinaciones de x objetos extraídos de n es igual al número de combinaciones de $(n - x)$ objetivos extraídos de n, donde $1 \leqslant x \leqslant (n - 1)$.
d) Si A y B son dos sucesos, la probabilidad de A, dado B, es igual que la probabilidad de B, dado A, si la probabilidad de A es igual que la probabilidad de B.
e) Si un suceso y su complementario son igual de probables, la probabilidad de ese suceso debe ser 0,5.
f) Si A y B son independientes, entonces $\bar{A}$ y $\bar{B}$ deben ser independientes.
g) Si A y B son mutuamente excluyentes, entonces $\bar{A}$ y $\bar{B}$ deben ser mutuamente excluyentes.

4.89. Explique detenidamente el significado de probabilidad condicionada. ¿Por qué es importante este concepto en el análisis de la probabilidad de que ocurra un suceso?

4.90. «El teorema de Bayes es importante, porque es una regla para pasar de una probabilidad *a priori* a una probabilidad *a posteriori*». Explique esta afirmación de manera que la entienda perfectamente un compañero que aún no haya estudiado probabilidad.

4.91. Indique si cada una de las afirmaciones siguientes es verdadera o falsa y arguméntelo:

a) La probabilidad de la unión de dos sucesos no puede ser menor que la probabilidad de su intersección.
b) La probabilidad de la unión de dos sucesos no puede ser mayor que la suma de sus probabilidades individuales.
c) La probabilidad de la intersección de dos sucesos no puede ser mayor que cualquiera de sus probabilidades individuales.
d) Un suceso y su complementario son mutuamente excluyentes.
e) Las probabilidades individuales de un par de sucesos no pueden sumar más de 1.
f) Si dos sucesos son mutuamente excluyentes, también deben ser colectivamente exhaustivos.
g) Si dos sucesos son colectivamente exhaustivos, también deben ser mutuamente excluyentes.

4.92. Distinga entre probabilidad conjunta, probabilidad marginal y probabilidad condicionada. Ponga algunos ejemplos para aclarar las distinciones.

4.93. Indique si cada una de las afirmaciones siguientes es verdadera o falsa y explique su respuesta:

a) La probabilidad condicionada de A, dado B, debe ser como mínimo tan grande como la probabilidad de A.
b) Un suceso debe ser independiente de su complementario.
c) La probabilidad de A, dado B, debe ser como mínimo tan grande como la probabilidad de la intersección de A y B.
d) La probabilidad de la intersección de dos sucesos no puede ser superior al producto de sus probabilidades individuales.
e) La probabilidad *a posteriori* de un suceso debe ser como mínimo tan grande como su probabilidad *a priori*.

4.94. Demuestre que la probabilidad de la unión de los sucesos A y B puede expresarse de la forma siguiente:

$$P(A \cup B) = P(A) + P(B)[1 - P(A \mid B)]$$

4.95. Una compañía de seguros estimó que el 30 por ciento de todos los accidentes de tráfico se debía en parte a las condiciones meteorológicas y que en el 20 por ciento había heridos. Además, el 40 por ciento de los accidentes en los que había heridos se debía en parte a las condiciones meteorológicas.

a) ¿Cuál es la probabilidad de que un accidente seleccionado aleatoriamente se debiera en parte a las condiciones meteorológicas y en él hubiera heridos?
b) ¿Son independientes los sucesos «debido en parte a las condiciones meteorológicas» y «hubo heridos»?
c) Si un accidente seleccionado aleatoriamente se debió en parte a las condiciones meteorológicas, ¿qué probabilidad hay de que hubiera heridos?

d) ¿Cuál es la probabilidad de que un accidente seleccionado aleatoriamente no se debiera en parte a las condiciones meteorológicas y en él no hubiera heridos?

4.96. Una empresa hace un pedido urgente de alambre de dos tipos de grosor que debe enviársele en cuanto se disponga de él. La experiencia dice que hay una probabilidad de 0,8 de que al menos uno de los pedidos llegue antes de una semana. También se estima que si el alambre más fino llega antes de una semana, hay una probabilidad de 0,4 de que el alambre más grueso también llegue antes de una semana. Se estima, además, que si el alambre más grueso llega antes de una semana, hay una probabilidad de 0,6 de que el más fino también llegue antes de una semana.

a) ¿Qué probabilidad hay de que el alambre más grueso llegue antes de una semana?
b) ¿Qué probabilidad hay de que el alambre más fino llegue antes de una semana?
c) ¿Qué probabilidad hay de que ambos pedidos lleguen antes de una semana?

4.97. Basándose en una encuesta realizada a estudiantes de una gran universidad, se estimó que el 35 por ciento bebe al menos una vez a la semana en los bares locales y que el 40 por ciento tiene una calificación media de notable o más. Además, el 30 por ciento de los que beben al menos una vez a la semana en bares locales tiene una calificación media de notable o más.

a) ¿Cuál es la probabilidad de que un estudiante seleccionado aleatoriamente beba al menos una vez a la semana en bares locales y tenga una calificación media de notable o más?
b) ¿Cuál es la probabilidad de que un estudiante seleccionado aleatoriamente que tenga una calificación media de notable o más beba al menos una vez a la semana en bares locales?
c) ¿Cuál es la probabilidad de que un estudiante seleccionado aleatoriamente tenga al menos una de estas características: «bebe al menos una vez a la semana en bares locales» y «tiene una calificación media de notable o más»?
d) ¿Cuál es la probabilidad de que un estudiante seleccionado aleatoriamente que no tiene una calificación media de notable o más no beba al menos una vez a la semana en bares locales?
e) ¿Son independientes los sucesos «bebe al menos una vez a la semana en bares locales» y «tiene una calificación media de notable o más»?
f) ¿Son mutuamente excluyentes los sucesos «bebe al menos una vez a la semana en bares locales» y «tiene una calificación media de notable o más»?
g) ¿Son colectivamente exhaustivos los sucesos «bebe al menos una vez a la semana en bares locales» y «tiene una calificación media de notable o más»?

4.98. En el comedor de un campus universitario se observó que el 35 por ciento de todos los clientes pedía platos calientes y el 50 por ciento eran estudiantes. Además, el 25 por ciento de todos los clientes que eran estudiantes pedía platos calientes.

a) ¿Cuál es la probabilidad de que un cliente seleccionado aleatoriamente fuera estudiante y pidiera platos calientes?
b) Si un cliente seleccionado aleatoriamente pedía platos calientes, ¿cuál es la probabilidad de que fuera estudiante?
c) ¿Cuál es la probabilidad de que un cliente seleccionado aleatoriamente no pidiera platos calientes y no fuera estudiante?
d) ¿Son independientes los sucesos «el cliente pide platos calientes» y «el cliente es estudiante»?
e) ¿Son mutuamente excluyentes los sucesos «el cliente pide platos calientes» y «el cliente es estudiante?
f) ¿Son colectivamente exhaustivos los sucesos «el cliente pide platos calientes» y «el cliente es estudiante»?

4.99. Se sabe que el 20 por ciento de todas las explotaciones agrícolas de una región tiene más de 160 acres y que el 60 por ciento de todas las explotaciones agrícolas de esa región pertenece a personas de más de 50 años. El 55 por ciento de todas las explotaciones agrícolas de la región de más de 160 acres es propiedad de personas de más de 50 años.

a) ¿Cuál es la probabilidad de que una explotación agrícola seleccionada aleatoriamente en esta región tenga más de 160 acres y sea propiedad de una persona de más de 50 años?
b) ¿Cuál es la probabilidad de que una explotación agrícola de esta región tenga más de 160 acres o sea propiedad de una persona de más de 50 años (o ambas cosas)?
c) ¿Cuál es la probabilidad de que una explotación agrícola de esta región, propiedad de una persona de más de 50 años, tenga más de 160 acres?
d) ¿Son estadísticamente independientes la extensión de la explotación y la edad del propietario en esta región?

4.100. En una gran empresa, el 80 por ciento de los empleados son hombres y el 20 por ciento son mujeres. Por lo que se refiere a los hombres, el 10 por ciento tiene estudios de postgrado, el 30 por ciento tiene una licenciatura y el 60 por ciento tiene estudios de secundaria. En el caso de las mujeres, el 15 por ciento tiene estudios de postgrado, el 40 por ciento tiene una licenciatura y el 45 por ciento tiene estudios de secundaria.

a) ¿Cuál es la probabilidad de que un empleado seleccionado aleatoriamente sea un hombre que sólo tiene estudios de secundaria?
b) ¿Cuál es la probabilidad de que un empleado seleccionado aleatoriamente tenga estudios de postgrado?
c) ¿Cuál es la probabilidad de que un empleado seleccionado aleatoriamente que tiene estudios de postgrado sea un hombre?
d) ¿Son el sexo y el nivel de estudios de los empleados de esta empresa estadísticamente independientes?
e) ¿Cuál es la probabilidad de que un empleado seleccionado aleatoriamente que no tiene estudios de postgrado sea una mujer?

4.101. Una gran empresa sometió a votación entre todos sus trabajadores un nuevo plan de primas. Se observó que era partidario del plan el 65 por ciento de todos los trabajadores del turno de noche y el 40 por ciento de todas las mujeres. Además, el 50 por ciento de todos los trabajadores estaba en el turno de noche y el 30 por ciento de todos eran mujeres. Por último, el 20 por ciento de todos los trabajadores del turno de noche eran mujeres.

a) ¿Cuál es la probabilidad de que un empleado seleccionado aleatoriamente sea una mujer partidaria del plan?
b) ¿Cuál es la probabilidad de que un empleado seleccionado aleatoriamente sea una mujer o un trabajador del turno de noche (o ambas cosas)?
c) ¿Es el sexo del trabajador independiente de que trabaje o no en el turno de noche?
d) ¿Cuál es la probabilidad de que una empleada trabaje en el turno de noche?
e) Si el 50 por ciento de todos los empleados varones es partidario del plan, ¿cuál es la probabilidad de que un empleado seleccionado aleatoriamente no trabaje en el turno de noche y no sea partidario del plan?

4.102. Hay que elegir a un jurado de 12 miembros de entre 8 hombres y 8 mujeres.

a) ¿Cuántas selecciones son posibles?
b) Si la selección se hace aleatoriamente, ¿cuál es la probabilidad de que la mayoría de los miembros del jurado sean hombres?

4.103. Un envío de 12 componentes electrónicos contiene 1 componente defectuoso. Se seleccionan aleatoriamente dos para probarlos.

a) ¿Cuántas combinaciones de 2 componentes podrían seleccionarse?
b) ¿Cuál es la probabilidad de que se seleccione el componente defectuoso para probarlo?

4.104. De 100 pacientes que padecían una determinada enfermedad, se eligieron 10 aleatoriamente para someterlos a un tratamiento farmacológico que aumenta la tasa de curación del 50 por ciento en el caso de los que no reciben el tratamiento al 75 por ciento en el caso de los que reciben el tratamiento.

a) ¿Cuál es la probabilidad de que un paciente seleccionado aleatoriamente se curara y recibiera el tratamiento?
b) ¿Cuál es la probabilidad de que un paciente que se curó hubiera recibido el tratamiento?
c) ¿Cuál es la probabilidad de que se eligiera un grupo específico de 10 pacientes para recibir el tratamiento? Exprese sus resultados en factoriales.

4.105. Las suscripciones a una revista se clasifican en regalos, renovaciones anteriores, correo directo o servicio de suscripción. En enero, el 8 por ciento de las suscripciones que expiraron eran regalos; el 41 por ciento eran renovaciones anteriores; el 6 por ciento era correo directo, y el 45 por ciento era servicio de suscripción. Los porcentajes de renovaciones en estas cuatro categorías eran 81, 79, 60 y 21 por ciento, respectivamente. En febrero de ese mismo año, el 10 por ciento de las suscripciones que expiraron eran regalos; el 57 por ciento eran renovaciones anteriores; el 24 por ciento era correo directo, y el 9 por ciento era servicio de suscripción. Los porcentajes de renovaciones eran 80, 76, 51 y 14 por ciento, respectivamente.

a) Halle la probabilidad de que una suscripción seleccionada aleatoriamente que expiraba en enero se renovara.
b) Halle la probabilidad de que una suscripción seleccionada aleatoriamente que expiraba en febrero se renovara.
c) Verifique que la probabilidad del apartado (b) es mayor que la del apartado (a). ¿Cree que los directores de esta revista deben con-

siderar que el cambio de enero a febrero es positivo o negativo?

4.106. En una gran ciudad, el 8 por ciento de los habitantes ha contraído una enfermedad. Se realiza una prueba y el resultado es positivo en el 80 por ciento de las personas que tienen la enfermedad y negativo en el 80 por ciento de las personas que no la tienen. ¿Cuál es la probabilidad de que tenga la enfermedad una persona cuya prueba ha dado un resultado positivo?

4.107. Un vendedor de seguros de vida observa que el 70 por ciento de las personas a las que vende un seguro ya tiene una póliza. También observa que el 50 por ciento de todas las personas con las que contacta y a las que no vende un seguro ya tiene una póliza. Además, consigue vender una póliza al 40 por ciento de las personas con las que contacta. ¿Cuál es la probabilidad de que venda una póliza a una persona que ya tiene una?

4.108. Un profesor observa que pone una calificación final de sobresaliente al 20 por ciento de los estudiantes. El 70 por ciento de los que obtienen una calificación final de sobresaliente obtuvo una calificación de sobresaliente en el examen parcial. Además, el 10 por ciento de los estudiantes que no obtiene una calificación final de sobresaliente obtuvo un sobresaliente en el examen parcial. ¿Cuál es la probabilidad de que un estudiante que obtuvo un sobresaliente en el examen parcial obtenga una calificación final de sobresaliente?

4.109. La tabla adjunta muestra el número de predicciones de los beneficios por acción de 1.000 empresas realizadas por analistas financieros y los resultados (en comparación con el año anterior) divididos en tres categorías.

	Predicción		
Resultado	**Mejores**	**Más o menos iguales**	**Peores**
Mejores	210	82	66
Más o menos iguales	106	153	75
Peores	75	84	149

a) Halle la probabilidad de que si se predice que los beneficios disminuirán se obtendrá este resultado.

b) Si se predice que los beneficios mejorarán, halle la probabilidad de que no se obtenga este resultado.

4.110. Un decano ha observado que el 62 por ciento de los estudiantes de primer año y el 78 por ciento de los estudiantes procedentes de programas de formación profesional acaban licenciándose. El 73 por ciento de todos los nuevos estudiantes son estudiantes de primer año y los restantes son estudiantes procedentes de programas de formación profesional.

a) ¿Cuál es la probabilidad de que un nuevo estudiante seleccionado aleatoriamente sea un estudiante de primer año que acabará licenciándose?

b) Halle la probabilidad de que un nuevo estudiante seleccionado aleatoriamente acabe licenciándose.

c) ¿Cuál es la probabilidad de que un nuevo estudiante seleccionado aleatoriamente sea un estudiante de primer año o acabe licenciándose (o ambas cosas)?

d) ¿Son independientes los sucesos «acaba licenciándose» y «procede de un programa de formación profesional»?

4.111. Un grupo de estudios de mercado se especializa en evaluar las perspectivas de los locales para abrir nuevas tiendas de ropa en centros comerciales. El grupo considera que las perspectivas son buenas, razonables o malas. Se han examinado las valoraciones realizadas por este grupo y se ha observado que en el caso de todas las tiendas que han tenido éxito, el grupo había dicho que las perspectivas eran buenas en el 70 por ciento, razonables en el 20 por ciento y malas en el 10 por ciento. De todas las tiendas que fracasaron, había dicho que las perspectivas eran buenas en el 20 por ciento, razonables en el 30 por ciento y malas en el 50 por ciento. Se sabe que el 60 por ciento de las nuevas tiendas de ropa tiene éxito y el 40 por ciento fracasa.

a) ¿Cuál es la probabilidad de que el grupo considere buenas las perspectivas de una tienda seleccionada aleatoriamente?

b) Si las perspectivas de una tienda se consideran buenas, ¿cuál es la probabilidad de que tenga éxito?

c) ¿Son estadísticamente independientes los sucesos «las perspectivas son buenas» y «la tienda tiene éxito»?

d) Suponga que se eligen aleatoriamente cinco tiendas. ¿Cuál es la probabilidad de que al menos una tenga éxito?

4.112. El director de un restaurante clasifica a los clientes en bien vestidos, vestidos normalmente y mal vestidos y observa que el 50, el 40 y el

10 por ciento de todos los clientes, respectivamente, pertenecen a estas categorías. Observa que el 70 por ciento de los clientes bien vestidos, el 50 por ciento de los que van vestidos normalmente y el 30 por ciento de los que van mal vestidos piden vino.

a) ¿Cuál es la probabilidad de que un cliente seleccionado aleatoriamente pida vino?
b) Si se pide vino, ¿cuál es la probabilidad de que la persona que lo pide vaya bien vestida?
c) Si se pide vino, ¿cuál es la probabilidad de que la persona que lo pide no vaya bien vestida?

4.113. El dueño de una tienda de discos divide a los clientes que entran en su tienda en clientes en edad escolar, clientes en edad universitaria y clientes mayores y observa que el 30, el 50 y el 20 por ciento de todos los clientes, respectivamente, pertenecen a estas categorías. También observa que compra discos el 20 por ciento de los clientes en edad escolar, el 60 por ciento de los clientes en edad universitaria y el 80 por ciento de los clientes mayores.

a) ¿Cuál es la probabilidad de que un cliente seleccionado aleatoriamente compre un disco?
b) Si un cliente seleccionado aleatoriamente compra un disco, ¿cuál es la probabilidad de que esté en edad escolar?

4.114. Obsérvese que este ejercicio representa una situación absolutamente imaginaria. Suponga que en una clase de estadística hay exactamente 8 hombres y 8 mujeres. Ha descubierto que el profesor ha decidido suspender a 5 personas en un examen extrayendo aleatoriamente los nombres de un sombrero. Ha llegado a la conclusión de que es más fácil que calificar todos los trabajos de curso y que todos sus estudiantes tienen los mismos conocimientos de estadística, pero alguien tiene que suspender. ¿Cuál es la probabilidad de que los 5 suspendidos sean hombres?

4.115. Se ha cometido un robo y se le ha encomendado la investigación a Maqueda, un sabueso en la lucha contra la delincuencia. Descubre que Sara Manosfrías fue vista portando guantes en las cercanías poco después del delito, por lo que llega a la conclusión de que debe ser detenida. Usted sabe por experiencia que el 50 por ciento de las personas que Maqueda dice que deben ser detenidas por robo son realmente culpables. Antes de realizar la detención, usted pide algunas investigaciones más. Observa que en una gran población de ladrones convictos el 60 por ciento llevaba guantes en el momento del delito y continuó llevándolos durante un tiempo después. Otra investigación revela que el 80 por ciento de las personas que se encontraban en las inmediaciones llevaba guantes en el momento del delito.

a) Basándose en el hecho de que Sara llevaba guantes, ¿cuál es la probabilidad de que Sara cometiera realmente el delito?
b) Si la acusara del delito, ¿cree que un jurado la condenaría basándose en la evidencia de los guantes? Explique por qué sí o por qué no.

4.116. Usted es responsable de detectar la fuente del error cuando falla el sistema informático. De su análisis se desprende que la fuente del error es la unidad de disco, la memoria o el sistema operativo. Sabe que el 50 por ciento de los errores son errores de la unidad de disco, el 30 por ciento son errores de la memoria y el resto son errores del sistema operativo. Según las especificaciones técnicas de los componentes, sabe que cuando el error es de la unidad de disco, la probabilidad de que falle el sistema informático es de 0,60; que cuando el error es de la memoria, la probabilidad de que falle el sistema informático es de 0,7; y que cuando el error es del sistema operativo, la probabilidad de que falle el sistema informático es de 0,4. Dada la información de las especificaciones técnicas de los componentes, ¿cuál es la probabilidad de que el error sea de la unidad de disco, dado que hubo un fallo en el sistema informático?

4.117. Tras reunirse con los directores regionales de ventas, Laura András, presidenta de una empresa de computadores, cree que la probabilidad de que aumenten las ventas un 10 por ciento el próximo año es de 0,70. Tras llegar a esa conclusión, recibe un informe de que Juan Candamo, presidente de una empresa de programas informáticos, acaba de anunciar un nuevo sistema operativo que estará a la venta dentro de 8 meses. Sabe por experiencia que en las situaciones en las que han acabado aumentando las ventas, se han anunciado sistemas operativos el 30 por ciento de las veces. Sin embargo, en las situaciones en las que las ventas no han acabado aumentando, se han anunciado nuevos sistemas operativos el 10 por ciento de las veces. Basándose en todos estos hechos, ¿cuál es la probabilidad de que las ventas crezcan un 10 por ciento?

Apéndice: permutaciones y combinaciones

Una dificultad práctica que se plantea a veces cuando se calcula la probabilidad de un suceso es contar el número de resultados básicos en el espacio muestral y el suceso de interés. En algunos problemas, puede ser útil el uso de *permutaciones* o de *combinaciones*.

1. Número de ordenaciones

Comenzamos con el problema de la ordenación. Supongamos que tenemos un número x de objetos que hay que ordenar. Cada uno sólo puede utilizarse una vez. ¿Cuántas series diferentes son posibles? Podemos imaginar que en este problema se nos pide que coloquemos cada uno de los objetos en cada una de las x cajas colocadas en fila.

Comenzando por la caja situada a la izquierda en la Figura 4.11, hay x formas de llenarla. Una vez que se coloca un objeto en esa caja, quedan $(x - 1)$ objetos, por lo que hay $(x - 1)$ formas de llenar la segunda caja. Es decir, para cada una de las x formas de colocar un objeto en la primera caja, hay $(x - 1)$ formas posibles de llenar la segunda caja, por lo que las dos primeras cajas pueden llenarse de un total de $x \times (x - 1)$ formas. Dado que las dos primeras cajas están llenas, ahora hay $(x - 2)$ formas de llenar la tercera, por lo que las tres primeras pueden llenarse de un total de $x \times (x - 1) \times (x - 2)$ formas. Cuando llegamos a la última caja, sólo queda un objeto para llenarla. Tenemos finalmente el número de ordenaciones posibles.

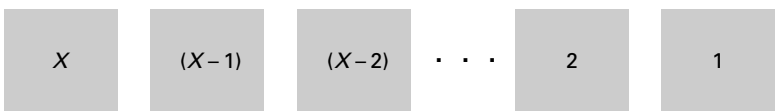


Figura 4.11. Las ordenaciones de x objetos.

Número de ordenaciones posibles

El número total de formas posibles de ordenar x objetos viene dado por

$$x(x - 1)(x - 2) \cdots (2)(1) = x! \qquad \textbf{(4.15)}$$

donde $x!$ es «x factorial».

2. Permutaciones

Supongamos que ahora tenemos un número n de objetos con los que podrían llenarse x cajas *ordenadas* (siendo $n > x$). Cada objeto sólo puede utilizarse una vez. El número de ordenaciones posibles se llama número de *permutaciones* de x objetos elegidos de n y se representa por medio del símbolo P_x^n.

Ahora podemos hacer el mismo razonamiento que antes, salvo que habrá n formas de llenar la primera caja, $(n - 1)$ formas de llenar la segunda, y así sucesivamente, hasta que llegar a la última caja. En ese momento, quedarán $(n - x + 1)$ objetos, cada uno de los cuales podría colocarse en esa caja, como muestra la Figura 4.12.

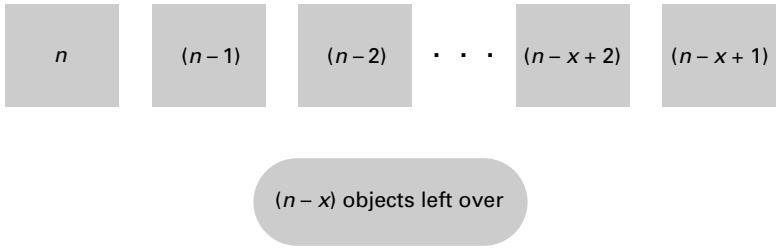


Figura 4.12. Las permutaciones de *x* objetos elegidos de *n*.

Permutaciones

El número total de **permutaciones** de *x* objetos elegidos de *n*, P_x^n, es el número de ordenaciones posibles cuando se seleccionan *x* objetos de un total de *n* y se ordenan.

$$P_x^n = n(n-1)(n-2)\cdots(n-x+1) \tag{4.16}$$

Multiplicando y dividiendo la ecuación 4.16 por

$$(n-x)(n-x-1)\cdots(2)(1) = (n-x)!$$

tenemos que

$$P_x^n = \frac{n(n-1)(n-2)\cdots(n-x+1)(n-x)(n-x-1)\cdots(2)(1)}{(n-x)(n-x-1)\cdots(2)(1)}$$

o sea

$$P_x^n = \frac{n!}{(n-x)!} \tag{4.17}$$

EJEMPLO 4.26. Cinco letras (permutaciones)

Supongamos que hay que seleccionar dos letras de A, B, C, D y E y colocarlas en orden. ¿Cuántas permutaciones son posibles?

Solución

El número de permutaciones, siendo $n = 5$ y $x = 2$, es

$$P_2^5 = \frac{5!}{3!} = 20$$

Éstas son

AB	AC	AD	AE	BC
BA	CA	DA	EA	CB
BD	BE	CD	CE	DE
DB	EB	DC	EC	ED

3. Combinaciones

Supongamos, por último, que nos interesa saber cuál es el número de formas en que pueden seleccionarse x objetos de n (donde ningún objeto puede elegirse más de una vez), pero *no nos interesa el orden*. Obsérvese que en el ejemplo 4.26 las entradas de la segunda fila y la cuarta son simplemente reordenaciones de las que se encuentran directamente encima de ellas, por lo que podemos dejarlas de lado. Por lo tanto, sólo hay 10 posibilidades de elegir 2 objetos de un grupo de 5 si el orden no es importante. El número de selecciones posibles se llama número de *combinaciones* y se representa por medio de C_x^n, donde hay que elegir x de n. Para hallar este número, obsérvese primero que el número de permutaciones posibles es P_x^n. Sin embargo, muchas son reordenaciones de los mismos x objetos, por lo que son irrelevantes. De hecho, como x objetos pueden ordenarse de $x!$ formas, sólo nos interesa una proporción $1/x!$ de las permutaciones. Eso nos lleva a un resultado antes formulado, a saber, la ecuación 4.5 del apartado 4.2, que repetimos aquí para que el análisis quede más completo.

Número de combinaciones

El **número de combinaciones**, C_x^n, de x objetos elegidos de n es el número de selecciones posibles que pueden realizarse. Este número es

$$C_x^n = \frac{P_x^n}{x!}$$

o simplemente

$$C_x^n = \frac{n!}{x!\,(n-x)!} \tag{4.18}$$

EJEMPLO 4.27. Probabilidad de selección de empleados (combinaciones)

Un jefe de personal tiene 8 candidatos para cubrir 4 puestos parecidos. Cinco son hombres y tres son mujeres. Si todas las combinaciones de candidatos tienen las mismas probabilidades de ser elegidas, ¿cuál es la probabilidad de que no se contrate a ninguna mujer?

Solución

En primer lugar, el número total de combinaciones posibles de 4 candidatos elegidos de 8 es

$$C_4^8 = \frac{8!}{4!\,4!} = 70$$

Ahora bien, para que no se contrate a ninguna mujer, los 4 candidatos seleccionados deben proceder de los 5 hombres. El número de esas combinaciones es

$$C_4^5 = \frac{5!}{4!\,1!} = 5$$

Por lo tanto, si al principio cada una de las 70 combinaciones posibles tenía la misma probabilidad de ser elegida, la probabilidad de que se eligiera 1 de las 5 combinaciones formadas únicamente por hombres es $5/70 = 1/14$.

Bibliografía

1. Bayes, Thomas, «Essay Towards Solving a Problem in the Doctrine of Chance», *Biometrika*, 1958, 45, págs. 293-315 (reproducción de un artículo de 1763).
2. Carlson, William L., «Alcohol Usage of the Night Driver», *Journal of Safety Research*, marzo, 1972, 4, n.º 1, págs. 12-29.
3. Carlson, William L. y Betty Thorne, *Applied Statistical Methods for Business and Economics*, Upper Saddle River, NJ, Prentice Hall, 1997.

Capítulo 5

Variables aleatorias discretas y distribuciones de probabilidad

Esquema del capítulo

Introducción

En el Capítulo 4 comenzamos nuestro análisis de la probabilidad para representar situaciones en las que los resultados son inciertos. En éste nos basamos en esas ideas para presentar modelos de probabilidad que ponen énfasis en las variables aleatorias discretas. En el 6 desarrollamos modelos de probabilidad para variables aleatorias continuas.

Los modelos de probabilidad tienen muchas aplicaciones en algunos problemas empresariales; aquí analizamos algunas de ellas. Supongamos que tenemos una tienda que alquila toda una variedad de equipo. Sabemos por experiencia —frecuencia relativa— que el 30 por ciento de las personas que entran en nuestra tienda quiere alquilar una caravana. Hoy tenemos tres caravanas. Cinco personas que no guardan ninguna relación entre sí entran en la tienda (la probabilidad de que una de ellas alquile una caravana es independiente de la de las demás). ¿Cuál es la probabilidad de que estas cinco personas quieran alquilar un total de cuatro o cinco caravanas? Si ocurre eso, perderemos oportunidades de alquilar caravanas y los clientes se irán decepcionados. La probabilidad de los sucesos (número de caravanas deseadas) puede calcularse utilizando el modelo binomial que presentamos en este capítulo.

5.1. Variables aleatorias

Cuando los resultados son valores numéricos, estas probabilidades pueden resumirse por medio del concepto de *variable aleatoria*.

Variable aleatoria

Una **variable aleatoria** es una variable que toma valores numéricos determinados por el resultado de un experimento aleatorio.

Es importante distinguir entre una variable aleatoria y los valores posibles que puede tomar. Hacemos la distinción utilizando letras mayúsculas, como X, para representar la variable aleatoria y la correspondiente letra minúscula, x, para representar un valor posible. Por ejemplo, antes de observar los resultados del lanzamiento de un dado al aire, podemos utilizar la variable aleatoria X para representar el resultado. Esta variable aleatoria puede tomar los valores específicos $x = 1$, $x = 2$, ..., $x = 6$, cada uno con una probabilidad $P(X = 2) = \cdots = P(X = 6) = \frac{1}{6}$.

También es importante distinguir entre *variables aleatorias discretas* y *variables aleatorias continuas*. El lanzamiento del dado al aire es un ejemplo de las primeras; sólo hay seis resultados posibles, cada uno con una probabilidad.

Variable aleatoria discreta

Una variable aleatoria es una **variable aleatoria discreta** si no puede tomar más que una cantidad numerable de valores.

De esta definición se deduce que cualquier variable aleatoria que sólo puede tomar un número finito de valores es discreta. Por ejemplo, el número de veces que sale cara cuando se lanza 10 veces al aire una moneda es una variable aleatoria discreta. Aunque el número de resultados posibles sea infinito pero numerable, la variable aleatoria es discreta. Un ejemplo es el número de veces que hay que lanzar una moneda al aire para que salga cara por primera vez. Los resultados posibles son 1, 2, 3 ..., cada uno con una probabilidad (en el apartado 5.6 se analizará una variable aleatoria discreta que puede tomar un número infinito numerable de valores). He aquí algunos otros ejemplos de variables aleatorias discretas:

1. El número de artículos defectuosos de una muestra de 20 artículos procedente de un gran envío.
2. El número de clientes que llegan a la caja de un supermercado en una hora.
3. El número de errores detectados en las cuentas de una empresa.
4. El número de reclamaciones en una póliza de seguro médico en un año.

Supongamos, por el contrario, que nos interesa saber cuál es la temperatura máxima del día. La variable aleatoria, «temperatura», se mide en un continuo y por eso se dice que es *continua*.

Variable aleatoria continua

Una variable aleatoria es una **variable aleatoria continua** si puede tomar cualquier valor de un intervalo.

En el caso de las variables aleatorias continuas, no podemos asignar probabilidades a valores específicos. Por ejemplo, la probabilidad de que la temperatura máxima de hoy sea exactamente 12,537 °C es 0. Naturalmente, la temperatura no será *exactamente* esa cifra. Sin embargo, es posible determinar la probabilidad correspondiente a intervalos, por lo que podemos asignar una probabilidad al suceso «la temperatura máxima de hoy estará entre 10° y 15 °C». He aquí algunos otros ejemplos de variables aleatorias continuas:

1. La renta anual de una familia.
2. La cantidad de petróleo importado en un mes.
3. La variación del precio de las acciones ordinarias de IBM en un mes.
4. El tiempo que transcurre desde que se instala un nuevo componente hasta que se avería.
5. El porcentaje de impurezas que hay en un lote de productos químicos.

Tal vez parezca bastante artificial la distinción que hemos hecho entre variables aleatorias discretas y variables aleatorias continuas. Al fin y al cabo, raras veces se mide realmente algo en un continuo. Por ejemplo, no podemos medir la temperatura máxima de un día con más precisión de lo que permite el instrumento de medición. Por otra parte, la renta anual de una familia es un número entero de centavos. Sin embargo, observaremos que es cómodo actuar como si las mediciones se hubieran realizado realmente en un continuo cuando las diferencias entre los valores adyacentes son insignificantes. La diferencia entre una renta familiar de 35.276,21 $ y una renta familiar de 35.276,22 $ no tiene mucha importancia y la asignación de probabilidades a cada una de ellas sería un ejercicio tedioso e inútil.

A efectos prácticos, consideramos que las variables aleatorias son discretas cuando tiene sentido asignar probabilidades a los resultados individuales posibles; todas las demás variables aleatorias se consideran continuas. Como consecuencia de esta distinción, analizamos las dos clases por separado: analizamos las variables aleatorias discretas y las variables aleatorias continuas en el Capítulo 6.

EJERCICIOS

Ejercicios básicos

5.1. Una tienda vende entre 0 y 12 computadores al día. ¿Es la venta diaria de computadores una variable aleatoria discreta o continua?

5.2. Un proceso de producción fabril produce un pequeño número de piezas defectuosas diariamente. ¿Es el número de piezas defectuosas una variable aleatoria discreta o continua?

5.3. Indique en cada uno de los casos siguientes cuál es la mejor definición: una variable aleatoria discreta o una variable aleatoria continua.

a) El número de automóviles que llegan diariamente a un taller de reparación en el que trabajan dos personas.
b) El número de automóviles producidos anualmente por General Motors.
c) Las ventas diarias totales de una tienda de comercio electrónico en dólares.
d) El número de pasajeros que se quedan sin plaza en una compañía aérea específica tres días antes de Navidad.

5.4. Un actor hace 100 representaciones al año. ¿Es su programa de trabajo (número de representaciones) una variable aleatoria discreta?

Ejercicios aplicados

5.5. Ponga cinco ejemplos de variables aleatorias discretas que podrían observarse en una nueva consultora.

5.6. Defina tres variables aleatorias continuas que debería examinar periódicamente un vicepresidente de marketing.

5.7. Una encuesta electoral entrevista a 2.000 personas seleccionadas aleatoriamente. ¿Debe analizarse el número de personas que apoyan al candidato A utilizando modelos de probabilidad discreta o continua?

5.8. Un vendedor entra diariamente en contacto con 20 personas y les pide que compren. ¿Debe analizarse el número de compras diarias utilizando modelos de probabilidad discreta o continua?

5.2. Distribuciones de probabilidad de variables aleatorias discretas

Supongamos que X es una variable aleatoria discreta y que x es uno de sus valores posibles. La probabilidad de que la variable aleatoria X tome el valor específico x se representa por medio de $P(X = x)$. La *función de probabilidad* de una variable aleatoria es una representación de las probabilidades de todos los resultados posibles. Esta representación podría ser algebraica, gráfica o tabular. En el caso de las variables aleatorias discretas, un sencillo método es enumerar las probabilidades de todos los resultados posibles de acuerdo con los valores de x.

Función de distribución de probabilidad

La **función de distribución de probabilidad**, $P(x)$, de una variable aleatoria discreta X expresa la probabilidad de que X tome el valor x, como una función de x. Es decir,

$$P(x) = P(X = x), \quad \text{para todos los valores de } x \tag{5.1}$$

En este libro utilizaremos la expresión *distribución de probabilidad* para representar las funciones de probabilidad siguiendo la práctica cada vez más habitual de utilizar estos términos indistintamente.

Como la función de probabilidad sólo toma valores distintos de 0 en puntos discretos x, a veces se denomina *función de masa de probabilidad*. Una vez que se han calculado las probabilidades, la función puede representarse gráficamente.

EJEMPLO 1.1. Lanzamiento de un dado al aire (gráfico de la función de probabilidad)

Represente gráficamente la función de probabilidad correspondiente al lanzamiento al aire de un dado equilibrado de seis caras.

Solución

Sea la variable aleatoria X el número resultante de un único lanzamiento al aire de un dado equilibrado de seis caras. Dado que

$$P(X = 1) = P(X = 2) = \cdots = P(X = 6) = \frac{1}{6}$$

la función de probabilidad es

$$P(x) = P(X = x) = \frac{1}{6} \quad \text{para } x = 1, 2, 3, ..., 6$$

La función toma el valor 0 en el caso de todos los demás valores de x, que no pueden ocurrir. La función de probabilidad se representa en la Figura 5.1, en la que las barras de altura $(X = 1) = P(X = 2) = \cdots = P(X = 6) = \frac{1}{6}$ representan masas de probabilidad en los puntos $x = 1$, $x = 2$, ..., $x = 6$.

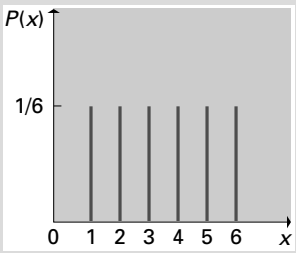


Figura 5.1. Gráfico de la función de probabilidad correspondiente al ejemplo 5.1.

La función de probabilidad de una variable aleatoria discreta debe satisfacer las dos propiedades siguientes.

Propiedades que deben satisfacer las funciones de probabilidad de variables aleatorias discretas

Sea X una variable aleatoria discreta que tiene una función de probabilidad $P(x)$. En ese caso,

1. $0 \leqslant P(x) \leqslant 1$ para cualquier valor x y
2. Las probabilidades individuales suman 1, es decir,

$$\sum_{x} P(x) = 1$$

donde la notación indica que el sumatorio abarca todos los valores posibles de x.

La propiedad 1 establece simplemente que las probabilidades no pueden ser negativas o mayores que 1. La propiedad 2 se deduce del hecho de que los sucesos «$X = x$», para todos los valores posibles de x, son mutuamente excluyentes y colectivamente exhaustivos. Las probabilidades de estos sucesos deben sumar, por lo tanto, 1. Este resultado puede verificarse directamente. Es una sencilla manera de afirmar que, cuando se realiza un experimento aleatorio, debe ocurrir algo.

También es útil otra representación de las distribuciones de probabilidad de variables aleatorias discretas.

Función de probabilidad acumulada

La **función de probabilidad acumulada**, $F(x_0)$, de una variable aleatoria X, expresa la probabilidad de que X no tenga un valor superior a x_0, como una función de x_0. Es decir,

$$F(x_0) = P(X \leqslant x_0) \qquad \textbf{(5.2)}$$

donde la función se evalúa en todos los valores de x_0.

EJEMPLO 5.2. Las ventas de automóviles (probabilidades)

Serrano Motor, S.A., es un concesionario de automóviles de una pequeña ciudad. Basándose en un análisis de su historial de ventas, sus directivos saben que en un día cualquiera el número de automóviles Vértigo A puede oscilar entre 0 y 5. ¿Cómo puede utilizarse la función de probabilidad mostrada en la Tabla 5.1 para planificar las existencias?

Tabla 5.1. Función de probabilidad de las ventas de automóviles.

x	$P(x)$	$F(x)$
0	0,15	0,15
1	0,30	0,45
2	0,20	0,65
3	0,20	0,85
4	0,10	0,95
5	0,05	1,00

Solución

La variable aleatoria, X, toma los valores de x indicados en la primera columna y la función de probabilidad, $P(x)$, se define en la segunda columna. La tercera columna contiene la distribución acumulada, $F(x)$. Este modelo podría utilizarse para planificar las existencias de automóviles. Por ejemplo, si sólo hay cuatro automóviles en existencias, Serrano Motor podría satisfacer las necesidades de los clientes de un automóvil el 95 por ciento de las veces. Pero si sólo hay dos automóviles en existencias, no se satisfarían las necesidades del 35 por ciento [$(1 - 0{,}65) \times 100$] de los clientes.

En el caso de las variables aleatorias discretas, la función de probabilidad acumulada a veces se denomina *función de masa acumulada*. Puede verse en la definición que, cuando x_0 aumenta, la función de probabilidad acumulada sólo cambia de valor en los puntos x_0 que puede tomar la variable aleatoria con una probabilidad positiva. Su evaluación en estos puntos se realiza por medio de la función de probabilidad.

Relación entre la función de probabilidad y la función de probabilidad acumulada

Sea X una variable aleatoria que tiene la función de probabilidad $P(x)$ y la función de probabilidad acumulada $F(x_0)$. Podemos demostrar que

$$F(x_0) = \sum_{x \leqslant x_0} P(x) \qquad \textbf{(5.3)}$$

donde la notación implica que el sumatorio abarca todos los valores posibles de x que son menores o iguales que x_0.

El resultado de la ecuación 5.3 es fácil de deducir, ya que el suceso «$X \leqslant x_0$» es la unión de los sucesos mutuamente excluyentes «$X = x$», para todos los valores posibles de x menores o iguales que x_0. La probabilidad de la unión es, pues, la suma de las probabilidades de esos sucesos individuales.

Propiedades de las funciones de probabilidad acumulada de variables aleatorias discretas

Sea X una variable aleatoria discreta que tiene una función de probabilidad acumulada $F(x_0)$. Podemos demostrar que

1. $0 \leqslant F(x_0) \leqslant 1$ para todo número x_0; y
2. Si x_0 y x_1 son dos números tales que $x_0 < x_1$, entonces $F(x_0) \leqslant F(x_1)$.

La propiedad 1 establece simplemente que una probabilidad no puede ser menor que 0 o mayor que 1. Obsérvense, por ejemplo, las probabilidades de la Figura 5.1 correspondientes al lanzamiento de un dado al aire. La propiedad 2 implica que la probabilidad de que una variable aleatoria no sea mayor que un determinado número no puede ser mayor que la probabilidad de que no sea mayor que cualquier número más alto.

EJERCICIOS

Ejercicios básicos

5.9. ¿Cuál es la función de probabilidad del número de caras cuando se lanza al aire una moneda equilibrada?

5.10. Muestre la función de probabilidad del número de caras en el lanzamiento al aire de una moneda equilibrada.

5.11. Muestre la función de probabilidad del número de caras cuando se lanzan al aire independientemente tres monedas equilibradas.

5.12. Suponga que la variable aleatoria representa el número de veces que faltará a clase este cuatrimestre. Elabore una tabla que muestre la función de probabilidad y la función de probabilidad acumulada.

Ejercicios aplicados

5.13. El número de computadores vendidos al día en una tienda viene definido por la siguiente distribución de probabilidad:

X	0	1	2	3	4	5	6
$P(x)$	0,05	0,10	0,20	0,20	0,20	0,15	0,10

a) ¿$P(3 \leqslant x < 6) = ?$

b) ¿$P(x > 3) = ?$

c) ¿$P(x \leqslant 4) = ?$

d) ¿$P(2 < x \leqslant 5) = ?$

5.14. Una compañía aérea le ha pedido que estudie los retrasos de los vuelos que se registraron en un aeropuerto la semana antes de las Navidades. La variable aleatoria X es el número de vuelos retrasados por hora.

X	0	1	2	3	4	5	6	7	8	9
$P(x)$	0,10	0,08	0,07	0,15	0,12	0,08	0,10	0,12	0,08	0,10

a) ¿Cuál es la distribución de probabilidad acumulada?

b) ¿Cuál es la probabilidad de que haya cinco o más vuelos retrasados?

c) ¿Cuál es la probabilidad de que haya entre tres y siete (inclusive) vuelos retrasados?

5.3. Propiedades de las variables aleatorias discretas

La distribución de probabilidad contiene toda la información sobre las propiedades probabilísticas de una variable aleatoria y el examen gráfico de esta distribución puede ser, desde luego, valioso. Sin embargo, a menudo es deseable disponer de alguna medida sintética de las características de la distribución.

Valor esperado de una variable aleatoria discreta

Para tener una medida del punto central de una distribución de probabilidad, introducimos el concepto de *esperanza* de una variable aleatoria. En el Capítulo 3 calculamos la media muestral como una medida del punto central de datos muestrales. El *valor esperado* es la medida correspondiente del punto central de una variable aleatoria. Antes de definirlo, mostramos el error de una medida alternativa que parece atractiva a primera vista.

Consideremos el ejemplo siguiente: en una revisión de los libros de texto de un segmento del campo de administración de empresas se observó que el 81 por ciento de todas las páginas no tenía ninguna errata, que el 17 por ciento contenía una errata y que el 2 por ciento restante contenía dos erratas. Utilizamos la variable aleatoria X para representar el número de erratas que hay en una página elegida aleatoriamente en uno de estos libros; sus valores posibles son 0, 1 y 2 y la función de probabilidad es

$$P(0) = 0{,}81 \qquad P(1) = 0{,}17 \qquad P(2) = 0{,}02$$

Podríamos considerar la posibilidad de utilizar la media simple de los valores como medida del punto central de una variable aleatoria. En este ejemplo, el número de erratas que puede haber en una página es 0, 1 y 2. Su media es, pues, una errata. Sin embargo, basta una breve reflexión para convencer al lector de que esta medida del punto central es absurda. Al calcular esta media, no hemos prestado atención al hecho de que el 81 por ciento de todas las páginas no contiene ninguna errata, mientras que sólo el 2 por ciento contiene dos erratas. Para obtener una medida sensata del punto central, *ponderamos* los distintos resultados posibles por las probabilidades de que ocurran.

Valor esperado

El **valor esperado**, $E(X)$, de una variable aleatoria discreta X se define de la forma siguiente:

$$E(X) = \mu = \sum_x xP(x) \qquad \textbf{(5.4)}$$

donde la notación indica que el sumatorio abarca todos los valores posibles de x.

El valor esperado de una variable aleatoria también se llama **media** y se representa por medio del símbolo μ.

El valor esperado puede expresarse por medio de frecuencias relativas a largo plazo. Supongamos que un experimento aleatorio se repite N veces y que el suceso «$X = x$» ocurre en N_x de estas pruebas. La media de los valores que toma la variable aleatoria en las N pruebas es la suma de los xN_x/N correspondientes a todos los valores posibles de x. Ahora bien, como el número de repeticiones, N, tiende a infinito, el cociente N_x/N tiende a la probabilidad de que ocurra el suceso «$X = x$», es decir, a $P(x)$. De ahí que la cantidad xN_x/N tienda a $xP(x)$. Por lo tanto, podemos concebir el valor esperado como el valor medio a largo plazo que toma una variable aleatoria cuando se realiza un gran número de pruebas. Recuérdese que en el Capítulo 3 utilizamos la palabra *media* para referirnos al promedio de un conjunto de observaciones numéricas. Utilizamos el mismo término para referirnos a la esperanza de una variable aleatoria.

EJEMPLO 5.3. Erratas de los libros de texto (valor esperado)

Supongamos que la función de probabilidad del número de erratas, X, que hay en las páginas de los libros de texto de administración de empresas es

$$P(0) = 0{,}81 \qquad P(1) = 0{,}17 \qquad P(2) = 0{,}02$$

Halle el número medio de erratas por página.

Solución

Tenemos que

$$\mu = E(X) = \sum_x xP(x) = (0)(0{,}81) + (1)(0{,}17) + (2)(0{,}02) = 0{,}21$$

De este resultado se deduce que, si se analiza un gran número de páginas, es de esperar que haya una media de 0,21 erratas por página. La Figura 5.2 muestra la función de probabilidad e indica dónde se encuentra la media.

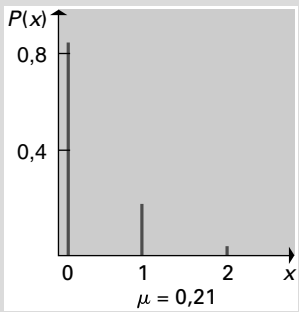


Figura 5.2. Función de probabilidad del número de erratas por página de los libros de texto de administración de empresas; localización de la media poblacional, μ, del ejemplo 5.3.

Varianza de una variable aleatoria discreta

En el Capítulo 3 observamos que la varianza muestral era una medida útil de la dispersión de un conjunto de observaciones numéricas. La varianza muestral es el promedio de los cuadrados de las diferencias entre las observaciones y la media. Nos basamos en esta misma idea para medir la dispersión de la distribución de probabilidad de una variable aleatoria. La *varianza* de una variable aleatoria es el promedio ponderado de los cuadrados de sus diferencias posibles con respecto a la media, $(x - \mu)$; la ponderación correspondiente a $(x - \mu)^2$ es la probabilidad de que la variable aleatoria tome el valor x. Puede considerarse, pues, que la varianza, definida en la ecuación 5.5, es el valor medio que tomará la función $(X - \mu)^2$ en un número muy grande de pruebas repetidas.

Varianza y desviación típica de una variable aleatoria discreta

Sea X una variable aleatoria discreta. La esperanza de los cuadrados de las diferencias con respecto a la media, $(X - \mu)^2$, se llama **varianza**, se representa por medio del símbolo σ^2 y viene dada por

$$\sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 P(x) \qquad \textbf{(5.5)}$$

La varianza de una variable aleatoria discreta X también puede expresarse de la forma siguiente:

$$\sigma^2 = E(X^2) - \mu^2 =$$
$$= \sum_x x^2 P(x) - \mu^2 \mu_x^2 \qquad \textbf{(5.6)}$$

La **desviación típica**, σ_X, es la raíz cuadrada positiva de la varianza.

El concepto de varianza puede ser muy útil para comparar las dispersiones de distribuciones de probabilidad. Consideremos, por ejemplo, que el rendimiento de una inversión en un año es una variable aleatoria. Aunque dos inversiones tengan los mismos rendimientos esperados, son muy diferentes si las varianzas de estos rendimientos son muy diferentes. Si la varianza es mayor, es más probable que los rendimientos sean considerablemente diferentes de la media que si la varianza es pequeña. En este contexto, pues, la varianza del rendimiento puede guardar relación con el concepto de riesgo de una inversión: cuanto mayor es la varianza, mayor es el riesgo.

Como señalamos en el Capítulo 3, tomando la raíz cuadrada de la varianza para hallar la desviación típica se obtiene una cantidad en las unidades originales de medición.

En algunas aplicaciones prácticas, es preferible una fórmula alternativa, pero equivalente, de la varianza para efectuar los cálculos. Esa fórmula alternativa se define en la ecuación 5.6, que puede verificarse algebraicamente (véase el apéndice del capítulo).

EJEMPLO 5.4. Valor esperado y varianza de las ventas de automóviles (valor esperado y varianza)

En el ejemplo 5.2, Serrano Motor, S.A., averiguó que el número de automóviles Vértigo A vendidos diariamente podía oscilar entre 0 y 5 y que las probabilidades se indicaban en la Tabla 5.1. Halle el valor esperado y la varianza de esta distribución de probabilidad.

Solución

Aplicando la ecuación 5.4, el valor esperado es

$$\mu = E(X) \sum_x xP(x) = 0(0{,}15) + 1(0{,}30) + \cdots + 5(0{,}05) = 1{,}95$$

Aplicando la ecuación 5.5, la varianza es

$$\sigma^2 = (0 - 1{,}95)^2(0{,}15) + (1 - 1{,}95)^2(0{,}3) + \cdots + (5 - 1{,}95)^2(0{,}05) = 1{,}9475$$

Cuando las distribuciones de probabilidad son más complejas, puede utilizarse el programa Excel para realizar estos cálculos. Las Figuras 5.3 y 5.4 muestran cómo se obtienen el valor esperado y la varianza de la distribución de la Tabla 5.1.

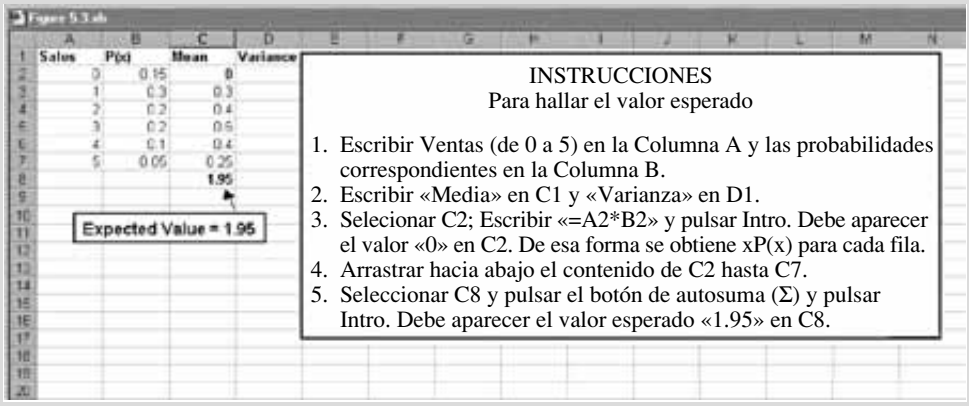
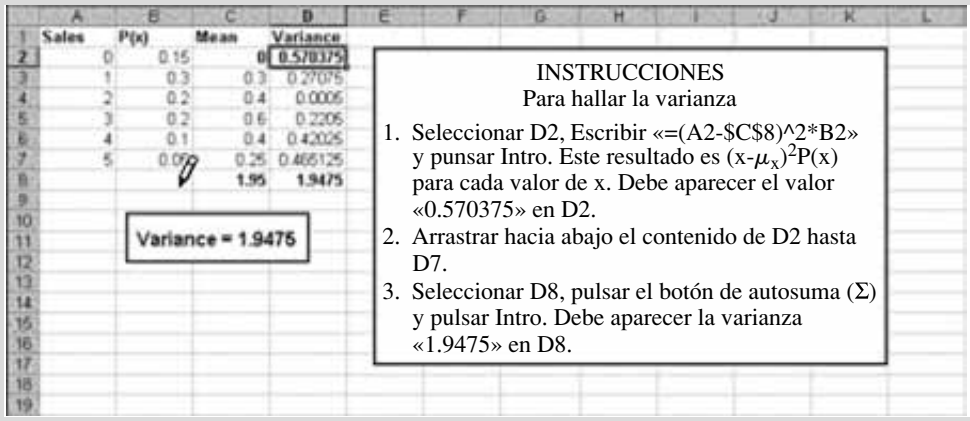


Figura 5.3. Valor esperado de la variable aleatoria discreta de la Tabla 5.1 calculado utilizando el programa Excel de Microsoft.



	A	B	C	D
1	Sales	P(x)	Mean	Variance
2	0	0.15	0	0.570375
3	1	0.3	0.3	0.27075
4	2	0.2	0.4	0.0005
5	3	0.2	0.6	0.2205
6	4	0.1	0.4	0.42025
7	5	0.05	0.25	0.465125
8			1.95	1.9475

Variance = 1.9475

INSTRUCCIONES
Para hallar la varianza

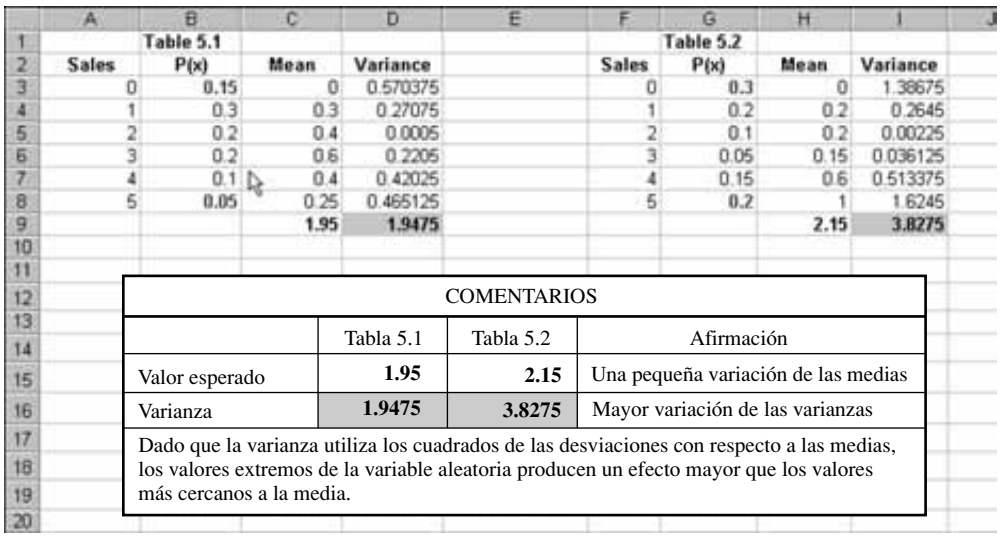
1. Seleccionar D2, Escribir «=(A2-$C$8)^2*B2» y punsar Intro. Este resultado es $(x-\mu_x)^2P(x)$ para cada valor de x. Debe aparecer el valor «0.570375» en D2.
2. Arrastrar hacia abajo el contenido de D2 hasta D7.
3. Seleccionar D8, pulsar el botón de autosuma (Σ) y pulsar Intro. Debe aparecer la varianza «1.9475» en D8.

Figura 5.4. Varianza de la variable aleatoria discreta de la Tabla 5.1 calculada utilizando el programa Excel de Microsoft.

Supongamos que modificamos la función de probabilidad de la Tabla 5.1 para que sea mayor la probabilidad tanto de que las ventas sean bajas como de que sean altas. La Tabla 5.2 muestra las nuevas probabilidades y la Figura 5.5 indica la variación de la media y de la varianza.

Tabla 5.2. Reconsideración de la función de probabilidad de las ventas de automóviles.

Ventas	$P(X)$
0	**0,30**
1	0,20
2	0,10
3	0,05
4	0,15
5	**0,20**



	A	B	C	D	E	F	G	H	I	J
1		Table 5.1					Table 5.2			
2	Sales	P(x)	Mean	Variance		Sales	P(x)	Mean	Variance	
3	0	0.15	0	0.570375		0	0.3	0	1.38675	
4	1	0.3	0.3	0.27075		1	0.2	0.2	0.2645	
5	2	0.2	0.4	0.0005		2	0.1	0.2	0.00225	
6	3	0.2	0.6	0.2205		3	0.05	0.15	0.036125	
7	4	0.1	0.4	0.42025		4	0.15	0.6	0.513375	
8	5	0.05	0.25	0.465125		5	0.2	1	1.6245	
9			1.95	1.9475				2.15	3.8275	

COMENTARIOS			
	Tabla 5.1	Tabla 5.2	Afirmación
Valor esperado	**1.95**	**2.15**	Una pequeña variación de las medias
Varianza	**1.9475**	**3.8275**	Mayor variación de las varianzas
Dado que la varianza utiliza los cuadrados de las desviaciones con respecto a las medias, los valores extremos de la variable aleatoria producen un efecto mayor que los valores más cercanos a la media.			

Figura 5.5. Comparación de las medias y las varianzas de la variable aleatoria discreta de la Tabla 5.2 calculadas utilizando el programa Excel de Microsoft.

Comentarios

- En la Tabla 5.2, la probabilidad de que las ventas sean 0 es mayor (0,30 en lugar de 0,15 de la Tabla 5.1). La probabilidad de que se vendan 5 automóviles también es mayor (0,20 en lugar de 0,05 de la Tabla 5.1).
- La varianza debería aumentar ya que la probabilidad de los valores extremos 0 y 5 aumenta.

Media y varianza de funciones lineales de una variable aleatoria

El concepto de esperanza no se limita a la propia variable aleatoria sino que puede aplicarse a cualquier función de la variable aleatoria. Por ejemplo, un contratista puede no saber cuánto tiempo tardará en realizar el trabajo estipulado en un contrato. Esta incertidumbre puede representarse por medio de una variable aleatoria cuyos valores posibles son el número de días que transcurren desde el inicio del trabajo estipulado en el contrato hasta su terminación. Sin embargo, lo que preocupa principalmente al contratista no es el tiempo que tardará sino, más bien, el coste de cumplir el contrato. Este coste es una función del tiempo que tardará, por lo que para hallar el valor esperado de la variable aleatoria «coste» es necesario hallar la esperanza de una función de la variable aleatoria «tiempo que se tardará».

Valor esperado de las funciones de variables aleatorias

Sea X una variable aleatoria cuya función de probabilidad es $P(x)$ y sea $g(X)$ una función de X. El valor esperado, $E[g(X)]$, de esa función se define de la forma siguiente:

$$E[g(X)] = \sum_x g(x)P(x) \qquad \textbf{(5.7)}$$

La ecuación 5.7 define la esperanza de una función de una variable aleatoria X. Es decir, la esperanza puede concebirse como el valor promedio que tomaría $g(X)$ en un número muy grande de repeticiones de un experimento. A continuación, desarrollamos el valor esperado y la varianza de funciones lineales de una variable aleatoria. Consideremos, en primer lugar, la función lineal $a + bX$, donde a y b son números fijos constantes. Sea X una variable aleatoria que toma el valor x con una probabilidad $P(x)$ y consideremos una nueva variable aleatoria Y, definida por

$$Y = a + bX$$

Cuando la variable aleatoria X toma el valor específico x, Y debe tomar el valor $a + bx$. A menudo se necesita la media y la varianza de esas variables. En el apéndice de este capítulo se desarrolla la media, la varianza y la desviación típica de una función lineal de una variable aleatoria. Los resultados se resumen en las ecuaciones 5.8 y 5.9.

Resumen de las propiedades de las funciones lineales de una variable aleatoria

Sea X una variable aleatoria de media μ_x y varianza σ_2^x y sean a y b unos números fijos constantes cualesquiera. Definamos la variable aleatoria Y como $a + bX$. Entonces, la **media y la varianza de *Y*** son

$$\mu_Y = E(a + bX) = a + b\mu_X \qquad \textbf{(5.8)}$$

y

$$\sigma_Y^2 = \text{Var}(a + bX) = b^2\sigma_X^2 \tag{5.9}$$

por lo que la desviación típica de Y es

$$\sigma_y = |b|\sigma_x \tag{5.10}$$

EJEMPLO 5.5. Coste total de un proyecto (cálculos de las funciones de variables aleatorias)

Un contratista está interesado en saber cuál es el coste total de un proyecto para el que pretende presentar una oferta. Estima que los materiales costarán 25.000 $ y su trabajo 900 $ al día. Si el proyecto tarda en realizarse X días, el coste laboral total será de $900X$ $ y el coste total del proyecto (en dólares) será

$$C = 25.000 + 900X$$

El contratista estima unas probabilidades subjetivas (Tabla 5.3) de la duración probable del proyecto.

a) Halle la media y la varianza de la duración X.
b) Halle la media, la varianza y la desviación típica del coste total C.

Tabla 5.3. Distribución de probabilidad de la duración.

Duración X (días)	10	11	12	13	14
Probabilidad	0,1	0,3	0,3	0,2	0,1

Solución

a) La media y la varianza de la duración X pueden hallarse mediante las ecuaciones 5.4 y 5.5.

$$\mu = E(X) = \sum_x x\,P(x) =$$

$$= (10)(0,1) + (11)(0,3) + (12)(0,3) + (13)(0,2) + (14)(0,1) = 11,9 \text{ días}$$

y

$$\sigma_x^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 P(x) =$$

$$= (10 - 11,9)^2(0,1) + (11 - 11,9)^2(0,3) + \cdots + (14 - 11,9)^2(0,1) = 1,29 \text{ días}$$

b) La media, la varianza y la desviación típica del coste total, C, se hallan mediante las ecuaciones 5.8, 5.9 y 5.10.

La media es

$$\mu_C = E(25.000 + 900X) = (25.000 + 900\mu_X)$$
$$= 25.000 + (900)(11,9) = 35.710 \text{ \$}$$

La varianza es

$$\sigma_C^2 = \text{Var}(25.000 + 900X) = (900)^2\sigma_X^2$$
$$= (810.000)(1,29) = 1.044,900$$

La desviación típica es

$$\sigma_C = \sqrt{\sigma_C^2} = 1.022,20 \text{ \$}$$

Hay tres ejemplos especiales de la función lineal $W = a + bX$ que son importantes. El primero considera una función constante, $W = a$, para cualquier constante a. En esta situación, el coeficiente $b = 0$. En el segundo ejemplo, $a = 0$, de donde $W = bX$. Las ecuaciones 5.11 y 5.12 definen el valor esperado y la varianza de estas funciones. El tercer ejemplo es importante en capítulos posteriores. Las ecuaciones 5.13 y 5.14 definen la media y la varianza de esta función lineal especial. Por lo tanto, restando de una variable aleatoria su media y dividiendo por su desviación típica se obtiene una variable aleatoria de media 0 y desviación típica 1.

Resultados sintéticos de la media y la varianza de funciones lineales especiales

a) Sea $b = 0$ en la función lineal $W = a + bX$. Entonces, $W = a$ (para cualquier constante a).

$$E(a) = a \qquad \text{y} \qquad \text{Var}(a) = 0 \tag{5.11}$$

Si una variable aleatoria siempre toma el valor a, tendrá una media a y una varianza 0.

b) Sea $a = 0$ en la función lineal $W = a + bX$. Entonces, $W = bX$.

$$E(bX) = b\mu_X \qquad \text{y} \qquad \text{Var}(bX) = b^2\sigma_X^2 \tag{5.12}$$

La media y la varianza de $Z = \dfrac{X - \mu_X}{\sigma_X}$

Sea $a = -\mu_X/\sigma_X$ y $b = 1/\sigma_X$ en la función lineal $Z = a + bX$. Entonces,

$$Z = a + bX = \frac{X - \mu_X}{\sigma_X}$$

de manera que

$$E\left(\frac{X - \mu_X}{\sigma_X}\right) = -\frac{\mu_X}{\sigma_X} + \frac{1}{\sigma_X}\mu_X = 0 \tag{5.13}$$

y

$$\text{Var}\left(\frac{X - \mu_X}{\sigma_X}\right) = \frac{1}{\sigma_X^2}\sigma_X^2 = 1 \tag{5.14}$$

EJERCICIOS

Ejercicios básicos

5.15. Considere la función de probabilidad

x	0	1
Probabilidad	0,40	0,60

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) Halle la media de la variable aleatoria X.
d) Halle la varianza de X.

5.16. Dada la función de probabilidad

x	0	1	2
Probabilidad	0,25	0,50	0,25

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) Halle la media de la variable aleatoria X.
d) Halle la varianza de X.

5.17. Considere la función de probabilidad

x	0	1
Probabilidad	0,50	0,50

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) Halle la media de la variable aleatoria X.
d) Halle la varianza de X.

5.18. Un concesionario de automóviles calcula la proporción de automóviles nuevos vendidos que se han devuelto varias veces para que se corrijan los defectos durante el periodo de garantía. La tabla adjunta muestra los resultados.

Número de devoluciones	0	1	2	3	4
Proporción	0,28	0,36	0,23	0,09	0,04

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) Halle la media del número de devoluciones de un automóvil para que se corrijan los defectos durante el periodo de garantía.
d) Halle la varianza del número de devoluciones de un automóvil para que se corrijan los defectos durante el periodo de garantía.

5.19. Una empresa está especializada en la instalación y el mantenimiento de calefacciones centrales. Antes de que empiece el invierno, las llamadas al servicio de mantenimiento pueden dar como resultado el pedido de una nueva caldera. La tabla adjunta muestra las probabilidades estimadas del número de pedidos de calderas nuevas generados de esta forma en las 2 últimas semanas de septiembre.

Número de pedidos	0	1	2	3	4	5
Probabilidad	0,10	0,14	0,26	0,28	0,15	0,07

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) Halle la probabilidad de que se hagan al menos tres pedidos en este periodo.
d) Halle la media del número de pedidos de una nueva caldera en este periodo de 2 semanas.
e) Halle la desviación típica del número de pedidos de una nueva caldera en este periodo de 2 semanas.

Ejercicios aplicados

5.20. Una empresa produce paquetes de clips. El número de clips por paquete varía, como indica la tabla adjunta.

Número de clips	47	48	49	50	51	52	53
Proporción de paquetes	0,04	0,13	0,21	0,29	0,20	0,10	0,03

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) ¿Cuál es la probabilidad de que un paquete seleccionado aleatoriamente contenga entre 49 y 51 clips (inclusive)?
d) Se seleccionan dos paquetes aleatoriamente. ¿Cuál es la probabilidad de que al menos uno de ellos contenga como mínimo 50 clips?
e) Utilice el programa Excel de Microsoft para hallar la media y la desviación típica del número de clips por paquete.
f) El coste (en centavos) de producir un paquete de clips es $16 + 2X$, donde X es el número de clips que hay en el paquete. Los ingresos generados por la venta del paquete, cualquiera que sea el número de clips que contenga, son de 1,50 $. Si los beneficios son la diferencia entre los ingresos y el coste, halle la media y la desviación típica de los beneficios por paquete.

5.21. Una empresa municipal de autobuses ha comenzado a dar servicio en un nuevo barrio. Se ha registrado el número de usuarios que hay en este barrio en el servicio de primera hora de la mañana. La tabla adjunta muestra la proporción de cada uno de los días de la semana.

Número de usuarios	0	1	2	3	4	5	6	7
Proporción	0,02	0,12	0,23	0,31	0,19	0,08	0,03	0,02

a) Trace la función de probabilidad.
b) Calcule y trace la función de probabilidad acumulada.
c) ¿Cuál es la probabilidad de que en un día seleccionado aleatoriamente haya al menos cuatro usuarios del barrio en este servicio?

d) Se seleccionan dos días aleatoriamente. ¿Cuál es la probabilidad de que en estos dos días haya menos de tres usuarios del barrio en este servicio?

e) Halle la media y la desviación típica del número de usuarios de este barrio en este servicio en un día de la semana.

f) Suponiendo que el coste de un viaje es de 50 centavos, halle la media y la desviación típica del total de pagos de los usuarios de este barrio en este servicio un día de la semana.

5.22. **a)** Un gran envío de piezas contiene un 10 por ciento de piezas defectuosas. Se seleccionan aleatoriamente dos y se prueban. Sea la variable aleatoria X el número de defectos encontrados. Halle la función de probabilidad de esta variable aleatoria.

b) Un envío de 20 piezas contiene dos defectuosas. Se seleccionan aleatoriamente dos y se prueban. Sea la variable aleatoria Y el número de defectos encontrados. Halle la función de probabilidad de esta variable aleatoria. Explique por qué su respuesta es diferente de la respuesta del apartado (a).

c) Halle la media y la varianza de la variable aleatoria X del apartado (a).

d) Halle la media y la varianza de la variable aleatoria Y del apartado (b).

5.23. Un estudiante necesita saber qué tareas ha puesto el profesor para el próximo día y decide llamar a algunos compañeros para obtener esa información. Cree que la probabilidad de obtener la información necesaria en una llamada cualquiera es 0,40. Decide continuar llamando a los compañeros hasta obtener la información. Sea la variable aleatoria X el número de llamadas necesarias para obtener la información.

a) Halle la función de probabilidad de X.

b) Halle la función de probabilidad acumulada de X.

c) Halle la probabilidad de que sean necesarias tres llamadas como mínimo.

5.24. Un jugador universitario de baloncesto que tiene un porcentaje de aciertos del 75 por ciento en sus tiros libres se sitúa en la línea de lanzamiento de «uno más uno» (si encesta a la primera, puede tirar otra vez, pero no en caso contrario; se anota un punto por cada enceste). Suponga que el resultado del segundo lanzamiento, si lo hay, es independiente del resultado del primero. Halle el número esperado de puntos resultantes del «uno más uno». Compárelo con el número esperado de puntos de una «falta de dos tiros libres», en la que se permite lanzar una segunda vez, cualquiera que sea el resultado del primer lanzamiento.

5.25. Un profesor tiene un numeroso grupo de alumnos y ha previsto un examen a las 7 de la tarde en un aula diferente. Estime en la tabla las probabilidades del número de estudiantes que lo llamarán a casa una hora antes del examen preguntándole en qué aula se realizará.

Número de llamadas	0	1	2	3	4	5
Probabilidad	0,10	0,15	0,19	0,26	0,19	0,11

Halle la media y la desviación típica del número de llamadas.

5.26. Se ha pedido a los estudiantes de una numerosa clase de contabilidad que valoren el curso en una escala de 1 a 5. Una puntuación mayor indica que los estudiantes dan un valor mayor al curso. La tabla adjunta muestra las proporciones de estudiantes que puntúan el curso en cada categoría.

Puntuación	1	2	3	4	5
Proporción	0,07	0,19	0,28	0,30	0,16

Halle la media y la desviación típica de las puntuaciones.

5.27. Un quiosquero tiene un periódico que a veces le pide un pequeño número de clientes. Cada ejemplar le cuesta 70 centavos y lo vende a 90 centavos. Los ejemplares que le quedan al final del día no tienen ningún valor y se destruyen. El quiosquero considera que por cada ejemplar que le piden y no puede vender porque se han agotado tiene una pérdida de clientela que valora en 5 centavos. La tabla adjunta muestra la distribución de probabilidad del número de demandas del periódico en un día. Si el beneficio diario total del quiosquero son los ingresos totales generados por las ventas del periódico menos los costes totales de los periódicos pedidos, menos la pérdida de clientela como consecuencia de las demandas insatisfechas, ¿cuántos ejemplares diarios debe pedir para maximizar los beneficios esperados?

Número de demandas	0	1	2	3	4	5
Probabilidad	0,12	0,16	0,18	0,32	0,14	0,08

5.28. El director de una fábrica está considerando la posibilidad de sustituir una máquina caprichosa. El historial de la máquina indica la siguiente distribución de probabilidad del número de averías registradas en una semana.

Número de averías	0	1	2	3	4
Probabilidad	0,10	0,26	0,42	0,16	0,06

a) Halle la media y la desviación típica del número de averías semanales.
b) Se estima que cada avería le cuesta a la empresa 1.500 $ de producción perdida. Halle la media y la desviación típica del coste semanal de las averías de esta máquina.

5.29. Un inversor está considerando tres estrategias para invertir 1.000 $. Se estima que los rendimientos probables son los siguientes:

- *Estrategia* 1: unos beneficios de 10.000 $ con una probabilidad de 0,15 y una pérdida de 1.000 $ con una probabilidad de 0,85.
- *Estrategia* 2: unos beneficios de 1.000 $ con una probabilidad de 0,50, unos beneficios de 500 $ con una probabilidad de 0,30 y una pérdida de 500 $ con una probabilidad de 0,20.
- *Estrategia* 3: unos beneficios seguros de 400 $.

¿Qué estrategia tiene el mayor beneficio esperado? ¿Aconsejaría necesariamente al inversor que adoptara esta estrategia?

5.4. Distribución binomial

A continuación, desarrollamos la distribución de probabilidad binomial que se utiliza mucho en numerosos problemas aplicados empresariales y económicos. Comenzamos desarrollando primero el modelo de Bernoulli, que es una pieza esencial de la distribución binomial. Consideramos un experimento aleatorio que puede dar lugar a dos resultados posibles mutuamente excluyentes y colectivamente exhaustivos, que por comodidad llamamos «éxito» y «fracaso». Sea P la probabilidad de éxito, por lo que la probabilidad de fracaso es $(1 - P)$. Definamos ahora la variable aleatoria X de manera que tome el valor 1 si el resultado del experimento es un éxito y 0 en caso contrario. La función de probabilidad de esta variable aleatoria es, entonces,

$$P(0) = (1 - P) \quad \text{y} \quad P(1) = P$$

Esta distribución se conoce con el nombre de *distribución de Bernoulli*. Su media y su varianza pueden hallarse aplicando directamente las ecuaciones del apartado 5.3.

Obtención de la media y la varianza de una variable aleatoria de Bernoulli

La **media** es

$$\mu = E(X) = \sum_x xP(x) = (0)(1 - P) + (1)P = P \tag{5.15}$$

y la **varianza** es

$$\begin{aligned} \sigma^2 &= E[(X - \mu)^2] = \sum_x (x - \mu)^2 P(x) \\ &= (0 - P)^2(1 - P) + (1 - P)^2 P = P(1 - P) \end{aligned} \tag{5.16}$$

EJEMPLO 5.6. Venta de un contrato (calcular la media y la varianza de Bernoulli)

Susana Ferrater, agente de seguros, cree que la probabilidad de vender un seguro en un contacto específico es 0,4. Si la variable aleatoria X toma el valor 1 si se vende un seguro y 0 en caso contrario, entonces X tiene una distribución de Bernoulli con una probabilidad de éxito P igual a 0,4. Halle la media y la varianza de la distribución.

Solución

La función de probabilidad de X es $P(0) = 0{,}6$ y $P(1) = 0{,}4$. La media de la distribución es $P = 0{,}40$ y la varianza es $\sigma^2 = P(1 - P) = (0{,}4)(0{,}6) = 0{,}24$.

Una importante generalización de la distribución de Bernoulli es el caso en el que se realiza varias veces un experimento aleatorio con dos resultados posibles y las repeticiones son independientes. En este caso, podemos hallar las probabilidades utilizando la distribución binomial. Supongamos de nuevo que la probabilidad de éxito en una única prueba es P y que se realizan n pruebas independientes, por lo que el resultado de cualquiera de ellas no influye en el resultado de las demás. El número de éxitos X resultantes de estas n pruebas podría ser cualquier número entero comprendido entre 0 y n y nos interesa saber cuál es la probabilidad de obtener exactamente $X = x$ éxitos en n pruebas.

Desarrollamos el resultado en dos fases. En primer lugar, observamos que el resultado de las n pruebas es una secuencia de n resultados, cada uno de los cuales debe ser un éxito (S) o un fracaso (F). Una secuencia con x éxitos y $(n - x)$ fracasos es

$$\underbrace{\text{S, S, ..., S}}_{(x \text{ veces})} \qquad \underbrace{\text{F, F, ..., F}}_{(n - x \text{ veces})}$$

En palabras, el resultado de las x primeras pruebas es un éxito, mientras que el del resto es un fracaso. Ahora bien, la probabilidad de éxito en una única prueba es P y la probabilidad de fracaso es $(1 - P)$. Dado que las n pruebas son independientes entre sí, la probabilidad de cualquier secuencia de resultados es, por la regla del producto de probabilidades (Capítulo 4), igual al producto de las probabilidades de los resultados individuales. Por lo tanto, la probabilidad de observar la secuencia específica de resultados que acabamos de describir es

$$\underbrace{[P \times P \times \cdots \times P]}_{(x \text{ veces})} \times \underbrace{[(1 - P) \times (1 - P) \times \cdots \times (1 - P)]}_{(n - x \text{ veces})} = P^x(1 - P)^{(n-x)}$$

Según este argumento, la probabilidad de observar *cualquier secuencia específica* que contenga x éxitos y $(n - x)$ fracasos es $P^x(1 - P)^{n-x}$. Supongamos, por ejemplo, que hay cinco pruebas independientes, cada una con una probabilidad de éxito $P = 0{,}60$, y hay que hallar la probabilidad de conseguir tres éxitos exactamente. Utilizando el signo + para representar un éxito y 0 para representar un fracaso, los resultados deseados pueden representarse de la forma siguiente:

$$+\,+\,+00 \qquad \text{o} \qquad +0+0+$$

La probabilidad de cualquiera de estos dos resultados específicos es $(0{,}6)^3(0{,}4)^2 = 0{,}03456$.

El problema original no era hallar la probabilidad de ocurrencia de una determinada secuencia sino la probabilidad de conseguir x éxitos exactamente, independientemente del orden de los resultados. Hay varias secuencias en las que podría haber x éxitos entre $(n - x)$ fracasos. De hecho, el número de esas posibilidades es precisamente el número de

combinaciones de x objetos elegidos de n, ya que se pueden seleccionar x posiciones de un total de n en las que colocar los éxitos y el número total de éxitos puede calcularse utilizando la ecuación 5.17. Volviendo al ejemplo de tres éxitos en cinco pruebas ($P = 0{,}60$), el número de diferentes secuencias con tres éxitos sería

$$C_3^5 = \frac{5!}{3!(3-5)!} = 10$$

La probabilidad de conseguir tres éxitos en cinco pruebas independientes de Bernoulli es, pues, 10 multiplicado por la probabilidad de cada una de las secuencias que tiene tres éxitos y, por lo tanto,

$$P(X = 3) = (10)(0{,}03456) = 0{,}3456$$

A continuación, generalizamos este resultado para cualquier combinación de n y x.

Número de secuencias con *x* éxitos en *n* pruebas

El **número de secuencias con *x* éxitos en *n* pruebas independientes** es

$$C_x^n = \frac{n!}{x!(n-x)!} \tag{5.17}$$

donde $n! = n \times (n-1) \times (n-2) \times \cdots \times 1$ y $0! = 1$.

Estas C_x^n secuencias son mutuamente excluyentes, ya que no pueden ocurrir dos al mismo tiempo. Este resultado se desarrolló en el apéndice del Capítulo 4.

El suceso «se obtienen x éxitos en n pruebas» puede ocurrir de C_x^n maneras mutuamente excluyentes, cada una con una probabilidad $P^x(1-P)^{n-x}$. Por lo tanto, por la regla de la suma de probabilidades (Capítulo 4), la probabilidad que buscamos es la suma de estas C_x^n probabilidades individuales. El resultado se obtiene mediante la ecuación 5.18.

La distribución binomial

Supongamos que un experimento aleatorio puede tener dos resultados posibles mutuamente excluyentes y colectivamente exhaustivos, «éxito» y «fracaso», y que P es la probabilidad de éxito en una única prueba. Si se realizan n pruebas independientes, la distribución del número de éxitos resultantes, x, se llama **distribución binomial**. Su función de probabilidad de la variable aleatoria binomial $X = x$ es

$$P(x \text{ éxitos en } n \text{ pruebas independientes}) = P(x) =$$

$$= \frac{n!}{x!(n-x)!} P^x(1-P)^{(n-x)} \quad \text{para } x = 0, 1, 2, \ldots, n \tag{5.18}$$

La media y la varianza se hallan en el apéndice del capítulo y los resultados se obtienen por medio de las ecuaciones 5.19 y 5.20.

Media y varianza de una distribución binomial

Sea X el número de éxitos en n repeticiones independientes, cada una con una probabilidad de éxito P. Entonces, X sigue una distribución binomial de **media**

$$\mu = E(X) = nP \tag{5.19}$$

y **varianza**

$$\sigma^2 = E[(X-\mu)^2] = nP(1-P) \tag{5.20}$$

La distribución binomial se utiliza mucho en aplicaciones empresariales y económicas en las que se quiere hallar la probabilidad de ocurrencias discretas. Antes de utilizar la distribución binomial, debe analizarse la situación específica para ver si

1. En la aplicación se realizan varias pruebas, cada una de las cuales sólo tiene dos resultados: sí o no, encendido o apagado, éxito o fracaso.
2. La probabilidad del resultado es la misma en cada prueba.
3. La probabilidad del resultado de una prueba no afecta a la probabilidad del resultado de otras pruebas.

En los siguientes ejemplos se muestran algunas aplicaciones representativas.
Las probabilidades de una distribución binomial pueden hallarse utilizando:

1. La ecuación 5.18 (buena cuando los valores de n son bajos); véase el ejemplo 5.7.
2. Las tablas del apéndice (buenas para un valor seleccionado de n y P); véase el ejemplo 5.8.
3. Probabilidades obtenidas por computador; véase el ejemplo 5.9.

EJEMPLO 5.7. Múltiples ventas de seguros (cálculos binomiales)

Suponga que Susana Ferrater, la agente de seguros del ejemplo 5.6, contacta con cinco personas y cree que la probabilidad de vender un seguro a cada una es de 0,40. Utilizando la ecuación 5.18:

a) Halle la probabilidad de que venda como máximo un seguro.
b) Halle la probabilidad de que venda entre dos y cuatro seguros (inclusive).
c) Represente gráficamente la función de probabilidad.

Solución

a) $P(\text{como máximo 1 venta}) = P(X \leqslant 1) = P(X = 0) + P(X = 1) = 0{,}078 + 0{,}259 = 0{,}337$, ya que

$$P(\text{ninguna venta}) = P(0) = \frac{5!}{0!5!}(0{,}4)^0(0{,}6)^5 = 0{,}078$$

$$P(1 \text{ venta}) = P(1) = \frac{5!}{1!4!}(0{,}4)^1(0{,}6)^4 = 5(0{,}4)(0{,}6)^4 = 0{,}259$$

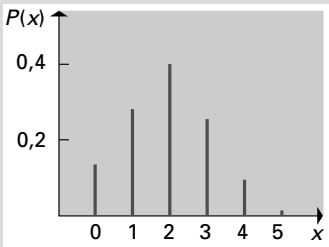
b) $P(2 \leqslant X \leqslant 4) = P(2) + P(3) + P(4) = 0{,}346 + 0{,}230 + 0{,}077 = 0{,}653$, ya que

$$P(2) = \frac{5!}{2!3!}(0{,}4)^2(0{,}6)^3 = 10(0{,}4)^2(0{,}6)^3 = 0{,}346$$

$$P(3) = \frac{5!}{3!2!}(0{,}4)^3(0{,}6)^2 = 10(0{,}4)^3(0{,}6)^2 = 0{,}230$$

$$P(4) = \frac{5!}{4!1!}(0{,}4)^4(0{,}6)^1 = 5(0{,}4)^4(0{,}6)^1 = 0{,}077$$

c) La Figura 5.6 muestra la función de probabilidad.



Comentarios

- Esta forma es representativa de las posibilidades binomiales cuando el valor de P no es ni muy grande ni muy pequeño.
- En los extremos (0 ventas o 5), las probabilidades son bastante pequeñas.

Figura 5.6. Gráfico de la función de probabilidad binomial del ejemplo 5.7 ($n = 5$, $P = 0{,}40$).

El cálculo de probabilidades binomiales tiende a ser muy tedioso, a menos que el número de pruebas n sea muy pequeño. Las probabilidades binomiales también pueden consultarse en las tablas del apéndice.

EJEMPLO 5.8. Admisiones en una universidad (cálculo de las probabilidades binomiales por medio de tablas)

A principios de agosto, una universidad descubre que puede admitir a algunos estudiantes más. La admisión de esos estudiantes aumentaría significativamente los ingresos sin incrementar los costes de explotación de la universidad; es decir, no habría que abrir nuevas clases. La universidad sabe por experiencia que el 40 por ciento de los estudiantes admitidos se matricula realmente.

a) ¿Cuál es la probabilidad de que se matriculen como máximo 6 estudiantes si la universidad admite a 10 estudiantes más?
b) ¿Cuál es la probabilidad de que se matriculen más de 12 estudiantes si admite a 20?
c) Si se matricula el 70 por ciento de los estudiantes admitidos, ¿cuál es la probabilidad de que se matriculen al menos 12 de 15 estudiantes admitidos?

Solución

a) Esta probabilidad puede hallarse utilizando la distribución de probabilidad binomial acumulada de la tabla 3 del apéndice. La probabilidad de que se matriculen como máximo 6 estudiantes si $n = 10$ y $P = 0{,}40$ es

$$P(X \leqslant 6 \,|\, n = 10,\ P = 0{,}40) = 0{,}945$$

b) $P(X > 12 \,|\, n = 20,\ P = 0{,}40) = 1 - P(X \leqslant 12) = 1 - 0{,}979 = 0{,}021$.
c) La probabilidad de que se matriculen al menos 12 de 15 estudiantes es igual que la probabilidad de que no se matriculen como máximo 3 de 15 estudiantes (la probabilidad de que no se matricule un estudiante es $1 - 0{,}70 = 0{,}30$).

$$P(X \geqslant 12 \,|\, n = 15,\ P = 0{,}70) = P(X \leqslant 3 \,|\, n = 15,\ P = 0{,}30) = 0{,}297$$

La mayoría de los paquetes informáticos buenos pueden calcular probabilidades binomiales y de otros tipos para diversas funciones de probabilidad. El ejemplo 5.9 muestra el método utilizando el programa Minitab, pero también pueden emplearse otros paquetes informáticos.

EJEMPLO 5.9. Ventas de plazas en una compañía aérea (cálculo de probabilidades binomiales por medio del programa Minitab)

¿Ha aceptado el lector alguna vez renunciar a un billete de avión a cambio de un billete gratis? ¿Ha buscado alguna vez el billete más barato para poder ir a ver a un amigo especial? El ejemplo siguiente permite analizar los casos en los que se venden más billetes que plazas hay en un avión y en los que se ofrecen tarifas más bajas en algunos vuelos.

Suponga que es responsable de la venta de las plazas de avión de una gran compañía aérea. Cuatro días antes de la fecha del vuelo, quedan 16 plazas libres. Sabemos por experiencia que el 80 por ciento de las personas que compran un billete en este periodo de tiempo se presenta el día del vuelo.

a) Si vende 20 billetes más, ¿cuál es la probabilidad de que el número de personas que se presentan sea mayor que el de plazas o de que haya al menos una plaza libre?

b) Si vende 18 billetes más, ¿cuál es la probabilidad de que el número de personas que se presentan sea mayor que el de plazas o de que haya al menos una plaza libre?

Solución

a) Para hallar $P(X > 16)$, dados $n = 20$ y $P = 0{,}80$, utilizamos el programa Minitab siguiendo las instrucciones de la Figura 5.7. Con el Minitab, el usuario debe seleccionar *o* Probability [como $P(X = 16)$] *o* Cumulative Probability [$P(X = 16)$], pero no las dos simultáneamente.

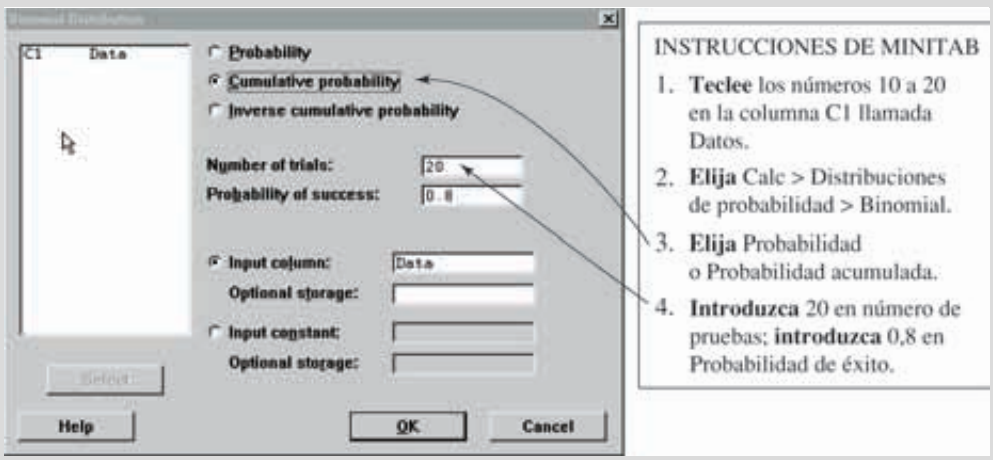


Figura 5.7. Cuadro de diálogo para una probabilidad binomial en la que $n = 20$, $P = 0{,}80$ utilizando el programa Minitab.

Tabla 5.4. Probabilidades binomiales obtenidas utilizando el programa Minitab, siendo $n = 20$, $P = 0{,}80$.

X	$P(X <= X)$
10	0,0026
11	0,0100
12	0,0321
13	0,0867
14	0,1958
15	0,3704
16	0,5886
17	0,7939
18	0,9308
19	0,9885
20	1,0000

Comentarios

- Hallar la probabilidad de que el número de pasajeros sea mayor que el de plazas vendidas,

$$P(X > 16) = 1 - P(X <= 16) = 1 - 0{,}589 = 0{,}411$$

- Si se venden 20 billetes, eso también significa que la probabilidad de que se presenten 15 personas o menos es

$$P(X <= 15) = 0{,}37$$

- Es decir, hay un 37 por ciento de probabilidades de que si se venden 20 billetes, ¡haya al menos una plaza libre!

b) Para hallar la probabilidad de que vendiendo 18 billetes, el número de personas que se presentan sea mayor que el de plazas vendidas, seguimos los mismos pasos que antes. La probabilidad de que el número de pasajeros sea mayor que el de plazas será del 10 por ciento solamente, ¡pero la probabilidad de que haya al menos una plaza libre aumentará a un 72,9 por ciento!

La dirección de la compañía aérea debe comparar, pues, el coste de ofrecer más billetes que plazas (facilitando billetes gratis) con el coste de quedarse con plazas libres que no generan ningún ingreso. Las compañías aéreas analizan los datos para averiguar el número de plazas que deben venderse a tarifas más bajas con el fin de maximizar los ingresos generados por los billetes en cada vuelo. Este análisis es complejo, pero tiene su punto de partida en análisis como el ejemplo que hemos presentado aquí.

EJERCICIOS

Ejercicios básicos

5.30. Dada una variable aleatoria de Bernoulli que tiene una probabilidad de éxito $P = 0{,}5$, calcule la media y la varianza.

5.31. Dada una función de probabilidad binomial en la que $P = 0{,}5$ y $n = 12$, halle la probabilidad de que el número de éxitos sea igual a 7 la probabilidad de que el número de éxitos sea menor que 6.

5.32. Dada una función de probabilidad binomial en la que $P = 0{,}3$ y $n = 14$, halle la probabilidad de que el número de éxitos sea igual a 7 y la probabilidad de que el número de éxitos sea menor que 6.

5.33. Dada una función de probabilidad binomial en la que $P = 0{,}4$ y $n = 20$, halle la probabilidad de que el número de éxitos sea igual a 9 y la probabilidad de que el número de éxitos sea menor que 7.

5.34. Dada una función de probabilidad binomial en la que $P = 0{,}7$ y $n = 18$, halle la probabilidad de que el número de éxitos sea igual a 12 y la probabilidad de que el número de éxitos sea menor que 6.

Ejercicios aplicados

5.35. Un director de producción sabe que el 5 por ciento de los componentes producidos en un determinado proceso de producción tiene algún defecto. Se examinan seis de estos componentes, cuyas características puede suponerse que son independientes entre sí.

a) ¿Cuál es la probabilidad de que ninguno de estos componentes tenga un defecto?
b) ¿Cuál es la probabilidad de que uno de estos componentes tenga un defecto?
c) ¿Cuál es la probabilidad de que al menos dos de estos componentes tengan un defecto?

5.36. Un político cree que el 25 por ciento de todos los macroeconomistas que ocupan altos cargos apoyará firmemente una propuesta que desea presentar. Suponga que esta creencia es correcta y que se seleccionan cinco macroeconomistas aleatoriamente.

a) ¿Cuál es la probabilidad de que al menos uno de los cinco apoye firmemente la propuesta?
b) ¿Cuál es la probabilidad de que la mayoría de los cinco apoye firmemente la propuesta?

5.37. Una organización de interés público contrata estudiantes para pedir donaciones por teléfono. Tras un breve periodo de formación, los estudiantes llaman a posibles donantes y cobran a comisión. La experiencia indica que al principio los estudiantes tienden a tener poco éxito y que el 70 por ciento deja el trabajo a las dos semanas. La organización contrata seis estudiantes, que pueden concebirse como una muestra aleatoria.

a) ¿Cuál es la probabilidad de que al menos dos de los seis dejen el trabajo en las dos primeras semanas?
b) ¿Cuál es la probabilidad de que al menos dos de los seis no dejen el trabajo en las dos primeras semanas?

5.38. Suponga que la probabilidad de que el valor del dólar estadounidense suba frente al yen japonés es de 0,5 y que el resultado de una semana es independiente del resultado de cualquier otra. ¿Cuál es la probabilidad de que el valor del dólar suba en relación con el yen japonés la mayoría de las semanas durante un periodo de 7 semanas?

5.39. Una empresa instala calefacciones centrales y ha observado que en el 15 por ciento de todas las instalaciones es necesario volver para hacer algunas modificaciones. Suponga que los resultados de estas instalaciones son independientes.

a) ¿Cuál es la probabilidad de que sea necesario volver en todos estos casos?
b) ¿Cuál es la probabilidad de que no sea necesario volver en ninguno de estos casos?
c) ¿Cuál es la probabilidad de que sea necesario volver en más de uno de estos casos?

5.40. Los Verdes van a jugar cinco partidos contra los Azules. Se estima que la probabilidad de que ganen los Verdes en cualquier partido es 0,4. Los resultados de los cinco partidos son independientes entre sí.

a) ¿Cuál es la probabilidad de que los Verdes ganen los cinco partidos?
b) ¿Cuál es la probabilidad de que los Verdes ganen la mayoría de los cinco partidos?
c) Si los Verdes ganan el primer partido, ¿cuál es la probabilidad de que ganen la mayoría de los cinco partidos?
d) Antes de que comiencen los partidos, ¿cuál es el número de partidos que se espera que ganen los Verdes?
e) Si los Verdes ganan el primer partido, ¿cuál es el número de partidos que se espera que ganen los Verdes?

5.41. Una pequeña compañía aérea tiene aviones que pueden llevar hasta ocho pasajeros. Ha calculado que la probabilidad de que no se presente un pasajero con un billete es de 0,2. Vende billetes para cada vuelo a las 10 primeras personas que piden un billete. La tabla adjunta muestra la distribución de probabilidad del número de billetes vendidos por vuelo. ¿En qué proporción de vuelos de la compañía es mayor el número de pasajeros que se presentan con billete que el número de plazas disponibles? Suponga que el número de billetes vendidos y la probabilidad de que se presente un pasajero con un billete son independientes.

Número de billetes	6	7	8	9	10
Probabilidad	0,25	0,35	0,25	0,10	0,05

5.42. Tras un ensayo, un entrenador de fútbol americano universitario tiene la opción de intentar «una conversión de 2 puntos», es decir, anotar 2 puntos más si el intento tiene éxito y ninguno si fracasa. El entrenador cree que la probabilidad de que su equipo tenga éxito en cualquier intento es 0,4 y que los resultados de los diferentes intentos son independientes entre sí. En un partido, el equipo logra cuatro ensayos y en cada uno intenta la conversión de 2 puntos.

a) ¿Cuál es la probabilidad de que tengan éxito al menos dos de estos intentos?
b) Halle la media y la desviación típica del número total de puntos resultantes de estos cuatro intentos.

5.43. Un concesionario de automóviles organiza una nueva campaña de promoción. Los compradores de nuevos automóviles pueden devolverlos en el plazo de 2 días si no están satisfechos y recuperar todo el dinero pagado. El coste que tiene para el concesionario la devolución del dinero es de 250 $. El concesionario estima que el 15 por ciento de todos los compradores devolverá los automóviles y recuperará el dinero. Suponga que se compran 50 automóviles durante la campaña.

a) Halle la media y la desviación típica del número de automóviles que se devolverán a cambio del dinero.
b) Halle la media y la desviación típica de los costes totales de la devolución del dinero de estas 50 compras.

5.44. Una sociedad de fondos de inversión tiene un servicio que permite a los clientes hacer transferencias de dinero de unas cuentas a otras por teléfono. Se estima que el 3,2 por ciento de los clientes que llaman se encuentra con que la línea está ocupada o se los mantiene tanto tiempo a la espera que cuelgan. La dirección estima que cualquier fallo de este tipo es una pérdida de clientela valorada en 10 $. Suponga que se intenta hacer 2.000 llamadas en un determinado periodo.

a) Halle la media y la desviación típica del número de personas que llaman y que se encuentran con la línea ocupada o cuelgan después de que se las mantenga a la espera.
b) Halle la media y la desviación típica de la pérdida total de clientela que experimenta la sociedad de fondos de inversión en estas 2.000 llamadas.

5.45. Hemos visto que en una distribución binomial con n pruebas, cada una de las cuales tiene una probabilidad de éxito P, la media es

$$\mu_X = E(X) = nP$$

Verifique este resultado con los datos del ejemplo 5.7 calculando la media directamente a partir de

$$\mu_X = \Sigma\, xP(x)$$

demostrando que en el caso de la distribución binomial las dos fórmulas dan la misma respuesta.

5.46. El jefe de la sección de recaudación del municipio de Callesanchas observa que, de todas las multas de aparcamiento que se ponen, se paga el 78 por ciento. La multa es de 2 $. En la semana más reciente, se han puesto 620 multas.

a) Halle la media y la desviación típica del número de multas que se pagan.
b) Halle la media y la desviación típica de la cantidad de dinero que se obtiene por el pago de estas multas.

5.47. Una empresa recibe un gran envío de componentes. Se comprobará una muestra aleatoria de 16 de estos componentes y se aceptará el envío si son defectuosos menos de 2 componentes de esta muestra. Halle cuál es la probabilidad de que se acepte un envío que contenga:

a) Un 5 por ciento de componentes defectuosos.
b) Un 15 por ciento de componentes defectuosos.
c) Un 25 por ciento de componentes defectuosos.

5.48. Están considerándose las dos reglas de aceptación siguientes para averiguar si se debe aceptar el envío de una gran remesa de componentes:

- Comprobar una muestra aleatoria de 10 componentes y aceptar el envío únicamente si ninguno de ellos es defectuoso.
- Comprobar una muestra aleatoria de 20 componentes y aceptar el envío únicamente si no hay más de uno defectuoso.

¿Con cuál de estas reglas de aceptación es menor la probabilidad de aceptar un envío que contenga un 20 por ciento de componentes defectuosos?

5.49. Una empresa recibe grandes envíos de piezas de dos fuentes. El 70 por ciento de los envíos procede de un proveedor cuyos envíos normalmente contienen un 10 por ciento de piezas defectuosas, mientras que el resto procede de un proveedor

cuyos envíos normalmente contienen un 20 por ciento de piezas defectuosas. Un directivo recibe un envío, pero desconoce la procedencia. Se comprueba una muestra aleatoria de 20 piezas de este envío y se observa que una de ellas es defectuosa. ¿Cuál es la probabilidad de que este envío proceda del proveedor más fiable? *Pista*: utilice el teorema de Bayes.

5.5. Distribución hipergeométrica

La distribución binomial presentada en el apartado 5.4 supone que los objetos se seleccionan independientemente y que la probabilidad de seleccionar uno es constante. En muchos problemas aplicados, estos supuestos pueden satisfacerse si se extrae una pequeña muestra de una gran población. Pero aquí examinamos una situación en la que es necesario seleccionar 5 empleados de un grupo de 15 igual de cualificados: una pequeña población. En el grupo de 15, hay 9 mujeres y 6 hombres. Supongamos que en el grupo de 5 empleados seleccionados, 3 son hombres y 2 son mujeres. ¿Cuál es la probabilidad de seleccionar ese grupo concreto si las selecciones se hacen aleatoriamente sin sesgo alguno? En el grupo inicial de 15, la probabilidad de seleccionar una mujer es 9/15. Si no se selecciona una mujer a la primera, la probabilidad de seleccionar una mujer a la segunda es 9/14. Por lo tanto, las probabilidades varían con cada selección. Como no se cumplen los supuestos de la distribución binomial, debe elegirse un modelo de probabilidad diferente. Esta distribución de probabilidad es la *distribución de probabilidad hipergeométrica*.

Podemos utilizar la distribución binomial en las situaciones que se denominan «muestreo con reposición». Si se repone el objeto seleccionado en la población, la probabilidad de seleccionar ese tipo de objeto sigue siendo la misma y se satisfacen los supuestos binomiales. En cambio, si no se reponen los objetos —«muestreo sin reposición»— las probabilidades varían con cada selección y, por lo tanto, el modelo de probabilidad que debe utilizarse es la distribución hipergeométrica. Si la población es grande ($N > 10.000$) y el tamaño de la muestra es pequeño ($< 1\%$), la variación de la probabilidad después de cada selección es muy pequeña. En esas situaciones, la distribución binomial es una aproximación muy buena y es la que se utiliza normalmente. La ecuación 5.21 muestra el modelo de probabilidad hipergeométrica.

Distribución hipergeométrica

Supongamos que se elige una muestra aleatoria de n objetos de un grupo de N objetos, de los cuales S son éxitos. La distribución del número de éxitos, X, en la muestra se llama **distribución hipergeométrica**. Su función de probabilidad es

$$P(x) = \frac{C_x^S C_{n-x}^{N-s}}{C_n^N} = \frac{\dfrac{S!}{x!(S-x)!} \times \dfrac{(N-S)!}{(n-x)!(N-S-n+x)!}}{\dfrac{N!}{n!(N-n)!}} \qquad \textbf{(5.21)}$$

donde x puede tomar valores enteros que van desde el mayor de 0 y $[n - (N - S)]$ hasta el menor de n y S.

En el apartado 4.3 explicamos la lógica de la distribución hipergeométrica utilizando la definición clásica de probabilidad y las fórmulas de recuento para las combinaciones. En la ecuación 5.21, los componentes son:

1. El número de formas en que pueden seleccionarse x éxitos en la muestra de un total de S éxitos contenidos en la población:

$$C_x^S = \frac{S!}{x!(S-x)!}$$

2. El número de formas en que pueden seleccionarse $n - x$ fracasos en la población que contiene $N - S$ fracasos:

$$C_{n-x}^{N-S} = \frac{(N-S)!}{(n-x)!(N-S-n+x)!}$$

3. Y, por último, el número total de muestras de tamaño n que pueden obtenerse en una población de tamaño N:

$$C_n^N = \frac{N!}{n!(N-n)!}$$

Cuando se combinan estos componentes utilizando la definición clásica de probabilidad, se obtiene la distribución de probabilidad hipergeométrica.

EJEMPLO 5.10. Envío de artículos (cálculo de la probabilidad hipergeométrica)

Una empresa recibe un envío de 20 artículos. Como es caro inspeccionarlos todos, tiene la política de comprobar una muestra aleatoria de 6 artículos de ese envío y, si no hay más de 1 artículo defectuoso en la muestra, no comprueba el resto. ¿Cuál es la probabilidad de que un envío de 5 artículos defectuosos no se someta a una comprobación adicional?

Solución

Si se identifica «artículo defectuoso» con «éxito» en este ejemplo, el envío contiene $N = 20$ artículos y $S = 5$ de los 20 que son éxitos. Se selecciona una muestra de $n = 6$ artículos. En ese caso, el número de éxitos, X, que hay en la muestra tiene una distribución hipergeométrica con la función de probabilidad

$$P(x) = \frac{C_x^S C_{n-x}^S}{C_n^N} = \frac{C_x^5 C_{6-x}^{15}}{C_6^{20}} = \frac{\dfrac{5!}{x!(5-x)!} \times \dfrac{15!}{(6-x)!(9+x)!}}{\dfrac{20!}{6!\,14!}}$$

El envío no se verifica más si la muestra contiene cero éxitos (artículos defectuosos) o uno, por lo que la probabilidad de que se acepte es

$$P(\text{envío aceptado}) = P(0) + P(1)$$

La probabilidad de que no haya artículos defectuosos en la muestra es

$$P(0) = \frac{\frac{5!}{0!\,5!} \times \frac{15!}{6!\,9!}}{\frac{20!}{6!\,14!}} = 0{,}129$$

La probabilidad de que haya 1 artículo defectuoso en la muestra es

$$P(1) = \frac{\frac{5!}{1!\,4!} \times \frac{15!}{5!\,10!}}{\frac{20!}{6!\,14!}} = 0{,}387$$

Por lo tanto, observamos que la probabilidad de que no se compruebe más el envío de 20 artículos que contenga 5 defectuosos es $P(\text{envío aceptado}) = P(0) + P(1) = = 0{,}129 + 0{,}387 = 0{,}516$. Esta tasa de error es alta e indica que es necesario mejorar el proceso.

Las probabilidades hipergeométricas también pueden calcularse utilizando programas informáticos mediante un método similar al empleado en el ejemplo 5.9 para calcular las probabilidades binomiales.

EJERCICIOS

Ejercicios básicos

5.50. Calcule la probabilidad de obtener 5 éxitos en una muestra aleatoria de tamaño $n = 12$ extraída de una población de tamaño $N = 50$ que contiene 25 éxitos.

5.51. Calcule la probabilidad de obtener 7 éxitos en una muestra aleatoria de tamaño $n = 14$ extraída de una población de tamaño $N = 60$ que contiene 25 éxitos.

5.52. Calcule la probabilidad de obtener 9 éxitos en una muestra aleatoria de tamaño $n = 20$ extraída de una población de tamaño $N = 80$ que contiene 42 éxitos.

5.53. Calcule la probabilidad de obtener 3 éxitos en una muestra aleatoria de tamaño $n = 5$ extraída de una población de tamaño $N = 40$ que contiene 25 éxitos.

5.54. Calcule la probabilidad de obtener 8 éxitos en una muestra aleatoria de tamaño $n = 15$ extraída de una población de tamaño $N = 400$ que contiene 200 éxitos.

Ejercicios de aplicación

5.55. Una empresa recibe un envío de 16 artículos. Se selecciona una muestra aleatoria de 4 y se rechaza el envío si cualquiera de estos artículos resulta defectuoso.

a) ¿Cuál es la probabilidad de que se acepte un envío que contiene 4 artículos defectuosos?

b) ¿Cuál es la probabilidad de que se acepte un envío que contiene 1 artículo defectuoso?

c) ¿Cuál es la probabilidad de que se rechace un envío que contiene 1 artículo defectuoso?

5.56. Hay que formar un comité de ocho miembros de un grupo de ocho hombres y ocho mujeres. Si los miembros del comité se eligen aleatoriamente, ¿cuál es la probabilidad de que exactamente la mitad sean mujeres?

5.57. Un analista de bonos recibió una lista de 12 bonos de empresa. Seleccionó de esa lista 3 cuya calificación creía que corría el riesgo de que se rebajara al año siguiente. En realidad, al año siguiente se rebajó la calificación de 4 de los 12 bonos. Suponga que el analista hubiera elegido

simplemente 3 bonos aleatoriamente de la lista. ¿Cuál es la probabilidad de que al menos 2 de los elegidos se encontraran entre los bonos cuya calificación se rebajó al año siguiente?

5.58. Un ejecutivo bancario recibe 10 solicitudes de crédito. Los perfiles de los solicitantes son similares, salvo que 5 pertenecen a minorías y 5 no. Al final, el ejecutivo autoriza 6 de las solicitudes. Si estas autorizaciones se eligen aleatoriamente del grupo de 10 solicitudes, ¿cuál es la probabilidad de que menos de la mitad de las autorizaciones sean autorizaciones de solicitudes de personas que pertenecen a minorías?

5.6. La distribución de Poisson

La *distribución de Poisson* fue propuesta por primera vez por Siméon Poisson (1781-1840) en un libro publicado en 1837. El número de aplicaciones comenzó a aumentar a principios del siglo XX y la aparición del computador ha permitido aumentarlas en el siglo XXI. La distribución de Poisson es una importante distribución de probabilidad discreta para algunas aplicaciones entre las que se encuentran las siguientes:

1. El número de fallos de un gran sistema informático en un día dado.
2. El número de pedidos de sustitución de una pieza recibido por una empresa en un mes dado.
3. El número de barcos que llegan a una terminal de carga durante un periodo de 6 horas.
4. El número de camiones de reparto que llegan a un almacén central en una hora.
5. El número de abolladuras, rasguños u otros defectos de un gran rollo de lámina de metal utilizada para fabricar filtros.
6. El número de clientes que llegan a tomar un vuelo cada 15 minutos entre las 3 y las 6 de la tarde durante los días de la semana.
7. El número de clientes que llegan a una caja en el supermercado local durante un determinado intervalo de tiempo.

Podemos utilizar la distribución de Poisson para hallar la probabilidad de cada una de estas variables aleatorias, que se caracterizan por ser el número de ocurrencias o de éxitos de un suceso en un intervalo continuo dado (como el tiempo, la superficie o la longitud).

La distribución de Poisson se basa en ciertos supuestos.

Supuestos de la distribución de Poisson

Supongamos que un intervalo está dividido en un gran número de subintervalos de manera que la probabilidad de que ocurra un suceso de cualquier subintervalo es muy pequeña. Los **supuestos de la distribución de Poisson** son los siguientes:

1. La probabilidad de que ocurra un suceso es constante en todos los subintervalos.
2. No puede haber más de una ocurrencia en cada subintervalo.
3. Las ocurrencias son independientes; es decir, las ocurrencias en intervalos que no se solapan son independientes entre sí.

Podemos formular directamente la ecuación para calcular las probabilidades de Poisson a partir de la distribución de probabilidad binomial tomando los límites matemáticos cuando $P \to 0$ y $n \to \infty$. Con estos límites, el parámetro $\lambda = nP$ es una constante que especifica el número medio de ocurrencias (éxitos) en un determinado tiempo y/o espacio. La ecuación 5.22 define la función de probabilidad de Poisson.

La función, la media y la varianza de la distribución de probabilidad de Poisson

Se dice que la variable aleatoria X sigue la **distribución de probabilidad de Poisson** si tiene la función de probabilidad

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{para } x = 0, 1, 2 \qquad \textbf{(5.22)}$$

donde

$P(x)$ = probabilidad de x éxitos en un tiempo o un espacio dados, dado λ
λ = número esperado de éxitos por unidad de tiempo o espacio; $\lambda > 0$
$e \cong 2{,}71828$ (la base de los logaritmos naturales)

La **media y la varianza de la distribución de probabilidad de Poisson** son

$$\mu = E(X) = \lambda \quad \text{y} \quad \sigma^2 = E[(X - \mu)^2] = \lambda$$

La suma de las variables aleatorias de Poisson también es una variable aleatoria de Poisson. Por lo tanto, la suma de K variables aleatorias de Poisson, cada una de media λ, es una variable aleatoria de Poisson de media $K\lambda$.

EJEMPLO 5.11. Fallos de los componentes de un sistema (probabilidades de Poisson)

Andrés Gutiérrez, director de un centro informático, informa de que su sistema informático ha experimentado tres fallos de componentes en los 100 últimos días.

a) ¿Cuál es la probabilidad de que no haya ningún fallo en un día dado?
b) ¿Cuál es la probabilidad de que haya uno o más fallos de componentes en un día dado?
c) ¿Cuál es la probabilidad de que haya al menos dos fallos en un periodo de tres días?

Solución

Un sistema informático moderno tiene un gran número de componentes, cada uno de los cuales puede fallar y provocar así un fallo del sistema informático. Para calcular la probabilidad de que haya fallos utilizando la distribución de Poisson, supongamos que cada uno de los millones de componentes tiene la misma pequeñísima probabilidad de fallar. Supongamos también que el primer fallo no afecta a la probabilidad de que haya un segundo fallo (en algunos casos, estos supuestos pueden no cumplirse, en cuyo caso se utilizarían distribuciones más complejas).

La experiencia dice que el número esperado de fallos al día es 3/100, o sea, $\lambda = 0{,}03$.

a) $P(\text{ningún fallo en un día dado}) = P(X = 0 \mid \lambda = 0{,}03)$

$$= \frac{e^{-0{,}03}\lambda^0}{0!} = 0{,}970446$$

b) La probabilidad de que haya al menos un fallo es el complementario de la probabilidad de que haya 0 fallos:

$$\begin{aligned} P(X \geqslant 1) &= 1 - P(X = 0) \\ &= 1 - \left[\frac{e^{-\lambda}\lambda^{x}}{x!}\right] = 1 - \left[\frac{e^{-0,03}\lambda^{0}}{0!}\right] \\ &= 1 - e^{-0,03} = 1 - 0,970446 = 0,029554 \end{aligned}$$

c) P(al menos dos fallos en un periodo de 3 días) $= P(X \geqslant 2 \mid \lambda = 0,09)$, donde la media en un periodo de 3 días es $\lambda = 3(0,03) = 0,09$:

$$\begin{aligned} P(X \geqslant 2 \mid \lambda = 0,09) &= 1 - P(X \leqslant 1) = 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - [0,913931 + 0,082254] \end{aligned}$$

y, por lo tanto,

$$P(X \geqslant 2 \mid \lambda = 0,09) = 1 - 0,996185 = 0,003815$$

Se ha observado que la distribución de Poisson es especialmente útil en los problemas de *listas de espera* o de *colas*. Ejemplos son el número de clientes que llegan a una caja de un supermercado, el número de camiones de reparto que llegan a un almacén central, el número de personas que se presentan a los vuelos, el número de estudiantes que aguardan a comprar libros de texto en la librería universitaria, etc. En la práctica, a menudo es posible representar los procesos de llegada de este tipo por medio de una distribución de Poisson.

EJEMPLO 5.12. Clientes de una fotocopiadora (probabilidad de Poisson)

Los clientes llegan a una fotocopiadora a una tasa media de dos cada 5 minutos. Suponga que estas llegadas son independientes, que la tasa de llegada es constante y que este problema sigue un modelo de Poisson, donde X representa el número de clientes que llegan en un periodo de 5 minutos y la media $\lambda = 2$. Halle la probabilidad de que lleguen más de dos clientes en un periodo de 5 minutos.

Solución

Como el número medio de llegadas en 5 minutos es dos, entonces $\lambda = 2$. Para hallar la probabilidad de que lleguen más de dos clientes, primero se calcula la probabilidad de que lleguen al menos dos en un periodo de 5 minutos y después se utiliza la regla del complementario.

Estas probabilidades pueden encontrarse en la Tabla 5 del apéndice o pueden calcularse por computador:

$$P(X = 0) = \frac{e^{-2}2^{0}}{0!} = e^{-2} = 0,1353$$

$$P(X = 1) = \frac{e^{-2}2^{1}}{1!} = 2e^{-2} = 0,2707$$

$$P(X = 2) = \frac{e^{-2}2^{2}}{2!} = 2e^{-2} = 0,2707$$

Por lo tanto, la probabilidad de que lleguen más de dos clientes en un periodo de 5 minutos es

$$P(X > 2) = 1 - P(X \leqslant 2) = 1 - [0{,}135335 + 0{,}27067 + 0{,}27067] = 0{,}323325$$

Aproximación de Poisson de la distribución binomial

Antes hemos señalado que la distribución de probabilidades de Poisson se obtiene partiendo de la distribución binominal, donde P tiende a 0 y n tiende a infinito. Por lo tanto, la distribución de Poisson puede utilizarse como aproximación de las probabilidades binomiales cuando el número de pruebas, n, es grande y al mismo tiempo la probabilidad, P, es pequeña (generalmente tal que $\lambda = nP \leqslant 7$). Ejemplos de situaciones que satisfarían estas condiciones son los siguientes:

- Una compañía de seguros tiene un gran número de pólizas de seguro de vida de individuos de una determinada edad y la probabilidad de que una póliza genere una reclamación durante el año es muy baja. En este caso, tenemos una distribución binomial con un valor de n grande y un valor de P pequeño.
- Una empresa puede tener un gran número de máquinas trabajando simultáneamente en un proceso. Si la probabilidad de que se averíe cualquiera de ellas en un día es pequeña, la distribución del número de averías diarias es binomial con un valor de n grande y un valor de P pequeño.

Aproximación de Poisson de la distribución binominal

Sea X el número de éxitos resultante de n pruebas independientes, cada una con una probabilidad de éxito P. La distribución del número de éxitos, X, es binomial de media nP. Si el número de pruebas, n, es grande y nP sólo tiene un tamaño moderado (preferiblemente $nP \leqslant 7$), es posible utilizar como **aproximación la distribución de Poisson**, en la que $\lambda = nP$. La función de probabilidad de la distribución aproximada es, pues,

$$P(x) = \frac{e^{-nP}(nP)^x}{x!} \quad \text{para } x = 0, 1, 2, \ldots \qquad \textbf{(5.23)}$$

EJEMPLO 5.13. Probabilidad de quiebra (probabilidad de Poisson)

Un analista ha predicho que el 3,5 por ciento de todas las pequeñas empresas quebrará el próximo año. Suponiendo que la predicción del analista es correcta, estime la probabilidad de que el próximo año quiebren al menos 3 pequeñas empresas de una muestra aleatoria de 100.

Solución

La distribución de X, el número de quiebras, es binomial, siendo $n = 100$ y $P = 0{,}035$, por lo que la media de la distribución es $\mu_x = nP = 3{,}5$. Utilizando la distribución de

Poisson como aproximación de la probabilidad de que haya al menos 3 quiebras, tenemos que

$$P(X \geqslant 3) = 1 - P(X \leqslant 2)$$

$$P(0) = \frac{e^{-3,5}(3,5)^0}{0!} = e^{-3,5} = 0,030197$$

$$P(1) = \frac{e^{-3,5}(3,5)^1}{1!} = (3,5)(0,030197) = 0,1056895$$

$$P(2) = \frac{e^{-3,5}(3,5)^2}{2!} = (6,125)(0,030197) = 0,1849566$$

Por lo tanto,

$$P(X \leqslant 2) = P(0) + P(1) + P(2) = 0,030197 + 0,1056895 + 0,1849566 = 0,3208431$$

$$P(X \geqslant 3) = 1 - 0,3208431 = 0,6791569$$

La probabilidad binominal de $X \geqslant 3$ es

$$P(X \geqslant 3) = 0,684093$$

La probabilidad de Poisson es simplemente una estimación de la probabilidad binomial efectiva.

Comparación de la distribución de Poisson y la distribución binomial

Llegados a este punto, debemos indicar que puede existir confusión a la hora de elegir la distribución binomial o la distribución de Poisson en una aplicación específica. En muchos casos, es más fácil elegir repasando atentamente los supuestos de las dos distribuciones de probabilidad. Por ejemplo, si el problema se basa en una pequeña muestra de observaciones, no es posible hallar una probabilidad límite cuando n es grande y, por lo tanto, la distribución binomial es la correcta. Además, si tenemos una pequeña muestra y la probabilidad de éxito en una única prueba está comprendida entre 0,05 y 0,95, hay más razones para elegir la distribución binomial. Si supiéramos o pudiéramos suponer que cada uno de 10 clientes seleccionados aleatoriamente en un concesionario de automóviles tienen la misma probabilidad de comprar un automóvil (supongamos que $0,05 \leqslant P \leqslant 0,95$), el número de compras de este grupo seguiría una distribución binomial. Sin embargo, si el conjunto de casos que podrían estar afectados es muy grande —por ejemplo, varios miles— y el número medio de «éxitos» en ese gran conjunto de casos es pequeño —por ejemplo, menos de 30—, hay muchas razones para elegir la distribución de Poisson. Si quisiéramos calcular la probabilidad de que haya un cierto número de piezas defectuosas en un grupo de 100.000 piezas cuando el número medio de 15 piezas defectuosas por 100.000 piezas representa un ciclo de producción representativo, utilizaríamos la distribución de Poisson.

En el análisis anterior, hemos señalado que cuando P es menor que 0,05 y n es grande, podemos utilizar la distribución de Poisson como aproximación de la distribución binomial. También puede demostrarse que cuando $n \geqslant 20$ y $P \leqslant 0,05$ y la media poblacional es la misma, se observa que los valores de la probabilidad son los mismos con la distribución binomial que con la distribución de Poisson.

EJERCICIOS

Ejercicios básicos

5.59. Halle la probabilidad de obtener 7 éxitos exactamente en el caso de una variable aleatoria que sigue una distribución de Poisson, siendo $\lambda = 3{,}5$.

5.60. Halle la probabilidad de obtener 4 éxitos exactamente en el caso de una variable aleatoria que sigue una distribución de Poisson, siendo $\lambda = 2{,}5$.

5.61. Halle la probabilidad de obtener más de 7 éxitos en el caso de una variable aleatoria que sigue una distribución de Poisson, siendo $\lambda = 4{,}5$.

5.62. Halle la probabilidad de obtener menos 6 éxitos en el caso de una variable aleatoria que sigue una distribución de Poisson, siendo $\lambda = 3{,}5$.

5.63. Halle la probabilidad de obtener menos de 9 éxitos o 9 éxitos en el caso de una variable aleatoria que sigue una distribución de Poisson, siendo $\lambda = 8{,}0$.

Ejercicios de aplicación

5.64. Los clientes llegan a una caja registradora ocupada a una tasa media de tres por minuto. Si las llegadas siguen una distribución de Poisson, halle la probabilidad de que en un minuto dado lleguen dos clientes o menos.

5.65. El número de accidentes que se producen en una fábrica tiene una distribución de Poisson con una media de 2,6 al mes.

a) ¿Cuál es la probabilidad de que haya menos de dos accidentes en un mes dado?

b) ¿Cuál es la probabilidad de que haya más de tres accidentes en un mes dado?

5.66. Un profesor recibe, por término medio, 4,2 llamadas telefónicas de los estudiantes el día antes del examen final. Si las llamadas siguen una distribución de Poisson, ¿cuál es la probabilidad de que reciba al menos tres llamadas ese día?

5.67. Los datos indican que en la hora punta de la mañana se producen, por término medio, 3,2 colisiones al día en una vía urbana. Suponga que la distribución es de Poisson.

a) Halle la probabilidad de que en un día dado se produzcan menos de dos colisiones en esta vía durante la hora punta de la mañana.

b) Halle la probabilidad de que en un día dado se produzcan más de cuatro colisiones en esta vía durante la hora punta de la mañana.

5.68. Hacienda ha informado de que el 5,5 por ciento de todos los contribuyentes comete errores al rellenar los impresos de declaración de la renta. Si se eligen aleatoriamente 100 declaraciones, ¿cuál es la probabilidad de que menos de 3 contengan errores? Utilice la aproximación de Poisson de la distribución binomial.

5.69. Una empresa tiene 250 computadores personales. La probabilidad de que uno cualquiera de ellos necesite una reparación en una semana dada es 0,01. Halle la probabilidad de que menos de 4 de los computadores personales necesiten una reparación en una semana dada. Utilice la aproximación de Poisson de la distribución binomial.

5.70. Una compañía de seguros tiene 6.000 pólizas de seguro contra las estafas con otras tantas empresas. En un año dado, la probabilidad de que una póliza genere una reclamación es de 0,001. Halle la probabilidad de que se presenten al menos tres reclamaciones en un año dado. Utilice la aproximación de Poisson de la distribución binomial.

5.71. Por ley, los automovilistas deben tener un seguro. Se ha estimado que, a pesar de la ley, el 7,5 por ciento de todos los automovilistas no tiene seguro. Se ha tomado una muestra aleatoria de 60 automovilistas. Utilice la aproximación de Poisson de la distribución binomial para estimar la probabilidad de que al menos 3 de los automovilistas de esta muestra no estén asegurados. Indique también qué cálculos tendría que hacer para hallar esta probabilidad exactamente si no utilizara la aproximación de Poisson.

5.72. Está diseñándose un nuevo almacén y hay que tomar una decisión sobre el número de zonas de carga. Hay dos modelos para el uso de este almacén, dado que para cargar un camión se necesita 1 hora. El almacén podría contratar a uno de los muchos miles de camioneros independientes que llegan aleatoriamente para recoger una carga y distribuirla. Se sabe que cada hora llega en promedio uno de estos camiones. La empresa también podría contratar una flota de 10 camiones dedicados a tiempo completo a transportar envíos de este almacén. Partiendo de ese supuesto, los camiones llegarían aleatoriamente, pero la probabilidad de que llegara uno durante una hora dada es 0,1. Halle la distribución de probabilidad adecuada para cada uno de estos supuestos y compare los resultados. Los valores de la distribución de probabilidad pueden consultarse en las Tablas 2 y 5 del Apéndice o calcularse por computador.

5.7. Distribución conjunta de variables aleatorias discretas

Las aplicaciones empresariales y económicas de estadística a menudo se refieren a las relaciones entre variables. Los precios de los productos de diferentes niveles de calidad se fijan a diferentes intervalos. Los grupos de edad tienen diferentes preferencias por la ropa, los automóviles y la música. Los rendimientos porcentuales de las acciones de dos empresas distintas pueden tender a estar relacionados y la probabilidad de que los rendimientos de las acciones de las dos sean más altos puede aumentar cuando el mercado está creciendo. También puede ocurrir que, cuando los rendimientos de las acciones de una empresa están aumentando, los de las acciones de la otra estén disminuyendo. Cuando trabajamos con modelos de probabilidad para resolver problemas en los que hay relaciones entre variables, es importante incluir en el modelo el efecto de estas relaciones. Supongamos, por ejemplo, que un concesionario de automóviles tiene en venta los siguientes automóviles: (1) un utilitario rojo de dos puertas, (2) un monovolumen azul y (3) un sedán plateado; la distribución de probabilidad de comprar un automóvil de una mujer que tiene entre 20 y 30 años no sería igual que la de una que tiene entre 30 y 40 y que la de una que tiene entre 50 y 60. Es importante, pues, que los modelos de probabilidad reflejen el efecto conjunto que producen las variables en las probabilidades.

En el apartado 4.6 analizamos las probabilidades conjuntas. Ahora consideraremos el caso en el que se examinan dos o más variables aleatorias discretas que pueden estar relacionadas. Cuando hay una única variable aleatoria, las probabilidades de todos los resultados posibles pueden resumirse en una función de probabilidad, mientras que ahora tenemos que definir las probabilidades de que las variables aleatorias que nos interesan tomen simultáneamente valores específicos. Consideremos el siguiente ejemplo que implica el uso de una distribución conjunta de variables aleatorias discretas.

EJEMPLO 5.14. Estudio de mercado (probabilidades conjuntas)

A Sara Perales, analista de mercado, le han pedido que desarrolle un modelo de probabilidad para la relación entre la venta de utensilios de cocina de lujo y el grupo de edad. Este modelo es importante para desarrollar una campaña de marketing para una nueva línea de utensilios de cocina de lujo. Cree que las pautas de compra de utensilios de cocina de lujo varían de unos grupos de edad a otros.

Solución

Para representar el mercado, Sara propone utilizar tres grupos de edad —de 16 a 25 años, de 26 a 45 años y de 46 a 65 años— y dos pautas de compra: «comprar» y «no comprar». A continuación, recoge una muestra aleatoria de personas de 16-65 años y anota su grupo de edad y su deseo de comprar. El resultado de este conjunto de datos es la distribución de probabilidad conjunta de la Tabla 5.5. Esta tabla es, pues, un resumen de la probabilidad de compra y el grupo de edad que será un recurso valioso para el estudio de mercado.

Tabla 5.5. Distribución de probabilidad conjunta del grupo de edad (X) frente a la decisión de compra (Y).

	Grupo de edad (X)			
Decisión de compra (Y)	**1 (16-25)**	**2 (26-45)**	**3 (46-65)**	$P(y)$
1 (comprar)	0,10	0,20	0,10	0,40
2 (no comprar)	0,25	0,25	0,10	0,60
$P(x)$	0,35	0,45	0,20	1,00

Función de probabilidad conjunta

Sean X e Y un par de variables aleatorias discretas. Su **función de probabilidad conjunta** expresa la probabilidad de que simultáneamente X tome el valor específico x e Y tome el valor y como función de x e y. Señalamos que este análisis es una extensión directa del apartado 4.4, en el que presentamos la probabilidad de la intersección de dos sucesos, $P(A_i \cap B_j)$. Aquí utilizamos variables aleatorias. La notación empleada es $P(x, y)$, de donde

$$P(x, y) = P(X = x \cap Y = y)$$

A menudo se desea formular las funciones de probabilidad de las variables aleatorias individuales cuando se analizan variables aleatorias distribuidas conjuntamente.

Obtención de la función de probabilidad marginal

Sean X e Y un par de variables aleatorias distribuidas conjuntamente. En este contexto, la función de probabilidad de la variable aleatoria X se llama **función de probabilidad marginal** y se obtiene sumando las probabilidades conjuntas correspondientes a todos los valores posibles; es decir,

$$P(x) = \sum_{y} P(x, y) \qquad \textbf{(5.24)}$$

Asimismo, la función de probabilidad marginal de la variable aleatoria Y es

$$P(y) = \sum_{x} P(x, y) \qquad \textbf{(5.25)}$$

En la fila inferior y la columna derecha de la Tabla 5.5 se muestra un ejemplo de estas funciones de probabilidad marginal.

Las funciones de probabilidad conjunta deben tener las siguientes propiedades.

Propiedades de las funciones de probabilidad conjunta de variables aleatorias discretas

Sean X e Y variables aleatorias discretas que tienen una función de probabilidad conjunta $P(x, y)$.

1. $0 < P(x, y) < 1$ para cualquier par de valores x e y.
2. La suma de las probabilidades conjuntas $P(x, y)$ correspondientes a todos los pares posibles de valores debe ser 1.

La *función de probabilidad condicionada* de una variable aleatoria, dados valores específicos de otra, es el conjunto de probabilidades condicionadas.

Función de probabilidad condicionada

Sean X e Y un par de variables aleatorias discretas distribuidas conjuntamente. La **función de probabilidad condicionada** de la variable aleatoria Y, dado que la variable aleatoria X toma el valor x, expresa la probabilidad de que Y tome el valor y en función de y cuando se especifica el valor x de X. Esta función se representa por medio de $P(y \mid x)$ y, por lo tanto, por la definición de probabilidad condicionada

$$P(y \mid x) = \frac{P(x, y)}{P(x)} \qquad \textbf{(5.26)}$$

Asimismo, la función de probabilidad condicionada de X, dado $Y = y$, es

$$P(x \mid y) = \frac{P(x,\ y)}{P(y)} \tag{5.27}$$

Por ejemplo, utilizando las probabilidades de la Tabla 5.5, podemos calcular la probabilidad condicionada de compra ($y = 1$), dado el grupo de edad 26-45 ($x = 2$), de la forma siguiente:

$$P(1 \mid 2) = \frac{P(2,\ 1)}{P(2)} = \frac{0{,}20}{0{,}45} = 0{,}44$$

En el Capítulo 4 analizamos la independencia de los sucesos. Este concepto se extiende directamente a las variables aleatorias.

Independencia de las variables aleatorias distribuidas conjuntamente

Se dice que las variables aleatorias distribuidas conjuntamente X e Y son **independientes** si y sólo si su función de probabilidad conjunta es el producto de sus funciones de probabilidad marginal; es decir, si y sólo si

$$P(x,\ y) = P(x)P(y)$$

para todos los pares posibles de valores x e y. Y k variables aleatorias son independientes si y sólo si

$$P(X_1,\ X_2,\ ...,\ X_k) = P(X_1)P(X_2) \cdots P(X_k) \tag{5.28}$$

De la definición de funciones de probabilidad condicionada se deduce que, si las variables aleatorias X e Y son independientes, la función de probabilidad condicionada de Y, dado X, es igual que la función de probabilidad marginal de Y; es decir,

$$P(y \mid x) = P(y)$$

Asimismo, se deduce que

$$P(x \mid y) = P(x)$$

En el ejemplo 5.15 se analizan los rendimientos porcentuales posibles de las acciones de dos empresas, A y B; muestra cómo se calculan las probabilidades marginales, se hace un contraste de la independencia y se calcula las medias y las varianzas de dos variables aleatorias distribuidas conjuntamente.

EJEMPLO 5.15. Los rendimientos de las acciones, la probabilidad marginal, la media, la varianza (probabilidades conjuntas)

Supongamos que Carlota Reina tiene acciones de dos empresas, A y B. Sean X e Y variables aleatorias de los rendimientos porcentuales posibles (0 por ciento, 5 por ciento, 10 por ciento y 15 por ciento) de las acciones de cada una de estas dos empresas; la Tabla 5.6 muestra la distribución de probabilidad conjunta.

a) Halle las probabilidades marginales.
b) Averigüe si X e Y son independientes.
c) Halle las medias y las varianzas tanto de X como de Y.

Tabla 5.6. Distribución de probabilidad conjunta de las variables aleatorias X e Y.

	Rendimiento de Y			
Rendimiento de X	**0%**	**5%**	**10%**	**15%**
0%	0,0625	0,0625	0,0625	0,0625
5%	0,0625	0,0625	0,0625	0,0625
10%	0,0625	0,0625	0,0625	0,0625
15%	0,0625	0,0625	0,0625	0,0625

Solución

a) Este problema se resuelve utilizando las definiciones presentadas en este capítulo. Obsérvese que para cada combinación de valores de X e Y, $P(x, y) = 0{,}0625$. Es decir, todas las combinaciones posibles de rendimientos x e y tienen un 6,25 por ciento de probabilidades. Para hallar la probabilidad marginal de que X tenga un rendimiento de 0 por ciento,

$$P(X = 0) = \sum_{y} P(0, y) = 0{,}0625 + 0{,}0625 + 0{,}0625 + 0{,}0625 = 0{,}25$$

Aquí todas las probabilidades marginales de X son del 25 por ciento. Obsérvese que la suma de las probabilidades marginales es 1. Los resultados son similares en el caso de las probabilidades marginales de Y.

b) Para contrastar la independencia, tenemos que comprobar si $P(x, y) = P(x)P(y)$ para todos los pares posibles de valores x e y.

$$P(x, y) = 0{,}0625 \text{ para todos los pares posibles de valores } x \text{ e } y$$
$$P(x) = 0{,}25 \text{ y } P(y) = 0{,}25 \text{ para todos los pares posibles de valores } x \text{ e } y$$
$$P(x, y) = 0{,}0625 = (0{,}25)\,(0{,}25) = P(x)P(y)$$

Por lo tanto, X e Y son independientes.

c) La media de X es

$$\mu_X = E(X) = \sum_{X} xP(x)$$
$$= 0(0{,}25) + 0{,}05(0{,}25) + 0{,}10(0{,}25) + 0{,}15(0{,}25) = 0{,}075$$

Asimismo, la media de Y es $\mu_Y = E(y) = 0{,}075$.

La varianza de X es

$$\sigma_X^2 = \sum_{X} (x - \mu_X)^2 P(x) = P(x) \sum_{X} (x - \mu_X)^2 = (0{,}25) \sum_{x} (x - \mu_X)^2$$
$$= (0{,}25)[(0 - 0{,}075)^2 + (0{,}05 - 0{,}075)^2(0{,}10 - 0{,}075)^2 + (0{,}15 - 0{,}075)^2]$$
$$= 0{,}003125$$

y la desviación típica de X es $\sigma_X = \sqrt{0{,}003125} = 0{,}0559016$, o sea, 5,59 por ciento. Para hallar la varianza y la desviación típica de Y se siguen los mismos pasos.

Aplicaciones informáticas

Actualmente no existe ningún modelo complementario específico que permita calcular fácilmente las probabilidades marginales, las medias y las varianzas de variables aleatorias distribuidas conjuntamente. Sin embargo, podemos desarrollar fórmulas en Excel para simplificar el trabajo. Para calcular probabilidades marginales, medias y varianzas de variables aleatorias distribuidas conjuntamente X e Y por medio del programa Microsoft Excel, sígase el ejemplo de la Figura 5.8.

		Y Return						
X Return	0%	5%	10%	15%	P(x)	Mean of	Var of X	StDev of
0%	0.0625	0.0625	0.0625	0.0625	0.25	0	0.0014063	
5%	0.0625	0.0625	0.0625	0.0625	0.25	0.0125	0.0001563	
10%	0.0625	0.0625	0.0625	0.0625	0.25	0.025	0.0001563	
15%	0.0625	0.0625	0.0625	0.0625	0.25	0.0375	0.0014063	
P(y)	0.25	0.25	0.25	0.25		0.075	0.003125	0.055902
Mean of Y	0	0.0125	0.025	0.0375	0.075			
Var of Y	0.00140625	0.00015625	0.00015625	0.00140625	0.003125			
StDev of Y					0.055902			

Figura 5.8. Probabilidades marginales, medias y varianzas de X e Y.

Covarianza

La *covarianza* es una medida de la variabilidad conjunta de dos variables aleatorias. Puede utilizarse para calcular la varianza de combinaciones lineales de variables aleatorias, como la varianza del valor total de la combinación de acciones de dos empresas en una cartera. La covarianza también se utiliza para calcular una medida estandarizada de la variabilidad conjunta llamada correlación. Primero definimos la covarianza y, a continuación, presentamos algunas aplicaciones importantes. Supongamos que X e Y son un par de variables aleatorias que no son estadísticamente independientes. Nos gustaría tener alguna medida de la naturaleza y el grado de relación entre ellas. Eso es bastante difícil de lograr, ya que es razonable pensar que las variables aleatorias pueden estar relacionadas de diversas formas. Para simplificar el análisis, nos limitamos a analizar la posibilidad de que tengan una relación lineal. Por ejemplo, un elevado valor de X podría ir acompañado, en promedio, de un elevado valor de Y y un bajo valor de X de un bajo valor de Y; en ese caso, si se representaran en un gráfico los valores relacionados entre sí, la línea recta que se trazara pasando por ellos sería una buena aproximación.

Supongamos que la variable aleatoria X tiene una media μ_X y la variable aleatoria Y tiene una media μ_Y y consideremos el producto $(X - \mu_X)(Y - \mu_Y)$. Si los valores altos de X tienden a ir acompañados de valores altos de Y, es de esperar que este producto sea positivo, y cuanto mayor sea la relación, mayor será la esperanza de $(X - \mu_X)(Y - \mu_Y)$, definida de la forma siguiente: $E[(X - \mu_X)(Y - \mu_Y)]$. En cambio, si los valores altos de X van acompañados de valores bajos de Y y los valores bajos de X van acompañados de valores altos de Y, el valor esperado de este producto, $E[(X - \mu_X)(Y - \mu_Y)]$, sería negativo. Una esperanza $E[(X - \mu_X)(Y - \mu_Y)]$ igual a 0 implicaría la ausencia de una relación lineal entre X e Y. Por lo tanto, se utilizará el valor esperado, $E[(X - \mu_X)(Y - \mu_Y)]$, como medida de la relación lineal que existe en la población.

Covarianza

Sea X una variable aleatoria de media μ_X e Y una variable aleatoria de media μ_Y. El valor esperado de $(X - \mu_X)(Y - \mu_Y)$ se llama **covarianza** entre X e Y y se representa por medio de $\text{Cov}(X, Y)$. En el caso de las variables aleatorias discretas,

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)P(x, y) \qquad \textbf{(5.29)}$$

Una expresión equivalente es

$$\text{Cov}(X, Y) = E[XY] - \mu_X\mu_Y = \sum_x \sum_y xyP(x, y) - \mu_X\mu_Y$$

Correlación

Aunque la covarianza indica el sentido de la relación entre variables aleatorias, no tiene un límite superior o inferior y su magnitud depende extraordiariamente de las unidades en las que se mide. Existe una estrecha relación lineal cuando los puntos de observación están cerca de una línea recta. Es difícil utilizar la covarianza para medir el grado de relación lineal, ya que no tiene límites. Una medida relacionada con ésta, el coeficiente de correlación, es una medida del grado de relación lineal entre dos variables cuyo valor sólo puede estar entre -1 y 1.

Correlación

Sean X e Y variables aleatorias distribuidas conjuntamente. La **correlación** entre X e Y es

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} \qquad \textbf{(5.30)}$$

La correlación es la covarianza dividida por las desviaciones típicas de las dos variables aleatorias. El resultado es una medida estandarizada de la relación que puede ir de -1 a $+1$. Son importantes las siguientes interpretaciones:

1. Una correlación de 0 indica que no existe ninguna relación lineal entre las dos variables aleatorias. Si las dos variables aleatorias son independientes, la correlación es igual a 0.
2. Una correlación positiva indica que, si una de las variables aleatorias es alta (baja), la otra tiene una probabilidad mayor de ser alta (baja) y decimos que las variables son dependientes positivamente. La dependencia lineal positiva perfecta se indica por medio de una correlación de $+1{,}0$.
3. Una correlación negativa indica que, si una de las variables aleatorias es alta (baja), la otra tiene una probabilidad mayor de ser baja (alta) y decimos que las variables son dependientes negativamente. La dependencia lineal negativa perfecta se indica por medio de una correlación de $-1{,}0$.

La correlación es más útil que la covarianza para describir relaciones. Con una correlación de $+1$, las dos variables aleatorias tienen una relación lineal positiva perfecta, y, por lo tanto, un valor específico de una variable, X, predice la otra, Y, exactamente. Una correlación de -1 indica la existencia de una relación lineal negativa perfecta entre dos variables; una de las variables, X, predice la negativa de la otra, Y. Una correlación de 0 indica

que no existe ninguna relación lineal entre las dos variables. Los valores intermedios indican que las variables tienden a estar relacionadas; las relaciones son más estrechas cuando el valor absoluto de la correlación tiende a 1.

También sabemos que el término correlación se ha convertido en una palabra de uso común. En muchos casos, se utiliza para indicar que existe una relación. Sin embargo, las variables que tienen relaciones no lineales no tienen un coeficiente de correlación cercano a 1,0. Esta distinción es importante para nosotros con el fin de evitar la confusión entre las variables aleatorias correlacionadas y las variables aleatorias que tienen relaciones no lineales.

EJEMPLO 5.16. Distribución conjunta de los precios de las acciones (cálculo de la covarianza y de la correlación)

Halle la covarianza y la correlación de las acciones de las empresas A y B del ejemplo 5.15 con la distribución de probabilidad conjunta de la Tabla 5.6.

Solución

El cálculo de la covarianza es tedioso incluso en un problema como éste, que se ha simplificado de manera que todas las probabilidades conjuntas, $P(x, y)$, sean 0,0625 para todos los pares de valores x e y. Por definición, tenemos que hallar

$$\begin{aligned}\text{Cov}(X, Y) &= \sum_x \sum_y xyP(x, y) - \mu_X\mu_Y \\ &= 0[(0)(0{,}0625) + (0{,}05)(0{,}0625) + (0{,}10)(0{,}0625) + (0{,}15)(0{,}0625)] \\ &\quad + \cdots + (0{,}15)[(0)(0{,}0625) + (0{,}05)(0{,}0625) + (0{,}10)(0{,}0625) \\ &\quad + (0{,}15)(0{,}0625)] - (0{,}075)(0{,}075) \\ &= 0{,}005625 - 0{,}005625 \\ &= 0\end{aligned}$$

Por lo tanto,

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = 0$$

Podemos utilizar el programa Microsoft Excel para realizar estos cálculos siguiendo atentamente el ejemplo de la Figura 5.9.

		Y Return						
X Return	0%	5%	10%	15%	P(x)	Mean of	Var of X	StDev of
0%	0.0625	0.0625	0.0625	0.0625	0.25	0	0.0014063	
5%	0.0625	0.0625	0.0625	0.0625	0.25	0.0125	0.0001563	
10%	0.0625	0.0625	0.0625	0.0625	0.25	0.025	0.0001563	
15%	0.0625	0.0625	0.0625	0.0625	0.25	0.0375	0.0014063	
P(y)	0.25	0.25	0.25	0.25		0.075	0.003125	0.055902
Mean of Y	0	0.0125	0.025	0.0375	0.075			
Var of Y	0.00140625	0.00015625	0.00015625	0.00140625	0.003125			
StDev of Y					0.055902			
xyP(x)	0.0000000	0.0009375	0.0018750	0.0028125	0.0056250			
ΣΣ xyP(x)	0.0000000							

Figura 5.9. Cálculo de la covarianza y la correlación por medio del programa Microsoft Excel.

Covarianza e independencia estadística

Si dos variables aleatorias son **estadísticamente independientes**, la **covarianza** entre ellas es 0. Sin embargo, lo contrario no es necesariamente cierto.

La razón por la que el hecho de que una covarianza sea 0 no implica necesariamente que las variables aleatorias sean estadísticamente independientes se halla en que la covarianza pretende medir una relación lineal y es posible que esta cantidad no detecte otros tipos de dependencia. Supongamos que la variable aleatoria X tiene la función de probabilidad

$$P(-1) = 1/4 \qquad P(0) = 1/2 \qquad P(1) = 1/4$$

Definamos la variable aleatoria Y de la forma siguiente:

$$Y = X^2$$

Así pues, para saber cuál es el valor de X hay que saber cuál es el valor de Y y, por lo tanto, estas dos variables aleatorias no son, desde luego, independientes. Siempre que $X = 0$, entonces $Y = 0$, y si X es -1 o 1, entonces $Y = 1$. La función de probabilidad conjunta de X e Y es

$$P(-1, 1) = 1/4 \qquad P(0, 0) = 1/2 \qquad P(1, 1) = 1/4$$

y la probabilidad de cualquier otra combinación de valores es igual a 0. Es sencillo entonces verificar que

$$E(X) = 0 \qquad E(Y) = 1/2 \qquad E(XY) = 0$$

La covarianza entre X e Y es 0.

Funciones lineales de variables aleatorias

Antes hemos definido la esperanza de una función de una única variable aleatoria. Ahora podemos extender esta definición a las funciones de varias variables aleatorias.

Valor esperado de las funciones de variables aleatorias distribuidas conjuntamente

Sean X e Y un par de variables aleatorias discretas que tienen la función de probabilidad conjunta $P(x, y)$. La esperanza de cualquier función $g(X, Y)$ de estas variables aleatorias se define de la forma siguiente:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)P(x, y) \qquad \textbf{(5.31)}$$

Para terminar el análisis de las distribuciones conjuntas, consideremos la media y la varianza de una variable aleatoria que es la suma o la diferencia de otras variables aleatorias. Estos resultados se resumen a continuación y pueden obtenerse por medio de la ecuación 5.31.

Resumen de los resultados relativos a las sumas y las diferencias de variables aleatorias

Sean X e Y un par de variables aleatorias que tienen las medias μ_X y μ_Y y las varianzas σ_X^2 y σ_Y^2. Se cumplen las siguientes propiedades:

1. El **valor esperado de su suma** es la suma de sus valores esperados:

$$E(X + Y) = \mu_X + \mu_Y \tag{5.32}$$

2. El **valor esperado de su diferencia** es la diferencia entre sus valores esperados:

$$E(X - Y) = \mu_X - \mu_Y \tag{5.33}$$

3. Si la covarianza entre X e Y es 0, la **varianza de su suma** es la suma de sus varianzas:

$$\text{Var}(X + Y) = \sigma_X^2 + \sigma_Y^2 \tag{5.34}$$

pero si la covarianza no es 0, entonces

$$\text{Var}(X + Y) = \sigma_X^2 + \sigma_Y^2 + 2\,\text{Cov}(X, Y)$$

4. Si la covarianza entre X e Y es 0, la **varianza de su diferencia** es la *suma* de sus varianzas:

$$\text{Var}(X - Y) = \sigma_X^2 + \sigma_Y^2 \tag{5.35}$$

pero si la covarianza no es 0, entonces

$$\text{Var}(X - Y) = \sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}(X, Y)$$

Sean X_1, X_2, ..., X_K K variables aleatorias que tienen las medias μ_1, μ_2, ..., μ_K y las varianzas σ_1^2, σ_2^2, ..., σ. Se cumplen las siguientes propiedades:

5. El valor esperado de su suma es

$$E(X_1 + X_2 + \cdots + X_K) = \mu_1 + \mu_2 + \cdots + \mu_K \tag{5.36}$$

6. Si la covarianza entre cada par de estas variables aleatorias es 0, la varianza de su suma es

$$\text{Var}(X_1 + X_2 + \cdots + X_K) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_K^2 \tag{5.37}$$

EJEMPLO 5.17. Sencilla cartera de inversión (medias y varianzas, funciones de variables aleatorias)

Un inversor tiene 1.000 $ para invertir y dos oportunidades de inversión, cada una de las cuales requiere un mínimo de 500 $. Los beneficios por cada 100 $ de la primera pueden representarse por medio de una variable aleatoria X, que tiene la siguiente función de probabilidad:

$$P(X = -5) = 0{,}4 \quad \text{y} \quad P(X = 20) = 0{,}6$$

El beneficio por cada 100 $ de la segunda viene dado por la variable aleatoria Y, cuya función de probabilidad es

$$P(Y = 0) = 0{,}6 \quad \text{y} \quad P(Y = 25) = 0{,}4$$

Las variables aleatorias X e Y son independientes. El inversor tiene las siguientes estrategias posibles:

a) 1.000 $ en la primera inversión
b) 1.000 $ en la segunda inversión
c) 500 $ en cada inversión

Halle la media y la varianza de los beneficios generados por cada estrategia.

Solución

La variable aleatoria X tiene la media

$$\mu_X = E(X) = \sum_x xP(x) = (-5)(0{,}4) + (20)(0{,}6) = 10 \$$$

y la varianza

$$\sigma_X^2 = E[(X - \mu_x)^2 = \sum_x (x - \mu_x)^2 P(x)$$

$$= (-5 - 10)^2(0{,}4) + (20 - 10)^2(0{,}6) = 150$$

La estrategia (a) tiene un beneficio medio de $E(10X) = 10E(X) = 100$ $ y una varianza de

$$\text{Var}\,(10X) = 100\,\text{Var}\,(X) = 15.000$$

La variable aleatoria Y tiene la media

$$\mu_Y = E(Y) = \sum_Y yP(y) = (0)(0{,}6) + (25)(0{,}4) = 10 \$$$

y la varianza

$$\sigma_Y^2 = E[(Y - \mu_Y)^2] = \sum_Y (y - \mu_Y)^2 P(y)$$

$$= (0 - 10)^2(0{,}6) + (25 - 10)^2(0{,}4) = 150$$

La estrategia (b) tiene un beneficio medio de $E(10Y) = 10E(Y) = 100$ $ y una varianza de

$$\text{Var}\,(10Y) = 100\,\text{Var}\,(Y) = 15.000$$

Consideremos ahora la estrategia (c): 500 $ en cada inversión. El rendimiento de la estrategia (c) es $5X + 5Y$, que tiene una media de

$$E(5X + 5Y) = E(5X) + E(5Y) = 5E(X) + 5E(Y) = 100 \$$$

Por lo tanto, las tres estrategias tienen el mismo beneficio esperado. Sin embargo, como X e Y son independientes y la covarianza es 0, la varianza del rendimiento de la estrategia (c) es

$$\text{Var}(5X + 5Y) = \text{Var}(5X) + \text{Var}(5Y) = 25\,\text{Var}(X) + 25\,\text{Var}(Y) = 7.500$$

Esta varianza es menor que las varianzas de las demás estrategias, debido a la disminución que experimenta el riesgo como consecuencia de la diversificación de una cartera de inversión. Este inversor debería preferir, desde luego, la estrategia (c), ya que genera el mismo rendimiento esperado que las otras dos, pero con un riesgo menor.

Análisis de carteras

Los gestores de inversiones realizan considerables esfuerzos para crear carteras de inversión formadas por un conjunto de instrumentos financieros que tengan cada uno de ellos unos rendimientos definidos por un modelo de distribución de probabilidad. Las carteras se utilizan para conseguir una inversión combinada que tenga un rendimiento y un riesgo esperados dados. Se pueden construir carteras de acciones de alto riesgo combinando varias acciones cuyos valores tiendan a subir o a bajar a la vez. Con una cartera de ese tipo, un inversor experimentará grandes ganancias o grandes pérdidas. Se pueden combinar acciones cuyos valores varíen en sentido contrario para crear una cartera que tenga un valor más estable, lo que implica menos riesgo. Los descensos del precio de una de las acciones son compensados por las subidas del precio de otra.

Este proceso de construcción y análisis de carteras se realiza utilizando modelos de probabilidad definidos mediante variables aleatorias y funciones de distribución de probabilidad. El valor medio de la cartera es la combinación lineal de los valores medios de sus dos acciones. La varianza del valor de la cartera se calcula utilizando la suma de las varianzas y la covarianza de la distribución conjunta de los valores de las acciones. Desarrollaremos el método poniendo un ejemplo de una cartera formada por acciones de dos empresas.

Consideremos una cartera formada por a acciones de la empresa A y b acciones de la empresa B. Es importante poder hallar la media y la varianza del valor de mercado, W, de una cartera, donde W es la función lineal $W = aX + bY$. La media y la varianza se obtienen en el apéndice del capítulo.

La media y la varianza del valor de mercado de una cartera

La variable aleatoria X es el precio de las acciones de A y la variable aleatoria Y es el precio de las acciones de B. El **valor de mercado de la cartera**, W, viene dado por la función lineal

$$W = aX + bY$$

donde a es el número de acciones de la empresa A y b es el número de acciones de la empresa B.

El **valor medio de W** es

$$\begin{aligned}\mu_W &= E[W] = E[aX + bY] \\ &= a\mu_X + b\mu_Y \end{aligned} \qquad \textbf{(5.38)}$$

La **varianza de *W*** es

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\mathrm{Cov}\,(X,\,Y) \tag{5.39}$$

o utilizando la correlación

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\mathrm{Corr}\,(X,\,Y)\sigma_X\sigma_Y$$

EJEMPLO 5.18. Análisis de carteras de acciones (medias y varianzas, funciones de variables aleatorias)

Jorge Téllez tiene 5 acciones de la empresa A y 10 de la empresa B; las variaciones de sus precios siguen el modelo de distribución de probabilidad de la Tabla 5.7. Halle la media y la varianza de la cartera.

Tabla 5.7. Precios de las acciones de A y B.

	Precio de las acciones de B			
Precio de las acciones de A	**40 $**	**50 $**	**60 $**	**70 $**
45 $	0,24	0,003333	0,003333	0,003333
50 $	0,003333	0,24	0,003333	0,003333
55 $	0,003333	0,003333	0,24	0,003333
60 $	0,003333	0,003333	0,003333	0,24

Solución

El valor, W, de la cartera puede representarse por medio de la combinación lineal $W = 5X + 10Y$. La media y la varianza de la acción de A son 53 $ y 31,3, respectivamente, mientras que las de la acción de B son 55 $ y 125, respectivamente. La covarianza es 59,17 y la correlación es 0,947. Estos resultados se han obtenido con el programa Microsoft Excel haciendo cálculos similares a los de la Figura 5.9.

El valor medio de la cartera es, pues,

$$\mu_W = E[W] = E[5X + 10Y] = 5(53) + (10)(55) = 815$$

La varianza del valor de la cartera es

$$\begin{aligned}\sigma_W^2 &= 5^2\sigma_X^2 + 10^2\sigma_Y^2 + 2 \times 5 \times 10 \times \mathrm{Cov}\,(X,\,Y)\\ &= 5^2 \times 31{,}3 + 10^2 \times 125 + 2 \times 5 \times 10 \times 59{,}17 = 19.199{,}5\end{aligned}$$

Jorge sabe que una elevada varianza implica un elevado riesgo. Cree que el riesgo de esta cartera es demasiado alto, por lo que nos pide que le preparemos una cartera que tenga menos riesgo. Tras algunas investigaciones, descubrimos un par distinto de acciones cuyos precios siguen el modelo de distribución de probabilidad de la Tabla 5.8.

La media de las acciones de la empresa C es de 53 $, igual que la de las acciones de la empresa A. Asimismo, la media de las acciones de la empresa D es de 55 $, igual que la de las acciones de la empresa B. Por lo tanto, el valor medio de la cartera no varía.

Tabla 5.8. Nueva cartera de acciones de C y D.

	Precio de las acciones de D			
Precio de las acciones de C	**40 $**	**50 $**	**60 $**	**70 $**
45 $	0,003333	0,003333	0,003333	0,24
50 $	0,003333	0,003333	0,24	0,003333
55 $	0,003333	0,24	0,003333	0,003333
60 $	0,24	0,003333	0,003333	0,003333

La varianza de las acciones de cada empresa también es la misma, pero ahora la covarianza es $-59{,}17$. Por lo tanto, la varianza de la nueva cartera incluye un término de *covarianza negativa* y es

$$\begin{aligned}\sigma_W^2 &= 5^2\sigma_X^2 + 10^2\sigma_Y^2 + 2 \times 5 \times 10 \times \text{Cov}(X, Y)\\ &= 5^2 \times 31{,}3 + 10^2 \times 125 + 2 \times 5 \times 10 \times (-59{,}17) = 7.365{,}5\end{aligned}$$

Vemos que la covarianza negativa reduce la varianza y, por lo tanto, el riesgo de la cartera.

La Figura 5.10 muestra cómo varía la varianza de la cartera y, por lo tanto, el riesgo con diferentes correlaciones entre los precios de las acciones. Obsérvese que la varianza de la cartera está relacionada linealmente con la correlación. Para ayudar a controlar el riesgo, los creadores de carteras de acciones seleccionan las acciones basándose en la correlación entre los precios.

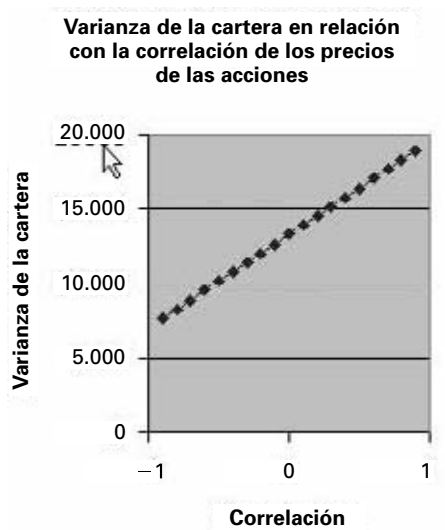


Figura 1.1. Varianza de la cartera en relación con la correlación de los precios de las acciones.

Como hemos visto en el ejemplo 5.18, la correlación entre los precios de las acciones o entre dos variables aleatorias cualesquiera produce importantes efectos en la variable aleatoria del valor de la cartera. La existencia de una correlación positiva indica que los dos precios, X e Y, suben o bajan a la vez. Por lo tanto, se magnifican los valores grandes o

pequeños de la cartera, por lo que el rango y la varianza son mayores que cuando la correlación es cero. Y, a la inversa, la existencia de una correlación negativa indica que, cuando sube el precio de X, baja el precio de Y. Como consecuencia, el rango y la varianza de la cartera son menores que cuando la correlación es cero. Seleccionando acciones que tienen determinadas combinaciones de correlaciones, los gestores de fondos pueden controlar la varianza y el riesgo de las carteras.

EJERCICIOS

Ejercicios básicos

5.73. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,25	0,25
	1	0,25	0,25

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.

5.74. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,20	0,25
	1	0,30	0,25

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.

5.75. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,25	0,25
	1	0,25	0,25

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = X + Y$.

5.76. Considere la distribución de probabilidad conjunta

		X	
		0	**1**
Y	**0**	0,30	0,20
	1	0,25	0,25

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = 2X + Y$.

5.77. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,70	0,0
	1	0,0	0,30

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = 3X + 4Y$.

5.78. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,25	0,25
	1	0,25	0,25

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = X + Y$.

5.79. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,30	0,20
	1	0,25	0,25

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = 2X + Y$.

5.80. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,0	0,60
	1	0,40	0,0

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = 2X - 4Y$.

5.81. Considere la distribución de probabilidad conjunta

		X	
		1	**2**
Y	**0**	0,70	0,0
	1	0,0	0,30

a) Calcule las distribuciones de probabilidad marginal de X e Y.
b) Calcule la covarianza y la correlación de X e Y.
c) Calcule la media y la varianza de la función lineal $W = 10X - 8Y$.

Ejercicios aplicados

5.82. Un investigador sospechaba que el número de tentempiés que tomaban en un día los estudiantes durante la época de exámenes finales dependía del número de exámenes que tenían que realizar ese día. La tabla adjunta muestra las probabilidades conjuntas, estimadas a partir de una encuesta.

	Número de exámenes (*X*)			
Número de tentempiés (*Y*)	**0**	**1**	**2**	**3**
0	0,07	0,09	0,06	0,01
1	0,07	0,06	0,07	0,01
2	0,06	0,07	0,14	0,03
3	0,02	0,04	0,16	0,04

a) Halle la función de probabilidad de X y, por lo tanto, el número medio de exámenes realizados por los estudiantes ese día.
b) Halle la función de probabilidad de Y y, por lo tanto, el número medio de exámenes realizados por los estudiantes ese día.
c) Halle e interprete la función de probabilidad condicionada de Y, dado $X = 3$.
d) Halle la covarianza entre X e Y.
e) ¿Son el número de tentempiés y el número de exámenes independientes entre sí?

5.83. Una agencia inmobiliaria tiene interés en saber cuál es la relación entre el número de líneas de un anuncio de prensa sobre un apartamento y el volumen de llamadas de interesados. Representemos el volumen de llamadas por medio de la variable aleatoria X, cuyo valor es 0 cuando el interés por el anuncio es escaso, 1 cuando es moderado y 2 cuando es grande. La agencia estimó la función de probabilidad conjunta mostrada en la tabla adjunta.

	Número de llamadas (*X*)		
Número de líneas (*Y*)	**0**	**1**	**2**
3	0,09	0,14	0,07
4	0,07	0,23	0,16
5	0,03	0,10	0,11

a) Halle la función de probabilidad acumulada conjunta en $X = 1$, $Y = 4$ e interprete su resultado.
b) Halle e interprete la función de probabilidad condicionada de Y, dado $X = 0$.
c) Halle e interprete la función de probabilidad condicionada de X, dado $Y = 5$.
d) Halle e interprete la covarianza entre X e Y.
e) ¿Son el número de líneas del anuncio y el volumen de llamadas independientes entre sí?

5.84. La tabla adjunta muestra las probabilidades conjuntas del número de tarjetas de crédito que poseen las personas que tienen entre una y tres tarjetas de crédito (X) y el número de compras semanales realizadas con tarjeta de crédito (Y).

Número de tarjetas de crétido (Y)	Número de compras semanales (X)				
	0	**1**	**2**	**3**	**4**
1	0,08	0,13	0,09	0,06	0,03
2	0,03	0,08	0,08	0,09	0,07
3	0,01	0,03	0,06	0,08	0,08

a) ¿Cuál es la función de probabilidad del número de compras semanales de una persona de este grupo elegida aleatoriamente?
b) ¿Cuál es la función de probabilidad del número de compras semanales de una persona de este grupo que tenga tres tarjetas?
c) ¿Son el número de tarjetas que posee una persona y el número de compras estadísticamente independientes?

5.85. Una empresa de estudios de mercado quiere saber si un nuevo modelo de computador personal que se anunciaba en un programa que se emitía de madrugada ha conseguido que sea una marca más conocida para las personas que veían el programa habitualmente que para las que no lo veían. Tras realizar una encuesta, observó que el 15 por ciento de todas las personas veía el programa habitualmente y podía identificar correctamente el producto. Además, el 16 por ciento de todas las personas veía habitualmente el programa y el 45 por ciento de todas las personas podía identificar correctamente el producto. Defina un par de variables aleatorias de la forma siguiente:

$X = 1$	si se ve habitualmente el programa	$X = 0$	en caso contrario
$Y = 1$	si se identifica correctamente el producto	$Y = 0$	en caso contrario

a) Halle la función de probabilidad conjunta de X e Y.
b) Halle la función de probabilidad condicionada de Y, dado $X = 1$.
c) Halle e interprete la covarianza entre X e Y.

5.86. Un vendedor de libros de texto universitarios llama a los despachos de los profesores y tiene la impresión de que los profesores tienden más a no estar en su despacho los viernes que los demás días laborales. Un repaso de las llamadas, de las cuales un quinto se realiza los viernes, indica que en el 16 por ciento de las llamadas realizadas los viernes, el profesor no está en su despacho, mientras que eso ocurre únicamente en el caso del 12 por ciento de las llamadas realizadas los demás días laborales. Defina las variables aleatorias de la forma siguiente:

$X = 1$	si la llamada se realiza los viernes	$X = 0$	en caso contrario
$Y = 1$	si el profesor no está en el despacho	$Y = 0$	en caso contrario

a) Halle la función de probabilidad conjunta de X e Y.
b) Halle la función de probabilidad condicionada de Y, dado $X = 0$.
c) Halle las funciones de probabilidad marginal de X e Y.
d) Halle e interprete la covarianza entre X e Y.

5.87. El director de un restaurante recibe quejas de vez en cuando sobre la calidad tanto de la comida como del servicio. La tabla adjunta muestra las funciones de probabilidad marginal del número de quejas semanales de cada categoría. Halle la función de probabilidad conjunta suponiendo que las quejas sobre la comida y el servicio son independientes entre sí.

Número de quejas sobre la comida	Probabilidad	Número de quejas sobre el servicio	Probabilidad
0	0,12	0	0,18
1	0,29	1	0,38
2	0,42	2	0,34
3	0,17	3	0,10

5.88. Vuelva a la información del ejercicio 5.87. Halle la media y la desviación típica del número total de quejas recibidas en una semana. Llegado a este punto, sospecha que el número de quejas sobre la comida y sobre el servicio no son independientes entre sí. Sin embargo, no tiene ninguna información sobre la naturaleza de su dependencia. ¿Qué puede decir ahora sobre la media y la desviación típica del número total de quejas recibidas en una semana?

5.89. Una empresa tiene 5 representantes que cubren grandes territorios y 10 que cubren territorios más pequeños. La tabla adjunta muestra las distribuciones de probabilidad del número de pedidos recibidos por cada uno de estos tipos de representantes en un día. Suponiendo que el número de pedidos que recibe cualquier representante es independiente del número que recibe cualquier otro, halle la media y la desviación típica del número total de pedidos recibidos por la empresa en un día.

Número de pedidos (territorio grande)	Probabilidad	Número de pedidos (territorio más pequeño)	Probabilidad
0	0,08	0	0,18
1	0,16	1	0,26
2	0,28	2	0,36
3	0,32	3	0,13
4	0,10	4	0,07
5	0,06		

RESUMEN

En este capítulo hemos presentado modelos de probabilidad discreta. Estos modelos se definen por medio de una variable aleatoria y una función de distribución de probabilidad. También hemos definido los valores esperados y las varianzas de estos modelos. Hemos presentado tres importantes modelos de probabilidad discreta —el binomial, el de Poisson y el hipergeométrico— junto con posibles aplicaciones. Por último, hemos desarrollado distribuciones de probabilidad discreta conjunta y hemos indicado cómo se calcula la covarianza de estos modelos. Hemos mostrado cómo pueden utilizarse para hallar la media y la varianza de combinaciones lineales de variables aleatorias, con una aplicación especial a las carteras de acciones.

TÉRMINOS CLAVE

análisis de carteras, 189
aproximación de Poisson de la distribución binomial, 176
correlación, 184
covarianza, 184
derivación de la función de probabilidad marginal, 180
desviación típica de una variable aleatoria discreta, 153
diferencias entre las variables aleatorias, 187
distribución binomial, 163
distribución hipergeométrica, 170
distribución de probabilidad de Poisson, 174
función de distribución de probabilidad, 148
función de probabilidad acumulada, 149
función de probabilidad condicionada, 180
función de probabilidad conjunta, 180
función de probabilidad marginal, 180
independencia de las variables aleatorias distribuidas conjuntamente, 181
media, 152
media de una distribución binomial, 163
media de funciones de variables aleatorias, 156
media y la varianza de una variable aleatoria de Bernoulli, 161
propiedades de las funciones de distribución de probabilidad, 149
propiedades de las funciones de probabilidad acumulada, 150
propiedades de las funciones de probabilidad conjunta, 180
relación entre la función de probabilidad y la función de probabilidad acumulada, 150
sumas de variables aleatorias, 187
valor esperado, 152
valor esperado de funciones de variables aleatorias, 156
valor esperado de funciones de variables aleatorias distribuidas conjuntamente, 186
valor de mercado de una cartera, 189
variable aleatoria, 146
variable aleatoria continua, 146
variable aleatoria discreta, 146
varianza de una distribución binomial, 163
varianza de funciones de variables aleatorias, 156
varianza de una variable aleatoria discreta, 153
varianza de una variable aleatoria discreta (fórmula alternativa), 153

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

5.90. Un asesor financiero le dice a un cliente que una inversión en un fondo tendrá (el próximo año) un rendimiento esperado más alto que una inversión en el mercado de dinero. El cliente le hace entonces las siguientes preguntas:

a) ¿Significa eso que el fondo de inversión tendrá con seguridad un rendimiento mayor que una inversión en el mercado de dinero?

b) ¿Quiere decir que debo invertir en el fondo de inversión y no en el mercado de dinero?

¿Qué le respondería?

5.91. Un contratista estima las probabilidades del número de días necesarios para terminar un cierto tipo de proyecto de construcción:

Tiempo (días)	1	2	3	4	5
Probabilidad	0,05	0,20	0,35	0,30	0,10

a) ¿Cuál es la probabilidad de que se tarde menos de 3 días en terminar un proyecto elegido aleatoriamente?

b) Halle el tiempo esperado de terminar un proyecto.

c) Halle la desviación típica del tiempo necesario para terminar un proyecto.

d) El coste del proyecto del contratista consta de dos partes: un coste fijo de 20.000 $ más 2.000 $ por cada día necesario para realizar el proyecto. Halle la media y la desviación típica del coste total del proyecto.

e) Si se realizan tres proyectos, ¿cuál es la probabilidad de que se tarde al menos 4 días en terminar al menos dos de ellos, suponiendo que los días que se tarda en terminar un proyecto y los que se tarda en terminar otro son independientes?

5.92. Un vendedor de automóviles estima que las probabilidades de vender un número de automóviles la próxima semana son:

Número de automóviles	0	1	2	3	4	5
Probabilidad	0,10	0,20	0,35	0,16	0,12	0,07

a) Halle el número esperado de automóviles que venderá en la semana.
b) Halle la desviación típica del número de automóviles que venderá en la semana.
c) El vendedor gana 250 $ a la semana más 300 $ más por cada automóvil que venda. Halle la media y la desviación típica de su sueldo semanal total.
d) ¿Cuál es la probabilidad de que el sueldo semanal del vendedor sea de más de 1.000 $?

5.93. Un examen de tipo test consta de nueve preguntas. En cada pregunta hay que elegir entre cuatro respuestas posibles. El alumno recibe un punto por cada respuesta correcta y no se restan puntos por las respuestas incorrectas. El profesor da un punto más si el estudiante deletrea su nombre correctamente. Un alumno que no ha estudiado para este examen decide elegir aleatoriamente una respuesta en cada pregunta.

a) Halle el número esperado de respuestas correctas del estudiante a estas nueve preguntas.
b) Halle la desviación típica del número de respuestas correctas del estudiante a estas nueve preguntas.
c) El estudiante deletrea su nombre correctamente:
  i. Halle la puntuación total esperada de este estudiante en el examen.
  ii. Halle la desviación típica de su puntuación total en el examen.

5.94. Ponga ejemplos realistas de pares de variables aleatorias en las que es de esperar que

a) La covarianza sea positiva.
b) La covarianza sea negativa.
c) La covarianza sea cero.

5.95. Una empresa de taxis de larga distancia posee cuatro vehículos. Éstos son de diferente antigüedad y tienen diferentes historiales de reparaciones. Las probabilidades de que en un día cualquiera cada uno esté listo para su uso son 0,95, 0,90, 0,90 y 0,80. El hecho de que un vehículo esté listo o no es independiente de que lo esté otro.

a) Halle la función de probabilidad del número de vehículos listos en un día dado.
b) Halle el número esperado de vehículos listos en un día dado.
c) Halle la desviación típica del número de vehículos listos en un día dado.

5.96. Los estudiantes de una universidad se clasificaron según el número de años que llevaban en la universidad (X) y el número de visitas que habían realizado a un museo el año anterior ($Y = 0$ en el caso en que no hubieran realizado ninguna visita, 1 en el caso en que hubieran realizado una y 2 en el caso en que hubieran realizado más de una). Se estimaron las probabilidades conjuntas de estas variables aleatorias que se muestran en la tabla adjunta.

Número de visitas (Y)	**Años en la universidad (X)**			
	1	**2**	**3**	**4**
0	0,07	0,05	0,03	0,02
1	0,13	0,11	0,17	0,15
2	0,04	0,04	0,09	0,10

a) Halle la probabilidad de que un estudiante elegido aleatoriamente no visitara un museo el año anterior.
b) Halle la media de las variables aleatorias X e Y.
c) Halle e interprete la covarianza entre las variables aleatorias X e Y.

5.97. La estrella de un equipo de baloncesto especialista en lanzamientos de 3 puntos realiza seis lanzamientos de 3 puntos en un partido. Históricamente, ha encestado el 40 por ciento de los lanzamientos de 3 puntos realizados en un partido. Indique al principio qué supuestos ha postulado.

a) Halle la probabilidad de que enceste al menos dos de sus lanzamientos.
b) Halle la probabilidad de que enceste exactamente tres de sus lanzamientos.
c) Halle la media y la desviación típica del número de encestes realizados.
d) Halle la media y la desviación típica del número total de puntos conseguidos como consecuencia de estos encestes.

5.98. Se estima que el 55 por ciento de los estudiantes que entran en una universidad se licencia a los cuatro años.

a) ¿Cuál es la probabilidad de que tres exactamente de una muestra de cinco se licencie a los cuatro años?

b) ¿Cuál es la probabilidad de que la mayoría de una muestra de cinco se licencie a los cuatro años?

c) Se eligen aleatoriamente 80 estudiantes que entran en la universidad. Halle la media y la desviación típica de la proporción de estos 80 que se licenciará a los cuatro años.

5.99. En un campeonato de baloncesto participan dos equipos, el A y el B. El primero que gane cuatro partidos, gana el campeonato. Suponga que el equipo A es el mejor, en el sentido de que tiene una probabilidad de 0,6 de ganar cualquier partido. Suponga también que el resultado de cualquier partido es independiente del de cualquier otro.

a) ¿Cuál es la probabilidad de que gane el campeonato el equipo A?

b) ¿Cuál es la probabilidad de que sea necesario un séptimo partido para decidir el ganador?

c) Suponga que de hecho cada equipo gana dos de los cuatro primeros partidos.

i. ¿Cuál es la probabilidad de que gane el campeonato el equipo A?

ii. ¿Cuál es la probabilidad de que sea necesario un séptimo partido para decidir el ganador?

5.100. Basándose en información detallada sobre el flujo de caja, un analista financiero sostiene que es capaz de decir qué compañías son candidatas probables a la quiebra. Recibe información de 15 empresas y le dicen que 5 quebraron. Selecciona cinco del grupo de 15 como candidatas a la quiebra. Tres de las cinco seleccionadas por el analista estaban de hecho entre las que quebraron. Evalúe los resultados de este test sobre la capacidad del analista para detectar las empresas que son candidatas probables a la quiebra.

5.101. Un equipo de cinco analistas está a punto de examinar las perspectivas de beneficios de 20 empresas. Cada uno estudiará 4 empresas. Estos analistas no son igual de competentes. De hecho, uno de ellos es una estrella y tiene un excelente historial de previsión de los cambios de tendencia. A la dirección le gustaría asignar a este analista las 4 empresas cuyos beneficios se alejarán más de las tendencias pasadas. Sin embargo, al carecer de esta información, reparte las empresas aleatoriamente entre los analistas. ¿Cuál es la probabilidad de que asigne al mejor analista al menos 2 de las 4 empresas cuyos beneficios se alejarán más de las tendencias pasadas?

5.102. Durante la hora punta, llegan, en promedio, al mostrador de facturación de una compañía aérea 2,4 clientes por minuto. Suponga que las llegadas siguen una distribución de Poisson.

a) ¿Cuál es la probabilidad de que no llegue nadie en un minuto?

b) ¿Cuál es la probabilidad de que lleguen más de tres clientes en un minuto?

5.103. Según una estimación reciente, el 6,5 por ciento de todas las personas y parejas que declaran una renta de más de 200.000 $ no pagó impuestos o pagó un tipo impositivo de menos del 15 por ciento. Se tomó una muestra aleatoria de 100 personas del grupo que declaró una renta de más de 200.000 $. ¿Cuál es la probabilidad de que más de 2 miembros de la muestra no pagaran ningún impuesto o pagaran un tipo impositivo de menos del 15 por ciento?

5.104. Una empresa tiene dos cadenas de montaje, cada una de las cuales se para una media de 2,4 veces a la semana según una distribución de Poisson. Suponga que el comportamiento de una de estas cadenas de montaje es independiente del de la otra. ¿Cuál es la probabilidad de que al menos una se pare al menos una vez en cualquier semana dada?

5.105. Jorge Alas le ha pedido que analice su cartera de acciones, que contiene 10 acciones de la empresa D y 5 de la empresa C. La Tabla 5.9 muestra la distribución de probabilidad conjunta

Tabla 5.9. Distribución de probabilidad conjunta de los precios de las acciones.

Precio de las acciones de C	Precio de las acciones de D			
	40 $	**50 $**	**60 $**	**70 $**
45 $	0,00	0,00	0,05	0,20
50 $	0,05	0,00	0,05	0,10
55 $	0,10	0,05	0,00	0,05
60 $	0,20	0,10	0,05	0,00

de los precios de las acciones. Calcule la media y la varianza del valor total de su cartera de acciones.

5.106. Considere un país que importa acero y exporta automóviles. El valor por unidad de automóviles exportados se expresa en unidades de miles de dólares por automóvil por medio de la variable aleatoria X. El valor por unidad de acero importado se expresa en unidades de miles de dólares por tonelada de acero por medio de la variable aleatoria Y. Suponga que el país exporta anualmente 10 automóviles y 5 toneladas de acero. Calcule la media y la varianza de la balanza comercial, donde la balanza comercial es el total de dólares recibidos por todos los automóviles exportados menos el total de dólares gastados en todo el acero importado. La Tabla 5.10 muestra la distribución de probabilidad conjunta de los precios de los automóviles y del acero.

Tabla 5.10. Distribución conjunta de los precios de los automóviles y del acero.

Precio del acero (Y)	**Precio de los automóviles (X)**		
	3 $	**4 $**	**5 $**
4 $	0,10	0,15	0,05
6 $	0,10	0,20	0,10
8 $	0,05	0,15	0,10

Apéndice: Verificaciones

1. Verificación de una fórmula alternativa de la varianza de una variable aleatoria discreta (ecuación 5.6)

Comenzamos con la definición original de varianza:

$$\sigma_X^2 = \sum_x (x - \mu_X)^2 P(x) = \sum_x (x^2 - 2\mu_X x + \mu_X^2) P(x)$$

$$= \sum_x x^2 P(x) - 2\mu_X \sum_x xP(x) + \mu_X^2 \sum_x P(x)$$

Pero hemos visto que

$$\sum_x xP(x) = \mu_X \quad \text{y} \quad \sum_x P(x) = 1$$

Por lo tanto,

$$\sigma_X^2 = \sum_x x^2 P(x) - 2\mu_X^2 + \mu_X^2$$

y, por último,

$$\sigma_X^2 = \sum_x x^2 P(x) - \mu_X^2$$

2. Verificación de la media y la varianza de una función lineal de una variable aleatoria (ecuaciones 5.8 y 5.9)

De la definición de esperanza se deduce que si Y toma los valores $a + bx$ con las probabilidades $P_X(x)$, su media es

$$E(Y) = \mu_Y = \sum_x (a + bx)P(x)$$
$$= a \sum_x P(x) + b \sum_x xP(x)$$

Entonces, dado que el primer sumatorio del segundo miembro de esta ecuación es 1 y que el segundo es la media de X, tenemos que

$$E(Y) = a + b\mu_X \qquad \text{como en la ecuación 5.8}$$

Además, la varianza de Y es, por definición,

$$\sigma_Y^2 = E[(Y - \mu_Y)^2] = \sum_x [(a + bx) - \mu_Y]^2 P(x)$$

Sustituyendo μ_Y por $a + b\mu_X$, tenemos que

$$\sigma_Y^2 = \sum_X (bx - b\mu_X)^2 P(x) = b^2 \sum_X (x - \mu_X)^2 P(x)$$

Dado que el sumatorio del segundo miembro de esta ecuación es, por definición, la varianza de X, es fácil deducir el resultado de la ecuación 5.9:

$$\sigma_Y^2 = \text{Var}\,(a + bX) = b^2\sigma_X^2$$

3. Verificación de la media y la varianza de la distribución binomial (ecuaciones 5.19 y 5.20)

Para hallar la media y la varianza de la distribución binomial, es útil volver a la distribución de Bernoulli. Consideremos n pruebas independientes, cada una de las cuales tiene una probabilidad de éxito P, y sea $X_i = 1$ si la i-ésima prueba tiene éxito y 0 en caso contrario. Las variables aleatorias X_1, X_2, ..., X_n son, por lo tanto, n variables de Bernoulli independientes, cada una de las cuales tiene una probabilidad de éxito P. Además, el número total de éxitos X es

$$X = X_1 + X_2 + \cdots + X_n$$

Por lo tanto, la variable aleatoria binomial es la suma de variables aleatorias de Bernoulli independientes.

La media y la varianza de variables aleatorias de Bernoulli pueden utilizarse para hallar la media y la varianza de la distribución binomial. Aplicando la ecuación 5.15, sabemos que

$$E(X_i) = P \quad \text{y} \quad \sigma_{xi}^2 = P(1 - P) \quad \text{para todo } i = 1, 2, ..., n$$

Entonces, en el caso de la distribución binomial,

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = nP$$

Dado que las variables aleatorias de Bernoulli son independientes, la covarianza entre cualquier par de ellas es cero y

$$\begin{aligned}\sigma_X^2 &= \sigma^2(X_1 + X_2 + \cdots + X_n)\\ &= \sigma^2(X_1) + \sigma^2(X_2) + \cdots + \sigma^2(X_n)\\ &= nP(1 - P)\end{aligned}$$

4. Verificación de la media y la varianza del valor de mercado, *W*, de una cartera (ecuaciones 5.38 y 5.39)

Recibimos una combinación lineal, W, de las variables aleatorias X e Y, donde $W = aX + bY$ y a y b son constantes. La media de W es

$$\begin{aligned}\mu_W &= E[W] = E[aX + bY]\\ &= a\mu_X + b\mu_Y\end{aligned}$$

y la varianza de W es

$$\begin{aligned}\sigma_W^2 &= E[(W - \mu_W)^2]\\ &= E[((aX + bY) - (a\mu_X + b\mu_Y))^2]\\ &= E[(a(X - \mu_X) + b(Y - \mu_Y))^2]\\ &= E[a^2(X - \mu_X)^2 + b^2(Y - \mu_Y)^2 + 2ab(X - \mu_X)(Y - \mu_Y)]\\ &= a^2E[(X - \mu_X)^2] + b^2E[(Y - \mu_Y)^2] + 2abE[(X - \mu_X)(Y - \mu_Y)]\\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\text{Cov}\,(X, Y)\end{aligned}$$

Capítulo 6

Variables aleatorias continuas y distribuciones de probabilidad

Esquema del capítulo

Introducción

En el Capítulo 5, presentamos las variables aleatorias discretas y sus distribuciones de probabilidad. Aquí extendemos los conceptos de probabilidad a las variables aleatorias continuas y a sus distribuciones de probabilidad. Los conceptos y las ideas sobre las variables aleatorias discretas también se aplican a las variables aleatorias continuas, por lo que nos basamos directamente en el capítulo anterior. Muchos indicadores económicos y empresariales como las ventas, la inversión, el consumo, los costes y los ingresos pueden representarse por medio de variables aleatorias continuas. Además, las medidas del tiempo, la distancia, la temperatura y el peso encajan en esta categoría. Las afirmaciones sobre la probabilidad de variables aleatorias continuas se especifican en intervalos. Un ejemplo representativo es la probabilidad de que las ventas se encuentren entre 140 y 190 o sean superiores a 200. La teoría matemática nos lleva a concluir que, en realidad, las variables aleatorias de todos los problemas aplicados son discretas, porque las mediciones se redondean a algún valor. Pero para nosotros lo importante es que las variables aleatorias continuas y sus distribuciones de probabilidad son buenas aproximaciones en muchos problemas aplicados. Por lo tanto, estos modelos son muy importantes y constituyen excelentes instrumentos para las aplicaciones empresariales y económicas.

6.1. Variables aleatorias continuas

Aquí, X es de nuevo una variable aleatoria y x es un valor específico de la variable aleatoria. Comenzamos definiendo la *función de distribución acumulada*. A continuación, definimos la función de densidad de probabilidad, que es análoga a la función de distribución de probabilidad utilizada para las variables aleatorias discretas.

Función de distribución acumulada

La **función de distribución acumulada**, $F(x)$, de una variable aleatoria continua X expresa la probabilidad de que X no sea mayor que el valor de x, en función de x

$$F(x) = P(X \leqslant x) \tag{6.1}$$

Explicamos la función de distribución acumulada utilizando una sencilla estructura de probabilidad. Consideremos una estación de servicio que tiene un depósito de 1.000 litros que se llena todas las mañanas al comienzo de la jornada laboral. El análisis de la historia pasada indica que no es posible predecir la cantidad de gasolina que se venderá en un día cualquiera, pero el límite inferior es 0 y el superior es, por supuesto, 1.000 litros, que es el tamaño del depósito. Además, la historia pasada indica que cualquier demanda comprendida en el intervalo 1 a 1.000 litros es igual de probable. La variable aleatoria X indica las ventas de gasolina de un día específico en litros. Nos interesa saber cuál es la probabilidad de algunos niveles de ventas diarias de gasolina, donde la probabilidad de que se venda un número específico de litros es la misma en el intervalo de 0 a 1.000 litros. Se dice que la distribución de X sigue una **distribución de probabilidad uniforme** y la distribución acumulada es

$$F(x) = \begin{cases} 0 & \text{si} \cdots x < 0 \\ 0{,}001x & \text{si} \cdots 0 \leqslant x \leqslant 1.000 \\ 1 & \text{si} \cdots x > 1.000 \end{cases}$$

Esta función se representa por medio de una línea recta entre 0 y 1.000, como se muestra en la Figura 6.1. Permite ver que la probabilidad de que se venda entre 0 y 400 litros es

$$P(X \leqslant 400) = F(400) = (0{,}001)(400) = 0{,}40$$

Para hallar la probabilidad de que una variable aleatoria continua X esté comprendida en un intervalo específico, calculamos la diferencia entre la probabilidad acumulada en el extremo superior del intervalo y la probabilidad acumulada en el extremo inferior del intervalo.

Probabilidad de un intervalo utilizando una función de distribución acumulada

Sea X una variable aleatoria continua que tiene una función de distribución acumulada $F(x)$ y sean a y b dos valores posibles de X, siendo $a < b$. La **probabilidad de que X se encuentre entre a y b** es

$$P(a < X < b) = F(b) - F(a) \tag{6.2}$$

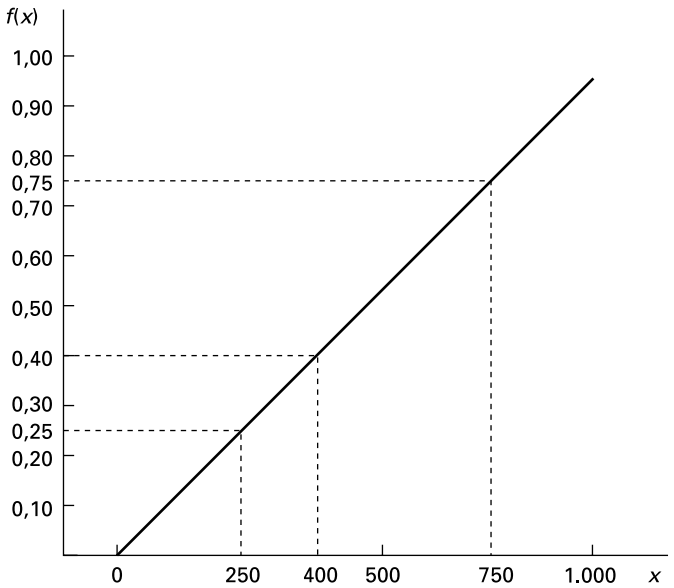


Figura 6.1. Función de distribución acumulada de una variable aleatoria que toma valores entre 0 y 1.000 con distribución de probabilidad uniforme.

En el caso de las variables aleatorias continuas, da lo mismo que escribamos «menor que» o «menor o igual que», ya que la probabilidad de que X sea exactamente igual a b es 0.

En el caso de la variable aleatoria que está distribuida uniformemente en el rango de 0 a 1.000, la función de distribución acumulada en ese rango es $F(x) = 0{,}001x$. Por lo tanto, si a y b son dos números comprendidos entre 0 y 1.000, siendo $a < b$,

$$P(a < X < b) = F(b) - F(a) = 0{,}001(b - a)$$

Por ejemplo, la probabilidad de que se venda entre 250 y 750 litros es

$$P(250 < X < 750) = (0{,}001)(750) - (0{,}001)(250) = 0{,}75 - 0{,}25 = 0{,}50$$

como muestra la Figura 6.1.

Hemos visto que la probabilidad de que una variable aleatoria continua se encuentre entre dos valores cualesquiera puede expresarse por medio de su función de distribución acumulada. Esta función contiene, pues, toda la información sobre la estructura de probabilidad de la variable aleatoria. Sin embargo, para muchos fines es más útil una función diferente. En el Capítulo 5 analizamos la función de probabilidad de las variables aleatorias discretas, que expresa la probabilidad de que una variable aleatoria discreta tome un valor específico cualquiera. Como la probabilidad de un valor específico es 0 en el caso de las variables aleatorias continuas, ese concepto no es directamente relevante aquí. Sin embargo, es posible construir una función relacionada con ésta, llamada *función de densidad de probabilidad*, para las variables aleatorias continuas, que permite la interpretación gráfica de su estructura de probabilidad.

Función de densidad de probabilidad

Sea X una variable aleatoria continua y x cualquier número situado en el rango de valores que puede tomar esta variable aleatoria. La **función de densidad de probabilidad**, $f(x)$, de la variable aleatoria es una función que tiene las siguientes propiedades:

1. $f(x) > 0$ para todos los valores de x.

2. El área situada debajo de la función de densidad de probabilidad, $f(x)$, cuando se abarcan todos los valores de la variable aleatoria, X, es igual a 1,0.
3. Supongamos que se representa gráficamente esta función de densidad. Sean a y b dos valores posibles de la variable aleatoria X, siendo $a < b$. En ese caso, la probabilidad de que X se encuentre entre a y b es el área situada debajo de la función de densidad entre estos puntos.
4. La función de distribución acumulada, $F(x_0)$, es el área situada debajo de la función de densidad de probabilidad, $f(x)$, hasta x_0:

$$F(x_0) = \int_{x_m}^{x_0} f(x)\, dx$$

donde x_m es el valor mínimo de la variable aleatoria X.

Es posible aproximarse a la función de densidad de probabilidad a partir de una distribución de probabilidad discreta en la que se consideran muchos valores cercanos entre sí, como se observa en la Figura 6.2.

La Figura 6.3 muestra una función de densidad de probabilidad arbitraria de una variable aleatoria continua. Se muestran dos valores posibles, a y b, y el área sombreada situada debajo de la curva entre estos puntos es la probabilidad de que la variable aleatoria se encuentre en el intervalo entre ellos (véase el apéndice del capítulo).

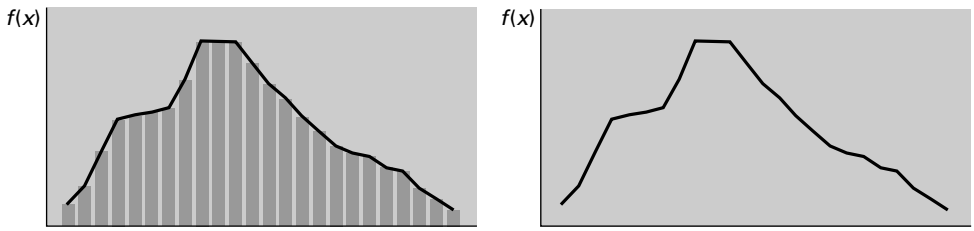


Figura 6.2. Aproximación de una función de densidad de probabilidad por medio de una distribución de probabilidad discreta.

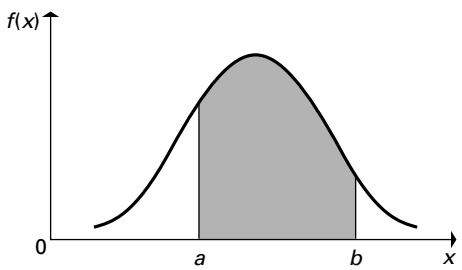


Figura 6.3. El área sombreada es la probabilidad de que X se encuentre entre a y b.

Áreas situadas debajo de funciones de probabilidad continua

Sea X una variable aleatoria continua que tiene una función de densidad de probabilidad $f(x)$ y una función de distribución acumulada $F(x)$. Se cumplen las siguientes propiedades:

1. El área total situada debajo de la curva $f(x)$ es 1.
2. El área situada debajo de la curva $f(x)$ a la izquierda de x_0 es $F(x_0)$, donde x_0 es cualquier valor que pueda tomar la variable aleatoria.

Estos resultados se muestran en la Figura 6.4; la 6.4(a) muestra que toda el área situada debajo de la función de densidad de probabilidad es igual a 1 y la 6.4(b) indica el área situada a la izquierda de x_0.

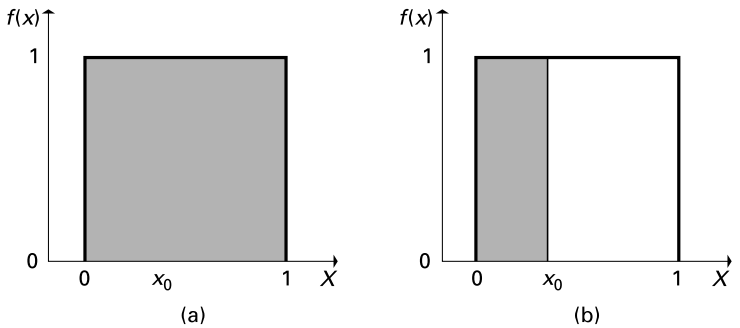


Figura 6.4. Propiedades de la función de densidad de probabilidad.

La distribución uniforme

A continuación, examinamos una función de densidad de probabilidad que representa una distribución de probabilidad en el rango de 0 a 1. La Figura 6.5 es una representación gráfica de la función de densidad de probabilidad uniforme. Ésta es la función de densidad de probabilidad del ejemplo de las ventas de gasolina. Dado que la probabilidad es la misma en cualquier intervalo de ventas que esté comprendido entre 0 y 1, deducimos que la función de densidad de probabilidad es constante en el rango de 0 a 1.000; una función como ésa se llama función de densidad de probabilidad uniforme y puede expresarse de la forma siguiente:

$$f(x) = \begin{cases} 0{,}001 & \text{si } 0 \leqslant x \leqslant 1.000 \\ 0 & \text{en caso contrario} \end{cases}$$

Cualquier variable aleatoria uniforme definida en el rango entre a y b tiene la siguiente función de densidad de probabilidad:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leqslant x \leqslant b \\ 0 & \text{en caso contrario} \end{cases}$$

Esta función de densidad de probabilidad puede utilizarse para hallar la probabilidad de que la variable aleatoria se encuentre dentro de un intervalo específico. Por ejemplo, la Figura 6.6 muestra la probabilidad de que se venda entre 250 litros y 750. Como la altura de la función de densidad es $f(x) = 0{,}001$, el área situada debajo de la curva entre 250 y 750 es igual a 0,50, que es la probabilidad que buscamos. Obsérvese que este resultado es igual que el que hemos obtenido antes con la función de probabilidad acumulada.

Hemos visto que la probabilidad de que una variable aleatoria se encuentre entre un par de valores es el área situada debajo de la función de densidad de probabilidad entre estos dos valores. Merece la pena señalar dos importantes resultados. El área situada debajo de toda la función de densidad de probabilidad es 1 y la probabilidad acumulada, $F(x_0)$, es el área situada debajo de la función de densidad a la izquierda de x_0.

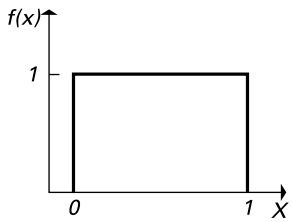


Figura 6.5. Función de densidad de probabilidad de una variable aleatoria uniforme que toma valores entre 0 y 1.

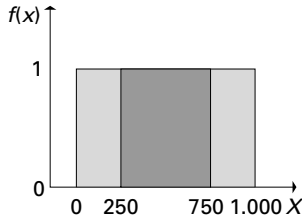


Figura 6.6. Función de densidad que muestra la probabilidad de que X se encuentre entre 250 y 750.

EJEMPLO 6.1. Probabilidad de que haya grietas en un oleoducto (función de distribución acumulada)

Un equipo de reparación es responsable de un tramo de un oleoducto de 2 kilómetros de largo. La distancia (en kilómetros) a la que surge cualquier grieta puede representarse por medio de una variable aleatoria distribuida uniformemente, con una función de densidad de probabilidad

$$f(x) = 0{,}5$$

Halle la función de distribución acumulada y la probabilidad de que surja cualquier grieta dada entre 0,5 kilómetros y 1,5 en este tramo del oleoducto.

Solución

La Figura 6.7 representa la función de densidad de probabilidad; el área sombreada representa $F(x_0)$, la función de distribución acumulada evaluada en x_0. Vemos, pues, que

$$F(x_0) = 0{,}5x_0 \qquad \text{para } 0 < x_0 < 2$$

La probabilidad de que surja una grieta entre 0,5 kilómetros y 1,5 en el oleoducto es

$$\begin{aligned} P(0{,}5 < X < 1{,}5) &= F(1{,}5) - F(0{,}5) \\ &= (0{,}5)(1{,}5) - (0{,}5)(0{,}5) = 0{,}5 \end{aligned}$$

Ésta es el área situada debajo de la función de densidad de probabilidad de $x = 0{,}5$ a $x = 1{,}5$.

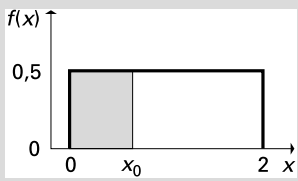


Figura 6.7. Función de densidad de probabilidad del ejemplo 6.1.

EJERCICIOS

Ejercicios básicos

6.1. Utilizando la función de densidad de probabilidad uniforme mostrada en la Figura 6.7, halle la probabilidad de que la variable aleatoria X esté entre 1,4 y 1,8.

6.2. Utilizando la función de densidad de probabilidad uniforme mostrada en la Figura 6.7, halle la probabilidad de que la variable aleatoria X esté entre 1,0 y 1,9.

6.3. Utilizando la función de densidad de probabilidad uniforme mostrada en la Figura 6.7, halle la probabilidad de que la variable aleatoria X sea menor que 1,4.

6.4. Utilizando la función de densidad de probabilidad uniforme mostrada en la Figura 6.7, halle la probabilidad de que la variable aleatoria X sea mayor que 1,3.

Ejercicios aplicados

6.5. Un analista dispone de dos predicciones, F_1 y F_2, de los beneficios por acción que obtendrá una empresa el próximo año. Pretende hacer una predicción intermedia que sea una media ponderada de las dos predicciones. Para hacer esa predicción, dará la ponderación X a la primera predicción y la ponderación $(1 - X)$ a la segunda, por lo que la predicción intermedia compromiso es $XF_1 + (1 - X)F_2$. El analista quiere elegir un valor entre 0 y 1 para la ponderación X, pero no sabe cuál es mejor. Suponga que lo que decide finalmente como mejor elección posible de la ponderación X puede concebirse como una variable aleatoria distribuida uniformemente entre 0 y 1, que tiene la función de densidad de probabilidad

$$f(x) = \begin{cases} 1 & \text{para } 0 \leqslant x \leqslant 1 \\ 0 & \text{para todos los demás valores de } x \end{cases}$$

a) Trace la función de densidad de probabilidad.
b) Halle y trace la función de distribución acumulada.
c) Halle la probabilidad de que la mejor elección de la ponderación X sea inferior a 0,25.
d) Halle la probabilidad de que la mejor elección de la ponderación X sea superior a 0,75.
e) Halle la probabilidad de que la mejor elección de la ponderación X esté entre 0,2 y 0,8.

6.6. Dentro de la jurisdicción de un equipo de salvamento se encuentran las emergencias que se produzcan en un tramo de un río que tiene 4 kilómetros de largo. La experiencia ha demostrado que la distancia, expresada en kilómetros desde el punto situado más al norte, a la que se produce una emergencia dentro de este tramo puede representarse por medio de una variable aleatoria distribuida uniformemente en el rango 0 a 4 kilómetros. En ese caso, si X representa la distancia (en kilómetros) a la que se produce una emergencia desde el punto situado más al norte de este tramo del río, su función de densidad de probabilidad es

$$f(x) = \begin{cases} 0{,}25 & \text{para } 0 < x < 4 \\ 0 & \text{para todos los demás valores de } x \end{cases}$$

a) Trace la función de densidad de probabilidad.
b) Halle y trace la función de distribución acumulada.
c) Halle la probabilidad de que se produzca una emergencia como máximo a un kilómetro del punto situado más al norte de este tramo del río.
d) La base del equipo de salvamento se encuentra en el punto medio de este tramo del río. Halle la probabilidad de que se produzca una emergencia a más de 1,5 kilómetros de esta base.

6.7. Las rentas de todas las familias de un barrio pueden representarse por medio de una variable aleatoria continua. Se sabe que la renta mediana de todas las familias de este barrio es de 60.000 $ y que el 40 por ciento de todas las familias del barrio tiene una renta de más de 72.000 $.

a) ¿Cuál es la probabilidad de que la renta de una familia elegida aleatoriamente esté comprendida entre 60.000 $ y 72.000 $?
b) Dado que no se dispone de más información, ¿qué puede decirse sobre la probabilidad de que una familia elegida aleatoriamente tenga una renta de menos de 65.000 $?

6.8. Al comienzo del invierno, la propietaria de un piso estima que la probabilidad de que su factura total de calefacción en los tres meses del invierno sea de menos de 380 $ es de 0,4. También estima que la probabilidad de que sea de menos de 460 $ es de 0,6.

a) ¿Cuál es la probabilidad de que la factura total esté comprendida entre 380 $ y 460 $?
b) Dado que no se dispone de más información, ¿qué puede decirse sobre la probabilidad de que la factura total sea de menos de 400 $?

6.2. Esperanzas de variables aleatorias continuas

En el apartado 5.2 presentamos los conceptos de valor esperado de una variable aleatoria discreta y de valor esperado de una función de esa variable aleatoria. Aquí extendemos esas ideas a las variables aleatorias continuas. Como la probabilidad de cualquier valor específico es 0 en el caso de una variable aleatoria continua, tenemos que utilizar la ecuación 6.3.

Justificación de las esperanzas de variables aleatorias continuas

Supongamos que en un experimento aleatorio se obtiene un resultado que puede representarse por medio de una variable aleatoria continua. Si se realizan N réplicas independientes de este experimento, el **valor esperado** de la variable aleatoria es la media de los valores obtenidos, cuando el número de réplicas tiende a infinito. El valor esperado de una variable aleatoria se representa de la siguiente manera: $E(X)$.

Asimismo, si $g(X)$ es cualquier función de la variable aleatoria X, el valor esperado de esta función es el valor medio obtenido en pruebas independientes repetidas, cuando el número de pruebas tiende a infinito. Esta esperanza se representa de la siguiente manera: $E[g(X)]$.

Utilizando el cálculo podemos definir los valores esperados de variables aleatorias continuas similares a los utilizados en el caso de las variables aleatorias discretas:

$$E[g(x)] = \int_x g(x) f(x)\, dx \qquad \textbf{(6.3)}$$

Estos conceptos pueden presentarse claramente si se sabe cálculo integral, como se muestra en el apéndice del capítulo. Utilizando la ecuación 6.3, podemos calcular la media y la varianza de variables aleatorias continuas. Las ecuaciones 6.4 y 6.5 presentan la media y la varianza de variables aleatorias continuas.

Media, varianza y desviación típica de variables aleatorias continuas

Sea X una variable aleatoria continua. Hay dos importantes valores esperados que se utilizan habitualmente para definir las distribuciones de probabilidad continua.

1. La **media de *X***, representada por μ_X, es el valor esperado de X:

$$\mu_X = E(X) \qquad \textbf{(6.4)}$$

2. La **varianza de *X***, representada por σ_X^2, es la esperanza del cuadrado de la diferencia entre la variable aleatoria y su media $(X - \mu_X)^2$:

$$\sigma_X^2 = E[(X - \mu_X)^2] \qquad \textbf{(6.5)}$$

Otra expresión es:

$$\sigma_X^2 = E(X^2) - \mu_X^2 \qquad \textbf{(6.6)}$$

La **desviación típica de *X***, σ_X, es la raíz cuadrada de la varianza.

La media y la varianza constituyen dos importantes indicadores sintéticos de una distribución de probabilidad. La media es una medida del centro de la distribución. Consideremos la siguiente interpretación física: recortemos el gráfico de una función de densidad de

probabilidad. El punto del eje de las x en el que la figura está exactamente en equilibrio sobre un dedo es la media de la distribución. Por ejemplo, en la Figura 6.4 la distribución uniforme es simétrica alrededor de $x = 0{,}5$ y, por lo tanto, $\mu_X = 0{,}5$ es la media de la variable aleatoria.

La varianza —o su raíz cuadrada, la desviación típica— es una medida de la dispersión de una distribución. Así, por ejemplo, si comparamos dos distribuciones uniformes que tienen la misma media, $\mu_X = 1$ —una en el rango 0,5 a 1,5 y la otra en el rango 0 a 2—, observaremos que la segunda tiene una varianza mayor porque se distribuye a lo largo de un intervalo mayor.

Para una ***distribución uniforme*** definida en el rango a a b, tenemos los siguientes resultados:

$$f(x) = \frac{1}{b-a}$$

$$a \leqslant X \leqslant b$$

$$\mu_x = E[X] = \frac{a+b}{2}$$

$$\sigma_x^2 = E[(X-\mu_x)^2] = \frac{(b-a)^2}{12}$$

En el apartado 5.3 mostramos cómo se calculan las medias y las varianzas de funciones lineales de variables aleatorias discretas. Los resultados son iguales en el caso de las variables aleatorias continuas, ya que se utiliza el operador del valor esperado. Repetimos aquí el resumen de los resultados del Capítulo 5.

Funciones lineales de variables aleatorias

Sea X una variable aleatoria continua de media μ_X y de varianza σ_X^2 y sean a y b unas constantes cualesquiera. Definiendo la variable aleatoria W,

$$W = a + bX$$

la media y la varianza de W son

$$\mu_W = E(a + bX) = a + b\mu_X \tag{6.7}$$

y

$$\sigma_W^2 = \mathrm{Var}\,(a + bX) = b^2\sigma_X^2 \tag{6.8}$$

y la desviación típica de W es

$$\sigma_W = |b|\sigma_X \tag{6.9}$$

Un importante caso especial de estos resultados es la variable aleatoria estandarizada

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{6.10}$$

de media 0 y varianza 1.

EJEMPLO 6.2. Costes de calefacción de una casa (media y desviación típica)

El propietario de un piso estima que dentro del rango de temperaturas probables, su factura de calefacción, Y, de enero en dólares será

$$Y = 290 - 5T$$

donde T es la temperatura media del mes, en grados Fahrenheit. Si la temperatura media de enero puede representarse por medio de una variable aleatoria que tiene una media de 24 y una desviación típica de 4, halle la media y la desviación típica de la factura de la calefacción de enero de este propietario.

Solución

La variable aleatoria T tiene una media $\mu_T = 24$ y una desviación típica $\sigma_T = 4$. Por lo tanto, la factura esperada de la calefacción es

$$\begin{aligned}\mu_Y &= 290 - 5\mu_T \\ &= 290 - (5)(24) = 170\ \$\end{aligned}$$

La desviación típica es

$$\sigma_Y = |-5|\,\sigma_T = (5)(4) = 20\ \$$$

EJERCICIOS

Ejercicios básicos

6.9. El coste total de un proceso de producción es de 1.000 $ más el doble del número de unidades producidas. La media y la varianza del número de unidades producidas son 500 y 900, respectivamente. Halle la media y la varianza del coste total.

6.10. El beneficio de un proceso de producción es de 1.000 $ menos el doble del número de unidades producidas. La media y la varianza del número de unidades producidas son 50 y 90, respectivamente. Halle la media y la varianza del beneficio.

6.11. El beneficio de un proceso de producción es de 2.000 $ menos el doble del número de unidades producidas. La media y la varianza del número de unidades producidas son 500 y 900, respectivamente. Halle la media y la varianza del beneficio.

6.12. El beneficio de un proceso de producción es de 6.000 $ menos el triple del número de unidades producidas. La media y la varianza del número de unidades producidas son 1.000 y 900, respectivamente. Halle la media y la varianza del beneficio.

Ejercicios aplicados

6.13. Un autor recibe de una editorial un contrato, según el cual recibirá una cantidad fija de 10.000 $ más 1,50 $ por cada ejemplar que se venda de su libro. Su incertidumbre sobre las ventas totales del libro pueden representarse por medio de una variable aleatoria que tiene una media de 30.000 y una desviación típica de 8.000. Halle la media y la desviación típica de la cantidad total de dinero que recibirá.

6.14. Un contratista presenta una oferta para realizar un proyecto, para el que hay que hacer más investigación y desarrollo. Se estima que el coste total del cumplimiento de las especificaciones del proyecto es de 20 millones de dólares más el coste de la investigación y el desarrollo adicionales. El contratista considera que el coste de este trabajo es una variable aleatoria que tiene una media de 4 millones de dólares y una desviación típica de 1 millón de dólares. El contratista desea presentar una oferta tal que su beneficio esperado sea un 10 por ciento de sus costes esperados. ¿Qué oferta debe presentar? Si se le acepta, ¿cuál será la desviación típica del beneficio generado por el proyecto?

6.15. Una organización benéfica solicita donaciones por teléfono. Los empleados reciben 60 $ más un 20 por ciento del dinero que generan las llamadas cada semana. La cantidad de dinero generada en una semana puede concebirse como una variable aleatoria que tiene una media de 700 $ y una desviación típica de 130 $. Halle la media y la desviación típica de la remuneración total de un empleado en una semana.

6.16. Un vendedor tiene un sueldo anual de 6.000 $ más un 8 por ciento del valor de los pedidos que reciba. El valor anual de estos pedidos puede representarse por medio de una variable aleatoria que tiene una media de 600.000 $ y una desviación típica de 180.000 $. Halle la media y la desviación típica de la renta anual del vendedor.

6.3. La distribución normal

En este apartado presentamos la distribución de probabilidad normal, que es la distribución de probabilidad de variables aleatorias continuas que se utiliza más a menudo en economía y en las aplicaciones empresariales. La Figura 6.8 muestra un ejemplo de la función de densidad de probabilidad normal.

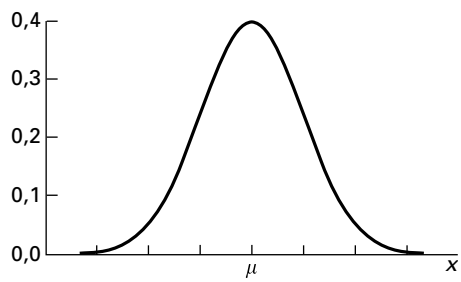


Figura 6.8. Función de densidad de probabilidad de una distribución normal.

Son muchas las razones por las que se utiliza frecuentemente.

1. La distribución normal es una aproximación muy buena de las distribuciones de probabilidad de una amplia variedad de variables aleatorias. Por ejemplo, las dimensiones de las piezas y el peso de los paquetes de alimentos a menudo siguen una distribución normal, por lo que tiene muchas aplicaciones en el control de calidad. Las ventas o la producción a menudo siguen una distribución normal, por lo que ésta tiene una gran cantidad de aplicaciones en el marketing y en la gestión de la producción. Las pautas de los precios de las acciones y de los bonos a menudo se analizan utilizando la distribución normal en grandes modelos informáticos de contratación financiera. Los modelos económicos utilizan la distribución normal para algunas medidas económicas.
2. Las distribuciones de las medias muestrales siguen una distribución normal, si el tamaño de la muestra es «grande».
3. El cálculo de probabilidades es directo e ingenioso.
4. La razón más importante es que la distribución de probabilidad normal ha llevado a tomar buenas decisiones empresariales en algunas aplicaciones.

La ecuación 6.11 define formalmente la función de densidad de probabilidad normal.

Función de densidad de probabilidad de la distribución normal

La **función de densidad de probabilidad de una variable aleatoria *X* que sigue una distribución normal *X*** es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{para } -\infty < x < \infty \tag{6.11}$$

donde μ y σ^2 son números tales que $-\infty < \mu < \infty$ y $0 < \sigma^2 < \infty$ y donde e y π son constantes físicas, $e = 2{,}71828...$ y $\pi = 3{,}14159...$

La distribución normal representa una gran familia de distribuciones, cada una con una especificación única de los parámetros μ y σ^2. Estos parámetros tienen una interpretación muy útil.

Propiedades de la distribución normal

Supongamos que la variable aleatoria X sigue una distribución normal cuyos parámetros son σ^2. En ese caso, se cumplen las siguientes propiedades:

1. La media de la variable aleatoria es μ:

$$E(X) = \mu$$

2. La varianza de la variable aleatoria es σ^2:

$$\text{Var}(X) = E[(X - \mu)^2] = \sigma^2$$

3. La forma de la función de densidad de probabilidad es una curva simétrica en forma de campana centrada en la media, μ, como muestra la Figura 6.8.
4. Si conocemos la media y la varianza, podemos definir la distribución normal utilizando la notación

$$X \sim N(\mu, \sigma^2)$$

La distribución normal tiene algunas características importantes para nuestros análisis estadísticos aplicados. Es simétrica. Las diferentes tendencias centrales son indicadas por las diferencias entre las μ. En cambio, las diferencias entre las σ^2 dan como resultado funciones de densidad de diferentes amplitudes. Seleccionando distintos valores de μ y σ^2, podemos definir una gran familia de funciones de densidad normales. Si cambia la media, se desplaza toda la distribución. Pero cambiando la varianza se obtienen distribuciones de diferentes amplitudes.

La media de la distribución es una medida de la tendencia central y la varianza es una medida de la dispersión en torno a la media. Por lo tanto, los parámetros μ y σ^2 producen diferentes efectos en la función de densidad de una variable aleatoria normal. La Figura 6.9(a) muestra funciones de densidad de dos distribuciones normales que tienen una varianza común y diferentes medias. Vemos que los aumentos de la media desplazan la distribución sin alterar su forma. En la Figura 6.9(b), las dos funciones de densidad tienen la misma media, pero diferentes varianzas. Las dos son simétricas en torno a la media común, pero la que tiene la mayor varianza es más dispersa.

Nuestra siguiente tarea es aprender a hallar las probabilidades de una distribución normal específica. Primero presentamos la *función de distribución acumulada*.

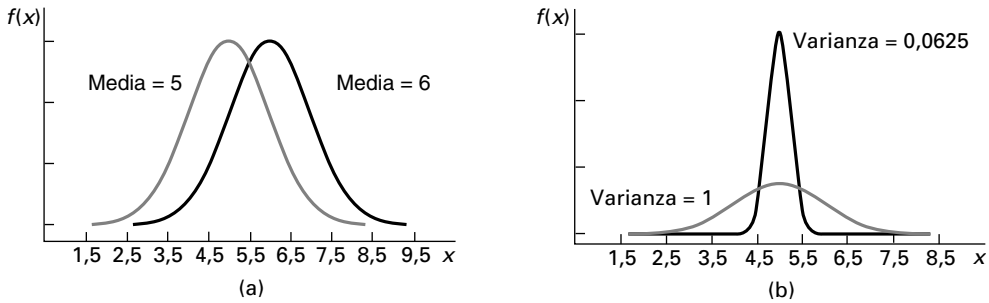


Figura 6.9. Efectos de μ y σ^2 en la función de densidad de una variable aleatoria normal:
(a) Dos distribuciones normales que tienen diferentes medias.
(b) Dos distribuciones normales que tienen diferentes varianzas y media = 5.

Función de distribución acumulada de la distribución normal

Supongamos que X es una variable aleatoria normal de media μ y varianza σ^2; es decir, $X \sim N(\mu, \sigma^2)$. En ese caso, la función de distribución acumulada es

$$F(x_0) = P(X \leqslant x_0)$$

Ésta es el área situada debajo de la función de densidad normal a la izquierda de x_0, como se muestra en la Figura 6.10. Al igual que ocurre en cualquier función de densidad, el área total situada debajo de la curva es 1; es decir,

$$F(\infty) = 1$$

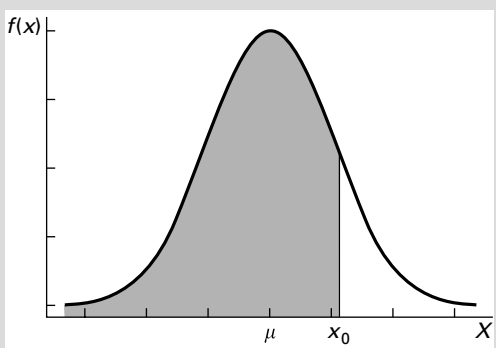


Figura 6.10. El área sombreada es la probabilidad de que X no sea mayor que x_0 en el caso de una variable aleatoria normal.

No tenemos una expresión algebraica sencilla para calcular la función de distribución acumulada de una variable aleatoria distribuida normalmente (véase el apéndice del capítulo). La Figura 6.11 muestra la forma general de la función de distribución acumulada. Se emplea la ecuación 6.12 para calcular las probabilidades normales utilizando la función de distribución acumulada.

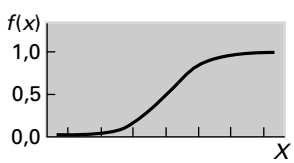


Figura 6.11. Distribución acumulada de una variable aleatoria normal.

Probabilidades de intervalos de variables aleatorias normales

Sea X una variable aleatoria normal que tiene una función de distribución acumulada $F(x)$ y sean a y b dos valores posibles de X, siendo $a < b$. Entonces,

$$P(a < X < b) = F(b) - F(a) \qquad \textbf{(6.12)}$$

La probabilidad es el área situada debajo de la correspondiente función de densidad entre a y b, como muestra la Figura 6.12.

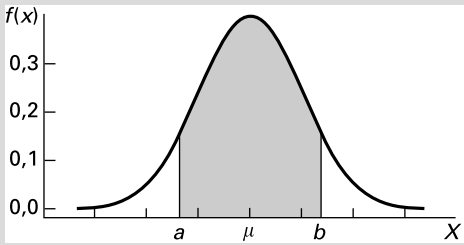


Figura 6.12. Función de densidad normal en la que el área sombreada indica la probabilidad de que X se encuentre entre a y b.

Es posible hallar cualquier probabilidad a partir de la función de distribución acumulada. Sin embargo, no disponemos de un método cómodo para calcular directamente la probabilidad de cualquier distribución normal que tenga una media y una varianza específicas. Podríamos utilizar métodos numéricos de integración por computador, pero ese método sería tedioso y pesado. Afortunadamente, podemos convertir cualquier distribución normal en una *distribución normal estándar* de media 0 y varianza 1.

La distribución normal estándar

Sea Z una variable aleatoria normal de media 0 y varianza 1; es decir,

$$Z \sim N(0, 1)$$

Decimos que Z sigue la **distribución normal estándar**.

Si la función de distribución acumulada es $F(z)$ y a y b son dos números tales que $a < b$, entonces,

$$P(a < Z < b) = F(b) - F(a) \qquad \textbf{(6.13)}$$

Podemos hallar las probabilidades de cualquier variable aleatoria distribuida normalmente convirtiendo primero la variable aleatoria en la variable aleatoria normal estándar, Z. Siempre existe una relación directa entre cualquier variable aleatoria distribuida normalmente y Z. Esa relación utiliza la transformación

$$Z = \frac{X - \mu}{\sigma}$$

donde X es una variable aleatoria distribuida normalmente:

$$X \sim N(\mu, \sigma^2)$$

Este importante resultado nos permite utilizar la tabla normal estándar para calcular las probabilidades de cualquier variable aleatoria distribuida normalmente. Veamos ahora cómo pueden calcularse las probabilidades de la variable aleatoria normal estándar Z.

La función de distribución acumulada de la distribución normal estándar se encuentra en la Tabla 1 del apéndice. Esta tabla da los valores de

$$F(z) = P(Z \leqslant z)$$

correspondientes a los valores no negativos de z. Por ejemplo, en la citada tabla vemos que la probabilidad acumulada de un valor de Z de 1,25 es

$$F(1{,}25) = 0{,}8944$$

Ésta es el área, representada en la Figura 6.13, correspondiente a los valores de Z inferiores a 1,25. Como consecuencia de la simetría de la distribución normal, la probabilidad de que $Z > -1{,}25$ también es igual a 0,8944. En general, los valores de la función de distribución acumulada correspondiente a los valores negativos de Z pueden deducirse utilizando la simetría de la función de densidad.

Para hallar la probabilidad acumulada de un valor negativo de Z (por ejemplo, $Z = -1{,}0$), que se define de la forma siguiente,

$$F(-Z_0) = P(Z \leqslant -z_0) = F(-1{,}0)$$

utilizamos el complemento de la probabilidad de $Z = +1$, mostrado en la Figura 6.14.

De la simetría podemos deducir que

$$F(-Z) = 1 - P(Z \leqslant +Z) = 1 - F(Z)$$

$$F(-1) = 1 - P(Z \leqslant +1) = 1 - F(1)$$

La Figura 6.15 indica la simetría de los valores positivos correspondientes de Z.

En la Figura 6.16 podemos ver que el área situada debajo de la curva a la izquierda de $Z = -1$ es igual al área situada a la derecha de $Z = +1$ debido a la simetría de la distribución normal. El área situada muy por debajo de $-Z$ a menudo se llama «cola inferior» y el área situada muy por encima de $+Z$ se llama «cola superior».

También podemos utilizar tablas normales que indican las probabilidades de los valores de Z de la mitad superior o positivos a partir de la distribución normal. Dentro de la portada del libro hay un ejemplo de este tipo de tabla. Este tipo de tabla normal se utiliza para hallar las probabilidades de la misma forma que antes. Cuando los valores de Z son positi-

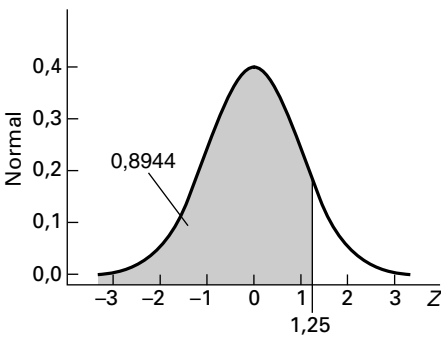


Figura 6.13. Probabilidad correspondiente a $Z = 1{,}25$ en una distribución normal estándar.

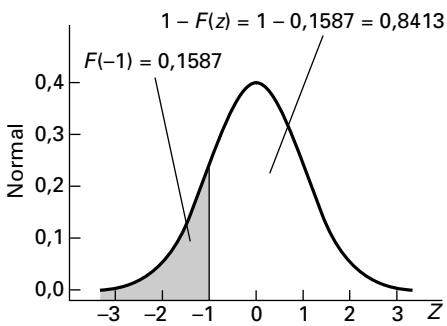


Figura 6.14. Distribución normal estándar cuando Z es igual a -1.

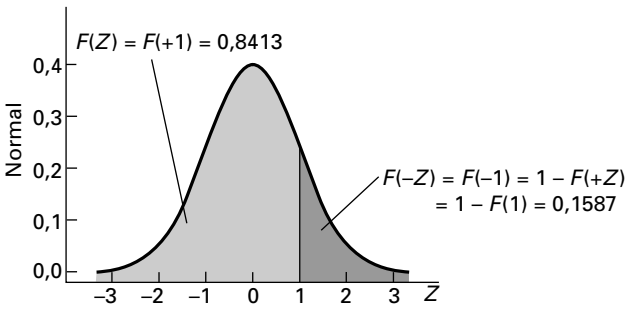


Figura 6.15. Distribución normal estándar cuando Z es igual a $+1$.

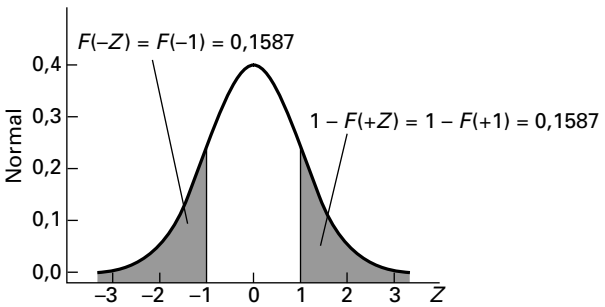


Figura 6.16. Valores superior e inferior simétricos en una función de densidad normal.

vos, sumamos 0,50 a los valores que se indican en la tabla que se encuentra dentro de la portada del libro. Cuando son negativos, utilizamos la simetría de la normal para hallar las probabilidades deseadas.

EJEMPLO 6.3. Probabilidades del valor de una cartera de inversión (probabilidades normales)

Un cliente tiene una cartera de inversión cuyo valor medio es de 500.000 $ y cuya desviación típica es 15.000 $. Le ha pedido que calcule la probabilidad de que el valor de su cartera esté entre 485.000 $ y 530.000 $.

Solución

El problema se muestra en la Figura 6.17. Para resolverlo, primero tenemos que hallar los valores correspondientes de Z de los límites de la cartera. El valor de Z correspondiente a 485.000 $ es

$$z_{485} = \frac{485.000 - 500.000}{15.000} = -1{,}0$$

Y el valor de X correspondiente al valor superior, 530.000 $, es

$$z_{530} = \frac{530.000 - 500.000}{15.000} = +2{,}0$$

Como muestra la Figura 6.17, la probabilidad de que el valor de la cartera, X, esté entre 485.000 $ y 530.000 $ es igual a la probabilidad de que Z esté entre -1 y $+2$.

Para hallar la probabilidad, primero calculamos las probabilidades de la cola inferior y de la cola superior y restamos estas probabilidades de 1. En términos algebraicos, el resultado es

$$P(485.000 \leqslant X \leqslant 530.000) = P(-1 \leqslant Z \leqslant +2) = 1 - P(Z \leqslant -1) - P(Z \geqslant +2)$$
$$= 1 - 0{,}1587 - 0{,}0228 = 0{,}8185$$

La probabilidad del intervalo indicado es, pues, 0,8185.

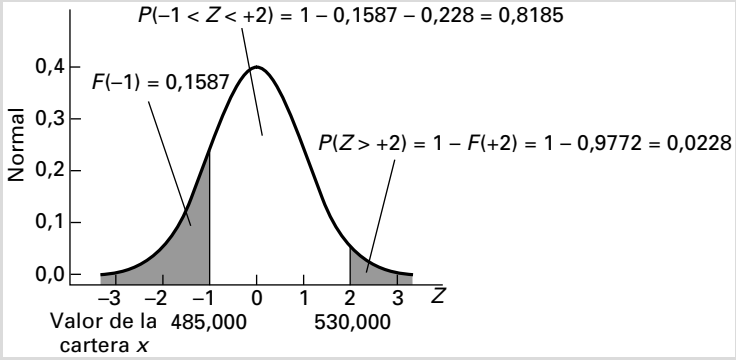


Figura 6.17. Distribución normal del ejemplo 6.3.

Recuérdese que en el Capítulo 2 presentamos la regla empírica que establece como una guía aproximada que $\mu \pm \sigma$ abarca alrededor del 68 por ciento del rango, mientras que $\mu \pm 2\sigma$ abarca alrededor del 95 por ciento del rango. A todos los efectos prácticos, casi ningún valor del rango se encuentra a más de 3σ de μ. Este útil instrumento de aproximación para las interpretaciones realizadas a partir de los estadísticos descriptivos se basa en la distribución normal.

Las probabilidades también pueden calcularse por medio de la ecuación 6.14.

Cómo se hallan las probabilidades de variables aleatorias distribuidas normalmente

Sea X una variable aleatoria distribuida normalmente de media μ y varianza σ^2. La variable aleatoria $Z = (X - \mu)/\sigma$ tiene una distribución normal estándar: $Z \sim N(0, 1)$.

Se deduce que si a y b son dos números tales que $a < b$, entonces

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$
$$= F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right) \qquad \textbf{(6.14)}$$

donde Z es la variable aleatoria normal estándar y F representa su función de distribución acumulada.

EJEMPLO 6.4. Distribución de probabilidad normal (probabilidades normales)

Si $X \sim N(15, 16)$, halle la probabilidad de que X sea mayor que 18.

Solución

Esta probabilidad puede calcularse de la forma siguiente:

$$\begin{aligned} P(x > 18) &= P\left(Z > \frac{18 - \mu}{\sigma}\right) \\ &= P\left(Z > \frac{18 - 15}{4}\right) \\ &= P(Z > 0{,}75) \\ &= 1 - P(Z < 0{,}75) \\ &= 1 - F(0{,}75) \end{aligned}$$

En la Tabla 1 del apéndice vemos que $F(0{,}75)$ es 0,7734 y, por lo tanto,

$$P(X > 18) = 1 - 0{,}7734 = 0{,}2266$$

EJEMPLO 6.5. La duración de una bombilla (probabilidades normales)

Una empresa produce bombillas cuya duración sigue una distribución normal que tiene una media de 1.200 horas y una desviación típica de 250 horas. Si elegimos una bombilla aleatoriamente, ¿cuál es la probabilidad de que dure entre 900 y 1.300 horas?

Solución

Sea X la duración en horas. Entonces,

$$\begin{aligned} P(900 < X < 1.300) &= P\left(\frac{900 - 1.200}{250} < Z < \frac{1.300 - 1.200}{250}\right) \\ &= P(-1{,}2 < Z < 0{,}4) \\ &= F(0{,}4) - F(-1{,}2) \\ &= 0{,}6554 - (1 - 0{,}8849) = 0{,}5403 \end{aligned}$$

Por lo tanto, la probabilidad de que una bombilla dure entre 900 y 1.300 horas es aproximadamente de 0,54.

EJEMPLO 6.6. Calificaciones de un examen (probabilidades normales)

Un grupo muy numeroso de estudiantes obtiene unas calificaciones (de 0 a 100) que siguen una distribución normal que tiene una media de 60 y una desviación típica de 15. ¿Qué proporción de los estudiantes obtiene una calificación de entre 85 y 95?

Solución

Sea X la calificación del examen. En ese caso, la probabilidad puede calcularse de la forma siguiente:

$$\begin{aligned} P(85 < X < 95) &= P\left(\frac{85 - 60}{15} < Z < \frac{95 - 60}{15}\right) \\ &= P(1{,}67 < Z < 2{,}33) \\ &= F(2{,}33) - F(1{,}67) \\ &= 0{,}9901 - 0{,}9525 = 0{,}0376 \end{aligned}$$

Es decir, el 3,76 por ciento de los estudiantes obtuvo una calificación comprendida entre 85 y 95.

EJEMPLO 6.7. Puntos de corte de las calificaciones de un examen (variables aleatorias normales)

Halle el punto de corte del 10 por ciento superior de todos los estudiantes correspondiente a las calificaciones del ejemplo 6.6.

Solución

Sea b el punto de corte. Para hallar el valor numérico del punto de corte, primero observamos que la probabilidad de superar b es 0,10 y, por lo tanto, la probabilidad de no llegar a b es 0,90. El valor de la cola superior de 0,10 se muestra en la Figura 6.18. Ahora podemos hallar la probabilidad a partir de la distribución acumulada de la forma siguiente:

$$\begin{aligned} 0{,}90 &= P\left(Z < \frac{b - 60}{15}\right) \\ &= F\left(\frac{b - 60}{15}\right) \end{aligned}$$

En la Tabla 1 del apéndice vemos que $Z = 1{,}28$ cuando $F(Z) = 0{,}90$. Por lo tanto, despejando b, tenemos que

$$\frac{b - 60}{15} = 1{,}28$$

$$b = 79{,}2$$

Llegamos, pues, a la conclusión de que el 10 por ciento de los estudiantes obtiene una calificación de más de 79,2, como muestra la Figura 6.18.

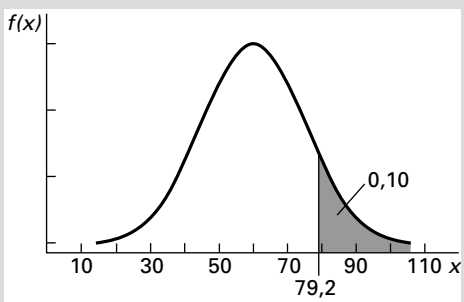


Figura 6.18. Distribución normal de media 60 y desviación típica 15 que muestra una probabilidad de la cola superior igual a 0,10.

Debe señalarse que las calificaciones de los exámenes, como las de los ejemplos 6.6 y 6.7, normalmente son números enteros y, por lo tanto, la distribución de las calificaciones es discreta. Sin embargo, debido al gran número de resultados posibles, la distribución normal es una aproximación muy buena de la distribución discreta. En la mayoría de los problemas empresariales y económicos aplicados, utilizamos, de hecho, la distribución normal como aproximación de una distribución discreta que tiene muchos resultados diferentes.

Gráficos de probabilidades normales

El modelo de probabilidad normal es el más utilizado por las razones antes señaladas. En los problemas aplicados, nos gustaría saber si los datos proceden de una distribución que se parece lo suficiente a una distribución normal para garantizar la validez del resultado. Buscamos, pues, pruebas que corroboren el supuesto de que la distribución normal es una aproximación cercana de la distribución desconocida efectiva. Los gráficos de probabilidades normales son útiles para contrastar este supuesto y averiguar si puede utilizarse el modelo normal. El uso es sencillo. Si los datos siguen una distribución normal, el gráfico es una línea recta.

La Figura 6.19 es una representación gráfica de probabilidades normales de una variable aleatoria de $n = 1.000$ observaciones extraídas de una distribución normal que tiene una $\mu = 100$ y una $\sigma = 25$. El gráfico se ha realizado utilizando el programa Minitab. El eje de abscisas indica los puntos de datos ordenados de menor a mayor. El de ordenadas indica las probabilidades normales acumuladas de los valores de los datos ordenados si los datos muestrales proceden de una población cuyas variables aleatorias siguen una distribución normal. Vemos que el eje de ordenadas tiene una escala normal transformada. El gráfico de la Figura 6.19 se parece a una línea recta incluso en el límite superior y en el inferior y ese resultado es una prueba sólida de que los datos siguen una distribución normal. Las líneas de trazo discontinuo constituyen un intervalo en el que se encontrarían los puntos de datos de una variable aleatoria distribuida normalmente en la mayoría de los casos. Por lo tanto, si los puntos representados se encuentran dentro de los límites establecidos por las líneas de trazo discontinuo, podemos concluir que los puntos de datos representan una variable aleatoria distribuida normalmente.

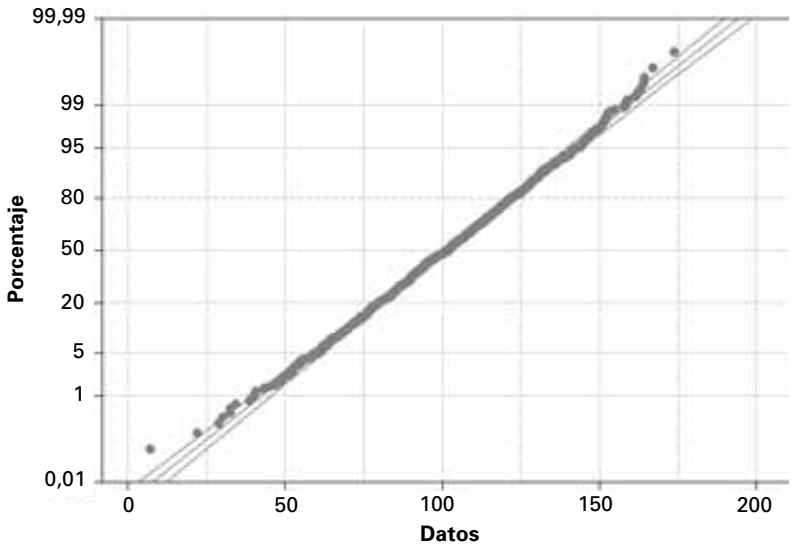


Figura 6.19. Gráfico de probabilidades normales de una distribución normal (salida Minitab).

A continuación, consideramos una muestra aleatoria de $n = 1.000$ observaciones extraídas de una *distribución uniforme* cuyos límites son 25 y 175. La Figura 6.20 muestra la representación gráfica de probabilidades normales. En este caso, la representación de los datos tiene una forma de **S** que se desvía claramente de una línea recta, por lo que los datos muestrales no siguen una distribución normal. Las grandes desviaciones en los valores altos y bajos extremos son un motivo de gran preocupación porque la inferencia estadística a menudo se basa en pequeñas probabilidades de valores extremos.

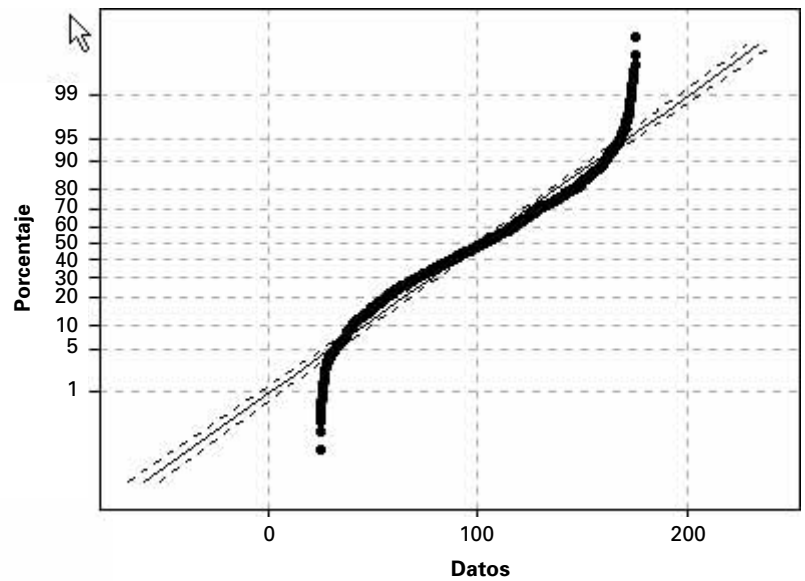


Figura 6.20. Gráfico de probabilidades normales de una distribución uniforme (salida Minitab).

Examinemos a continuación una distribución discreta muy sesgada, como la que muestra la Figura 6.21. En la 6.22 vemos la representación gráfica de probabilidades normales de esta distribución muy sesgada. Vemos, de nuevo, que no es una línea recta sino que tiene una desviación considerable en los valores altos y bajos extremos. Esta representación indica claramente que los datos no proceden de una distribución normal.

Los ejemplos anteriores nos dan una indicación de los resultados posibles de una representación de probabilidades normales. Si la representación de un problema es similar a la

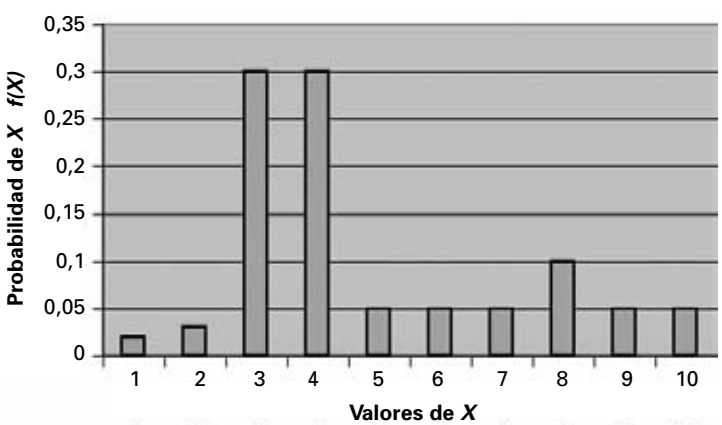


Figura 6.21. Función de la distribución de probabilidades discretas sesgadas.

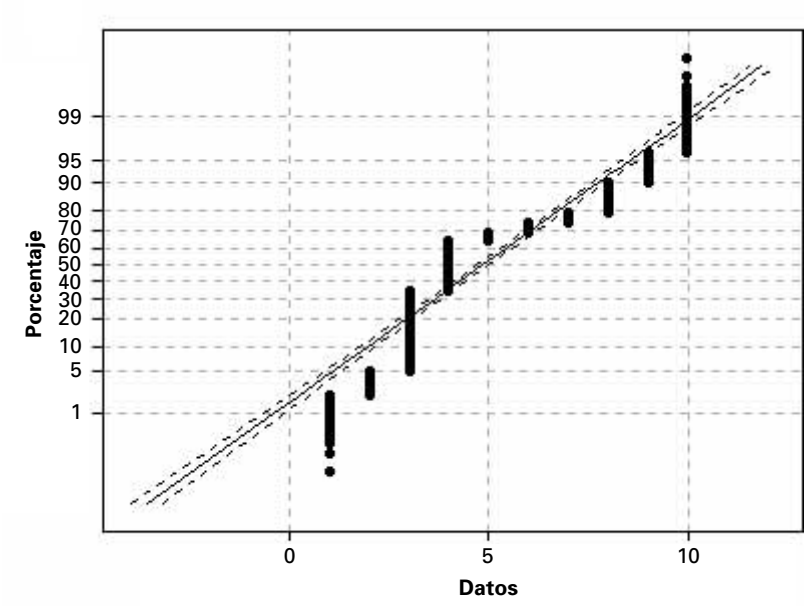


Figura 6.22. Gráfico de probabilidades normales de una distribución muy sesgada (salida Minitab).

de la Figura 6.19, se puede suponer sin riesgo a equivocarse que el modelo normal es una buena aproximación. Obsérvese, sin embargo, que si se desvía de una línea recta, como ocurre en las Figuras 6.20 y 6.22, no debe utilizarse la distribución normal.

EJERCICIOS

Ejercicios básicos

6.17. Suponga que la variable aleatoria Z sigue una distribución normal estándar.

a) Halle $P(Z < 1{,}20)$
b) Halle $P(Z > 1{,}33)$
c) Halle $P(Z < -1{,}70)$
d) Halle $P(Z > -1{,}00)$
e) Halle $P(1{,}20 < Z < 1{,}33)$
f) Halle $P(-1{,}70 < Z < 1{,}20)$
g) Halle $P(-1{,}70 < Z < -1{,}00)$

6.18. Suponga que la variable aleatoria Z sigue una distribución normal estándar.

a) La probabilidad de que Z sea inferior a _____ es 0,70.
b) La probabilidad de que Z sea inferior a _____ es 0,25.
c) La probabilidad de que Z sea superior a _____ es 0,2.
d) La probabilidad de que Z sea superior a _____ es 0,6.

6.19. Suponga que la variable aleatoria X sigue una distribución normal que tiene una $\mu = 50$ y una $\sigma^2 = 64$.

a) Halle la probabilidad de que X sea superior a 60.
b) Halle la probabilidad de que X sea superior a 35 e inferior a 62.
c) Halle la probabilidad de que X sea inferior a 55.
d) La probabilidad de que X sea superior a _____ es 0,2.
e) La probabilidad de que X esté en el intervalo simétrico en torno a la media entre _____ y _____ es 0,05.

6.20. Suponga que la variable aleatoria X sigue una distribución normal que tiene una $\mu = 80$ y una $\sigma^2 = 100$.

a) Halle la probabilidad de que X sea superior a 60.
b) Halle la probabilidad de que X sea superior a 72 e inferior a 82.
c) Halle la probabilidad de que X sea inferior a 55.
d) La probabilidad de que X sea superior a _____ es 0,1.
e) La probabilidad de que X esté en el intervalo simétrico en torno a la media entre _____ y _____ es 0,08.

6.21. Suponga que la variable aleatoria X sigue una distribución normal que tiene una $\mu = 0{,}2$ y una $\sigma^2 = 0{,}0025$.

a) Halle la probabilidad de que X sea superior a 0,4.
b) Halle la probabilidad de que X sea superior a 0,15 e inferior a 0,28.
c) Halle la probabilidad de que X sea inferior a 0,10.
d) La probabilidad de que X sea superior a _____ es 0,2.
e) La probabilidad de que X esté en el intervalo simétrico en torno a la media entre _____ y _____ es 0,05.

Ejercicios aplicados

6.22. Se sabe que la cantidad de dinero que gastan los estudiantes en libros de texto en un año en una universidad sigue una distribución normal que tiene una media de 380 $ y una desviación típica de 50 $.

a) ¿Cuál es la probabilidad de que un estudiante elegido aleatoriamente gaste menos de 400 $ en libros de texto en un año?
b) ¿Cuál es la probabilidad de que un estudiante elegido aleatoriamente gaste más de 360 $ en libros de texto en un año?
c) Explique gráficamente por qué las respuestas de los apartados (a) y (b) son iguales.
d) ¿Cuál es la probabilidad de que un estudiante elegido aleatoriamente gaste entre 300 $ y 400 $ en libros de texto en un año?
e) Quiere hallar un intervalo de gasto en libros de texto que incluya el 80 por ciento de todos los estudiantes de esta universidad. Explique por qué podría encontrarse cualquier número de intervalos que lo incluya y halle el más corto.

6.23. La demanda de consumo de un producto prevista para el próximo mes puede representarse por medio de una variable aleatoria normal que tiene una media de 1.200 unidades y una desviación típica de 100 unidades.

a) ¿Cuál es la probabilidad de que las ventas superen las 1.000 unidades?
b) ¿Cuál es la probabilidad de que las ventas se encuentren entre 1.100 y 1.300 unidades?
c) La probabilidad de que las ventas sean de más de _____ unidades es de 0,10.

6.24. La duración de una determinada marca de neumáticos sigue una distribución normal que tiene una media de 35.000 kilómetros y una desviación típica de 4.000 kilómetros.

a) ¿Qué proporción de estos neumáticos tiene una duración de más de 38.000 kilómetros?
b) ¿Qué proporción de estos neumáticos tiene una duración de menos de 38.000 kilómetros?
c) ¿Qué proporción de estos neumáticos tiene una duración de entre 32.000 y 38.000 kilómetros?
d) Represente gráficamente la función de densidad de las duraciones mostrando:
i. Por qué las respuestas de los apartados (a) y (b) son iguales.
ii. Por qué las respuestas de los apartados (a), (b) y (c) suman 1.

6.25. Una cartera de inversión contiene acciones de un gran número de empresas. El año pasado, las tasas de rendimiento de estas acciones siguieron una distribución normal que tenía una media de 12,2 por ciento y una desviación típica de 7,2 por ciento.

a) ¿De qué proporción de estas empresas fue la tasa de rendimiento de más del 20 por ciento?
b) ¿De qué proporción de estas empresas fue la tasa de rendimiento negativa?
c) ¿De qué proporción de estas empresas fue la tasa de rendimiento de entre el 5 y el 15 por ciento?

6.26. Una empresa produce sacos de un producto químico y le preocupa la cantidad de impurezas que contienen. Se cree que el peso de las impurezas por saco sigue una distribución normal que tiene una media de 12,2 gramos y una desviación típica de 2,8 gramos. Se elige aleatoriamente un saco.

a) ¿Cuál es la probabilidad de que contenga menos de 10 gramos de impurezas?
b) ¿Cuál es la probabilidad de que contenga más de 15 gramos de impurezas?
c) ¿Cuál es la probabilidad de que contenga entre 12 y 15 gramos de impurezas?
d) Es posible deducir, sin realizar los cálculos detallados, cuál de las respuestas a los apartados (a) y (b) es mayor. ¿Cómo?

6.27. Un contratista considera que el coste de cumplir un contrato es una variable aleatoria que sigue una distribución normal que tiene una media de 500.000 $ y una desviación típica de 50.000 $.

a) ¿Cuál es la probabilidad de que el coste de cumplir el contrato esté entre 460.000 $ y 540.000 $?

b) La probabilidad de que el coste de cumplir el contrato cueste menos de _____ es 0,2.
c) Halle el intervalo más corto tal que la probabilidad de que el coste de cumplir el contrato esté en este intervalo sea 0,95.

6.28. Las calificaciones de un examen siguen una distribución normal. ¿Cuál es la probabilidad de que un estudiante seleccionado aleatoriamente obtenga una calificación mayor que la media más de 1,5 desviaciones típicas?

6.29. Se va a estrenar una nueva serie de televisión. Un ejecutivo de la cadena cree que su incertidumbre sobre el índice de audiencia que tendrá este programa durante el primer mes puede representarse por medio de una distribución normal que tiene una media de 18,2 y una desviación típica de 1,6. Según este ejecutivo, la probabilidad de que la audiencia sea de menos de _____ es 0,1.

6.30. Un ejecutivo de una cadena de televisión está revisando las perspectivas de una nueva serie televisiva. En su opinión, la probabilidad de que la serie tenga una audiencia de más de 17,8 es 0,25 y la probabilidad de que tenga una audiencia de más de 19,2 es 0,15. Si la incertidumbre del ejecutivo sobre la audiencia puede representarse por medio de una distribución normal, ¿cuáles son la media y la varianza de esa distribución?

6.31. Las calificaciones de un examen realizado por un gran número de estudiantes siguen una distribución normal que tiene una media de 700 y una desviación típica de 120.

a) Se concede un sobresaliente por una calificación de más de 820. ¿Qué proporción de todos los estudiantes obtiene un sobresaliente?
b) Se concede un notable por las calificaciones comprendidas entre 730 y 820. Un profesor tiene un subgrupo de 100 estudiantes que puede considerarse que son una muestra aleatoria de todos los estudiantes del grupo grande. Halle el número esperado de estudiantes de este grupo pequeño que obtendrán un notable.
c) Se decide suspender al 5 por ciento de los estudiantes que tienen las calificaciones más bajas. ¿Cuál es la calificación mínima necesaria para evitar el suspenso?

6.32. Estoy considerando dos inversiones distintas. No estoy seguro en ninguno de los dos casos del rendimiento porcentual, pero creo que mi incertidumbre puede representarse por medio de distribuciones normales que tienen las medias y las desviaciones típicas mostradas en la tabla adjunta. Quiero hacer la inversión que tenga más probabilidades de generar un rendimiento de al menos un 10 por ciento. ¿Cuál debo elegir?

	Media	Desviación típica
Inversión A	10,4	1,2
Inversión B	11,0	4,0

6.33. Una empresa puede comprar una materia prima a dos proveedores y le preocupa la cantidad de impurezas que contiene. El examen de los datos de cada proveedor indica que los niveles porcentuales de impurezas de los envíos de la materia prima recibidos siguen distribuciones normales que tienen las medias y las desviaciones típicas indicadas en la tabla adjunta. La empresa tiene especial interés en que el nivel de impurezas de un envío no supere el 5 por ciento y quiere comprar al proveedor que tenga más probabilidades de cumplir esa condición. ¿Qué proveedor debe elegir?

	Media	Desviación típica
Proveedor A	4,4	0,4
Proveedor B	4,2	0,6

6.34. Un profesor ha observado que el tiempo que dedican los estudiantes a hacer un trabajo de curso sigue una distribución normal que tiene una media de 150 minutos y una desviación típica de 40 minutos.

a) La probabilidad de que un estudiante elegido aleatoriamente dedique más de _____ minutos a este trabajo es 0,9.
b) La probabilidad de que un estudiante elegido aleatoriamente dedique menos de _____ minutos a este trabajo es 0,8.
c) Se eligen aleatoriamente dos estudiantes. ¿Cuál es la probabilidad de que al menos uno de ellos dedique al menos 2 horas a este trabajo?

6.35. Una empresa se dedica a reparar fotocopiadoras. El examen de sus registros muestra que el tiempo que tarda en hacer una reparación puede representarse por medio de una variable aleatoria normal que tiene una media de 75 minutos y una desviación típica de 20 minutos.

a) ¿Qué proporción de reparaciones lleva menos de 1 hora?
b) ¿Qué proporción de reparaciones lleva más de 90 minutos?

c) Explique gráficamente por qué las respuestas de los apartados (a) y (b) son iguales.
d) La probabilidad de que una reparación lleve más de _____ minutos es de 0,1.

6.36. Se sabe que las calificaciones de un examen siguen una distribución normal que tiene una media de 420 y una desviación típica de 80.
a) ¿Cuál es la probabilidad de que una persona elegida aleatoriamente obtenga una calificación de entre 400 y 480?
b) ¿Cuál es la calificación mínima necesaria para estar en el 10 por ciento superior de todas las personas que realizan el examen?
c) Indique, sin realizar los cálculos, en cuál de los intervalos siguientes es más probable que se encuentre la calificación de una persona elegida aleatoriamente: 400-439, 440-479, 480-519 o 520-559.
d) ¿En cuál de los intervalos enumerados en el apartado (c) es menos probable que se encuentre la calificación de esta persona?
e) Se eligen aleatoriamente dos personas que realizan el examen. ¿Cuál es la probabilidad de que al menos una de ellas tenga una calificación de más de 500 puntos?

6.37. Se estima que el tiempo que está una conocida banda de rock, Living Ingrates, en el escenario en sus conciertos sigue una distribución normal que tiene una media de 200 minutos y una desviación típica de 20 minutos.
a) ¿Qué proporción de conciertos de esta banda dura entre 180 y 200 minutos?
b) Uno de los espectadores introduce a escondidas en un concierto de Living Ingrates una grabadora con cintas que tienen una capacidad de 245 minutos. ¿Cuál es la probabilidad de que esta capacidad sea insuficiente para grabar todo el concierto?
c) Si la desviación típica de la duración de los conciertos fuera de 15 minutos solamente, indique, sin realizar los cálculos, si la probabilidad de que un concierto dure más de 245 minutos es mayor, menor o igual que la que ha calculado en el apartado (b). Represente gráficamente su respuesta.
d) La probabilidad de que un concierto de Living Ingrates dure menos de _____ minutos es 0,1. Suponga como antes que la desviación típica poblacional es de 20 minutos.

6.38. Un numeroso grupo de estudiantes realiza un examen de economía. Las calificaciones siguen una distribución normal que tiene una media de 70 y la probabilidad de que un estudiante elegido aleatoriamente obtenga una calificación de menos de 85 es de 0,9332. Se eligen aleatoriamente cuatro estudiantes. ¿Cuál es la probabilidad de que al menos uno de ellos tenga una calificación de más de 80 puntos en este examen?

6.4. La distribución normal como aproximación de la distribución binomial

En este apartado mostramos cómo puede utilizarse la distribución normal como aproximación de las variables aleatorias discretas binomiales y proporcionales que se emplean frecuentemente en el mundo empresarial y en la economía. Esta aproximación puede utilizarse para calcular las probabilidades de muestras de mayor tamaño cuando no es fácil disponer de tablas. La distribución normal como aproximación de la distribución binomial también es útil para resolver problemas aplicados. Vemos que los métodos basados en la distribución normal también pueden utilizarse en problemas en los que hay variables aleatorias binomiales y proporcionales. Por lo tanto, es posible reducir el número de métodos estadísticos que es necesario aprender para resolver problemas empresariales.

Examinemos un problema con n pruebas independientes, cada una de las cuales tiene una probabilidad de éxito P. En el apartado 5.4 vimos que la variable aleatoria binomial X podía expresarse por medio de la suma de n variables aleatorias de Bernoulli independientes:

$$X = X_1 + X_2 + \cdots + X_n$$

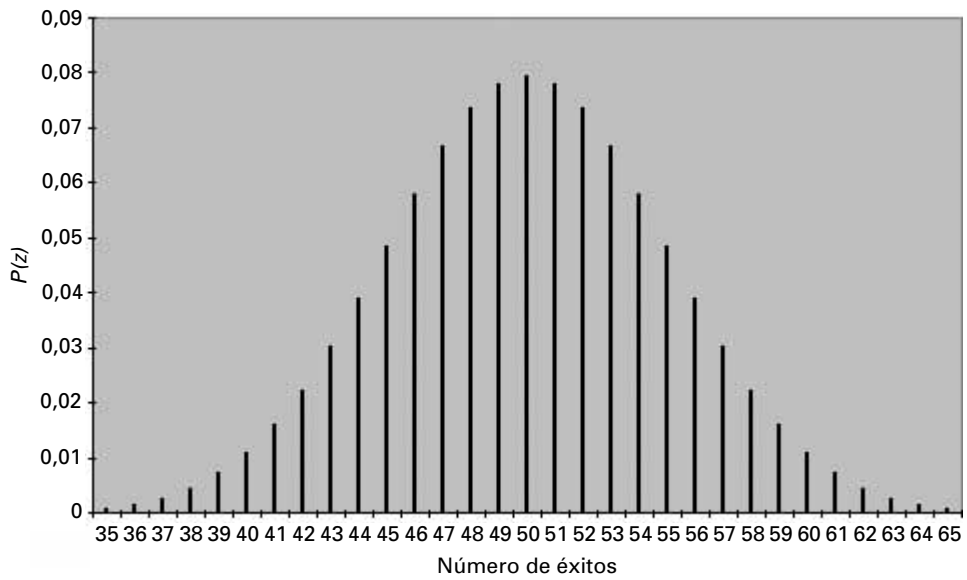


Figura 6.23. Una distribución binomial en la que $n = 100$ y $P = 0{,}50$.

donde la variable aleatoria X_i toma el valor 1 si el resultado de la i-ésima prueba es un «éxito» y 0 en caso contrario, con las probabilidades respectivas P y $1 - P$. El número X de éxitos resultante sigue una distribución binomial de media y varianza

$$E(X) = \mu = nP$$

$$\text{Var}\,(X) = \sigma^2 = nP(1 - P)$$

La representación de una distribución binomial cuando $P = 0{,}5$ y $n = 100$ de la Figura 6.23 nos muestra que la distribución binomial tiene la misma forma que la normal. Esta evidencia visual de que la distribución binomial puede aproximarse con una distribución normal de la misma media y la misma varianza también ha sido demostrada por estadísticos matemáticos. Una buena regla para nosotros es que la distribución normal es una buena aproximación de la distribución binomial cuando $nP(1 - P) > 9$.

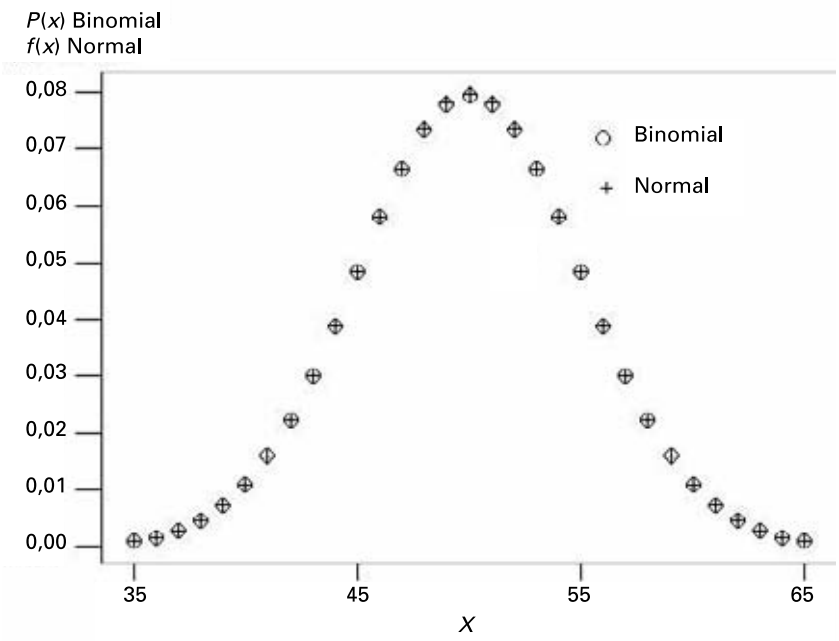
Para comprender mejor la aproximación de la distribución binomial por medio de la distribución normal, consideremos las Figuras 6.24(a) y (b). Tanto en (a) como en (b), mostramos puntos de una función de densidad normal comparados con las probabilidades correspondientes de una distribución binomial utilizando gráficos realizados con el programa Minitab. En la parte (a), observamos que el valor de la regla de aproximación es

$$nP(1 - P) = 100(0{,}5)(1 - 0{,}5) = 25 > 9$$

y que la distribución normal es una buena aproximación de la distribución binomial. En cambio, en el ejemplo de la parte (b) el valor de la regla de aproximación es

$$nP(1 - P) = 25(0{,}2)(1 - 0{,}2) = 4 < 9$$

y la distribución normal no es una buena aproximación de la distribución binomial. La evidencia como la que contiene la Figura 6.24 es la razón por la que se utiliza mucho la distribución normal como aproximación de la distribución binomial. A continuación, explicamos el método para aplicarla.



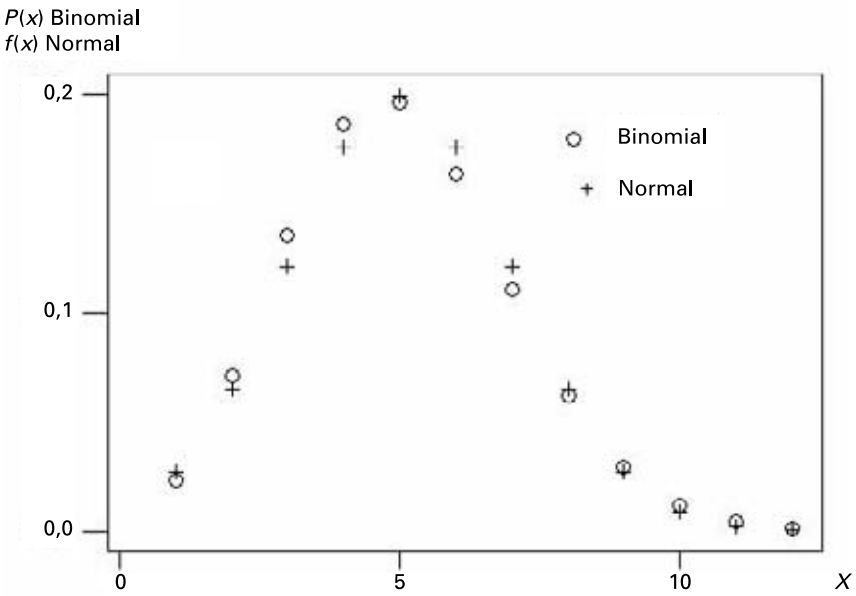


Figura 6.24. Comparación de las aproximaciones binomial y normal (salida Minitab): (a) Binomial en la que $P = 0{,}50$ y $n = 100$ y normal de $\mu = 50$ y $\sigma = 5$. (b) Binomial en la que $P = 0{,}20$ y $n = 25$ y normal de $\mu = 5$ y $\sigma = 2$.

Utilizando la media y la varianza de la distribución binomial, observamos que si el número de pruebas n es grande —tal que $nP(1 - P) > 9$— la distribución de la variable aleatoria

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}\,(X)}} = \frac{X - nP}{\sqrt{nP(1 - P)}}$$

es aproximadamente normal estándar.

Este resultado es muy importante, porque nos permite hallar, cuando n es grande, la probabilidad de que el número de éxitos se encuentre dentro de un intervalo dado. Si queremos hallar la probabilidad de que el número de éxitos se encuentre entre a y b, inclusive, tenemos que

$$P(a \leqslant X \leqslant b) = P\left(\frac{a - nP}{\sqrt{nP(1-P)}} \leqslant \frac{X - nP}{\sqrt{nP(1-P)}} \leqslant \frac{b - nP}{\sqrt{nP(1-P)}}\right)$$
$$= P\left(\frac{a - nP}{\sqrt{nP(1-P)}} \leqslant Z \leqslant \frac{b - nP}{\sqrt{nP(1-P)}}\right)$$

Cuando n es grande, la normal estándar es una buena aproximación de Z y podemos hallar la probabilidad utilizando los métodos del apartado 6.3.

EJEMPLO 6.8. Ventas a clientes (probabilidades normales)

Una vendedora se pone en contacto por teléfono con posibles clientes en un intento de averiguar si es probable que merezca la pena ir a su casa a verlos. Su experiencia sugiere que en el 40 por ciento de los contactos iniciales acaba yendo a casa del cliente. Si se pone en contacto con 100 personas por teléfono, ¿cuál es la probabilidad de que vaya a ver a entre 45 y 50 personas?

Solución

Sea X el número de personas a las que va a ver la vendedora. Entonces, X tiene una distribución binomial tal que $n = 100$ y $P = 0{,}40$. Aproximando la probabilidad que buscamos, tenemos que

$$P(45 \leqslant X \leqslant 50) \cong P\left(\frac{45 - (100)(0{,}4)}{\sqrt{(100)(0{,}4)(0{,}6)}} \leqslant Z \leqslant \frac{50 - (100)(0{,}4)}{\sqrt{(100)(0{,}4)(0{,}6)}}\right)$$
$$= P(1{,}02 \leqslant Z \leqslant 2{,}04)$$
$$= F(2{,}04) - F(1{,}02)$$
$$= 0{,}9793 - 0{,}8461 = 0{,}1332$$

Esta probabilidad está representada por el área situada debajo de la curva normal estándar de la Figura 6.25.

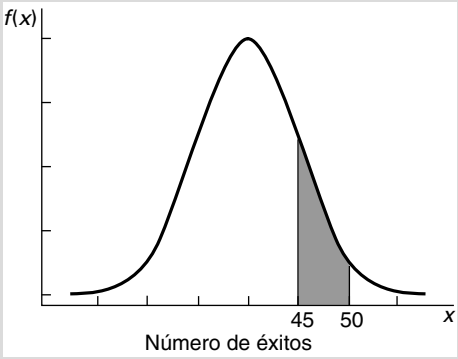


Figura 6.25. Probabilidad de conseguir entre 45 y 50 éxitos en una distribución binomial en la que $n = 100$ y $P = 0{,}4$.

Variable aleatoria proporcional

En algunos problemas aplicados, tenemos que calcular probabilidades de intervalos proporcionales o porcentuales. Podemos calcularlas utilizando una extensión directa de la aproximación de la distribución binomial por medio de la distribución normal. Una variable aleatoria proporcional, P, puede calcularse dividiendo el número de éxitos, X, por el tamaño de la muestra, n.

$$P = \frac{X}{n}$$

Utilizando la transformación lineal de variables aleatorias, podemos calcular la media y la varianza de P de la forma siguiente:

$$\mu = P$$

$$\sigma^2 = \frac{P(1 - P)}{n}$$

Podemos utilizar la media y la varianza resultantes con la distribución normal para calcular la probabilidad deseada.

EJEMPLO 6.9. Predicciones electorales (probabilidades proporcionales)

A menudo hemos observado el éxito de las cadenas de televisión en la predicción de los resultados electorales. Éste es un buen ejemplo del fructífero uso de los métodos de probabilidad en los problemas aplicados. Veamos cómo pueden predecirse los resultados electorales utilizando muestras relativamente pequeñas en un ejemplo simplificado. Un experto en predicciones electorales ha obtenido una muestra aleatoria de 900 votantes, en la que 500 declaran que votarán a Susana Cinca. ¿Debe prever Susana que ganará las elecciones?

Solución

En este problema suponemos que sólo hay dos candidatos y, por lo tanto, si más del 50 por ciento de la población apoya a Susana, ésta ganará las elecciones. Calculamos la probabilidad de que 500 votantes o más de una muestra de 900 apoyen a Susana suponiendo que exactamente el 50 por ciento, $P = 0{,}50$, de toda la población apoya a Susana.

$$\begin{aligned} P(X \geqslant 500 \,|\, n = 900,\ P = 0{,}50) &\approx P(X \geqslant 500 \,|\, \mu = 450,\ \sigma^2 = 225) \\ &= P\left(Z \geqslant \frac{500 - 450}{\sqrt{225}}\right) \\ &= P(Z \geqslant 3{,}33) \\ &= 0{,}000 \end{aligned}$$

La probabilidad de tener 500 éxitos en 900 pruebas si $P = 0{,}50$ es muy pequeña y, por lo tanto, concluimos que P debe ser mayor de 0,50. Por lo tanto, predecimos que Susana Cinca ganará las elecciones.

También podríamos calcular la probabilidad de que más del 55,6 por ciento (500/900) de la muestra declare su apoyo a Susana si la proporción correspondiente a la población es $P = 0{,}50$. Utilizando la media y la varianza de variables aleatorias proporcionales

$$\mu = P = 0{,}50$$

$$\sigma^2 = \frac{P(1-P)}{n} = \frac{0{,}50(1-0{,}50)}{900}$$

$$\sigma = 0{,}0167$$

$$P(P \geqslant 0{,}556 \,|\, n = 900,\ P = 0{,}50) \approx P(P \geqslant 0{,}556 \,|\, \mu = 0{,}50,\ \sigma = 0{,}0167)$$

$$= P\left(Z \geqslant \frac{0{,}556 - 0{,}50}{0{,}0167}\right)$$

$$= P(Z \geqslant 3{,}33)$$

$$= 0{,}000$$

Obsérvese que la probabilidad es exactamente igual que la de la variable aleatoria binomial correspondiente. Eso siempre es así porque cada valor proporcional o porcentual está relacionado directamente con un número específico de éxitos. Como el término porcentaje es más frecuente que el término proporción en el lenguaje empresarial y económico, tenderemos a utilizarlo más a menudo en los ejercicios y los análisis de este libro de texto.

EJERCICIOS

Ejercicios básicos

6.39. Dada una muestra aleatoria de tamaño $n = 900$ de una distribución de probabilidad binomial en la que $P = 0{,}50$,

a) Halle la probabilidad de que el número de éxitos sea superior a 500.
b) Halle la probabilidad de que el número de éxitos sea inferior a 430.
c) Halle la probabilidad de que el número de éxitos esté entre 440 y 480.
d) El número de éxitos es inferior a _____ con una probabilidad de 0,10.
e) El número de éxitos es superior a _____ con una probabilidad de 0,08.

6.40. Dada una muestra aleatoria de tamaño $n = 1.600$ de una distribución de probabilidad binomial en la que $P = 0{,}40$,

a) Halle la probabilidad de que el número de éxitos sea superior a 1.650.
b) Halle la probabilidad de que el número de éxitos sea inferior a 1.530.
c) Halle la probabilidad de que el número de éxitos esté entre 1.550 y 1.650.
d) El número de éxitos es inferior a _____ con una probabilidad de 0,09.
e) El número de éxitos es superior a _____ con una probabilidad de 0,20.

6.41. Dada una muestra aleatoria de tamaño $n = 900$ de una distribución de probabilidad binomial en la que $P = 0{,}10$,

a) Halle la probabilidad de que el número de éxitos sea superior a 110.
b) Halle la probabilidad de que el número de éxitos sea inferior a 53.
c) Halle la probabilidad de que el número de éxitos esté entre 55 y 120.
d) El número de éxitos es inferior a _____ con una probabilidad de 0,10.
e) El número de éxitos es superior a _____ con una probabilidad de 0,08.

6.42. Dada una muestra aleatoria de tamaño $n = 1.600$ de una distribución de probabilidad binomial en la que $P = 0{,}40$,

a) Halle la probabilidad de que el porcentaje de éxitos sea superior a 0,45.
b) Halle la probabilidad de que el porcentaje de éxitos sea inferior a 0,36.
c) Halle la probabilidad de que el porcentaje de éxitos esté entre 0,37 y 0,44.
d) El porcentaje de éxitos es inferior a _____ con una probabilidad de 0,20.
e) El porcentaje de éxitos es superior a _____ con una probabilidad de 0,09.

6.43. Dada una muestra aleatoria de tamaño $n = 400$ de una distribución de probabilidad binomial en la que $P = 0{,}20$,
a) Halle la probabilidad de que el porcentaje de éxitos sea superior a 0,25.
b) Halle la probabilidad de que el porcentaje de éxitos sea inferior a 0,16.
c) Halle la probabilidad de que el porcentaje de éxitos esté entre 0,17 y 0,24.
d) El porcentaje de éxitos es inferior a _____ con una probabilidad de 0,15.
e) El porcentaje de éxitos es superior a _____ con una probabilidad de 0,11.

Ejercicios aplicados

6.44. Una compañía de alquiler de automóviles ha observado que la probabilidad de que un automóvil necesite una reparación en un mes cualquiera dado es 0,2. La compañía tiene 900 automóviles.
a) ¿Cuál es la probabilidad de que más de 200 automóviles necesiten una reparación en un mes determinado?
b) ¿Cuál es la probabilidad de que menos de 175 automóviles necesiten una reparación en un mes determinado?

6.45. Se sabe que el 10 por ciento de todos los artículos que salen de un determinado proceso de producción tiene un defecto. Se eligen aleatoriamente 400 artículos de un elevado volumen de producción de un día.
a) ¿Cuál es la probabilidad de que al menos 35 de los artículos seleccionados tenga un defecto?
b) ¿Cuál es la probabilidad de que entre 40 y 50 de los artículos seleccionados tenga un defecto?
c) ¿Cuál es la probabilidad de que entre 34 y 48 de los artículos seleccionados tenga un defecto?
d) Sin realizar los cálculos, indique cuál de los siguientes intervalos de artículos defectuosos tiene la probabilidad más alta: 38-39, 40-41, 42-43, 44-45, 46-47.

6.46. Se encuesta a una muestra de 100 obreros de una gran empresa para saber qué piensan de un nuevo plan de trabajo propuesto. Si el 60 por ciento de todos los obreros de esta empresa es partidario de este nuevo plan, ¿cuál es la probabilidad de que menos de 50 de los miembros de la muestra sea partidario del plan?

6.47. Un hospital observa que el 25 por ciento de sus facturas tienen al menos 1 mes de retraso. Se toma una muestra aleatoria de 450 facturas.
a) ¿Cuál es la probabilidad de que menos de 100 facturas de la muestra tenga al menos 1 mes de retraso?
b) ¿Cuál es la probabilidad de que el número de facturas de la muestra que tienen al menos 1 mes de retraso esté entre 120 y 150 (inclusive)?

6.48. La duración de una marca de neumáticos puede representarse (como en el ejercicio 6.24) por medio de una distribución normal que tiene una media de 35.000 km y una desviación típica de 4.000 km. Se toma una muestra de 100 neumáticos. ¿Cuál es la probabilidad de que más de 25 tengan una duración de más de 38.000 km?

6.49. Los sacos de un producto químico de una empresa tienen un peso de impurezas que puede representarse por medio de una distribución normal que tiene una media de 12,2 gramos y una desviación típica de 2,8 gramos. Se toma una muestra aleatoria de 400 de estos sacos. ¿Cuál es la probabilidad de que al menos 100 contengan menos de 10 gramos de impurezas?

6.5. La distribución exponencial

A continuación, introducimos una distribución continua, la *distribución exponencial*, que se ha observado que es especialmente útil para resolver problemas de listas de espera o colas. En muchos problemas sobre el tiempo que se dedica a la realización de un servi-

cio, éste puede representarse por medio de una distribución exponencial. Debemos señalar que la distribución exponencial se diferencia de la normal en dos importantes aspectos: se limita a las variables aleatorias que tienen valores positivos y su distribución no es simétrica.

La distribución exponencial

La variable aleatoria exponencial $T(t > 0)$ tiene una función de densidad

$$f(t) = \lambda e^{-\lambda t} \quad \text{para } t > 0 \tag{6.15}$$

donde λ es el número medio de ocurrencias por unidad de tiempo, t es el número de unidades de tiempo hasta la siguiente ocurrencia y $e = 2{,}71828...$ Se dice que T sigue una **distribución de probabilidad exponencial**. Puede demostrarse que λ es el mismo parámetro utilizado para la distribución de Poisson en el apartado 5.6 y que el tiempo medio entre las ocurrencias es $1/\lambda$.

La función de distribución acumulada es

$$F(t) = 1 - e^{-\lambda t} \quad \text{para } t > 0 \tag{6.16}$$

La distribución tiene una media de $1/\lambda$ y una varianza de $1/\lambda^2$.

La variable aleatoria T puede utilizarse para representar el tiempo que transcurre hasta que se termina de realizar un servicio o hasta la siguiente llegada a un proceso de cola, comenzando en un tiempo arbitrario 0. Los supuestos del modelo son iguales que los de la distribución de Poisson. Obsérvese que la distribución de Poisson indica la probabilidad de que haya X éxitos o llegadas durante una unidad de tiempo. En cambio, la distribución exponencial indica la probabilidad de que haya un éxito o una llegada durante un intervalo de tiempo t. La Figura 6.26 muestra la función de densidad de una distribución exponencial que tiene una $\lambda = 0{,}2$. El área situada a la izquierda de 10 indica la probabilidad de que una tarea se realice antes del tiempo 10. Esta área puede hallarse evaluando la función $1 - e^{-\lambda t}$ para el valor dado de $t = 10$. La función puede calcularse por medio de una calculadora electrónica. Veamos ahora un ejemplo para mostrar la aplicación de la distribución exponencial.

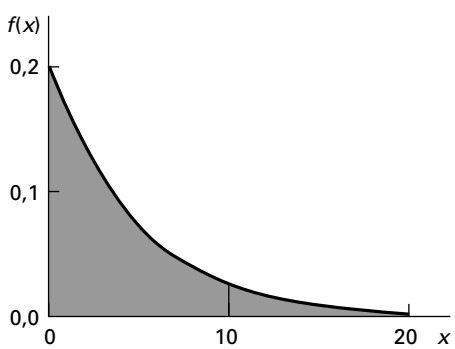


Figura 6.26. Función de densidad de una distribución exponencial en la que $\lambda = 0{,}2$.

EJEMPLO 6.10. Tiempo que se dedica a atender al público en el mostrador de información de una biblioteca (probabilidades exponenciales)

El tiempo que se dedica a atender al público en el mostrador de información de una biblioteca puede representarse por medio de una distribución exponencial que tiene un tiempo medio de atención de 5 minutos. ¿Cuál es la probabilidad de que el tiempo de atención al público sea de más de 10 minutos?

Solución

Sea t el tiempo de atención en minutos. La tasa de atención es $\lambda = 1/5 = 0{,}2$ por minuto y la función de densidad es

$$f(t) = \lambda e^{-\lambda t}$$

que se muestra en la Figura 6.26. La probabilidad que buscamos puede calcularse de la forma siguiente:

$$\begin{aligned} P(T > 10) &= 1 - P(T < 10) \\ &= 1 - F(10) \\ &= 1 - (1 - e^{-(0{,}20)(10)}) \\ &= e^{-2{,}0} = 0{,}1353 \end{aligned}$$

Por lo tanto, la probabilidad de que el tiempo de atención sea de más de 10 minutos es 0,1353.

EJEMPLO 6.11. Tiempo que transcurre entre los accidentes en las fábricas británicas representativas (probabilidades exponenciales)

En Gran Bretaña, una fábrica de 2.000 asalariados tiene un número semanal medio de accidentes con baja igual a $\lambda = 0{,}4$ y el número de accidentes sigue una distribución de Poisson. ¿Cuál es la probabilidad de que el tiempo que transcurre entre los accidentes sea de menos de 2 semanas?

Solución

En este problema, señalamos que el intervalo de tiempo se mide en semanas y nuestra tasa es $\lambda = 0{,}4$ a la semana, lo que da un tiempo medio entre accidentes de $\mu = 1/(0{,}4) = 2{,}5$ semanas. Entonces, la probabilidad de que el tiempo que transcurre entre accidentes sea de menos de 2 semanas es

$$\begin{aligned} P(T < 2) = F(2) &= 1 - e^{-(0{,}4)(2)} \\ &= 1 - e^{-0{,}8} \\ &= 1 - 0{,}4493 = 0{,}5507 \end{aligned}$$

Por lo tanto, la probabilidad de que transcurran menos de 2 semanas entre los accidentes es de alrededor del 55 por ciento.

EJERCICIOS

Ejercicios básicos

6.50. Dado un proceso de llegada en el que $\lambda = 1{,}0$, ¿cuál es la probabilidad de que se produzca una llegada en las primeras $t = 2$ unidades de tiempo?

6.51. Dado un proceso de llegada en el que $\lambda = 8{,}0$, ¿cuál es la probabilidad de que se produzca una llegada en las primeras $t = 7$ unidades de tiempo?

6.52. Dado un proceso de llegada en el que $\lambda = 5{,}0$, ¿cuál es la probabilidad de que se produzca una llegada en las primeras $t = 7$ unidades de tiempo?

6.53. Dado un proceso de llegada en el que $\lambda = 6{,}0$, ¿cuál es la probabilidad de que se produzca una llegada en las primeras $t = 5$ unidades de tiempo?

6.54. Dado un proceso de llegada en el que $\lambda = 3{,}0$, ¿cuál es la probabilidad de que se produzca una llegada en las primeras $t = 2$ unidades de tiempo?

Ejercicios aplicados

6.55. Un profesor atiende a los estudiantes durante las horas normales de despacho. El tiempo que dedica a los estudiantes sigue una distribución exponencial que tiene una media de 10 minutos.

a) Halle la probabilidad de que un estudiante dado pase menos de 20 minutos con el profesor.

b) Halle la probabilidad de que un estudiante dado pase más de 5 minutos con el profesor.

c) Halle la probabilidad de que un estudiante dado pase entre 10 y 15 minutos con el profesor.

6.56. El tiempo que se tarda en recoger información preliminar sobre los pacientes que entran en una clínica sigue una distribución exponencial que tiene una media de 15 minutos. Halle la probabilidad de que se tarde más de 18 minutos en el caso de un paciente elegido aleatoriamente.

6.57. Se sabe que el número de fallos que experimenta el sistema informático de un laboratorio durante un mes sigue una distribución de Poisson que tiene una media de 0,8. El sistema acaba de fallar. Halle la probabilidad de que pasen al menos 2 meses antes de que falle de nuevo.

6.58. Suponga que el tiempo que transcurre entre sucesivas ocurrencias de un suceso sigue una distribución exponencial que tiene una media de $1/\lambda$ minutos. Suponga que ocurre un suceso.

a) Demuestre que la probabilidad de que transcurran más de 3 minutos antes de la ocurrencia del siguiente suceso es $e^{-3\lambda}$.

b) Demuestre que la probabilidad de que transcurran más de 6 minutos antes de la ocurrencia del siguiente suceso es $e^{-6\lambda}$.

c) Utilizando los resultados de los apartados (a) y (b), demuestre que si ya han transcurrido 3 minutos, la probabilidad de que transcurran otros 3 antes de la siguiente ocurrencia es $e^{-3\lambda}$. Explique su respuesta en palabras.

6.6. Distribución conjunta de variables aleatorias continuas

En el apartado 5.7 introdujimos las distribuciones conjuntas de variables aleatorias discretas. Aquí mostramos que muchos de los conceptos y los resultados de las variables aleatorias discretas también se aplican a las variables aleatorias continuas. Muchas variables aleatorias continuas pueden representarse utilizando variables aleatorias que siguen una distribución conjunta. Los valores de mercado de los precios de varias acciones se representan normalmente como variables aleatorias conjuntas. En los estudios de las pautas de producción y de ventas de varias empresas e industrias se utilizan variables aleatorias continuas que siguen una distribución conjunta. El número de unidades vendidas por unos grandes almacenes durante una semana y el precio por unidad pueden representarse por medio de variables aleatorias conjuntas. En los estudios sobre la conducta de las importaciones y de las exportaciones de varios países normalmente se utilizan variables aleatorias conjuntas.

Después de presentar algunos conceptos básicos, ponemos algunos ejemplos para mostrar la importancia de los métodos y ver cómo se analizan las variables aleatorias continuas que siguen una distribución conjunta.

Función de distribución acumulada conjunta

Sean X_1, X_2, ..., X_k variables aleatorias continuas.

1. Su **función de distribución acumulada conjunta**, $F(x_1, x_2, ..., x_k)$ define la probabilidad de que simultáneamente X_1 sea menor que x_1, X_2 sea menor que x_2, y así sucesivamente; es decir,

$$F(x_1, x_2, ..., x_k) = P(X_1 < x_1 \cap X_2 < x_2 \cap \cdots \cap X_k < x_k) \quad \textbf{(6.17)}$$

2. Las funciones de distribución acumulada —$F(x_1)$, $F(x_2)$, ..., $F(x_k)$— de las variables aleatorias individuales se llaman **funciones de distribución marginal**. Para cualquier i, $F(x_i)$ es la probabilidad de que la variable aleatoria X_i no sea mayor que el valor específico x_i.
3. Las variables aleatorias son *independientes* si y sólo si

$$F(x_1, x_2, ..., x_k) = F(x_1)F(x_2) \cdots F(x_k) \quad \textbf{(6.18)}$$

El concepto de independencia es en este caso exactamente igual que en el caso discreto. La independencia de un conjunto de variables aleatorias implica que en la distribución de probabilidad de cualquiera de ellas no influyen los valores que tomen las demás. Así, por ejemplo, la afirmación de que las variaciones diarias consecutivas del precio de las acciones de una empresa son independientes entre sí implica que la información sobre las variaciones pasadas del precio carece de valor para saber qué ocurrirá probablemente mañana.

El concepto de esperanza se extiende a las funciones de variables aleatorias continuas que siguen una distribución conjunta. Al igual que ocurre en el caso de las variables aleatorias discretas, tenemos el concepto de *covarianza*, que se utiliza para evaluar las relaciones lineales entre pares de variables aleatorias.

Covarianza

Sean X e Y un par de variables aleatorias continuas que tienen las medias μ_x y μ_Y, respectivamente. El valor esperado de $(X - \mu_X)(Y - \mu_Y)$ se denomina **covarianza** (Cov) entre X e Y. Es decir,

$$\text{Cov}\,(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \quad \textbf{(6.19)}$$

Otra expresión alternativa, pero equivalente, es

$$\text{Cov}\,(X, Y) = E(XY) = -\mu_X\mu_Y \quad \textbf{(6.20)}$$

Si las variables aleatorias X e Y son independientes, la covarianza entre ellas es 0. Sin embargo, lo contrario no es necesariamente cierto.

En el apartado 5.7 también presentamos la *correlación* como una medida estandarizada de la relación entre dos variables aleatorias discretas. Los resultados son los mismos en el caso de las variables aleatorias continuas.

Correlación

Sean X e Y variables aleatorias distribuidas conjuntamente. La **correlación** (Corr) entre X e Y es

$$\rho = \text{Corr}\,(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X\sigma_Y} \quad \textbf{(6.21)}$$

En el apartado 5.7 presentamos las medias y las varianzas de sumas y diferencias de variables aleatorias discretas. Los resultados son los mismos en el caso de las variables aleatorias continuas, ya que se obtienen utilizando esperanzas, por lo que no influye el hecho de que las variables aleatorias sean discretas o continuas.

Sumas de variables aleatorias

Sean X_1, X_2, ..., X_K K variables aleatorias que tienen las medias μ_1, μ_2, ..., μ_K y las varianzas σ_1^2, σ_2^2, ..., σ_K^2. Se cumplen las siguientes propiedades:

1. La media de su suma es la suma de sus medias; es decir,

$$E(X_1 + X_2 + \cdots + X_K) = \mu_1 + \mu_2 + \cdots + \mu_K \quad \textbf{(6.22)}$$

2. Si la covarianza entre cada par de estas variables aleatorias es 0, entonces la varianza de su suma es la suma de sus varianzas; es decir,

$$\text{Var}(X_1 + X_2 + \cdots + X_K) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_K^2 \quad \textbf{(6.23)}$$

Sin embargo, si las covarianzas entre pares de variables aleatorias no son 0, la varianza de su suma es

$$\text{Var}(X_1 + X_2 + \cdots + X_K) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_K^2 + 2\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\text{Cov}(X_i, X_j) \quad \textbf{(6.24)}$$

Diferencias entre un par de variables aleatorias

Sean X e Y un par de variables aleatorias que tienen las medias μ_X y μ_Y y las varianzas σ_X^2 y σ_Y^2. Se cumplen las siguientes propiedades:

1. La media de su diferencia es la diferencia de sus medias; es decir,

$$E(X - Y) = \mu_X - \mu_Y \quad \textbf{(6.25)}$$

2. Si la covarianza entre X e Y es 0, entonces la varianza de su diferencia es

$$\text{Var}(X - Y) = \sigma_X^2 + \sigma_Y^2 \quad \textbf{(6.26)}$$

3. Si la covarianza entre X e Y no es 0, entonces la varianza de su diferencia es

$$\text{Var}(X - Y) = \sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}(X, Y) \quad \textbf{(6.27)}$$

EJEMPLO 6.12. Costes totales de un proyecto (media y desviación típica)

Un contratista no está seguro de cuáles son exactamente los costes totales de los materiales o de la mano de obra de un proyecto. Además, la línea total de crédito para financiar el proyecto es de 260.000 $ y el contratista quiere saber cuál es la probabilidad de que los costes totales sean de más de 260.000 $. Se cree que los costes de los materiales pueden representarse por medio de una variable aleatoria distribuida normalmente que tiene una media de 100.000 $ y una desviación típica de 10.000 $. Los costes laborales son de 1.500 $ al día y el número de días necesarios para realizar el proyecto puede representarse por medio de una variable aleatoria distribuida normalmente que tiene una

media de 80 y una desviación típica de 12. Suponiendo que los costes de los materiales y de la mano de obra son independientes, ¿cuáles son la media y la desviación típica del coste total del proyecto (materiales más mano de obra)? Además, ¿cuál es la probabilidad de que el coste total del proyecto sea de más de 260.000 $?

Solución

Sean las variables aleatorias X_1 y X_2 los costes de los materiales y de la mano de obra, respectivamente. Entonces, la media de X_1 es $\mu_1 = 100.000$ y su desviación típica es $\sigma_1 = 10.000$. En el caso de la variable aleatoria X_2,

$$\mu_2 = (1.500)(80) = 120.000 \quad \text{y} \quad \sigma_2 = (1.500)(12) = 18.000$$

El coste total del proyecto es $W = X_1 + X_2$ y el coste medio

$$\mu_W = \mu_1 + \mu_2 = 100.000 + 120.000 = 220.000\ \$$$

y dado que X_1 y X_2 son independientes, la varianza de su suma es

$$\sigma_W^2 = \sigma_1^2 + \sigma_2^2 = (10.000)^2 + (18.000)^2 = 424.000.000$$

Tomando la raíz cuadrada, observamos que la desviación típica es 20.591 $.

Dado que X_1 y X_2 siguen una distribución normal, puede demostrarse que su suma, W, también sigue una distribución normal. Por lo tanto, la media y la varianza de W pueden utilizarse para calcular una variable aleatoria normal estándar, Z, y la probabilidad de que W sea superior a 260.000 $.

$$Z = \frac{260.000 - 220.000}{20.591} = 1{,}94$$

Utilizando la tabla de la probabilidad normal acumulada, observamos que la probabilidad de que el coste total sea de más de 260.000 $ es 0,0262. Como esta probabilidad es pequeña, el contratista tiene una cierta seguridad de que el proyecto puede realizarse con la línea de crédito de que dispone.

EJEMPLO 6.13. Riesgo de una cartera de inversión (media y varianza de una función lineal)

Enrique Chamizo le ha pedido ayuda para crear una cartera que contenga acciones de dos empresas. Enrique tiene 1.000 $, que puede repartir en cualquier proporción entre las acciones de dos empresas. Los rendimientos por dólar de estas inversiones son las variables aleatorias X e Y. Las dos son independientes y tienen la misma media y la misma varianza. Enrique desea saber cuál es el riesgo de diversas posibilidades de asignar el dinero. Le señala que el riesgo está relacionado directamente con la varianza y que, por lo tanto, podría saber la respuesta si supiera cuál es la varianza de algunas posibilidades de asignar el dinero.

Solución

La cantidad de dinero asignada a la primera inversión es α y, por lo tanto, el resto, $1.000 - \alpha$, se asignará a la segunda. El rendimiento total de la inversión es

$$R = \alpha X + (1.000 - \alpha)Y$$

Esta variable aleatoria tiene un valor esperado de

$$\begin{aligned} E(R) &= \alpha E(X) + (1.000 - \alpha)E(Y) \\ &= \alpha\mu + (1.000 - \alpha)\mu = 1.000\mu \end{aligned}$$

Vemos, pues, que el rendimiento esperado de todas las asignaciones del dinero es el mismo.

Sin embargo, el riesgo o varianza es otra historia.

$$\begin{aligned} \text{Var}\,(R) &= \alpha^2\,\text{Var}\,(X) + (1.000 - \alpha)^2\,\text{Var}\,(Y) \\ &= \alpha^2\sigma^2 + (1.000 - \alpha)^2\sigma^2 \\ &= (2\alpha^2 - 2.000\ \alpha + 1.000.000)\sigma^2 \end{aligned}$$

Si α es igual a 0 o a 1.000, de manera que toda la cartera se asigna solamente a las acciones de una de las empresas, la varianza del rendimiento total es $1.000.000\sigma^2$. Sin embargo, si se asignan 500 $ a cada inversión, la varianza del rendimiento total es $500.000\sigma^2$, que es la varianza más pequeña posible. Repartiendo su inversión entre las acciones de dos empresas, Enrique puede reducir el efecto que puede producir el hecho de que los rendimientos de las acciones de una de las empresas sean altos o bajos. Por lo tanto, es posible obtener el mismo rendimiento esperado con una variedad de niveles de riesgo.

Combinaciones lineales de variables aleatorias

En el Capítulo 5 desarrollamos la media y la varianza de combinaciones lineales de variables aleatorias discretas. Estos resultados también se aplican a las variables aleatorias continuas, ya que su desarrollo se basa en operaciones con valores esperados y no depende de las distribuciones de probabilidad. Las ecuaciones 6.28 a 6.31 indican las propiedades importantes de las combinaciones lineales.

Combinaciones lineales de variables aleatorias

La combinación lineal de dos variables aleatorias, X e Y, es

$$W = aX + bY \tag{6.28}$$

donde a y b son constantes.

El valor medio de W es

$$\mu_W = E[W] = E[aX + bY] \tag{6.29}$$

La varianza de W es

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\text{Cov}\,(X, Y) \tag{6.30}$$

o, utilizando la correlación,

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\text{Corr}\,(X,\,Y)\sigma_X\sigma_Y \tag{6.31}$$

Si la combinación lineal de la ecuación 6.28 es una diferencia, es decir, si

$$W = aX - bY \tag{6.32}$$

entonces la media y la varianza son

$$\begin{aligned}\mu_w &= E[W] = E[aX - bY] \\ &= a\mu_X - b\mu_Y\end{aligned} \tag{6.33}$$

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 - 2ab\,\text{Cov}\,(X,\,Y) \tag{6.34}$$

o, utilizando la correlación,

$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 - 2ab\,\text{Corr}\,(X,\,Y)\sigma_X\sigma_Y \tag{6.35}$$

Estos resultados proceden directamente de las ecuaciones 6.28 a 6.31 sustituyendo el coeficiente *b* por un valor negativo en las ecuaciones.

Si tanto *X* como *Y* son variables aleatorias distribuidas normalmente, entonces la variable aleatoria resultante, *W*, también sigue una distribución normal que tiene la media y la varianza mostradas. Este resultado nos permite averiguar la probabilidad de que la combinación lineal, *W*, esté dentro de un intervalo específico.

EJEMPLO 6.14. Análisis de cartera (probabilidad de una cartera)

Cristina Juárez, gestora de cuentas de la sociedad de valores Norte, tiene una cartera que contiene 20 acciones de Sistemas Informáticos Albertina y 30 de Ciberanálisis Beta. Las dos empresas producen dispositivos de acceso a la web que compiten en el mercado de consumidores. El precio de las acciones de Albertina sigue una distribución normal que tiene una media $\mu_X = 25$ y una varianza $\sigma_X^2 = 81$. El precio de las acciones de Beta también sigue una distribución normal de media $\mu_Y = 40$ y varianza $\sigma_Y^2 = 121$. Los precios de las acciones tienen una correlación negativa, $\rho_{XY} = -0{,}40$. Cristina le ha pedido que halle la probabilidad de que el valor de la cartera sea de más de 2.000.

Solución

El valor de la cartera de Cristina, *W*, viene definido por la combinación lineal

$$W = 20X + 30Y$$

y *W* sigue una distribución normal. El valor medio de su cartera de acciones es

$$\begin{aligned}\mu_W &= 20\mu_X + 30\mu_Y \\ &= 20 \times 25 + 30 \times 40 = 1.700\end{aligned}$$

La varianza del valor de la cartera es

$$\begin{aligned}\sigma_W^2 &= 20^2\sigma_X^2 + 30^2\sigma_Y^2 + 2 \times 20 \times 30\,\text{Corr}\,(X,\,Y)\sigma_X\sigma_Y \\ &= 20^2 \times 81 + 30^2 \times 121 + 2 \times 20 \times 30 \times (-0{,}40) \times 9 \times 11 = 93{,}780\end{aligned}$$

y la desviación típica del valor de la cartera es

$$\sigma_W = 306{,}24$$

La normal estándar Z de 2.000 es

$$Z_W = \frac{2.000 - 1.700}{306{,}24} = 0{,}980$$

Y la probabilidad de que el valor de la cartera sea de más de 2.000 es 0,1635. De la simetría de la distribución normal se deduce que la probabilidad de que el valor de la cartera sea de menos de 1.400 también es 0,1635.

Si los precios de las acciones de las dos empresas tuvieran una correlación positiva, $\rho = +0{,}40$, la media sería la misma, pero la varianza y la desviación típica serían

$$\begin{aligned}\sigma_W^2 &= 20^2\sigma_X^2 + 30^2\sigma_Y^2 + 2 \times 20 \times 30\,\text{Corr}\,(X, Y)\sigma_X\sigma_Y \\ &= 20^2 \times 81 + 30^2 \times 121 + 2 \times 20 \times 30 \times (+0{,}40) \times 9 \times 11 = 188.820 \\ \sigma_W &= 434{,}53\end{aligned}$$

La normal estándar Z de 2.000 es

$$Z_W = \frac{2.000 - 1.700}{434{,}53} = 0{,}690$$

La probabilidad de que el valor de su cartera sea de más de 2.000 es 0,2451 y la probabilidad de que sea de menos de 1.400 también es 0,2451.

Vemos, pues, que cuando la correlación entre los precios de las acciones es positiva, la varianza y el riesgo son mayores. En este ejemplo, el riesgo aumenta la probabilidad de que el valor de la cartera sea de más de 2.000, de 0,1635 a 0,2451. Eso también implica una variación similar de la probabilidad de que el valor de la cartera sea de menos de 1.400. Cuando el riesgo es mayor, también es mayor la probabilidad de que el valor de la cartera sea mayor o menor en comparación con la opción en la que el riesgo es menor.

El ejemplo anterior ilustra un principio fundamental muy importante en la creación de carteras de inversión. Recuérdese que el riesgo de una inversión está relacionado directamente con la varianza de su valor. En el ejemplo anterior, hemos mostrado que si los valores de los precios de las acciones de dos empresas están correlacionados positivamente, la cartera resultante tiene una varianza mayor y, por lo tanto, un riesgo mayor. Y si los precios están correlacionados negativamente, la cartera resultante tiene una varianza menor y, por lo tanto, un riesgo menor. Los gestores de fondos utilizan a menudo el término *cobertura* para describir este fenómeno. Este importante principio en el caso de una cartera de acciones de dos empresas se extiende directamente a una cartera de acciones de un gran número de empresas, pero en ese caso los cálculos son más complejos y normalmente se realizan utilizando un complejo programa informático. Los gestores de fondos de inversión utilizan este principio para seleccionar combinaciones de muchas acciones distintas para hallar el valor y el riesgo que se desea que tenga la cartera y que son los objetivos de un fondo de inversión.

EJERCICIOS

Ejercicios básicos

6.59. Una variable aleatoria X sigue una distribución normal de media 100 y varianza 100 y una variable aleatoria Y sigue una distribución normal de media 200 y varianza 400. Las variables aleatorias tienen un coeficiente de correlación igual a 0,5. Halle la media y la varianza de la variable aleatoria

$$W = 5X + 4Y$$

6.60. Una variable aleatoria X sigue una distribución normal de media 100 y varianza 100 y una variable aleatoria Y sigue una distribución normal de media 200 y varianza 400. Las variables aleatorias tienen un coeficiente de correlación igual a $-0,5$. Halle la media y la varianza de la variable aleatoria

$$W = 5X + 4Y$$

6.61. Una variable aleatoria X sigue una distribución normal de media 100 y varianza 100 y una variable aleatoria Y sigue una distribución normal de media 200 y varianza 400. Las variables aleatorias tienen un coeficiente de correlación igual a 0,5. Halle la media y la varianza de la variable aleatoria

$$W = 5X - 4Y$$

6.62. Una variable aleatoria X sigue una distribución normal de media 500 y varianza 100 y una variable aleatoria Y sigue una distribución normal de media 200 y varianza 400. Las variables aleatorias tienen un coeficiente de correlación igual a 0,5. Halle la media y la varianza de la variable aleatoria

$$W = 5X - 4Y$$

6.63. Una variable aleatoria X sigue una distribución normal de media 100 y varianza 500 y una variable aleatoria Y sigue una distribución normal de media 200 y varianza 400. Las variables aleatorias tienen un coeficiente de correlación igual a $-0,5$. Halle la media y la varianza de la variable aleatoria

$$W = 5X - 4Y$$

Ejercicios aplicados

6.64. Un inversor planea repartir 200.000 $ entre dos inversiones. La primera genera un beneficio seguro del 10 por ciento, mientras que la segunda genera un beneficio que tiene un valor esperado de 18 por ciento y una desviación típica de 6 por ciento. Si el inversor reparte el dinero por igual entre estas dos inversiones, halle la media y la desviación típica del beneficio total.

6.65. El propietario de una vivienda ha instalado un nuevo sistema de calefacción de bajo consumo. Se estima que este sistema reducirá los costes de calefacción durante un año en una cantidad que puede considerarse una variable aleatoria que tiene una media de 200 $ y una desviación típica de 60 $. Indicando los supuestos que necesite postular, halle la media y la desviación típica de la reducción total del coste de calefacción en un periodo de 5 años.

6.66. Un consultor está comenzando a trabajar en tres proyectos cuyos beneficios esperados son 50.000 $, 72.000 $ y 40.000 $. Las desviaciones típicas correspondientes son 10.000 $, 12.000 $ y 9.000 $. Suponiendo que los resultados son independientes, halle la media y la desviación típica de los beneficios totales de estos tres proyectos.

6.67. Un consultor tiene tres fuentes de ingresos: unos cursos breves, la venta de programas informáticos y la consultoría. Los ingresos anuales que espera obtener de estas fuentes son 20.000 $, 25.000 $ y 15.000 $ y las desviaciones típicas respectivas son 2.000 $, 5.000 $ y 4.000 $. Suponiendo que son independientes, halle la media y la desviación típica de sus ingresos anuales totales.

6.68. Cinco inspectores tienen la responsabilidad de verificar la calidad de los componentes que produce una cadena de montaje. El número de componentes que puede verificar cada inspector en un turno puede representarse por medio de una variable aleatoria que tiene una media de 120 y una desviación típica de 16. Sea X el número de componentes comprobados por un inspector en un turno. Entonces, el número total comprobado es $5X$, que tiene una media de 600 y una desviación típica de 80. ¿Dónde está el error en este razonamiento? Suponiendo que los rendimientos de los inspectores son independientes entre sí, halle la media y la desviación típica del número total de componentes comprobados en un turno.

6.69. Se estima que conduciendo normalmente por una autopista, el número de kilómetros que pueden recorrer los automóviles de un determinado mo-

delo con 1 litro de gasolina puede representarse por medio de una variable aleatoria que tiene una media de 28 y una desviación típica de 2,4. Se conducen independientemente 16 automóviles de este modelo, cada uno con 1 litro de gasolina. Halle la media y la desviación típica del número medio de kilómetros que pueden recorrer estos automóviles.

6.70. Sara Jonás, gestora de cartera, le ha pedido que analice una cartera recién adquirida para hallar su valor medio y su variabilidad. La cartera consta de 50 acciones de Xilófonos Reunidos y 40 de Talleres Yunque. El análisis de la historia pasada indica que el precio de las acciones de Xilófonos tiene una media de 25 y una varianza de 121. Un análisis similar indica que el precio de las acciones de Yunque tiene una media de 40 y una varianza de 225. Los mejores datos de los que se dispone indican que los precios de las acciones tienen una correlación de $+0{,}5$.

a) Calcule la media y la varianza de la cartera.

b) Suponga que la correlación entre los precios de las acciones fuera realmente de $-0{,}5$. ¿Cuáles son ahora la media y la varianza de la cartera?

6.71. Cereales Flores de la Pradera tiene unos ingresos anuales por ventas de 400.000.000 \$. Jorge Severino, vicepresidente de 58 años, es responsable de la producción y de las ventas del producto Cereales Afrutados con Nueces. La producción diaria en cajas sigue una distribución normal que tiene una media de 100 y una varianza de 625. Las ventas diarias en cajas también siguen una distribución normal que tiene una media de 100 y una desviación típica de 8. Las ventas y la producción tienen una correlación de 0,60. El precio de venta por caja es de 10 \$. El coste variable de producción por caja es de 7 \$. Los costes fijos de producción por día son de 250 \$.

a) ¿Cuál es la probabilidad de que el ingreso total sea mayor que los costes totales un día cualquiera?

b) Construya un intervalo de aceptación del 95 por ciento para los ingresos totales por ventas menos los costes totales.

6.72. Olecarl, país situado en el Pacífico Sur, le ha pedido que analice las pautas de comercio internacional. Primero descubre que todos los años exporta 10 unidades e importa 10 unidades de un paño maravilloso. El precio de las exportaciones es una variable aleatoria que tiene una media de 100 y una varianza de 100. El precio de las importaciones es una variable aleatoria que tiene una media de 90 y una varianza de 400. Descubre, además, que los precios de las importaciones y las exportaciones tienen una correlación de $\rho = -0{,}40$. Los precios de las exportaciones y de las importaciones siguen una distribución normal. La balanza comercial es la diferencia entre los ingresos totales generados por las exportaciones y los costes totales de las importaciones.

a) ¿Cuáles son la media y la varianza de la balanza comercial?

b) ¿Cuál es la probabilidad de que la balanza comercial sea negativa?

6.73. Le han pedido que halle la probabilidad de que el «margen de contribución» (la diferencia entre el ingreso total y el coste variable total) de una determinada línea de productos sea mayor que el coste fijo de 2.000 \$. El número total de unidades vendidas es una variable aleatoria que sigue una distribución normal que tiene una media de 400 y una varianza de 900 $X \sim N$ (400, 900). El precio de venta por unidad es de 10 \$. El número total de unidades producidas es una variable aleatoria que sigue una distribución normal que tiene una media de 400 y una varianza de 1.600 $Y \sim N$ (400, 1.600). El coste variable de producción es de 4 \$ por unidad. La producción y las ventas tienen una correlación positiva de 0,50.

6.74. El país de Waipo ha creado recientemente un plan de desarrollo económico que incluye un aumento de las exportaciones y de las importaciones. Ha realizado una serie de extensos estudios de la economía mundial y de la capacidad económica de Waipo, tras un extenso programa decenal de mejora de la educación. El modelo resultante indica que el próximo año las exportaciones seguirán una distribución normal de media 100 y varianza 900 (en miles de millones de yuanes de Waipo). Además, se espera que las importaciones sigan una distribución normal de media 105 y varianza 625 en las mismas unidades. Se espera que la correlación entre las exportaciones y las importaciones sea de $+0{,}70$. La balanza comercial es igual a las exportaciones menos las importaciones.

a) Halle la media y la varianza de la balanza comercial (exportaciones menos importaciones) suponiendo que los parámetros del modelo dados antes son verdaderos.

b) ¿Cuál es la probabilidad de que la balanza comercial sea positiva?

RESUMEN

En el Capítulo 6 hemos desarrollado modelos de probabilidad de variables aleatorias continuas que siguen una pauta similar a la que utilizamos en el caso de los modelos de probabilidad de variables aleatorias discretas en el Capítulo 5. Hemos desarrollado dos modelos de distribución de probabilidad paramétricos, el normal y el exponencial. Hemos mostrado, además, que puede utilizarse la distribución normal como aproximación de la binomial cuando el tamaño de la muestra es grande. Por último, hemos presentado las distribuciones conjuntas de variables aleatorias continuas. Hemos ampliado los modelos de combinaciones de variables aleatorias para mostrar cómo podemos utilizar la media y la varianza para calcular la probabilidad de que la cartera total esté en un intervalo específico, basándonos en el modelo de probabilidad normal. Éstas y otras extensas aplicaciones constituyen una sólida base para utilizar las variables aleatorias continuas.

TÉRMINOS CLAVE

áreas situadas debajo de funciones de probabilidad continua, 204
combinaciones lineales de variables aleatorias, 238
cómo se hallan probabilidades de intervalos de variables aleatorias normales, 214
correlación, 235
covarianza, 235
desviación típica de una variable aleatoria continua, 208
diferencias entre pares de variables aleatorias, 236
distribución normal como aproximación de la distribución binomial, 225
distribución normal estándar, 214
distribución de probabilidad exponencial, 232
distribución de probabilidad uniforme, 202
función de densidad de la distribución normal, 212
función de densidad de probabilidad, 203
función de distribución acumulada, 202
función de distribución acumulada conjunta, 235
función de distribución acumulada de la distribución normal, 213
funciones de distribución marginal, 235
media de una variable aleatoria continua, 208
probabilidad de un intervalo utilizando una función de distribución acumulada, 202
probabilidades de intervalos de variables aleatorias normales, 214
propiedades de la distribución normal, 212
sumas de variables aleatorias, 236
valor esperado de variables aleatorias continuas, 208
varianza, 208

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

6.75. Un consultor sabe que le costará 10.000 $ cumplir un contrato. El contrato se va a sacar a subasta y cree que la oferta más baja, excluida la suya, puede representarse por medio de una distribución que es uniforme entre 8.000 $ y 20.000 $. Por lo tanto, si la variable aleatoria X representa la oferta más baja de todas las demás (en miles de dólares), su función de densidad es

$$f_x(x) = \begin{cases} 1/12 & \text{para } 8 < x < 20 \\ 0 & \text{para todos los demás valores de } x \end{cases}$$

a) ¿Cuál es la probabilidad de que la oferta más baja de todas las demás sea menor que la estimación del coste de 10.000 $ del consultor?
b) Si el consultor presenta una oferta de 12.000 $, ¿cuál es la probabilidad de que consiga el contrato?
c) El consultor decide presentar una oferta de 12.000 $. ¿Cuál es el beneficio esperado de esta estrategia?
d) Si el consultor quiere presentar una oferta que le permita obtener el máximo beneficio esperado posible, explique qué debe hacer para tomar esta decisión.

6.76. Las edades de un grupo de ejecutivos que asisten a un congreso están distribuidas de una manera uniforme entre 35 y 65 años. Si la variable aleatoria X representa las edades en años, la función de densidad es

$$f_x(x) = \begin{cases} 1/30 & \text{para } 35 < x < 65 \\ 0 & \text{para todos los demás valores de } x \end{cases}$$

a) Trace la función de densidad de X.
b) Halle y trace la función de distribución acumulada de X.

c) Halle la probabilidad de que la edad de un ejecutivo de este grupo elegido aleatoriamente esté entre 40 y 50 años.
d) Halle la edad media de los ejecutivos del grupo.

6.77. La variable aleatoria X tiene la función de densidad

$$f_x(x) = \begin{cases} x & \text{para } 0 < x < 1 \\ 2 - x & \text{para } 1 < x < 2 \\ 0 & \text{para todos los demás valores de } x \end{cases}$$

a) Trace la función de densidad de X.
b) Demuestre que la densidad tiene las propiedades de una función de densidad bien definida.
c) Halle la probabilidad de que X tome un valor entre 0,5 y 1,5.

6.78. Un inversor coloca 2.000 $ en una cuenta que tiene una tasa de rendimiento fija del 10 por ciento al año. Invierte otros 1.000 $ en un fondo que tiene una tasa esperada de rendimiento del 16 por ciento y una desviación típica del 8 por ciento al año.
a) Halle el valor esperado de la cantidad total de dinero que tendrá el inversor después de un año.
b) Halle la desviación típica de la cantidad total que tendrá después de un año.

6.79. Una hamburguesería vende hamburguesas a 1,45 $ cada una. Las ventas diarias tienen una distribución de media de 530 y desviación típica de 69.
a) Halle los ingresos totales diarios medios generados por la venta de hamburguesas.
b) Halle la desviación típica de los ingresos totales generados por la venta de hamburguesas.
c) Los costes diarios (en dólares) vienen dados por

$$C = 100 + 0{,}95X$$

donde X es el número vendido de hamburguesas. Halle la media y la desviación típica de los beneficios diarios generados por las ventas.

6.80. Una analista predice los beneficios empresariales y su trabajo se evalúa comparando los beneficios efectivos con los predichos. Sean

$$\text{Beneficios efectivos} = \text{beneficios predichos} + \text{error de predicción}$$

Demuestre que si los beneficios predichos y el error de predicción son independientes entre sí, la varianza de los beneficios predichos es menor que la varianza de los beneficios efectivos.

6.81. Sean X_1 y X_2 un par de variables aleatorias. Demuestre que la covarianza entre las variables aleatorias $(X_1 + X_2)$ y $(X_1 - X_2)$ es 0 si y sólo si X_1 y X_2 tienen la misma varianza.

6.82. Las calificaciones medias de los estudiantes de una gran universidad siguen una distribución normal que tiene una media de 2,6 y una desviación típica de 0,5.
a) Se elige aleatoriamente un estudiante de esta universidad. ¿Cuál es la probabilidad de que tenga una calificación media de más de 3,0?
b) Se elige aleatoriamente un estudiante de esta universidad. ¿Cuál es la probabilidad de que tenga una calificación media de entre 2,25 y 2,75?
c) Cuál es la calificación media mínima necesaria para que la calificación media de un estudiante esté entre el 10 por ciento más alto de la universidad?
d) Se eligen aleatoriamente una muestra de 400 estudiantes de esta universidad. ¿Cuál es la probabilidad de que al menos 80 de estos estudiantes tengan una calificación media de más de 3,0?
e) Se eligen aleatoriamente dos estudiantes de esta universidad. ¿Cuál es la probabilidad de que al menos uno de ellos tenga una calificación media de más de 3,0?

6.83. Una empresa repara aparatos de aire acondicionado. Se sabe que el tiempo que tarda en repararlos sigue una distribución normal que tiene una media de 60 minutos y una desviación típica de 10 minutos.
a) ¿Cuál es la probabilidad de que tarde en reparar un aparato más de 65 minutos?
b) ¿Cuál es la probabilidad de que tarde en reparar un aparato entre 50 y 70 minutos?
c) La probabilidad de que tarde más de _____ minutos en reparar un aparato es 0,025.
d) Halle el intervalo más corto de tiempos que incluya el 50 por ciento de todos los avisos de reparación.
e) Se toma una muestra aleatoria de cuatro reparaciones de aparatos. ¿Cuál es la probabilidad de que el tiempo de reparación exactamente de dos de ellos sea de más de 65 minutos?

6.84. Se ha observado que el tiempo que tarda la gente en rellenar un impreso de declaración de impuestos sigue una distribución normal que tiene una media de 100 minutos y una desviación típica de 30 minutos.

a) ¿Cuál es la probabilidad de que una persona elegida aleatoriamente tarde menos de 85 minutos en rellenar este impreso?

b) ¿Cuál es la probabilidad de que una persona elegida aleatoriamente tarde entre 70 y 130 minutos en rellenar este impreso?

c) El 5 por ciento de todas las personas tarda más de _____ minutos en rellenar este impreso.

d) Se eligen aleatoriamente dos personas. ¿Cuál es la probabilidad de que al menos una de ellas tarde más de una hora en rellenar este impreso?

e) Se eligen aleatoriamente cuatro personas. ¿Cuál es la probabilidad de que exactamente dos de ellas tarden más de una hora en rellenar este impreso?

f) Indique en el caso de una persona elegida aleatoriamente en cuál de los intervalos siguientes (expresados en minutos) es más probable que esté el tiempo que tarda en rellenar el impreso.

70-89 90-109 110-129 130-149

g) Indique en el caso de una persona elegida aleatoriamente en cuál de los intervalos siguientes (expresados en minutos) es menos probable que esté el tiempo que tarda en rellenar el impreso.

70-89 90-109 110-129 130-149

6.85. Una pizzería tiene un servicio de reparto de pizzas en una residencia de estudiantes. Los tiempos de entrega siguen una distribución normal que tiene una media de 20 minutos y una desviación típica de 4 minutos.

a) ¿Cuál es la probabilidad de que tarde en entregar una pizza entre 15 y 25 minutos?

b) La pizzería no cobra la pizza si tarda más de 30 minutos en entregarla. ¿Cuál es la probabilidad de conseguir una pizza gratis en un único pedido?

c) En la época de los exámenes finales, un estudiante planea pedir pizza cinco noches seguidas. Suponga que los tiempos de entrega son independientes entre sí. ¿Cuál es la probabilidad de que el estudiante consiga al menos una pizza gratis?

d) Halle el intervalo más corto de tiempos que contenga el 40 por ciento de todas las entregas.

e) Indique en cuál de los intervalos siguientes (expresados en minutos) es más probable que esté el tiempo de entrega de un único pedido.

18-20 19-21 20-22 21-23

f) Indique en cuál de los intervalos siguientes (expresados en minutos) es menos probable que esté el tiempo de entrega de un único pedido.

18-20 19-21 20-22 21-23

6.86. Una cadena de videoclubs estima que los gastos anuales de los socios siguen una distribución normal que tiene una media de 100 $. También se ha observado que el 10 por ciento de todos los socios gasta más de 130 $ al año. ¿Qué porcentaje de socios gasta más de 140 $ al año?

6.87. Se estima que la cantidad de dinero que gastan en gasolina los clientes de una estación de servicio sigue una distribución normal que tiene una desviación típica de 2,50 $. También se ha observado que el 10 por ciento de todos los clientes gasta más de 25 $. ¿Qué porcentaje de los clientes gasta menos de 20 $?

6.88. Una empresa de estudios de mercado ha observado que el 40 por ciento de todos los clientes de los supermercados se niega a cooperar cuando le preguntan sus encuestadores. Si éstos abordan a 1.000 compradores, ¿cuál es la probabilidad de que menos de 500 se nieguen a cooperar?

6.89. Una organización que da seminarios habitualmente sobre métodos para vender más observa que el 60 por ciento de sus clientes ha asistido a otros seminarios anteriores. ¿Cuál es la probabilidad de que más de la mitad de una muestra de 400 clientes haya asistido a otros seminarios anteriores?

6.90. Un servicio de grúa de emergencia recibe una media de 70 llamadas al día. ¿Cuál es la probabilidad de que en un día cualquiera reciba menos de 50 llamadas?

6.91. En unos grandes almacenes, el departamento de atención al cliente recibe, en promedio, seis reclamaciones por hora sobre la calidad del servicio. La distribución es de Poisson.

a) ¿Cuál es la probabilidad de que se reciban en cualquier hora seis reclamaciones exactamente?

b) ¿Cuál es la probabilidad de que transcurran más de 20 minutos entre una reclamación y otra?

c) ¿Cuál es la probabilidad de que transcurran menos de 5 minutos entre una reclamación y otra?

d) El director de los grandes almacenes observa el departamento de atención al cliente durante

un periodo de 30 minutos, en el que no se recibe ninguna reclamación. Llega a la conclusión de que una charla que dio al personal sobre el tema «El cliente siempre tiene razón» ha surtido claramente un efecto beneficioso. Suponga que la charla no ha surtido, en realidad, ningún efecto. ¿Cuál es la probabilidad de que el director observe un periodo de 30 minutos o más sin ninguna reclamación?

6.92. Una emisora de radio cree que el 40 por ciento de su audiencia tiene menos de 25 años. Se eligen aleatoriamente 600 oyentes.

a) Si lo que cree la emisora es cierto, ¿cuál es la probabilidad de que más de 260 de estos oyentes tenga menos de 25 años?

b) Si lo que cree la emisora es cierto, la probabilidad de que más de _____ de estos 600 oyentes tenga menos de 25 años es 0,6.

6.93. Se estima que el tiempo de duración de un partido de béisbol sigue una distribución normal que tiene una media de 132 minutos y una desviación típica de 12 minutos.

a) ¿Qué proporción de todos los partidos dura entre 120 y 150 minutos?

b) El 33 por ciento de todos los partidos dura más de _____ minutos.

c) ¿Qué proporción de todos los partidos dura menos de 120 minutos?

d) Si se eligen aleatoriamente 100 partidos, ¿cuál es la probabilidad de que al menos 25 duren menos de 120 minutos?

6.94. Un consultor de empresas observó que la cantidad diaria de tiempo que dedicaban los ejecutivos a realizar tareas que podían ser realizadas igual de bien por subordinados seguía una distribución normal que tenía una media de 2,4 horas. También observó que el 10 por ciento de los ejecutivos dedicaba más de 3,5 horas al día a realizar tareas de este tipo. Halle la probabilidad de que más de 80 ejecutivos de una muestra de 400 dedique más de 3 horas al día a tareas de este tipo.

6.95. Gestores Financieros S.A. compra y vende normalmente acciones de un gran número de empresas para los distintos fondos que gestiona. La gestora de carteras Andrea Colson le ha pedido ayuda para analizar un fondo cuya cartera está formada en parte por 10 acciones de la empresa A y 8 de la B. El precio de las acciones de A tiene una media de 10 y una varianza de 16, mientras que el de las acciones de B tiene una media de 12 y una varianza de 9. La correlación entre los precios es 0,3.

a) ¿Cuáles son la media y la varianza del valor de la cartera?

b) Le han pedido a Andrea que reduzca la varianza (el riesgo) de la cartera. Propone vender las 10 acciones de la empresa A y recibe dos ofertas de las que puede seleccionar una: 10 acciones de la empresa 1 con un precio medio de 10, una varianza de 25 y una correlación con el precio de las acciones de B igual a $-0,2$; o 10 acciones de la empresa 2 con un precio medio de 10, una varianza de 9 y una correlación con el precio de las acciones de B igual a $+0,6$. ¿Qué oferta debe seleccionar?

6.96. Gestores Financieros S.A. compra y vende normalmente acciones de un gran número de empresas para los distintos fondos que gestiona. La gestora de cartera Sara Barco le ha pedido ayuda para analizar un fondo cuya cartera está formada en parte por 10 acciones de la empresa A y 89 de la empresa B. El precio de las acciones de A tiene una media de 12 y una varianza de 14, mientras que el precio de las acciones de B tiene una media de 10 y una varianza de 12. La correlación entre los precios es 0,5.

a) ¿Cuáles son la media y la varianza del valor de la cartera?

b) Le han pedido a Sara que reduzca la varianza (el riesgo) de la cartera. Propone vender las 10 acciones de la empresa A y recibe dos ofertas de las que puede seleccionar una: 10 acciones de la empresa 1 con un precio medio de 12, una varianza de 25 y una correlación con el precio de las acciones de B igual a $-0,2$; o 10 acciones de la empresa 2 con un precio medio de 10, una varianza de 9 y una correlación con el precio de las acciones de B igual a 0,6. ¿Qué oferta debe seleccionar?

6.97. Construcciones El Clavo está construyendo un gran centro de estudiantes para una famosa universidad. Durante el proyecto, Cristina Vilches, la directora del proyecto, pide que se eche un montón de arena que pesa entre 138.000 kilos y 141.000 en el camino recién construido. Le ha pedido que halle la probabilidad de que la arena entregada satisfaga la petición de Cristina. Usted ha ordenado que se utilice un camión grande y uno pequeño para llevar la arena. La cantidad de arena que lleva el camión grande sigue una distribución normal que tiene una media de 80.000 y una varianza de 1.000.000 y la que lleva el camión pequeño también sigue una distribución normal que tienen un peso medio de 60.000 kilos

y una varianza de 810.000. Sabe por experiencia que el peso de la arena de los dos camiones tiene una correlación de 0,40. ¿Cuál es la probabilidad de que el montón de arena resultante pese entre 138.000 y 141.000 kilos?

6.98. La compañía aérea Vuelos Nocturnos tiene un vuelo regular de Minneapolis a Francfort que sale a las 18 horas los días laborables. Basándose en una compleja relación entre Vuelos Nocturnos y Vuelos Cercanos, una compañía local que vuela a algunas pequeñas ciudades, se reservan 100 plazas para los pasajeros de dos de los vuelos de Vuelos Cercanos que llegan diariamente a las 17 horas. El número de pasajeros del vuelo procedente de Tri-mountain (Montana) sigue una distribución normal que tiene una media de 40 pasajeros y una varianza de 100. El número de pasajeros del otro vuelo, procedente de Bighog (Iowa), también sigue una distribución normal que tiene una media de 35 pasajeros y una varianza de 144. Los números de pasajeros de estos dos vuelos tienen una correlación de 0,6.

a) ¿Qué probabilidad hay de que se ocupen las 100 plazas del vuelo de Francfort?

b) ¿Qué probabilidad hay de que se ocupen entre 75 y 90 plazas?

Apéndice

1. Los lectores que tengan conocimientos de cálculo reconocerán que la probabilidad de que una variable aleatoria se encuentre en un intervalo dado es la integral de la función de densidad entre los puntos extremos del intervalo; es decir,

$$P(a < X < b) = \int_a^b f(x)\,dx$$

2. En términos formales, utilizando la notación del cálculo integral,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

La función de distribución acumulada es, pues, la integral

$$F(x_0) = \int_{-\infty}^{x_0} f(x)\,dx$$

Se deduce, pues, que la función de densidad es la derivada de la función de distribución acumulada; es decir,

$$f(x) = \frac{dF(x)}{dx}$$

3. En términos formales, utilizando el cálculo integral expresamos el valor esperado de la variable aleatoria X de la forma siguiente:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$$

y el valor esperado de la función $g(X)$:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

Obsérvese que, en la formación de estas esperanzas, la integral desempeña el mismo papel que el operador de los sumatorios en el caso discreto.

4. La integral

$$F(x_0) = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}\, dx$$

no tiene una sencilla forma algebraica.

5. Utilizando el cálculo integral, vemos que

$$\begin{aligned} P(t \leqslant T) &= \int_0^T \lambda e^{-\lambda t}\, dt \\ &= 1 - e^{-\lambda T} \end{aligned}$$

Capítulo 7

Muestreo y distribuciones en el muestreo

Esquema del capítulo

Introducción

En los Capítulos 4, 5 y 6 desarrollamos modelos de probabilidad que pueden utilizarse para representar la variabilidad subyacente de algunos procesos empresariales y económicos. En el 3 presentamos estadísticos descriptivos que pueden utilizarse para resumir muestras de datos procedentes de estos distintos procesos. En este capítulo relacionamos estos conceptos. Esta combinación nos permite construir modelos de probabilidad para distintos estadísticos calculados a partir de datos muestrales. Estos modelos de probabilidad se llaman *distribuciones en el muestro* y se utilizarán para desarrollar diversos métodos de inferencia estadística en el resto de este libro.

Los métodos estadísticos centran la atención en la realización de inferencias sobre grandes poblaciones de objetos utilizando una pequeña muestra de los objetos. Ejemplos representativos de poblaciones son:

1. *Todas* las familias que viven en una ciudad.
2. *Todas* las acciones que cotizan en una bolsa de valores.
3. El conjunto de *todas* las reclamaciones de cobertura de un seguro de accidentes de tráfico recibidas durante un año.
4. *Todos* los automóviles de un determinado modelo.
5. *Todas* las cuentas pendientes de cobro de un gran proveedor de piezas de automóvil.

Podría interesarnos conocer características medidas específicas de individuos de estas poblaciones. Por ejemplo, podríamos querer hacer una inferencia sobre la media y la varianza de la distribución poblacional de las rentas de las familias de una ciudad o sobre la proporción de todas las familias de una ciudad que tienen una renta anual de menos de 20.000 $.

7.1. Muestreo de una población

A menudo utilizamos muestras en lugar de toda la población porque el coste y el tiempo necesarios para medir todos los miembros de la población serían prohibitivos. Además, en algunos casos la medición requiere la destrucción de miembros. En general, se consigue una precisión mayor extrayendo con cuidado una muestra aleatoria de la población que dedicando los recursos a medir todos los miembros. La precisión es mayor por dos razones. En primer lugar, a menudo es muy difícil obtener y medir todos los miembros de una población e, incluso cuando es posible, el coste es muy alto cuando la población es grande. Por ejemplo, los estadísticos saben perfectamente que en el censo que se realiza cada 10 años en Estados Unidos algunos grupos tienen una representación muy inferior a la que les corresponde (véase la referencia bibliográfica 2). En segundo lugar, como vemos en este capítulo, pueden utilizarse muestras bien seleccionadas para realizar estimaciones medidas de las características de la población que son muy cercanas a los valores reales. La muestra ideal para este fin es la *muestra aleatoria simple*.

Muestra aleatoria simple

Supongamos que queremos seleccionar una muestra de n objetos de una población de N objetos. Se selecciona una **muestra aleatoria simple** tal que todos los objetos tienen la misma probabilidad de ser seleccionados y se seleccionan independientemente, es decir, la selección de un objeto no altera la probabilidad de que sean seleccionados otros objetos.

Las muestras aleatorias simples son el ideal. En algunos estudios por muestreo del mundo real, los analistas desarrollan métodos alternativos para reducir los costes del muestreo. Pero la base para saber si estas estrategias alternativas son aceptables es el grado en que los resultados se aproximan a los de una muestra aleatoria simple.

Es importante que una muestra represente al conjunto de la población. Si un director de marketing quiere evaluar las reacciones a un nuevo producto alimenticio, no muestrea únicamente a sus amigos y vecinos. Es improbable que las opiniones de esos grupos representen las de toda la población y es probable que estén concentradas en un intervalo más reducido. Para evitar estos problemas, seleccionamos una muestra aleatoria simple. El muestreo aleatorio es nuestra póliza de seguro contra la posibilidad de que los sesgos personales influyan en la selección.

El muestreo aleatorio simple puede realizarse de muchas formas. Podemos colocar los N miembros de la *población* —por ejemplo, bolas de colores— en un gran tonel y mezclarlos perfectamente. A continuación, podemos seleccionar en este tonel de bolas perfectamente mezcladas bolas de diferentes partes del tonel. En la práctica, solemos utilizar números aleatorios para seleccionar objetos a los que podemos asignar un valor numérico. Por ejemplo, los grupos de estudios de mercado pueden utilizar números aleatorios para seleccionar números telefónicos a los que llamar y preguntar por las preferencias por un producto. Algunos paquetes estadísticos y hojas de cálculo tienen rutinas para obtener números aleatorios, que se utilizan generalmente en la mayoría de los estudios por muestreo. Estos números aleatorios generados por ordenador tienen las propiedades necesarias para elaborar muestras aleatorias. Las organizaciones que necesitan muestras aleatorias de grandes poblaciones humanas —por ejemplo, los candidatos políticos que tratan de averiguar las preferencias de los votantes— recurren a empresas profesionales de muestreo, que se dedican a seleccionar y gestionar el proceso de muestreo. Un buen muestreo exige mucho trabajo y tiene un elevado coste.

Aquí centramos la atención en los métodos para analizar los resultados de muestras aleatorias simples con el fin de obtener información sobre la población. Este proceso, sobre el que nos extenderemos en los cinco capítulos siguientes, se conoce con el nombre de inferencia clásica. Estos métodos suponen generalmente que se utilizan muestras aleatorias simples. Sin embargo, existen otros métodos de muestreo, que es posible que en algunas circunstancias se prefieran a otros métodos de muestreo.

Las muestras aleatorias protegen contra la posibilidad de que algún grupo de la población esté subrepresentado en la muestra. Si una población se muestrea repetidamente utilizando métodos de muestreo aleatorio, ningún subgrupo específico está sobrerrepresentado o subrepresentado en las muestras. Además, el concepto de distribución en el muestreo nos permite averiguar la probabilidad de obtener una determinada muestra.

Utilizamos la información muestral para hacer inferencias sobre la población de la que procede la muestra. La distribución de todos los valores de interés de esta población puede representarse por medio de una variable aleatoria. Sería demasiado ambicioso intentar describir toda la distribución poblacional basándonos en una pequeña muestra aleatoria de observaciones. Sin embargo, podemos muy bien hacer inferencias bastante sólidas sobre importantes características de la distribución poblacional, como la media y la varianza poblacionales. Por ejemplo, dada una muestra aleatoria del consumo de combustible de 20 automóviles de un determinado modelo, podemos utilizar la media y la varianza muestrales para hacer inferencias sobre la media y la varianza poblacionales del consumo de combustible. Esta inferencia se basará en la información muestral. Podemos hacer preguntas como la siguiente: «Si el consumo de combustible, en kilómetros por litro, de la población de todos los automóviles de un determinado modelo tiene una media de 25 y una desviación típica de 2, ¿cuál es la probabilidad de que el consumo medio muestral de combustible de los automóviles de una muestra aleatoria de 20 sea de menos de 24 kilómetros por litro?» A continuación, podemos utilizar la distribución de la media muestral en el muestreo para responder a esta pregunta.

Necesitamos distinguir entre los atributos de la población y los atributos de la muestra aleatoria. En el párrafo anterior, la población de mediciones del consumo de combustible de todos los automóviles de un determinado modelo sigue una distribución que tiene una determinada media. Esta media, un atributo de la población, es un número fijo (pero desconocido). Hacemos inferencias sobre este atributo extrayendo una muestra aleatoria de la población y calculando la media muestral. Cada muestra que extraigamos tendrá una media muestral distinta y la media muestral puede considerarse como una variable aleatoria con una distribución de probabilidad. La distribución de las medias muestrales posibles constituye la base para realizar inferencias sobre la muestra. En este capítulo, examinamos las propiedades de las *distribuciones en el muestreo*.

Distribuciones en el muestreo

Consideremos una muestra aleatoria extraída de una población que se utiliza para realizar una inferencia sobre alguna característica de la población, como la media poblacional, μ, utilizando un estadístico muestral, como la media muestral, $\bar{x}$. La inferencia se basa en la comprensión de que cada muestra aleatoria tiene una $\bar{x}$ distinta y de que, por lo tanto, $\bar{x}$ es una variable aleatoria. La **distribución en el muestreo** de este estadístico es la distribución de probabilidad de las medias muestrales obtenidas de estas muestras posibles del mismo número de observaciones extraídas de la población.

Ilustramos el concepto de distribución en el muestreo examinando la posición de un supervisor que tiene seis empleados, cuyos años de experiencia son

2 4 6 6 7 8

Hay que elegir aleatoriamente dos de estos empleados para formar un grupo de trabajo. La media de los años de experiencia de esta población de seis empleados es

$$\mu = \frac{2 + 4 + 6 + 6 + 7 + 8}{6} = 5{,}5$$

Examinemos ahora el número medio de años de experiencia de los dos empleados elegidos aleatoriamente de la población de seis. Podrían seleccionarse 15 muestras aleatorias. La Tabla 7.1 presenta todas las muestras posibles y las medias muestrales correspondientes. Obsérvese que algunas muestras (como 2, 6) aparecen dos veces porque hay dos empleados en la población que tienen seis años de experiencia.

Todas las 15 muestras de la Tabla 7.1 tienen la misma probabilidad, 1/15, de ser seleccionadas. Obsérvese que aparece varias veces la misma media muestral. Por ejemplo, la media muestral 5,0 aparece tres veces y, por lo tanto, la probabilidad de obtener una media muestral de 5,0 es 3/15. La Tabla 7.2 presenta la distribución en el muestreo de las medias muestrales de la población y la Figura 7.1 representa gráficamente la función de probabilidad.

Tabla 7.1. Muestras y medias muestrales de la muestra poblacional de trabajadores de tamaño $n = 2$.

Muestra	Media muestral	Muestra	Media muestral
2,4	3,0	4,8	6,0
2,6	4,0	6,6	6,0
2,6	4,0	6,7	6,5
2,7	4,5	6,8	7,0
2,8	5,0	6,7	6,5
4,6	5,0	6,8	7,0
4,6	5,0	7,8	7,5
4,7	5,5		

Tabla 7.2. Distribución de las medias muestrales en el muestreo correspondiente a la muestra poblacional de trabajadores de tamaño $n = 2$.

Media muestral $\bar{X}$	Probabilidad de $\bar{X}$
3,0	1/15
4,0	2/15
4,5	1/15
5,0	3/15
5,5	1/15
6,0	2/15
6,5	2/15
7,0	2/15
7,5	1/15

Vemos que, mientras que el número de años de experiencia de los seis trabajadores va de 2 a 8, los valores posibles de la media muestral van de 3,0 a 7,5 solamente. Además, la mayoría de los valores se encuentran en la parte central del rango.

La Tabla 7.3 muestra que los resultados son parecidos cuando el tamaño de la muestra es $n = 5$ y la Figura 7.2 representa gráficamente la distribución en el muestreo. Obsérvese

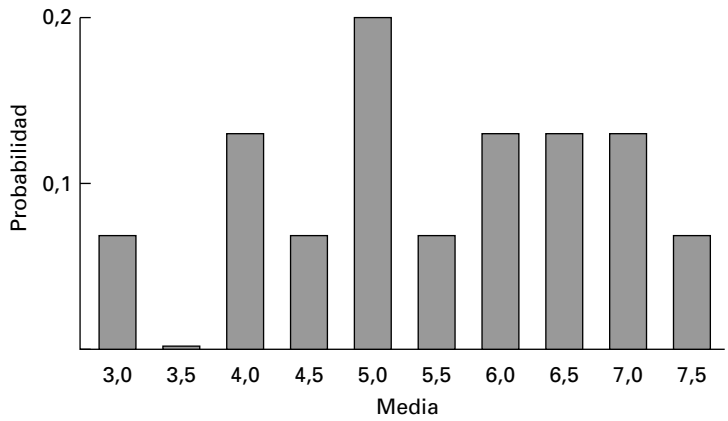


Figura 7.1. Función de probabilidad de la distribución de las medias muestrales en el muestreo correspondiente a la muestra poblacional de trabajadores de tamaño $n = 2$.

Tabla 7.3. Distribución de las medias muestrales en el muestreo correspondiente a la muestra poblacional de trabajadores de tamaño $n = 5$.

Muestra	$\bar{x}$	Probabilidad
2, 4, 6, 6, 7	5,0	1/6
2, 4, 6, 6, 8	5,2	1/6
2, 4, 6, 7, 8	5,4	1/3
2, 6, 6, 7, 8	5,8	1/6
4, 6, 6, 7, 8	6,2	1,/6

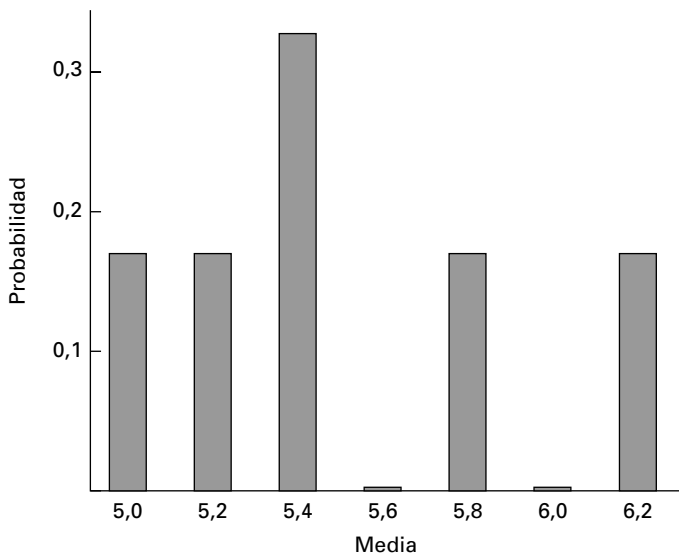


Figura 7.2. Función de probabilidad de la distribución de las medias muestrales en el muestreo correspondiente a la muestra poblacional de trabajadores de tamaño $n = 5$.

que las medias están concentradas en un rango más reducido. Estas medias muestrales están todas más cerca de la media poblacional, $\mu = 5{,}5$. Veremos que eso siempre es cierto: la distribución en el muestreo está más concentrada en torno a la media poblacional a medida que aumenta el tamaño de la muestra. Este importante resultado constituye un importante fundamento para la inferencia estadística. En los apartados y los capítulos siguientes presentaremos un conjunto de rigurosos instrumentos analíticos que se basan en este fundamento.

En este apartado hemos presentado el concepto básico de distribuciones en el muestreo. Los ejemplos procedían de una distribución discreta sencilla en la que es posible definir todas las muestras posibles de un tamaño dado. Hemos calculado la media muestral de cada muestra posible y hemos construido la distribución de probabilidad de todas las medias muestrales posibles. Siguiendo este sencillo método, hemos descubierto que, cuando aumenta el tamaño de la muestra, la distribución de las medias muestrales —la distribución en el muestreo— está más concentrada en torno a la media poblacional. En la mayoría de los estudios estadísticos aplicados, las poblaciones son muy grandes y no es práctico o racional construir la distribución de todas las muestras posibles de un tamaño dado. Pero valiéndonos de lo que hemos aprendido sobre las variables aleatorias, podemos mostrar que las distribuciones en el muestreo de muestras de todas las poblaciones tienen las mismas características que las de nuestra población discreta sencilla. Ese resultado constituye la base de las numerosas y útiles aplicaciones que presentaremos en capítulos posteriores.

EJERCICIOS

Ejercicios básicos

7.1. Suponga que lanza un par de dados al aire y anota el valor de las caras de cada uno.

a) ¿Cuál es la distribución poblacional de un dado?
b) Halle la distribución en el muestreo de las medias muestrales obtenidas lanzando dos dados al aire.

7.2. Suponga que tiene una moneda equilibrada y que le asigna el valor 1 a la cara y el valor 0 a la cruz.

a) Ahora lanza dos veces la moneda al aire y anota el valor numérico obtenido en cada lanzamiento. Sin tirar realmente la moneda al aire, anote la distribución de las medias muestrales en el muestreo.
b) Repita el apartado (a) lanzando cuatro veces la moneda al aire.
c) Repita el apartado (a) lanzando 10 veces la moneda al aire.

Ejercicios aplicados

7.3. Una población contiene 6 millones de 0 y 4 millones de 1. ¿Cuál es la distribución aproximada de la media muestral en el muestreo cuando

a) El tamaño de la muestra es $n = 5$?
b) El tamaño de la muestra es $n = 100$?

Nota: hay una forma fácil y una forma difícil de responder a esta pregunta. Le recomendamos la primera.

7.4. Suponga que un matemático dijera que es imposible obtener una muestra aleatoria simple de una población del mundo real. Por lo tanto, es inútil toda la base para aplicar los métodos estadísticos a los problemas reales. ¿Qué respondería?

7.2. Distribuciones de las medias muestrales en el muestreo

A continuación, mostramos algunas propiedades importantes de la distribución de las medias muestrales en el muestreo. Nuestro análisis comienza con una muestra aleatoria de n observaciones de una población muy grande que tiene una media μ y una varianza σ^2; las observaciones muestrales se representan por medio de X_1, X_2, ..., X_n. Antes de observar la muestra, existe incertidumbre sobre los resultados. Esta incertidumbre se recoge concibiendo las observaciones como variables aleatorias extraídas de una población que tiene una media μ y una varianza σ^2. Lo que nos interesa principalmente es hacer inferencias sobre la media poblacional μ. Un punto de partida obvio es la *media muestral*.

Media muestral

Sean las variables aleatorias X_1, X_2, ..., X_n una muestra aleatoria de una población. La **media muestral** de estas variables aleatorias es

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Consideremos la distribución de la variable aleatoria $\bar{X}$ en el muestreo. De momento no podemos averiguar la forma de la distribución en el muestreo, pero sí su media y su varianza a partir de las definiciones básicas que aprendimos en los Capítulos 5 y 6. En primer lugar, hallamos la media de la distribución. En los citados capítulos vimos que la esperanza de una combinación lineal de variables aleatorias es la combinación lineal de las esperanzas:

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) = \frac{n\mu}{n} = \mu$$

Por lo tanto, la media de la distribución de las medias muestrales en el muestreo es la media poblacional. Si se extrae repetida e independientemente muestras de n observaciones aleatorias e independientes de una población, entonces a medida que aumenta el número de muestras, la media de las medias muestrales se aproxima a la verdadera media poblacional. Este resultado del muestreo aleatorio es importante e indica la protección que dan las muestras aleatorias contra las muestras poco representativas. Una única media muestral podría ser mayor o menor que la media poblacional. Sin embargo, en promedio, no hay razones para esperar que una media muestral sea mayor o menor que la media poblacional. Más adelante en este apartado, se demuestra este resultado utilizando muestras aleatorias obtenidas por ordenador.

EJEMPLO 7.1. Valor esperado de la media muestral (valor esperado)

Calcule el valor esperado de la media muestral del ejemplo del grupo de empleados antes analizado.

Solución

La Tabla 7.2 y la Figura 7.1 muestran la distribución de las medias muestrales en el muestreo. Partiendo de esta distribución, podemos calcular el valor esperado de la media muestral de la forma siguiente:

$$E(\bar{X}) = \sum \bar{x}P(\bar{x}) = (3{,}0)\left(\frac{1}{15}\right) + (4{,}0)\left(\frac{2}{15}\right) + \cdots + (7{,}5)\left(\frac{1}{15}\right) = 5{,}5$$

que es la media poblacional, μ. Se puede hacer un cálculo parecido para obtener el mismo resultado utilizando la distribución en el muestreo de la Tabla 7.3.

Una vez demostrado que la distribución de las medias muestrales está concentrada en torno a la media poblacional, es necesario hallar la varianza de la distribución de medias muestrales. Supongamos que el consumo medio de combustible de una muestra aleatoria de 20 automóviles es de 24 kilómetros por litro. Podemos utilizar la media muestral como estimación de la media poblacional. Pero también queremos saber en qué medida es la me-

dia muestral $\bar{x} = 24$ una buena aproximación de la media poblacional. Para saberlo utilizamos la varianza de la distribución de las medias muestrales en el muestreo.

Si la población es muy grande en comparación con el tamaño de la muestra, las distribuciones de los miembros de muestras aleatorias son aproximadamente independientes entre sí. En los Capítulos 5 y 6 vimos que la varianza de una combinación lineal de variables aleatorias independientes es la suma de los cuadrados de los coeficientes lineales multiplicados por la varianza de las variables aleatorias. Por lo tanto,

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right) = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 \sigma_i^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

La varianza de la distribución de $\bar{X}$ en el muestreo disminuye a medida que aumenta el tamaño de la muestra n. Eso quiere decir, en efecto, que cuanto mayor es el tamaño de la muestra, más concentrada está la distribución en el muestreo. El sencillo ejemplo del apartado anterior muestra este resultado. Por lo tanto, cuanto mayor es la muestra, más seguros estamos de nuestra inferencia de la media poblacional, como cabía esperar. A medida que obtenemos más información de una población —de una muestra mayor— podemos conocer mejor las características de la población, como la media poblacional. La varianza de la media muestral se representa por medio de $\sigma_{\bar{x}}^2$ y la desviación típica correspondiente, llamada error típico de $\bar{X}$, se halla de la siguiente manera:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Si el tamaño de la muestra, n, no es una pequeña proporción del tamaño de la población, N, los miembros de la muestra no están distribuidos independientemente unos de otros. Como un miembro de la población no puede incluirse más de una vez en una muestra, la probabilidad de que un miembro específico de una muestra sea la segunda observación depende del miembro de la muestra elegido como primera observación. Por lo tanto, las observaciones no se seleccionan independientemente. Puede demostrarse en este caso que la varianza de la media muestral es

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

El término $(N - n)/(N - 1)$ a menudo se llama *factor de corrección en el caso de una población finita.*

INTERPRETACIÓN

Hemos presentado ya la media y la varianza de la distribución de $\bar{X}$ en el muestreo. En la mayoría de las aplicaciones, la media y la varianza definen la distribución en el muestreo. Estos resultados de la media y la varianza de la distribución en el muestreo se aplican a cualquier distribución de probabilidad que defina la pauta de los valores existentes en la población. Si fuera imposible generalizar más estos resultados, podrían ser interesantes desde el punto de vista teórico, pero apenas tendrían valor para las aplicaciones prácticas. Afortunadamente, veremos que con algún análisis más estos resultados pueden ser muy poderosos para muchas aplicaciones prácticas. En primer lugar, examinamos estos resultados suponiendo que la población subyacente sigue una distribución normal. A continuación, analizamos las distribuciones de la media muestral en el muestreo a medida que aumenta el tamaño de la muestra. Este segundo caso nos permite obtener algunos resultados muy importantes para muchas aplicaciones empresariales y económicas prácticas.

En primer lugar, examinamos los resultados suponiendo que la población de la que procede la muestra sigue una distribución normal. Esta población es la población de interés de la que se extrae la muestra aleatoria. Si sigue una distribución normal, la distribución de las medias muestrales en el muestreo también sigue una distribución normal. Esta conclusión intuitiva procede del resultado perfectamente demostrado de que las funciones lineales de variables aleatorias que siguen una distribución normal también siguen una distribución normal. En el Capítulo 6 vimos aplicaciones en los problemas de carteras. Con la distribución en el muestreo como una distribución de probabilidad normal, podemos calcular la normal estándar Z de la media muestral. En el Capítulo 6 vimos que podemos utilizar la normal estándar Z para calcular las probabilidades de cualquier variable aleatoria que siga una distribución normal. Ese resultado también se aplica a la media muestral.

Distribución normal estándar de las medias muestrales

Siempre que la distribución de las medias muestrales en el muestreo es una distribución normal, podemos calcular una **variable aleatoria normal estandarizada, *Z***, que tiene una media de 0 y una varianza de 1:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \tag{7.1}$$

Por último, resumimos los resultados de este apartado.

Resultados de la distribución de las medias muestrales en el muestreo

Sea $\bar{X}$ la media muestral de una muestra aleatoria de n observaciones de una población que tiene una media μ_X y una varianza σ^2. En ese caso,

1. La distribución de $\bar{X}$ en el muestreo tiene la media

$$E(\bar{X}) = \mu \tag{7.2}$$

2. La distribución de $\bar{X}$ en el muestreo tiene la desviación típica

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{7.3}$$

   Se llama error típico de $\bar{X}$.

3. Si el tamaño de la muestra, n, no es pequeño en comparación con el tamaño de la población, N, el error típico de $\bar{X}$ es

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N - 1}} \tag{7.4}$$

4. Si la distribución de la población de la que procede la muestra es normal y, por lo tanto, la distribución de las medias muestrales en el muestreo es normal, la variable aleatoria

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \tag{7.5}$$

   sigue una distribución normal estándar de media 0 y de varianza 1.

La Figura 7.3 muestra la distribución en el muestreo de las medias muestrales de muestras de tamaño $n = 25$ y $n = 100$ extraídas de una población que sigue una distribución normal. Las dos distribuciones están centradas en la media, pero a medida que aumenta el tamaño de la muestra, están más concentradas en torno a la media poblacional, ya que el error típico de la media muestral disminuye a medida que aumenta el tamaño de la muestra. Por lo tanto, la probabilidad de que una media muestral se encuentre a una determinada distancia de la media poblacional disminuye a medida que aumenta el tamaño de la muestra.

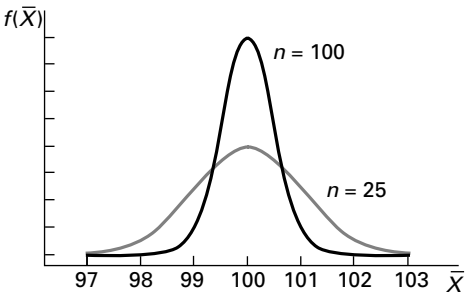


Figura 7.3. Funciones de densidad de medias muestrales de una población de $\mu = 100$ y $\sigma = 5$.

EJEMPLO 7.2. Distribuciones de los sueldos de los ejecutivos (probabilidad normal)

Suponga que las subidas salariales porcentuales anuales de los directores generales de todas las empresas de tamaño medio siguen una distribución normal que tiene una media de 12,2 por ciento y una desviación típica de 3,6 por ciento. Se extrae una muestra aleatoria de nueve observaciones de esta población y se calcula la media muestral. ¿Cuál es la probabilidad de que la media muestral sea inferior a un 10 por ciento?

Solución

Sabemos que

$$\mu = 12{,}2 \qquad \sigma = 3{,}6 \qquad n = 9$$

Sea $\bar{x}$ la media muestral y calculemos su error típico

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3{,}6}{\sqrt{9}} = 1{,}2$$

A continuación, podemos calcular

$$P(\bar{x} < 10) = P\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{10 - 12{,}2}{1{,}2}\right) = P(Z < -1{,}83) = 0{,}0336$$

donde Z sigue una distribución normal estándar y la probabilidad resultante se obtiene en la Tabla 1 del apéndice utilizando los métodos desarrollados en el Capítulo 6.

Este análisis nos permite extraer la conclusión de que la probabilidad de que la media muestral sea inferior a un 10 por ciento es de 0,0336 solamente. Si la media muestral fuera realmente de menos del 10 por ciento, podríamos comenzar a sospechar que la media poblacional es de menos del 12,2 por ciento.

EJEMPLO 7.3. Duración de las bujías (probabilidad normal)

Un fabricante de bujías sostiene que la duración de sus bujías sigue una distribución normal que tiene una media de 36.000 kilómetros y una desviación típica de 4.000 kilómetros. Una muestra aleatoria de 16 bujías tenía una duración media de 34.500 kilómetros. Si la afirmación del fabricante es correcta, ¿cuál es la probabilidad de obtener una media muestral de 34.500 o menos?

Solución

Para calcular la probabilidad, hay que hallar primero el error típico de la media muestral

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4.000}{\sqrt{16}} = 1.000$$

La probabilidad deseada es

$$P(\bar{x} < 34.500) = P\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{34.500 - 36.000}{1.000}\right) = P(Z < -1{,}50) = 0{,}0668$$

La Figura 7.4(a) muestra la función de densidad de $\bar{X}$; el área sombreada indica la probabilidad de que la media muestral sea de menos de 34.500. En la Figura 7.4(b) vemos la función de densidad normal estándar y el área sombreada indica la probabilidad de que Z sea de menos de $-1{,}5$. Obsérvese, que cuando comparamos estas figuras, vemos que a cada valor de $\bar{X}$ le corresponde un valor de Z y las afirmaciones comparables sobre la probabilidad dan el mismo resultado.

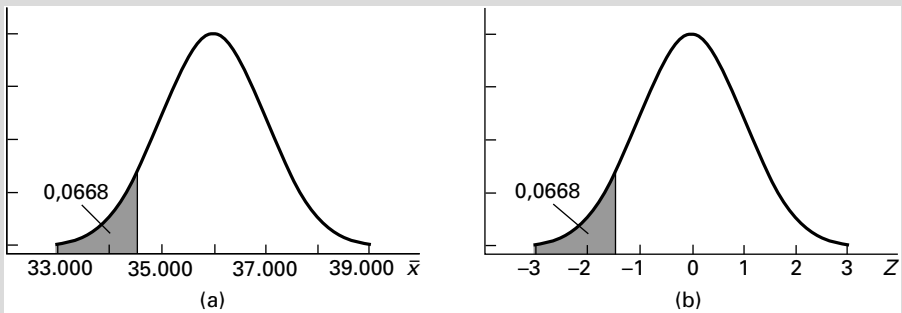


Figura 7.4. (a) Probabilidad de que la media muestral sea de menos de 34.500; (b) probabilidad de que la variable aleatoria normal estándar sea inferior a $-1{,}5$.

Utilizando la Z normal estándar, los valores de la probabilidad normal de la Tabla 1 del apéndice y los métodos del Capítulo 6, observamos que la probabilidad de que $\bar{X}$ sea de menos de 34.500 es 0,0668. Esta probabilidad sugiere que, si las afirmaciones del fabricante —$\mu = 36.000$ y $\sigma = 4.000$— son ciertas, una media muestral de 34.500 o menos tiene una pequeña probabilidad. Por lo tanto, dudamos de las afirmaciones del fabricante. Este importante concepto —la utilización de la probabilidad de estadísticos muestrales para poner en duda el supuesto original— se analizará más extensamente en el Capítulo 10.

Teorema del límite central

En el apartado anterior hemos visto que la media muestral, $\bar{x}$ de una muestra aleatoria de tamaño n extraída de una población que sigue una distribución normal que tiene una media μ y una varianza σ^2, también sigue una distribución normal que tiene una media μ y una varianza σ^2/n. En este apartado presentamos el *teorema del límite central*, que establece que la media de una muestra aleatoria, extraída de una población que tiene cualquier distribución de probabilidad, sigue aproximadamente una distribución normal que tiene una media μ y una varianza σ^2/n, dado un tamaño de la muestra suficientemente grande.

Este importante resultado nos permite utilizar la distribución normal para calcular las probabilidades de medias muestrales extraídas de muchas poblaciones diferentes. En estadística aplicada, a menudo no se conoce la distribución de probabilidad de la población de la que se realiza un muestreo y, en particular, no es posible concluir que la distribución subyacente es normal.

En los análisis estadísticos aplicados, muchas de las variables aleatorias utilizadas pueden caracterizarse por medio de la suma o la media de un gran número de variables aleatorias. Por ejemplo, las ventas diarias totales de una tienda son el resultado de toda una serie de ventas a diferentes clientes, cada una de las cuales puede considerarse que es una variable aleatoria. El gasto de inversión nacional total en un mes es la suma de muchas decisiones de inversión de empresas específicas. Por lo tanto, si X_1, X_2, ..., X_n representa el resultado de sucesos aleatorios, la variable aleatoria observada

$$X = X_1 + X_2 + \cdots + X_n$$

Como vimos en el Capítulo 5,

$$E(X) = n\mu \qquad \text{Var}\,(X) = n\sigma^2$$

El teorema del límite central establece que la suma resultante, X, sigue una distribución normal y puede utilizarse para calcular una variable aleatoria, Z, que tiene una media de 0 y una varianza de 1:

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}\,(X)}} = \frac{X - n\mu}{\sqrt{n\sigma^2}}$$

Además, si dividimos X por n para obtener una media $\bar{X}$, también podemos calcular una Z correspondiente que tiene una media de 0 y una varianza de 1:

$$Z = \frac{\bar{X} - \mu_X}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{\dfrac{\sigma_X}{\sqrt{n}}}$$

Utilizando estos resultados, tenemos el teorema del límite central.

Formulación del teorema del límite central

Sea X_1, X_2, ..., X_n un conjunto de n variables aleatorias independientes que tienen distribuciones idénticas con una media μ y una varianza σ^2. Sea X la suma y $\bar{X}$ la media de estas variables aleatorias. A medida que aumenta n, el **teorema del límite central** establece que la distribución de

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{X - n\mu_X}{\sqrt{n\sigma^2}} \qquad \textbf{(7.6)}$$

tiende a la distribución normal estándar.

El teorema del límite central constituye la base de muchos análisis estadísticos aplicados. Como hemos indicado, muchas variables aleatorias pueden analizarse como sumas o medias de variables aleatorias independientes. Por este teorema, la distribución normal a menudo constituye una buena aproximación de la verdadera distribución. Por lo tanto, la distribución normal estándar puede utilizarse para calcular los valores de la probabilidad de muchas medias muestrales o sumas observadas.

El teorema del límite central puede aplicarse tanto a las variables aleatorias discretas como a las continuas. En el apartado 7.3 utilizamos este teorema con variables aleatorias discretas cuando desarrollamos las probabilidades de variables aleatorias proporcionales utilizando métodos similares a los empleados en el caso de las medias muestrales.

El teorema del límite central es el resultado de una demostración matemática formal que queda fuera del alcance de este libro. Este teorema es un resultado fundamental en el que se basan muchas aplicaciones estadísticas. Los resultados procedentes de simulaciones realizadas mediante muestras aleatorias también pueden utilizarse para demostrarlo. Además, se pueden plantear problemas para hacer caso que permitan al lector realizar un análisis experimental mayor. A continuación, presentamos algunos resultados utilizando simulaciones de Monte Carlo para obtener distribuciones en el muestreo. Para obtener cada uno de estos resultados, seleccionamos 1.000 muestras aleatorias de tamaño n y representamos las distribuciones en el muestreo en histogramas y gráficos de probabilidad normal. En el apéndice del capítulo mostramos el método para obtener distribuciones de las medias muestrales en el muestreo de cualquier distribución de probabilidad. En este apéndice y en el disco de datos incluimos una macro de Minitab para que el lector obtenga fácilmente sus propias distribuciones en el muestreo.

Examinemos primero una distribución de probabilidad uniforme en el rango de 1 a 10. La Figura 7.5 muestra la distribución de probabilidad. Es evidente que los valores de la variable aleatoria no siguen una distribución normal, ya que son uniformes en el rango de 1 a 10. A continuación, mostramos los resultados de las simulaciones por ordenador que generaron muestras aleatorias de diversos tamaños a partir de esta distribución de probabilidad, calcularon la media de cada muestra y prepararon en un histograma la distribución de esas medias muestrales en un muestreo. Este proceso construye distribuciones empíricas de las medias muestrales en el muestreo. Obsérvense los histogramas de las Figura 7.6 y 7.7, que utilizan 1.000 muestras que tienen primero un tamaño $n = 2$ y después un tamaño $n = 25$. Se representa una función de densidad normal con la misma media y la misma varianza sobre cada histograma a modo de comparación.

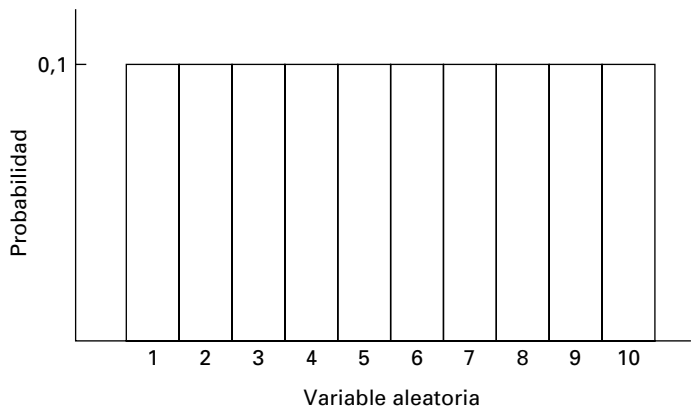


Figura 7.5. Distribución de probabilidad de una variable aleatoria uniforme.

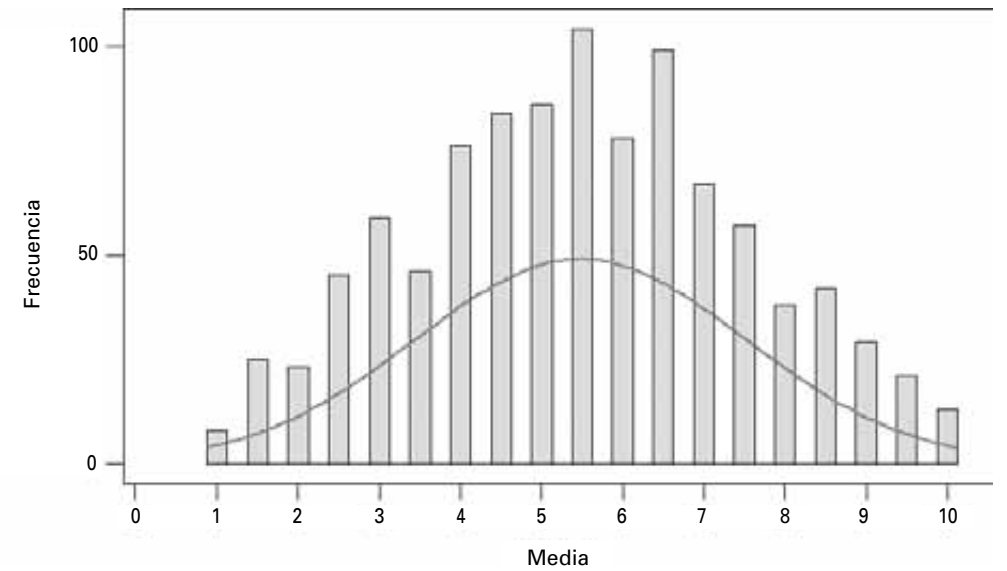


Figura 7.6.
Distribución de las medias muestrales en el muestreo de una distribución uniforme siendo $n = 2$.

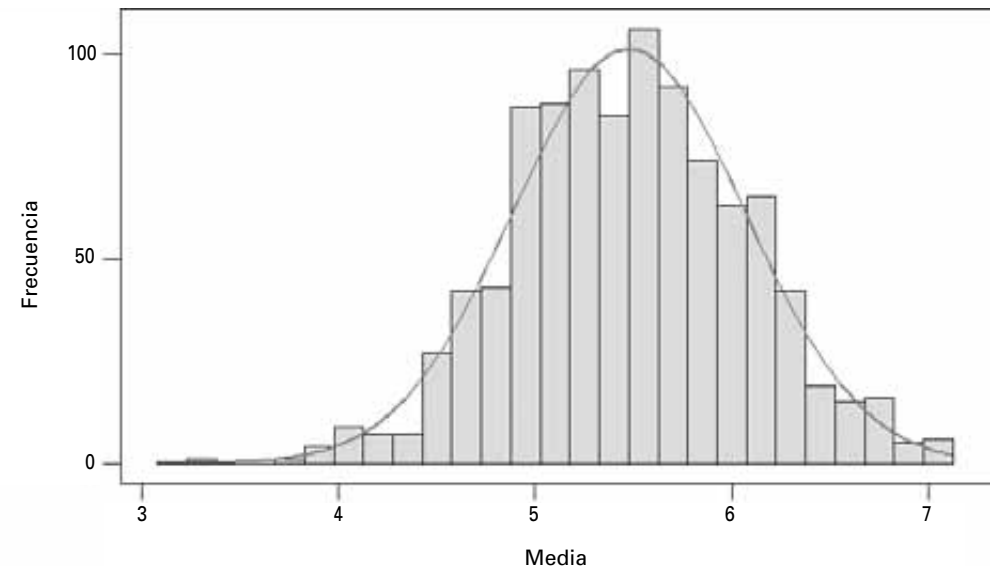


Figura 7.7.
Distribución de las medias muestrales en el muestreo de una distribución uniforme siendo $n = 25$.

Vemos en los histogramas que las medias de muestras de tamaño 2 tienden hacia los valores centrales. Sin embargo, en el caso de las muestras de tamaño 25, el histograma es simétrico y similar a los histogramas muestrales que se obtendrían a partir de una distribución normal. Generalmente, en el caso de la distribución de las medias muestrales de distribuciones uniformes o simétricas, puede utilizarse como aproximación la distribución normal, con muestras de tamaño 25 o más.

Examinemos a continuación una población que tiene una distribución de probabilidad sesgada hacia la derecha. En el Capítulo 2 vimos que las distribuciones de observaciones de muchos procesos empresariales y económicos están sesgadas. Por ejemplo, las rentas familiares y los precios de la vivienda de una ciudad, una región o un país suelen estar sesgados hacia la derecha. Normalmente, hay un pequeño porcentaje de familias que tienen

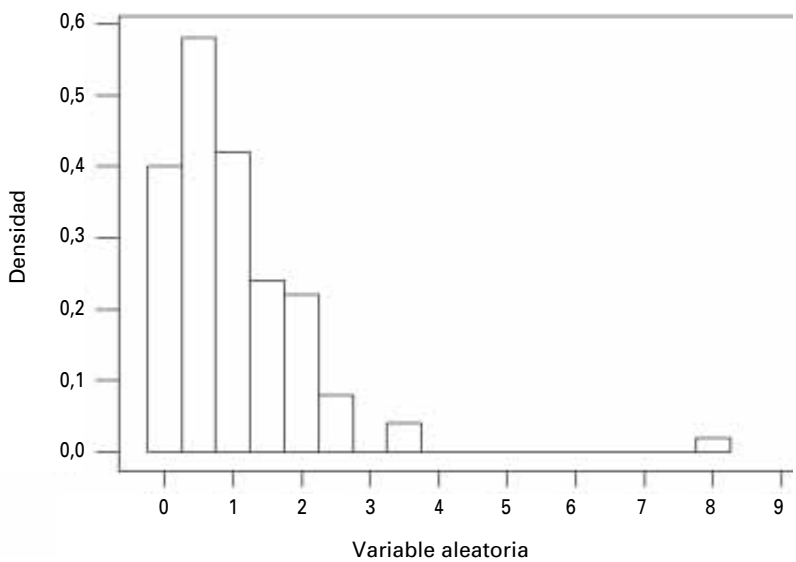


Figura 7.8. Distribución de probabilidad de una distribución sesgada.

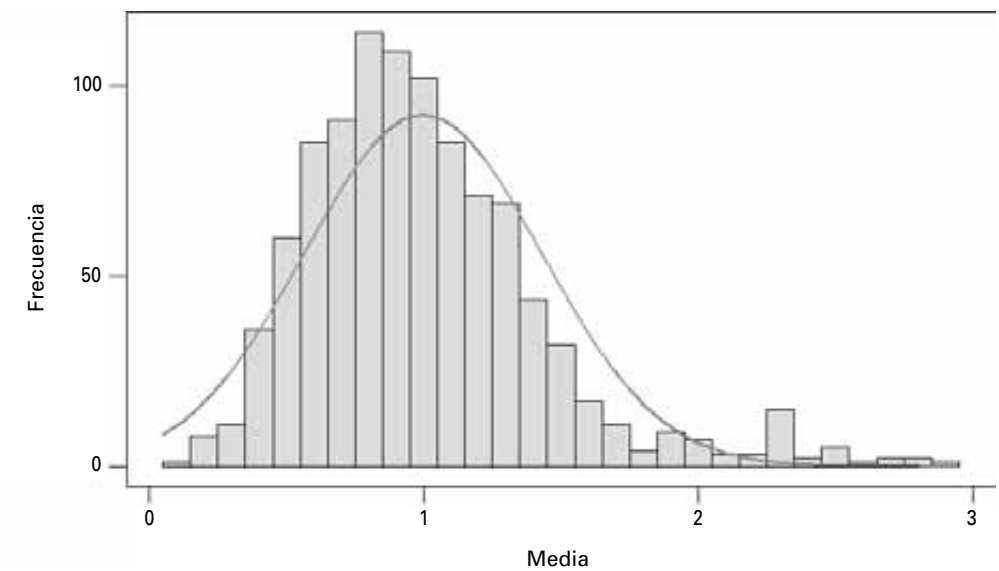


Figura 7.9. Distribución de las medias muestrales en el muestreo: distribución sesgada con una muestra de tamaño $n = 10$.

una renta muy alta y estas familias tienden a vivir en viviendas caras. Consideremos la distribución de probabilidad discreta que muestra la Figura 7.8. Podría ser una distribución de rentas familiares de un país en vías de desarrollo. Supongamos que queremos comparar la renta media de ese país con las medias de un grupo mayor de países que tienen niveles de estudios similares.

Las distribuciones de las rentas medias en el muestreo se comparan utilizando una muestra aleatoria de la distribución de probabilidad. La Figura 7.9 muestra un histograma de 1.000 muestras de tamaño $n = 10$ y la 7.10 muestra un histograma de 1.000 muestras de tamaño $n = 25$. Si utilizamos una muestra aleatoria de tamaño $n = 10$ y suponemos que la media muestral sigue una distribución normal, existen muchas posibilidades de estimar incorrectamente las probabilidades. Estos errores de las estimaciones de las probabilidades

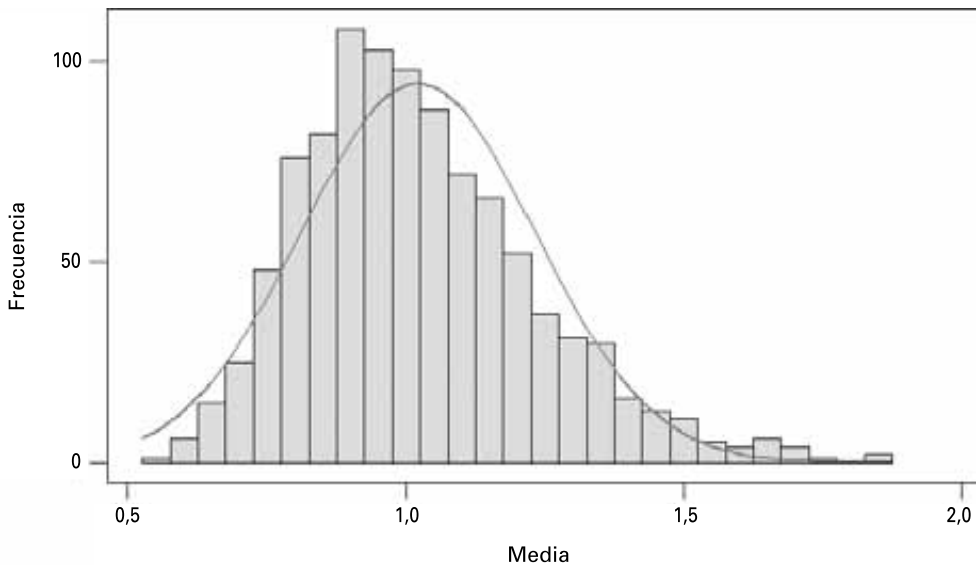


Figura 7.10. Distribución de las medias muestrales en el muestreo: distribución sesgada con una muestra de tamaño $n = 25$.

son especialmente grandes en el caso de las medias muestrales de la cola superior de la distribución. Obsérvese que el histograma es diferente del que se obtendría con una distribución normal. Pero si utilizamos una muestra aleatoria de tamaño $n = 25$, los resultados son mucho mejores. Obsérvese que el segundo histograma —$n = 25$— se parece mucho más a una distribución normal. Si obtuviéramos distribuciones en el muestreo de muestras mayores, los resultados serían incluso mejores. Así pues, incluso cuando la distribución de observaciones está muy sesgada, la distribución de las medias muestrales en el muestreo se parece mucho a una distribución normal cuando $A \geqslant 25$.

En el Capítulo 6 vimos que la variable aleatoria binomial sigue una distribución normal aproximada cuando aumenta el tamaño de la muestra. En el estudio del muestreo aleatorio de este capítulo y en el estudio anterior de la distribución binomial, tenemos pruebas adicionales para demostrar el teorema del límite central. Muchos estadísticos han realizado en numerosas ocasiones demostraciones parecidas, por lo que existen abundantes datos empíricos que sustentan la aplicación del teorema del límite central no sólo a los resultados teóricos sino también a las aplicaciones estadísticas reales.

En el análisis aplicado, la cuestión es saber cuál es el tamaño de la muestra necesario para que las medias muestrales sigan una distribución normal. Sabemos por numerosas investigaciones y por la experiencia que, si las distribuciones son simétricas, la distribución normal es una buena aproximación de las medias de las muestras de tamaño $n = 20$ a 25. En el caso de las distribuciones sesgadas, el tamaño de la muestra generalmente tiene que ser algo mayor. Pero obsérvese que en los ejemplos anteriores que utilizan una distribución sesgada, un tamaño de la muestra de $n = 25$ producía una distribución de las medias muestrales en el muestreo que seguía en gran medida una distribución normal.

En este capítulo hemos comenzado nuestro análisis del importante problema estadístico que se plantea cuando se hacen inferencias sobre una población basándose en los resultados de una muestra. A menudo se calcula la media muestral o la proporción muestral para hacer inferencias sobre medias o proporciones poblacionales. Utilizando el teorema del límite central, tenemos un argumento para aplicar las técnicas que presentaremos en futuros capítulos a una amplia variedad de problemas. Los ejemplos siguientes muestran importantes aplicaciones de este teorema.

EJEMPLO 7.4. Estudio de mercado para Cafés Antílope (probabilidad normal)

Cafés Antílope, S.A., está considerando la posibilidad de abrir una tienda de cafés en Villalegre, ciudad de 50.000 habitantes. Según algunos estudios de mercado realizados anteriormente, sus tiendas tendrán éxito en las ciudades de ese tamaño si la renta anual per cápita es de más de 60.000 $. También se sabe que la desviación típica de la renta es de 5.000 $.

Se ha obtenido una muestra aleatoria de 36 personas y la renta media es de 62.300 $. ¿Constituye esta muestra una prueba para concluir que debe abrirse una tienda?

Solución

Se sabe que la distribución de las rentas está sesgada, pero el teorema del límite central nos permite concluir que la media muestral sigue aproximadamente una distribución normal. Para responder a esta pregunta necesitamos hallar la probabilidad de obtener una media muestral de al menos $\bar{x} = 62.300$ si la media poblacional es $\mu = 60.000$.

Primero calculamos el estadístico Z normal estandarizado,

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{62.300 - 60.000}{\frac{5.000}{\sqrt{36}}} = 2{,}76$$

En la tabla normal estándar observamos que la probabilidad de que el valor de Z sea 2,76 o mayor es 0,0029. Como esta probabilidad es muy baja, podemos concluir que es probable que la renta media de la población no sea de 60.000 $ sino mayor. Este resultado es una poderosa prueba de que la renta media de la población es de más de 60.000 $ y de que la tienda de café probablemente será un éxito. En este ejemplo, podemos ver la importancia de las distribuciones en el muestreo y del teorema del límite central para resolver problemas.

Intervalos de aceptación

En muchas aplicaciones estadísticas, nos gustaría hallar el intervalo en el que es probable que se encuentren las medias muestrales. La determinación de esos intervalos es una aplicación directa de los conceptos de distribución en el muestreo que hemos analizado. Un **intervalo de aceptación** es un intervalo en el que es muy probable que se encuentre una media muestral, dado que conocemos la media y la varianza poblacionales. Si la media muestral se encuentra dentro de ese intervalo, podemos aceptar la conclusión de que la muestra aleatoria procede de la población que tiene la media y la varianza poblacionales conocidas. Es posible calcular la probabilidad de que la media muestral se encuentre dentro de un determinado intervalo si las medias muestrales siguen una distribución aproximadamente normal.

Los intervalos de aceptación basados en la distribución normal vienen definidos por la media y la varianza de la distribución. Sabemos por el teorema del límite central que la distribución de las medias muestrales en el muestreo a menudo es aproximadamente normal y, por lo tanto, los intervalos de aceptación basados en la distribución normal tienen

muchas aplicaciones. Suponiendo que conocemos la media poblacional μ y la varianza poblacional σ^2, podemos construir un intervalo de aceptación simétrico:

$$\mu \pm Z_{\alpha/2}\sigma_{\bar{x}}$$

siempre que $\bar{x}$ siga una distribución normal y $z_{\alpha/2}$ sea la normal estándar cuando la probabilidad de la cola superior es $\alpha/2$. La probabilidad de que la media muestral $\bar{x}$ esté incluida en el intervalo es $1 - \alpha$.

Los intervalos de aceptación se emplean mucho para el control de calidad de muchos procesos de producción y servicios. Se representa el intervalo

$$\mu \pm Z_{\alpha/2}\sigma_{\bar{x}}$$

con respecto al tiempo (el resultado se llama gráfico X-barra), que nos da los límites de la media muestral $\bar{x}$, dada la media poblacional μ. Normalmente, el valor de α es muy bajo ($\alpha < 0{,}01$) y en las empresas estadounidenses normalmente se emplea $z = 3$. Si la media muestral está fuera del intervalo de aceptación, sospechamos que la media poblacional no es μ. Generalmente, los ingenieros siguen varios pasos para lograr una pequeña varianza para realizar importantes mediciones de los productos que están relacionadas directamente con su calidad. Una vez que el proceso se ha ajustado de manera que la varianza es pequeña, se establece un intervalo de aceptación para una media muestral —llamado *intervalo de control*— en forma de gráfico de control. A continuación, se obtienen muestras aleatorias periódicas y se comparan con el intervalo de control. Si la media muestral está dentro del intervalo de control, se concluye que el proceso está funcionando bien y no se toma ninguna medida. Pero si la media muestral está fuera del intervalo de control, se concluye que el proceso no está funcionando bien y se toman medidas para corregirlo. En el Capítulo 18 analizamos los gráficos de control mucho más extensamente.

EJEMPLO 7.5. Control de las reclamaciones presentadas en una compañía de seguros médicos (intervalo de aceptación)

Carlota Reina, vicepresidenta financiera de una gran compañía de seguros médicos, quiere controlar los desembolsos diarios por reclamaciones para averiguar si el número medio de reclamaciones por suscriptor se mantiene estable, está aumentando o está disminuyendo. El número de reclamaciones varía de un día al siguiente y sería ingenuo extraer conclusiones o cambiar las operaciones basándose en estas variaciones diarias. Pero en un momento dado los cambios son sustanciales y deben señalarse. Le ha pedido que desarrolle un método para controlar el nivel de reclamaciones.

Solución

Nuestra investigación inicial indica que las reclamaciones están muy sesgadas y que el número de grandes reclamaciones es pequeño. Para averiguar si ha habido cambios, primero tenemos que hallar la media y la varianza históricas de las reclamaciones. Tras algunas investigaciones, también observamos que la media de muestras aleatorias de $n = 100$ reclamaciones sigue una distribución normal. Basándonos en la historia, el nivel medio de reclamaciones, μ, es 4.000 $ con una desviación típica de $\sigma = 2.000$.

Basándonos en esta información, procedemos a desarrollar un sistema de control de las reclamaciones que obtiene una muestra aleatoria de 100 reclamaciones diarias y calcula la media muestral. La compañía ha establecido un intervalo de aceptación del

95 por ciento para el control de las reclamaciones. Un intervalo definido para la normal estándar utilizando $Z = \pm 1{,}96$ incluye el 95 por ciento de los valores. A partir de este resultado, calculamos el intervalo de aceptación del 95 por ciento para las reclamaciones de la forma siguiente:

$$4.000 \pm 1{,}96 \frac{2.000}{\sqrt{100}} = 4.000 \pm 392$$

Cada día se calcula la media muestral de 100 reclamaciones seleccionadas aleatoriamente y se compara con el intervalo de aceptación. Si la media muestral está fuera del intervalo 3.608 a 4.392, Carlota Reina puede concluir que las reclamaciones están desviándose del patrón histórico. Le explicamos que esta conclusión será correcta el 95 por ciento de las veces. La media muestral podría estar fuera del intervalo con una probabilidad de 0,05 incluso con una media poblacional de 4.000. En esos casos, la conclusión de Carlota Reina de que el nivel medio de reclamaciones ha cambiado con respecto al patrón histórico sería errónea. Para simplificar el análisis, damos a los analistas instrucciones para que representen la media diaria de reclamaciones en un gráfico de control, mostrado en la Figura 7.11. Utilizando este gráfico, Carlota Reina y su equipo pueden estudiar las pautas de las medias muestrales y averiguar si hay tendencias y si las medias están fuera de los límites que indica la conducta histórica de las reclamaciones.

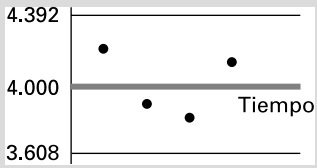


Figura 7.11. Intervalo de aceptación del 95 por ciento para las reclamaciones al seguro médico.

EJEMPLO 7.6. Peso de las cajas de cereales de Flores de la Pradera (intervalos de aceptación)

Cereales Flores de la Pradera, S.A., quiere que el peso de sus cajas de cereales sea correcto. Las cajas indican que su peso es de 440 gramos y la empresa tiene interés en controlar el proceso para garantizar que el peso de las cajas es estable.

Solución

Se recoge una muestra aleatoria de cinco cajas cada 30 minutos y se pesa electrónicamente cada una. A continuación, se representa el peso medio en un gráfico de control X-barra como el de la Figura 7.12. Cuando se utiliza un gráfico X-barra para controlar los límites de la calidad de un producto —y muchas empresas prósperas lo hacen— el teorema del límite central constituye la razón para utilizar la distribución normal a fin de establecer los límites de las pequeñas medias muestrales. Así pues, una importante teoría estadística impulsa un proceso clave de gestión.

En este gráfico, SL es la desviación típica de la media muestral. Los límites superior e inferior se fijan en $\pm 3\sigma_{\bar{X}}$ en lugar de $\pm 1{,}96\sigma_{\bar{X}}$, o sea, un 95 por ciento, que es el intervalo de aceptación utilizado en el ejemplo anterior. El intervalo $\bar{X} \pm 3\sigma_{\bar{X}}$ (el programa Minitab pone dos barras cuando se refiere a la media de toda la población: $\bar{\bar{X}}$)

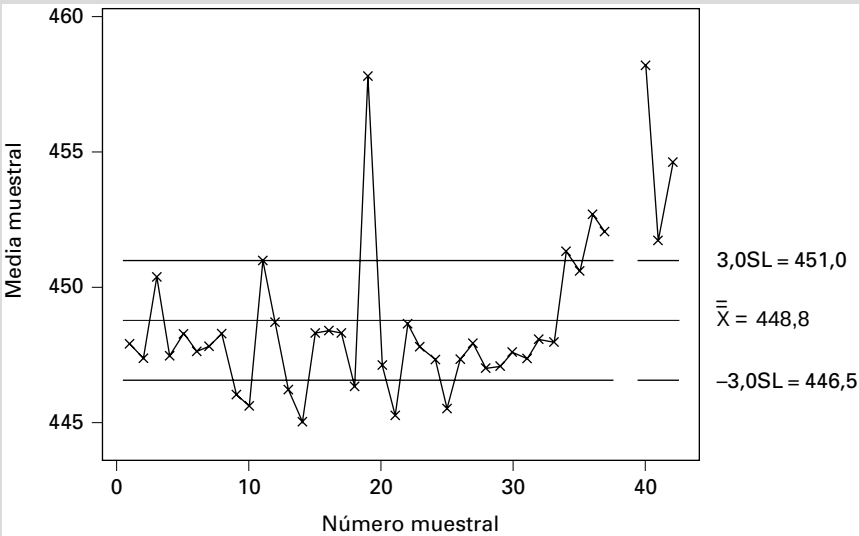


Figura 7.12. Gráfico X-barra del peso de las cajas de cereales.

comprende casi todas las medias muestrales cuando la distribución es normal, siempre que la media y la varianza sean estables. Por lo tanto, una media muestral situada fuera de los límites de control indica que algo ha cambiado y que deben hacerse correcciones. Dado el número de puntos situados fuera del intervalo de aceptación, recomendamos que el proceso se detenga y se ajuste.

EJERCICIOS

Ejercicios básicos

7.5. Dada una población de media $\mu = 100$ y varianza $\sigma^2 = 81$, el límite central se aplica cuando el tamaño de la muestra es $n \geqslant 25$. Se obtiene una muestra aleatoria de tamaño $n = 25$.

a) ¿Cuáles son la media y la varianza de la distribución de las medias muestrales en el muestreo?
b) ¿Cuál es la probabilidad de que $\bar{x} > 102$?
c) ¿Cuál es la probabilidad de que $98 \leqslant \bar{x} \leqslant 101$?
d) ¿Cuál es la probabilidad de que $\bar{x} \leqslant 101{,}5$?

7.6. Dada una población de media $\mu = 100$ y varianza $\sigma^2 = 900$, el límite central se aplica cuando el tamaño de la muestra es $n \geqslant 25$. Se obtiene una muestra aleatoria de tamaño $n = 30$.

a) ¿Cuáles son la media y la varianza de la distribución de las medias muestrales en el muestreo?
b) ¿Cuál es la probabilidad de que $\bar{x} > 109$?
c) ¿Cuál es la probabilidad de que $96 \leqslant \bar{x} \leqslant 110$?
d) ¿Cuál es la probabilidad de que $\bar{x} \leqslant 107$?

7.7. Dada una población de media $\mu = 200$ y varianza $\sigma^2 = 625$, el límite central se aplica cuando el tamaño de la muestra es $n \geqslant 25$. Se obtiene una muestra aleatoria de tamaño $n = 25$.

a) ¿Cuáles son la media y la varianza de la distribución de las medias muestrales en el muestreo?
b) ¿Cuál es la probabilidad de que $\bar{x} > 209$?
c) ¿Cuál es la probabilidad de que $198 \leqslant \bar{x} \leqslant 211$?
d) ¿Cuál es la probabilidad de que $\bar{x} \leqslant 202$?

7.8. Dada una población de media $\mu = 400$ y varianza $\sigma^2 = 1.600$, el límite central se aplica cuando el tamaño de la muestra es $n \geqslant 25$. Se obtiene una muestra aleatoria de tamaño $n = 35$.

a) ¿Cuáles son la media y la varianza de la distribución de las medias muestrales en el muestreo?
b) ¿Cuál es la probabilidad de que $\bar{x} > 412$?
c) ¿Cuál es la probabilidad de que $393 \leqslant \bar{x} \leqslant 407$?
d) ¿Cuál es la probabilidad de que $\bar{x} \leqslant 389$?

Ejercicios aplicados

7.9. Cuando un proceso de producción funciona correctamente, el número de unidades producidas por hora sigue una distribución normal que tiene una media de 92,0 y una desviación típica de 3,6. Se ha tomado una muestra aleatoria de cuatro horas distintas.

a) Halle la media de la distribución de las medias muestrales en el muestreo.
b) Halle la varianza de la media muestral.
c) Halle el error típico de la media muestral.
d) ¿Cuál es la probabilidad de que la media muestral sea de más de 93,0 unidades?

7.10. La duración de las bombillas de un fabricante tiene una media de 1.200 horas y una desviación típica de 400 horas. La población sigue una distribución normal. Suponga que compra nueve bombillas, que puede considerarse que son una muestra aleatoria de la producción del fabricante.

a) ¿Cuál es la media de la media muestral de la duración?
b) ¿Cuál es la varianza de la media muestral?
c) ¿Cuál es el error típico de la media muestral?
d) ¿Cuál es la probabilidad de que esas nueve bombillas tengan, en promedio, una duración de menos de 1.050 horas?

7.11. El consumo de combustible, en kilómetros por litro, de todos los automóviles de un determinado modelo tiene una media de 25 y una desviación típica de 2. Puede suponerse que la distribución poblacional es normal. Se toma una muestra aleatoria de estos automóviles.

a) Halle la probabilidad de que la media muestral del consumo de combustible sea inferior a 24 kilómetros por litro suponiendo que

i. Se toma una muestra de 1 observación.
ii. Se toma una muestra de 4 observaciones.
iii. Se toma una muestra de 16 observaciones.

b) Explique por qué las tres respuestas del apartado (a) son diferentes. Trace un gráfico para explicar su razonamiento.

7.12. El precio medio de venta de las viviendas nuevas fue en una ciudad de 115.000 $ durante un año. La desviación típica poblacional fue de 25.000 $. Se extrajo una muestra aleatoria de 100 ventas de viviendas nuevas de esta ciudad.

a) ¿Cuál es la probabilidad de que la media muestral de los precios de venta fuera de más de 110.000 $?
b) ¿Cuál es la probabilidad de que la media muestral de los precios de venta estuviera comprendida entre 113.000 $ y 117.000 $?
c) ¿Cuál es la probabilidad de que la media muestral de los precios de venta estuviera comprendida entre 114.000 $ y 116.000 $?
d) Indique sin realizar los cálculos en cuál de los intervalos siguientes es más probable que se encuentre la media muestral de los precios de venta:

113.000 $-115.000 $	114.000 $-116.000 $
115.000 $-117.000 $	116.000 $-118.000 $

e) Suponga que, una vez realizados estos cálculos, un amigo le dijera que es casi seguro que la distribución poblacional de los precios de venta de las viviendas nuevas de esta ciudad no sea normal. ¿Qué respondería?

7.13. Los aspirantes a bomberos tienen que aprobar un examen escrito de aptitud. Las calificaciones de este examen siguen una distribución normal que tiene una media de 280 y una desviación típica de 60. Se ha tomado una muestra aleatoria de nueve calificaciones.

a) ¿Cuál es el error típico de la media muestral de las calificaciones?
b) ¿Cuál es la probabilidad de que la media muestral de las calificaciones sea superior a 270?
c) ¿Cuál es la probabilidad de que la media muestral de las calificaciones sea inferior a 250?
d) Suponga que la desviación típica poblacional es, en realidad, de 40 en lugar de 60. Indique sin realizar los cálculos cómo cambiaría eso sus respuestas a los apartados (a), (b) y (c). Ilustre gráficamente sus conclusiones.

7.14. Se ha tomado una muestra aleatoria de 16 directivos de empresas de una gran ciudad para estimar el tiempo medio que tardan diariamente en desplazarse al trabajo. Suponga que el tiempo poblacional sigue una distribución normal que tiene una media de 87 minutos y una desviación típica de 22 minutos.

a) ¿Cuál es el error típico de la media muestral de los tiempos de desplazamiento?
b) ¿Cuál es la probabilidad de que la media muestral sea de menos de 100 minutos?
c) ¿Cuál es la probabilidad de que la media muestral sea de más de 80 minutos?
d) ¿Cuál es la probabilidad de que la media muestral esté fuera del intervalo 85-95 minutos?
e) Suponga que se toma una segunda muestra aleatoria (independiente) de 50 directivos. In-

dique sin realizar los cálculos si las probabilidades de los apartados (b), (c) y (d) serían mayores, menores o iguales que en el caso de la segunda muestra. Ilustre sus respuestas gráficamente.

7.15. Una empresa produce cereales de desayuno. El verdadero peso medio de sus cajas de cereales es de 200 gramos y la desviación típica es de 60 gramos. La distribución poblacional del peso es normal. Suponga que compra cuatro cajas, que puede considerarse que son una muestra aleatoria de todas las que se producen.

a) ¿Cuál es el error típico de la media muestral del peso?

b) ¿Cuál es la probabilidad de que el contenido de estas cuatro cajas pese, en promedio, menos de 197 gramos?

c) ¿Cuál es la probabilidad de que el contenido de estas cuatro cajas pese, en promedio, más de 206 gramos?

d) ¿Cuál es la probabilidad de que el contenido de estas cuatro cajas pese, en promedio, entre 195 y 205 gramos?

e) Se eligen aleatoriamente dos de las cuatro cajas. ¿Cuál es la probabilidad de que el contenido de estas dos cajas pese, en promedio, entre 195 y 205 gramos?

7.16. Suponga que la desviación típica de los alquileres mensuales que pagan los estudiantes en una ciudad es de 40 $. Se toma una muestra aleatoria de 100 estudiantes para estimar el alquiler mensual medio que paga toda la población estudiantil.

a) ¿Cuál es el error típico de la media muestral de los alquileres mensuales?

b) ¿Cuál es la probabilidad de que la media muestral sea más de 5 $ superior a la media poblacional?

c) ¿Cuál es la probabilidad de que la media muestral sea más de 4 $ inferior a la media poblacional?

d) ¿Cuál es la probabilidad de que la media muestral difiera más de 3 $ de la media poblacional?

7.17. El tiempo que dedican los estudiantes a estudiar la semana antes de los exámenes finales sigue una distribución normal que tiene una desviación típica de 8 horas. Se toma una muestra aleatoria de 4 estudiantes para estimar el tiempo medio de estudio de la población total de estudiantes.

a) ¿Cuál es la probabilidad de que la media muestral sea más de 2 horas superior a la media poblacional?

b) ¿Cuál es la probabilidad de que la media muestral sea más de 3 horas inferior a la media poblacional?

c) ¿Cuál es la probabilidad de que la media muestral difiera más de 4 horas de la media poblacional?

d) Suponga que se toma una segunda muestra aleatoria (independiente) de 10 estudiantes. Indique sin realizar los cálculos si las probabilidades de los apartados (a), (b) y (c) serían mayores, menores o iguales que en el caso de la segunda muestra.

7.18. Un proceso industrial produce lotes de un producto químico cuyos niveles de impurezas siguen una distribución normal que tiene una desviación típica de 1,6 gramos por 100 gramos de producto químico. Se selecciona una muestra aleatoria de 100 lotes para estimar el nivel de impureza medio poblacional.

a) La probabilidad de que la media muestral del nivel de impurezas sea _____ mayor que la media poblacional es de 0,05.

b) La probabilidad de que la media muestral del nivel de impurezas sea _____ menor que la media poblacional es de 0,10.

c) La probabilidad de que la media muestral del nivel de impurezas difiera en _____ de la media poblacional es de 0,15.

7.19. Las relaciones precio-beneficio de todas las empresas cuyas acciones cotizan en bolsa siguen una distribución normal que tiene una desviación típica de 3,8. Se selecciona una muestra aleatoria de estas empresas para estimar la relación precio-beneficio media poblacional.

a) ¿Cuál debe ser el tamaño de la muestra para garantizar que la probabilidad de que la media muestral difiera más de 1,0 de la media poblacional es de menos de 0,10?

b) Indique sin realizar los cálculos si sería necesaria una muestra mayor o menor que la del apartado (a) para garantizar que la probabilidad de que la media muestral difiera en más de 1,0 de la media poblacional es de menos de 0,05.

c) Indique sin realizar los cálculos si sería necesaria una muestra mayor o menor que la del apartado (a) para garantizar que la probabilidad de que la media muestral difiera en más de 1,5 de la media poblacional es de menos de 0,05.

7.20. El número de horas que dedican los estudiantes de una gran universidad a estudiar la semana antes de los exámenes finales sigue una distribución normal que tiene una desviación típica de

8,4 horas. Se toma una muestra aleatoria de estos estudiantes para estimar el número medio de horas de estudio de esta población.

a) ¿De qué tamaño tiene que ser la muestra para garantizar que la probabilidad de que la media muestral difiera en más de 2,0 horas de la media poblacional es de menos de 0,05?

b) Indique sin realizar los cálculos si sería necesaria una muestra mayor o menor que la del apartado (a) para garantizar que la probabilidad de que la media muestral difiera en más de 2,0 horas de la media poblacional es de menos de 0,10.

c) Indique sin realizar los cálculos si sería necesaria una muestra mayor o menor que la de la parte (a) para garantizar que la probabilidad de que la media muestral difiera en más de 1,5 horas de la media poblacional es de menos de 0,05.

7.21. En la Tabla 7.1 y en el ejemplo 7.1, examinamos muestras de $n=2$ observaciones de una población de $N=6$ valores de años de experiencia de los empleados. La media poblacional es $\mu=5{,}5$ años.

a) Confirme con los seis valores de la población que la varianza poblacional es

$$\sigma^2 = 3{,}92$$

b) Confirme, siguiendo el método del ejemplo 7.1, que la varianza de la distribución de la media muestral en el muestreo es

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{15} (\bar{x}_i - \mu)^2 P(x_i) = 1{,}57$$

c) Verifique en este ejemplo que

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

7.22. En una muestra de n observaciones de una población de N miembros, la varianza de la distribución de las medias muestrales en el muestreo es

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

La cantidad $\dfrac{(N-n)}{(N-1)}$ se llama *factor de corrección en el caso de una población finita.*

a) Para hacerse una idea de las magnitudes posibles del factor de corrección en el caso de una población finita, calcúlelo para muestras de $n = 20$ observaciones de poblaciones de $N = 20$, 40, 100, 1.000 y 10.000 miembros.

b) Explique por qué el resultado correspondiente a $N = 20$, que ha obtenido en el apartado (a), es precisamente el que sería de esperar intuitivamente.

c) Dados los resultados del apartado (a), analice la importancia práctica de la utilización del factor de corrección en el caso de una población finita con muestras de 20 observaciones de poblaciones de diferentes tamaños.

7.23. Una ciudad tiene 500 agencias inmobiliarias. El valor medio de las propiedades vendidas en un año por estas agencias es de 800.000 $ y la desviación típica es de 300.000 $. Se selecciona una muestra aleatoria de 100 agencias y se anota el valor de las propiedades que venden en un año.

a) ¿Cuál es el error típico de la media muestral?

b) ¿Cuál es la probabilidad de que la media muestral sea de más de 825.000 $?

c) ¿Cuál es la probabilidad de que la media muestral sea de más de 780.000 $?

d) ¿Cuál es la probabilidad de que la media muestral esté comprendida entre 790.000 $ y 820.000 $?

7.24. En un curso de economía hay 250 estudiantes. Se pide a cada miembro de una muestra aleatoria de 50 de estos estudiantes que estime la cantidad de tiempo que ha dedicado a hacer los ejercicios que puso el profesor la semana pasada. Suponga que la desviación típica poblacional es de 30 minutos.

a) ¿Cuál es la probabilidad de que la media muestral sea más de 2,5 minutos superior a la media poblacional?

b) ¿Cuál es la probabilidad de que la media muestral sea más de 5 minutos inferior a la media poblacional?

c) ¿Cuál es la probabilidad de que la media muestral difiera en más de 10 minutos de la media poblacional?

7.25. El tiempo medio de desplazamiento de 600 personas que asistieron a un concierto fue de 32 minutos y la desviación típica fue de 10 minutos. Se tomó una muestra aleatoria de 150 asistentes.

a) ¿Cuál es la probabilidad de que la media muestral del tiempo de desplazamiento fuera de más de 31 minutos?

b) ¿Cuál es la probabilidad de que la media muestral del tiempo de desplazamiento fuera de menos de 33 minutos?

c) Explique gráficamente por qué las respuestas a los apartados (a) y (b) son iguales.

d) ¿Cuál es la probabilidad de que la media muestral del tiempo de desplazamiento no esté comprendida entre 31 y 33 minutos?

7.3. Distribuciones de proporciones muestrales en el muestreo

En el apartado 5.4 dijimos que la distribución binomial era la suma de n variables aleatorias de Bernoulli independientes, cada una de las cuales tenía una probabilidad de éxito P. Para caracterizar la distribución, necesitamos saber cuál es el valor de P. Aquí indicamos cómo podemos utilizar la proporción muestral para hacer inferencias sobre la proporción poblacional. La variable aleatoria proporcional tiene muchas aplicaciones, entre las cuales se encuentran la cuota porcentual de mercado, el porcentaje de inversiones empresariales que tienen éxito y los resultados electorales.

Proporción muestral

Sea X el número de éxitos en una muestra binomial de n observaciones cuyo parámetro es P. El parámetro es la proporción de miembros de la población que tienen una característica de interés. La **proporción muestral** es

$$\hat{P} = \frac{X}{n} \tag{7.7}$$

X es la suma de un conjunto de n variables aleatorias de Bernoulli independientes, cada una de las cuales tiene una probabilidad de éxito P. Por lo tanto, $\hat{P}$ es la media de un conjunto de variables aleatorias independientes y se aplican los resultados que hemos obtenido en los apartados anteriores para las medias muestrales. Además, puede utilizarse el teorema del límite central para sostener que la distribución de probabilidad de $\hat{P}$ puede considerarse una variable aleatoria que sigue una distribución normal.

En el apartado 6.4 mostramos que el número de éxitos en una distribución binomial y la proporción de éxitos tienen una distribución de la que la distribución normal es una buena aproximación (véanse las Figuras 6.23 y 6.24). La aproximación es muy buena cuando $nP(1 - P) > 9$.

La media y la varianza de la distribución de la proporción muestral $\hat{P}$ en el muestreo pueden hallarse a partir de la media y la varianza del número de éxitos, X.

$$E(X) = nP \qquad \text{Var}(X) = nP(1 - P)$$

y, por lo tanto,

$$E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = P$$

Vemos que la media de la distribución de $\hat{P}$ es la proporción poblacional, P.

La varianza de $\hat{P}$ es la varianza de la distribución poblacional de las variables aleatorias de Bernoulli dividida por n.

$$\sigma^2_{\hat{P}} = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2}\text{Var}(X) = \frac{P(1 - P)}{n}$$

La desviación típica de $\hat{P}$, que es la raíz cuadrada de la varianza, se llama su error típico.

Dado que la distribución de la proporción muestral es aproximadamente normal cuando el tamaño de la muestra es grande, podemos obtener una variable aleatoria normal estándar restando P de $\hat{P}$ y dividiendo por el error típico.

Distribución de la proporción muestral en el muestreo

Sea $\hat{P}$ la proporción muestral de éxitos en una muestra aleatoria extraída de una población en la que la proporción de éxitos es P. En ese caso,

1. La distribución de $\hat{P}$ en el muestreo tiene una media P:

$$E(\hat{P}) = P \tag{7.8}$$

2. La distribución de p en el muestreo tiene una desviación típica

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} \tag{7.9}$$

3. Si el tamaño de la muestra es grande, la variable aleatoria

$$Z = \frac{\hat{P} - P}{\sigma_{\hat{P}}} \tag{7.10}$$

está distribuida aproximadamente como una normal estándar. Esta aproximación es buena si

$$nP(1-P) > 9$$

Vemos que, al igual que en el apartado anterior, el error típico de la proporción muestral, $\hat{P}$, disminuye a medida que aumenta el tamaño de la muestra y la distribución está más concentrada, como se observa en la Figura 7.13. Este resultado es de esperar, ya que la proporción muestral es una media muestral. Cuando el tamaño de la muestra es mayor, nuestras inferencias sobre la proporción poblacional mejoran. Sabemos por el teorema del límite central que la distribución normal puede utilizarse como aproximación de la distribución binomial con las correspondientes media y varianza. Vemos este resultado en los siguientes ejemplos.

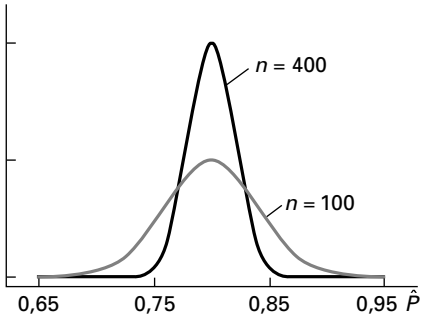


Figura 7.13. Funciones de densidad de proporciones muestrales.

EJEMPLO 7.7. Evaluación del estado de la instalación eléctrica de las viviendas (probabilidad de la proporción muestral)

Se ha extraído una muestra aleatoria de 250 viviendas de una gran población de viviendas antiguas para estimar la proporción cuya instalación eléctrica es peligrosa. Si el 30 por ciento de las viviendas tiene realmente una instalación eléctrica peligrosa, ¿cuál es la probabilidad de que la proporción de edificios de la muestra que tienen una instalación eléctrica peligrosa esté comprendida entre el 25 y el 35 por ciento?

Solución

En este problema, tenemos que

$$P = 0{,}30 \qquad n = 250$$

Podemos calcular la desviación típica de la proporción muestral, $\hat{P}$:

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0{,}30(1-0{,}30)}{250}} = 0{,}029$$

La probabilidad que buscamos es

$$\begin{aligned} P(0{,}25 < \hat{P} < 0{,}35) &= P\left(\frac{0{,}25-P}{\sigma_{\hat{P}}} < \frac{\hat{P}-P}{\sigma_{\hat{P}}} < \frac{0{,}35-P}{\sigma_{\hat{P}}}\right) \\ &= P\left(\frac{0{,}25-0{,}30}{0{,}029} < Z < \frac{0{,}35-0{,}30}{0{,}029}\right) \\ &= P(-1{,}72 < Z < 1{,}72) \\ &= 0{,}9146 \end{aligned}$$

donde la probabilidad del intervalo Z se obtiene en la Tabla 1 del apéndice.

Vemos, pues, que la probabilidad de que la proporción muestral esté comprendida en el intervalo 0,25 a 0,35, dado $P = 0{,}30$, es 0,9146. Este intervalo puede denominarse intervalo de aceptación del 91,46 por ciento. También podemos señalar que, si la proporción muestral estuviera realmente fuera de este intervalo, podríamos comenzar a sospechar que la proporción poblacional, P, no es 0,30.

EJEMPLO 7.8. Selección de una asignatura en un programa de administración de empresas (probabilidad de la proporción muestral)

Se ha estimado que el 43 por ciento de los licenciados en administración de empresas cree que la asignatura de ética empresarial es muy importante para impartir valores éticos a los estudiantes (véase la referencia bibliográfica 1). Halle la probabilidad de que más de la mitad de una muestra aleatoria de 80 licenciados crea eso.

Solución

Tenemos que

$$P = 0{,}43 \qquad n = 80$$

Primero calculamos la desviación típica de la proporción muestral:

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0{,}43(1-0{,}43)}{80}} = 0{,}055$$

A continuación calculamos la probabilidad que buscamos:

$$\begin{aligned} P(\hat{P} > 0{,}50) &= P\left(\frac{\hat{P}-P}{\sigma_{\hat{P}}} > \frac{0{,}50-P}{\sigma_{\hat{P}}}\right) \\ &= P\left(Z > \frac{0{,}50-0{,}43}{0{,}055}\right) \\ &= P(Z > 1{,}27) \\ &= 0{,}1020 \end{aligned}$$

Esta probabilidad, mostrada en la Figura 7.14, se ha obtenido en la Tabla 1 del apéndice. La probabilidad de que la mitad de la muestra crea en el valor de la asignatura de ética empresarial es aproximadamente de 0,1.

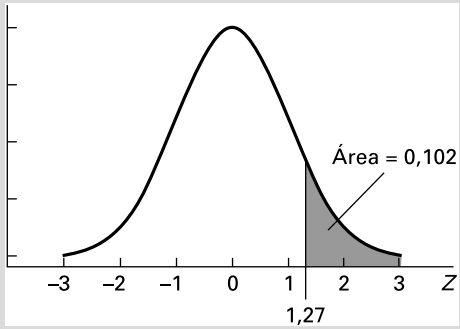


Figura 7.14. Probabilidad de que una variable aleatoria normal estándar sea de más de 1,27.

EJERCICIOS

Ejercicios básicos

7.26. Suponga que tenemos una población con una proporción $P = 0{,}40$ y una muestra aleatoria de tamaño $n = 100$ extraída de la población.

a) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,45?

b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,29?

c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,35 y 0,51?

7.27. Suponga que tenemos una población con una proporción $P = 0{,}25$ y una muestra aleatoria de tamaño $n = 200$ extraída de la población.

a) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,31?

b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,14?

c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,24 y 0,40?

7.28. Suponga que tenemos una población con una proporción $P = 0{,}60$ y una muestra aleatoria de tamaño $n = 100$ extraída de la población.

a) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,66?

b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,48?

c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,52 y 0,66?

7.29. Suponga que tenemos una población con una proporción $P = 0{,}50$ y una muestra aleatoria de tamaño $n = 900$ extraída de la población.

a) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,52?

b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,46?

c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,47 y 0,53?

Ejercicios aplicados

7.30. En 1992, los canadienses votaron en un referéndum sobre una nueva constitución. En la provincia de Quebec, el 42,4 por ciento de los que votaron estaba a favor de la nueva constitución. Se extrajo una muestra aleatoria de 100 votantes de la provincia.

a) ¿Cuál es la media de la distribución de la proporción muestral a favor de una nueva constitución?

b) ¿Cuál es la varianza de la proporción muestral?

c) ¿Cuál es el error típico de la proporción muestral?

d) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,5?

7.31. Según la Agencia Tributaria, el 75 por ciento de todas las declaraciones de la renta da lugar a una devolución. Se ha tomado una muestra aleatoria de 100 declaraciones de la renta.

a) ¿Cuál es la media de la distribución de la proporción muestral de declaraciones que dan lugar a una devolución?

b) ¿Cuál es la varianza de la proporción muestral?

c) ¿Cuál es el error típico de la proporción muestral?
d) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,8?

7.32. El propietario de una tienda de discos observa que el 20 por ciento de los clientes que entran en su tienda efectúa una compra. Una mañana entran en la tienda 180 personas que pueden considerarse una muestra aleatoria de todos los clientes.

a) ¿Cuál es la media de la distribución de la proporción muestral de clientes que realizan una compra?
b) ¿Cuál es la varianza de la proporción muestral?
c) ¿Cuál es el error típico de la proporción muestral?
d) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,15?

7.33. Un gerente de un gran grupo de hospitales cree que el 30 por ciento de todos los pacientes genera facturas que se cobran con 2 meses de retraso como mínimo. Se toma una muestra aleatoria de 200 pacientes.

a) ¿Cuál es el error típico de la proporción muestral que generará facturas que se cobrarán con 2 meses de retraso como mínimo?
b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,25?
c) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,33?
d) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,27 y 0,33?

7.34. Una empresa recibe 120 solicitudes de trabajo de personas recién licenciadas en administración de empresas. Suponiendo que estos demandantes de empleo pueden considerarse una muestra aleatoria de todos esos licenciados, ¿cuál es la probabilidad de que entre el 35 y el 45 por ciento de ellos sean mujeres si el 40 por ciento de todas las personas recién licenciadas en administración de empresas son mujeres?

7.35. Una institución benéfica ha observado que el 42 por ciento de todas las personas que donaron el año pasado volverán a donar este año. Se ha tomado una muestra aleatoria de 300 donantes del año pasado.

a) ¿Cuál es el error típico de la proporción muestral que donará de nuevo este año?
b) ¿Cuál es la probabilidad de que más de la mitad de estos miembros de la muestra done de nuevo este año?
c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,40 y 0,45?
d) Indique sin realizar los cálculos en cuál de los intervalos es más probable que se encuentre la proporción muestral: 0,39-0,41, 0,41-0,43, 0,43-0,45, 0,45-0,47.

7.36. Una empresa está considerando la posibilidad de sacar una nueva emisión de bonos convertibles. La dirección cree que los términos de la oferta serán atractivos para el 20 por ciento de todos sus accionistas actuales. Suponga que está en lo cierto. Se toma una muestra aleatoria de 130 accionistas actuales.

a) ¿Cuál es el error típico de la proporción muestral que piensa que esta oferta es atractiva?
b) ¿Cuál es la probabilidad de que la proporción muestral sea superior a 0,15?
c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,18 y 0,22?
d) Suponga que se hubiera tomado una muestra de 500 accionistas actuales. Indique sin realizar los cálculos si las probabilidades de los apartados (b) y (c) habrían sido mayores, menores o iguales que las obtenidas.

7.37. Una tienda ha observado que el 30 por ciento de todos los compradores de cortacéspedes también contrata un servicio de mantenimiento. En 1 mes se venden 280 cortacéspedes a clientes que pueden considerarse una muestra aleatoria de todos los compradores.

a) ¿Cuál es el error típico de la proporción muestral de clientes que contratarán un servicio de mantenimiento?
b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,32?
c) Indique sin realizar los cálculos en cuál de los siguientes intervalos es más probable que se encuentre la proporción muestral: 0,29-0,31, 0,30-0,32, 0,31-0,33, 0,32-0,34.

7.38. Se toma una muestra aleatoria de 100 votantes para estimar la proporción del electorado que está a favor de una subida del impuesto sobre la gasolina a fin de obtener más ingresos para reparar las autopistas. ¿Cuál es el valor más alto que puede tomar el error típico de la proporción muestral que está a favor de esta medida?

7.39. Vuelva al ejercicio 7.38 y suponga que se decide que una muestra de 100 votantes es demasiado pequeña para obtener una estimación suficientemente fiable de la proporción poblacional. Se exige, por el contrario, que la probabilidad de que la proporción muestral difiera de la propor-

ción poblacional (cualquiera que sea su valor) en más de 0,03 no sea superior a 0,05. ¿De qué tamaño debe ser la muestra para que se cumpla este requisito?

7.40. Una empresa quiere estimar la proporción de personas que es probable que compren maquinillas de afeitar eléctricas y que ven los partidos de fútbol que se retransmiten los fines de semana. Se toma una muestra aleatoria de 120 personas que se consideraron probables compradoras de maquinillas de afeitar eléctricas. Suponga que la proporción de probables compradoras de maquinillas eléctricas en la población que ve los partidos retransmitidos es 0,25.

a) La probabilidad de que la proporción muestral que ve los partidos retransmitidos sea _____ mayor que la proporción poblacional es de 0,10.

b) La probabilidad de que la proporción muestral sea _____ menor que la proporción poblacional es 0,05.

c) La probabilidad de que la proporción muestral se diferencie en _____ de la proporción poblacional es 0,30.

7.41. Suponga que el 50 por ciento de todos los ciudadanos adultos de un país cree que es esencial revisar profundamente el sistema sanitario nacional. ¿Cuál es la probabilidad de que más del 56 por ciento de una muestra aleatoria de 150 adultos tenga esta opinión?

7.42. Suponga que el 50 por ciento de todos los ciudadanos adultos de un país cree que el déficit presupuestario público actual será perjudicial a largo plazo para la economía nacional. ¿Cuál es la probabilidad de que más del 58 por ciento de una muestra aleatoria de 250 adultos tenga esta opinión?

7.43. Un periodista quería conocer las opiniones de los directores generales de las 500 mayores empresas de Estados Unidos sobre la contratación informatizada de acciones. En el tiempo de que disponía sólo pudo contactar con una muestra aleatoria de 81 de estos directores generales. Si el 55 por ciento de todos los miembros de la población cree que la contratación informatizada debe prohibirse, ¿cuál es la probabilidad de que menos de la mitad de los miembros de la muestra tenga esta opinión?

7.44. Una pequeña universidad tiene 528 alumnos de primer curso, de los cuales 211 llevan su propio ordenador personal al campus. Se ha tomado una muestra aleatoria de 120 estudiantes de primer curso.

a) ¿Cuál es el error típico de la proporción muestral que lleva su propio ordenador personal al campus?

b) ¿Cuál es la probabilidad de que la proporción muestral sea de menos de 0,33?

c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,5 y 0,6?

7.45. Una fábrica tiene 438 obreros, de los cuales 239 están preocupados por las futuras prestaciones sanitarias. Se ha pedido a una muestra aleatoria de 80 de estos obreros que estime la proporción poblacional preocupada por las futuras prestaciones sanitarias.

a) ¿Cuál es el error típico de la proporción muestral preocupada?

b) ¿Cuál es la probabilidad de que la proporción muestral sea inferior a 0,5?

c) ¿Cuál es la probabilidad de que la proporción muestral esté comprendida entre 0,5 y 0,6?

7.46. Las subidas salariales porcentuales anuales de los directores generales de todas las medianas empresas siguen una distribución normal que tiene una media de 12,2 por ciento y una desviación típica de 3,6 por ciento. Se ha tomado una muestra aleatoria de 81 de estos directores generales. ¿Cuál es la probabilidad de que la mitad de los miembros de la muestra tenga subidas salariales de menos del 10 por ciento?

7.4. Distribuciones de las varianzas muestrales en el muestreo

Una vez analizadas las distribuciones de las medias muestrales y de las proporciones muestrales en el muestreo, examinaremos las distribuciones de las varianzas muestrales en el muestreo. A medida que las empresas y la industria ponen más énfasis en la producción de productos que satisfagan los criterios de calidad de los clientes, es mayor la necesidad de calcular y reducir la varianza poblacional. Cuando la varianza es alta en un proceso, algunas características importantes de los productos pueden tomar una gama más amplia

de valores, como consecuencia de la cual hay más productos que no tienen un nivel de calidad aceptable. Al fin y al cabo, a un cliente le da lo mismo que un producto funcione bien «en promedio». Lo que le interesa es que funcione el que ha comprado. Se pueden obtener productos de calidad en un proceso de producción si éste tiene una baja varianza poblacional, de manera que es menor el número de unidades que tienen un nivel de calidad inferior al deseado. Comprendiendo la distribución de las varianzas muestrales en el muestreo, podemos hacer inferencias sobre la varianza poblacional. Por lo tanto, es posible identificar y corregir los procesos que tienen una elevada varianza. Además, cuando la varianza poblacional es menor, podemos hacer mejores inferencias sobre las medias poblacionales utilizando medias muestrales.

Comenzamos examinando una muestra aleatoria de n observaciones procedentes de una población que tiene una media μ y una varianza σ^2 desconocidas. Representamos los miembros de la muestra por medio de $x_1, x_2, ..., x_n$. La varianza poblacional es la esperanza

$$\sigma^2 = E[(X - \mu)^2]$$

que sugiere que consideremos la media de $(x_i - \bar{x})^2$ de n observaciones. Dado que la μ es desconocida, utilizaremos la media muestral $\bar{x}$ para calcular la varianza muestral.

Varianza muestral

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria de observaciones procedentes de una población. La cantidad

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

se llama **varianza muestral** y su raíz cuadrada, s, se llama *desviación típica muestral*. Dada una muestra aleatoria específica, podríamos calcular la varianza muestral y ésta sería diferente para cada muestra aleatoria, debido a las diferencias entre las observaciones muestrales.

Tal vez nos sorprenda al principio el uso de $(n - 1)$ como divisor en la definición anterior. Una sencilla explicación es que en una muestra aleatoria de n observaciones tenemos n valores o grados de libertad independientes. Pero una vez que conocemos la media muestral calculada, sólo hay $n - 1$ valores diferentes que pueden definirse de forma independiente. Puede demostrarse, además, que el valor esperado de la varianza muestral calculado de esta forma es la varianza poblacional. Este resultado se demuestra en el apéndice del capítulo y se cumple cuando el tamaño de la muestra, n, es una pequeña proporción del tamaño de la población N:

$$E(s^2) = \sigma^2$$

La conclusión de que el valor esperado de la varianza muestral es la varianza poblacional es bastante general. Pero para hacer una inferencia estadística nos gustaría saber más sobre la distribución en el muestreo. Si podemos suponer que la distribución poblacional subyacente es normal, podemos demostrar que la varianza muestral y la varianza poblacional están relacionadas a través de una distribución de probabilidad que se conoce con el nombre de *distribución ji-cuadrado*.

Distribución ji-cuadrado de varianzas muestrales y poblacionales

Dada una muestra aleatoria de n observaciones procedentes de una población que sigue una distribución normal cuya varianza poblacional es σ^2 y cuya varianza muestral resultante es s^2, puede demostrarse que

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$

tiene una distribución conocida con el nombre de **distribución χ^2 (ji-cuadrado)** con $n-1$ grados de libertad.

La familia de distribuciones ji-cuadrado se utiliza en el análisis estadístico aplicado porque establece una relación entre las varianzas muestrales y las varianzas poblacionales. La distribución ji-cuadrado con $n-1$ grados de libertad es la distribución de la suma de los cuadrados de $n-1$ variables aleatorias normales estándar independientes. La distribución ji-cuadrado anterior y las probabilidades calculadas resultantes de varios valores de s^2 requieren que la distribución poblacional sea normal. Por lo tanto, el supuesto de la existencia de una distribución normal subyacente es más importante para hallar las probabilidades de las varianzas muestrales que para hallar las probabilidades de las medias muestrales.

La distribución se define únicamente para valores positivos, ya que las varianzas son todas ellas valores positivos. La Figura 7.15 muestra un ejemplo de la función de densidad. La función de densidad es asimétrica y tiene una larga cola positiva. Podemos caracterizar un miembro de la familia de distribuciones ji-cuadrado mediante un único parámetro denominado grados de libertad y representado por medio del símbolo v. Una distribución χ_2 con v grados de libertad se representa de la siguiente manera: χ^2_v. La media y la varianza de esta distribución son iguales al número de grados de libertad y el doble del número de grados de libertad.

$$E(\chi^2_v) = v \quad \text{y} \quad \text{Var}(\chi^2_v) = 2v$$

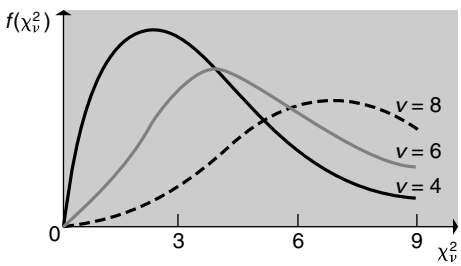


Figura 7.15. Funciones de densidad de la distribución ji-cuadrado con 4, 6 y 8 grados de libertad.

Utilizando estos resultados de la media y la varianza de la distribución ji-cuadrado, tenemos que

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = (n-1)$$

$$\frac{(n-1)}{\sigma^2}E(s^2) = (n-1)$$

$$E(s^2) = \sigma^2$$

Para hallar la varianza de s^2, tenemos que

$$\text{Var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1)$$

$$\frac{(n-1)^2}{\sigma^4}\text{Var}\,(s^2) = 2(n-1)$$

$$\text{Var}\,(s^2) = \frac{2\sigma^4}{(n-1)}$$

Podemos utilizar las propiedades de la distribución χ^2 para hallar la varianza de la distribución de la varianza muestral en el muestreo cuando la población de la que procede la muestra es normal.

El parámetro v de la distribución χ^2 se llama *grados de libertad*. Para ayudar a comprender el concepto de grados de libertad, consideremos primero que la varianza muestral es la suma de los cuadrados de n valores de la forma $(x_i - \bar{x})$. Estos n valores no son independientes, porque su suma es cero (como podemos demostrar utilizando la definición de media). Por lo tanto, si conocemos cualesquiera $n - 1$ de los valores $(x_i - \bar{x})$,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

$$x_n - \bar{x} = \sum_{i=1}^{n-1} (x_i - \bar{x})$$

Dado que podemos hallar la n-ésima cantidad si conocemos las $n - 1$ cantidades restantes, decimos que hay $n - 1$ grados de libertad —valores independientes— para calcular s^2. En cambio, si conociéramos μ, podríamos calcular una estimación de σ^2 utilizando las cantidades

$$(x_1 - \mu),\ (x_2 - \mu),\ \ldots,\ (x_n - \mu)$$

cada una de las cuales es independiente. En ese caso, tendríamos n grados de libertad de las n observaciones muestrales independientes, x_i. Sin embargo, μ no se conoce, por lo que debemos utilizar su estimación $\bar{x}$ para calcular la estimación de σ^2. Como consecuencia, se pierde un grado de libertad al calcular la media muestral y tenemos $n - 1$ grados de libertad para calcular s^2.

En muchas aplicaciones en las que interviene la varianza poblacional, hay que hallar los valores de la distribución acumulada de χ^2, sobre todo la cola superior y la inferior de la distribución; por ejemplo,

$$P(\chi^2_{10} < K) = 0{,}05$$

$$P(\chi^2_{10} > K) = 0{,}05$$

Para ello tenemos la distribución de la variable aleatoria que sigue una distribución ji-cuadrado calculada en la Tabla 7 del apéndice. En esa tabla, los grados de libertad se indican en la columna de la izquierda y los valores críticos de K correspondientes a los diferentes niveles de probabilidad se indican en las demás columnas. Así, por ejemplo, con 10 grados de libertad el valor de K correspondiente al intervalo inferior es 3,94. Este resultado se encuentra mirando la fila de 10 grados de libertad en la columna de la izquierda y la columna correspondiente a la probabilidad 0,950. El valor de la ji-cuadrado es 3,94. Asimismo, en el caso del intervalo superior de 0,05, el valor de K es 18,31. Este resultado se encuentra mirando la fila de 10 grados de libertad en la columna de la izquierda y la columna

correspondiente a la probabilidad 0,050. El valor de la ji-cuadrado es 18,31. Estas probabilidades se muestran esquemáticamente en la Figura 7.16.

$$P(\chi^2_{10} < 3{,}94) = 0{,}05$$
$$P(\chi^2_{10} > 18{,}31) = 0{,}05$$

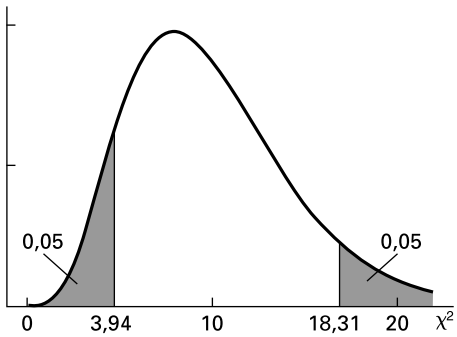


Figura 7.16. Probabilidades superior e inferior de una χ^2_{10} con 10 grados de libertad.

El recuadro siguiente resume los resultados de la distribución en el muestreo.

Distribución de las varianzas muestrales en el muestreo

Sea s^2 la varianza muestral de una muestra aleatoria de n observaciones procedentes de una población que tiene una varianza σ^2. En ese caso,

1. La distribución de s^2 en el muestreo tiene una media σ^2:

$$E(s^2) = \sigma^2 \tag{7.11}$$

2. La varianza de la distribución de s^2 en el muestreo depende de la distribución de la población subyacente. Si esa distribución es normal, entonces

$$\text{Var}\,(s^2) = \frac{2\sigma^4}{n-1} \tag{7.12}$$

3. Si la distribución de la población es normal, entonces $\dfrac{(n-1)s^2}{\sigma^2}$ se distribuye como una $\chi^2_{(n-1)}$.

Por lo tanto, si tenemos una muestra aleatoria procedente de una población que sigue una distribución normal, podemos hacer inferencias sobre la varianza muestral σ^2 utilizando s^2 y la distribución ji-cuadrado. Este proceso se muestra en los siguientes ejemplos.

EJEMPLO 7.9. Proceso de control de la calidad de Electrónica Integrada (probabilidad de la varianza muestral)

Jorge Sánchez es responsable de la garantía de calidad de Electrónica Integrada. Le ha pedido que cree un proceso de control de la calidad para la fabricación de un mecanismo de control A. La variabilidad de la resistencia eléctrica, expresada en ohmios, es fundamental para este mecanismo. Las normas de fabricación especifican una desviación típica de 3,6 y la distribución poblacional de las medidas de la resistencia es nor-

mal. El proceso de control requiere que se obtenga una muestra aleatoria de $n = 6$ observaciones de la población de mecanismos y que se calcule la varianza muestral. Halle un límite superior de la varianza muestral tal que la probabilidad de que se supere este límite, dada una desviación típica poblacional de 3,6, sea inferior a 0,05.

Solución

En este problema tenemos que $n = 6$ y $\sigma^2 = (3{,}6)^2 = 12{,}96$. Utilizando la distribución ji-cuadrado, podemos establecer que

$$P(s^2 > K) = P\left(\frac{(n-1)s^2}{12{,}96} > \chi^2_5\right) = 0{,}05$$

donde K es el límite superior deseado y $\chi^2_5 = 11{,}07$ es el valor crítico superior correspondiente al nivel 0,05 de la distribución ji-cuadrado con 5 grados de libertad de la fila 5 de la Tabla 7. El límite superior de s^2 que buscamos —representado por K— puede hallarse resolviendo

$$\frac{(n-1)K}{12{,}96} = 11{,}07$$

$$K = \frac{(11{,}07)(12{,}96)}{(6-1)} = 28{,}69$$

Si la varianza muestral, s^2, procedente de una muestra aleatoria de tamaño $n = 6$ es superior a 28,69, existen pruebas contundentes para sospechar que la varianza poblacional es superior a 12,96 y que hay que detener el proceso de producción y realizar los debidos ajustes.

EJEMPLO 7.10. Análisis del proceso de producción de Alimentos Valleverde (probabilidad de la varianza muestral)

Susana Méndez es la directora de garantía de calidad de Alimentos Valleverde, una empaquetadora de verduras congeladas. Susana quiere estar segura de que la variación del peso de las bolsas de verduras es pequeña, de manera que la empresa no produzca una elevada proporción de bolsas que tengan un peso inferior al indicado. Le ha pedido que halle el límite superior e inferior del cociente entre la varianza muestral y la varianza poblacional de una muestra aleatoria de $n = 20$ observaciones. Los límites son tales que la probabilidad de que el conciente sea inferior al límite inferior es 0,025 y la probabilidad de que sea superior al límite superior es 0,025. Por lo tanto, el 95 por ciento de los cocientes estará entre estos límites. Puede suponerse que la distribución poblacional es normal.

Solución

Se nos pide que hallemos los valores K_L y K_U tales que

$$P\left(\frac{s^2}{\sigma^2} < K_L\right) = 0{,}025 \quad \text{y} \quad P\left(\frac{s^2}{\sigma^2} > K_U\right) = 0{,}025$$

dado que se utiliza una muestra aleatoria de tamaño $n = 20$ para calcular la varianza muestral. En el caso del límite inferior, podemos establecer que

$$0{,}025 = P\left[\frac{(n-1)s^2}{\sigma^2} < (n-1)K_L\right] = P[\chi^2_{19} < (n-1)K_L]$$

En el caso del límite superior, podemos establecer que

$$0{,}975 = P\left[\frac{(n-1)s^2}{\sigma^2} > (n-1)K_U\right] = P[\chi^2_{19} > (n-1)K_U]$$

Estos límites superior e inferior de la ji-cuadrado definen un intervalo tal que si la ji-cuadrado calculada con la muestra está dentro de ese intervalo, aceptamos el supuesto de que la varianza del proceso se encuentra en el valor supuesto. Este intervalo se denomina *intervalo de aceptación*.

Utilizando los límites inferior y superior del intervalo de aceptación basados en la ji-cuadrado, podemos calcular los límites del intervalo de aceptación, K_L y K_U, del cociente entre la varianza muestral y la varianza poblacional. Los valores superior e inferior de la distribución ji-cuadrado pueden hallarse en la Tabla 7:

$$\chi^2_{19L} = 8{,}91$$
$$\chi^2_{19U} = 32{,}85$$
$$0{,}025 = P[\chi^2_{19L} < (n-1)K_L] = P[8{,}91 < (19)K_L]$$

Por lo tanto,

$$K_L = 0{,}469$$

En el caso del límite superior, tenemos que

$$0{,}975 = P[\chi^2_{19U} > (n-1)K_U] = P[32{,}85 > (19)K_U]$$

y, por lo tanto,

$$K_U = 1{,}729$$

El intervalo de aceptación del 95 por ciento del cociente entre la varianza muestral y la varianza poblacional es

$$P\left(0{,}469 \leqslant \frac{s^2}{\sigma^2} \leqslant 1{,}729\right) = 0{,}95$$

Por lo tanto, la varianza muestral se encuentra entre 46,9 por ciento y 172,9 por ciento de la varianza poblacional con una probabilidad de 0,95.

Es importante subrayar aquí que en los métodos empleados para hacer inferencias sobre la varianza poblacional influye mucho el supuesto de que la población sigue una distribución normal. En las inferencias sobre la media poblacional basadas en la media muestral no influyen mucho las desviaciones con respecto a la distribución normal. Además, las inferencias basadas en la media muestral pueden utilizar el teorema del límite central, que

establece que las medias muestrales generalmente siguen una distribución normal si el tamaño de la muestra es razonablemente grande. Las inferencias basadas en la media muestral son, pues, robustas con respecto al supuesto de la normalidad. Desgraciadamente, las inferencias basadas en varianzas muestrales no lo son.

Sabemos que en muchas aplicaciones la varianza poblacional tiene un interés directo para el investigador. Pero cuando utilizamos los métodos que hemos mostrado, debemos tener presente que si sólo se dispone de un número moderado de observaciones muestrales, la existencia de serias desviaciones con respecto a la normalidad en la población de la que procede la muestra puede invalidar gravemente las conclusiones de los análisis. En estas circunstancias, el analista cauto deberá tener bastante cuidado al hacer inferencias.

EJERCICIOS

Ejercicios básicos

7.47. Se obtiene una muestra aleatoria de tamaño $n = 16$ de una población que sigue una distribución normal de media $\mu = 100$ y varianza $\sigma^2 = 25$.

a) ¿Cuál es la probabilidad de que $\bar{x} > 101$?

b) ¿Cuál es la probabilidad de que la varianza muestral sea superior a 45?

c) ¿Cuál es la probabilidad de que la varianza muestral sea superior a 60?

7.48. Se obtiene una muestra aleatoria de tamaño $n = 25$ de una población que sigue una distribución normal de media $\mu = 198$ y varianza $\sigma^2 = 100$.

a) ¿Cuál es la probabilidad de que la media muestral sea superior a 200?

b) ¿Cuál es el valor de la media muestral tal que el 5 por ciento de las varianzas muestrales sería inferior a este valor?

c) ¿Cuál es el valor de la media muestral tal que el 5 por ciento de las varianzas muestrales sería superior a este valor?

7.49. Se obtiene una muestra aleatoria de tamaño $n = 18$ de una población que sigue una distribución normal de media $\mu = 46$ y varianza $\sigma^2 = 50$.

a) ¿Cuál es la probabilidad de que la media muestral sea superior a 50?

b) ¿Cuál es el valor de la varianza muestral tal que el 5 por ciento de las varianzas muestrales sería inferior a este valor?

c) ¿Cuál es el valor de la varianza muestral tal que el 5 por ciento de las varianzas muestrales sería superior a este valor?

7.50. Un proceso produce lotes de un producto químico cuyas concentraciones de impurezas siguen una distribución normal de varianza 1,75. Se elige una muestra aleatoria de 20 lotes. Halle la probabilidad de que la varianza muestral sea superior a 3,10.

7.51. Las tasas mensuales de rendimiento de las acciones de una empresa son independientes de las de otra y siguen una distribución normal que tiene una desviación típica de 1,7. Se toma una muestra de 12 meses.

a) Halle la probabilidad de que la desviación típica muestral sea inferior a 2,5.

b) Halle la probabilidad de que la desviación típica muestral sea superior a 1,0.

7.52. Se cree que los sueldos que perciben durante el primer año los contables recién titulados siguen una distribución normal que tiene una desviación típica de 2.500 $. Se toma una muestra aleatoria de 16 observaciones.

a) Halle la probabilidad de que la desviación típica muestral sea superior a 3.000 $.

b) Halle la probabilidad de que la desviación típica muestral sea inferior a 1.500 $.

Ejercicios aplicados

7.53. Se va a realizar a todos los estudiantes de primer año un examen de matemáticas con 100 preguntas de tipo test. Se ha hecho primero un estudio piloto en el que se ha realizado el examen a una muestra aleatoria de 20 estudiantes de primer año. Suponga que la distribución del número de respuestas correctas de la población de todos los estudiantes de primer año es normal con una varianza de 250.

a) ¿Cuál es la probabilidad de que la varianza muestral sea inferior a 100?

b) ¿Cuál es la probabilidad de que la varianza muestral sea superior a 500?

7.54. En una gran ciudad se ha observado que durante el verano las facturas del consumo de electricidad siguen una distribución normal que tiene una desviación típica de 100 $. Se ha tomado una muestra aleatoria de 25 facturas.

a) Halle la probabilidad de que la desviación típica muestral sea inferior a 75 $.

b) Halle la probabilidad de que la desviación típica muestral sea superior a 150 $.

7.55. El número de horas que dedican a ver la televisión los estudiantes la semana anterior a los exámenes finales sigue una distribución normal que tiene una desviación típica de 4,5 horas. Se ha tomado una muestra aleatoria de 30 estudiantes.

a) ¿Es superior a 0,95 la probabilidad de que la desviación típica muestral sea de más de 3,5 horas?

b) ¿Es superior a 0,95 la probabilidad de que la desviación típica muestral sea de menos de 6 horas?

7.56. En la Tabla 7.1 hemos examinado las 15 muestras posibles de dos observaciones procedentes de una población de $N = 6$ valores de años de experiencia de los trabajadores. La varianza poblacional de estos seis valores es

$$\sigma^2 = \frac{47}{12}$$

Calcule para cada una de las 15 muestras posibles la varianza muestral. Halle la media de estas 15 varianzas muestrales, confirmando así que el valor esperado de la varianza muestral no es igual a la varianza poblacional cuando el número de miembros de la muestra no es una pequeña proporción del número de miembros de la población [de hecho, como puede verificar aquí, $E(s^2) = N\sigma^2/(N - 1)$].

7.57. Un proceso de producción fabrica componentes electrónicos que emiten señales cuya duración sigue una distribución normal. Se ha tomado una muestra aleatoria de seis componentes y se ha medido la duración de las señales que emiten.

a) La probabilidad de que la varianza muestral sea superior a _____ por ciento de la varianza poblacional es 0,05.

b) La probabilidad de que la varianza muestral sea inferior a _____ por ciento de la varianza poblacional es 0,10.

7.58. Se ha tomado una muestra aleatoria de 10 fondos de inversión. Suponga que las tasas de rendimiento de la población de todos los fondos de inversión siguen una distribución normal.

a) La probabilidad de que la varianza muestral sea superior a _____ por ciento de la varianza poblacional es 0,10.

b) Halle cualquier par de números, a y b, que completen la frase siguiente: la probabilidad de que la varianza muestral esté comprendida entre a por ciento y b por ciento de la varianza poblacional es de 0,95.

c) Suponga que se hubiera tomado una muestra de 20 fondos de inversión. Indique sin hacer los cálculos cómo cambiaría eso su respuesta al apartado (b).

7.59. Se pide a cada uno de los miembros de una muestra aleatoria de 15 economistas que prediga la tasa de inflación del próximo año. Suponga que las predicciones de toda la población de economistas sigue una distribución normal que tiene una desviación típica de 1,8 por ciento.

a) La probabilidad de que la desviación típica muestral sea superior a _____ es 0,01.

b) La probabilidad de que la desviación típica muestral sea inferior a _____ es 0,025.

c) Halle cualquier par de números tal que la probabilidad de que la desviación típica muestral se encuentre entre esos números sea de 0,90.

7.60. Se comprueba un instrumento de precisión realizando 12 lecturas de la misma cantidad. La distribución poblacional de las lecturas es normal.

a) La probabilidad de que la varianza muestral sea superior a _____ por ciento de la varianza poblacional es 0,95.

b) La probabilidad de que la varianza muestral sea superior a _____ por ciento de la varianza poblacional es 0,90.

c) Halle cualquier par de números, a y b, que completen la frase siguiente: la probabilidad de que la varianza muestral esté comprendida entre a por ciento y b por ciento de la varianza poblacional es de 0,95.

7.61. Una compañía farmacéutica produce píldoras que contienen un principio activo. A la compañía le preocupa el peso medio de este principio por píldora, pero también quiere que la varianza (en miligramos cuadrados) no sea superior a 1,5. Se selecciona una muestra aleatoria de 20 píldoras y se observa que la varianza muestral es de 2,05. ¿Qué probabilidad hay de que la varianza muestral sea tan alta o más que ésta si la varianza poblacional es de hecho de 1,5? Suponga que la distribución de la población es normal.

7.62. Un fabricante ha comprado materias primas a un proveedor cuyos envíos tienen unos niveles de impureza con una varianza de 15,4 (en kilos cuadrados). Un proveedor rival sostiene que puede suministrar esta materia prima con el mismo nivel medio de impurezas, pero con una varianza menor. En una muestra aleatoria de 25 envíos del segundo proveedor se ha observado que la varianza de los niveles de impureza era de 12,2. ¿Cuál es la probabilidad de que el valor de la varianza muestral sea tan bajo o más si la verdadera varianza poblacional es de hecho de 15,4? Suponga que la distribución de la población es normal.

RESUMEN

En el Capítulo 7 hemos presentado el concepto de distribuciones en el muestreo, que son las distribuciones de probabilidad de estadísticos muestrales. Las distribuciones en el muestreo nos permiten hallar la probabilidad de un estadístico muestral, dado un modelo específico de distribución de probabilidad para la distribución en el muestreo. Hemos relacionado, pues, los estadísticos muestrales analizados en el Capítulo 3 con las distribuciones de probabilidad examinadas en el 5 y el 6. En futuros capítulos veremos que esta relación nos permite utilizar nuestros estadísticos muestrales para extraer algunas conclusiones o hacer algunas inferencias sobre el sistema y el proceso que desarrollan una población de datos de la que procede nuestra muestra. Ésta es la base de las decisiones objetivas basadas en datos muestrales. En nuestro análisis, hemos incluido el importante concepto de intervalo de aceptación. Los intervalos de aceptación definen un intervalo, con una probabilidad dada, para los estadísticos muestrales basados en una función de distribución de probabilidad supuesta. Si el estadístico muestral está dentro de ese intervalo, «aceptamos» el modelo supuesto de probabilidad y lo consideramos correcto.

TÉRMINOS CLAVE

distribución ji-cuadrado, 279
distribución en el muestreo, 251
distribución de las medias muestrales en el muestreo, 254
distribución normal estándar de medias muestrales, 257
distribución de las proporciones muestrales en el muestreo, 273
distribución de las varianzas muestrales en el muestreo, 281
factor de corrección en el caso de una población finita, 256
intervalo de aceptación, 265
media muestral, 255
muestra aleatoria simple, 250
proporción muestral, 272
teorema del límite central, 260
variable aleatoria normal estandarizada, 257
varianza muestral, 278

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

7.63. ¿Qué quiere decir la afirmación de que la media muestral tiene una distribución en el muestreo?

7.64. Un inversor está considerando seis fondos de inversión distintos. El número medio de días al vencimiento de cada uno de estos fondos es

41 39 35 35 33 38

Se eligen aleatoriamente dos de estos fondos.

a) ¿Cuántas muestras posibles de dos fondos hay?
b) Enumere todas las muestras posibles.
c) Halle la función de probabilidad de la distribución de las medias muestrales en el muestreo.
d) Verifique directamente que la media de la distribución de las medias muestrales en el muestreo es igual a la media poblacional.

7.65. ¿Qué importancia tiene el teorema del límite central para la distribución de las medias muestrales en el muestreo?

7.66. Las calificaciones de todos los estudiantes que realizan un examen de aptitud que se exige para entrar en una facultad de derecho siguen una distribución normal que tiene una media de 420 y una desviación típica de 100. Se toma una muestra aleatoria de 25 calificaciones.

a) Halle la probabilidad de que la media muestral de las calificaciones sea superior a 450.

b) Halle la probabilidad de que la media muestral de las calificaciones esté comprendida entre 400 y 450.
c) La probabilidad de que la media muestral de las calificaciones sea superior a _____ es 0,10.
d) La probabilidad de que la media muestral de las calificaciones sea inferior a _____ es 0,10.
e) La probabilidad de que la desviación típica muestral de las calificaciones sea superior a _____ es 0,05.
f) La probabilidad de que la desviación típica muestral de las calificaciones sea inferior a _____ es 0,05.
g) Si se hubiera tomado una muestra de 50 calificaciones, ¿sería la probabilidad de que la media muestral de las calificaciones sea superior a 450 menor, mayor o igual que la respuesta correcta al apartado (a)? No es necesario hacer aquí los cálculos detallados. Ilustre gráficamente su razonamiento.

7.67. Una empresa repara aparatos de aire acondicionado. Se ha observado que la duración de las reparaciones sigue una distribución normal que tiene una media de 60 minutos y una desviación típica de 10 minutos. Se ha tomado una muestra aleatoria de la duración de las reparaciones.

a) ¿Cuál es la probabilidad de que la media muestral de la duración de las reparaciones sea de más de 65 minutos?
b) La probabilidad de que la media muestral de la duración de las reparaciones sea de menos de _____ minutos es 0,10.
c) La probabilidad de que la desviación típica muestral de la duración de las reparaciones sea de más de _____ minutos es 0,10.
d) La probabilidad de que la desviación típica muestral de la duración de las reparaciones sea de menos de _____ minutos es 0,10.
e) ¿Cuál es la probabilidad de que más de dos de estas reparaciones duren más de 65 minutos?

7.68. Un año las tasas porcentuales de rendimiento de los fondos de inversión siguieron una distribución normal de media 14,8 y desviación típica 6,3. Se tomó una muestra aleatoria de nueve de estos fondos.

a) ¿Cuál es la probabilidad de que la media muestral de las tasas porcentuales de rendimiento sea de más de 19,0?
b) ¿Cuál es la probabilidad de que la media muestral de las tasas porcentuales de rendimiento esté comprendida entre 10,6 y 19,0?
c) La probabilidad de que la media muestral de las tasas porcentuales de rendimiento sea de menos de _____ es 0,25.
d) La probabilidad de que la desviación típica muestral de las tasas porcentuales de rendimiento sea de más de _____ es 0,10.
e) Si se tomara una muestra de 20 de estos fondos, indique si la probabilidad de que la media muestral de las tasas porcentuales de rendimiento fuera de más de 19,0 sería mayor, menor o igual que la respuesta correcta del apartado (a). Represente gráficamente su razonamiento.

7.69. Se sabe que la duración de un componente electrónico sigue una distribución normal que tiene una media de 1.600 horas y una desviación típica de 400 horas.

a) Halle la probabilidad de que la media muestral de una muestra aleatoria de 16 componentes sea de más de 1.500 horas.
b) La probabilidad de que la media muestral de la duración de una muestra aleatoria de 16 componentes sea de más de _____ horas es 0,15.
c) La probabilidad de que la desviación típica muestral de la duración de una muestra aleatoria de 16 componentes sea de más de _____ horas es 0,10.

7.70. Utilice el apéndice del capítulo para hallar la media de la distribución de las varianzas muestrales en el muestreo de una muestra de n observaciones procedentes de una población de N miembros cuando la varianza poblacional es σ^2. Modificando convenientemente el argumento sobre las varianzas del apéndice del capítulo, demuestre que

$$E(s^2) = N\sigma^2/(N - 1)$$

Obsérvese la verosimilitud intuitiva de este resultado cuando $n = N$.

7.71. Se ha observado que el tiempo que tarda la gente en rellenar un impreso de declaración de impuestos sigue una distribución normal que tiene una media de 100 minutos y una desviación típica de 30 minutos. Se ha tomado una muestra aleatoria de nueve personas que han rellenado este impreso.

a) ¿Cuál es la probabilidad de que la media muestral del tiempo que se tarda sea de más de 120 minutos?
b) La probabilidad de que la media muestral del tiempo que se tarda sea de menos de _____ minutos es 0,20.

c) La probabilidad de que la desviación típica muestral del tiempo que se tarda sea de menos de _____ minutos es 0,05.

7.72. Se ha observado que el 80 por ciento de los estudiantes de último año de una universidad acepta una oferta de trabajo antes de licenciarse. La distribución de los salarios de los que aceptan ofertas era normal y tiene una media de 29.000 $ y una desviación típica de 4.000 $.

a) ¿Cuál es la probabilidad de que menos del 70 por ciento de una muestra aleatoria de 60 estudiantes de último año acepte una oferta?
b) ¿Cuál es la probabilidad de que menos del 70 por ciento de una muestra aleatoria de 6 estudiantes de último año acepte una oferta?
c) ¿Cuál es la probabilidad de que el salario medio de una muestra aleatoria de 6 estudiantes de último año que aceptan una oferta fuera de más de 30.000 $?
d) Se elige aleatoriamente un estudiante de último año. ¿Cuál es la probabilidad de que haya aceptado una oferta de trabajo con un salario de más de 30.000 $?

7.73. Las bolsas de plástico utilizadas para empaquetar productos se fabrican de tal manera que su resistencia a los golpes sigue una distribución normal que tiene una desviación típica de 1,8 kilos por centímetro cuadrado. Se selecciona una muestra aleatoria de 16 bolsas.

a) La probabilidad de que la desviación típica muestral de la resistencia a los golpes sea de más de _____ es 0,01.
b) La probabilidad de que la media muestral sea _____ mayor que la media poblacional es 0,15.
c) La probabilidad de que la media muestral difiera en _____ de la media poblacional es 0,05.

7.74. Un director de control de calidad tenía interés en conocer la variabilidad de la cantidad de principio activo que contenían las píldoras producidas por un determinado proceso. Se tomó una muestra aleatoria de 21 píldoras. ¿Cuál es la probabilidad de que la varianza muestral de la cantidad de principio activo fuera más del doble de la varianza poblacional?

7.75. Se toma una muestra de 100 estudiantes para averiguar qué marca de cerveza se prefiere en una cata ciega de dos marcas. Suponga que el 50 por ciento de toda la población de estudiantes prefiere la marca A.

a) ¿Cuál es la probabilidad de que más del 60 por ciento de los miembros de la muestra prefiera la marca A?
b) ¿Cuál es la probabilidad de que entre el 45 y el 55 por ciento de los miembros de la muestra prefiera la marca A?
c) Suponga que sólo se dispone de una muestra de 10 estudiantes. Indique en qué diferiría el método de cálculo de las probabilidades en comparación con las soluciones de los apartados (a) y (b).

7.76. Las calificaciones de un examen realizado por un gran grupo de estudiantes sigue una distribución normal que tiene una desviación típica de 40 puntos. Se toma una muestra aleatoria de 16 calificaciones para estimar la calificación media de la población. Sea $\bar{X}$ la media muestral. ¿Cuál es la probabilidad de que el intervalo $(\bar{X} - 10) - (\bar{X} + 10)$ contenga la verdadera media de la población?

7.77. Un fabricante de detergente líquido sostiene que el peso medio del líquido que contienen los envases vendidos es al menos de 300 gramos. Se sabe que la distribución poblacional del peso es normal y tiene una desviación típica de 13 gramos. Para comprobar la afirmación del fabricante, se examina una muestra aleatoria de 16 envases. La afirmación se pondrá en duda si la media muestral del peso es de menos de 295 gramos. ¿Cuál es la probabilidad de que se ponga en duda la afirmación si la media poblacional del peso es en realidad de 300 gramos?

7.78. Un año el 40 por ciento de las ventas de viviendas fue financiado parcialmente por el vendedor. Se examina una muestra aleatoria de 250 ventas.

a) La probabilidad de que la proporción muestral sea de más de _____ es 0,8.
b) La probabilidad de que la proporción muestral sea de menos de _____ es 0,9.
c) La probabilidad de que la proporción muestral difiera en _____ de la proporción poblacional es de 0,7.

7.79. Un candidato a la presidencia tiene intención de hacer campaña si inicialmente lo apoya más de un 30 por ciento de los votantes. Se toma una muestra aleatoria de 300 votantes y se decide hacer campaña si la proporción muestral que apoya al candidato es de más de 0,28.

a) ¿Cuál es la probabilidad de que se decida no hacer campaña si el nivel inicial de apoyo es, en realidad, del 20 por ciento?

b) ¿Cuál es la probabilidad de que se decida no hacer campaña si el nivel inicial de apoyo es, en realidad, del 40 por ciento?

7.80. Se sabe que las rentas de los suscriptores de una revista siguen una distribución normal que tiene una desviación típica de 6.600 $. Se toma una muestra aleatoria de 25 suscriptores.

a) ¿Cuál es la probabilidad de que la desviación típica muestral de sus rentas sea de más de 4.000 $?

b) ¿Cuál es la probabilidad de que la desviación típica muestral de sus rentas sea de menos de 8.000 $?

7.81. Un proceso de producción fabrica lotes de productos químicos. Se seleccionan muestras de 20 lotes para examinarlos. Si la desviación típica del porcentaje de impurezas de los lotes de las muestras es de más del 2,5 por ciento, el proceso de producción se revisa minuciosamente. Suponga que la distribución poblacional de los porcentajes de impurezas es normal. ¿Cuál es la probabilidad de que el proceso de producción se revise minuciosamente si la desviación típica poblacional de los porcentajes de impurezas es del 2 por ciento?

Apéndice

1. Realización de simulaciones muestrales de Monte Carlo por medio del programa Minitab

En el apartado 7.2 presentamos los resultados de las simulaciones muestrales de Monte Carlo para demostrar el teorema del límite central. En este apéndice mostramos cómo pueden realizarse simulaciones similares para una distribución de probabilidad. La simulación puede realizarse utilizando una macro de Minitab llamada Centlimit.mac, que se encuentra en el disco que acompaña al libro de texto. Para utilizar esta macro, cópiela en el directorio

```
MTBWIN\MACROS\
```

utilizando el Windows Explorer. Esta macro se almacenará entonces con otras macros del paquete Minitab. Cuando se almacena la macro en este directorio, puede ejecutarse directamente en Minitab. También se puede almacenar en otro directorio y escribir el nombre completo para ejecutar la macro. Para realizar la simulación muestral, siga los pasos siguientes:

1. Almacene en la columna 1 un conjunto de valores que tengan la frecuencia indicada por la distribución de probabilidad que tenga interés en simular. Normalmente, almacenamos 100 valores, pero podría almacenarse cualquier número. Por ejemplo, para almacenar una distribución binomial con $P = 0{,}40$, almacenaríamos 40 1 y 60 0 en la columna 1. También podríamos almacenar una distribución empírica de números de una población estudiada. Otro método para obtener los valores muestrales es utilizar el comando

```
CALC > RANDOM DATA > "SELECT PROBABILITY DISTRIBUTION"
```

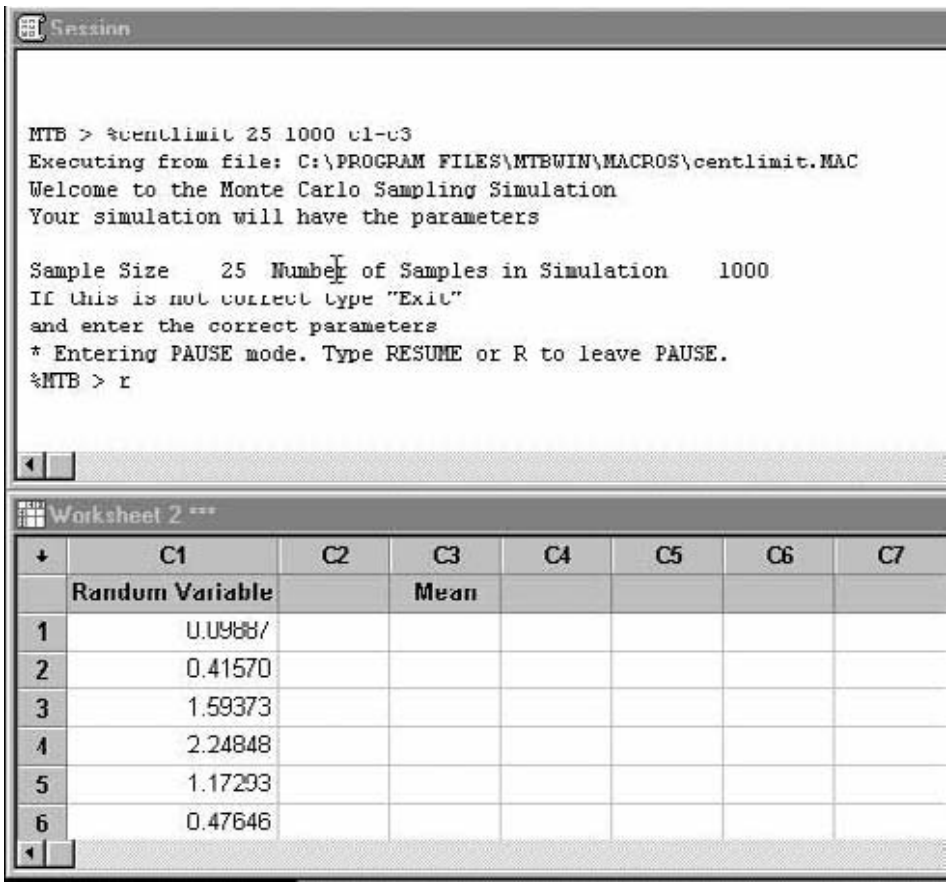
De esa forma, obtenemos una muestra aleatoria de una de las posibles distribuciones de probabilidad habituales.

2. En el Minitab Session Window, pulse el comando

```
MTB > %CENTLIMIT N1 N2 C1-C3
```

donde N1 es el tamaño de la muestra de las muestras que están simulándose y N2 es el número de muestras cuyas medias van a obtenerse en la simulación. Generalmente, entre 500 y 1.000 muestras dan lugar a una buena distribución muestral,

Figura 7.17. Simulación muestral de Monte Carlo en Minitab.



+	C1	C2	C3	C4	C5	C6	C7
	Random Variable		Mean				
1	0.0987						
2	0.41570						
3	1.59373						
4	2.24848						
5	1.17293						
6	0.47646						

pero se puede seleccionar cualquier valor razonable. Obsérvese que cuanto mayor sea el número de muestras, más se tardará en realizar la simulación. C1 a C3 son las columnas utilizadas por Minitab para realizar la simulación y la distribución de probabilidad de interés está en la columna 1. El lector puede utilizar las columnas que quiera con tal de que la distribución de probabilidad esté en la columna 1.

La Figura 7.17 muestra el resultado de una simulación muestral.

La simulación genera muestras en la columna 2 y calcula la media muestral. La media de cada muestra se almacena en la columna 3, titulada «Mean». Se calculan estadísticos descriptivos e histogramas para los valores de la variable aleatoria («random variable») de la columna 1 y para las medias muestrales de la columna 3. Pinchando en el comando del menú

```
WINDOWS>TILE
```

se puede obtener la pantalla de la Figura 7.18, que es útil para comparar la distribución inicial y la distribución muestral con una normal comparable.

En la Figura 7.18, vemos claramente que la distribución de la variable aleatoria no es normal sino que está muy sesgada hacia la derecha. En cambio, la distribución muestral de las medias se parece mucho a una distribución normal. La Figura 7.19 muestra una copia de la macro Centlimit.mac de Minitab, que también se encuentra en el disco de datos del libro de texto. Los usuarios familiarizados con las macros de Minitab pueden modificar esta macro para obtener salidas diferentes.

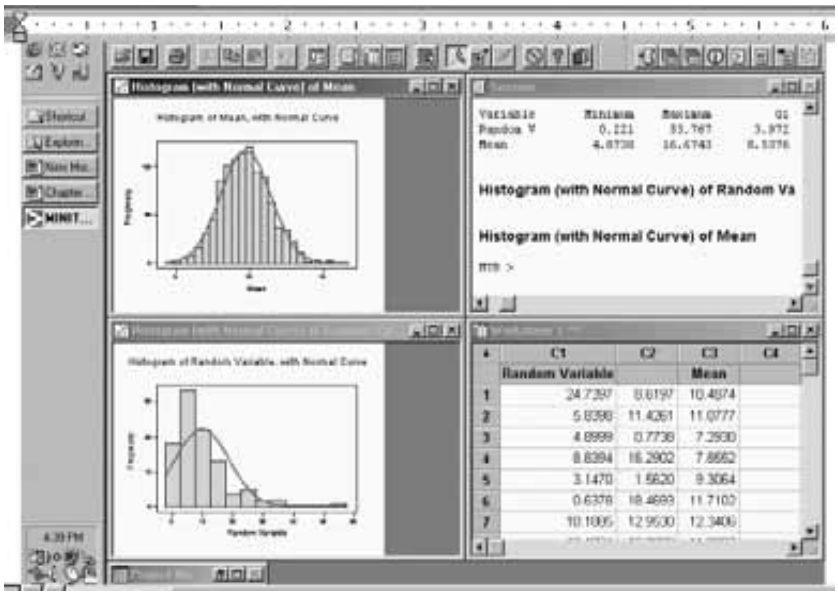


Figura 7.18. Resultados de la simulación muestral de Monte Carlo.

Figura 7.19. Copia de la macro Centlimit.mac de Minitab.

```
Macro
Centlimit n1,n2,Dist,Samp,Xbar
#  Dr.William L. Carlson
#  Professor of Economics
#  St Olaf College
#  Northfield MN   55057
#  Carlson@Stolaf.edu
#  To Execute this Macro in Minitab  Type
#  %Centlimit "sample size"  "Number of Samples"  C1 C2 C3
#
#The output includes a histogram and a normal probability plot for the
original #distribution and a histogram and normal probability plot for the
sampling #distribution of sample means
#Macro is Stored as a text file in C:\program
files\mtbwin\macros\centlimit.mac
#
#Definition of Variables
#
# n1  Sample size obtained from probability distribution
# n2  Number of samples of size n1 obtained in this simulation
# Dist       Column that contains an empirical distribution from which the
random #          sample is obtained.
# Xbar      Column that contains the sample means from each of the n2 samples
#           obtained in the simulation
# Samp      Column that will be used to generate each of the samples.
#
#
Mconstant  n1 n2 k1 k2
Mcolumn Dist Xbar Samp c11 c12 c13 c14
Name Dist 'Random Variable' Xbar 'Mean'
Let c11="Sample Size"
Let c12= n1
Let c13="Number of Samples in Simulation"
Let c14=n2
Note Welcome to the Monte Carlo Sampling Simulation
Note Your simulation will have the parameters
Write 'Terminal' c11-c14
Note If this is not correct type "exit"
Note and enter the correct parameters
Pause
Brief 0
Do k1=1:n2
Sample n1 Dist Samp;
Replace.
Mean Samp k2
Let xbar(k1)=k2
Enddo
Brief
Describe Dist Xbar;
GNHist.
Endmacro
```

2. Media de la distribución de las varianzas muestrales en el muestreo

En este apéndice, mostramos que la media de la distribución de las varianzas muestrales en el muestreo es la varianza poblacional. Comenzamos hallando la esperanza de la suma de los cuadrados de las diferencias entre cada miembro de la muestra y la media muestral; es decir, la esperanza de

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}[(X_i-\mu)^2-2(\bar{X}-\mu)(X_i-\mu)+(\bar{X}-\mu)^2]\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\sum_{i=1}^{n}(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)^2+n(\bar{X}-\mu)^2\\ &= \sum(X_i-\mu)^2-n(\bar{X}-\mu)^2\end{aligned}$$

Tomando esperanzas, tenemos que

$$\begin{aligned}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] &= E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-nE[(\bar{X}-\mu)^2]\\ &= \sum_{i=1}^{n}E[(X_i-\mu)^2]-nE[(\bar{X}-\mu)^2]\end{aligned}$$

Ahora bien, la esperanza de cada $(X_i-\mu)^2$ es la varianza poblacional, σ^2, y la esperanza de $(\bar{X}-\mu)^2$ es la varianza de la media muestral, σ^2/n. Por lo tanto, tenemos que

$$E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=n\sigma^2-\frac{n\sigma^2}{n}=(n-1)\sigma^2$$

Por último, el valor esperado de la varianza muestral es

$$\begin{aligned}E(s^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\\ &= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\\ &= \frac{1}{n-1}(n-1)\sigma^2=\sigma^2\end{aligned}$$

Éste es el resultado que queríamos demostrar.

Bibliografía

1. David, F. R., L. M. Anderson y K. W. Lawrimore, «Perspectives on Business Ethics in Management Education», *S. A. M. Advanced Management Journal*, 55, n.º 4, 1990, págs. 26-32.
2. Hogan, H., «The 1990 Post-enumeration Survey: An Overview», *American Statistician*, 46 (1992), págs. 261-269.

Capítulo 8

Estimación: una población

Esquema del capítulo

Introducción

En este capítulo hacemos hincapié en las afirmaciones inferenciales sobre la estimación de un parámetro poblacional, basadas en la información que contiene una muestra aleatoria. Centramos la atención en los métodos para estimar una media poblacional o una proporción de los miembros de la población que poseen una determinada característica. Por ejemplo, podemos querer una estimación de la demanda semanal media de una determinada marca de zumo de naranja o una estimación de la proporción de empleados de una empresa que son partidarios de que se modifique el plan de pluses.

En este capítulo presentamos dos métodos de estimación. En primer lugar, estimamos un parámetro poblacional desconocido por medio de un único número llamado estimación puntual. En el apartado 8.1 examinamos las propiedades de esta estimación puntual. En la mayoría de los problemas prácticos no basta con una estimación puntual. Para comprender mejor el proceso que generó la población también se necesita una medida de la variabilidad. En el resto del capítulo analizamos un segundo método, que tiene en cuenta esta variación estableciendo un intervalo de valores en el que es probable que se encuentre la cantidad que queremos estimar.

En el Capítulo 9 examinamos la estimación de la diferencia entre las medias o las proporciones de dos poblaciones y la estimación de la varianza.

8.1. Propiedades de los estimadores puntuales

Cualquier inferencia extraída de la población se basa en estadísticos muestrales. La elección de los estadísticos adecuados dependerá de cuál sea el parámetro poblacional que interese. El valor de ese parámetro será desconocido y uno de los objetivos del muestreo es estimar su valor. Debe hacerse una distinción entre los términos *estimador* y *estimación*.

Estimador y estimación

Un **estimador** de un parámetro poblacional es una variable aleatoria que depende de la información de la muestra; su valor proporciona aproximaciones a este parámetro desconocido. Un valor específico de esa variable aleatoria se llama **estimación**.

Hildebrand y Ott (véase la referencia bibliográfica 4) señalan que existe «una distinción técnica entre un *estimador* como una función de variables aleatorias y una *estimación* como un único número. Es la distinción entre un proceso (el estimador) y el resultado de ese proceso (la estimación)». Para aclarar esta distinción entre estimador y estimación, consideremos la estimación de las ventas semanales medias de una determinada marca de zumo de naranja. Un *estimador* posible de la media poblacional es la media muestral. Si se observa que la media de una muestra aleatoria de ventas semanales es de 3.280 litros, entonces 3.280 litros es una *estimación* de la media poblacional de las ventas semanales. Otro *estimador* posible de las ventas semanales medias podría ser la mediana muestral.

En el Capítulo 3 estudiamos otros estimadores, como la varianza muestral, s^2, y el coeficiente de correlación muestral, r. Si el valor de la varianza muestral de la demanda semanal de zumo de naranja es de 300 litros, entonces s^2 es el estimador y 300 es la estimación.

Cuando se analiza la estimación de un parámetro desconocido, deben considerarse dos posibilidades. En primer lugar, puede calcularse un *único número* a partir de la muestra y considerar que es el más representativo del parámetro poblacional desconocido. Éste se llama *estimación puntual*. Un ejemplo es la estimación de 3.280 litros de zumo de naranja. También podríamos hallar el intervalo o rango que es más probable que contenga el valor del parámetro poblacional. Por ejemplo, la demanda semanal media de esta marca de zumo de naranja en esta tienda se encuentra, con un grado especificado de confianza, entre 2.500 y 3.500 litros. Esta estimación por intervalos es un ejemplo de un tipo de *intervalo de confianza* que analizaremos en este capítulo.

Estimador puntual y estimación puntual

Consideremos un parámetro poblacional como la media poblacional μ o la proporción poblacional P. Un **estimador puntual** de un parámetro poblacional es una función de la información de la muestra que genera un único número llamado **estimación puntual**. Por ejemplo, la media muestral $\bar{X}$ es un estimador puntual de la media poblacional, μ, y el valor que toma $\bar{X}$ para un conjunto dado de datos se llama estimación puntual, $\bar{x}$.

Debe señalarse desde el principio que no existe ningún único mecanismo para saber cuál es el «mejor» estimador puntual en todas las circunstancias. Lo que existe es un conjunto de criterios con los que pueden evaluarse los estimadores. La media muestral también da una estimación puntual de la media poblacional, μ. Sin embargo, más adelante mostramos que la mediana no es el mejor estimador de la media de algunas distribuciones.

Evaluaremos los estimadores basándonos en tres importantes propiedades: ausencia de sesgo, consistencia y eficiencia.

Estimador insesgado

Para buscar un estimador de un parámetro poblacional, lo primero que debe ser es un estimador *insesgado*.

Estimador insesgado

Se dice que un estimador puntual es un **estimador insesgado** de un parámetro poblacional si su valor esperado es igual a ese parámetro; es decir, si

$$E(\hat{\theta}) = \theta$$

entonces $\hat{\theta}$ es un estimador insesgado de θ.

Obsérvese que el hecho de que un estimador sea insesgado no significa que un *determinado* valor de $\hat{\theta}$ tenga que ser exactamente el valor correcto de θ; lo que significa es que tiene «la capacidad de estimar el parámetro poblacional correctamente en promedio. Un estimador insesgado es correcto en promedio. Podemos concebir el valor esperado de $\hat{\theta}$ como la media de los valores de $\hat{\theta}$ para todas las muestras posibles o como la media a largo plazo de los valores de $\hat{\theta}$ para muestras repetidas. La condición de que el estimador $\hat{\theta}$ debe ser insesgado quiere decir que el valor *medio* de $\hat{\theta}$ es exactamente correcto. No quiere decir que un determinado valor de $\hat{\theta}$ es exactamente correcto» (véase la referencia bibliográfica 4).

A veces $\hat{\theta}$ sobreestima el parámetro y otras veces lo subestima, pero del concepto de esperanza se deduce que si se repite muchas veces el método de muestreo, entonces, en promedio, el valor de un estimador insesgado que se obtenga es igual al parámetro poblacional. Parece razonable afirmar que, manteniéndose todo lo demás constante, es deseable que un estimador puntual tenga la propiedad de ser insesgado. La Figura 8.1 ilustra las funciones de densidad de dos estimadores, $\hat{\theta}_1$ y $\hat{\theta}_2$, del parámetro θ. Debería ser evidente que $\hat{\theta}_1$ es un estimador insesgado de θ y $\hat{\theta}_2$ no lo es.

La media muestral, la varianza muestral y la proporción muestral son estimadores insesgados de sus correspondientes parámetros poblacionales:

1. La media muestral es un estimador insesgado de μ, $[E(\bar{x}) = \mu]$.
2. La varianza muestral es un estimador insesgado de σ^2, $[E(s^2) = \sigma^2]$.
3. La proporción muestral es un estimador insesgado de P, $[E(\hat{p}) = P]$.

Un estimador que no es insesgado es *sesgado*. El grado de sesgo es la diferencia entre la media del estimador y el verdadero parámetro.

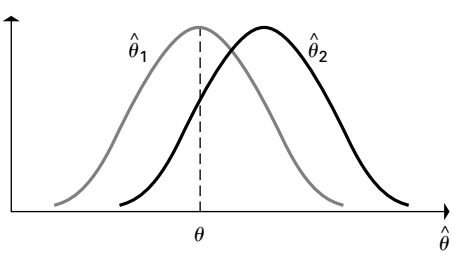


Figura 8.1. Funciones de densidad de los estimadores $\hat{\theta}_1$ (insesgado) y $\hat{\theta}_2$ (sesgado).

Sesgo

Sea $\hat{\theta}$ un estimador de θ. El **sesgo** de $\hat{\theta}$ es la diferencia entre su media y θ; es decir,

$$\text{Sesgo}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Se deduce que el sesgo de un estimador insesgado es 0.

No sólo es deseable que un estimador sea insesgado. Puede haber varios estimadores insesgados de un parámetro poblacional. Por ejemplo, si la población sigue una distribución normal, tanto la media muestral como la mediana son estimadores insesgados de la media poblacional.

Estimador consistente

Examinamos a continuación otra propiedad llamada *consistencia*.

Estimador consistente

Se dice que un estimador puntual $\hat{\theta}$ es un **estimador consistente** del parámetro θ si la diferencia entre el valor esperado del estimador y el parámetro disminuye a medida que aumenta el tamaño de la muestra. Es lo mismo que decir que el sesgo disminuye conforme aumenta el tamaño de la muestra.

Se utilizan estimadores consistentes en los casos en los que es difícil o imposible obtener estimadores insesgados, lo cual ocurre en algunos estudios econométricos avanzados. No todos los estimadores insesgados son consistentes y, por supuesto, no todos los estimadores consistentes son insesgados. Si la varianza muestral se calculara de la forma siguiente

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

sería un estimador sesgado de la varianza poblacional. Sin embargo, es consistente, ya que a medida que aumenta el tamaño de la muestra, tiende al estimador insesgado

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

En términos generales, el uso de un estimador consistente con una cantidad infinita de información sobre la muestra da el resultado correcto. En cambio, el uso de un estimador inconsistente no da el resultado correcto ni siquiera con una cantidad infinita de información sobre la muestra. Por este motivo, la inconsistencia de un estimador puntual se considera negativa.

Estimador eficiente

En muchos problemas prácticos, pueden obtenerse diferentes estimadores insesgados y es necesario encontrar algún método para elegir entre ellos. En esta situación, es lógico preferir el estimador cuya distribución esté más concentrada en torno al parámetro poblacional que se pretende estimar. Es menos probable que los valores de ese estimador difieran, en cualquier cantidad fija, del parámetro que se pretende estimar que los de sus competidores. Utilizando la varianza como medida de la concentración, introducimos la *eficiencia* de un estimador como criterio para preferir uno a otro.

Estimador más eficiente y eficiencia relativa

Si hay varios estimadores insesgados de un parámetro, el estimador insesgado que tiene la menor varianza es el **estimador más eficiente** o el **estimador insesgado de varianza mínima**. Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores insesgados de θ, basados en el mismo número de observaciones muestrales. En ese caso.

1. Se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$.
2. La **eficiencia relativa** de $\hat{\theta}_1$ con respecto a $\hat{\theta}_2$ es el cociente entre sus varianzas; es decir,

$$\text{Eficiencia relativa} = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$$

EJEMPLO 8.1. Selección entre estimadores insesgados rivales (eficiencia relativa)

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria extraída de una población que sigue una distribución normal de media μ y varianza σ^2. ¿Debe utilizarse la media muestral o la mediana muestral para estimar la media poblacional?

Solución

Suponiendo que la población sigue una distribución normal y es de gran tamaño en comparación con el tamaño de la muestra, la media muestral, $\bar{X}$, es un estimador insesgado de la media poblacional y tiene una varianza:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

También podría utilizarse como estimador la mediana de las observaciones muestrales. Puede demostrarse que este estimador también es insesgado con respecto a μ y que, cuando n es grande, su varianza es

$$\text{Var}(\text{Mediana}) = \frac{\pi}{2} \times \frac{\sigma^2}{n} = \frac{1{,}57\sigma^2}{n}$$

La media muestral es más eficiente que la mediana; la eficiencia relativa de la media con respecto a la mediana es

$$\text{Eficiencia relativa} = \frac{\text{Var}(\text{Mediana})}{\text{Var}(\bar{X})} = 1{,}57$$

La varianza de la mediana muestral es un 57 por ciento mayor que la de la media muestral. Para que la mediana muestral tuviera una varianza tan pequeña como la media muestral, tendría que basarse en un 57 por ciento más de observaciones. Una de las ventajas de la mediana frente a la media es que da menos peso a las observaciones extremas. Un posible inconveniente de la utilización de la mediana muestral como medida de la tendencia central se encuentra en su eficiencia relativa.

Subrayamos la importancia de la utilización de un gráfico de probabilidad normal para averiguar si hay alguna evidencia de ausencia de normalidad. Si la población no sigue una distribución normal, la media muestral puede no ser el estimador más eficiente de la media poblacional. En concreto, si los casos atípicos afectan mucho a la distribución poblacional, la media muestral es menos eficiente que otros estimadores (como

la mediana). La Tabla 8.1 resume algunas propiedades de algunos estimadores puntuales. No contiene ni una lista exhaustiva de estimadores ni una lista exhaustiva de las propiedades que posee un estimador.

Tabla 8.1. Propiedades de algunos estimadores puntuales.

Parámetro poblacional	Estimador puntual	Propiedades
Media, μ	$\overline{X}$	Insesgado, consistente, de máxima eficiencia (suponiendo la existencia de normalidad)
Media, μ	Mediana	Insesgado (suponiendo la existencia de normalidad), pero no de máxima eficiencia
Proporción, P	$\hat{p}$	Insesgado, consistente, de máxima eficiencia (suponiendo la existencia de normalidad)
Varianza, σ^2	s^2	Insesgado, consistente, de máxima eficiencia (suponiendo la existencia de normalidad)

EJEMPLO 8.2. Relaciones precio-beneficio (estimadores)

Supongamos que un día seleccionamos aleatoriamente una muestra de acciones que cotizan en la bolsa y observamos que las relaciones precio-beneficio de estas acciones son

10 16 13 11 12 14 12
15 14 14 13 13 13

¿Sugiere el gráfico de probabilidad normal la ausencia de normalidad? Halle estimaciones puntuales de la media y la varianza. Analice las propiedades de estos estimadores.

Solución

En el gráfico de probabilidad normal de la Figura 8.2, no se observa nada que indique ausencia de normalidad. Suponiendo que la distribución es normal, una estimación de las relaciones medias precio-beneficio es la media muestral, 13.1, y una estimación de la varianza es $s^2 = 2{,}58$. Tanto $\overline{X}$ como s^2 son estimadores puntuales insesgados, consistentes y eficientes de μ y σ^2, respectivamente.

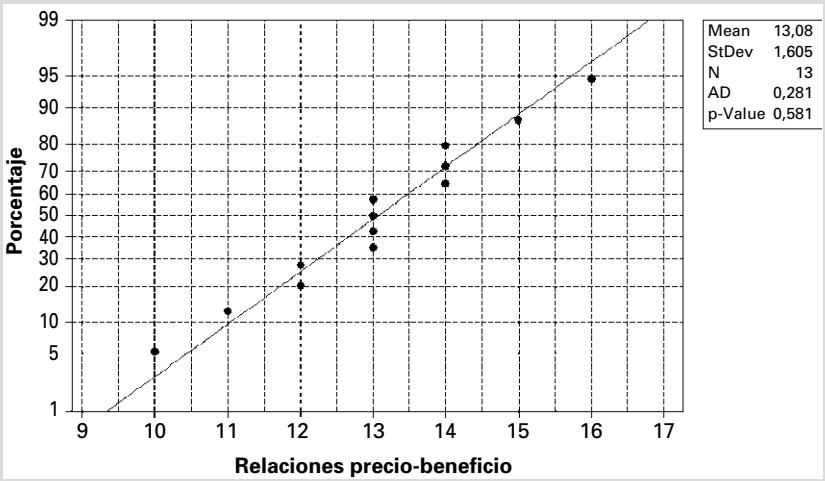


Figura 8.2. Ejemplo de relaciones precio-beneficio (Minitab).

Un problema que se plantea a menudo en la práctica es cómo elegir un estimador puntual adecuado de un parámetro poblacional. Una posibilidad atractiva es elegir el estimador insesgado más eficiente de todos. Sin embargo, a veces hay problemas de estimación en los que no es muy satisfactorio ningún estimador insesgado o situaciones en las que no siempre es posible encontrar un estimador insesgado de varianza mínima. También es posible que los datos no sigan una distribución normal. En estas situaciones, no es fácil seleccionar el mejor estimador puntual y la selección plantea considerables dificultades matemáticas que están fuera del alcance de este libro.

EJERCICIOS

Ejercicios básicos

8.1. Considere los datos siguientes:

6 8 7 10 3 5 9 8

a) Busque pruebas de la ausencia de normalidad.
b) Halle una estimación puntual de la media poblacional que sea insesgada, eficiente y consistente.
c) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de la media muestral.

8.2. Una muestra aleatoria de ocho viviendas de un barrio tenía los siguientes precios de venta (en miles de dólares):

92 83 112 127 109 96 102 90

a) Busque pruebas de la ausencia de normalidad.
b) Halle una estimación puntual de la media poblacional que sea insesgada y eficiente.
c) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de la media muestral.
d) Utilice un estimador insesgado para estimar la proporción de viviendas de este barrio que se venden por menos de 92.500 $.

8.3. Una muestra aleatoria de 10 economistas ha realizado las siguientes predicciones del crecimiento porcentual del producto interior bruto real del próximo año:

2,2 2,8 3,0 2,5 2,4 2,6 2,5 2,4 2,7 2,6

Utilice métodos de estimación insesgados para hallar las estimaciones puntuales de:

a) La media poblacional.
b) La varianza poblacional.
c) La varianza de la media muestral.
d) La proporción poblacional de economistas que han predicho un crecimiento del producto interior bruto real de al menos un 2,5 por ciento.
e) La varianza de la proporción muestral de economistas que han predicho un crecimiento del producto interior bruto real de al menos un 2,5 por ciento.

8.4. Una muestra aleatoria de 12 obreros de una gran fábrica encontró las siguientes cifras sobre el número de horas extraordinarias realizadas el mes anterior:

22 16 28 12 18 36 23 11 41 29 26 31

Utilice métodos de estimación insesgados para hallar estimaciones puntuales de

a) La media poblacional.
b) La varianza poblacional.
c) La varianza de la media muestral.
d) La proporción poblacional de obreros que trabajaron más de 30 horas extraordinarias en esta fábrica el mes anterior.
e) La varianza de la proporción muestral de obreros que trabajaron más de 30 horas extraordinarias en esta fábrica el mes anterior.

Ejercicios aplicados

8.5. Project Romanian Rescue (PRR) es una fundación rumana registrada que atiende las necesidades de los niños trágicamente desfavorecidos de Constanta (Rumanía) (véase la referencia bibliográfica 7). Las actividades de PRR, como misión cristiana interconfesional que es, son un programa de contacto, un centro de día, un albergue de niños (Casa Charis), un albergue de niñas (Casa Chara) y ayuda educativa individualizada para los niños de familias pobres. PRR planea abrir un centro en la vecina Kogalniceanu para albergar a más niños de la calle. Supongamos que Daniel Mercado, fundador del proyecto, y Camelia Vilcoci, directora ejecutiva del proyecto, disponen de información como el número de almuerzos repartidos diariamente entre los niños de la calle, el nú-

mero de niños que asisten al centro de día y la edad de los niños, y supongamos que el fichero de datos **PRR** contiene una muestra aleatoria de esa información.

a) Compruebe cada variable para averiguar si los datos siguen una distribución normal.

b) Halle estimaciones insesgadas de la media poblacional y de la varianza poblacional.

8.6. Suponga que x_1 y x_2 son muestras aleatorias de observaciones extraídas de una población de media μ y varianza σ^2. Considere los tres estimadores puntuales siguientes, X, Y, Z de μ:

$$X = \frac{1}{2}x_1 + \frac{1}{2}x_2 \qquad Y = \frac{1}{4}x_1 + \frac{3}{4}x_2$$

$$Z = \frac{1}{3}x_1 + \frac{2}{3}x$$

a) Demuestre que los tres estimadores son insesgados.

b) ¿Cuál de los estimadores es más eficiente?

c) Halle la eficiencia relativa de X con respecto a cada uno de los otros dos estimadores.

8.7. Al Fiedler, director de planta de LDS Vacuum Products, que se encuentra en Altamonte Springs (Florida), aplica la teoría estadística en su centro de trabajo. LDS, importante proveedor de los fabricantes de automóviles, quiere estar seguro de que la tasa de incidencia de fugas (en centímetros cúbicos por segundo) de los enfriadores del aceite de la transmisión (TOC) satisface los límites de especificación establecidos. Se comprueba una muestra aleatoria de 50 TOC y se anotan las tasas de incidencia de fugas en el fichero llamado **TOC** (véase la referencia bibliográfica 3).

a) ¿Existen pruebas de que los datos no siguen una distribución normal?

b) Halle una estimación puntual insesgada de varianza mínima de la media poblacional.

c) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de la media muestral.

8.8. La demanda de agua embotellada aumenta durante la temporada de huracanes en Florida. El director de operaciones de una planta que embotella agua potable quiere estar seguro de que el proceso de embotellado de las botellas de 1 galón está funcionando bien. Actualmente, la compañía está comprobando el volumen de las botellas de 1 galón. Suponga que se comprueba una muestra aleatoria de 75 botellas y que se anotan las mediciones en el fichero de datos **Water**.

a) ¿Existen pruebas de que los datos no siguen una distribución normal?

b) Halle una estimación puntual insesgada de varianza mínima de la media poblacional.

c) Halle una estimación puntual insesgada de varianza mínima de la varianza poblacional.

8.2. Intervalos de confianza de la media: varianza poblacional conocida

Primero suponemos que se toma una muestra aleatoria de una población que sigue una distribución normal y que tiene una media desconocida y una varianza *conocida*. Nuestro objetivo es hallar un intervalo de valores, en lugar de un único número, para estimar una media poblacional. *Este problema a veces es poco realista*, *ya que en raras ocasiones se conoce exactamente la varianza poblacional y*, *sin embargo*, *la media es desconocida.* A veces sí ocurre, sin embargo, que se han hecho tantas muestras a poblaciones similares que puede suponerse que la varianza de la población de interés se conoce bastante bien por experiencia. Cuando el tamaño de la muestra n es bastante grande, pueden utilizarse los métodos desarrollados para el caso en el que se conoce la varianza poblacional si hay que estimar esa varianza a partir de la muestra. No obstante, la principal ventaja de comenzar con este problema se halla en que permite hacer una exposición bastante fácil de los métodos necesarios para hallar intervalos de confianza.

El número medio de automóviles producidos diariamente en una fábrica es una importante medida. Si ese número es a menudo muy diferente, por encima o por debajo, de la media, la fábrica puede tener excesivos costes en existencias o pérdidas de ventas. Se necesita un estimador y una estimación que tengan en cuenta esta variación y que den un

intervalo de valores en el que parece probable que se encuentre la cantidad que se pretende estimar. En este apartado, explicamos el formato general de esos estimadores.

Cuando se hace un muestreo de una población, manteniéndose todo lo demás constante, se obtiene una *información* más segura sobre esa población con una muestra relativamente grande que con una muestra más pequeña. Sin embargo, este factor no se refleja en las estimaciones puntuales. Por ejemplo, la estimación puntual de la proporción de piezas defectuosas que hay en un envío sería la misma si se encontrara 1 pieza defectuosa en una muestra de 10 piezas que si se encontraran 100 piezas defectuosas en una muestra de 1.000 piezas. El grado de precisión de nuestra información sobre los parámetros poblacionales se refleja en las *estimaciones de intervalos de confianza*; concretamente, cuanto mayor es el tamaño de la muestra, menores son, manteniéndose todo lo demás constante, las estimaciones de intervalos que reflejan nuestra incertidumbre sobre el verdadero valor de un parámetro.

Estimador de intervalos de confianza

Un **estimador de un intervalo de confianza** de un parámetro poblacional es una regla para hallar (basándose en la información muestral) un intervalo que es probable que incluya ese parámetro. La estimación correspondiente se llama **estimación de un intervalo de confianza**.

Hasta ahora hemos dicho que es «probable» o «muy probable» que los estimadores de intervalos de confianza incluyan el valor verdadero, pero desconocido, del parámetro poblacional. Para que nuestro análisis sea más preciso, es necesario expresar esas afirmaciones en términos probabilísticos. Supongamos que se ha tomado una muestra aleatoria y que, basándose en la información muestral, es posible hallar dos variables aleatorias, A y B, y que A es menor que B. Si los valores muestrales específicos de las variables aleatorias A y B son a y b, el intervalo de a a b incluye el parámetro o no lo incluye. No lo sabemos realmente con seguridad.

Supongamos, sin embargo, que se toman repetidamente muestras aleatorias de la población y se hallan de esta misma forma intervalos similares. A largo plazo, un cierto porcentaje de estos intervalos (por ejemplo, el 95 o el 98 por ciento) contendrá el valor desconocido. Según el concepto de probabilidad basado en la frecuencia relativa, esos intervalos pueden interpretarse de la manera siguiente: *si se hacen repetidos muestreos de una población y se calculan intervalos de esta forma, a largo plazo el* 95 *por ciento* (*o algún otro porcentaje*) *de los intervalos contendrá el verdadero valor del parámetro desconocido*. Se dice entonces que el intervalo A a B es un estimador de un intervalo de confianza al 95 por ciento de la proporción poblacional. Este resultado puede generalizarse de inmediato.

Intervalo de confianza y nivel de confianza

Sea θ un parámetro desconocido. Supongamos que, basándose en la información muestral, se hallan variables aleatorias A y B tales que $P(A < \theta < B) = 1 - \alpha$, donde α es cualquier número comprendido entre 0 y 1. Si los valores muestrales específicos de A y B son a y b, entonces el intervalo de a a b se llama **intervalo de confianza** de θ al $100(1 - \alpha)\%$. La cantidad $100(1 - \alpha)\%$ se llama **nivel de confianza** del intervalo.

Si se extraen repetidamente muestras aleatorias de la población, el verdadero valor del parámetro θ se encontrará en el $100(1 - \alpha)\%$ de los intervalos calculados de esta forma. El intervalo de confianza calculado de esta forma se expresa de la manera siguiente: $a < \theta < b$ a un nivel de confianza del $100(1 - \alpha)\%$.

Conviene tener presente que siempre que se extrae una muestra aleatoria, existe la posibilidad de que haya una diferencia entre el valor de un estimador y el verdadero valor del parámetro. El verdadero valor de un parámetro desconocido podría ser algo mayor o algo menor que el valor hallado incluso por medio del mejor estimador puntual. No es sorprendente que, en muchos problemas de estimación, una estimación de intervalos de confianza del parámetro desconocido adopte la forma siguiente: mejor estimación puntual $\pm$ un factor de error.

Intervalos basados en la distribución normal

Sea $x_1, x_2, ..., x_n$ una muestra aleatoria de n observaciones extraídas de una población que sigue una distribución normal de media μ desconocida y varianza conocida σ^2. Supongamos que queremos un intervalo de confianza de la media poblacional al $100(1 - \alpha)\%$. En el Capítulo 7 vimos que

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

sigue una distribución normal estándar y $z_{\alpha/2}$ es el valor de la distribución normal estándar tal que la probabilidad de la cola superior es $\alpha/2$. Utilizamos el álgebra básica para hallar

$$\begin{aligned} 1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &= P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \\ &= P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

En el caso de un nivel de confianza del 95 por ciento, se deduce que

$$P\left(\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

La Figura 8.3 muestra que la probabilidad de que una variable aleatoria normal estándar se encuentre entre los números $-1{,}96$ y $1{,}96$ es $0{,}95$.

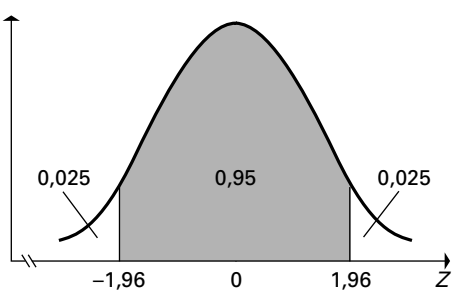


Figura 8.3. $P(-1{,}96 < z < 1{,}96) = 0{,}95$, donde z es una variable aleatoria normal estándar.

Intervalos de confianza de la media de una población que sigue una distribución normal: varianza poblacional conocida

Consideremos una muestra aleatoria de n observaciones extraídas de una población que sigue una distribución normal de media μ y varianza σ^2. Si la media muestral es $\bar{x}$, entonces el **intervalo de confianza** al 100 $(1 - \alpha)$% **de la media poblacional, cuando la varianza es conocida,** viene dado por

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.1}$$

o, lo que es lo mismo,

$$\bar{x} \pm ME$$

donde *ME*, el **margen de error** (también llamado **error de muestreo**), es

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.2}$$

La **amplitud**, *w*, es igual al doble del margen de error:

$$w = 2(ME) \tag{8.3}$$

El **límite superior de confianza, *LSC*,** es

$$LSC = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.4}$$

El **límite inferior de confianza, *LIC*,** es

$$LIC = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{8.5}$$

Es necesario interpretar exactamente los intervalos de confianza. Si se extraen repetida e independientemente muestras aleatorias de n observaciones de la población y se calculan intervalos de confianza al $100(1 - \alpha)$% mediante la ecuación 8.1, entonces, en un elevado número de pruebas repetidas, el $100(1 - \alpha)$% de estos intervalos contendrá el verdadero valor de la media poblacional.

La Tabla 8.2 muestra los valores de $Z_{\alpha/2}$, llamados a veces **factor de fiabilidad**, correspondientes a algunos niveles de confianza. En el caso del intervalo de confianza al 90 por ciento, la ecuación 8.1 se convierte en

$$\bar{x} - 1{,}645\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}645\frac{\sigma}{\sqrt{n}}$$

En el caso del intervalo de confianza al 95 por ciento, la ecuación 8.1 se convierte en

$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}$$

Tabla 8.2. Algunos niveles de confianza y los valores de $Z_{\alpha/2}$ correspondientes.

Nivel de confianza	90%	95%	98%	99%
α	0,10	0,05	0,02	0,01
$z_{\alpha/2}$	1,645	1,96	2,33	2,58

EJEMPLO 8.3. Tiempo en la tienda de alimentación (intervalo de confianza)

Supongamos que el tiempo que permanecen los clientes en una tienda local de alimentación sigue una distribución normal. Una muestra aleatoria de 16 clientes tenía un tiempo medio de 25 minutos. Supongamos que $\sigma = 6$ minutos. Halle el error típico, el margen de error y la amplitud del intervalo de confianza de la media poblacional, μ, al 95 por ciento.

Solución

El error típico y el margen de error son

$$\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{16}} = 1{,}5$$

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96(1{,}5) = 2{,}94$$

Por lo tanto, la amplitud es igual a $2(2{,}94) = 5{,}88$ y el intervalo de confianza al 95 por ciento es $22{,}06 < \mu < 27{,}94$.

¿Cómo debe interpretarse ese intervalo de confianza? Basándonos en una muestra de 16 observaciones, el intervalo de confianza de la media poblacional desconocida al 95 por ciento va de alrededor de 22 minutos a alrededor de 28 minutos. Ahora bien, esta muestra no es más que una de las muchas que podrían haberse extraído de la población. Si comenzamos de nuevo y tomamos una segunda muestra de 16 compradores, es casi seguro que la media de la segunda muestra será diferente de la de la primera. Por lo tanto, si se calcula el intervalo de confianza al 95 por ciento a partir de los resultados de la segunda muestra, probablemente será diferente del intervalo anterior. Imaginemos que tomamos un número muy grande de muestras aleatorias independientes de 16 observaciones de esta población y que a partir de cada resultado muestral calculamos el intervalo de confianza al 95 por ciento. *El nivel de confianza del intervalo implica que a largo plazo el 95 por ciento de los intervalos obtenidos de esta forma contiene el verdadero valor de la media poblacional*. Es en este sentido en el que se dice que hay una confianza del 95 por ciento en nuestra estimación del intervalo. Sin embargo, no se sabe si nuestro intervalo es uno de los que pertenecen al 95 por ciento de los buenos o al 5 por ciento de los malos sin conocer μ.

La Figura 8.4 muestra la distribución en el muestreo de la media muestral de n observaciones procedentes de una población que sigue una distribución normal de media μ y desviación típica σ. Esta distribución en el muestreo sigue una distribución normal de media μ y desviación típica $\sigma/\sqrt{n}$. El intervalo de confianza de la media poblacional se basará en el valor observado de la media muestral, es decir, en una observación extraída de nuestra distribución en el muestreo.

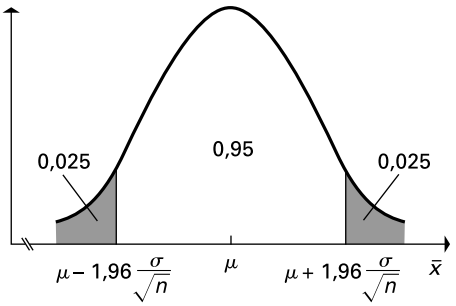


Figura 8.4. Distribución en el muestreo de la media muestral de n observaciones procedentes de una distribución normal de media μ, varianza σ^2 y un nivel de confianza del 95 por ciento.

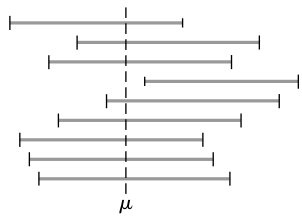


Figura 8.5. Descripción esquemática de intervalos de confianza al 95 por ciento.

La Figura 8.5 muestra una descripción esquemática de una secuencia de intervalos de confianza al 95 por ciento, obtenidos de muestras independientes extraídas de la población. Los centros de estos intervalos, que son simplemente las medias muestrales observadas, a menudo estarán muy cerca de la media poblacional, μ. Sin embargo, algunos pueden diferir mucho de μ. Se deduce que el 95 por ciento de un gran número de estos intervalos contendrá la media poblacional.

EJEMPLO 8.4. Azúcar refinado (intervalo de confianza)

Un proceso produce bolsas de azúcar refinado. El peso del contenido de estas bolsas sigue una distribución normal que tiene una desviación típica de 12 gramos. El contenido de una muestra aleatoria de 25 bolsas tiene un peso medio de 198 gramos. Halle el límite superior de confianza y el inferior del intervalo de confianza al 99 por ciento del verdadero peso medio de todas las bolsas de azúcar producidas por el proceso.

Solución

En el caso del intervalo de confianza al 99 por ciento, el factor de fiabilidad es

$$z_{0,005} = 2{,}58$$

y con una media muestral de 198, $n = 25$, y una desviación típica de 12, los límites de confianza son

$$LSC = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 198 + 2{,}58\,\frac{12}{\sqrt{25}} = 204{,}2$$

$$LIC = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 198 - 2{,}58\,\frac{12}{\sqrt{25}} = 191{,}8$$

Reducción del margen de error

¿Puede reducirse el margen de error (y, por consiguiente, la amplitud) de un intervalo de confianza? Consideremos los factores que afectan al margen de error: la desviación típica poblacional, el tamaño de la muestra n y el nivel de confianza.

Manteniendo todos los demás factores constantes, cuanto más puede reducirse la desviación típica poblacional, σ, menor es el margen de error. Las empresas se esfuerzan en

reducir la variabilidad en la medición de los productos (Capítulo 18). Cuando es posible, ése debe ser el primer paso para reducir la amplitud. Sin embargo, a veces no es posible reducir la desviación típica poblacional.

INTERPRETACIÓN

Otra manera de reducir el margen de error es aumentar el tamaño de la muestra. De esa forma se reduce la desviación típica de la distribución de la media muestral en el muestreo y, por lo tanto, el margen de error. Es decir, manteniendo constantes todos los demás factores, un aumento del tamaño de la muestra n reduce el margen de error. Cuanta más información se obtiene de una población, más precisa debe ser la inferencia sobre su media. Cuando se examine la ecuación del margen de error, obsérvese que la amplitud del intervalo es directamente proporcional a $1/\sqrt{n}$. Por ejemplo, si se multiplica por 4 el tamaño de la muestra, la amplitud del intervalo se reduce a la mitad. Si el tamaño de la muestra fuera inicialmente de 100, un aumento de su tamaño de 400 provocaría una reducción de la amplitud del intervalo inicial de confianza a la mitad (manteniendo todos los demás factores constantes). El inconveniente del aumento del tamaño de la muestra es el incremento de los costes.

Por último, manteniendo todos los demás factores constantes, si se reduce el nivel de confianza $(1 - \alpha)$, el margen de error disminuye. Por ejemplo, un intervalo de confianza al 95 por ciento es menor que un intervalo de confianza al 99 por ciento basado en la misma observación. *Advertencia*: la reducción del nivel de confianza reduce la probabilidad de que el intervalo contenga el valor del verdadero parámetro poblacional. La Figura 8.6 muestra algunos de los efectos del tamaño de la muestra n, la desviación típica poblacional σ y el nivel de confianza $(1 - \alpha)$ en los intervalos de confianza de la media de una población que sigue una distribución normal; la media muestral es en todos los casos 19,80.

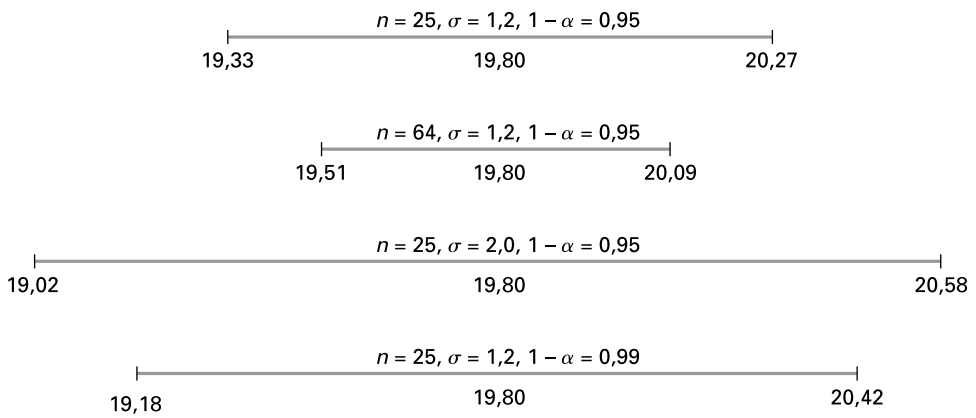


Figura 8.6. Efectos del tamaño de la muestra, la desviación típica poblacional y el nivel de confianza en los intervalos de confianza.

EJERCICIOS

Ejercicios básicos

8.9. Halle el factor de fiabilidad, $z_{\alpha/2}$, en cada uno de los casos siguientes:

a) un nivel de confianza del 96 por ciento
b) un nivel de confianza del 88 por ciento
c) un nivel de confianza del 85 por ciento
d) $\alpha = 0{,}07$
e) $\alpha/2 = 0{,}07$

8.10. Calcule el margen de error para estimar la media poblacional, μ, en los casos siguientes:

a) un nivel de confianza del 98 por ciento; $n = 64$; $\sigma^2 = 144$
b) un nivel de confianza del 99 por ciento; $n = 120$; $\sigma = 100$.

8.11. Calcule la amplitud para estimar la media poblacional, μ, en los casos siguientes:

a) un nivel de confianza del 90 por ciento; $n = 100$; $\sigma^2 = 169$
b) un nivel de confianza del 95 por ciento; $n = 120$; $\sigma = 25$

8.12. Calcule el *LIC* y el *LSC* de
a) $\bar{x} = 50$; $n = 64$; $\sigma = 40$; $\alpha = 0{,}05$
b) $\bar{x} = 85$; $n = 225$; $\sigma^2 = 400$; $\alpha = 0{,}01$
c) $\bar{x} = 510$; $n = 485$; $\sigma = 50$; $\alpha = 0{,}10$

Ejercicios aplicados

8.13. Un director de personal ha observado que históricamente las puntuaciones de los tests de aptitud realizados a los solicitantes de empleo en los niveles de entrada siguen una distribución normal con una desviación típica de 32,4 puntos. Una muestra aleatoria de nueve puntuaciones del grupo actual de solicitantes tenía una puntuación media de 187,9 puntos.
a) Halle el intervalo de confianza al 80 por ciento de la media poblacional de las puntuaciones del grupo actual de solicitantes.
b) Basándose en estos resultados muestrales, un estadístico ha hallado para la media poblacional un intervalo de confianza que va de 165,8 a 210,0 puntos. Halle el nivel de confianza de este intervalo.

8.14. Se sabe que la desviación típica de los volúmenes de las botellas de 710 ml de agua mineral embotellada por una empresa es de 6 ml. Se ha tomado una muestra aleatoria de 90 botellas y se han medido.
a) Halle el factor de fiabilidad de un intervalo de confianza al 92 por ciento de la media poblacional de los volúmenes.
b) Calcule el error típico de la media.
c) Calcule la amplitud de un intervalo de confianza al 92 por ciento de la media poblacional de los volúmenes.

8.15. La secretaría de admisiones en un programa de máster en administración de empresas ha observado que históricamente los solicitantes tienen una calificación media en los estudios de licenciatura que sigue una distribución normal con una desviación típica de 0,45. Se ha extraído una muestra aleatoria de 25 solicitudes cuya calificación media ha resultado ser 2,90.
a) Halle el intervalo de confianza de la media poblacional al 95 por ciento.
b) Basándose en estos resultados muestrales, un estadístico calcula para la media poblacional el intervalo de confianza que va de 2,81 a 2,99. Halle el nivel de confianza correspondiente a este intervalo.

8.16. Se sabe que el peso de los ladrillos que produce una fábrica sigue una distribución normal con una desviación típica de 0,12 kilos. Una muestra aleatoria de 16 ladrillos de la producción de hoy tenía un peso medio de 4,07 kilos.
a) Halle el intervalo de confianza al 99 por ciento del peso medio de todos los ladrillos producidos hoy.
b) Explique sin realizar los cálculos si el intervalo de confianza al 95 de la media poblacional tendría más amplitud, menos o igual que la obtenida en el apartado (a).
c) Se decide que mañana se tomará una muestra de 20 ladrillos. Explique sin realizar los cálculos si el intervalo de confianza al 99 por ciento del peso medio de la producción de mañana calculado correctamente tendría más amplitud, menos o igual que la obtenida en el apartado (a).
d) Suponga que la desviación típica poblacional de la producción de hoy es de 0,15 kilos (no 0,12 kilos). Explique sin realizar los cálculos si el intervalo de confianza al 99 por ciento del peso medio de la producción de hoy calculado correctamente tendría más amplitud, menos o igual que la obtenida en el apartado (a).

8.3. Intervalos de confianza de la media: varianza poblacional desconocida

En el apartado anterior hemos explicado los intervalos de confianza de la media de una población normal cuando se conoce la varianza poblacional. A continuación, estudiamos el caso en el que no se conoce el valor de la varianza poblacional y que tiene considerable importancia práctica. Por ejemplo:

1. Los ejecutivos de cadenas de establecimientos minoristas pueden querer estimar las ventas diarias medias de sus tiendas.

2. Los fabricantes pueden querer estimar la productividad media, en unidades por hora, de lo trabajadores que utilizan un determinado proceso de producción.
3. Los fabricantes de automóviles y de camiones pueden querer estimar el consumo medio de combustible, expresado en kilómetros por litro, de un determinado modelo.

En estos tipos de situaciones, es probable que no exista ninguna información histórica sobre la media poblacional o sobre la varianza poblacional. Para avanzar es necesario introducir una nueva clase de distribuciones de probabilidad que desarrolló William Sealy Gosset, estadístico irlandés que trabajó en la Guinness Brewery de Dublín a principios de la década de 1900 (véase la referencia bibliográfica 5).

Distribución *t* de Student

Gosset trató de desarrollar una distribución de probabilidad, cuando no se conoce la varianza poblacional σ^2, de una variable aleatoria que sigue una distribución normal. En aquella época, estaba comenzando a realizarse tests de laboratorio y a aplicarse el método científico en la industria cervecera. Gosset, cuyos trabajos aparecieron con el pseudónimo de «Student», influyó mucho en el desarrollo moderno del pensamiento estadístico y de la variación de los procesos. «Las circunstancias en las que se elabora la cerveza, con sus variables materias primas y su sensibilidad a los cambios de temperatura [...] subrayan la necesidad de disponer de un método correcto para tratar muestras pequeñas. No fue, pues, la casualidad, sino las circunstancias de su trabajo, las que llevaron a Student a centrar la atención en este problema y a descubrir la distribución de la desviación típica muestral» (véase la referencia bibliográfica 6). Gosset demostró la conexión entre la investigación estadística y los problemas prácticos. La distribución aún se conoce con el nombre de «distribución *t* de Student». La distribución *t* desarrollada por Gosset es el cociente entre dos distribuciones, la distribución normal estándar y la raíz cuadrada de la distribución ji-cuadrado dividida por sus grados de libertad, v (véase el apéndice del capítulo).

El apartado 8.2 se basaba en el hecho de que la variable aleatoria, Z, que viene dada por

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

sigue una distribución normal estándar. En el caso en el que la desviación típica poblacional es desconocida, este resultado no puede utilizarse directamente. En esas circunstancias, es lógico considerar la variable aleatoria obtenida sustituyendo la σ desconocida por la desviación típica muestral, s, lo que nos da

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Esta variable aleatoria no sigue una distribución normal estándar. Sin embargo, su distribución se conoce y es, de hecho, un miembro de una familia de distribuciones llamadas *t* de Student.

Distribución *t* de Student

Dada una muestra aleatoria de n observaciones, de media $\bar{x}$ y desviación típica s, extraída de una población que sigue una distribución normal de media μ, la variable aleatoria t sigue la **distribución *t* de Student** con $(n - 1)$ grados de libertad y viene dada por

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Un miembro específico de la familia de distribuciones t de Student se caracteriza por el número de grados de libertad. Utilizaremos el parámetro v para representar los grados de libertad y t_v para representar una variable aleatoria t de Student con v grados de libertad. La forma de la distribución t de Student es bastante parecida a la de la distribución normal estándar. Ambas distribuciones tienen una media de 0 y las funciones de densidad de las dos son simétricas en torno a sus medias. Sin embargo, la función de densidad de la distribución t de Student tiene una dispersión mayor (reflejada en una varianza mayor) que la distribución normal estándar, como puede verse en la Figura 8.7, que muestra las funciones de densidad de la distribución normal estándar y de la distribución t de Student con 3 grados de libertad.

La dispersión mayor de la distribución t de Student se debe a la incertidumbre adicional provocada por la sustitución de la desviación típica poblacional conocida por su estimador muestral. A medida que aumenta el número de grados de libertad, la distribución t de Student es cada vez más parecida a la distribución normal estándar. Cuando el número de grados de libertad es alto, las dos distribuciones son casi idénticas. Es decir, la distribución t de Student converge hacia $N(0, 1)$, que es bastante parecida a la t si n es grande. Este resultado es intuitivamente razonable y se deduce del hecho de que cuando la muestra es grande, la desviación típica muestral es un estimador muy preciso de la desviación típica poblacional.

Para basar las inferencias sobre una media poblacional en la distribución t de Student, se necesitan valores críticos análogos a $z_{\alpha/2}$. De la misma forma que $z_{\alpha/2}$ es el valor de la distribución normal estándar tal que la probabilidad de la cola superior es $\alpha/2$, $t_{v,\,\alpha/2}$ es el valor de la distribución t de Student para v (grados de libertad) tal que la probabilidad de la cola superior es $\alpha/2$, como muestra la Figura 8.8.

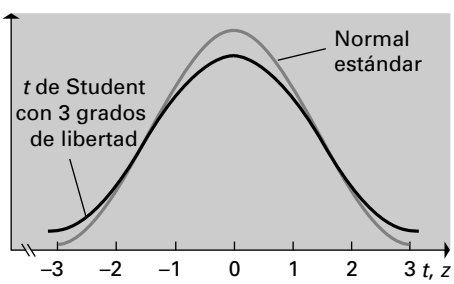


Figura 8.7. Funciones de densidad de la distribución normal estándar y la distribución t de Student con 3 grados de libertad.

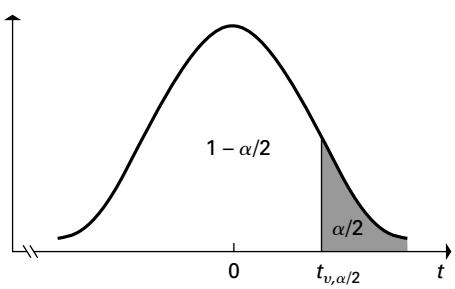


Figura 8.8. $P(t_v > t_{v,\,\alpha/2}) = \alpha/2$, donde t_v es una variable aleatoria t de Student con v grados de libertad.

Notación

Una variable aleatoria que tiene la distribución *t* de Student con *v* grados de libertad se representa por medio de t_v. $t_{v,\alpha/2}$ es el factor de fiabilidad, que es el número para el que

$$P(t_v > t_{v,\alpha/2}) = \alpha/2$$

Supongamos que tenemos que hallar un número tal que una variable aleatoria que sigue una *t* de Student con 15 grados de libertad lo supera con una probabilidad de 0,05. Es decir,

$$P(t_{15} > t_{15,0,05}) = 0{,}05$$

Consultando directamente la tabla de la distribución *t* de Student, tenemos que

$$t_{15,0,05} = 1{,}753$$

También pueden utilizarse muchos programas informáticos para hallar estos valores.

Intervalos basados en la distribución *t* de Student

Nos encontraremos con muchas situaciones en las que no se conoce la varianza poblacional. Para hallar el intervalo de confianza al $100(1-\alpha)\%$ para este tipo de problema se sigue exactamente el mismo razonamiento que en el apartado 8.2. La terminología es análoga.

Intervalos de confianza de la media de una población normal: varianza poblacional desconocida

Supongamos que tenemos una muestra aleatoria de *n* observaciones extraídas de una *distribución normal* de media μ y varianza desconocida. Si la media y la desviación típica muestrales son, respectivamente, $\bar{x}$ y *s*, entonces los grados de libertad $v = n - 1$ y el **intervalo de confianza** al $100(1-\alpha)\%$ **de la media poblacional, cuando la varianza es desconocida**, viene dado por

$$\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \qquad \textbf{(8.6)}$$

o, lo que es lo mismo,

$$\bar{x} \pm ME$$

donde *ME*, el **margen de error**, es

$$ME = t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \qquad \textbf{(8.7)}$$

Supongamos que tenemos una muestra aleatoria de *n* observaciones extraída de una población normal de media μ y varianza desconocida y que hay que hallar intervalos de confianza de la media poblacional. El gráfico de probabilidad normal es uno de los métodos para averiguar si los datos no siguen una distribución normal. En este capítulo hemos supuesto en todos los casos que los datos siguen una distribución normal. En las aplicaciones relacionadas con el mundo empresarial y el político y con las investigaciones médicas y de otros tipos, hay que comprobar primero si los datos siguen o no una distribución normal. La terminología de los intervalos de confianza de una media poblacional cuando la varianza es desconocida es similar a la terminología que se emplea cuando la varianza es conocida.

EJEMPLO 8.5. Camiones: consumo de gasolina (intervalo de confianza)

Los precios de la gasolina experimentaron una vertiginosa subida en los primeros años de este siglo. Supongamos que se ha realizado recientemente un estudio con camioneros que tenían más o menos el mismo número de años de experiencia para comprobar el comportamiento de 24 camiones de un determinado modelo en la misma autopista. Estime la media poblacional del consumo de combustible de este modelo de camión con una confianza del 90 por ciento suponiendo que el consumo de combustible, en millas por galón, de estos 24 camiones es

15,5	21,0	18,5	19,3	19,7	16,9	20,2	14,5
16,5	19,2	18,7	18,2	18,0	17,5	18,5	20,5
18,6	19,1	19,8	18,0	19,8	18,2	20,3	21,8

Trucks

Los datos se encuentran en el fichero de datos **TRUCKS**.

Solución

En el gráfico de probabilidad normal de la Figura 8.9 no se observa nada que indique la ausencia de normalidad. Calculando la media y la desviación típica, tenemos que

$$\bar{x} = 18{,}68 \qquad s = 1{,}69526 \qquad t_{n-1,\,\alpha/2} = t_{23,\,0{,}05} = 1{,}714$$

Aplicando la ecuación 8.6, el intervalo de confianza al 90 por ciento es

$$\bar{x} \pm t_{n-1,\,\alpha/2}\frac{s}{\sqrt{n}} = 18{,}68 \pm t_{23,\,0{,}05}\frac{1{,}69526}{\sqrt{24}} = 18{,}68 \pm (1{,}714) \times (0{,}3460)$$
$$= 18{,}68 \pm 0{,}5930$$

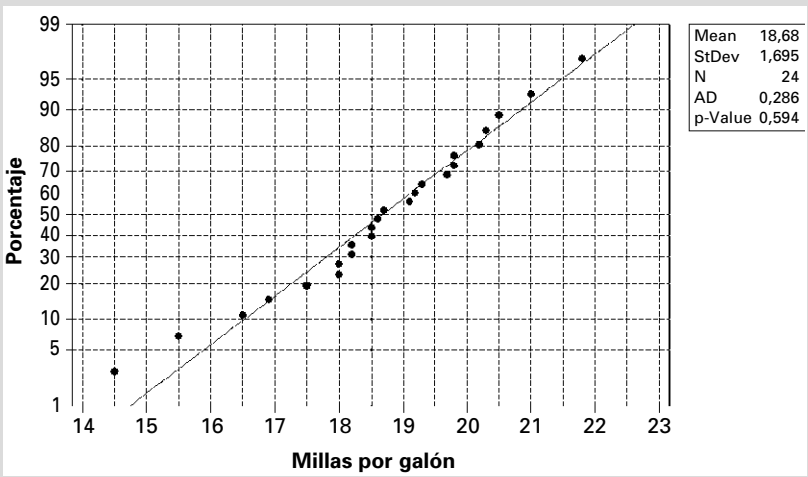


Figura 8.9. Gráfico de probabilidad normal.

El intervalo de confianza es, pues, $18{,}1 < \mu < 19{,}3$. Cuando el conjunto de datos es mayor, se utiliza el computador. La Figura 8.10 es la salida Minitab y la 8.11 es la salida Excel generadas para el ejemplo 8.5.

La interpretación del intervalo de confianza es importante. Si se seleccionan repetidamente muestras aleatorias independientes de 24 camiones de la población y se calcu-

lan intervalos de confianza de cada una de las muestras, en un número muy grande de pruebas repetidas el 90 por ciento de estos intervalos contendrá el valor del verdadero consumo medio de combustible de este modelo de camión. Sin embargo, en la práctica no se extraen repetidamente muestras independientes.

```
Variable   N      Mean     StDev    SE Mean          90% cr
MPG       24   18,6792    1,6953     0,3460   (18,0861. 19,2722)
```

Figura 8.10. Salida para el fichero de datos **Trucks** del ejemplo 8.5 (salida Minitab).

Media	**18,67917**
Error típico	0,346043
Mediana	18,65
Moda	18,5
Desviación típica	1,695257
Varianza muestral	2,873895
Curtosis	0,624798
Sesgo	–0,60902
Rango	7,3
Mínimo	14,5
Máximo	21,8
Suma	448,3
Número de casos	24
Nivel de confianza (90,0%)	**0,593073**

Figura 8.11. Salida para el fichero de datos **Trucks** del ejemplo 8.5 (salida Excel).

EJERCICIOS

Ejercicios básicos

8.17. Halle el factor de fiabilidad, $t_{v,\alpha/2}$, para estimar la media poblacional, μ, en los casos siguientes:

a) $n = 20$; nivel de confianza del 90%
b) $n = 7$; nivel de confianza del 98%
c) $n = 16$; nivel de confianza del 95%
d) $n = 23$; nivel de confianza del 99%

8.18. Halle el margen de error en los casos siguientes:

a) $n = 20$; nivel de confianza del 90%; $s = 36$
b) $n = 7$; nivel de confianza del 98%; $s = 16$
c) $n = 16$; nivel de confianza del 95%; $s^2 = 43$
d) nivel de confianza del 99%; $x_1 = 15$; $x_2 = 17$; $x_3 = 13$; $x_4 = 11$

8.19. El tiempo (en minutos) que tarda una muestra aleatoria de cinco personas en desplazarse al trabajo es

30 42 35 40 45

a) Calcule el error típico.
b) Halle $t_{v,\alpha/2}$ correspondiente a el intervalo de confianza de la verdadera media poblacional al 95 por ciento.
c) Calcule la amplitud de un intervalo de confianza al 95 por ciento de la media poblacional del tiempo que se tarda en desplazarse al trabajo.

8.20. Halle el *LIC* y el *LSC* en los casos siguientes:

a) $\alpha = 0{,}05$; $n = 25$; $\bar{x} = 560$; $s = 45$
b) $\alpha/2 = 0{,}05$; $n = 9$; $\bar{x} = 160$; $s^2 = 36$
c) $1 - \alpha = 0{,}98$; $n = 22$; $\bar{x} = 58$; $s = 15$

8.21. Calcule el margen de error para estimar la media poblacional, μ, en los casos siguientes:

a) un nivel de confianza del 98%; $n = 64$; $s^2 = 144$
b) un nivel de confianza del 99%; $n = 120$; $s^2 = 100$
c) un nivel de confianza del 95%; $n = 200$; $s^2 = 40$

8.22. Calcule la amplitud en los casos siguientes:

a) $n = 6$; $s = 40$; $\alpha = 0{,}05$
b) $n = 22$; $s^2 = 400$; $\alpha = 0{,}01$
c) $n = 25$; $s = 50$; $\alpha = 0{,}10$

Ejercicios aplicados

8.23. Al Fiedler, director de planta de LDS Vacuum Products de Altamonte Springs (Florida), aplica la teoría estadística en su centro de trabajo. LDS, importante proveedor de los fabricantes de automóviles, quiere estar seguro de que la tasa de incidencia de fugas (en centímetros cúbicos por segundo) de los enfriadores del aceite de la transmisión (TOC) satisface los límites de especificación establecidos. Se comprueba una muestra aleatoria de 50 TOC y se anotan las tasas de incidencia de fugas en el fichero llamado **TOC** (véase la referencia bibliográfica 3).

a) Estime con una confianza del 95 por ciento la tasa media de fugas de este producto.
b) Estime con una confianza del 98 por ciento la tasa media de fugas de este producto.

8.24. Está estudiándose una empaquetadora de cajas de cereales azucarados de 18 onzas (510 gramos). El peso de una muestra aleatoria de 100 cajas de cereales empaquetadas por esta máquina se encuentra en el fichero de datos **Sugar**.

a) Halle el intervalo de confianza al 90 por ciento de la media poblacional del peso de los cereales.
b) Indique sin hacer los cálculos si el intervalo de confianza al 80 por ciento de la media poblacional sería mayor, menor o igual que la respuesta del apartado (a).

8.25. Una tienda de ropa tiene interés en saber cuánto gastan los estudiantes universitarios en ropa durante el primer mes del año escolar. El gasto medio de una muestra aleatoria de nueve estudiantes es de 157,82 $ y la desviación típica muestral es de 38,89 $. Suponiendo que la población sigue una distribución normal, halle el margen de error del intervalo de confianza al 95 por ciento de la media poblacional.

8.26. Preocupa la velocidad a la que se conduce en un determinado tramo de una autopista. El radar indica la siguiente velocidad de una muestra aleatoria de siete automóviles en kilómetros por hora:

79 73 68 77 86 71 69

Suponiendo que la población sigue una distribución normal, halle el margen de error del intervalo de confianza al 95 por ciento de la velocidad media de todos los automóviles que circulan por este tramo de la autopista.

8.27. Una clínica ofrece un programa de adelgazamiento. Según sus historiales, una muestra aleatoria de 10 pacientes había experimentado las siguientes pérdidas de peso en kilos al término del programa:

18 25 6 11 15 20 16 19 12 17

a) Halle el intervalo de confianza de la media poblacional al 99 por ciento.
b) Explique sin realizar los cálculos si el intervalo de confianza de la media poblacional al 90 por ciento sería mayor, menor o igual que el obtenido en el apartado (a).

8.28. El director de la oficina de colocación de una escuela de administración de empresas quiere estimar los sueldos anuales medios que perciben los licenciados cinco años después. Una muestra aleatoria de 25 licenciados tenía una media muestral de 42.740 $ y una desviación típica muestral de 4.780 $. Halle el intervalo de confianza de la media poblacional al 90 por ciento, suponiendo que la población sigue una distribución normal.

8.29. Una empresa de alquiler de automóviles tiene interés en saber cuánto tiempo permanecen sus vehículos en el taller de reparaciones. Formule todos los supuestos y halle el intervalo de confianza al 90 por ciento del número anual medio de días que todos los vehículos de la flota de la empresa permanecen en el taller de reparaciones si una muestra aleatoria de nueve automóviles mostró el siguiente número de días que había permanecido cada uno en el taller de reparaciones:

16 10 21 22 8 17 19 14 19

8.4. Intervalos de confianza de proporciones de la población (grandes muestras)

¿Qué porcentaje de rumanos son partidarios de la entrada de su país en la Unión Europea? ¿Piensan las autoridades académicas que las notas de selectividad son un buen indicador del éxito académico en la universidad? ¿A qué proporción de los estudiantes de una uni-

versidad le gustaría que hubiera clase los sábados? En cada uno de estos casos, interesa la proporción de miembros de la población que posee una característica específica. Si se toma una muestra aleatoria de la población, la proporción muestral constituye un estimador puntual natural de la proporción de la población. En este apartado, se desarrollan intervalos de confianza para la proporción de la población.

Utilizando el modelo binomial, sea $\hat{p}$ la proporción de «éxitos» en n pruebas independientes, cada una de las cuales tiene una probabilidad de éxito P. Ya hemos visto en este libro que, si el número n de miembros de la muestra es grande, la distribución de la variable aleatoria

$$Z = \frac{\hat{p} - P}{\sqrt{\dfrac{P(1-P)}{n}}}$$

es aproximadamente una distribución normal estándar. Si el tamaño de la muestra es lo suficientemente grande para que $(n)(P)(1 - P) > 9$, se obtiene una buena aproximación si se sustituye P por el estimador puntual $\hat{p}$ en el denominador; es decir,

$$\sqrt{\frac{P(1-P)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Por lo tanto, cuando el tamaño de la muestra es grande, la distribución de la variable aleatoria

$$Z = \frac{\hat{p} - P}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

es aproximadamente una distribución normal estándar. Ahora puede utilizarse este resultado para obtener intervalos de confianza de la proporción de la población. Se obtienen de manera parecida a los ejemplos anteriores.

$$\begin{aligned}
1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} < \frac{\hat{p} - P}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \hat{p} - P < z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \\
&= P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)
\end{aligned}$$

Por lo tanto, si la proporción muestral observada es $\hat{p}$, se obtiene un intervalo de confianza aproximado de la proporción de la población al $100(1 - \alpha)\%$ por medio de la ecuación 8.8 siguiente.

Intervalos de confianza de la proporción de la población (grandes muestras)

Sea $\hat{p}$ la proporción observada de «éxitos» en una muestra aleatoria de n observaciones procedentes de una población que tiene una proporción de éxitos P. En ese caso, si n es lo suficientemente grande para que $(n)(P)(1 - P) > 9$, el **intervalo de confianza al 100(1 − α)% de la proporción de la población** viene dado por:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \textbf{(8.8)}$$

o, lo que es lo mismo,

$$\hat{p} \pm ME$$

donde ME, el **margen de error**, es

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \textbf{(8.9)}$$

Las investigaciones recientes sugieren que existen otros intervalos alternativos al intervalo de confianza de la ecuación 8.8. Esos intervalos ajustados son útiles tanto con muestras grandes como con muestras pequeñas (véanse las referencias bibliográficas 1 y 2). Se recomiendan estas lecturas para estudios más avanzados.

Los intervalos de confianza de la proporción de la población están centrados en la proporción muestral. Puede observarse también que, manteniéndose todo lo demás constante, cuanto mayor es el tamaño de la muestra, n, menor es la amplitud del intervalo de confianza, debido a que la información sobre la proporción poblacional obtenida es más precisa a medida que es mayor el tamaño de la muestra.

EJEMPLO 8.6. Plan de pluses modificado (intervalo de confianza)

La dirección quiere una estimación de la proporción de los empleados de la empresa que es partidaria de un plan de pluses modificado. Se ha observado que en una muestra aleatoria de 344 empleados, 261 están a favor de este plan. Halle una estimación del intervalo de confianza al 90 por ciento de la verdadera proporción de la población que es partidaria de este plan modificado.

Solución

Si P representa la verdadera proporción de la población y $\hat{p}$ la proporción muestral, los intervalos de confianza de la proporción de la población se obtienen por medio de la ecuación 8.8:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde, en el caso de un intervalo de confianza al 90 por ciento, $\alpha = 0{,}10$, por lo que a partir de la distribución normal estándar,

$$\alpha/2 = 0{,}05 \quad \text{y} \quad z_{\alpha/2} = z_{0{,}05} = 1{,}645$$

Se deduce que

$$n = 344 \qquad \hat{p} = 261/344 = 0{,}759$$

y

$$z_{\alpha/2} = 1{,}645$$

Por lo tanto, el intervalo de confianza de la proporción de la población al 90 por ciento es

$$0{,}759 - 1{,}645\sqrt{\frac{(0{,}759)(0{,}241)}{344}} < P < 0{,}759 + 1{,}645\sqrt{\frac{(0{,}759)(0{,}241)}{344}}$$

o sea, $0{,}721 < P < 0{,}797$. En rigor, ¿qué implican estos números? Podríamos decir que a largo plazo alrededor del 76 por ciento (con un margen de error del 4 por ciento al nivel de confianza del 90 por ciento) de la población de todos los empleados de esta empresa es partidario del plan modificado.

En las Figuras 8.12 y 8.13 comparamos los intervalos de confianza al 90 y el 99 por ciento, respectivamente.

Confidence Interval for Proportion

Sample	X	N	Sample **p**	**90.0% CI**	Z-Value	P-Value
1	261	344	0.758721	**(0.720776, 0.796665)**	9.60	0.000

Figura 8.12. Plan de pluses modificado del ejemplo 8.6: 90 por ciento (salida Minitab).

Confidence Interval for Proportion

Sample	X	N	Sample p	**99% CI**	Z-Value	P-Value
1	261	344	0.758721	**(0.699300, 0.818142)**	9.60	0.000

Figura 8.13. Plan de pluses modificado del ejemplo 8.6: 90 por ciento (salida Minitab).

Vemos que, aumentando el nivel de confianza del 90 al 99 por ciento, también aumenta el margen de error (y la amplitud). El intervalo de confianza al 90 por ciento va aproximadamente de 72,1 a 79,7 por ciento, lo que da un margen de error del 3,8 por ciento, mientras que el intervalo de confianza al 99 por ciento va aproximadamente de 69,9 a 81,8 por ciento, lo que da un margen de error del 5,95 por ciento. Cuando más amplios son los intervalos, dada α, mayor es la imprecisión con que conocemos la proporción poblacional. Se pueden obtener intervalos de confianza más reducidos tomando muestras mayores.

EJERCICIOS

Ejercicios básicos

8.30. Halle el error típico de la proporción en los casos siguientes:

a) $n = 250$; $\hat{p} = 0{,}3$
b) $n = 175$; $\hat{p} = 0{,}45$
c) $n = 400$; $\hat{p} = 0{,}05$

8.31. Halle el margen de error en los casos siguientes:

a) $n = 250$; $\hat{p} = 0{,}3$; $\alpha = 0{,}05$
b) $n = 175$; $\hat{p} = 0{,}45$; $\alpha = 0{,}08$
c) $n = 400$; $\hat{p} = 0{,}05$; $\alpha = 0{,}04$

8.32. Halle el intervalo de confianza para estimar la proporción de la población en los casos siguientes:

a) un nivel de confianza del 92,5 por ciento; $n = 650$; $\hat{p} = 0{,}10$
b) un nivel de confianza del 99 por ciento; $n = 140$; $\hat{p} = 0{,}01$
c) $\alpha = 0{,}09$; $n = 365$; $\hat{p} = 0{,}50$

Ejercicios aplicados

8.33. Suponga que se preguntó a una muestra aleatoria de 142 responsables de las admisiones en programas de postgrado qué papel desempeñan las calificaciones obtenidas en exámenes normalizados en la consideración de un candidato. En esta muestra, 87 miembros respondieron «un papel muy importante». Halle el intervalo de confianza al 95 por ciento de la proporción poblacional de responsables que tienen esta opinión.

8.34. En una muestra aleatoria de 95 empresas manufactureras, 67 han indicado que su empresa ha obtenido la certificación ISO en los dos últimos años. Halle el intervalo de confianza al 99 por ciento de la proporción poblacional de empresas que han recibido la certificación en los dos últimos años.

8.35. En un estudio reciente de una biblioteca universitaria, se preguntó a los estudiantes si pensaban que la biblioteca tenía una colección suficiente de libros. Los resultados de la encuesta se encuentran en el fichero de datos llamado **Library**.

a) Halle una estimación puntual insesgada de la proporción de estudiantes que piensa que la colección es suficiente (las respuestas se han codificado de tal forma que 1 significa «sí» y 2, «no»).

b) Halle el intervalo de confianza al 90 por ciento de la proporción de estudiantes que piensan que la colección de libros de la biblioteca es suficiente.

8.36. La escuela de administración de empresas de la Universidad de Michigan publica cuatro veces al año el índice estadounidense de satisfacción de los clientes (ACSI) (véase la referencia bibliográfica 1). Desde 1994 y basándose en miles de encuestas a clientes, se recogen índices de satisfacción de los clientes basados en una escala de 0 a 100 en comercios minoristas, supermercados, servicios financieros, servicios de mensajería, compañías aéreas, etc. «Comercios con escasez de personal, dependientes desinformados, líneas telefónicas automatizadas que van remitiendo unas a otras» son algunas de las razones por las que la puntuación de la mayoría de las empresas bajó entre 1995 y 2000 (véase la referencia bibliográfica 8). Preocupado por este informe, el director de una tienda local de una cadena minorista nacional encuestó a una muestra aleatoria de 320 clientes. La encuesta indicó que 80 clientes pensaban que el servicio de atención al cliente también estaba empeorando en esta tienda. ¿Qué conclusiones pueden extraerse de estos datos? Indique el nivel de confianza.

8.37. En una muestra aleatoria de 400 posibles votantes de una ciudad, 320 indicaron que en las siguientes elecciones votarían a favor de una política propuesta.

a) Calcule el *LIC* de una estimación del intervalo de confianza al 98 por ciento de la proporción de la población que está a favor de esta política.

b) Calcule la amplitud de la estimación del intervalo de confianza al 90 por ciento de la proporción de la población que está a favor de esta política.

8.38. En una muestra aleatoria de 198 estudiantes de marketing, 98 consideraron que no era ético inflar las calificaciones. Basándose en esta información (véase la referencia bibliográfica 2), un estadístico calculó el intervalo de confianza de la proporción poblacional que iba de 0,445 a 0,545. ¿Cuál es el nivel de confianza de este intervalo?

8.39. En un año de elecciones presidenciales, los candidatos quieren saber qué votarán los votantes de diferentes partes del país. Suponga que se pregunta a 420 posibles votantes del noreste si votarían a un determinado candidato si las elecciones fueran hoy. En esta muestra, 223 indicaron que votarían a favor de este candidato. ¿Cuál es el margen de error? Halle la estimación del intervalo de confianza al 95 por ciento del apoyo con que cuenta este candidato en el noreste.

8.39. Suponga que las autoridades sanitarias creen que este año la epidemia de gripe será menor que durante el mismo periodo del año pasado. Se ha preguntado a los residentes de una zona metropolitana si esta noticia los disuadiría de vacunarse contra la gripe. Si sólo 40 personas de una muestra aleatoria de 246 declararan que ahora no se vacunarían, estime con una confianza del 98 por ciento la proporción de todos los residentes de la zona metropolitana que ahora consideran innecesario vacunarse contra la gripe.

8.41. Es importante que las compañías aéreas respeten las horas programadas de salida de los vuelos. Suponga que una compañía ha examinado recientemente las horas de salida de una muestra aleatoria de 246 vuelos y ha observado que 10 vuelos se retrasaron debido al mal tiempo, 4 por razones de mantenimiento y el resto salió a su hora.

a) Estime el porcentaje de vuelos que salieron a su hora utilizando un nivel de confianza del 98 por ciento.

b) Estime el porcentaje de vuelos que se retrasaron debido al mal tiempo utilizando un nivel de confianza del 98 por ciento.

RESUMEN

En este capítulo hemos hecho hincapié en los estimadores y en los intervalos de confianza. Hemos analizado tres propiedades de los estimadores, a saber, la ausencia de sesgo, la consistencia y la eficiencia. Tanto la media muestral como la varianza muestral son estimadores insesgados, consistentes y eficientes de la media poblacional y de la varianza poblacional, respectivamente. Hemos desarrollado estimaciones de intervalos de confianza de parámetros como (1) la media poblacional de una población que sigue una distribución normal cuando la varianza poblacional se conoce o no se conoce y (2) la proporción poblacional con grandes muestras. Generalmente, sumando y restando el error de muestreo del estimador puntual se obtienen intervalos de confianza. Sin embargo, en el Capítulo 9 veremos que no ocurre así en el caso de la varianza poblacional. En este capítulo hemos utilizado dos tablas, la tabla de la Z normal estándar y la tabla de la t de Student para desarrollar los intervalos de confianza.

TÉRMINOS CLAVE

eficiencia relativa, 299
error de muestreo, 305
estimación, 296
estimación puntual, 296
estimador, 296
estimador consistente, 298
estimador eficiente, 298
estimador insesgado, 297
estimador insesgado
de varianza mínima, 299
estimador más eficiente, 299
estimador puntual, 296
factor de fiabilidad, 305
intervalo de confianza, 303
estimación, 303
estimador, 303
de la media, cuando la varianza es conocida, 305
de la media, cuando la varianza, es desconocida, 312
de la proporción, 317
límite inferior de confianza, 305
límite superior de confianza, 305
amplitud, 305
margen de error, 305
nivel de confianza, 303
sesgo, 298
t de Student, 310

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

8.42. Existen varios medicamentos para tratar la diabetes. Un experto en ventas de una importante compañía farmacéutica necesita una estimación del número de nuevas prescripciones de su nuevo medicamento contra la diabetes que se hicieron durante un determinado mes. El número de nuevas prescripciones en una muestra de 10 distritos de ventas es

210	240	190	275	290
265	312	284	261	243

a) Halle el intervalo de confianza al 90 por ciento del número medio de prescripciones de este nuevo medicamento en todos los distritos de ventas. Indique los supuestos.
b) Calcule la amplitud de los intervalos de confianza al 95 y el 98 por ciento.

8.43. Suponga que Braulio Mateo, directivo de la Compañía Lechera Occidental, quiere estimar el número medio de litros de leche que se venden en un día representativo. Braulio comprobó los datos de ventas de una muestra aleatoria de 16 días y observó que el número medio de litros vendidos es de 150 litros al día; la desviación típica muestral es de 12 litros. Estime con una confianza del 95 por ciento el número de litros que debería tener diariamente en existencias.

8.44. Todo el mundo sabe que el ejercicio físico es importante. Recientemente, se ha encuestado y se ha preguntado a los residentes de una comunidad cuántos minutos dedican diariamente a hacer algún tipo de ejercicio riguroso. En una muestra aleatoria de 50 residentes, el tiempo medio dedicado diariamente a hacer algún tipo de ejercicio riguroso era de media hora. Se observó que la desviación típica era de 4,2 minutos. Halle una estimación del intervalo al 90 por ciento del tiempo que dedican diariamente estos residentes a hacer algún tipo de ejercicio riguroso.

8.45. Los datos siguientes representan el número de pasajeros por vuelo de una muestra aleatoria de 50 vuelos entre Amsterdam y Viena en una compañía aérea:

163	165	094	137	123	095	170	096	117	129
152	138	147	119	166	125	148	180	152	149
167	120	129	159	150	119	113	147	169	151
116	150	110	110	143	090	134	145	156	165
174	133	128	100	086	148	139	150	145	100

Estime el número medio de pasajeros por vuelo, así como el intervalo de confianza al 95 por ciento.

8.46. El supervisor de una planta embotelladora de botellas de plástico extrajo una muestra aleatoria para averiguar si estaba presente alguno de los defectos siguientes: abolladuras, falta de etiquetado, etiquetado incorrecto y color erróneo. Los tipos de defectos se encuentran en el fichero de datos **Defects**.

a) Estime la proporción de defectos que se deben a un etiquetado incorrecto. Utilice un riesgo del 5 por ciento.

b) Estime el porcentaje de defectos que se deben a la falta de etiquetado. Utilice un intervalo de confianza al 90 por ciento.

8.47. Se han comprobado ocho lotes de un producto químico seleccionados aleatoriamente para averiguar la concentración de impurezas. Los niveles porcentuales de impurezas encontrados en esta muestra son

3,2 4,3 2,1 2,8 3,2 3,6 4,0 3,8

a) Halle las estimaciones más eficientes de la media y la varianza poblacionales.

b) Estime la proporción de lotes que tienen unos niveles de impurezas de más del 3,75 por ciento.

8.48. Un ayudante de estudios de mercado de un hospital veterinario encuestó a una muestra aleatoria de 457 propietarios de animales domésticos. Les pidió que indicaran el número de veces que van al veterinario al año. La media muestral de las respuestas fue de 3,59 y la desviación típica muestral fue de 1,045. Basándose en estos resultados, se calculó el intervalo de confianza de la media poblacional de 3,49 a 3,69. Halle la probabilidad que corresponde a este intervalo.

8.49. Se ha preguntado a una muestra aleatoria de 174 estudiantes universitarios por el número de horas semanales que navegan por Internet en busca de información personal o de material para realizar los trabajos de curso. La media muestral de las respuestas es de 6,06 horas y la desviación típica muestral es de 1,43 horas. Basándose en estos resultados, se ha calculado el intervalo de confianza de la media poblacional que va de 5,96 a 6,16. Halle el nivel de confianza de este intervalo.

8.50. Una muestra de 33 estudiantes de contabilidad anotó el número de horas que dedicaban a estudiar un examen final. Los datos se encuentran en el fichero de datos **Study**.

a) Ponga un ejemplo de estimador insesgado, consistente y eficiente de la media poblacional.

b) Halle el error de muestreo correspondiente a una estimación del número medio de horas dedicadas a estudiar este examen con un intervalo de confianza al 95 por ciento.

8.51. El doctor Miguel Savedra quiere estimar la duración media de una estancia hospitalaria (el número de días) de los pacientes que padecen una determinada enfermedad contagiosa. En una muestra aleatoria de 25 historiales de pacientes observa que el número medio de días que permanecen esos pacientes en el hospital es de 6 días con una desviación típica de 1,8 días.

a) Halle el factor de fiabilidad de una estimación de la media poblacional de la duración de la estancia con un intervalo de confianza al 95 por ciento.

b) Halle el *LIC* de una estimación de la media poblacional de la duración de la estancia con un intervalo de confianza al 99 por ciento.

8.52. Suponga que se les preguntó a los aficionados a la carrera Daytona 500 de NASCAR de esta semana si era la primera vez que asistían a la carrera. En una muestra aleatoria de 250 aficionados, 100 respondieron afirmativamente.

a) Halle el error típico para estimar la proporción de la población que asistía por primera vez.

b) Halle el error de muestreo para estimar la proporción de la población que asistía por primera vez.

c) Estime la proporción de aficionados que ya habían asistido antes con un nivel de confianza del 92 por ciento.

8.53. Los datos siguientes representan el número de pasajeros por vuelo en una muestra aleatoria de 20 vuelos de Viena a Cluj-Napoca (Rumanía) con una nueva compañía aérea:

63	65	94	37	83	95	70	96	47	29
52	38	47	79	66	25	48	80	52	49

a) ¿Cuál es el factor de fiabilidad de la estimación del número medio de pasajeros por vue-

lo con un intervalo de confianza al 90 por ciento?

b) Halle el *LIC* de la estimación del número medio de pasajeros por vuelo con un intervalo de confianza al 99 por ciento.

8.54. Un grupo de estudiantes de administración de empresas realizó una encuesta en su campus universitario para averiguar la demanda estudiantil de un producto, un suplemento proteínico para los batidos de frutas (Smoothies en inglés). Uno de los primeros pasos fue extraer una muestra aleatoria de 113 estudiantes y obtener datos que pudieran ser útiles para elaborar su estrategia de marketing. Las respuestas a esta encuesta se encuentran en el fichero de datos **Smoothies**.

a) Halle una estimación de la proporción poblacional de estudiantes a los que les gustarían suplementos como proteínas, creatina o suplementos energéticos con un intervalo de confianza al 95 por ciento.

b) Estime la proporción poblacional de estudiantes que consideran que se preocupan por su salud con un nivel de confianza del 98 por ciento.

c) De los 113 encuestados, 77 indicaron que consumen batidos por la tarde. Halle con una confianza del 90 por ciento una estimación de la proporción poblacional que consume batidos por la tarde.

8.55. Se ha extraído una muestra aleatoria de 100 estudiantes de una pequeña universidad a los que se les ha realizado una serie de preguntas como su situación como su nacionalidad, la especialidad cursada, el sexo, la edad, el curso en el que están y su nota media hasta ese momento. Se les han formulado otras preguntas sobre el nivel de satisfacción con el aparcamiento del campus, las residencias del campus y los comedores del campus. Por último, se les ha preguntado si, cuando se gradúen, tienen intención de seguir estudios de postgrado en un plazo de cinco años. Estos datos se encuentran en el fichero de datos **Finstad and Lie Study**.

a) Estime la nota media de la población con un nivel de confianza del 95 por ciento.

b) Estime la proporción poblacional de estudiantes que estaban muy insatisfechos (código de respuesta 1) o moderadamente insatisfechos (código de respuesta 2) con los servicios de aparcamiento del campus. Utilice un nivel de confianza del 90 por ciento.

c) Estime la proporción poblacional de estudiantes que estaban al menos moderadamente satisfechos (códigos de respuesta 4 y 5) con el servicio de comedores del campus.

8.56. En el Capítulo 1 propusimos varias preguntas que podían ser de interés para el director de Florin's Flower Mart. Utilice los datos del fichero de datos **Florin** para responder a cada una de las siguientes preguntas propuestas en el Capítulo 1.

a) Estime la edad media de los clientes de la tienda.

b) Estime la proporción poblacional de clientes que están insatisfechos con el sistema de reparto de la tienda.

c) Estime la media poblacional de las cantidades cargadas a una tarjeta de crédito Visa.

8.57. ¿Cuál es el método más frecuente para renovar el permiso de circulación de los vehículos? Examinando una muestra aleatoria de 500 renovaciones en una provincia, la consejería de hacienda observó que 200 se realizaron por correo, 160 se pagaron en persona y el resto se pagó por Internet. Esta operación no podía realizarse por teléfono.

a) Estime la proporción poblacional que paga la renovación en persona en las oficinas de la consejería de hacienda. Utilice un nivel de confianza del 90 por ciento.

b) Estime la proporción poblacional de renovaciones por Internet. Utilice un nivel de confianza del 95 por ciento.

8.58. Considere los datos del ejercicio 8.57. Suponga que calculáramos para la proporción poblacional que paga la renovación por correo un intervalo de confianza que fuera de 0,34 a 0,46. ¿Cuál es el nivel de confianza de este intervalo?

8.59. Considere los datos del ejercicio 8.57. Se ha dicho en un periódico local que menos de un tercio (entre el 23,7 y el 32,3 por ciento) de la población prefiere pagar por Internet. ¿Cuál es el nivel de confianza de ese intervalo?

8.60. La consejería de hacienda también quiere información sobre las tarjetas de aparcamiento de minusválidos. Suponga que en una muestra de 350 transacciones relacionadas con estas tarjetas se observó que 250 se pagaron electrónicamente.

a) ¿Cuál es el margen de error de una estimación de la proporción poblacional de tarjetas pagadas electrónicamente considerando un intervalo de confianza al 99 por ciento?

b) Indique sin realizar los cálculos si es el margen de error de una estimación similar a la anterior pero con un nivel de confianza del

95 por ciento es mayor, menor o igual que el obtenido en el apartado (a) en el que el nivel de confianza era del 99 por ciento.

8.61. ¿Cuál es la edad representativa de una persona que renueva su carné de conducir por Internet? En una muestra aleatoria de 460 renovaciones del carné de conducir, la edad media era de 42,6 y la desviación típica era de 5,4. Calcule la estimación de la edad media de los conductores que renuevan el carné de conducir por Internet con un intervalo de confianza al 98 por ciento.

Apéndice

Distribución *t* de Student

Gosset trató de desarrollar una distribución de probabilidad de las variables aleatorias que siguen una distribución normal que no incluyera la varianza poblacional σ^2. Para ello, tomó el cociente entre Z, una variable aleatoria normal estándar, y la raíz cuadrada de χ^2 dividida por sus grados de libertad, v. Utilizando la notación matemática,

$$t = \frac{Z}{\sqrt{\chi^2/v}}$$

$$t = \frac{(x - \mu)/\sigma}{\sqrt{s^2(n-1)/\sigma^2(n-1)}} = \frac{(x - \mu)}{s}$$

El estadístico t resultante tiene $n - 1$ grados de libertad. Obsérvese que la distribución de probabilidad de la t se basa en variables aleatorias que siguen una distribución normal. En las aplicaciones, se utiliza la normal Z cuando se dispone de la varianza poblacional σ^2 y se utiliza la t de Student cuando sólo se dispone de la varianza muestral s^2. Las investigaciones estadísticas que utilizan muestras aleatorias generadas por computador han demostrado que puede utilizarse la t para estudiar la distribución de medias muestrales aunque la distribución de las variables aleatorias no sea normal.

Bibliografía

1. *American Customer Satisfaction Index*, Ann Arbor, University of Michigan Business School, 2000.
2. Dabholkar, P. A. y J. J. Kellaris, «Toward Understanding Marketing Students' Ethical Judgment of Controversial Personal Selling Practices», *Journal of Business Research*, 24, 1992, págs. 313-329.
3. Fiedler, Alfred W., director de planta, «Machine Reading Leak Rate Repeatability Studies Conducted at LDS Vacuum Products», Altamonte Springs, FL, febrero, 1999.
4. Hildebrand, David y A. L. Ott, *Statistical Thinking for Managers*, Nueva York, Brooks/Cole, 1998.
5. Pearson, Egon Sharpe y R. L. Plackett (comps.), *Student*: *A Statistical Biography of William Sealy Gosset*, Oxford, Inglaterra, Clarendon Press, 1990.
6. Pearson, Egon Sharpe y John Wishart (comps.), *Development of Statistics*: *Student's Collected Papers*, Cambridge, 1958. Prólogo de Launce McMullen. Información facilitada a los autores por Teresa O'Donnell, encargada del archivo de Guinness (GIG), 13 de septiembre de 2000.
7. «Proyect Romanian Rescue: Headline News», octubre, 2000.
8. Wessel, Harry, «Lousy Service? Get Used to It», *Orlando Sentinel*, 24 de noviembre de 2000, pág. A1.

Capítulo 9

Estimación: otros temas

Esquema del capítulo

Introducción

En este capítulo analizamos algunos otros temas relacionados con la estimación. En el Capítulo 8 presentamos métodos basados en intervalos de confianza para estimar algunos parámetros de *una* población. En éste examinamos métodos basados en intervalos de confianza para estimar algunos parámetros de *dos* poblaciones. Un importante problema en la inferencia estadística es la comparación de *dos medias* de poblaciones que siguen una distribución normal o la comparación de *dos proporciones* de grandes poblaciones. Por ejemplo:

1. Los ejecutivos de las cadenas minoristas pueden querer estimar la diferencia entre las ventas diarias medias de dos de sus establecimientos.
2. Los fabricantes pueden querer comparar la productividad media, en unidades por hora, de los trabajadores del turno de día y del turno de noche de una planta.
3. El director de campaña de un candidato presidencial puede querer comparar el índice de popularidad de este candidato en dos regiones del país.
4. En el North American Fareston versus Tamoxifen Adjuvant Trial for Breast Cancer (véase la referencia bibliográfica 5) están comparándose las tasas de recurrencia de los carcinomas de las supervivientes al cáncer de mama que toman un nuevo medicamento, el Fareston, con las tasas de recurrencia de las supervivientes que toman el Tamoxifen.
5. Una compañía química recibe envíos de dos proveedores. Se seleccionan muestras aleatorias independientes de lotes procedentes de los dos proveedores y se comparan los niveles de impurezas de los dos lotes.

En este capítulo, también presentamos métodos para estimar la varianza de una población y hacemos una introducción a la elección del tamaño de la muestra, que analizamos más extensamente en el Capítulo 20.

9.1. Intervalos de confianza de la diferencia entre las medias de dos poblaciones normales

Para comparar medias de dos poblaciones, se extraen muestras aleatorias de las dos poblaciones. El método que empleamos para seleccionar las muestras determina el método que debemos utilizar para analizar inferencias basadas en los resultados muestrales. En este apartado presentamos dos sistemas de muestreo, uno para las muestras *dependientes* y otro para las muestras *independientes* cuando las varianzas poblacionales son *conocidas*. En el apartado 9.2 centramos la atención en los sistemas de muestreo para muestras independientes cuando no podemos suponer que las varianzas poblacionales son iguales.

Muestras dependientes

Consideramos que las muestras son *dependientes* si en los valores de una de las muestras influyen los de la otra. En este sistema, los miembros de la muestra se eligen por pares, uno de cada población, por lo que este método se conoce a menudo con el nombre de *datos pareados*.

La idea es que, aparte del factor estudiado, los miembros de estos pares deben parecerse lo más posible para poder hacer directamente la comparación que interesa. Supongamos que se quiere medir la eficacia de un curso de lectura rápida. Uno de los enfoques posibles sería anotar el número de palabras por minuto que lee una muestra de estudiantes *antes* de hacer el curso y compararlo con el número de palabras por minuto que leen esos mismos estudiantes *después* de hacer el curso. En este caso, cada par de observaciones consiste en las mediciones realizadas «antes» y «después» de la asistencia de un estudiante al curso.

A continuación, explicamos cómo se estiman los intervalos en el caso general de n pares de observaciones, representadas por (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), extraídas de poblaciones de medias μ_X y μ_Y.

Intervalos de confianza de dos medias: muestras dependientes (datos pareados)

Supongamos que tenemos una muestra aleatoria de n pares de observaciones enlazadas procedentes de distribuciones normales de medias μ_X y μ_Y. Es decir, sean $x_1, x_2, ..., x_n$ los valores de las observaciones de la población que tiene la media μ_X; e $y_1, y_2, ..., y_n$ los valores correspondientes de la población que tiene la media μ_Y. Sean $\bar{d}$ y s_d la media y la desviación típica muestrales observadas de las n diferencias $d_i = x_i - y_i$. Si se supone que la distribución poblacional de las diferencias es normal, entonces se obtiene un **intervalo de confianza al 100(1 − α)% de la diferencia entre las medias** ($\mu_d = \mu_X - \mu_Y$) de la forma siguiente:

$$\bar{d} - t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} \qquad \textbf{(9.1)}$$

o, lo que es lo mismo,

$$\bar{d} \pm ME$$

La desviación típica de las diferencias, s_d, y el margen de error, ME, son

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n_d}}$$

$$ME = t_{n-1,\alpha/2}\frac{s_d}{\sqrt{n}} \qquad \textbf{(9.2)}$$

y $t_{n-1,\,\alpha/2}$ es el número para el que

$$P(t_{n-1} > t_{n-1,\,\alpha/2}) = \frac{\alpha}{2}$$

La variable aleatoria, t_{n-1}, tiene una distribución t de Student con $(n-1)$ grados de libertad.

EJEMPLO 9.1. Estudio sobre la reducción del colesterol (intervalo de confianza)

Se realiza un estudio médico para comparar la diferencia de eficacia de dos medicamentos para reducir el nivel de colesterol. El grupo de investigación utiliza un enfoque de datos pareados para controlar la variación de la reducción que podría deberse a factores distintos del medicamento. Los miembros de cada par tienen las mismas características de edad, peso, estilo de vida y otros factores pertinentes. Se administra el medicamento X a una persona seleccionada aleatoriamente en cada par y el medicamento Y a la otra persona del par. Tras un determinado periodo de tiempo, se mide de nuevo el nivel de colesterol de cada persona. Supongamos que se selecciona de las grandes poblaciones de participantes una muestra aleatoria de ocho pares de pacientes que tienen problemas conocidos de colesterol. La Tabla 9.1 muestra el número de puntos en que se ha reducido el nivel de colesterol de cada persona, así como las diferencias, $d_i = x_i - y_i$, correspondientes a cada par. Estime con un nivel de confianza del 99 por ciento la diferencia media de eficacia entre los dos medicamentos, X e Y, para reducir el colesterol.

Tabla 9.1. Reducción del colesterol.

Par	Medicamento X	Medicamento Y	Diferencia ($d_i = x_i - y_i$)
1	29	26	3
2	32	27	5
3	31	28	3
4	32	27	5
5	32	30	2
6	29	26	3
7	31	33	−2
8	30	36	−6

Solución

A partir de la Tabla 9.1, calculamos la media muestral, $\bar{d}$, y la desviación típica muestral observada, s_d, de las diferencias de reducción del colesterol:

$$\bar{d} = 1{,}625 \quad \text{y} \quad s_d = 3{,}777$$

Vemos en la tabla de la distribución t de Student que $t_{n-1,\alpha/2} = t_{7,0,005} = 3{,}499$. Utilizamos la ecuación 9.1 y obtenemos el intervalo de confianza al 99 por ciento de la diferencia entre las medias poblacionales:

$$\bar{d} - \frac{t_{n-1,\alpha/2} s_d}{\sqrt{n}} < \mu_x - \mu_y < \bar{d} + \frac{t_{n-1,\alpha/2} s_d}{\sqrt{n}}$$

$$1{,}625 - \frac{(3{,}499)(3{,}777)}{\sqrt{8}} < \mu_x - \mu_y < 1{,}625 + \frac{(3{,}499)(3{,}777)}{\sqrt{8}}$$

$$-3{,}05 < \mu_x - \mu_y < 6{,}30$$

Como el intervalo de confianza contiene el valor de cero, podemos concluir que $\mu_x - \mu_y$ podría ser positivo, lo que sugeriría que el medicamento X es más eficaz; que $\mu_x - \mu_y$ podría ser negativo, lo que sugeriría que el medicamento Y es más eficaz; o que $\mu_x - \mu_y$ podría ser cero, lo que sugeriría que el medicamento X y el Y son igual de eficaces. Por lo tanto, no es posible saber si uno de los dos medicamentos es más eficaz para reducir el nivel de colesterol. En el apéndice del capítulo se presenta un breve análisis de los datos pareados cuando hay valores perdidos.

Muestras independientes, varianzas poblacionales conocidas

En este sistema, se extraen muestras *independientemente* de las dos poblaciones que siguen una distribución normal y tienen *varianzas poblacionales conocidas*, por lo que en la pertenencia a una de las muestras no influye la pertenencia a la otra.

Consideremos el caso en el que se extraen de las dos poblaciones de interés muestras independientes, no necesariamente del mismo tamaño. Supongamos que tenemos una muestra aleatoria de n_x observaciones procedentes de una población de media μ_X y varianza σ_X^2 y una muestra aleatoria independiente de n_y observaciones procedentes de una población de media μ_Y y varianza σ_Y^2. Sean las medias muestrales respectivas $\bar{x}$ e $\bar{y}$.

Examinemos, en primer lugar, la situación en la que las dos distribuciones poblacionales son normales y tienen varianzas conocidas. Como lo que nos interesa es la diferencia entre las dos medias poblacionales, es lógico basar una inferencia en la diferencia entre las medias muestrales correspondientes. Esta variable aleatoria tiene una media

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_x - \mu_y$$

y como las muestras son independientes,

$$\text{Var}\,(\bar{X} - \bar{Y}) = \text{Var}\,(\bar{X}) + \text{Var}\,(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$$

Puede demostrarse, además, que su distribución es normal. Se deduce, pues, que la variable aleatoria

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_x} + \dfrac{\sigma_Y^2}{n_y}}}$$

sigue una distribución normal estándar. A continuación, puede utilizarse un argumento paralelo al del Capítulo 8 para hallar el intervalo de confianza de la diferencia entre las medias poblacionales.

Intervalos de confianza de la diferencia entre medias: muestras independientes (distribuciones normales y varianzas poblacionales conocidas)

Supongamos que tenemos dos **muestras aleatorias independientes** de n_X y n_Y observaciones procedentes de poblaciones que siguen una distribución normal de medias μ_X y μ_Y y varianzas σ_X^2 y σ_Y^2. Si las medias muestrales observadas son $\bar{x}$ e $\bar{y}$, entonces obtenemos un intervalo de confianza al $100(1-\alpha)\%$ de (μ_X y μ_Y) de la forma siguiente:

$$(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}} < \mu_X - \mu_Y < (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}} \qquad \textbf{(9.3)}$$

o, lo que es lo mismo,

$$(\bar{x}-\bar{y}) \pm ME$$

donde el margen de error, *ME*, es

$$ME = z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}} \qquad \textbf{(9.4)}$$

En algunas aplicaciones, pueden utilizarse las varianzas históricas de estudios similares como las verdaderas varianzas poblacionales.

EJEMPLO 9.2. ¿Qué materia exige más tiempo de preparación de las clases? (intervalo de confianza)

Se pide a muestras aleatorias independientes de profesores de contabilidad y de profesores de sistemas de información que indiquen el número de horas que dedican a preparar cada clase. La muestra de 321 profesores de sistemas de información tiene un tiempo medio de 3,01 horas de preparación y la muestra de 94 profesores de contabilidad tiene un tiempo medio de 2,88 horas. Basándose en estudios similares anteriores, se supone que la desviación típica poblacional de los profesores de sistemas de información es 1,09 y que la desviación típica poblacional de los profesores de contabilidad es 1,01. Representando la media poblacional de los profesores de sistemas de información por medio de μ_X y la media poblacional de los profesores de contabilidad por medio de μ_Y, halle el intervalo de confianza al 95 por ciento de (μ_X y μ_Y).

Solución

Utilizamos la ecuación 9.3,

$$(\bar{x}-\bar{y}) - z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}} < \mu_x - \mu_y < (\bar{x}-\bar{y}) + z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

siendo

$$n_x = 321 \qquad \bar{x} = 3{,}01 \qquad \sigma_x = 1{,}09$$
$$n_y = 94 \qquad \bar{y} = 2{,}88 \qquad \sigma_y = 1{,}01$$

y para obtener el intervalo de confianza al 95 por ciento

$$z_{\alpha/2} = z_{0,025} = 1,96$$

El intervalo de confianza es, pues,

$$(3,01 - 2,88) - 1,96\sqrt{\frac{(1,09)^2}{321} + \frac{(1,01)^2}{94}} < \mu_x - \mu_y < (3,01 - 2,88) + 1,96\sqrt{\frac{(1,09)^2}{321} + \frac{(1,01)^2}{94}}$$

o sea

$$-0,11 < \mu_x - \mu_y < 0,37$$

Este intervalo incluye cero, lo que indica que no existen pruebas contundentes de que las medias poblacionales sean diferentes.

EJERCICIOS

Ejercicios básicos

9.1. Una muestra aleatoria dependiente extraída de dos poblaciones que siguen una distribución normal da los siguientes resultados:

$$n = 15 \qquad \bar{d} = 25,4 \qquad s_d = 2,8$$

a) Halle el intervalo de confianza al 95 por ciento de la diferencia entre las medias de las dos poblaciones.

b) Halle el margen de error del intervalo de confianza al 95 por ciento de la diferencia entre las medias de las dos poblaciones.

9.2. Se desea hallar el intervalo de confianza de la diferencia entre las medias de dos poblaciones que siguen una distribución normal basándose en las siguientes muestras dependientes:

Antes	Después
6	8
12	14
8	9
10	13
6	7

a) Halle el margen de error a un nivel de confianza del 90 por ciento.

b) Halle el *LSC* y el *LIC* a un nivel de confianza del 90 por ciento.

c) Halle la amplitud del intervalo de confianza al 95 por ciento.

9.3. El muestreo aleatorio independiente de de dos poblaciones que siguen una distribución normal da los siguientes resultados:

$$n_x = 64 \qquad \bar{x} = 400 \qquad \sigma_x = 20$$
$$n_y = 36 \qquad \bar{y} = 360 \qquad \sigma_y = 25$$

Halle una estimación del intervalo de confianza al 90 por ciento de la diferencia entre las medias de las dos poblaciones.

Ejercicios aplicados

9.4. Se elige una muestra aleatoria de 10 pares de viviendas idénticas de una gran ciudad y se instala un sistema pasivo de calefacción solar en uno de los miembros de cada par. Se obtienen las facturas totales de combustible (en dólares) de tres meses de invierno de estas casas que se muestran en la tabla adjunta. Suponiendo que las poblaciones siguen una distribución normal, halle el intervalo de confianza al 90 por ciento de la diferencia entre las dos medias poblacionales.

Par	Sin calefacción solar	Con calefacción solar	Par	Sin calefacción solar	Con calefacción solar
1	485	452	6	386	380
2	423	386	7	426	395
3	515	502	8	473	411
4	425	376	9	454	415
5	653	605	10	496	441

9.5. Se controla a una muestra aleatoria de seis vendedores que han asistido a un curso sobre técnicas de venta durante los tres meses anteriores y posteriores al curso. La tabla muestra los valores de las ventas (en miles de dólares) realizadas por estos seis vendedores en los dos periodos. Suponga que las distribuciones poblacionales son normales. Halle el intervalo de confianza al 80 por ciento de la diferencia entre las dos medias poblacionales.

Vendedores	Antes del curso	Después del curso
1	212	237
2	282	291
3	203	191
4	327	341
5	165	192
6	198	180

9.6. Un fabricante sabe que los números de artículos producidos por hora por la máquina A y por la máquina B siguen una distribución normal con una desviación típica de 8,4 piezas en el caso de la máquina A y una desviación típica de 11,3 piezas en el de la máquina B. La cantidad media por hora producida por la máquina A en una muestra aleatoria de 40 horas es de 130 unidades; la cantidad media por hora producida por la máquina B en una muestra aleatoria de 36 horas es de 120 unidades. Halle el intervalo de confianza al 95 por ciento de la diferencia entre los números medios de artículos producidos por hora por estas dos máquinas.

9.2. Intervalos de confianza de la diferencia entre las medias de dos poblaciones normales cuando las varianzas poblacionales son desconocidas

Parece razonable pensar que, si no conocemos las medias poblacionales, lo más probable es que tampoco conozcamos las varianzas poblacionales. Por lo tanto, en este apartado centramos la atención en esta situación más frecuente. Existen dos posibilidades: o bien se supone que las varianzas poblacionales desconocidas son iguales, o bien *no* se supone que sean iguales. Presentamos las dos situaciones, pero dejamos para el Capítulo 11 la explicación de cómo se averigua si las varianzas poblacionales son iguales.

Muestras independientes, varianzas poblacionales que se supone que son iguales

Supongamos de nuevo que tenemos dos muestras aleatorias independientes de n_x y n_y observaciones procedentes de poblaciones que siguen una distribución normal de medias μ_X y μ_Y y que las poblaciones tienen una varianza común (desconocida) σ^2, es decir, $\sigma_x^2 = \sigma_y^2 = \sigma^2$. La inferencia sobre las medias poblacionales se basa en la diferencia $(\bar{x} - \bar{y})$ entre las dos medias muestrales. Esta variable aleatoria sigue una distribución normal de media $(\mu_x - \mu_y)$ y varianza

$$\begin{aligned}\text{Var}\,(\bar{X} - \bar{Y}) &= \text{Var}\,(\bar{X}) + \text{Var}\,(\bar{Y}) \\ &= \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\end{aligned}$$

Se deduce, pues, que la variable aleatoria,

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma^2}{n_x} + \dfrac{\sigma^2}{n_y}}}$$

sigue una distribución normal estándar. Sin embargo, este resultado no puede utilizarse tal como está porque no se conoce la varianza poblacional.

Dado que $\sigma_x^2 = \sigma_y^2 = \sigma^2$, entonces tanto s_x^2 como s_y^2 son estimadores de la varianza poblacional común σ^2. Si se utilizara solamente s_x^2 o solamente s_y^2 para estimar la varianza común, no se tendría en cuenta información de la otra muestra. Si las muestras son del mismo tamaño ($n_x = n_y$), entonces podría utilizarse la media de s_x^2 y s_y^2 para estimar la varianza común. Sin embargo, en la situación más general en la que las muestras no son del mismo tamaño, se necesita una estimación que reconozca el hecho de que se obtiene más información sobre la varianza común de la muestra de mayor tamaño. Por lo tanto, se utiliza una media ponderada de s_x^2 y s_y^2. Este estimador, s_p^2, agrupa los dos conjuntos de información muestral y se obtiene mediante la ecuación 9.7.

Intervalos de confianza de dos medias: varianzas poblacionales desconocidas que se supone que son iguales

Supongamos que tenemos dos muestras aleatorias independientes de n_X y n_Y observaciones procedentes de poblaciones que siguen una distribución **normal** de medias μ_X y μ_Y y una **varianza poblacional común, pero desconocida**. Si las medias muestrales observadas son $\bar{x}$ e $\bar{y}$ y las varianzas muestrales observadas son s_x^2 y s_y^2, entonces se obtiene un intervalo de confianza al 100(1 − α)% de ($\mu_X - \mu_Y$) de la forma siguiente:

$$(\bar{x} - \bar{y}) - t_{n_x+n_y-2,\,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + t_{n_x+n_y-2,\,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} \qquad \textbf{(9.5)}$$

o, lo que es lo mismo,

$$(\bar{x} - \bar{y}) \pm ME$$

donde el **margen de error**, *ME*, es

$$ME = t_{n_x+n_y-2,\,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} \qquad \textbf{(9.6)}$$

y la **varianza muestral agrupada**, s_p^2, es

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} \qquad \textbf{(9.7)}$$

$t_{n_x+n_y-2,\,\alpha/2}$ es el número para el que

$$P(t_{n_x+n_y-2} > t_{n_x+n_y-2,\,\alpha/2}) = \frac{\alpha}{2}$$

La variable aleatoria, *t*, es aproximadamente una distribución *t* de Student con $n_x + n_{y-2}$ grados de libertad y *t* es

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

EJEMPLO 9.3. Multas de tráfico (intervalo de confianza)

Los residentes de Orange City se quejan de que las multas de tráfico por exceso de velocidad son más altas en su ciudad que las que se imponen en la vecina DeLand. Las autoridades acordaron estudiar el problema para ver si las quejas eran razonables. Se obtuvieron muestras aleatorias independientes de las multas pagadas por los residentes de cada una de las dos ciudades durante tres meses. Las cuantías de estas multas eran

Orange City	100	125	135	128	140	142	128	137	156	142
DeLand	95	87	100	75	110	105	85	95		

Suponiendo que la varianza poblacional es igual, halle el intervalo de confianza al 95 por ciento de la diferencia entre los costes medios de las multas de estas dos ciudades.

Solución

Sea X la población de Orange City e Y la población de DeLand. En primer lugar, utilizamos un paquete estadístico como Minitab y concluimos que los gráficos de probabilidad normal de ambas muestras no indican que las poblaciones no sigan una distribución normal.

$$n_x = 10 \qquad \bar{x} = 133{,}30\ \$ \qquad s_x^2 = 218{,}0111$$
$$n_y = 8 \qquad \bar{y} = 94{,}00\ \$ \qquad s_y^2 = 129{,}4286$$

Utilizando la ecuación 9.7, tenemos que la varianza muestral agrupada es

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{(10 - 1)(218{,}011) + (8 - 1)(129{,}4286)}{10 + 8 - 2} = 179{,}2562$$

y

$$(\bar{x} - \bar{y}) = (133{,}30 - 94{,}00) = 39{,}30\ \$$$

Los grados de libertad son $n_x + n_y - 2 = 16$ y $t_{(16,\,0{,}025)} = 2{,}12$.

El intervalo de confianza se obtiene por medio de la ecuación 9.5:

$$(\bar{x} - \bar{y}) - t_{n_x+n_y-2,\,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + t_{n_x+n_y-2,\,\alpha/2}\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

$$39{,}3 - (2{,}12)\sqrt{\frac{179{,}2562}{10} + \frac{179{,}2562}{8}} < \mu_x - \mu_y < 39{,}3 + (2{,}12)\sqrt{\frac{179{,}2562}{10} + \frac{179{,}2562}{8}}$$

$$39{,}3 \pm 13{,}46$$

La Figura 9.1 es la salida Minitab de este ejemplo.

A largo plazo, hay una diferencia entre el coste de las multas de Orange City y el de las multas de DeLand. El coste medio de una multa impuesta en Orange City es entre 25,84 $ y 52,76 $ más alto que el coste medio de una multa similar impuesta en DeLand.

	N	Mean	StDev	SE Mean
Orange City	10	133.3	14.8	4.7
DeLand	8	94.0	11.4	4.0

Difference = mu Orange City - mu DeLand
Estimate for difference: 39.30
95% CI for difference: (25.84, 52.76)

Figura 9.1. Multas de tráfico (salida Minitab).

Muestras independientes, varianzas poblacionales que no se supone que sean iguales

En muchas aplicaciones, no es razonable suponer que las varianzas poblacionales son iguales. En ese caso, no necesitamos una varianza muestral agrupada. Cuando las varianzas poblacionales no se conocen y no se supone que sean iguales, los grados de libertad aproximados se obtienen aplicando la ecuación 9.9 y se conocen con el nombre de aproximación de Satterthwaite (véanse las referencias bibliográficas 6 y 7). La mayoría de los paquetes estadísticos contienen ambos métodos (con y sin varianzas iguales) para hallar intervalos de confianza de las diferencias entre las medias de muestras independientes.

Intervalos de confianza de dos medias: varianzas poblacionales desconocidas, no se supone que sean iguales

Supongamos que tenemos dos **muestras aleatorias independientes** de n_X y n_Y observaciones procedentes de poblaciones que siguen una distribución **normal** de medias μ_X y μ_Y y supongamos que las varianzas poblacionales no son iguales. Si las medias y las varianzas muestrales observadas son $\bar{x}$ e $\bar{y}$ y s_x^2 y s_y^2, entonces se obtiene un intervalo de confianza al $100(1-\alpha)\%$ de $(\mu_X - \mu_Y)$ de la forma siguiente:

$$(\bar{x} - \bar{y}) - t_{(v,\,\alpha/2)}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} < \mu_x - \mu_y < (\bar{x} - \bar{y}) + t_{(v,\,\alpha/2)}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \qquad \textbf{(9.8)}$$

donde el **margen de error,** *ME*, es

$$ME = t_{(v,\,\alpha/2)}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \qquad \textbf{(9.9)}$$

y los grados de libertad, *v*, son

$$v = \frac{\left[\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2/(n_y - 1)} \qquad \textbf{(9.10)}$$

Si las muestras son del mismo tamaño, entonces los grados de libertad se reducen a

$$v = \left(1 + \frac{2}{\frac{s_x^2}{s_y^2} + \frac{s_y^2}{s_x^2}}\right) \times (n - 1) \qquad \textbf{(9.11)}$$

EJEMPLO 9.4. Auditores (intervalo de confianza)

La empresa de auditoría Master's Accounting Firm tomó una muestra aleatoria de facturas pendientes de pago de las oficinas este y oeste de Amalgamated Distributors. Quería estimar con estas dos muestras independientes la diferencia entre los valores medios de las facturas pendientes de pago. Los estadísticos muestrales obtenidos fueron los siguientes:

	Oficina Este (población *X*)	**Oficina Oeste (población *Y*)**
Media muestral	290 $	250 $
Tamaño de la muestra	16	11
Desviación típica muestral	15 $	50

No suponemos que las varianzas poblacionales desconocidas son iguales. Estime la diferencia entre los valores medios de las facturas pendientes de pago de las dos oficinas. Utilice un nivel de confianza del 95 por ciento.

Solución

Primero calculamos los grados de libertad por medio de la ecuación 9.10:

$$v = \frac{\left[\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2/(n_y - 1)} = \frac{[(225/16 + 2.500/11]^2}{\left(\frac{255}{16}\right)^2/15 + \left(\frac{2.500}{11}\right)^2/10} \approx 11$$

Ahora hallamos el margen de error utilizando la ecuación 9.9:

$$ME = t_{(v,\,\alpha/2)}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = t_{(11,\,0{,}025)}\sqrt{\frac{225}{16} + \frac{2.500}{11}} = 2{,}201(15{,}534967) = 34{,}19$$

Aplicando la ecuación 9.8, el intervalo de confianza al 95 por ciento es

$$(290 - 250) \pm 34{,}19 =$$
$$5{,}81\ \$ < \mu_x - \mu_y < 74{,}19\ \$$$

La Figura 9.2 es la salida Minitab de estos datos.

Two-Sample T-Test and CI

```
Sample      N     Mean     StDev     SE Mean
1          16    290.0      15.0         3.8
2          11    250.0      50.0          15

Difference = mu (1) - mu (2)
Estimate for difference: 40.000
95% CI for difference: (5.8078, 74.1922)
T-Test of difference = 0 (vs not =): T-Value-3.57  P-Value = 0.026  DF = 11
```

Figura 9.2. Diferencia entre las facturas pendientes de pago de las oficinas este y oeste (salida Minitab).

A largo plazo, el valor medio de las facturas pendientes de pago de la oficina este son entre 5,81 $ y 74,19 $ mayores que el valor de las facturas pendientes de pago de la oficina oeste.

EJERCICIOS

Ejercicios básicos

9.7. Suponiendo que las varianzas poblacionales son iguales, halle el número de grados de libertad en los casos siguientes:

a) $n_1 = 12, \quad s_1^2 = 30; \quad n_2 = 14, \quad s_2^2 = 36$

b) $n_1 = 6, \quad s_1^2 = 30; \quad n_2 = 7, \quad s_2^2 = 36$

c) $n_1 = 9, \quad s_1^2 = 16; \quad n_2 = 12, \quad s_2^2 = 25$

9.8. Suponiendo que las varianzas poblacionales son iguales, calcule la varianza muestral agrupada, s_p^2, para los apartados (a) a (c) del ejercicio 9.7.

9.9. Suponiendo que las varianzas poblacionales no son iguales, halle el número de grados de libertad en los casos siguientes:

a) $n_1 = 12, \quad s_1^2 = 6; \quad n_2 = 14, \quad s_2^2 = 10$

b) $n_1 = 6, \quad s_1^2 = 30; \quad n_2 = 10, \quad s_2^2 = 36$

c) $n_1 = 9, \quad s_1^2 = 16; \quad n_2 = 12, \quad s_2^2 = 25$

d) $n_1 = 6, \quad s_1^2 = 30; \quad n_2 = 7, \quad s_2^2 = 36$

9.10. Halle el margen de error del intervalo de confianza al 95 por ciento de la diferencia entre las medias poblacionales en los casos siguientes (suponga que las varianzas poblacionales son iguales):

a) $n_1 = 12, \quad s_1^2 = 6; \quad \bar{x}_1 = 200$
$n_2 = 14, \quad s_2^2 = 10; \quad \bar{x}_2 = 160$

b) $n_1 = 6, \quad s_1^2 = 6; \quad \bar{x}_1 = 200$
$n_2 = 7, \quad s_2^2 = 10; \quad \bar{x}_2 = 160$

c) Los tamaños de las muestras del apartado (a) son el doble de los del (b). Comente sus respuestas al apartado (a) en comparación con sus respuestas al apartado (b).

Ejercicios aplicados

9.11. Se observa que en una muestra aleatoria de seis estudiantes de un curso de introducción a la economía financiera que utiliza técnicas de aprendizaje de grupo la calificación media es de 76,12 y la desviación típica muestral es de 2,53. En una muestra aleatoria independiente de nueve estudiantes de otro curso de introducción a la economía financiera que no utiliza técnicas de aprendizaje de grupo, la media y la desviación típica muestrales de las calificaciones de los exámenes son 74,61 y 8,61, respectivamente. Estime con una confianza del 95 por ciento la diferencia entre las dos calificaciones medias poblacionales. Suponga que las varianzas poblacionales no son iguales.

9.12. Prairie Flower Cereal Inc. es un fabricante pequeño, pero en expansión, de cereales de desayuno que sólo deben calentarse para comerlos. Gordon Thorson, próspero agricultor que cultiva cereales, creó la empresa en 1910 (véase la referencia bibliográfica 3). Se utilizan dos máquinas para empaquetar cajas de cereales de trigo azucarados de 18 onzas (510 gramos). Estime la diferencia entre los pesos medios de las cajas de este tipo de cereales empaquetados por las dos máquinas. Utilice un nivel de confianza del 95 por ciento y el fichero de datos **Sugar Coated Wheat**. Explique sus respuestas.

9.13. Se encuesta a personas recién licenciadas en administración de empresas que trabajan a tiempo completo y que declaran que su origen socioeconómico es relativamente alto o bajo. La remuneración total media de una muestra aleatoria de 16 personas de origen socioeconómico alto es de 34.500 $ y la desviación típica muestral es de 8.520 $. La remuneración total media de una muestra aleatoria independiente de 9 personas de origen socioeconómico bajo es de 31.499 $ y la desviación típica muestral es de 7.521 $. Halle el intervalo de confianza al 90 por ciento de la diferencia entre las dos medias poblacionales.

9.14. Suponga que en una muestra aleatoria de 200 empresas que revaluaron sus activos fijos, el cociente medio entre la deuda y los activos tangibles era de 0,517 y la desviación típica muestral era de 0,148. En una muestra aleatoria independiente de 400 empresas que no revaluaron sus activos fijos, el cociente medio entre la deuda y los activos tangibles era de 0,489 y la desviación típica muestral era de 0,159. Halle el intervalo de confianza al 99 por ciento de la diferencia entre las dos medias poblacionales.

9.15. Un investigador planea estimar el efecto que produce un medicamento en las puntuaciones que obtienen los sujetos humanos que realizan una tarea de coordinación psicomotriz. Administra el medicamento antes de la prueba a los miembros de una muestra aleatoria de 9 sujetos. Su puntuación media es de 9,78 y la varianza muestral es de 17,64. Utiliza una muestra aleatoria independiente de 10 sujetos como grupo de control y le administra un placebo antes de la prueba. La puntuación media de este grupo de control es de 15,10 y la varianza muestral es de 27,01. Suponiendo que las distribuciones poblacionales son normales y tienen varianzas iguales, halle el intervalo de confianza al 90 por ciento de la diferencia entre las medias poblacionales de las puntuaciones.

9.3. Intervalos de confianza de la diferencia entre dos proporciones poblacionales (grandes muestras)

En el Capítulo 8 explicamos cómo se obtienen intervalos de confianza de una proporción poblacional. A menudo interesa comparar dos proporciones poblacionales. Por ejemplo, podría interesarnos comparar la proporción de residentes de una ciudad que declaran que votarán a favor de un determinado candidato presidencial con la proporción de residentes de otra ciudad que declaran lo mismo. En este apartado, examinamos los intervalos de confianza de la diferencia entre dos proporciones poblacionales con grandes muestras independientes extraídas de estas dos poblaciones.

Supongamos que una muestra aleatoria de n_x observaciones procedentes de una población que tiene la proporción P_x de «éxitos» genera la proporción muestral $\hat{p}_x$ y que una muestra aleatoria independiente de n_y observaciones procedentes de una población que tiene la proporción P_y de «éxitos» genera la proporción muestral $\hat{p}_y$. Como lo que nos interesa es la diferencia poblacional $(P_x - P_y)$, es lógico examinar la variable aleatoria $(\hat{p}_X - \hat{p}_Y)$. Ésta tiene la media

$$E(\hat{p}_X - \hat{p}_Y) = E(\hat{p}_X) - E(\hat{p}_Y) = P_X - P_Y$$

y como las muestras se toman independientemente, la varianza

$$\begin{aligned} \text{Var}\,(\hat{p}_X - \hat{p}_Y) &= \text{Var}\,(\hat{p}_X) + \text{Var}\,(\hat{p}_Y) \\ &= \frac{P_X(1 - P_X)}{n_x} + \frac{P_Y(1 - P_Y)}{n_y} \end{aligned}$$

Además, si el tamaño de las muestras es grande, la distribución de esta variable aleatoria es aproximadamente normal, por lo que restando su media y dividiéndola por su desviación típica, obtenemos una variable aleatoria estándar normal. Además, cuando las muestras son de gran tamaño, esta aproximación sigue siendo válida cuando las proporciones poblacionales desconocidas P_x y P_y se sustituyen por las correspondientes cantidades muestrales. Por lo tanto, la variable aleatoria

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (P_x - P_y)}{\sqrt{\dfrac{\hat{p}_x(1 - \hat{p}_x)}{n_x} + \dfrac{\hat{p}_y(1 - \hat{p}_y)}{n_y}}}$$

sigue aproximadamente una distribución normal estándar. Este resultado permite hallar intervalos de confianza para la diferencia entre las dos proporciones poblacionales cuando las muestras son de gran tamaño.

Intervalos de confianza de la diferencia entre proporciones poblacionales (grandes muestras)

Sea $\hat{p}_x$ la proporción observada de éxitos en una muestra aleatoria de n_x observaciones procedentes de una población que tiene una proporción P_X de éxitos y sea $\hat{p}_y$ la proporción de éxitos observada en una muestra aleatoria independiente de n_y observaciones procedentes de una población que tiene una proporción P_Y de éxitos. En ese caso, si las muestras son de gran ta-

maño (generalmente al menos 40 observaciones en cada una), se obtiene un **intervalo de confianza al 100(1 − a)% de la diferencia entre proporciones poblacionales**, $(P_X - P_Y)$, de la forma siguiente:

$$(\hat{p}_x - \hat{p}_y) \pm ME \tag{9.12}$$

donde el **margen de error, *ME***, es

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \tag{9.13}$$

EJEMPLO 9.5. Preferencias electorales (intervalo de confianza)

Durante un año de elecciones generales, se realizan muchos pronósticos para averiguar cómo perciben los votantes a un determinado candidato. En una muestra aleatoria de 120 posibles votantes del distrito A, 107 declararon que apoyaban al candidato en cuestión. En una muestra aleatoria independiente de 141 posibles votantes del distrito B, sólo 73 declararon que apoyaban a ese candidato. Si las proporciones poblacionales respectivas se representan por medio de P_A y P_B, halle el intervalo de confianza al 95 por ciento de la diferencia poblacional, $(P_A - P_B)$.

Solución

De la información muestral se deduce que

$$n_A = 120 \quad \text{y} \quad \hat{p}_A = 107/120 = 0{,}892; \quad n_B = 141 \quad \text{y} \quad \hat{p}_B = 73/141 = 0{,}518$$

En el caso de un intervalo de confianza al 95 por ciento, $\alpha = 0{,}05$ y, por lo tanto,

$$z_{\alpha/2} = z_{0{,}025} = 1{,}96$$

El intervalo que queremos obtener es, pues,

$$(0{,}892 - 0{,}518) - 1{,}96\sqrt{\frac{(0{,}892)(0{,}108)}{120} + \frac{(0{,}518)(0{,}482)}{141}}$$

$$< P_A - P_B < (0{,}892 - 0{,}518) + 1{,}96\sqrt{\frac{(0{,}892)(0{,}108)}{120} + \frac{(0{,}518)(0{,}482)}{141}}$$

o

$$0{,}275 < P_A - P_B < 0{,}473$$

El hecho de que cero esté muy fuera de este intervalo sugiere que existe una diferencia entre las proporciones poblacionales de posibles votantes del distrito A y del distrito B que apoyan a este candidato. A largo plazo, se estima que la diferencia es nada menos que de entre un 27,5 por ciento y un 47,3 por ciento.

La Figura 9.3 es la salida Minitab del ejemplo 9.5. Los datos sugieren que hay una diferencia entre las proporciones poblacionales de posibles votantes del distrito A y del

distrito B que apoyan a este candidato presidencial. A largo plazo, alrededor del 95 por ciento de todos esos intervalos contendría el verdadero valor de la diferencia.

Sample	X	N	Sample p
1	107	120	0.891667
2	73	141	0.517730

Estimate for p(1) - p(2): 0.373936
95% CI for p(1) - p(2): (**0.274463, 0.473409**)

Figura 9.3. Preferencias electorales del ejemplo 9.5 (salida Minitab).

EJERCICIOS

Ejercicios básicos

9.16. Calcule el margen de error en los casos siguientes:

a) $n_1 = 260, \quad \hat{p}_1 = 0{,}75; \quad n_2 = 200, \quad \hat{p}_2 = 0{,}68$
b) $n_1 = 400, \quad \hat{p}_1 = 0{,}60; \quad n_2 = 500, \quad \hat{p}_2 = 0{,}68$
c) $n_1 = 500, \quad \hat{p}_1 = 0{,}20; \quad n_2 = 375, \quad \hat{p}_2 = 0{,}25$

9.17. Calcule el intervalo de confianza al 95 por ciento de la diferencia entre las proporciones poblacionales en los casos siguientes:

a) $n_1 = 370, \quad \hat{p}_1 = 0{,}65; \quad n_2 = 200, \quad \hat{p}_2 = 0{,}68$
b) $n_1 = 220, \quad \hat{p}_1 = 0{,}48; \quad n_2 = 270, \quad \hat{p}_2 = 0{,}52$
c) $n_1 = 500, \quad \hat{p}_1 = 0{,}30; \quad n_2 = 325, \quad \hat{p}_2 = 0{,}25$

Ejercicios aplicados

9.18. En una muestra aleatoria de 120 grandes minoristas, 85 utilizan la regresión como método de predicción. En una muestra aleatoria independiente de 163 pequeños minoristas, 78 utilizan la regresión como método de predicción. Halle el intervalo de confianza al 98 por ciento de la diferencia entre las dos proporciones poblacionales.

9.19. ¿Tienen los estudiantes de último año y los de primer año las mismas ideas sobre la colección de libros que hay en la biblioteca de la universidad? Utilizando el fichero de datos **Library**, estime la diferencia entre las proporciones de estudiantes de último año y de primer año que piensan que la biblioteca de la universidad tiene una colección suficiente de libros. Utilice un nivel de confianza del 90 por ciento.

9.20. «¿Iría más a la biblioteca si se ampliara su horario de apertura?» En una muestra aleatoria de 138 estudiantes de primer año, 80 declararon que irían más a la biblioteca de la universidad si se ampliara su horario. En una muestra aleatoria independiente de 96 estudiantes de segundo año, 73 respondieron que irían más si se ampliara su horario. Estime la diferencia entre las proporciones de estudiantes de primer año y de segundo año que respondieron afirmativamente a esta pregunta. Utilice un nivel de confianza del 95 por ciento.

9.21. Una muestra aleatoria de 100 hombres contenía 61 a favor de la introducción de una enmienda constitucional para reducir la tasa de crecimiento de los impuestos sobre bienes inmuebles. Una muestra aleatoria independiente de 100 mujeres contenía 54 a favor de esta enmienda. Se calculó el intervalo de confianza

$$0{,}04 < P_x - P_y < 0{,}10$$

de la diferencia entre las proporciones poblacionales. ¿Cuál es el nivel de confianza de este intervalo?

9.22. Se observó a los clientes de un supermercado y se les encuestó inmediatamente después de que colocaran un artículo en el carro. En una muestra aleatoria de 510 clientes que eligieron un producto al precio ordinario, 320 afirmaron que comprobaban el precio en el momento en el que lo elegían. En una muestra aleatoria independiente de 332 que eligieron un producto a un precio especial, 200 hicieron esta afirmación. Halle el intervalo de confianza al 90 por ciento de la diferencia entre las dos proporciones poblacionales.

9.4. Intervalos de confianza de la varianza de una distribución normal

A veces se necesitan estimaciones del intervalo de confianza de la varianza de una población. Como cabría esperar, esas estimaciones se basan en la varianza muestral.

Supongamos que de una población que sigue una distribución normal de varianza σ^2 se extrae una muestra aleatoria de n observaciones cuya varianza es s^2. La variable aleatoria

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

sigue una distribución ji-cuadrado con $(n-1)$ grados de libertad. Este resultado constituye la base para hallar intervalos de confianza de la varianza poblacional cuando se extrae una muestra de una población que sigue una distribución normal.

Para desarrollar una fórmula que permita calcular intervalos de confianza de la varianza, se necesita una notación adicional, que ilustramos en la Figura 9.4.

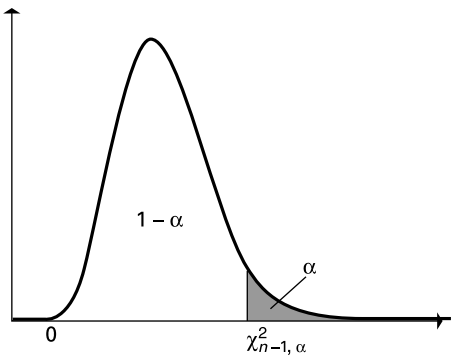


Figura 9.4. Distribución ji-cuadrado.

Notación

Una variable aleatoria que tiene la distribución ji-cuadrado con $v = n - 1$ grados de libertad se representa por medio de χ^2_v o simplemente χ^2_{n-1}. Sea $\chi^2_{n-1,\alpha}$ el número para el que

$$P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$$

Dada una probabilidad específica α, se necesita un número ji-cuadrado con $n-1$ grados de libertad, es decir, $\chi^2_{n-1,\alpha}$. Éste puede hallarse a partir de los valores de la función de distribución acumulada de una variable aleatoria ji-cuadrado. Supongamos, por ejemplo, que se necesita saber cuál es el número que es superado con una probabilidad 0,05 por una variable aleatoria ji-cuadrado con 6 grados de libertad; es decir,

$$P(\chi^2_6 > \chi^2_{6,0,05}) = 0{,}05$$

Vemos en la tabla 7 del apéndice que $\chi^2_{6,0,05} = 12{,}59$. Asimismo,

$$P(\chi^2_{n-1} > \chi^2_{n-1,\alpha/2}) = \frac{\alpha}{2}$$

Se deduce que $\chi^2_{n-1,\,1-\alpha/2}$ viene dado por

$$P(\chi^2_{n-1} > \chi^2_{n-1,\,1-\alpha/2}) = 1 - \frac{\alpha}{2}$$

y, por lo tanto,

$$P(\chi^2_{n-1} < \chi^2_{n-1,\,1-\alpha/2}) = \frac{\alpha}{2}$$

Por último,

$$P(\chi^2_{n-1,\,1-\alpha/2} < \chi^2_{n-1} < \chi^2_{n-1,\,\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

Esta probabilidad se muestra en la Figura 9.5.

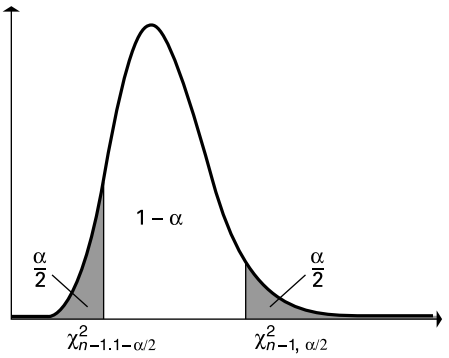


Figura 9.5. Distribución ji-cuadrado con $n - 1$ grados de libertad y un nivel de confianza de $(1 - \alpha)\%$.

Supongamos que se necesita un par de números tal que la probabilidad de que una variable aleatoria ji-cuadrado con 6 grados de libertad se encuentre entre estos números es 0,90. En ese caso, $\alpha = 0{,}10$ y

$$P(\chi^2_{6,\,0{,}95} < \chi^2_6 < \chi^2_{6,\,0{,}05}) = 0{,}90$$

Antes hemos observado que $\chi^2_{6,\,0{,}05} = 12{,}59$. En la tabla 7 del apéndice vemos que $\chi^2_{6,\,0{,}95} = 1{,}64$. La probabilidad de que esta variable aleatoria ji-cuadrado esté entre 1,64 y 12,59 es 0,90.

Para hallar intervalos de confianza de la varianza poblacional,

$$\begin{aligned} 1 - \alpha &= P(\chi^2_{n-1,\,1-\alpha/2} < \chi^2_{n-1} < \chi^2_{n-1,\,\alpha/2}) \\ &= P\left(\chi^2_{n-1,\,1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,\,\alpha/2}\right) \\ &= P\left(\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}\right) \end{aligned}$$

Intervalos de confianza de la varianza de una población normal

Supongamos que hay una muestra aleatoria de n observaciones extraídas de una población que sigue una distribución normal de varianza σ^2. Si la varianza muestral observada es s^2, entonces se obtiene un **intervalo de confianza al 100 $(1-\alpha)$% de la varianza poblacional** de la siguiente manera:

$$\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}} \tag{9.14}$$

donde $\chi^2_{n-1,\,\alpha/2}$ es el número para el que

$$P(\chi^2_{n-1} > \chi^2_{n-1,\,\alpha/2}) = \frac{\alpha}{2}$$

y $\chi^2_{n-1,\,1-\alpha/2}$ es el número para el que

$$P(\chi^2_{n-1} < \chi^2_{n-1,\,1-\alpha/2}) = \frac{\alpha}{2}$$

y la variable aleatoria χ^2_{n-1} sigue una distribución ji-cuadrado con $(n-1)$ grados de libertad.

Aunque se supone en este apartado que la población sigue una distribución normal, siempre debemos comprobar que se cumple este supuesto. Obsérvese que el intervalo de confianza de la ecuación 9.14 no tiene la forma habitual de ser el estimador puntual muestral $\pm$ margen de error.

EJEMPLO 9.6. Comparación de las varianzas de la temperatura (intervalo de confianza)

El director de Aceros Norte, S.A., quiere evaluar la variación de la temperatura en el nuevo horno eléctrico de la empresa. Se obtiene una muestra aleatoria de 25 temperaturas durante 1 semana y se observa que la varianza muestral es $s^2 = 100$. Halle el intervalo de confianza al 95 por ciento de la varianza poblacional de la temperatura.

Solución

En este ejemplo, $n = 25$ y $s^2 = 100$ y en el caso de un intervalo de confianza al 95 por ciento, $\alpha = 0{,}05$. De la Figura 9.6 basada en la tabla 7 del apéndice de la distribución ji-cuadrado se deduce que

$$\chi^2_{n-1,\,1-\alpha/2} = \chi^2_{24,\,0{,}975} = 12{,}40 \quad \text{y} \quad \chi^2_{n-1,\,\alpha/2} = \chi^2_{24,\,0{,}025} = 39{,}36$$

El intervalo de confianza al 95 por ciento de la varianza poblacional es

$$\frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}$$

Sustituyendo, tenemos que

$$\frac{(24)(100)}{39{,}36} < \sigma^2 < \frac{(24)(100)}{12{,}40} =$$

$$60{,}97 < \sigma^2 < 193{,}53$$

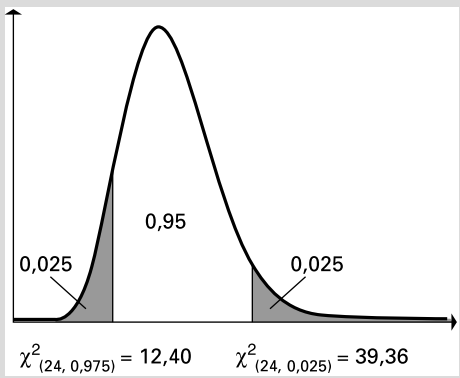


Figura 9.6. Producción diaria de cereales de trigo.

Es peligroso seguir el método que acabamos de mostrar cuando la población no sigue una distribución normal. La validez del estimador de un intervalo de la varianza depende mucho más del supuesto de la normalidad que la del estimador de un intervalo de la media poblacional.

EJERCICIOS

Ejercicios básicos

9.23. Halle el límite inferior de confianza para cada una de las siguientes poblaciones normales:

a) $n = 21$; $\alpha = 0,025$; $s^2 = 16$
b) $n = 16$; $\alpha = 0,05$; $s = 8$
c) $n = 28$; $\alpha = 0,01$; $s = 15$

9.24. Halle el límite superior de confianza para los apartados (a) a (c) del ejercicio 9.23.

9.25. Considere la siguiente muestra aleatoria extraída de una población normal:

12 16 8 10 9

a) Halle el intervalo de confianza al 90 por ciento de la varianza poblacional.
b) Halle el intervalo de confianza al 95 por ciento de la varianza poblacional.

Ejercicios aplicados

9.26. LDS quiere estar seguro de que la tasa de incidencia de fugas (en centímetros cúbicos por segundo) de los enfriadores del aceite de la transmisión (TOC) satisface los límites de especificación establecidos. Se comprueba una muestra aleatoria de 50 TOC y se anotan las tasas de incidencia de fugas en el fichero llamado **TOC** (véase la referencia bibliográfica 4). Estime la varianza de la tasa de incidencia de fugas con un nivel de confianza del 95 por ciento (compruebe la normalidad).

9.27. Una clínica ofrece un programa de adelgazamiento. Según sus historiales, una muestra aleatoria de 10 pacientes había experimentado las siguientes pérdidas de peso al término del programa:

18,2 25,9 6,3 11,8 15,4 20,3 16,8 19,5 12,3 17,2

Halle el intervalo de confianza al 90 por ciento de la varianza poblacional de las pérdidas de peso de los clientes de este programa de adelgazamiento.

9.28. El director de control de calidad de una empresa química ha extraído una muestra aleatoria de veinte sacos de fertilizante de 100 kilos para estimar la varianza de los kilos de impurezas. Se ha observado que la varianza muestral es de 6,62. Halle el intervalo de confianza al 95 por ciento de la varianza poblacional de los kilos de impurezas.

9.29. Un psicólogo quiere estimar la varianza de las puntuaciones obtenidas por los empleados en un test. Una muestra aleatoria de 18 puntuaciones tenía una desviación típica muestral de 10,4. Halle el intervalo de confianza al 90 por ciento de la varianza poblacional. ¿Cuáles son los supuestos, si los hay, para estimar este intervalo?

9.30. Un fabricante está preocupado por la variabilidad de los niveles de impurezas de los envíos de una materia prima de un proveedor. Una muestra aleatoria de 15 envíos ha mostrado una desviación típica de 2,36 en la concentración de los niveles de impurezas. Suponga que la población sigue una distribución normal.

a) Halle el intervalo de confianza al 95 por ciento de la varianza poblacional.

b) ¿Sería el intervalo de confianza al 99 por ciento de esta varianza mayor o menor que el obtenido en el apartado (a)?

9.31. Un fabricante se dedica a recubrir con plástico superficies de metal. Se toma una muestra aleatoria de nueve observaciones sobre el grosor del recubrimiento de plástico de la producción de una semana; el grosor (en milímetros) de estas observaciones es el siguiente:

19,8 21,2 18,6 20,4 21,6 19,8 19,9 20,3 20,8

Halle el intervalo de confianza al 90 por ciento de la varianza poblacional suponiendo que la población sigue una distribución normal.

9.5. Elección del tamaño de la muestra

Hemos explicado cómo se obtienen intervalos de confianza de parámetros poblacionales basándonos en la información que contiene una muestra. Después de ese proceso, puede que pensemos que el intervalo de confianza resultante es demasiado amplio, por lo que el grado de incertidumbre sobre el parámetro estimado es excesivo. Normalmente, una de las maneras de obtener un intervalo más pequeño con un nivel de confianza dado es tomar una muestra mayor.

En algunas circunstancias, podemos fijar por adelantado la amplitud del intervalo de confianza, eligiendo una muestra lo suficientemente grande para garantizar esa amplitud. En este apartado vemos cómo puede elegirse el tamaño de la muestra de esta forma para dos problemas de estimación de intervalos. Para resolver otros problemas pueden utilizarse métodos similares. En el Capítulo 20 centraremos la atención en poblaciones que no son necesariamente grandes.

Media de una población que sigue una distribución normal, varianza poblacional conocida

Si se toma una muestra aleatoria de n observaciones de una población que sigue una distribución normal de media μ y varianza conocida σ^2, en el Capítulo 8 vimos que se obtiene un intervalo de confianza al $100(1 - \alpha)\%$ de la media poblacional de la siguiente manera:

$$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

donde $\bar{x}$ es la media muestral observada y $Z_{\alpha/2}$ es el punto de corte adecuado de la distribución normal estándar. Este intervalo está centrado en la media muestral y su amplitud es B, el margen de error,

$$ME = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

a cada lado de la media muestral, de manera que ME es la mitad de la amplitud del intervalo. Supongamos ahora que el investigador quiere fijar el margen de error, ME, por adelantado. Del álgebra básica se deduce que si

$$ME = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

entonces

$$\sqrt{n} = \frac{z_{\alpha/2}\sigma}{ME}$$

Elevando al cuadrado los dos miembros de la ecuación, el tamaño de la muestra n es

$$n = \frac{z_{\alpha/2}^2\sigma^2}{ME^2}$$

Esta elección del tamaño de la muestra garantiza que la amplitud del intervalo de confianza es el doble de *ME*.

Tamaño de la muestra para estimar la media de una población que sigue una distribución normal cuando la varianza poblacional es conocida

Supongamos que se selecciona una muestra aleatoria de una población que sigue una distribución normal de varianza conocida σ^2. En ese caso, el intervalo de confianza al $100(1-\alpha)\%$ de la media poblacional tiene una amplitud *ME* (llamado a veces **error de muestreo**) a cada lado de la media muestral si el tamaño de la muestra, *n*, es

$$n = \frac{z_{\alpha/2}^2\sigma^2}{ME^2} \qquad \textbf{(9.15)}$$

Naturalmente, el número de observaciones muestrales debe ser necesariamente un entero. Si el número n resultante de la fórmula del tamaño de la muestra no es un entero, entonces debe *redondearse* al siguiente número entero para garantizar que el intervalo de confianza no es superior a la amplitud deseada.

EJEMPLO 9.7. Longitud de las barras de metal (tamaño de la muestra)

La longitud de las barras de metal producidas por un proceso industrial sigue una distribución normal que tiene una desviación típica de 1,8 milímetros. Basándose en una muestra aleatoria de nueve observaciones extraídas de esta población, se ha hallado el intervalo de confianza al 99 por ciento

$$194{,}65 < \mu < 197{,}75$$

de la media poblacional de la longitud. Supongamos que un director de producción cree que el intervalo es demasiado amplio para que tenga utilidad práctica y pide un intervalo de confianza al 99 por ciento cuya amplitud a cada lado de la media muestral no sea de más de 0,50 milímetros. ¿De qué tamaño debe ser la muestra para lograr ese intervalo?

Solución

Dado que

$$ME = 0{,}50 \qquad \sigma = 1{,}8 \qquad \text{y} \qquad z_{\alpha/2} = z_{0{,}005} = 2{,}576$$

la muestra debe tener el tamaño

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{ME^2}$$
$$= \frac{(2{,}576)^2(1{,}8)^2}{(0{,}5)^2} \approx 86$$

Por lo tanto, para satisfacer la exigencia del director, se necesita una muestra de 86 observaciones como mínimo. Este gran aumento del tamaño de la muestra representa el coste adicional de lograr una precisión mayor en la estimación del verdadero valor de la media poblacional, reflejada en un intervalo de confianza más estrecho. Se utiliza el valor 2,576, en lugar de 2,58, para hallar el tamaño de la muestra necesario.

Proporción poblacional

En el Capítulo 8 vimos que para una muestra aleatoria de n observaciones, el intervalo de confianza al $100(1 - \alpha)\%$ de la proporción poblacional P es

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < P < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

donde $\hat{p}$ es la proporción muestral observada. Este intervalo está centrado en la proporción muestral y su margen de error es:

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

a cada lado de la proporción muestral. Ahora bien, este resultado *no puede* utilizarse directamente para hallar el tamaño de la muestra necesario para obtener un intervalo de confianza de una determinada amplitud, ya que depende de la proporción muestral, que no se conoce de antemano. Sin embargo, cualquiera que sea el resultado, $\hat{p}(1 - \hat{p})$ no puede ser superior a 0,25, que es su valor cuando la proporción muestral es 0,5. Por lo tanto, el *mayor* valor que puede tener el margen de error, ME, es

$$ME = z_{\alpha/2}\sqrt{\frac{0{,}25}{n}} = \frac{(0{,}5)z_{\alpha/2}}{\sqrt{n}}$$

Supongamos, pues, que se elige una muestra lo suficientemente grande para *garantizar* que el intervalo de confianza no tiene una amplitud mayor que ME a cada lado de la proporción muestral. De nuevo, utilizando el álgebra básica, tenemos que

$$\sqrt{n} = \frac{0{,}5z_{\alpha/2}}{ME}$$

y elevando al cuadrado los dos miembros de esta igualdad, tenemos que

$$n = \frac{0{,}25(z_{\alpha/2})^2}{(ME)^2}$$

Tamaño de la muestra para estimar la proporción poblacional

Supongamos que se selecciona una muestra aleatoria de una población. Puede garantizarse entonces un intervalo de confianza al $100(1-\alpha)\%$ de la proporción poblacional, que tiene una amplitud máxima *ME* a cada lado de la proporción muestral si el tamaño de la muestra es

$$n = \frac{0{,}25(z_{\alpha/2})^2}{ME^2} \tag{9.16}$$

EJEMPLO 9.8. Personal responsable de las admisiones en programas de postgrado (tamaño de la muestra)

En el ejercicio 8.33 calculamos el intervalo de confianza al 95 por ciento de la proporción de responsables de las admisiones en programas de postgrado que pensaban que las calificaciones obtenidas en exámenes normalizados eran muy importantes en la consideración de un candidato. Basándose en 142 observaciones, se obtuvo un intervalo de

$$0{,}533 < P < 0{,}693$$

Suponga que ahora debe garantizarse que el intervalo de confianza al 95 por ciento de la proporción poblacional tiene una amplitud máxima de 0,96 a cada lado de la proporción muestral. ¿De qué tamaño debe ser la muestra?

Solución

Sabemos que

$$ME = 0{,}06 \quad \text{y} \quad z_{\alpha/2} = z_{0{,}025} = 1{,}96$$

Por lo tanto, el número de observaciones muestrales necesario es

$$n = \frac{0{,}25z_{\alpha/2}^2}{(ME)^2} = \frac{0{,}25(1{,}96)^2}{(0{,}06)^2} = 266{,}78$$

Para lograr este intervalo de confianza más estrecho, se necesita un mínimo de 267 observaciones muestrales (un aumento significativo con respecto a las 142 observaciones iniciales).

Los medios de comunicación a menudo publican los resultados de encuestas de opinión sobre cuestiones de actualidad, como el índice de aprobación del presidente en cuestiones nacionales o en política exterior o las opiniones de la gente sobre alguna propuesta fiscal. Estas encuestas generalmente representan las opiniones de algún subgrupo de la población. Normalmente, dan estimaciones del porcentaje de la población que tiene determinadas opiniones y suelen concluir con afirmaciones como «con un error de muestreo de más o menos 3 por ciento» o «la encuesta tiene un margen de error del 3 por ciento». Concretamente, estos intervalos son el porcentaje muestral, más o menos el error de muestreo o margen de error indicado. Sin embargo, debemos hacer hincapié en que el margen de error no incluye los errores que se deben a que la muestra es sesgada o es inadecuada por otras razones.

EJEMPLO 9.9. Sistema electoral (tamaño de la muestra)

Supongamos que se realiza una encuesta de opinión tras unas elecciones generales sobre las opiniones de una muestra de ciudadanos en edad de votar acerca de un cambio del sistema electoral. Se dice que la encuesta tiene un «margen de error del 3 por ciento». Eso quiere decir que el intervalo de confianza al 95 por ciento de la proporción poblacional que tiene una determinada opinión es la proporción muestral más o menos un 3 por ciento como máximo. ¿Cuántos ciudadanos en edad de votar debe tener la muestra para obtener este margen de error del 3 por ciento?

Solución

Aplicando la ecuación 9.16,

$$n = \frac{0{,}25z_{\alpha/2}^2}{(ME)^2} = \frac{(0{,}25)(1{,}96)^2}{(0{,}03)^2} = 1.067{,}111$$

Por lo tanto, la muestra debe contener 1.068 ciudadanos en edad de votar para obtener el resultado deseado.

EJERCICIOS

Ejercicios básicos

9.32. ¿De qué tamaño debe ser una muestra para estimar la media de una población que sigue una distribución normal en los casos siguientes?

a) $ME = 5$; $\sigma = 40$; $\alpha = 0{,}01$
b) $ME = 10$; $\sigma = 40$; $\alpha = 0{,}01$
c) Compare y comente las respuestas a los apartados (a) y (b).

9.33. ¿De qué tamaño debe ser una muestra para estimar la proporción poblacional en los casos siguientes?

a) $ME = 0{,}03$; $\alpha = 0{,}05$
b) $ME = 0{,}05$; $\alpha = 0{,}05$
c) Compare y comente las respuestas a los apartados (a) y (b).

9.34. ¿De qué tamaño debe ser una muestra para estimar la proporción poblacional en los casos siguientes?

a) $ME = 0{,}05$; $\alpha = 0{,}01$
b) $ME = 0{,}05$; $\alpha = 0{,}10$
c) Compare y comente las respuestas a los apartados (a) y (b).

Ejercicios aplicados

9.35. Un grupo de investigación quiere estimar la proporción de consumidores que planea comprar un escáner para su PC durante los tres próximos meses.

a) ¿De qué tamaño debe ser la muestra para que el error de muestreo sea como máximo de 0,04 con un intervalo de confianza al 90 por ciento?
b) ¿De qué tamaño debe ser la muestra si se eleva la confianza al 95 por ciento manteniendo el error de muestreo?
c) ¿De qué tamaño debe ser la muestra si el grupo de investigación amplía el error de muestreo a 0,05 y quiere un nivel de confianza del 98 por ciento?

9.36. Un político quiere estimar la proporción de electores que defienden una controvertida medida legislativa. Suponga que se necesita un intervalo de confianza al 99 por ciento que tenga una amplitud de 0,05 como máximo a cada lado de la proporción muestral. ¿Cuántas observaciones muestrales se necesitan?

9.37. La delegación de estudiantes de una universidad quiere estimar el porcentaje de estudiantes que es partidario de que se introduzca un cambio en el calendario académico de la universidad el próximo año académico. ¿Cuántos estudiantes deben encuestarse si se desea un intervalo de confianza al 90 por ciento y el margen de error debe ser de un 3 por ciento solamente?

RESUMEN

En el Capítulo 8 centramos la atención en la estimación de intervalos de confianza de parámetros basada en una población. En éste hemos centrado la atención en otros intervalos de confianza. Hemos presentado cuatro intervalos de confianza para comparar las medias de dos poblaciones que siguen una distribución normal basándonos en los siguientes sistemas de muestreo: (1) las muestras son dependientes (datos pareados); (2) las muestras son independientes y las varianzas poblacionales se conocen; (3) las muestras son independientes y las varianzas poblacionales no se conocen, pero se supone que son iguales; y (4) las muestras son independientes y las varianzas poblacionales no se conocen, pero no se supone que las varianzas sean iguales. También hemos analizado la estimación de intervalos de confianza de la diferencia entre dos proporciones poblacionales en el caso en el que las muestras son grandes, así como la estimación de intervalos de confianza de la varianza de una población que sigue una distribución normal.

Generalmente, sumando y restando el error de muestreo del estimador puntual se obtienen intervalos de confianza. Sin embargo, no ocurre así en el caso de la varianza poblacional. En este capítulo hemos utilizado tres tablas, la tabla de la Z normal estándar, la tabla de la t de Student y la tabla de la ji-cuadrado, para analizar los intervalos de confianza. Por último, hemos hecho una introducción a la elección del tamaño de la muestra para dos estimaciones de intervalos. En el Capítulo 20 analizaremos otras cuestiones relacionadas con el muestreo.

TÉRMINOS CLAVE

amplitud, 344
error de muestreo, 345
intervalo de confianza, 326
- de dos medias, independientes, 329
- de dos medias, pareados, 326
- de dos medias con varianzas que se supone que son iguales, 332
- de dos medias con varianzas que no se supone que sean iguales, 334
- de dos proporciones, 337
- de la varianza, 342

mitad de la amplitud del intervalo, 345
t de Student, 327
tamaño de la muestra para estimar la media cuando la varianza es conocida, 345
tamaño de la muestra para estimar la proporción, 347
varianza muestral agrupada, 332

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

9.38. Las muestras aleatorias independientes procedentes de dos poblaciones que siguen una distribución normal dan los siguientes resultados:

$$n_x = 15 \quad \bar{x} = 400 \quad s_x = 20$$
$$n_y = 13 \quad \bar{y} = 360 \quad s_y = 25$$

Suponga que las varianzas poblacionales desconocidas son iguales y halle el intervalo de confianza al 90 por ciento de la diferencia entre las medias poblacionales.

9.39. Las muestras aleatorias independientes procedentes de dos poblaciones que siguen una distribución normal dan los siguientes resultados:

$$n_x = 15 \quad \bar{x} = 400 \quad s_x = 10$$
$$n_y = 13 \quad \bar{y} = 360 \quad s_y = 40$$

Si no suponemos que las varianzas poblacionales desconocidas son iguales, ¿cuál es el intervalo de confianza al 90 por ciento de la diferencia entre las medias poblacionales?

9.40. Las muestras aleatorias independientes procedentes de dos poblaciones que siguen una distribución normal dan los siguientes resultados:

$$n_x = 10 \quad \bar{x} = 480 \quad s_x = 30$$
$$n_y = 12 \quad \bar{y} = 520 \quad s_y = 25$$

a) Si suponemos que las varianzas poblacionales desconocidas son iguales, ¿cuál es el intervalo de confianza al 90 por ciento de la diferencia entre las medias poblacionales?

b) Si suponemos que las varianzas poblacionales desconocidas son iguales, ¿cuál es el intervalo de confianza al 95 por ciento de la diferencia entre las medias poblacionales?

9.41. Una empresa envía una muestra aleatoria de 12 vendedores a un curso destinado a aumentar su

motivación y, por lo tanto, probablemente su eficacia. Un año más tarde, estas personas generan unas ventas que tienen un valor medio de 435.000 $ y una desviación típica muestral de 56.000 $. Durante ese mismo periodo, una muestra aleatoria elegida independientemente y formada por 15 vendedores que no asisten al curso genera unas ventas que tienen un valor medio de 408.000 $ y una desviación típica muestral de 43.000 $. Suponga que las dos distribuciones de la población son normales y tienen la misma varianza. Halle el intervalo de confianza al 95 por ciento de la diferencia entre sus medias.

9.42. Los estudiantes de un curso de introducción a la economía fueron asignados a clases de prácticas impartidas por distintos profesores ayudantes. Los 21 estudiantes de la clase de uno de los profesores ayudantes obtuvieron una calificación media de 72,1 en el examen final y una desviación típica de 11,3. Los 18 del segundo obtuvieron una calificación media en el examen final de 73,8 y una desviación típica de 10,6. Suponga que estos datos pueden considerarse muestras aleatorias independientes procedentes de poblaciones que siguen una distribución normal y tienen una varianza común. Halle el intervalo de confianza al 80 por ciento de la diferencia entre las medias poblacionales.

9.43. Existen varios medicamentos para tratar la diabetes. Un experto en ventas de una importante compañía farmacéutica tomó una muestra aleatoria de los archivos de 10 distritos de ventas para estimar el número de nuevas prescripciones del nuevo medicamento de la compañía contra la diabetes que se hicieron durante un determinado mes. El número de nuevas prescripciones era

210 240 190 275 290 265 312 284 261 243

a) Halle el intervalo de confianza al 90 por ciento del número medio de nuevas prescripciones de este nuevo medicamento que se hicieron en todos los distritos de ventas. ¿Cuáles son los supuestos?

b) Suponiendo que el nivel de confianza se mantiene constante, ¿de qué tamaño debe ser la muestra para reducir a la mitad el margen de error del intervalo de confianza del apartado (a)?

9.44. Se va a someter a votación una nueva subida de los impuestos de 1 centavo para apoyar la investigación sobre el cáncer. Se hace una encuesta a los residentes de dos ciudades para recabar su opinión. En una de ellas, una encuesta realizada recientemente a 225 residentes muestra que 140 apoyan la propuesta, 35 no saben y el resto se opone. En la ciudad vecina, según los resultados de una muestra aleatoria de 210 residentes, 120 apoyan la subida, 30 se oponen y el resto no sabe. Estime la diferencia entre los porcentajes de residentes de estas dos ciudades que apoyan esta propuesta. Utilice un nivel de confianza del 95 por ciento.

9.45. ¿Es la cantidad media que gastan cuatrimestralmente en libros de texto los estudiantes de contabilidad muy diferente de la cantidad media que gastan cuatrimestralmente en libros de texto los estudiantes de administración de empresas? Responda a esta pregunta con un intervalo de confianza al 90 por ciento utilizando los datos siguientes de muestras aleatorias de estudiantes de contabilidad o de administración de empresas. Analice los supuestos.

	Contabilidad	**Administración de empresas**
Media	340 $	285 $
Desviación típica	20 $	30 $
Tamaño de la muestra	40 $	50 $

9.46. El supervisor de una empresa embotelladora de zumo de naranja está considerando la posibilidad de comprar una nueva máquina para embotellar botellas de medio litro de zumo de naranja puro del 100 por ciento y quiere una estimación de la diferencia entre los pesos medios de las botellas que se llenan con la nueva máquina y los de las botellas que se llenan con la antigua. Se han tomado muestras aleatorias de botellas de zumo de naranja embotelladas por las dos máquinas. ¿Indican los datos siguientes que existe una diferencia entre el peso medio de las botellas llenadas con la nueva máquina y el de las botellas llenadas con la antigua? Analice los supuestos.

	Máquina nueva	**Máquina antigua**
Media	470 ml	460 ml
Desviación típica	5 ml	7 ml
Tamaño de la muestra	15	12

9.47. A Remedios Pazos, que trabaja en una gran sociedad de inversión, le gustaría estimar el porcentaje de nuevos clientes que realizarán un determinado tipo de inversión. Si quiere que el error de muestreo sea de menos de un 2,5 por ciento y que el nivel de confianza sea del 90 por ciento, ¿cuántos clientes debe tener la muestra? ¿De qué tamaño debe ser la muestra para que el nivel de confianza sea del 85 por ciento?

9.48. Una academia ofrece a los estudiantes cursos de preparación para el examen de admisión en un programa de postgrado. En un experimento para evaluar las virtudes del curso, se eligieron 12 estudiantes y se dividieron en seis pares cuyos miembros tenían parecido expediente académico. Antes de realizar el examen, se eligió aleatoriamente un miembro de cada par para que realizara el curso de preparación y el otro no realizó ningún curso. Las calificaciones obtenidas en el examen se encuentran en el fichero de datos **Student Pair**. Suponiendo que las diferencias entre las calificaciones siguen una distribución normal, halle el intervalo de confianza al 98 por ciento de la diferencia entre las calificaciones medias de los que asistieron al curso y las de los que no asistieron.

9.49. La política del gobierno en asuntos internos ha recibido un índice de aprobación del 65 por ciento en una encuesta reciente. Se ha dicho que el margen de error era de 0,035. ¿De qué tamaño era la muestra utilizada para hacer esta encuesta si suponemos que el nivel de confianza era del 95 por ciento?

9.50. Según un artículo de prensa, el 75 por ciento de 400 personas encuestadas en una ciudad se opone a una decisión judicial reciente. Según ese mismo artículo, sólo el 45 por ciento de 500 personas encuestadas en otra se opone a esa decisión. Construya el intervalo de confianza al 95 por ciento de la diferencia entre las proporciones poblacionales basándose en los datos.

Apéndice

1. La distribución *t* de Student

Gosset trató de desarrollar una distribución de probabilidad de las variables aleatorias que siguen una distribución normal que no incluyera la varianza poblacional σ^2. Para ello, tomó el cociente entre Z, una variable aleatoria normal estándar, y la raíz cuadrada de χ^2 dividida por sus grados de libertad, v. Utilizando la notación matemática,

$$t = \frac{Z}{\sqrt{\chi^2/v}}$$

$$t = \frac{(x-\mu)/\sigma}{\sqrt{s^2(n-1)/\sigma^2(n-1)}} = \frac{(x-\mu)}{s}$$

El estadístico t resultante tiene $n-1$ grados de libertad. Obsérvese que la distribución de probabilidad de la t se basa en variables aleatorias que siguen una distribución normal. En las aplicaciones, se utiliza la normal Z cuando se dispone de la varianza poblacional σ^2 y se utiliza la t de Student cuando sólo se dispone de la varianza muestral s^2. Las investigaciones estadísticas que utilizan muestras aleatorias generadas por computador han demostrado que puede utilizarse la t para estudiar la distribución de medias muestrales aunque la distribución de las variables aleatorias no sea normal.

2. Contraste de la *t* de Student para medias con varianzas poblacionales desconocidas que no se supone que sean iguales

Considerando la diferencia entre dos poblaciones, tenemos que

$$Z = \frac{(\bar{x}-\bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}} \quad \text{y} \quad \chi^2 = \chi_X^2 + \chi_Y^2$$

es la suma de dos variables aleatorias ji-cuadrado independientes extraídas de las dos muestras aleatorias independientes:

$$\chi_X^2 = \frac{(n_x - 1)s_x^2}{\sigma_X^2}$$

$$\chi_Y^2 = \frac{(n_y - 1)s_y^2}{\sigma_Y^2}$$

con $(n_x - 1)$ y $(n_y - 1)$ grados de libertad, respectivamente. Los grados de libertad de la χ^2 son la suma de los grados de libertad de los componentes, $v = (n_x - 1) + (n_y - 1) = n_x + n_y - 2$.

Reuniendo estos componentes, tenemos que

$$t = \frac{[(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)]/\sqrt{\sigma_X^2/n_x + \sigma_Y^2/n_y}}{\sqrt{[(n_x - 1)s_x^2/\sigma_x^2 + (n_y - 1)s_y^2/\sigma_y^2]/(n_x + n_y - 2)}}$$

Si $\sigma_X^2 = \sigma_Y^2$, entonces la expresión se reduce a

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$$

3. Datos pareados con valores perdidos

Consideremos el caso en que hay datos pareados con valores perdidos. Supongamos que se pierde al menos uno de los valores de la primera muestra y que hay *exactamente* el mismo número de valores perdidos en la segunda muestra (aunque no de las mismas observaciones). En este caso, los cálculos realizados con Excel darán resultados incorrectos. Primero hay que eliminar todos los casos de cualquiera de las dos muestras que contienen valores perdidos. También hay que realizar este mismo método en el Capítulo 11 cuando examinemos contrastes de hipótesis realizados con datos pareados.

Bibliografía

1. Agresti, A. y B. A. Coull, «Approximate Is Better than ''Exact'' for Interval Estimation of Binomial Proportions», *American Statistician*, 52, 1998, págs. 119-126.
2. Agresti, A. y B. Caffo, «Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures», *American Statistician*, 54, 2000, págs. 280-288.
3. Carlson, William L., *Cases in Managerial Data Analysis*, Belmont, CA, Wadsworth Publishing Company, 1997.
4. Fiedler, Alfred W., director de planta, «Machine Reading Leak Rate Repeatability Studies Conducted at LDS Vacuum Products», Altamonte Springs, FL, febrero, 1999.
5. North American Fareston versus Tamoxifen Adjuvant Trial for Breast Cancer: A Phase III Study of Tamoxifen Versus Toremifene as Adjuvant Therapy for Women with Carcinoma of the Breast, *www.naftatrial.com*, 31 de mayo de 2004.
6. Satterthwaite, F. E. (1946), «An approximate distribution of estimates of variance components», *Biometrics Bulletin*, 2, págs. 110-114.
7. Winer, B. J., *Statistical Principles in Experimental Design*, Nueva York, McGraw-Hill, 1971, 2.ª ed.

Capítulo 10

Contraste de hipótesis

Esquema del capítulo

Introducción

En este capítulo desarrollamos métodos para contrastar hipótesis que nos permiten contrastar la validez de una conjetura o de una afirmación utilizando datos muestrales. Este tipo de inferencia contrasta con los métodos de estimación desarrollados en los Capítulos 8 y 9 y los complementa. El proceso comienza cuando un investigador formula una hipótesis sobre la naturaleza de una población. La formulación de esta hipótesis implica claramente la elección entre dos opciones; a continuación, el investigador selecciona una opción basándose en los resultados de un estadístico calculado a partir de una muestra aleatoria de datos. He aquí algunos ejemplos de problemas representativos:

1. Cereales Malteados, S.A., fabricante de cereales de desayuno, sostiene que sus cajas de cereales pesan al menos 16 onzas. La empresa puede contrastar esta afirmación recogiendo una muestra aleatoria de cajas de cereales, pesando cada una y calculando el peso medio de los datos de la muestra.
2. Un fabricante de piezas de automóvil quiere verificar su proceso de producción para garantizar que el diámetro de los pistones cumple las especificaciones sobre tolerancia. Podría obtener muestras aleatorias cada 2 horas de la línea de producción y utilizarlas para averiguar si están cumpliéndose las normas.

Estos ejemplos se basan en un tema común. Formulamos una hipótesis sobre un parámetro poblacional y utilizamos datos muestrales para contrastar la validez de nuestra hipótesis.

10.1. Conceptos del contraste de hipótesis

Aquí presentamos un modelo general para contrastar hipótesis utilizando estadísticos calculados a partir de muestras aleatorias. Dado que estos estadísticos tienen una distribución en el muestreo, tomamos nuestra decisión en presencia de una cierta variación aleatoria. Por lo tanto, necesitamos unas reglas claras de decisión para elegir entre las dos opciones.

El proceso que desarrollamos aquí tiene una analogía directa con un juicio con jurado. En un juicio con jurado, suponemos que el acusado es inocente y el jurado decide que una persona es culpable sólo si existen pruebas muy contundentes en contra de la presunción de inocencia. Ese proceso para elegir entre la culpabilidad y la inocencia tiene:

1. Rigurosos procedimientos para presentar y evaluar la evidencia
2. Un juez para aplicar las reglas
3. Un proceso de decisión que supone que el acusado es inocente a menos que exista evidencia que demuestre su culpabilidad más allá de una duda razonable.

Obsérvese que este proceso no condena a algunas personas que, en realidad, son culpables. Pero si se rechaza la inocencia de una persona y se la halla culpable, tenemos la firme convicción de que es culpable.

Comenzamos el método del contraste de hipótesis considerando un valor de un parámetro de la distribución de probabilidad de una población, por ejemplo, la media, μ, la varianza, σ^2, o la proporción, P. Nuestro método empieza con una hipótesis sobre el parámetro —llamada **hipótesis nula**— que mantendremos a menos que existan pruebas contundentes en contra de ella. Si rechazamos la hipótesis nula, entonces aceptaremos la segunda hipótesis, llamada **hipótesis alternativa**. Sin embargo, si no rechazamos la hipótesis nula, no podemos concluir necesariamente que es correcta. Si no la rechazamos, o bien es correcta la hipótesis nula, o bien es correcta la hipótesis alternativa, pero nuestro método de contraste no es suficientemente fuerte para rechazar la hipótesis nula.

Utilizando nuestro ejemplo del fabricante de cereales, podríamos comenzar suponiendo que el peso medio de los paquetes es de 16 onzas, por lo que nuestra hipótesis nula es:

$$H_0 : \mu = 16$$

Una hipótesis, ya sea nula o alternativa, puede especificar un único valor —en este caso, $\mu = 16$— para el parámetro poblacional μ. Decimos que esta hipótesis es una **hipótesis simple**, que se lee de la siguiente manera: «la hipótesis nula es que el parámetro poblacional μ es igual a un valor específico de 16». En este ejemplo de los cereales, una hipótesis alternativa posible es que el peso medio de los paquetes se encuentra en el intervalo de valores superiores a 16 onzas:

$$H_1 : \mu > 16$$

Esta hipótesis alternativa se llama **hipótesis alternativa compuesta unilateral**. Otra posibilidad sería contrastar la hipótesis nula frente a la **hipótesis alternativa compuesta bilateral**:

$$H_1 : \mu \neq 16$$

Elegimos estas hipótesis de manera que una o la otra tenga que ser cierta. En este libro, representamos la hipótesis nula por medio del símbolo H_0 y la hipótesis alternativa por medio del símbolo H_1.

Al igual que ocurre en un juicio con jurado, seguimos un riguroso método para elegir una hipótesis o la otra. Utilizamos un estadístico calculado a partir de una muestra aleatoria, como una media muestral, $\bar{x}$, una varianza muestral, s^2, o una proporción muestral, $\hat{p}$. El estadístico tendrá una distribución en el muestreo conocida, basada en el método de muestreo y el valor del parámetro especificado por la hipótesis nula. A partir de esta distribución en el muestreo, hallamos los valores del estadístico que tienen una pequeña probabilidad de ocurrir si la hipótesis nula es verdadera. Si el estadístico tiene un valor que tiene una pequeña probabilidad de ocurrir cuando la hipótesis nula es verdadera, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Sin embargo, si el estadístico no tiene una pequeña probabilidad de ocurrir cuando la hipótesis nula es verdadera, no rechazaremos la hipótesis nula. La especificación de la hipótesis nula y de la hipótesis alternativa depende del problema, como indican los siguientes ejemplos.

1. Cereales Malteados quiere averiguar si el peso medio de las cajas es mayor de lo que éstas indican. Sea μ el peso medio poblacional (en onzas) de los cereales por caja. La hipótesis nula compuesta es que esta media es de 16 onzas como máximo:

$$H_0 : \mu \leqslant 16$$

y la alternativa evidente es que el peso medio es de más de 16 onzas:

$$H_1 : \mu > 16$$

En este problema, buscaríamos pruebas contundentes de que el peso medio de las cajas es de más de 16 onzas. Por ejemplo, una empresa querría evitar que se emprendieran acciones legales contra ella porque el peso de las cajas fuera bajo. Tendría confianza en su creencia si tuviera pruebas contundentes que permitieran rechazar H_0.

2. Una fábrica de pistones para automóviles ha propuesto un proceso para controlar periódicamente el diámetro de los pistones. Cada 2 horas se seleccionaría una muestra aleatoria de $n = 6$ pistones del proceso de producción y se medirían sus diámetros. Se calcularía el diámetro medio de los 6 pistones y se utilizaría para contrastar la hipótesis nula simple:

$$H_0 : \mu = 3{,}800$$

frente a la hipótesis alternativa:

$$H_1 : \mu \neq 3{,}800$$

En este caso, la empresa continuaría funcionando a menos que se rechazara la hipótesis nula en favor de la hipótesis alternativa. La existencia de pruebas contundentes de que los pistones no están cumpliendo las normas de tolerancia llevaría a interrumpir el proceso de producción.

Una vez que hemos especificado la hipótesis nula y la hipótesis alternativa y hemos recogido datos muestrales, debemos tomar una decisión sobre la hipótesis nula. Podemos rechazarla y aceptar la hipótesis alternativa o no rechazarla. Hay buenas razones por las que muchos estadísticos prefieren no decir «aceptamos la hipótesis nula» en lugar de «no rechazamos la hipótesis nula». Cuando no rechazamos la hipótesis nula, o bien ésta es verdadera, o bien nuestro método de contraste no es suficientemente fuerte para rechazarla y

hemos cometido un error. Para seleccionar la hipótesis —nula o alternativa— desarrollamos una regla de decisión basada en la evidencia muestral. Más adelante en este capítulo presentamos reglas de decisión específicas para varios problemas. En muchos casos, la forma de la regla es bastante obvia. Para contrastar la hipótesis nula de que el peso medio de las cajas de cereales es de menos de 16 onzas, obtenemos una muestra aleatoria de cajas y calculamos la media muestral. Si la media muestral es considerablemente superior a 16 onzas, podemos rechazar la hipótesis nula y aceptar la hipótesis alternativa. En general, cuanto más distante de 16 sea la media muestral, mayor será la probabilidad de rechazar la hipótesis nula. Más adelante desarrollamos reglas de decisión específicas.

En nuestro análisis de las distribuciones en el muestreo del Capítulo 7, vimos que la media muestral es diferente de la media poblacional. Con una media muestral solamente, no podemos estar seguros del valor de la media poblacional. Por lo tanto, sabemos que la regla de decisión adoptada tiene alguna probabilidad de extraer una conclusión errónea. La Tabla 10.1 resume los tipos posibles de error. El **error de Tipo I** es la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera. Definimos nuestra regla de decisión de tal forma que la probabilidad de rechazar una hipótesis nula verdadera, representada por α, es «pequeña». α es el nivel de significación del contraste. La probabilidad de no rechazar la hipótesis nula cuando es verdadera es $(1 - \alpha)$. También existe otro error posible, llamado **error de Tipo II**, que se comete cuando no se rechaza una hipótesis nula falsa. En una regla de decisión específica, la probabilidad de cometer ese error cuando la hipótesis nula es falsa se representa por medio de β. La probabilidad de rechazar una hipótesis nula falsa es $(1 - \beta)$ y se denomina potencia del contraste.

Tabla 10.1. Estados de la naturaleza y decisiones sobre la hipótesis nula, con las probabilidades de tomar las decisiones, dados los estados de la naturaleza.

Decisiones sobre la hipótesis nula	**Estados de la naturaleza**	
	La hipótesis nula es verdadera	**La hipótesis nula es falsa**
No rechazar H_0	Decisión correcta Probabilidad = $1 - \alpha$	Error de Tipo II Probabilidad = β
Rechazar H_0	Error de Tipo I Probabilidad = α (α se llama nivel de significación)	Decisión correcta Probabilidad = $1 - \beta$ ($1 - \beta$ se llama potencia del contraste)

Ilustraremos estas ideas por medio del ejemplo anterior. El director de una fábrica está tratando de averiguar si la media poblacional del peso de las cajas es mayor de lo que indican éstas. La hipótesis nula es que en la población el peso medio de las cajas es inferior o igual al de 16 onzas que indican éstas. Se contrasta esta hipótesis nula frente a la hipótesis alternativa de que el peso medio de las cajas es de más de 16 onzas. Para contrastar la hipótesis, tomamos una muestra aleatoria independiente de cajas de cereales y calculamos la media muestral. Si ésta es muy superior a 16 onzas, rechazamos la hipótesis nula. En caso contrario, no la rechazamos. Sea $\bar{x}$ la media muestral. Una regla de decisión posible es

$$\text{Rechazar } H_0 \quad \text{si } \bar{x} > 16{,}13$$

Supongamos ahora que la hipótesis nula es verdadera. Podríamos observar, aun así, que la media muestral es superior a 16,13 y, según nuestra regla de decisión, la hipótesis nula se rechazaría. En ese caso, habríamos cometido un error de Tipo I. La probabilidad de rechazo cuando la hipótesis nula es verdadera es el nivel de significación α.

Supongamos, por el contrario, que la hipótesis nula es falsa y que la media poblacional del peso de las cajas es de más de 16. Podríamos observar, aun así, que la media muestral es inferior a 16,13 y, según nuestra regla de decisión, la hipótesis nula no se rechazaría. Por lo tanto, habríamos cometido un error de Tipo II. La probabilidad de cometer ese error dependerá de la cuantía exacta en que la media poblacional sea superior a 16. Veremos que es más probable que se rechace la hipótesis nula, dado el tamaño de la muestra, si la media poblacional es 16,5 que si es 16,1.

En teoría, nos gustaría que las probabilidades de los dos tipos de error fueran lo más pequeñas posible. Sin embargo, existe una disyuntiva entre las probabilidades de los dos tipos de errores. Dada una muestra, cualquier reducción de la probabilidad de cometer un error de Tipo I, α, provocará un aumento de la probabilidad de cometer un error de Tipo II, β, y viceversa. Debemos hacer hincapié aquí en que no existe una sustitución lineal directa (por ejemplo, una reducción de α de 0,02 no provoca normalmente un aumento de β de 0,02). Por lo tanto, en el ejemplo anterior, la probabilidad de cometer un error de Tipo I, α, podría reducirse cambiando la regla de decisión por:

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} > 16{,}23$$

Pero es más probable que no se rechace la hipótesis nula, aunque sea falsa. Como consecuencia, aumentaría la probabilidad de cometer un error de Tipo II. En la práctica, seleccionamos una pequeña probabilidad de cometer un error de Tipo I (por ejemplo, de menos de 0,10) y utilizamos esa probabilidad para fijar la regla de decisión. A continuación, hallamos la probabilidad de cometer un error de Tipo II, como muestra la Figura 10.1.

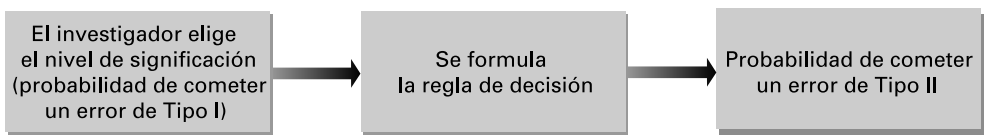


Figura 10.1. Consecuencias de la fijación del nivel de significación de un contraste.

Supongamos que el director de la fábrica quisiera averiguar si el verdadero peso medio de las cajas de cereales es de más de 16 onzas. Comenzaría el análisis fijando primero la probabilidad de cometer un error de Tipo I, lo cual es en cierto sentido como decidir las reglas de un partido de béisbol o de fútbol antes de que comience en lugar de ir estableciendo las reglas a medida que se juega el partido. Tras analizar la naturaleza del proceso e decisión, podría decidir que la regla de decisión debe tener una probabilidad de 0,05 o menos de rechazar la hipótesis nula cuando es verdadera. Lo haría seleccionando un número apropiado, K, en la regla de decisión: «rechazar la hipótesis nula si la media muestral es superior a K onzas». En los apartados siguientes explicamos el método para elegir K. Una vez elegido el número K, puede calcularse la probabilidad de cometer un error de Tipo II —para un valor de μ incluido en H_1— utilizando los métodos que desarrollamos en el apartado 10.5.

Otro concepto que se utiliza en el contraste de hipótesis es la **potencia** del contraste, que es la probabilidad de rechazar H_0 cuando H_1 es verdadera. Se calcula para valores específicos de μ que satisfacen la hipótesis nula. La potencia normalmente es diferente para cada valor de μ. Consideremos el problema de los cereales, en el que

$$H_0 : \mu = 16$$
$$H_1 : \mu > 16$$

Así, para cualquier valor de μ contenido en la hipótesis nula, H_1,

$$\text{Potencia} = P(\text{Rechazar } H_0 \mid \mu, (\mu \subset H_1))$$

Dado que la regla de decisión depende del nivel de significación elegido para el contraste, el concepto de potencia no afecta directamente a la decisión de rechazar o no rechazar una hipótesis nula. Sin embargo, calculando la potencia del contraste para niveles de significación y valores de μ específicos incluidos en H_1, tendremos valiosa información sobre las propiedades de la regla de decisión. Por ejemplo, veremos que aumentando el tamaño de la muestra, la potencia del contraste aumentará para un nivel dado de significación, α. Por lo tanto, sopesaremos el incremento de los costes que implica un aumento del tamaño de la muestra y los beneficios de aumentar la potencia del contraste. El cálculo de la potencia también es útil cuando, dado el tamaño de la muestra, podemos elegir entre dos o más contrastes que tienen los mismos niveles de significación. En ese caso, sería adecuado elegir el contraste que tenga la menor probabilidad de cometer un error de Tipo II, es decir, el contraste que tenga la mayor potencia.

En los apartados 10.2 a 10.4 mostramos cómo pueden formularse reglas de decisión, dados unos niveles de significación, para algunas clases importantes de problemas de contraste de hipótesis. En el 10.5 mostramos cómo puede calcularse la potencia de un contraste. A continuación, resumimos los términos y las ideas importantes que hemos presentado hasta ahora.

Resumen de la terminología del contraste de hipótesis

Hipótesis nula H_0: hipótesis que se mantiene que es verdadera, a menos que se obtenga suficiente evidencia en contra.

Hipótesis alternativa H_1: hipótesis frente a la que se contrasta la hipótesis nula y que se mantiene que es verdadera si se rechaza la hipótesis nula.

Hipótesis simple: hipótesis que especifica un único valor para un parámetro poblacional de interés.

Hipótesis compuesta: hipótesis que especifica un rango de valores para un parámetro poblacional.

Hipótesis alternativa unilateral: hipótesis alternativa que implica todos los valores posibles de un parámetro poblacional a un lado o al otro (es decir, mayores o menores) del valor especificado por una hipótesis nula simple.

Hipótesis alternativa bilateral: hipótesis alternativa que implica todos los valores posibles de un parámetro poblacional distintos del valor especificado por una hipótesis nula simple.

Decisiones de un contraste de hipótesis: se formula una regla de decisión que lleva al investigador a rechazar o no la hipótesis nula basándose en la evidencia muestral.

Error de Tipo I: rechazo de una hipótesis nula verdadera.

Error de Tipo II: aceptación de una hipótesis nula falsa.

Nivel de significación: probabilidad de rechazar una hipótesis nula que es verdadera. Esta probabilidad a veces se expresa en porcentaje, por lo que un contraste de nivel de significación α se denomina contraste de nivel $100\alpha\%$.

Potencia: probabilidad de rechazar una hipótesis nula que es falsa.

En los resúmenes formales de los resultados de los contrastes, utilizamos los términos *rechazar* y *no rechazar* posibles decisiones sobre una hipótesis nula. Veremos que estos términos no reflejan correctamente la asimetría de los estatus de hipótesis nula e hipótesis alternativa o las consecuencias de un método en el que el nivel de significación es fijo y la probabilidad de cometer un error de Tipo II no se controla. La hipótesis nula tiene el estatus de una hipótesis que se mantiene —que se sostiene que es verdadera— a menos que los

datos contengan pruebas contundentes para rechazarla. Fijando un bajo nivel de significación, α, tenemos una pequeña probabilidad de rechazar una hipótesis nula verdadera. Cuando la rechazamos, la probabilidad de cometer un error es el nivel de significación, α. Pero si sólo hay una pequeña muestra, rechazamos la hipótesis nula solamente cuando es totalmente errónea. A medida que aumenta el tamaño de la muestra, también aumenta la probabilidad de rechazar una hipótesis nula falsa. Pero si no se rechaza una hipótesis nula, es mucho mayor la incertidumbre, porque no sabemos cuál es la probabilidad de cometer un error de Tipo II. Por lo tanto, si no rechazamos una hipótesis nula, o bien es verdadera, o bien nuestro método para detectar una hipótesis nula falsa no tiene suficiente potencia, por ejemplo, el tamaño de la muestra es demasiado pequeño. Cuando rechazamos la hipótesis nula, tenemos pruebas contundentes de que no es verdadera y, por lo tanto, de que la hipótesis alternativa es verdadera. Si buscamos pruebas contundentes a favor de un determinado resultado, ese resultado es la hipótesis alternativa, H_1, y el otro es la hipótesis nula, H_0. Se denomina **argumento contrafactual**. Cuando rechazamos H_0, existen pruebas contundentes a favor de H_1 y estamos seguros de que nuestra decisión es correcta. Pero si no rechazamos la hipótesis nula, tenemos una gran incertidumbre. En los siguientes apartados vemos muchas aplicaciones de esta idea.

La analogía con un juicio es evidente. El acusado goza de la presunción de inocencia (la hipótesis nula) a menos que existan pruebas contundentes que indiquen que es culpable más allá de una duda razonable (rechazo de la hipótesis nula). El acusado puede ser declarado inocente bien porque lo es, bien porque las pruebas no son lo bastante poderosas para condenarlo. La carga de la prueba está en los datos muestrales.

EJERCICIOS

Ejercicios básicos

10.1. María Arnaldo quiere utilizar los resultados de un estudio de mercado basado en una muestra aleatoria para buscar pruebas contundentes de que su marca de cereales de desayuno tiene al menos un 20 por ciento de todo el mercado. Formule la hipótesis nula y la hipótesis alternativa utilizando P como proporción poblacional.

10.2. El banco central tiene que decidir si baja o no los tipos de interés para estimular el crecimiento económico. Formule la hipótesis nula y la hipótesis alternativa sobre el crecimiento económico que formularía el banco central para tomar su decisión.

10.3. Juan Estévez, vicepresidente de una empresa, está buscando pruebas contundentes que apoyen su opinión de que los nuevos métodos operativos han reducido el porcentaje de cajas de cereales que pesan menos de lo indicado. Formule la hipótesis nula y la hipótesis alternativa e indique los resultados que constituirían una prueba contundente.

Ejercicios aplicados

10.4. Durante 2000 y 2001, muchos europeos se negaron a comprar alimentos modificados genéticamente y producidos por agricultores estadounidenses. Los agricultores estadounidenses sostenían que no existía ninguna prueba científica que llevara a concluir que estos productos no eran saludables. Los europeos sostenían que, aun así, podían plantear problemas.

a) Formule la hipótesis nula y la hipótesis alternativa desde el punto de vista de los europeos.

b) Formule la hipótesis nula y la hipótesis alternativa desde el punto de vista de los agricultores estadounidenses.

10.5. El resultado de las elecciones presidenciales que se celebraron en 2000 en Estados Unidos fue muy ajustado y el resultado dependía de lo que se votara en el estado de Florida. El Tribunal Supremo de Estados Unidos declaró finalmente la victoria de George W. Bush frente a Al Gore, afirmando que no era adecuado contar a mano

los votos que habían sido rechazados por las máquinas de votar en varios condados. En ese momento, Bush tenía una pequeña ventaja basada en los votos que se habían contado. Imagine que fuera un abogado de George W. Bush. Formule su hipótesis nula y su hipótesis alternativa sobre el total de votos de cada candidato. Dadas sus hipótesis, ¿qué diría sobre los resultados del recuento propuesto si se hubiera realizado realmente?

10.2. Contrastes de la media de una distribución normal: varianza poblacional conocida

En este apartado y en los siguientes, presentamos métodos específicos para desarrollar y realizar contrastes de hipótesis que pueden aplicarse a problemas empresariales y económicos. Utilizamos una muestra aleatoria de n observaciones que siguen una distribución normal $x_1, x_2, ..., x_n$ procedentes de una población de μ y de varianza σ^2 conocida. Contrastaremos una hipótesis sobre la media poblacional desconocida. Más adelante abandonaremos en muchos casos nuestro supuesto de la normalidad debido al teorema del límite central.

En el análisis del contraste de hipótesis del apartado 10.1, hemos señalado que, si se rechaza una hipótesis nula utilizando un contraste con un nivel de significación α, se conoce la probabilidad de cometer un error. En este caso, o bien la decisión es correcta, o bien hemos cometido un error de Tipo I. Pero si no rechazamos una hipótesis nula, no sabemos cuál es la probabilidad de cometer un error. Por lo tanto, tenemos pruebas contundentes para apoyar una postura específica si elegimos la hipótesis nula y la hipótesis alternativa de tal manera que el rechazo de la hipótesis nula y la aceptación de la hipótesis alternativa llevan a apoyar nuestra postura específica. Lo demostramos en el siguiente ejemplo.

Consideremos nuestro ejemplo anterior sobre el peso de las cajas de cereales. Supongamos que las normas del sector dicen que si la media poblacional del peso de las cajas es de 16,1 onzas o menos en una población de cajas que indican que su peso es de 16 onzas, entonces se presentará una demanda contra el fabricante. Por lo tanto, nuestro objetivo es conseguir pruebas contundentes de que el peso medio de las cajas, μ, es superior a 16,1 onzas. En este caso, nuestra hipótesis nula sería

$$H_0 : \mu = \mu_0 = 16{,}1$$

y la hipótesis alternativa,

$$H_1 : \mu > \mu_0 = 16{,}1$$

Formulando nuestra regla de contraste con un nivel de significación α, sabemos que el rechazo de la hipótesis nula constituye una prueba contundente de que el peso medio es de más de 16,1 onzas, ya que la probabilidad de cometer un error tiene un valor pequeño, α.

Nuestro contraste de la media poblacional utiliza la media muestral $\bar{x}$. Si la media muestral es considerablemente superior a $\mu_0 = 16{,}1$, entonces rechazamos la hipótesis nula. Para obtener el valor de decisión adecuado, utilizamos el hecho de que la variable aleatoria estandarizada

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

sigue una distribución normal estándar de media 0 y de varianza 1, dado que H_0 es verdadera. Si α es la probabilidad de cometer un error de Tipo I y Z es grande de tal manera que

$$P(Z > z_\alpha) = \alpha$$

entonces, para contrastar la hipótesis nula, podemos utilizar la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

Se deduce que la probabilidad de rechazar la hipótesis nula, H_0, cuando es verdadera es el nivel de significación α.

Obsérvese que, mediante una sencilla manipulación algebraica, también podríamos formular la regla de decisión de la forma siguiente:

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n}$$

El valor $\bar{x}_c$ a menudo se llama **valor crítico** de la decisión. Obsérvese que para todo valor z_α procedente de la distribución normal estándar, también hay un valor $\bar{x}_c$ y cualquiera de las dos reglas de decisión anteriores da exactamente el mismo resultado.

Supongamos que en este problema la desviación típica poblacional es $\sigma = 0{,}4$ y obtenemos una muestra aleatoria de tamaño 25. Para realizar un contraste de hipótesis unilateral con un nivel de significación $\alpha = 0{,}05$, vemos en la tabla de la distribución normal estándar que el valor de z_α es 1,645. En este caso, nuestra regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 16{,}1}{0{,}4/\sqrt{25}} > 1{,}645$$

En otras palabras, la regla es

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n} = 16{,}1 + 1{,}645 \times (0{,}4/\sqrt{25}) = 16{,}232$$

Si rechazamos H_0 utilizando esta regla, aceptamos la hipótesis alternativa de que el peso medio es de más de 16 onzas con la probabilidad de cometer un error de Tipo I de 0,05 o menos. Ésta es una prueba contundente en la que apoyar nuestra conclusión. Pero el hecho de no rechazar la hipótesis nula nos lleva a concluir que o bien H_0 es verdadera, o bien el método de contraste seleccionado no es suficientemente sensible para rechazar H_0. Las reglas de decisión se muestran en la Figura 10.2. A continuación, resumimos el contraste de hipótesis para una hipótesis nula simple sobre la media poblacional.

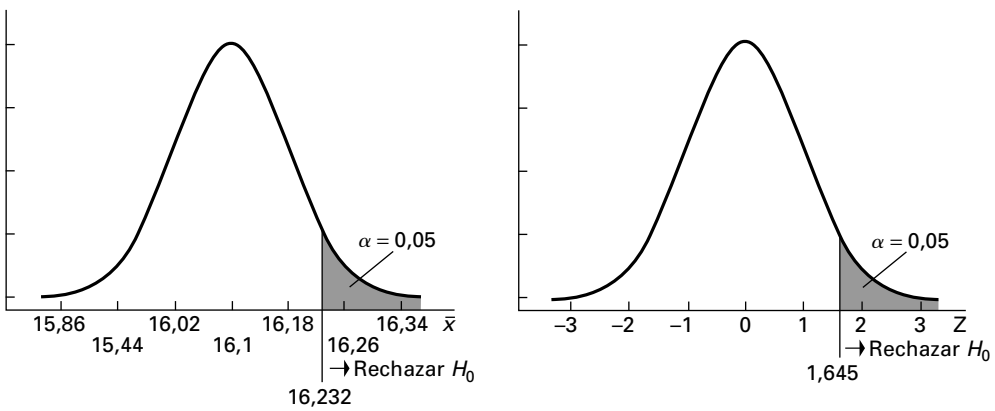


Figura 10.2. Función de densidad normal que muestra los valores tanto de Z como de $\bar{X}$ para la regla de decisión para contrastar la hipótesis nula $H_0 : \mu = 16{,}1$ frente a $H_1 : \mu > 16{,}1$.

Un contraste de la media de una población normal: varianza conocida

Tenemos una muestra aleatoria de n observaciones procedentes de una población que sigue una distribución normal de media μ y varianza conocida σ^2. Si la media muestral observada es $\bar{x}$, se obtiene un contraste con un nivel de significación α de la hipótesis nula

$$H_0 : \mu = \mu_0$$

frente a la alternativa

$$H_1 : \mu > \mu_0$$

utilizando la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \tag{10.1}$$

o, lo que es lo mismo,

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} > \mu_0 + z_\alpha \sigma/\sqrt{n}$$

donde z_α es el número para el que

$$P(Z > z_\alpha) = \alpha$$

y Z es la variable aleatoria normal estándar.

Detengámonos un momento a ver qué se entiende por rechazo de una hipótesis nula. En el problema de la caja de cereales, la hipótesis de que la media poblacional es 16,1 se rechazaría con un nivel de significación de 0,05 si $\bar{x} > 16{,}232$. Eso no significa, desde luego, que tengamos una prueba de que la media poblacional del peso sea superior a 16,1 unidades. Dada únicamente la información muestral, nunca podemos estar seguros sobre un parámetro poblacional. Podríamos concluir, más bien, que los datos han puesto en tela de juicio la veracidad de la hipótesis nula. Si ésta fuera verdadera, vemos que si en una muestra encontramos una media muestral, por ejemplo, de $\bar{x} = 16{,}3$ (observemos que $16{,}3 > 16{,}232$), ésta representaría una única observación improbable extraída de una distribución normal de media 16,1 y desviación típica

$$\frac{\sigma}{\sqrt{n}} = \frac{0{,}4}{\sqrt{25}} = 0{,}08$$

Lo que estamos preguntándonos realmente es qué probabilidad habría de observar un valor tan extremo si la hipótesis nula fuera en realidad verdadera. Hemos visto que la probabilidad de observar un valor medio superior a 16,232 es 0,05. Por lo tanto, al rechazar la hipótesis nula, o bien ésta es falsa, o bien hemos observado un suceso improbable, un suceso que sólo ocurriría con una probabilidad inferior a la que especifica el nivel de significación. Éste es el sentido en el que la información muestral ha suscitado dudas sobre la hipótesis nula.

p-valor

Existe otro conocido método para examinar el contraste de la hipótesis nula. Obsérvese que en nuestro problema de los cereales se rechaza la hipótesis nula al nivel de significación de 0,05, pero no se habría rechazado al nivel más bajo de 0,01. Si utilizáramos un

nivel de significación más bajo, reduciríamos la probabilidad de rechazar una hipótesis nula verdadera. Eso modificaría nuestra regla de decisión para que fuera menos probable que rechazáramos la hipótesis nula independientemente de que fuera verdadera o no. Evidentemente, cuanto menor es el nivel de significación al que rechazamos una hipótesis nula, mayores son las dudas sobre su veracidad. En lugar de contrastar hipótesis a los niveles preasignados de significación, los investigadores a menudo hallan el nivel menor de significación al que puede rechazarse una hipótesis nula.

El *p*-valor es la probabilidad de obtener un valor del estadístico del contraste igual de extremo o más que el valor efectivo obtenido cuando la hipótesis nula es verdadera. Por lo tanto, el *p*-valor es el menor nivel de significación al que puede rechazarse una hipótesis nula, dado el estadístico muestral observado. Supongamos, por ejemplo, que en el problema de las cajas de cereales con una media poblacional igual a 16,1, $\sigma = 0{,}4$ y $n = 25$ y partiendo de la hipótesis nula hemos obtenido una media muestral de 16,3 onzas. En ese caso, el *p*-valor sería

$$P(\bar{x} > 16{,}3 \mid H_0 : \mu = 16{,}1) = P\left(Z > \frac{16{,}3 - 16{,}1}{0{,}08} = 2{,}5\right)$$
$$= 0{,}0062$$

En la tabla de probabilidad normal vemos que la probabilidad de obtener una media muestral de 16,3 o más si tomamos una distribución normal de media poblacional 16,1 y desviación típica de la media muestral 0,08 es igual a 0,0062. Por lo tanto, el *p*-valor de este contraste es 0,0062. Ahora bien, el *p*-valor (0,0062) representa el menor nivel de significación, α, que llevaría a rechazar la hipótesis nula. Cuando calculamos el *p*-valor, podemos contrastar la hipótesis nula utilizando la regla

$$\text{Rechazar } H_0 \text{ si} \quad p\text{-valor} < \alpha$$

Esta regla lleva a la misma conclusión que la que se obtiene utilizando la ecuación 10.1.

Existe otra razón más importante por la que se utiliza a menudo el *p*-valor. El *p*-valor suministra información más precisa sobre la fuerza del rechazo de la hipótesis nula resultante de la media muestral observada. Supongamos que en el contraste del peso de las cajas de cereales hubiéramos fijado el nivel de significación en $\alpha = 0{,}05$, nivel que se elige frecuentemente. En ese caso, con una media muestral igual a 16,3, diríamos que la hipótesis nula se ha rechazado con un nivel de significación de 0,05. Sin embargo, en realidad, ese resultado muestral apunta a una conclusión mucho más fuerte. Podríamos haber rechazado la hipótesis nula a un nivel de significación de $\alpha = 0{,}0063$. Supongamos, por el contrario, que el *p*-valor calculado basándose en una media muestral diferente hubiera sido 0,07. En ese caso, no podríamos rechazar la hipótesis nula, pero también sabríamos que casi la rechazaríamos. En cambio, un *p*-valor de 0,30 nos diría que distaríamos mucho de rechazar la hipótesis nula. El *p*-valor se utiliza frecuentemente porque no sólo indica que se ha aceptado o se ha rechazado la hipótesis nula a un determinado nivel de significación. A continuación resumimos el *p*-valor.

Interpretación del valor de la probabilidad o *p*-valor

El valor de la probabilidad o *p*-valor es el nivel de significación más bajo al que puede rechazarse la hipótesis nula. Consideremos una muestra aleatoria de n observaciones procedente de una población que sigue una distribución normal de media μ y desviación típica σ y la media muestral calculada resultante, $\bar{x}$. Se ha contrastado la hipótesis nula

$$H_0 : \mu = \mu_0$$

frente a la hipótesis alternativa

$$H_1 : \mu > \mu_0$$

El *p*-valor del contraste es

$$p\text{-valor} = P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geqslant z_p \,|\, H_0 : \mu = \mu_0\right) \qquad \textbf{(10.2)}$$

donde z_p es el valor normal estándar correspondiente al menor nivel de significación al que puede rechazarse la hipótesis nula. La mayoría de los programas informáticos estadísticos calculan normalmente el *p*-valor; éste suministra más información sobre el contraste, basándose en la media muestral observada, por lo que es un instrumento que se utiliza frecuentemente en muchas aplicaciones estadísticas.

Es importante señalar que el *p*-valor es una variable aleatoria observada que es diferente en el caso de cada muestra aleatoria obtenida para realizar un contraste estadístico. Por lo tanto, dos analistas diferentes podrían obtener sus propias muestras aleatorias y sus propias medias muestrales de una misma población y, por lo tanto, calcular cada uno un *p*-valor diferente.

EJEMPLO 10.1. Evaluación de un nuevo proceso de producción (contraste de hipótesis)

El director de producción de Ventanas Norte, S.A., le ha pedido que evalúe un nuevo método propuesto para producir su línea de ventanas de doble hoja. El proceso actual tiene una producción media de 80 unidades por hora con una desviación típica poblacional de $\sigma = 8$. El director indica que no quiere sustituirlo por otro método, a menos que existan pruebas contundentes de que el nivel medio de producción es mayor con el nuevo método.

Solución

El director sólo adoptará el nuevo método si existen pruebas contundentes a su favor. Por lo tanto, la hipótesis nula es

$$H_0 : \mu \leqslant 80$$

y la hipótesis alternativa,

$$H_0 : \mu > 80$$

Vemos que si fijamos el nivel de significación $\alpha = 0{,}5$ y llegamos a la conclusión de que el nuevo método es más productivo, nuestra probabilidad de error es de 0,05 o menos. Eso implica que existen pruebas contundentes a favor de nuestra recomendación.

Obtenemos una muestra aleatoria de $n = 25$ horas de producción utilizando el nuevo método propuesto y calculamos la media muestral $\bar{x}$, a menudo utilizando un computador. Con un nivel de significación de $\alpha = 0{,}05$, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - 80}{8/\sqrt{25}} > 1{,}645$$

donde se obtiene $z_{0,05} = 1,645$ en la tabla de la normal estándar. También podríamos utilizar la regla

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} > \mu_0 + z_\alpha \sigma/\sqrt{n} = 80 + 1,645 \times (8/\sqrt{25}) = 82,63$$

Supongamos que la media muestral resultante fuera $\bar{x} = 83$. Basándonos en ese resultado,

$$z = \frac{83 - 80}{8/\sqrt{25}} = 1,875 > 1,645$$

rechazaríamos la hipótesis nula y concluiríamos que tenemos pruebas contundentes para apoyar la conclusión de que el nuevo método aumenta la productividad. Dada esta media muestral, también podríamos calcular el *p*-valor:

$$p\text{-valor} = P(z_p > 1,875) = 0,03$$

Podríamos recomendar, pues, el nuevo método al director de producción.

Un contraste de la media de una distribución normal (varianza conocida): hipótesis nula y alternativa compuestas

El método adecuado para contrastar a un nivel de significación α la hipótesis nula

$$H_0 : \mu \leqslant \mu_0$$

frente a la hipótesis alternativa

$$H_1 : \mu > \mu_0$$

es precisamente igual que el que se emplea cuando la hipótesis nula es $H_0 : \mu = \mu_0$. Además, los *p*-valores también se calculan exactamente de la misma forma.

Consideremos nuestro ejemplo anterior sobre el peso de las cajas de cereales. Supongamos que las normas del sector establecen que, si el peso medio de las cajas no es de 16 onzas o más en una población de cajas que indican que pesan 16 onzas, se presentará una demanda contra la empresa. En esta situación, el organismo regulador sólo podría demandarla si encontrara pruebas contundentes de que el peso medio de las cajas es de menos de 16 onzas. Por lo tanto, su objetivo es demostrar que el peso medio de las cajas, μ, no es de 16,0 onzas o más. En este caso, la hipótesis nula simple sería

$$H_0 : \mu = \mu_0 = 16,0$$

o, utilizando la hipótesis compuesta,

$$H_0 : \mu \geqslant \mu_0 = 16,0$$

y la hipótesis alternativa sería

$$H_1 : \mu < \mu_0 = 16,0$$

para la hipótesis simple o para la hipótesis compuesta. Formulando nuestra regla de contraste con un nivel de significación α, sabemos que, si rechazamos la hipótesis nula, tenemos pruebas contundentes de que el peso medio es de menos de 16,0 onzas, ya que la probabilidad de cometer un error de Tipo I tiene un pequeño valor, α.

Nuestro contraste de la media poblacional utiliza la media muestral, $\bar{x}$. Si la media muestral es considerablemente inferior a $\mu_0 = 16{,}0$, rechazamos la hipótesis nula. Para obtener el valor de decisión adecuado, utilizamos el hecho de que la variable aleatoria estándar

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

sigue una distribución normal estándar de media 0 y varianza 1 cuando la media poblacional es μ_0. Si z tiene un elevado valor negativo tal que

$$P(Z < -z_\alpha) = \alpha$$

entonces, para contrastar la hipótesis nula, podemos utilizar la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

Se deduce que la probabilidad de rechazar la hipótesis nula, H_0, cuando es verdadera es el nivel de significación α.

Obsérvese que realizando una sencilla manipulación algebraica, también podríamos formular la siguiente regla de decisión:

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} < \bar{x}_c = \mu_0 - z_\alpha \sigma/\sqrt{n}$$

El valor $\bar{x}_c$ es el «valor crítico» de la decisión. Obsérvese que para todo valor $-z_\alpha$ obtenido de la distribución normal estándar, también hay un valor $\bar{x}_c$ y cualquiera de las reglas de decisión anteriores da exactamente el mismo resultado.

Supongamos que en este problema la desviación típica poblacional es $\sigma = 0{,}4$ y obtenemos una muestra aleatoria de tamaño 25. En el caso de un contraste de hipótesis con un nivel de significación $\alpha = 0{,}05$, vemos en la tabla de la distribución normal estándar que el valor de $z_\alpha = 1{,}645$. En este caso, nuestra regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 16{,}0}{0{,}4/\sqrt{25}} < -1{,}645$$

o podríamos utilizar la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} < \bar{x}_c = \mu_0 - z_\alpha \sigma/\sqrt{n} = 16{,}0 - 1{,}645 \times (0{,}4/\sqrt{25}) = 15{,}868$$

Si rechazamos H_0 utilizando esta regla, aceptamos la hipótesis alternativa de que el peso medio es de menos de 16,0 onzas con la probabilidad de cometer un error de Tipo I de 0,05 o menos. Ésta es una prueba contundente a favor de nuestra conclusión. Esta regla de decisión se muestra en la Figura 10.3.

Obsérvese que este contraste de hipótesis es el complemento del primer ejemplo. Las reglas del contraste de hipótesis de las hipótesis alternativas que se refieren a la cola inferior son imágenes gemelas de las reglas de contraste de las hipótesis que se refieren a la cola superior de la distribución. También pueden calcularse los *p*-valores utilizando las probabilidades de la cola inferior en lugar de las probabilidades de la cola superior. Este resultado se resume en la ecuación 10.3.

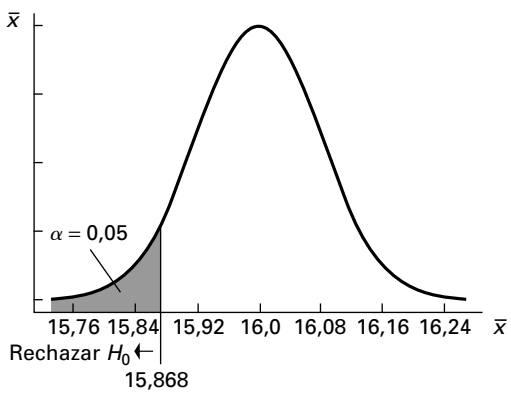


Figura 10.3. Función de densidad normal que muestra los valores de $\bar{x}$ correspondientes a la regla de decisión para contrastar la hipótesis nula $H_0: \mu \geqslant 16{,}0$ frente a $H_1: \mu < 16{,}0$.

Los ejemplos de los cereales tenían dos objetivos distintos. En el primer caso, queríamos encontrar pruebas contundentes de que el peso medio era de más de 16,1 onzas, por lo que la hipótesis nula era

$$H_0: \mu \leqslant 16{,}1$$

En el segundo caso, queríamos encontrar pruebas contundentes de que la media era de menos de 16 onzas, por lo que la hipótesis nula era

$$H_0: \mu \geqslant 16$$

Este tipo de posibilidades está presente en muchas situaciones en las que hay que tomar decisiones y el responsable de tomarlas tiene que saber qué opción debe utilizar en el problema en cuestión.

Un contraste de la media de una distribución normal (varianza conocida): hipótesis nula y alternativa compuestas o simples

El método adecuado para contrastar al nivel de significación α la hipótesis nula

$$H_0: \mu = \mu_0 \quad \text{o} \quad \mu \geqslant \mu_0$$

frente a la hipótesis alternativa

$$H_1: \mu < \mu_0$$

utiliza la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \quad Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

o, lo que es lo mismo,

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} < \bar{x}_c = \mu_0 - z_\alpha \sigma/\sqrt{n} \tag{10.3}$$

donde $-z_\alpha$ es el número para el que

$$P(Z < -z_\alpha) = \alpha$$

y Z es la variable aleatoria normal estándar.

Además, pueden calcularse los p-valores utilizando las probabilidades de la cola inferior.

EJEMPLO 10.2. Producción de rodamientos (contraste de hipótesis)

El director de producción de Rodamientos Niquelados, S.A., le ha pedido ayuda para evaluar un proceso modificado de producción de rodamientos. Cuando el proceso funciona correctamente, produce rodamientos cuyo peso sigue una distribución normal de media poblacional 5 onzas y desviación típica poblacional 0,1 onzas. Se ha recurrido a un nuevo proveedor de materia prima para un lote reciente de producción y el director quiere saber si, como consecuencia del cambio, el peso medio de los rodamientos es menor. No hay razón alguna para sospechar que el nuevo proveedor plantea problemas y el director continuará recurriendo a él a menos que existan pruebas contundentes de que están produciéndose rodamientos de menor peso que antes.

Solución

En este caso, nos interesa saber si existen pruebas contundentes para concluir que están produciéndose rodamientos de menor peso. Por lo tanto, contrastamos la hipótesis nula

$$H_0 : \mu = \mu_0 = 5$$

frente a la hipótesis alternativa

$$H_1 : \mu < 5$$

Obsérvese cómo nos lleva el concepto de pruebas contundentes a elegir la hipótesis nula y la hipótesis alternativa. Sólo emprendemos acciones si se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Se especifica un nivel de significación $\alpha = 0{,}05$ y, por lo tanto, el valor de la variable aleatoria normal estándar correspondiente a la cola inferior es $z_\alpha = -1{,}645$ según la tabla de la distribución normal. En este problema, obtenemos una muestra aleatoria de $n = 16$ observaciones y la media muestral es 4,962. Nuestra regla de decisión para este problema es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -1{,}645$$

o

$$\text{Rechazar } H_0 \text{ si } \bar{x} < 4{,}962$$
$$4{,}962 < 4{,}959$$

Vemos que no podemos rechazar la hipótesis nula, H_0, ya que $\dfrac{4{,}962 - 5{,}0}{0{,}1/\sqrt{16}} = -1{,}52$ y $5 - 1{,}645\left(\dfrac{0{,}1}{\sqrt{16}}\right) = 4{,}959$ y, por lo tanto, concluimos que no tenemos pruebas contundentes de que el proceso de producción esté produciendo rodamientos de menor peso que antes.

También podríamos calcular el *p*-valor correspondiente a este resultado muestral señalando que en el caso de la distribución normal estándar

$$p\text{-valor} = P(z_p < -1{,}52) = 0{,}0643$$

Hipótesis alternativa bilateral

Hay algunos problemas en los que las desviaciones demasiado altas o demasiado bajas tienen la misma importancia. Por ejemplo, el diámetro de un pistón de un automóvil no puede ser demasiado grande o demasiado pequeño. En esas situaciones, consideramos el contraste de la hipótesis nula

$$H_0 : \mu = \mu_0$$

frente a la hipótesis alternativa

$$H_1 : \mu \neq \mu_0$$

En este caso, no tenemos razones contundentes para sospechar que hay desviaciones por encima o por debajo de la media poblacional postulada como hipótesis, μ_0. Dudaríamos de la hipótesis nula si la media muestral fuera mucho mayor o mucho menor que μ_0. De nuevo, si la variable aleatoria sigue una distribución normal con una varianza conocida σ, obtenemos un contraste con un nivel de significación α utilizando el resultado de que según la hipótesis nula

$$P(Z > z_{\alpha/2}) = \frac{\alpha}{2} \quad \text{y} \quad P(Z < -z_{\alpha/2}) = \frac{\alpha}{2}$$

En este caso, hemos dividido el nivel de significación α por igual entre las dos colas de la distribución normal. Por lo tanto, la probabilidad de que Z sea superior a $z_{\alpha/2}$ o inferior a $-z_{\alpha/2}$ es α. La regla de decisión de un contraste con un nivel de significación α es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

es superior a $z_{\alpha/2}$ o inferior a $-z_{\alpha/2}$. Estos resultados se resumen en la ecuación 10.4.

Un contraste de la media de una distribución normal frente a una hipótesis alternativa bilateral (varianza conocida)

El método adecuado para contrastar a un nivel de significación α la hipótesis nula

$$H_0 : \mu = \mu_0$$

frente a la hipótesis alternativa

$$H_1 : \mu \neq \mu_0$$

utiliza la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \quad \text{o} \quad \text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2} \qquad \textbf{(10.4)}$$

o, lo que es lo mismo,

$$\text{Rechazar } H_0 \text{ si } \bar{x} < \mu_0 - z_{\alpha/2}\sigma/\sqrt{n} \quad \text{o} \quad \text{Rechazar } H_0 \text{ si } \bar{x} > \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$$

Además, pueden calcularse los *p*-valores observando que la probabilidad de la cola correspondiente se duplicaría para reflejar un *p*-valor que se refiere a la suma de las probabilidades de la

cola superior y la cola inferior para los valores positivos y negativos de Z. El p-valor correspondiente al contraste de dos colas es

$$p\text{-valor} = 2P\left(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{p/2} \,|\, H_0 : \mu = \mu_0\right) \qquad \textbf{(10.5)}$$

donde $z_{p/2}$ es el valor normal estándar correspondiente a la menor probabilidad de rechazar la hipótesis nula en cualquiera de las dos colas de la distribución de probabilidad.

EJEMPLO 10.3. Análisis del diámetro de los taladros (contraste de hipótesis)

El director de producción de Circuitos Ilimitados le ha pedido ayuda para analizar un proceso de producción. Este proceso consiste en hacer taladros cuyo diámetro sigue una distribución normal de media poblacional 2 centímetros y desviación típica poblacional 0,06 centímetros. Una muestra aleatoria de nueve mediciones tenía una media muestral de 1,95 centímetros. Utilice un nivel de significación de $\alpha = 0{,}05$ para averiguar si la media muestral observada es excepcional y sugiere que debe ajustarse la taladradora.

Solución

En este caso, el diámetro podría ser demasiado grande o demasiado pequeño. Por lo tanto, realizamos un contraste de hipótesis de dos colas planteando la siguiente la hipótesis nula:

$$H_0 : \mu = 2{,}0$$

y la hipótesis alternativa

$$H_1 : \mu \neq 2{,}0$$

La regla de decisión es rechazar H_0 en favor de H_1 si

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \quad \text{o} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

y en este problema

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1{,}95 - 2{,}0}{0{,}06/\sqrt{9}} = -2{,}50$$

para un contraste de nivel del 5 por ciento, $\alpha = 0{,}05$ y $z_{\alpha/2} = z_{0{,}05/2} = 1{,}96$. Por lo tanto, como $-2{,}50$ es menor que $-1{,}96$, rechazamos la hipótesis nula y concluimos que es necesario ajustar la taladradora.

Para calcular el p-valor, primero observamos que en la tabla de distribución normal la probabilidad de obtener una Z inferior a $-2{,}50$ es 0,0062. Aquí queremos el p-valor para un contraste de dos colas y debemos duplicar el valor de una cola. Por lo tanto, el p-valor de este contraste es 0,0124 y la hipótesis nula se habría rechazado para un nivel de significación superior a 1,24 por ciento.

En la Figura 10.10, que se encuentra en el resumen del capítulo, hemos resumido las distintas alternativas para contrastar hipótesis analizadas en este apartado.

EJERCICIOS

Ejercicios básicos

10.6. Se obtiene una muestra aleatoria de una población que tiene una varianza $\sigma^2 = 625$ y se calcula la media muestral. Contraste la hipótesis nula $H_0: \mu = 100$ frente a la hipótesis alternativa $H_1: \mu \geqslant 100$ con $\alpha = 0{,}05$. Calcule el valor crítico $\bar{x}_c$ y formule su regla de decisión para las siguientes opciones.

a) Tamaño de la muestra $n = 25$
b) Tamaño de la muestra $n = 16$
c) Tamaño de la muestra $n = 44$
d) Tamaño de la muestra $n = 32$

10.7. Se obtiene una muestra aleatoria de tamaño $n = 25$ de una población que tiene una varianza σ^2 y se calcula la media muestral. Contraste la hipótesis nula $H_0: \mu = 100$ frente a la hipótesis alternativa $H_1: \mu \geqslant 100$ con $\alpha = 0{,}05$. Calcule el valor crítico $\bar{x}_c$ y formule su regla de decisión para las siguientes opciones.

a) La variable poblacional es $\sigma^2 = 225$.
b) La variable poblacional es $\sigma^2 = 900$.
c) La variable poblacional es $\sigma^2 = 400$.
d) La variable poblacional es $\sigma^2 = 600$.

10.8. Utilizando los resultados de los dos ejercicios anteriores, indique cómo influye el tamaño de la muestra en el valor crítico $\bar{x}_c$. A continuación, indique cómo influye la varianza poblacional σ^2 en el valor crítico.

10.9. Se obtiene una muestra aleatoria de una población que tiene una varianza $\sigma^2 = 400$ y se calcula la media muestral $\bar{x}_c = 70$. Considere la hipótesis nula $H_0: \mu = 80$ frente a la hipótesis alternativa $H_1: \mu \leqslant 80$. Calcule el p-valor para las siguientes opciones.

a) Tamaño de la muestra $n = 25$
b) Tamaño de la muestra $n = 16$
c) Tamaño de la muestra $n = 44$
d) Tamaño de la muestra $n = 32$

10.10. Se obtiene una muestra aleatoria de tamaño $n = 25$ de una población que tiene la varianza σ^2 y se calcula la media muestral $\bar{x}_c = 70$. Considere la hipótesis nula $H_0: \mu = 80$ frente a la hipótesis alternativa $H_1: \mu \leqslant 80$. Calcule el p-valor para las siguientes opciones.

a) La varianza poblacional es $\sigma^2 = 225$.
b) La varianza poblacional es $\sigma^2 = 900$.
c) La varianza poblacional es $\sigma^2 = 400$.
d) La varianza poblacional es $\sigma^2 = 600$.

Ejercicios aplicados

10.11. Un fabricante de detergente sostiene que los contenidos de las cajas que vende pesan, en promedio, 16 onzas como mínimo. Se sabe que la distribución del peso es normal y tiene una desviación típica de 0,4 onzas. Una muestra aleatoria de 16 cajas ha dado un peso medio muestral de 15,84 onzas. Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la media poblacional del peso es al menos de 16 onzas.

10.12. Una empresa que recibe envíos de pilas comprueba una muestra aleatoria de nueve antes de aceptar un envío. Quiere que la verdadera duración media de todas las pilas del envío sea al menos de 50 horas. Sabe por experiencia que la distribución poblacional de la duración es normal y tiene una desviación típica de 3 horas. La duración media de una muestra de nueve pilas de un envío es de 48,2 horas. Contraste al nivel del 10 por ciento la hipótesis nula de que la media poblacional de la duración es al menos de 50 horas.

10.13. Una empresa farmacéutica quiere que la concentración de impurezas de sus píldoras no supere el 3 por ciento. Se sabe que la concentración de impurezas de un lote sigue una distribución normal con una desviación típica del 0,4 por ciento. Se comprueba una muestra aleatoria de 64 píldoras de un lote y se observa que la media muestral de la concentración de impurezas es de 3,07 por ciento.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional de la concentración de impurezas es del 3 por ciento frente a la alternativa de que es de más del 3 por ciento.
b) Halle el p-valor para este contraste.
c) Suponga que la hipótesis alternativa hubiera sido bilateral en lugar de unilateral (con una hipótesis nula $H_0: \mu = 3$). Indique sin hacer los cálculos si el p-valor del contraste sería mayor, menor o igual que el obtenido en el apartado (b). Represente gráficamente su razonamiento.
d) Explique por qué en este problema es más adecuada una hipótesis alternativa unilateral que una bilateral.

10.3. Contrastes de la media de una distribución normal: varianza poblacional desconocida

En este apartado analizamos el mismo conjunto de contrastes de hipótesis que hemos analizado en el apartado 10.2. La única diferencia estriba en que la variable poblacional es desconocida y, por lo tanto, debemos utilizar contrastes basados en la distribución *t* de Student. En el apartado 8.3 presentamos la distribución *t* de Student y mostramos su aplicación para desarrollar intervalos de confianza. Recuérdese que la distribución *t* de Student depende de los grados de libertad para calcular la varianza muestral, $n - 1$. Además, va pareciéndose cada vez más a la distribución normal a medida que aumenta el tamaño de la muestra. Por lo tanto, cuando el tamaño de la muestra es de más de 100, la distribución de probabilidad normal es una buena aproximación de la distribución *t* de Student. Utilizando la media muestral y la varianza muestral, sabemos que la variable aleatoria

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

sigue una distribución *t* de Student. Los métodos para realizar contrastes de hipótesis utilizando la varianza muestral se definen en las ecuaciones 10.6, 10.7 y 10.8.

Contrastes de la media de una distribución normal: varianza poblacional desconocida

Tenemos una muestra aleatoria de *n* observaciones procedentes de una población normal que tiene una media μ. Utilizando la media muestral y la desviación típica muestral, $\bar{x}$ y *s*, respectivamente, podemos utilizar los siguientes contrastes con el nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \mu = \mu_0 \quad \text{o} \quad H_0: \mu \leqslant \mu_0$$

frente a la alternativa

$$H_1: \mu > \mu_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si} \quad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha}$$

o, lo que es lo mismo,

$$\text{Rechazar } H_0 \text{ si} \quad \bar{x} > \bar{x}_c = \mu_0 + t_{n-1,\alpha} s/\sqrt{n} \tag{10.6}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \mu = \mu_0 \quad \text{o} \quad H_0: \mu \geqslant \mu_0$$

frente a la alternativa

$$H_1: \mu < \mu_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha} \tag{10.7}$$

o, lo que es lo mismo,

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} < \bar{x}_c = \mu_0 - t_{n-1,\alpha} s/\sqrt{n}$$

3. Para contrastar la hipótesis nula

$$H_0 : \mu = \mu_0$$

frente a la hipótesis nula

$$H_1 : \mu \neq \mu_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha/2} \quad \text{o}$$

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha/2} \tag{10.8}$$

o, lo que es lo mismo,

$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} < \mu_0 - t_{n-1,\alpha/2} s/\sqrt{n} \quad \text{o}$$
$$\text{Rechazar } H_0 \text{ si } \quad \bar{x} > \mu_0 + t_{n-1,\alpha/2} s/\sqrt{n}$$

donde $t_{n-1,\alpha/2}$ es el valor de la t de Student con $n-1$ grados de libertad y probabilidad $\alpha/2$.

Los *p*-valores de estos contrastes se calculan de la misma forma que en el caso de los contrastes con varianza conocida, con la salvedad de que el valor de la Z normal se sustituye por el valor de la t de Student. Para hallar el *p*-valor, a menudo necesitamos interpolar valores con la tabla de la t o utilizar un paquete informático como el Minitab.

EJEMPLO 10.4. Análisis de las ventas semanales de brócoli congelado (contraste de hipótesis)

Broccoli

Grand Junction Vegetables es un fabricante de una amplia variedad de verduras congeladas. El presidente de la empresa le ha pedido que averigüe si las ventas semanales de las bolsas de brócoli congelado de 16 onzas han aumentado. En los 6 últimos meses, se ha vendido una media semanal de 2.400 bolsas. Ha obtenido una muestra aleatoria de datos de ventas de 134 tiendas para realizar el estudio. Los datos se encuentran en el fichero **Broccoli**.

Solución

Dados los objetivos del proyecto, decidimos que hay que contrastar la hipótesis nula de que la media poblacional de las ventas es 2.400 frente a la alternativa de que las ventas han aumentado utilizando un nivel de significación $\alpha = 0{,}05$. La hipótesis nula es

$$H_0 : \mu = 2.400$$

frente a la hipótesis alternativa

$$H_1 : \mu > 2.400$$

La Figura 10.4 muestra la salida Minitab que contiene la media muestral y la varianza muestral. En la salida Minitab vemos que la media muestral es mucho mayor que la mediana y que el cuartil superior tiene un rango muy amplio. Es evidente, pues, que la distribución de las observaciones es normal. Pero el tamaño de la muestra es grande y, por lo tanto, aplicando el teorema del límite central del Capítulo 7, podemos suponer que la distribución de la media muestral en el muestreo es normal; por lo tanto, sería adecuado un contraste basado en la t de Student para el contraste de hipótesis. Vemos que la media muestral es 3.593 y la desviación típica muestral es 4.919. El estadístico t es

$$t = \frac{3.593 - 2.400}{4.919/\sqrt{134}} = 2{,}81$$

Descriptive Statistics: Broccoli

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Broccoli	134	0	3593	425	4919	156	707	2181	2300	27254

Figura 10.4. Estadísticos descriptivos de las ventas de brócoli (salida Minitab).

El valor de t con $n - 1 = 133$ grados de libertad y $\alpha = 0{,}05$ en el caso de la cola superior es aproximadamente 1,645. Basándonos en este resultado, rechazamos la hipótesis nula y concluimos que las ventas medias han aumentado.

Los contrastes presentados en este apartado se resumen en la Figura 10.10, que se encuentra en el resumen del capítulo.

EJERCICIOS

Ejercicios básicos

10.14. Contraste las hipótesis

$$H_0 : \mu \leqslant 100$$
$$H_1 : \mu > 100$$

utilizando una muestra aleatoria de tamaño $n = 25$, una probabilidad de cometer un error de Tipo I igual a 0,05 y los siguientes estadísticos muestrales.

a) $\bar{x} = 106$; $s = 15$
b) $\bar{x} = 104$; $s = 10$
c) $\bar{x} = 95$; $s = 10$
d) $\bar{x} = 92$; $s = 18$

10.15. Contraste las hipótesis

$$H_0 : \mu = 100$$
$$H_1 : \mu < 100$$

utilizando una muestra aleatoria de tamaño $n = 36$, una probabilidad de cometer un error de Tipo I igual a 0,05 y los siguientes estadísticos muestrales.

a) $\bar{x} = 106$; $s = 15$
b) $\bar{x} = 104$; $s = 10$
c) $\bar{x} = 95$; $s = 10$
d) $\bar{x} = 92$; $s = 18$

Ejercicios aplicados

10.16. Un centro de investigación de ingeniería sostiene que, utilizando un nuevo sistema de control informático, los automóviles deben lograr, en promedio, 3 kilómetros más por litro de gasolina. Se ha utilizado una muestra aleatoria de 100 automóviles para evaluar este producto. La media muestral del aumento de los kilómetros por litro logrados es de 2,4 y la desviación típica muestral es de 1,8 kilómetros por litro. Contraste la hipótesis de que la media poblacional es al menos de 3 kilómetros por litro. Halle

el *p*-valor de este contraste e interprete sus resultados.

10.17. Una muestra aleatoria de 1.562 estudiantes universitarios matriculados en un curso de ética empresarial debe responder en una escala de 1 (totalmente en desacuerdo) a 7 (totalmente de acuerdo) a esta proposición: «A los altos ejecutivos de las empresas les preocupa la justicia social». La media muestral de las respuestas es 4,27 y la desviación típica muestral es 1,32. Contraste al nivel del 1 por ciento la hipótesis nula de que la media poblacional es 4 frente a la hipótesis alternativa bilateral.

10.18. Le han pedido que evalúe la respuesta de las empresas a una nueva obligación legal de incrementar las prestaciones sanitarias que ofrecen a sus empleados. Tiene una muestra aleatoria de 76 cambios porcentuales de las prestaciones sanitarias prometidas. La media muestral de los cambios porcentuales es 0,078 y la desviación típica muestral es 0,201. Halle e interprete el *p*-valor de un contraste de la hipótesis nula de que la media poblacional de los cambios porcentuales es 0 frente a la hipótesis alternativa bilateral.

10.19. Se pide a una muestra aleatoria de 172 estudiantes de marketing que valoren en una escala de 1 (nada importante) a 5 (muy importante) las prestaciones sanitarias complementarias como característica del empleo. La media muestral de las valoraciones es 3,31 y la desviación típica muestral es 0,70. Contraste al nivel de significación del 1 por ciento la hipótesis nula de que la media poblacional de las valoraciones es como máximo de 3,0 frente a la hipótesis alternativa de que es superior a 3,0.

10.20. Se plantea a una muestra aleatoria de 170 personas un problema de predicción. Cada miembro de la muestra tiene que predecir de dos formas el próximo valor de una variable relacionada con las ventas al por menor. Se les presentan los 20 valores anteriores tanto en términos numéricos como en forma de puntos en un gráfico. Se les pide que predigan el próximo valor. Se miden los errores absolutos de predicción. La muestra consta, pues, de 170 diferencias entre los errores absolutos de predicción (numéricos menos gráficos). La media muestral de estas diferencias es $-2{,}91$ y la desviación típica muestral es 11,33. Halle e interprete el *p*-valor de un contraste de la hipótesis nula de que la media poblacional de las diferencias es 0 frente a la hipótesis alternativa de que es negativa (la hipótesis alternativa puede ser la hipótesis de que en conjunto la gente tiene más éxito en la predicción gráfica que en la numérica).

10.21. Las cuentas de una empresa muestran que, en promedio, las facturas pendientes de cobro ascienden a 125,32 $. Un auditor comprueba una muestra aleatoria de 16 cuentas. La media muestral es de 131,78 $ y la desviación típica muestral es 25,41 $. Suponga que la distribución poblacional es normal. Contraste al nivel de significación del 5 por ciento la hipótesis nula de que la media poblacional es 125,32 $ frente a la hipótesis alternativa bilateral.

10.22. Basándose en una muestra aleatoria, se contrasta la hipótesis nula

$$H_0 : \mu = \mu_0$$

frente a la alternativa

$$H_1 : \mu > \mu_0$$

y la hipótesis nula no se rechaza al nivel de significación del 5 por ciento.

a) ¿Implica eso necesariamente que μ_0 está contenida en el intervalo de confianza al 95 por ciento de μ?

b) ¿Implica eso necesariamente que μ_0 está contenida en el intervalo de confianza al 90 por ciento de μ si la media muestral observada es mayor que μ_0?

10.23. Una empresa que vende licencias de un nuevo programa informático de comercio electrónico anuncia que las empresas que lo utilizan obtienen, en promedio, durante el primer año un rendimiento del 10 por ciento por sus inversiones iniciales. Una muestra aleatoria de 10 de estas franquicias generó los siguientes rendimientos durante el primer año:

6,1 9,2 11,5 8,6 12,1 3,9 8,4 10,1 9,4 8,9

Suponiendo que los rendimientos poblacionales siguen una distribución normal, contraste la afirmación de la empresa.

10.24. Un proceso que produce botes de champú, cuando funciona correctamente, produce botes cuyo contenido pesa, en promedio, 200 gramos. Una muestra aleatoria de nueve botes procedentes de un lote tiene el siguiente peso (en gramos):

21,4 19,7 19,7 20,6 20,8 20,1 19,7 20,3 20,9

Suponiendo que la distribución poblacional es normal, contraste al nivel del 5 por ciento la

hipótesis nula de que el proceso funciona correctamente frente a una hipótesis alternativa bilateral.

10.25. Un profesor de estadística tiene interés en conocer la capacidad de los estudiantes para evaluar la dificultad de un examen que han hecho. Este examen se ha realizado a un gran grupo de estudiantes y la calificación media ha sido de 78,5. Se pide a una muestra aleatoria de ocho estudiantes que predigan la calificación media. Sus predicciones son

72 83 78 65 69 77 81 71

Suponiendo que la distribución es normal, contraste la hipótesis nula de que la media poblacional de las predicciones es 78,5. Utilice la hipótesis alternativa bilateral y un nivel de significación del 10 por ciento.

10.26. Un distribuidor de cerveza sostiene que una nueva presentación, que consiste en una foto de tamaño real de un conocido cantante de rock, aumentará las ventas del producto en los supermercados en una media de 50 cajas en una semana. En una muestra aleatoria de 20 supermercados, las ventas medias aumentaron en 41,3 cajas y la desviación típica muestral fue de 12,2 cajas. Contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional del aumento de las ventas es al menos de 50 cajas, indicando los supuestos que postule.

10.27. En las negociaciones con los representantes sindicales, una empresa sostiene que con el nuevo sistema de incentivos los ingresos semanales medios de todos los trabajadores de los servicios de atención al cliente son al menos de 400 $. Un representante sindical toma una muestra aleatoria de 15 trabajadores y observa que sus ingresos semanales tienen una media de 381,35 $ y una desviación típica de 48,60 $. Suponga que la distribución es normal.

a) Contraste la afirmación de la empresa.

b) Si se hubieran obtenido los mismos resultados muestrales con una muestra aleatoria de 50 trabajadores, ¿podría rechazarse la afirmación de la empresa a un nivel de significación más bajo que el utilizado en el apartado (a)?

10.4. Contrastes de la proporción poblacional (grandes muestras)

Otro importante conjunto de problemas empresariales y económicos consiste en contrastar proporciones poblacionales. Los ejecutivos tienen interés en saber cuál es la cuota porcentual de mercado de sus productos y las autoridades tienen interés en saber cuál es el porcentaje de la población que apoya una nueva propuesta. Por lo tanto, la inferencia sobre la proporción poblacional basada en proporciones muestrales es una importante aplicación del contraste de hipótesis.

En los Capítulos 6 y 7 vimos que la distribución normal es una aproximación bastante precisa de la distribución de la proporción muestral. En esta aproximación, P representa la proporción poblacional y $\hat{p}$ la proporción muestral. Por lo tanto, la proporción muestral $\hat{p}$ estimada a partir de una muestra aleatoria de tamaño n sigue una distribución normal aproximada de P y varianza $P(1 - P)/n$. El estadístico normal estándar es

$$Z = \frac{\hat{p} - P}{\sqrt{P(1 - P)/n}}$$

Si la hipótesis nula es que la proporción poblacional es

$$H_0 : P = P_0$$

se deduce que, cuando esta hipótesis es verdadera, la variable aleatoria

$$Z = \frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}}$$

sigue aproximadamente una distribución normal estándar. Utilizando ese resultado, podemos definir los contrastes.

Contrastes de la proporción poblacional (grandes muestras)

Comenzamos con una muestra aleatoria de n observaciones procedentes de una población que tiene una proporción P cuyos miembros poseen un determinado atributo. Si $P(1 - P) > 9$ y la proporción muestral es $\hat{p}$, los siguientes contrastes tienen el nivel de significación α:

1. Para contrastar cualquiera de las dos hipótesis

$$H_0 : P = P_0 \quad \text{o} \quad H_0 : P \leqslant P_0$$

frente a la alternativa

$$H_1 : P > P_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}} > z_\alpha \tag{10.9}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : P = P_0 \quad \text{o} \quad H_0 : P \geqslant P_0$$

frente a la alternativa

$$H_1 : P < P_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}} < -z_\alpha \tag{10.10}$$

3. Para contrastar la hipótesis nula

$$H_0 : P = P_0$$

frente a la alternativa bilateral

$$H_1 : P \neq P_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}} > z_{\alpha/2} \quad \text{o} \quad \frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}} < -z_{\alpha/2} \tag{10.11}$$

En todos estos contrastes, el *p*-valor es el nivel de significación más bajo al que puede rechazarse la hipótesis nula.

Los contrastes presentados aquí se resumen en la Figura 10.11, página 389.

EJEMPLO 10.5. Información de los clientes de un supermercado sobre el precio (contraste de hipótesis utilizando proporciones)

Una empresa de estudios de mercado quiere saber si los compradores son sensibles a los precios de los artículos que se venden en un supermercado. Obtiene una muestra aleatoria de 802 compradores y observa que 378 son capaces de decir el precio correcto de un artículo inmediatamente después de colocarlo en el carro. Contraste al nivel del 7 por ciento la hipótesis nula de que al menos la mitad de todos los compradores son capaces de decir el precio correcto.

Solución

Sea P la proporción poblacional de compradores de los supermercados que son capaces de decir el precio correcto en estas circunstancias. Contraste la hipótesis nula

$$H_0 : P \geqslant P_0 = 0{,}50$$

frente a la alternativa

$$H_1 : P < 0{,}50$$

La regla de decisión es rechazar la hipótesis nula en favor de la alternativa si

$$\frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}} < -z_\alpha$$

En este ejemplo, $n = 802$ y $\hat{p} = 378/802 = 0{,}471$. En un contraste al nivel del 7 por ciento, $\alpha = 0{,}07$ y $z_\alpha = -1{,}474$, según la tabla de distribución normal.

El estadístico del contraste es

$$\frac{\hat{p} - P_0}{\sqrt{P_0(1 - P_0)/n}} = \frac{0{,}471 - 0{,}50}{\sqrt{0{,}50(1 - 0{,}50)/802}} = -1{,}64$$

Dado que $-1{,}64$ es menor que $-1{,}474$, rechazamos la hipótesis nula y concluimos que menos de la mitad de los compradores puede decir correctamente el precio inmediatamente después de colocar un artículo en el carro. Utilizando el valor del estadístico del contraste calculado de $-1{,}64$, también observamos que el p-valor del contraste es 0,051.

EJERCICIOS

Ejercicios básicos

10.28. Se obtiene una muestra aleatoria de mujeres y se pregunta a cada una de ellas si compraría un nuevo modelo de zapatos. Para averiguar si las ventas de este nuevo modelo llegarían a superar la cifra del 25 por ciento para cumplir así los objetivos de beneficios de la empresa, se realiza el siguiente contraste de hipótesis al nivel $\alpha = 0{,}03$ utilizando la proporción muestral de mujeres que contestaron afirmativamente, $\hat{p}$.

$$H_0 : P \leqslant 0{,}25$$
$$H_1 : P > 0{,}25$$

¿Qué valor tiene que tener la proporción muestral, $\hat{p}$, para rechazar la hipótesis nula, dados los siguientes tamaños de la muestra?

a) $n = 400$
b) $n = 225$
c) $n = 625$
d) $n = 900$

10.29. Una empresa está tratando de averiguar si debe seguir fabricando un modelo de zapatos que antes tenía mucha aceptación. Se obtiene una muestra aleatoria de mujeres a las que se les pregunta si comprarían este modelo. Para averiguar si se debe seguir fabricando ese modelo, se realiza el siguiente contraste de hipótesis a un nivel $\alpha = 0{,}05$ utilizando la proporción muestral de mujeres que contestó afirmativamente, $\hat{p}$.

$$H_0 : P \geqslant 0{,}25$$
$$H_1 : P < 0{,}25$$

¿Qué valor debe tener la proporción muestral, $\hat{p}$, para rechazar la hipótesis nula, dados los siguietes tamaños de la muestra?

a) $n = 400$
b) $n = 225$
c) $n = 625$
d) $n = 900$

Ejercicios aplicados

10.30. En una muestra aleatoria de 361 propietarios de pequeñas empresas que se habían declarado en quiebra, 105 declararon que no habían hecho ningún estudio de mercado antes de abrir el negocio. Contraste al nivel $\alpha = 0{,}05$ la hipótesis de que el 25 por ciento como máximo de todos los miembros de esta población no realizó estudios de mercado antes de abrir el negocio.

10.31. En una muestra aleatoria de 998 adultos de Estados Unidos, el 17,3 por ciento de los miembros discrepa de la siguiente afirmación: «La globalización es más que un sistema comercial económico; incluye las instituciones y la cultura». Contraste al nivel del 5 por ciento la hipótesis de que al menos el 25 por ciento de todos los adultos estadounidenses discreparía de esta afirmación.

10.32. En una muestra aleatoria de 160 estudiantes de administración de empresas, 72 miembros se mostraron en alguna medida de acuerdo con la siguiente afirmación: «Las calificaciones de un examen de selectividad son menos importantes para las posibilidades de éxito académico de un estudiante que las calificaciones obtenidas en el bachillerato». Contraste la hipótesis nula de que la mitad de todos los estudiantes de administración de empresas estaría de acuerdo con esta afirmación frente a la hipótesis alternativa bilateral. Halle e interprete el p-valor del contraste.

10.33. En una muestra aleatoria de 199 auditores, 104 se mostraron en alguna medida de acuerdo con la siguiente afirmación: «El flujo de caja es un importante indicador de la rentabilidad». Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la mitad de los miembros de esta población estaría de acuerdo con esta afirmación frente a la alternativa bilateral. Halle e interprete también el p-valor de este contraste.

10.34. Se ha preguntado a una muestra aleatoria de 50 responsables de la admisión en programas de postgrado por lo que se espera en las entrevistas que se realizan a los solicitantes. En esta muestra aleatoria, 28 miembros estaban de acuerdo en que el entrevistador normalmente espera que el entrevistado haya realizado labores de voluntariado. Contraste al nivel $\alpha = 0{,}05$ la hipótesis nula de que la mitad de todos los entrevistadores tienen esta expectativa frente a la alternativa de que la proporción poblacional es de más de la mitad.

10.35. En una muestra aleatoria de 172 profesores de enseñanza primaria, 118 declararon que el apoyo de los padres era la fuente más importante de éxito de un niño. Contraste al nivel $\alpha = 0{,}05$ la hipótesis de que el apoyo de los padres es la fuente más importante de éxito de un niño al menos para el 75 por ciento de los profesores de enseñanza primaria frente a la alternativa de que el porcentaje poblacional es inferior al 75 por ciento.

10.36. Se ha preguntado a una muestra aleatoria de 202 profesores de una escuela de administración de empresas si debe exigirse a los estudiantes que asistan a un curso de lengua extranjera. En esta muestra, 140 miembros piensan que sí debe exigirse. Contraste al nivel $\alpha = 0{,}05$ la hipótesis de que al menos el 75 por ciento de todos los profesores defiende esta idea.

10.5. Valoración de la potencia de un contraste

En los apartados 10.2 a 10.4 hemos presentado varios contrastes de hipótesis con un nivel de significación α. En todos estos contrastes, hemos formulado reglas de decisión para rechazar la hipótesis nula en favor de una hipótesis alternativa. Cuando realizamos estos contrastes, sabemos que la probabilidad de cometer un error de Tipo I cuando rechazamos la hipótesis nula es como máximo igual a un determinado valor α que suele ser pequeño. Además, también podemos calcular el *p*-valor del contraste y, por lo tanto, sabemos cuál es el nivel mínimo de significación al que puede rechazarse la hipótesis nula. Cuando rechazamos la hipótesis nula, concluimos que existen pruebas contundentes para apoyar nuestra conclusión. Pero si no rechazamos la hipótesis nula, sabemos que la hipótesis nula es verdadera o hemos cometido un error de Tipo II al no rechazar la hipótesis nula cuando la alternativa es verdadera.

En este apartado examinamos las características de algunos de nuestros contrastes cuando la hipótesis nula no es verdadera. Aprendemos a calcular la probabilidad de cometer un error de Tipo II y a averiguar la potencia de un contraste de hipótesis. Naturalmente, sólo puede cometerse un error de Tipo II si la hipótesis alternativa es verdadera. Por lo tanto, consideraremos el error de Tipo II y la potencia que se dan cuando el parámetro poblacional adopta valores específicos que están incluidos en la hipótesis alternativa.

Contrastes de la media de una distribución normal: variable poblacional conocida

Siguiendo los métodos del apartado 10.2, queremos contrastar la hipótesis nula de que la media de una población normal es igual a un valor específico, μ_0.

Determinación de la probabilidad de cometer un error de Tipo II

Consideremos el contraste

$$H_0 : \mu = \mu_0$$

frente a la alternativa

$$H_1 : \mu > \mu_0$$

Utilizando la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha \quad \text{o} \quad \bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n}$$

hallamos los valores de la media muestral que llevan a no rechazar la hipótesis nula. Ahora bien, dado cualquier valor de la media poblacional definido por la hipótesis alternativa, H_1, hallamos la probabilidad de que la media muestral esté en la región de no rechazo de la hipótesis nula. Ésta es la probabilidad de cometer un error de Tipo II. Por lo tanto, consideramos una $\mu = \mu^*$ tal que $\mu^* > \mu_0$. Entonces para μ^* la probabilidad de cometer un error de Tipo II es

$$\begin{aligned} \beta &= P(\bar{x} < \bar{x}_c \,|\, \mu = \mu^*) \\ &= P\left(z < \frac{\bar{x}_c - \mu^*}{\sigma/\sqrt{n}}\right) \end{aligned} \qquad \textbf{(10.12)}$$

y

$$\text{Potencia} = 1 - \beta$$

El valor de β y la potencia serán diferentes para todo μ^*.

Consideremos un ejemplo en el que contrastamos la hipótesis nula de que la media poblacional del peso de los rodamientos de un proceso de producción es de 5 onzas frente a la hipótesis alternativa de que es de más de 5 onzas. Realizamos el contraste con una muestra aleatoria de 16 observaciones y un nivel de significación del 0,05. Se supone que la distribución poblacional es una distribución normal que tiene una desviación típica de 0,1 onzas. Por lo tanto, la hipótesis nula es

$$H_0 : \mu = 5$$

frente a la hipótesis alternativa

$$H_1 : \mu > 5$$

y la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - 5}{0{,}1/\sqrt{16}} > 1{,}645 \quad \text{o} \quad \bar{x} > 5 + 1{,}645(0{,}1/\sqrt{16}) = 5{,}041$$

Ahora bien, si la media muestral es menor o igual que 5,041, entonces, utilizando nuestra regla, no rechazaremos la hipótesis nula.

Supongamos que queremos hallar la probabilidad de que no se rechace la hipótesis nula si el verdadero peso medio es de 5,05 onzas. Es evidente que la hipótesis alternativa es correcta y queremos hallar la probabilidad de que no rechacemos la hipótesis nula y, por lo tanto, cometamos un error de Tipo II. Es decir, queremos hallar la probabilidad de que la media muestral sea de menos de 5,041 si la media poblacional es realmente 5,05. Utilizando las 16 observaciones, calculamos la probabilidad de cometer un error de Tipo II:

$$\begin{aligned}\beta &= P(\bar{X} \leqslant 5{,}041 \mid \mu = 5{,}05) = P\left(Z \leqslant \frac{5{,}041 - 5{,}05}{0{,}1/\sqrt{16}}\right) \\ &= P(Z \leqslant -0{,}36) \\ &= 1 - 0{,}6406 = 0{,}3594\end{aligned}$$

Por lo tanto, utilizando la regla de decisión anterior, podemos demostrar que la probabilidad, β, de cometer un error de Tipo II cuando la media poblacional es de 5,05 onzas es 0,3594. Dado que la potencia de un contraste es 1 menos la probabilidad de cometer un error de Tipo II, tenemos que cuando la media poblacional es 5,05,

$$\text{Potencia} = 1 - \beta = 1 - 0{,}3594 = 0{,}6406$$

Estos cálculos de la potencia se muestran en la Figura 10.5. En la parte (a) vemos que, cuando la media poblacional es 5, la probabilidad de que la media muestral sea superior a 5,041 es 0,05, que es el nivel de significación del contraste. La parte (b) de la figura muestra la función de densidad de la distribución de la media muestral en el muestreo cuando la media poblacional es 5,05. El área sombreada de esta figura muestra la probabilidad de que la media muestral sea superior a 5,041 cuando la media poblacional es 5,05: la poten-

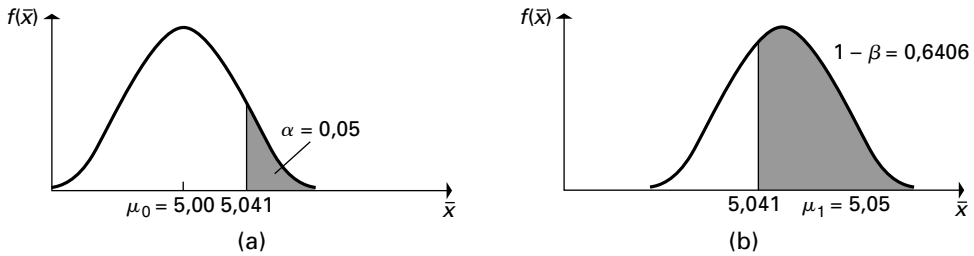


Figura 10.5. Distribución de la media muestral en el muestreo de 16 observaciones cuando $\sigma = 0{,}1$.

cia del contraste. Podrían realizarse unos cálculos similares para hallar la potencia y la probabilidad de cometer un error de Tipo II con cualquier valor de μ superior a 5,0.

Calculando la potencia de un contraste para todos los valores de μ incluidos en la hipótesis nula, puede generarse la función de potencia, mostrada en la Figura 10.6. La función de potencia tiene las siguientes características:

1. Cuanto más lejos se encuentra la verdadera media de la media postulada μ_0, mayor es la potencia del contraste, manteniéndose todo lo demás constante. La Figura 10.6 ilustra este resultado.
2. Cuanto menor es el nivel de significación (α) del contraste, menor es la potencia, manteniéndose todo lo demás constante. Por lo tanto, la reducción de la probabilidad de cometer un error de Tipo I (α) aumenta la probabilidad de cometer un error de Tipo II (β), pero la reducción de α en 0,01 no aumenta generalmente β en 0,01; los cambios no son lineales.
3. Cuanto mayor es la varianza poblacional, menor es la potencia del contraste, manteniéndose todo lo demás constante.
4. Cuanto mayor es el tamaño de la muestra, mayor es la potencia del contraste, manteniéndose todo lo demás constante. Obsérvese que las muestras de mayor tamaño reducen la varianza de la media poblacional y, por lo tanto, aumentan las posibilidades de que rechacemos H_0 cuando no es correcta. La Figura 10.7 presenta un conjunto de curvas de potencia correspondientes a los tamaños de la muestra de 4, 9 y 16 que ilustran el efecto.
5. La potencia del contraste al valor crítico es igual a 0,5 porque la probabilidad de que una media muestral sea superior a ($\mu_0 = \bar{x}_c$) es, por supuesto, 0,50.

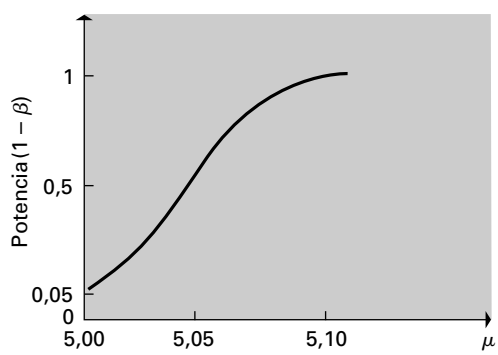


Figura 10.6. Función de potencia del contraste $H_0: \mu = 5$ frente a $H_1: \mu > 5$ ($\alpha = 0{,}05$, $\sigma = 0{,}1$, $n = 16$).

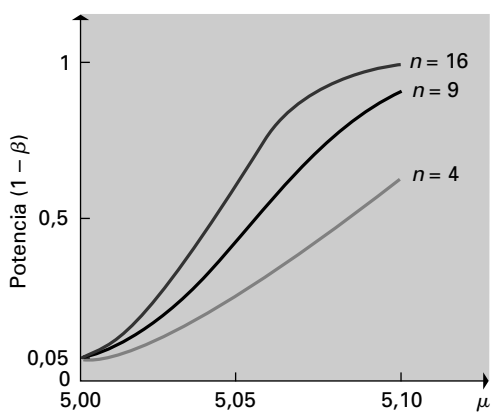


Figura 10.7. Funciones de potencia del contraste $H_0: \mu = 5$ frente a $H_1: \mu > 5$ ($\alpha = 0{,}05$, $\sigma = 0{,}1$) para los tamaños de la muestra 4, 9 y 16.

Muchos paquetes estadísticos tienen rutinas programadas que permiten calcular la potencia de un contraste. Por ejemplo, la Figura 10.8 muestra la salida Minitab del ejemplo analizado. Las pequeñas diferencias entre los valores de la potencia son el resultado del error de redondeo.

Figura 10.8. Cálculo de la potencia por computador (salida Minitab).

```
Power and Sample Size

1-Sample Z test

Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 0.1


                Sample
Difference        Size        Power
0.05                16     0.638760
```

```
Minitab steps
1. stat
2. Power and Sample Size
3. 1 Sample Z
4. Enter Sample Size 16
5. Difference 0.05
6. Standard Deviation 0.1.
7. Options Greater than
```

Potencia de los contrastes de proporciones poblacionales (grandes muestras)

En el apartado 10.4 hemos presentado contrastes de hipótesis y reglas de decisión para contrastar si la proporción poblacional tenía ciertos valores. Utilizando métodos parecidos a los del apartado anterior, también podemos hallar la probabilidad de cometer un error de Tipo II para los contrastes de proporciones. La probabilidad, β, de cometer un error de Tipo II dada una proporción poblacional P_1 incluida en H_1 se halla de la forma siguiente:

1. Partiendo de la regla de decisión del contraste, se halla el intervalo de valores de la proporción muestral que llevan a no rechazar la hipótesis nula.
2. Utilizando el valor P_1 de la proporción poblacional —donde P_1 está incluida en la hipótesis alternativa— se halla la probabilidad de que la proporción muestral esté en el intervalo de no rechazo hallado en el paso (1) para muestras de n observaciones cuando la proporción poblacional es P_1.

En el siguiente ejemplo mostramos cómo se utiliza este método.

EJEMPLO 10.6. Predicciones de los beneficios de Inversores Electrónicos S.A. (potencia y error de Tipo II)

El presidente de Inversores Electrónicos le ha pedido que analice las predicciones de los beneficios empresariales por acción realizadas por un grupo de analistas financieros. Estos analistas tenían interés en saber tanto cuál era la proporción de predicciones que eran superiores al nivel efectivo de beneficios como la proporción de predicciones que eran inferiores al nivel efectivo de beneficios.

Solución

Comencemos nuestro análisis construyendo un contraste de hipótesis para averiguar si existen pruebas contundentes que permitan concluir que la proporción de predicciones que son superiores a los beneficios efectivos es diferente del 50 por ciento. Representando por medio de P la proporción de predicciones superiores al nivel efectivo, la hipótesis nula es

$$H_0 : P = P_0 = 0{,}50$$

y la hipótesis alternativa es

$$H_1 : P \neq 0{,}50$$

La regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\hat{p}_x - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}} > z_{\alpha/2} \quad \text{o} \quad \frac{\hat{p}_x - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}} < -z_{\alpha/2}$$

Se obtiene una muestra aleatoria de $n = 600$ predicciones y se averigua que 382 son superiores a los beneficios efectivos. Utilizando un nivel de significación de $\alpha = 0{,}05$, la regla de decisión es rechazar la hipótesis nula si

$$\frac{\hat{p}_x - 0{,}50}{\sqrt{\dfrac{(0{,}50)(0{,}50)}{600}}} > 1{,}96 \quad \text{o} \quad \frac{\hat{p}_x - 0{,}50}{\sqrt{\dfrac{(0{,}50)(0{,}50)}{600}}} < -1{,}96$$

También se rechaza H_0 si

$$\hat{p}_x > 0{,}50 + 1{,}96\sqrt{\frac{(0{,}50)(0{,}50)}{600}} = 0{,}50 + 0{,}04 = 0{,}54$$

o sea

$$\hat{p}_x < 0{,}50 - 0{,}04 = 0{,}46$$

La proporción muestral observada es

$$\hat{p}_x = \frac{382}{600} = 0{,}637$$

y, por lo tanto, se rechaza la hipótesis nula al nivel del 5 por ciento.

Ahora queremos hallar la probabilidad de cometer un error de Tipo II cuando se utiliza esta regla de decisión. Supongamos que la verdadera proporción poblacional es $P_1 = 0{,}55$. Queremos hallar la probabilidad de que la proporción muestral se encuentre entre 0,46 y 0,54 si la proporción poblacional es 0,55. Por lo tanto, la probabilidad de cometer un error de Tipo II es

$$P(0{,}46 \leqslant \hat{p}_x \leqslant 0{,}54 \mid P = 0{,}55) = P\left[\frac{0{,}46 - P_1}{\sqrt{\dfrac{P_1(1-P_1)}{n}}} \leqslant Z \leqslant \frac{0{,}54 - P_1}{\sqrt{\dfrac{P_1(1-P_1)}{n}}}\right]$$

$$= P\left[\frac{0{,}46 - 0{,}55}{\sqrt{\dfrac{(0{,}55)(0{,}45)}{600}}} \leqslant Z \leqslant \frac{0{,}54 - 0{,}55}{\sqrt{\dfrac{0{,}55(0{,}45)}{600}}}\right]$$

$$= P(-4{,}43 \leqslant Z \leqslant -0{,}49) = 0{,}3121$$

Dada la regla de decisión, la probabilidad de cometer un error de Tipo II si no se rechaza la hipótesis nula cuando la verdadera proporción es 0,55 es $\beta = 0{,}3121$. La potencia del contraste con este valor de la proporción poblacional es

$$\text{Potencia} = 1 - \beta = 0{,}6879$$

Esta probabilidad puede calcularse para cualquier proporción P_1. La Figura 10.9 muestra la función de potencia de este ejemplo. Como la hipótesis alternativa es bilateral, la función de potencia tiene una forma distinta a la de la Figura 10.6. Aquí, estamos considerando valores posibles de la proporción poblacional a cualquiera de los dos lados del valor postulado, 0,50. Como vemos, la probabilidad de rechazar la hipótesis nula cuando es falsa aumenta cuanto más lejos esté la proporción poblacional verdadera del valor postulado.

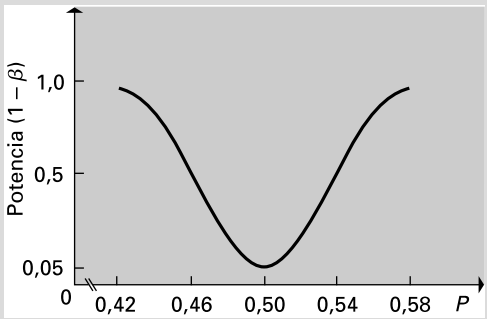


Figura 10.9. Función de potencia del contraste de $H_0: P = 0{,}50$ frente a $H_1: P \neq 0{,}50$ ($\alpha = 0{,}05$, $n = 600$).

EJERCICIOS

Ejercicios básicos

10.37. Considere el siguiente contraste de hipótesis

$$H_0: \mu = 5$$
$$H_1: \mu > 5$$

y la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \frac{x - 5}{0{,}1/\sqrt{16}} > 1{,}645 \quad \text{o}$$

$$\bar{x} > 5 + 1{,}645(0{,}1/\sqrt{16}) = 5{,}041$$

Calcule la probabilidad de cometer un error de Tipo II y la potencia en el caso de las medias poblacionales verdaderas.

a) $\mu = 5{,}10$
b) $\mu = 5{,}03$
c) $\mu = 5{,}15$
d) $\mu = 5{,}07$

10.38. Considere el ejemplo 10.6 en el que la hipótesis nula es

$$H_0: P = P_0 = 0{,}50$$

y la hipótesis alternativa es

$$H_0: P \neq 0{,}50$$

La regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\hat{p}_x - 0{,}50}{\sqrt{\frac{(0{,}50)(0{,}50)}{600}}} > 1{,}96 \quad \text{o}$$

$$\frac{\hat{p}_x - 0{,}50}{\sqrt{\frac{(0{,}50)(0{,}50)}{600}}} < -1{,}96$$

y el tamaño de la muestra es $n = 600$. Calcule la probabilidad de cometer un error de Tipo II si la proporción poblacional efectiva es

a) $P = 0{,}52$
b) $P = 0{,}58$
c) $P = 0{,}53$
d) $P = 0{,}48$
e) $P = 0{,}43$

Ejercicios aplicados

10.39. Una empresa que recibe envíos de pilas contrasta una muestra aleatoria de nueve de ellas antes de aceptar un envío. Quiere que la verdadera duración media de todas las pilas del envío sea al menos de 50 horas. Sabe por experiencia que la distribución poblacional de la duración es normal y tiene una desviación típica de 3 horas. La duración media de una muestra de nueve pilas de un envío es de 48,2 horas.

a) Contraste al nivel del 10 por ciento la hipótesis nula de que la media poblacional de la duración es al menos de 50 horas.

b) Halle la potencia de un contraste al nivel del 10 por ciento cuando la verdadera duración media de las pilas es de 49 horas.

10.40. Una empresa farmacéutica quiere que la concentración de impurezas de sus píldoras no supere el 3 por ciento. Se sabe que la concentración de impurezas de un lote sigue una distribución normal que tiene una desviación típica del 0,4 por ciento. Se comprueba una muestra aleatoria de 64 píldoras de un lote y se observa que la media muestral de la concentración de impurezas es del 3,07 por ciento.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional de la concentración de impurezas es del 3 por ciento frente a la alternativa de que es de más del 3 por ciento.

b) Halle la probabilidad de que un contraste rechace al nivel del 5 por ciento la hipótesis nula cuando la verdadera concentración media de impurezas es del 3,10 por ciento.

10.41. Una muestra aleatoria de 1.562 estudiantes universitarios matriculados en un curso de ética empresarial debe responder en una escala de 1 (totalmente en desacuerdo) a 7 (totalmente de acuerdo) a esta proposición: «A los altos ejecutivos de las empresas les preocupa la justicia social». La media muestral de las respuestas es 4,27 y la desviación típica muestral es 1,32.

a) Contraste al nivel del 1 por ciento la hipótesis nula de que la media poblacional es 4 frente a la hipótesis alternativa bilateral.

b) Halle la probabilidad de que un contraste acepte al nivel del 1 por ciento la hipótesis nula cuando la verdadera respuesta media es 3,95.

10.42. En una muestra aleatoria de 802 compradores en supermercados había 378 que preferían las marcas genéricas si su precio era más bajo. Contraste al nivel del 10 por ciento la hipótesis nula de que al menos la mitad de todos los compradores prefería las marcas genéricas frente a la alternativa de que la proporción poblacional es de menos de la mitad. Halle la potencia de un contraste al nivel del 10 por ciento si el 45 por ciento de los compradores es capaz realmente de indicar el precio correcto de un artículo inmediatamente después de colocarlo en el carro.

10.43. En una muestra aleatoria de 998 adultos de Estados Unidos, el 17,3 por ciento de los miembros discrepaba de la siguiente afirmación: «La globalización es más que un sistema comercial económico: incluye las instituciones y la cultura».

a) Contraste al nivel del 5 por ciento la hipótesis de que al menos el 25 por ciento de todos los adultos estadounidenses discreparía de esta afirmación.

b) Halle la probabilidad de rechazar la hipótesis nula con un contraste al nivel del 5 por ciento si el 20 por ciento de todos los adultos estadounidenses discrepara realmente de esta afirmación.

10.44. En una muestra aleatoria de 199 auditores, 104 se mostraron en alguna medida de acuerdo con la siguiente afirmación: «El flujo de caja es un importante indicador de la rentabilidad».

a) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la mitad de los miembros de esta población estaría de acuerdo con esta afirmación frente a la alternativa bilateral. Halle e interprete también el *p*-valor de este contraste.

b) Halle la probabilidad de aceptar la hipótesis nula con un contraste al nivel del 10 por ciento si el 60 por ciento de todos los auditores estuviera realmente de acuerdo en que el flujo de caja es un importante indicador de la rentabilidad.

10.45. Una cadena de comida rápida comprueba diariamente que el peso medio de sus hamburguesas es de 320 gramos como mínimo. La hipótesis alternativa es que el peso medio es de menos de 320 gramos, lo que indica que es necesario utilizar nuevos métodos. Puede suponerse que el peso de las hamburguesas sigue una

distribución normal que tiene una desviación típica de 30 gramos. La regla de decisión adoptada es rechazar la hipótesis nula si la media muestral del peso es de menos de 308 gramos.

a) Si se seleccionan muestras aleatorias de $n = 36$ hamburguesas, ¿cuál es la probabilidad de que se cometa un error de Tipo I utilizando esta regla de decisión?

b) Si se seleccionan muestras aleatorias de $n = 9$ hamburguesas, ¿cuál es la probabilidad de que se cometa un error de Tipo I utilizando esta regla de decisión? Explique por qué su respuesta es diferente de la respuesta del apartado (a).

c) Suponga que el verdadero peso medio es de 310 gramos. Si se seleccionan muestras aleatorias de 36 hamburguesas, ¿cuál es la probabilidad de que se cometa un error de Tipo II utilizando esta regla de decisión?

10.46. Un vinicultor sostiene que la proporción de clientes que no saben distinguir su producto del zumo de uva congelada es como máximo de 0,10. Decide contrastar esta hipótesis nula frente a la alternativa de que la verdadera proporción es de más de 0,10. La regla de decisión adoptada es rechazar la hipótesis nula si la proporción muestral que no sabe distinguir entre los dos sabores es de más de 0,14.

a) Si se elige una muestra aleatoria de 100 clientes, ¿cuál es la probabilidad de que se cometa un error de Tipo I utilizando esta regla de decisión?

b) Si se elige una muestra aleatoria de 400 clientes, ¿cuál es la probabilidad de que se cometa un error de Tipo I utilizando esta regla de decisión? Explique verbal y gráficamente por qué su respuesta es diferente de la respuesta del apartado (a).

c) Suponga que la verdadera proporción de clientes que no saben distinguir entre estos sabores es de 0,20. Si se elige una muestra aleatoria de 100 clientes, ¿cuál es la probabilidad de que se cometa un error de Tipo II?

d) Suponga que, en lugar de utilizar la regla de decisión dada, se decide rechazar la hipótesis nula si la proporción muestral de clientes que no saben distinguir entre los dos sabores es de más de 0,16. Se selecciona una muestra aleatoria de 100 clientes.

i. Indique sin realizar los cálculos si la probabilidad de cometer un error de Tipo I será mayor, menor o igual que en el apartado (a).

ii. Si la verdadera proporción es 0,20, ¿será la probabilidad de cometer un error de Tipo II mayor, menor o igual que en el apartado (c)?

RESUMEN

En este capítulo hemos presentado la metodología para realizar contrastes clásicos de hipótesis, comenzando con los argumentos para tomar decisiones en condiciones de incertidumbre. Se definen decisiones que implican la elección entre dos opciones. Las decisiones se toman rechazando una hipótesis nula si hay pruebas contundentes a favor de la hipótesis alternativa. Pueden cometerse dos tipos de error: un error de Tipo I, que se comete cuando se rechaza la hipótesis nula cuando es verdadera, y un error de Tipo II, que se comete cuando no se rechaza la hipótesis nula cuando no es verdadera. Hemos presentado diversos métodos y reglas de decisión específicos para realizar contrastes. Son contrastes de la media cuando las varianzas son conocidas y desconocidas y contrastes de proporciones. Hemos analizado los métodos para hallar la potencia y la probabilidad de cometer un error de Tipo II partiendo de diferentes supuestos sobre la media o la proporción poblacionales efectivas.

Las reglas de decisión se resumen en las Figuras 10.10 y 10.11. En la 10.10, se presentan reglas de decisión para contrastar hipótesis relacionadas con una media poblacional, μ. Obsérvese que se examinan contrastes de los tres tipos de hipótesis y de los casos en los que se conoce y se desconoce la variable poblacional. En la 10.11, se formulan reglas de decisión para contrastar hipótesis relacionadas con una proporción poblacional, P. Obsérvese de nuevo que se examinan contrastes de tres tipos de hipótesis.

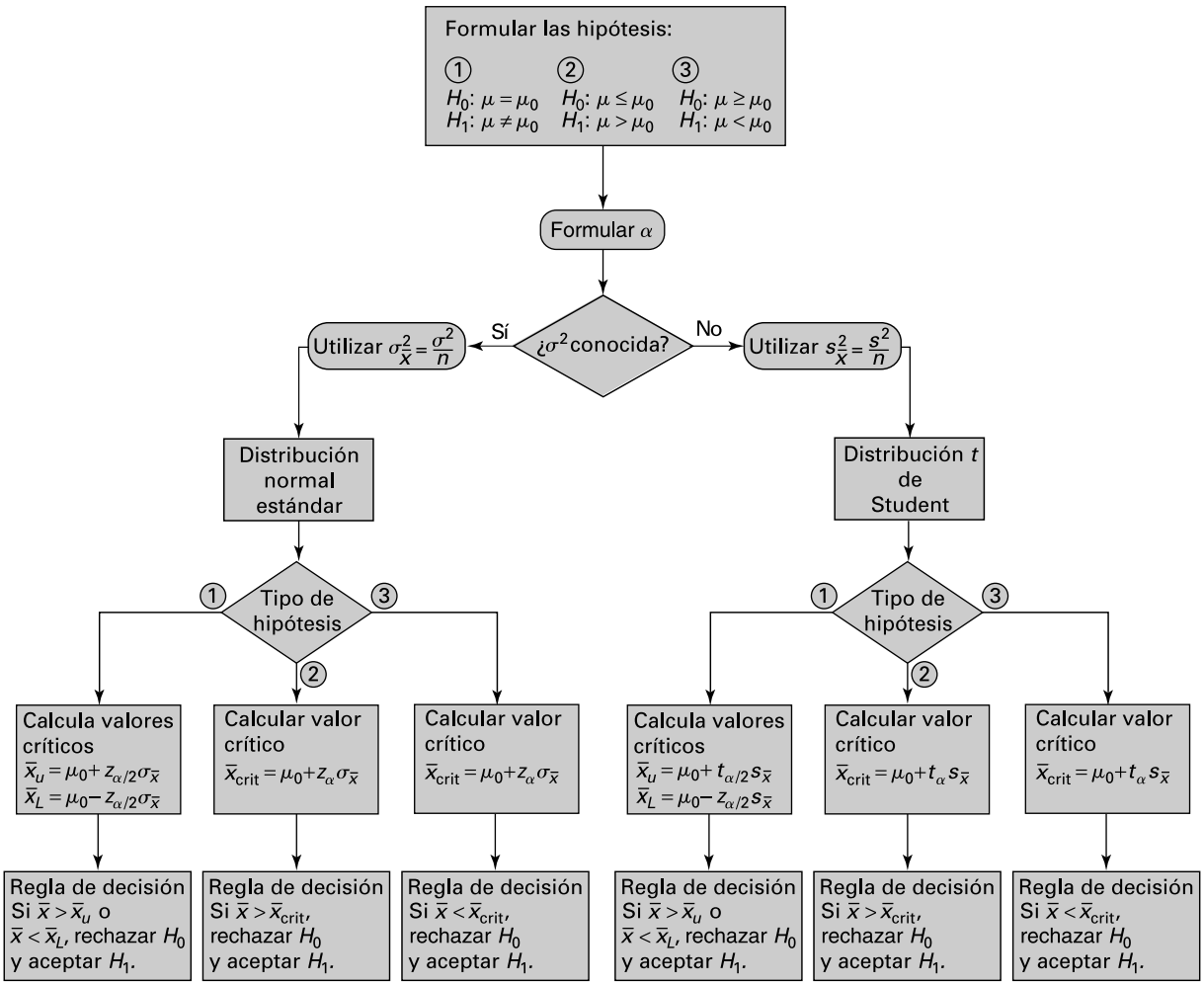


Figura 10.10. Directrices para elegir la regla de decisión adecuada para una media poblacional.

TÉRMINOS CLAVE

argumento contrafactual, 359
contraste de la media de una distribución normal (varianza conocida): hipótesis nula y alternativa compuestas, 365
contraste de la media de una distribución normal (varianza conocida): hipótesis nula y alternativa compuestas o simples, 367
contraste de la media de una población normal: varianza conocida, 362
contraste de la media de una distribución normal frente a una hipótesis alternativa bilateral: varianza conocida, 369
contrastes de la media de una distribución normal: varianza poblacional desconocida, 372
contrastes de la proporción poblacional (grandes muestras), 377
determinación de la probabilidad de cometer un error de Tipo II, 380
error de Tipo I, 356
error de Tipo II, 356
estados de la naturaleza y decisiones sobre la hipótesis nula, 356
función de potencia, 383
interpretación del valor de la probabilidad o *p*-valor, 363
hipótesis alternativa, 354
hipótesis compuesta, 354
hipótesis nula, 354
hipótesis simple, 354
potencia, 357
terminología del contraste de hipótesis, 358
valor crítico, 361

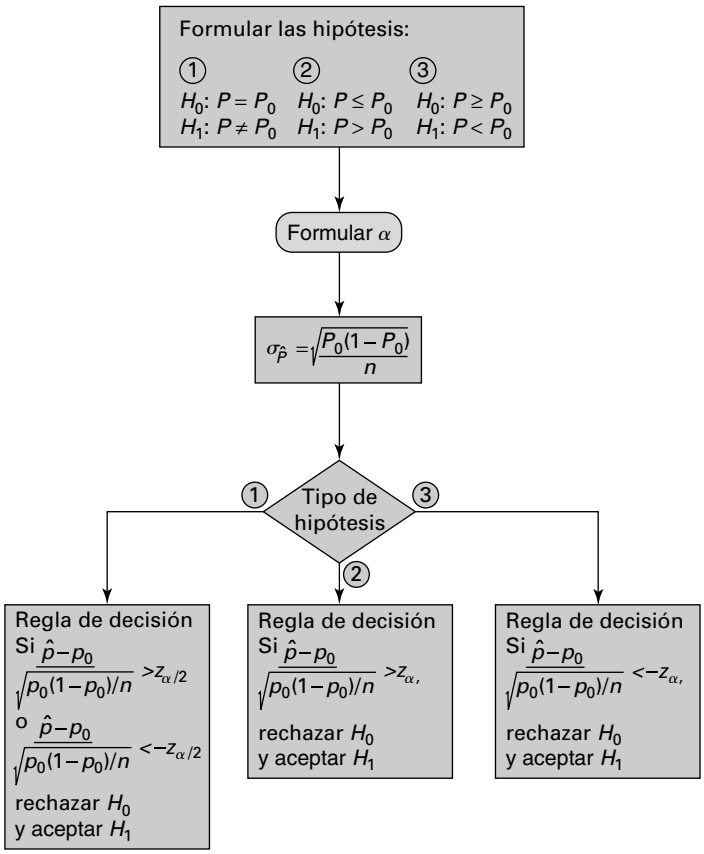


Figura 10.11. Directrices para elegir la regla de decisión adecuada para una proporción poblacional.

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

10.47. Explique detenidamente la distinción entre cada uno de los pares de términos siguientes:

a) Hipótesis nula e hipótesis alternativa
b) Hipótesis simple e hipótesis compuesta
c) Alternativa unilateral y bilateral
d) Errores de Tipo I y de Tipo II
e) Nivel de significación y potencia

10.48. Explique detenidamente qué significa el *p*-valor de un contraste y analice el uso de este concepto en el contraste de hipótesis.

10.49. Una muestra aleatoria de 10 estudiantes contiene las siguientes observaciones en horas sobre el tiempo que dedican a estudiar durante la semana antes de los exámenes finales.

28 57 42 35 61 39 55 46 49 38

Suponga que la distribución poblacional es normal.

a) Halle la media muestral y la desviación típica muestral.
b) Contraste al nivel de significación del 5 por ciento la hipótesis nula de que la media poblacional es 40 horas frente a la alternativa de que es mayor.

10.50. Indique si cada una de las afirmaciones siguientes es verdadera o falsa.

a) El nivel de significación de un contraste es la probabilidad de que la hipótesis nula sea falsa.
b) Se comete un error de Tipo I cuando se rechaza una hipótesis nula verdadera.
c) Se rechaza una hipótesis nula al nivel de 0,025, pero no se rechaza al nivel de 0,01. Eso significa que el *p*-valor del contraste se encuentra entre 0,01 y 0,025.
d) La potencia de un contraste es la probabilidad de aceptar una hipótesis nula que es verdadera.
e) Si se rechaza una hipótesis nula frente a una alternativa al nivel del 5 por ciento, entonces

utilizando los mismos datos debe rechazarse frente a la alternativa al nivel del 1 por ciento.

f) Si se rechaza una hipótesis nula frente a una alternativa al nivel del 1 por ciento, entonces utilizando los mismos datos debe rechazarse frente a la alternativa al nivel del 5 por ciento.

g) El *p*-valor de un contraste es la probabilidad de que la hipótesis nula sea verdadera.

10.51. Una compañía de seguros tiene agentes a comisión. Sostiene que el primer año de trabajo los agentes perciben una comisión media de 40.000 $ como mínimo y que la desviación típica poblacional no supera los 6.000 $. Considerando la comisión percibida el primer año, se observa que en una muestra aleatoria de nueve agentes,

$$\sum_{i=1}^{9} x_i = 333 \quad \text{y} \quad \sum_{i=1}^{9} (x_i - \bar{x})^2 = 312$$

donde x_i se expresa en miles de dólares y puede suponerse que la distribución de la población es normal. Contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional es de 40.000 $ como mínimo.

10.52. Los defensores de un nuevo molino de viento afirman que puede generar como mínimo una media de 800 kilovatios diarios de energía. Se supone que la generación diaria de energía sigue una distribución normal que tiene una desviación típica de 120 kilovatios. Se toma una muestra aleatoria de 100 días para contrastar esta afirmación frente a la hipótesis alternativa de que la verdadera media es de menos de 800 kilovatios. La afirmación no se rechaza si la media muestral es de 776 kilovatios o más y se rechaza en caso contrario.

a) ¿Cuál es la probabilidad α de que se cometa un error de Tipo I utilizando la regla de decisión si la media poblacional es, en realidad, de 800 kilovatios diarios?

b) ¿Cuál es la probabilidad β de que se cometa un error de Tipo II utilizando la regla de decisión si la media poblacional es, en realidad, de 740 kilovatios diarios?

c) Suponga que se utiliza la misma regla de decisión, pero con una muestra de 200 días en lugar de 100.

i. ¿Sería el valor de α mayor, menor o igual que el obtenido en el apartado (a)?

ii. ¿Sería el valor de β mayor, menor o igual que el obtenido en el apartado (b)?

d) Suponga que se toma una muestra de 100 observaciones, pero que se cambia la regla de decisión, de manera que la afirmación no se rechaza si la media muestral es de al menos 765 kilovatios.

i. ¿Sería el valor de α mayor, menor o igual que el obtenido en el apartado (a)?

ii. ¿Sería el valor de β mayor, menor o igual que el obtenido en el apartado (b)?

10.53. En una muestra aleatoria de 545 contables dedicados a elaborar presupuestos municipales, 117 indicaron que la tarea más difícil era estimar el flujo de caja.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que al menos el 25 por ciento de todos los contables considera que la tarea más difícil es estimar el flujo de caja.

b) Basándose en el método utilizado en el apartado (a), calcule la probabilidad de que la hipótesis nula se rechace si el verdadero porcentaje de contables que consideran que la tarea más difícil es estimar el flujo de caja es del

i. 20 por ciento

ii. 25 por ciento

iii. 30 por ciento

10.54. En una ocasión se preguntó a una muestra aleatoria de 104 vicepresidentes de marketing de grandes empresas de la lista de 500 empresas de la revista *Fortune* por la futura situación del clima empresarial. De los miembros de la muestra, 50 declararon que estaban de acuerdo en alguna medida con la siguiente afirmación: «Las empresas concentrarán sus esfuerzos en el flujo de caja más que en los beneficios». ¿Cuál es el nivel de significación más bajo al que puede rechazarse la hipótesis nula, según la cual la verdadera proporción de ejecutivos que estaría de acuerdo con esta afirmación es la mitad, frente a la hipótesis alternativa bilateral?

10.55. En una muestra aleatoria de 99 partidos de la liga profesional de béisbol, el equipo de casa ganó 57 partidos. Contraste la hipótesis nula de que el equipo de casa gana la mitad de todos los partidos frente a la hipótesis alternativa de que gana la mayoría.

10.56. En una muestra aleatoria de 150 licenciados en administración de empresas, 50 estaban de acuerdo o muy de acuerdo en que las empresas deben concentrar sus esfuerzos en buscar estrategias innovadoras de comercio electrónico.

Contraste al nivel del 5 por ciento la hipótesis nula de que el 25 por ciento como máximo de todos los licenciados en administración de empresas estaría de acuerdo con esta afirmación.

10.57. En una muestra aleatoria de 142 responsables de la admisión de estudiantes en programas de postgrado, 39 declararon que dedican en promedio 15 minutos o menos a estudiar cada solicitud. Contraste la hipótesis nula de que el 20 por ciento a lo más de todos los responsables dedican tan poco tiempo a estudiar las solicitudes.

10.58. Franquicias Nororientales, S.A., tiene algunos clientes que utilizan su proceso para producir cenas noruegas exóticas para clientes de todo el mundo. El coste de explotación del proceso franquiciado tiene un coste fijo de 1.000 $ a la semana más 5 $ por cada unidad producida. Recientemente, algunos dueños de restaurantes que utilizan el proceso se han quejado de que el modelo de costes ya no es válido y de que los costes semanales son, en realidad, más altos. Su trabajo es averiguar si existen pruebas contundentes que apoyen la afirmación de los dueños de los restaurantes. Obtiene una muestra aleatoria de $n = 25$ restaurantes y averigua sus costes. También sabe que el número de unidades producidas en cada restaurante sigue una distribución normal de media $\mu = 400$ y varianza $\sigma^2 = 625$. La media de los costes semanales obtenida con la muestra aleatoria ($n = 25$) es de 3.050 $. Elabore y aplique un análisis para averiguar si existen pruebas contundentes que permitan concluir que los costes son mayores de lo que predice el modelo de costes.

10.59. Prairie Flower Cereal Inc. le ha pedido que estudie la variabilidad del peso de las cajas de cereales producidas en la planta 2 que se encuentra en una zona rural de Malasia. Se sabe que el peso de las cajas sigue una distribución normal. Utilizando una muestra aleatoria de tamaño $n = 71$, observa que la media muestral del peso es 40 y la varianza muestral es 50.

El vicepresidente de marketing sostiene que existe una probabilidad muy pequeña de que la media poblacional del peso sea de menos de 39. Utilizando un análisis estadístico adecuado, comente su afirmación.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

10.60. Se pide a dos analistas financieros que predigan los beneficios por acción que tendrá una muestra aleatoria de 12 empresas el próximo año. Para evaluar la calidad de sus predicciones se utiliza como indicador el error porcentual absoluto de predicción, que se define de la forma siguiente:

$$100 \cdot \left(\frac{|\text{Efectivos} - \text{Predichos}|}{\text{Efectivos}} \right)$$

Los errores porcentuales absolutos de predicción cometidos se encuentran en el fichero de datos **Analyst Prediction**. Indique los supuestos que postule y contraste la hipótesis nula de que la media poblacional de los errores porcentuales absolutos de predicción de los dos analistas financieros es la misma frente a la hipótesis alternativa bilateral.

10.61. BBW Ltd. controla la calidad de las barras de pan que produce. El fichero de datos llamado **BBWltd**, que se encuentra en su disco de datos o en el sistema informático local, contiene datos recogidos como parte de su análisis del mercado. Las variables del fichero son:

1. «Dbread», que contiene una muestra aleatoria del peso en gramos de su pan negro recogida en los expositores de los supermercados.
2. «Sbread», que contiene una muestra aleatoria del peso en gramos de su pan especial recogida en los expositores de los supermercados.
3. «Csbread», que contiene una muestra aleatoria del peso en gramos del pan especial de su competidor recogida en los expositores de los supermercados.

La empresa garantiza que su pan negro tiene un peso de 100 gramos o más. Basándose en la muestra, ¿tiene la empresa pruebas contundentes, $\alpha = 0{,}05$, de que cumple la garantía? Aporte como prueba un resultado basado en un contraste de hipótesis.

10.62. Big River Inc., importante transformador de pescado de Alaska, está intentando averiguar el peso del salmón del río Northwest Green. Se obtiene una muestra aleatoria de salmón y se pesa. Los datos se encuentran en el fichero titulado **Bigfish**. Utilice un contraste clásico de hipótesis para averiguar si existen pruebas contundentes que permitan concluir que la media poblacional del peso del pescado es superior a 40. Utilice una probabilidad de cometer un error de Tipo I igual a 0,05.

Trace una curva de potencia del contraste. *Pista*: halle los valores de la media poblacional correspondientes a $\beta = 0{,}50$, $\beta = 0{,}25$, $\beta = 0{,}10$ y $\beta = 0{,}05$ y represente esas medias en relación con la potencia del contraste.

Capítulo 11

Contraste de hipótesis II

Esquema del capítulo

Introducción

En este capítulo presentamos métodos para contrastar las diferencias entre las medias o proporciones de dos poblacionales y para contrastar varianzas. Este tipo de inferencia contrasta con los métodos de estimación presentados en el Capítulo 9 y los complementa. El análisis de este capítulo es paralelo al del Capítulo 10 y se supone que el lector está familiarizado con el método para contrastar hipótesis desarrollado en el apartado 10.1. El proceso para comparar dos poblaciones comienza con la formulación de una hipótesis sobre la naturaleza de las dos poblaciones y la diferencia entre sus medias o proporciones. La formulación de la hipótesis implica claramente la elección entre dos opciones sobre la diferencia; a continuación, se toma una decisión basándose en los resultados de un estadístico calculado a partir de muestras aleatorias de datos de las dos poblaciones. Los contrastes de hipótesis relativos a las varianzas son cada vez más importantes, ya que las empresas tratan de reducir la variabilidad de los procesos con el fin de garantizar que todas las unidades producidas son de alta calidad. He aquí dos ejemplos de problemas representativos.

1. Un profesor tiene interés en saber si las calificaciones que obtienen sus estudiantes en los exámenes mejoran cuando da trabajos para realizar en casa. Podría poner trabajos para casa a un grupo y a otro no. En ese caso, recogiendo datos de las dos clases, podría averiguar si existen pruebas contundentes de que las calificaciones mejoran cuando pone trabajos para casa.

Supongamos que el profesor supone que la realización de trabajos en casa no aumenta la calificación total. Sea μ_1 la calificación media del examen final en la clase en la que el profesor da trabajos para casa y μ_2 la calificación media del examen final en la clase en la que no da trabajos para casa. La hipótesis nula es la hipótesis compuesta

$$H_0: \mu_1 - \mu_2 \leqslant 0$$

La alternativa de interés es que la realización de trabajos en casa aumenta realmente la calificación media y, por lo tanto, la hipótesis alternativa es

$$H_1: \mu_1 - \mu_2 > 0$$

En este problema, el profesor decidiría dar trabajos para casa sólo si existen pruebas contundentes de que eso mejora la calificación media de los exámenes. El rechazo de H_0 y la aceptación de H_1 es una prueba contundente.

2. Un periodista quiere saber si una reforma tributaria atrae de la misma forma a los hombres que a las mujeres. Para contrastarlo, recaba la opinión de una muestra aleatoria de hombres y mujeres y utiliza estos datos para obtener una respuesta. El periodista podría afirmar, como hipótesis de trabajo, que una nueva propuesta tributaria atrae por igual a los hombres y a las mujeres. Si P_1 es la proporción de hombres que defienden la propuesta y P_2 es la proporción de mujeres que defienden la propuesta, la hipótesis nula es

$$H_0: P_1 - P_2 = 0$$

Si el periodista no tiene ninguna razón de peso para sospechar que el apoyo a la propuesta proviene principalmente de los hombres o de las mujeres, contrastaría esta hipótesis nula frente a la hipótesis alternativa compuesta bilateral

$$H_1: P_1 - P_2 \neq 0$$

En este ejemplo, el rechazo de H_0 sería una prueba contundente de que hay una diferencia entre los hombres y las mujeres en su respuesta a la propuesta tributaria.

Una vez especificada la hipótesis nula y la hipótesis alternativa y una vez recogidos datos muestrales, debe tomarse una decisión sobre la hipótesis nula. Se puede rechazar y aceptar la hipótesis alternativa o no rechazar la hipótesis nula. Cuando no se rechaza la hipótesis nula, o bien es verdadera, o bien nuestro método para realizar el contraste no es lo suficientemente fuerte para rechazarla y se ha cometido un error. Para rechazar la hipótesis nula hay que formular una regla de decisión basada en evidencia muestral. Más adelante en este capítulo, presentamos reglas de decisión específicas para varios problemas.

11.1. Contrastes de la diferencia entre dos medias poblacionales

Existen algunas aplicaciones en las que queremos extraer conclusiones sobre las diferencias entre medias poblacionales en lugar de conclusiones sobre los niveles absolutos de las medias. Por ejemplo, podemos querer comparar la producción de dos procesos diferentes cuyas medias poblacionales no se conocen. También podemos querer saber si una estrategia de marketing aumenta las ventas más que otra sin conocer la media poblacional de las ventas de ninguna de las dos. Estas cuestiones pueden abordarse eficazmente mediante

algunos métodos de contraste de hipótesis. Como vimos en el apartado 9.1, cuando se calculan intervalos de confianza de las diferencias entre dos medias poblacionales, pueden postularse varios supuestos. Estos supuestos llevan generalmente a utilizar métodos específicos para calcular la varianza poblacional de la diferencia entre medias muestrales. Hay contrastes de hipótesis paralelos que implican la utilización de métodos similares para calcular la varianza. Nuestro análisis de los distintos métodos para contrastar hipótesis es paralelo a las estimaciones de los intervalos de confianza del apartado 9.1.

Dos medias, datos pareados

Aquí suponemos que se obtiene una muestra aleatoria de n pares de observaciones enlazadas procedentes de poblaciones que tienen las medias μ_x y μ_y. Las observaciones se representan de la forma siguiente: (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Cuando tenemos observaciones pareadas y los pares están correlacionados positivamente, la varianza de la diferencia entre las medias muestrales

$$\bar{d} = \bar{x} - \bar{y}$$

es menor que cuando se utilizan muestras independientes, debido a que algunas de las características de los pares son similares y, por lo tanto, esa parte de la variabilidad desaparece de la variabilidad total de las diferencias entre las medias. Por ejemplo, cuando examinamos medidas de la conducta humana, las diferencias entre los gemelos normalmente son menores que las diferencias entre dos personas seleccionadas aleatoriamente. En general, las dimensiones de dos piezas producidas en la misma máquina son más parecidas que las dimensiones de las piezas producidas en dos máquinas diferentes seleccionadas aleatoriamente. Por lo tanto, siempre que sea posible, preferiríamos utilizar observaciones pareadas cuando comparemos dos poblaciones porque la varianza de la diferencia es menor. Al ser menor, es mayor la probabilidad de que rechacemos H_0 cuando la hipótesis nula no es verdadera. Este principio se formuló en el apartado 10.5 cuando se analizó la potencia de un contraste. Las reglas de decisión específicas de diferentes tipos de contraste de hipótesis se resumen en las ecuaciones 11.1, 11.2 y 11.3.

Contrastes de la diferencia entre medias poblacionales: datos pareados

Supongamos que tenemos una muestra aleatoria de n pares de observaciones enlazadas de distribuciones que tienen las medias μ_x y μ_y. Sean $\bar{d}$ y s_d la media muestral y la desviación típica muestral observadas de las n diferencias $(x_i - y_i)$. Si la distribución poblacional de las diferencias es una distribución normal, los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \leqslant D_0$$

frente a la hipótesis alternativa

$$H_1 : \mu_x - \mu_y > D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{d} - D_0}{s_d/\sqrt{n}} > t_{n-1,\alpha} \qquad \textbf{(11.1)}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0: \mu_x - \mu_y \geqslant D_0$$

frente a la hipótesis alternativa

$$H_1: \mu_x - \mu_y < D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{d} - D_0}{s_d/\sqrt{n}} < -t_{n-1,\alpha} \tag{11.2}$$

3. Para contrastar la hipótesis nula

$$H_0: \mu_x - \mu_y = D_0$$

frente a la hipótesis alternativa bilateral

$$H_1: \mu_x - \mu_y \neq D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{d} - D_0}{s_d/\sqrt{n}} < t_{n-1,\alpha/2} \quad \text{o} \quad \frac{\bar{d} - D_0}{s_d/\sqrt{n}} > t_{n-1,\alpha/2} \tag{11.3}$$

Aquí, $t_{n-1,\alpha}$ es el número para el que

$$P(t_{n-1} > t_{n-1,\alpha}) = \alpha$$

donde la variable aleatoria t_{n-1} sigue una distribución t de Student con $(n-1)$ grados de libertad.

Cuando queremos contrastar la hipótesis nula de que las dos medias poblacionales son iguales, igualamos D_0 a 0 en las fórmulas.

Los p-valores de estos contrastes son la probabilidad de obtener un valor al menos tan extremo con el obtenido, dada la hipótesis nula.

EJEMPLO 11.1. Actividad cerebral y recuerdo de la publicidad televisiva

Unos investigadores realizaron un estudio para estimar la relación entre la actividad cerebral de un sujeto mientras veía un anuncio de televisión y su capacidad posterior para recordar su contenido. Se mostró a los sujetos dos anuncios comerciales de 10 productos. Se midió la capacidad para recordar cada anuncio 24 horas después y se llamó a cada miembro de un par de anuncios comerciales vistos por un sujeto específico «bien recordado» o «mal recordado». La Tabla 11.1 muestra un índice de la cantidad total de actividad cerebral de la muestra aleatoria de sujetos mientras veían estos anuncios. Los investigadores querían saber si la actividad de las ondas cerebrales era mayor en el caso de los anuncios bien recordados que en el de los anuncios mal recordados.

Solución

Sea μ_x la media poblacional de los anuncios bien recordados y μ_y la media poblacional de los anuncios mal recordados. Entonces, las diferencias d_i $(i = 1, ..., 10)$ son una muestra aleatoria de 10 observaciones procedentes de una población que tiene una me-

Tabla 11.1. Actividad cerebral de los sujetos que ven 10 pares de anuncios de televisión.

Observación del producto	*X* bien recordado	*Y* mal recordado
1	141	55
2	139	116
3	87	83
4	129	88
5	51	36
6	50	68
7	118	91
8	161	115
9	61	90
10	148	113

dia ($\mu_x - \mu_y$). Partiendo de estos supuestos, podemos contrastar la hipótesis nula de que no existe ninguna diferencia entre los niveles de actividad del cerebro

$$H_0 : \mu_x - \mu_y = 0$$

frente a la alternativa de que la actividad cerebral es, en promedio, mayor en el caso de los anuncios bien recordados; es decir,

$$H_1 : \mu_x - \mu_y > 0$$

Response to Commercials

En este contraste, calculamos la desviación típica muestral de las diferencias y, por lo tanto, utilizamos la distribución *t* de Student para realizar el contraste.

La pauta de los datos pareados se encuentra en la Tabla 11.1 y en el fichero de datos **Response to Commercials**. A cada sujeto se le asignó un anuncio bien recordado y uno mal recordado y los datos se enlazaron por medio del número de observación. La Figura 11.1 muestra la salida Minitab de este problema. El contraste se basa en el estadístico

$$t = \frac{\bar{d} - D_0}{s_d/\sqrt{n}} = \frac{23}{33{,}0/\sqrt{10}} = 2{,}21$$

Vemos en la Tabla 8 del apéndice que el valor $t_{9,0,05} = 1{,}833$. Como 2,21 es mayor que este valor, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Por lo tanto, llegamos a la conclusión de que existen considerables pruebas para concluir que la actividad cerebral es mayor en el caso de los anuncios bien recordados que en el de los

Paired T-Test and CI: X, Y

```
Paired T for X - Y

                N       Mean      St Dev     SE Mean
X              10    108.500     42.506      13.441
Y              10     85.500     26.471       8.371
Difference     10    23.0000    32.9848     10.4307
95% lower bound for mean difference: 3.8793
T-Test of mean difference = 0 (vs > 0): T-Value = 2.21  P-Value = 0.027
```

Minitab Instructions
1. Stat
2. Basic statistics
3. Paired *t*

Figura 11.1. Contraste de la hipótesis de las diferencias entre ondas cerebrales (salida Minitab).

mal recordados. Observamos que el p-valor de este contraste es 0,027, como muestra la salida Minitab.

Por último, debemos señalar que la existencia de datos perdidos es un problema que suele plantearse en los estudios estadísticos aplicados. Supongamos, por ejemplo, que la medición de las ondas cerebrales se perdiera en el caso de uno de los dos anuncios vistos por un sujeto. Normalmente, se eliminaría toda la observación y se realizaría el análisis con nueve observaciones pareadas.

Dos medias, muestras independientes, varianzas poblacionales conocidas

A continuación, analizamos el caso en el que tenemos muestras aleatorias independientes procedentes de dos poblaciones que siguen una distribución normal. La primera población tiene una media μ_x y una varianza σ_x^2 y obtenemos una muestra aleatoria de tamaño n_x. La segunda población tiene una media μ_y y una varianza σ_y^2 y obtenemos una muestra aleatoria de tamaño n_y.

En el apartado 9.1 demostramos que si las medias muestrales son $\bar{x}$ e $\bar{y}$, la variable aleatoria

$$Z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}}$$

sigue una distribución normal estándar. Si se conocen las dos varianzas poblacionales, los contrastes de la diferencia entre las medias poblacionales pueden basarse en este resultado, utilizando los mismos argumentos que antes. Generalmente, nos conformamos con utilizar varianzas poblacionales conocidas si el proceso estudiado se ha mantenido estable durante un tiempo y hemos obtenido mediciones similares de la varianza durante este tiempo. Y como consecuencia del teorema del límite central, los resultados presentados aquí son válidos cuando las muestras son de gran tamaño aunque las poblaciones no sean normales. Cuando las muestras son de gran tamaño, la aproximación es bastante satisfactoria cuando se utilizan las varianzas muestrales en lugar de las varianzas poblacionales. Naturalmente, también podemos realizar un contraste de hipótesis de la varianza, como se muestra en el apartado 11.3. Eso nos permite realizar contrastes que tienen numerosas aplicaciones y que se resumen en las ecuaciones 11.4, 11.5 y 11.6.

Contrastes de la diferencia entre medias poblacionales: muestras independientes (varianzas conocidas)

Supongamos que tenemos muestras aleatorias independientes de n_x y n_y observaciones procedentes de distribuciones normales que tienen las medias μ_x y μ_y y las varianzas σ_x^2 y σ_y^2, respectivamente. Si las medias muestrales observadas son $\bar{x}$ e $\bar{y}$, entonces los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \leqslant D_0$$

frente a la hipótesis alternativa

$$H_1 : \mu_x - \mu_y > D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}} > z_\alpha \tag{11.4}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \geqslant D_0$$

frente a la hipótesis alternativa

$$H_1 : \mu_x - \mu_y < D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}} < -z_\alpha \tag{11.5}$$

3. Para contrastar la hipótesis nula

$$H_0 : \mu_x - \mu_y = D_0$$

frente a la hipótesis alternativa bilateral

$$H_1 : \mu_x - \mu_y \neq D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}} < -z_{\alpha/2} \quad \text{o} \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}} > z_{\alpha/2} \tag{11.6}$$

Si los tamaños de las muestras son grandes ($n > 100$), puede obtenerse una buena aproximación al nivel de significación α si se sustituyen las varianzas poblacionales por las varianzas muestrales. Además, el teorema del límite central permite obtener buenas aproximaciones aunque las poblaciones no sigan una distribución normal.

Los *p*-valores de todos estos contrastes son la probabilidad de obtener un valor al menos tan extremo como el obtenido, dada la hipótesis nula.

EJEMPLO 11.2. Comparación de dos fertilizantes (contraste de hipótesis de diferencias entre medias)

Sara Briones, economista agraria, quiere comparar el uso de estiércol de vaca con el de pavo como fertilizantes. Históricamente, los agricultores han utilizado estiércol de vaca en los maizales. Recientemente, un importante criador de pavos vende el estiércol a un precio favorable. Los agricultores han decidido que sólo utilizarán este nuevo fertilizante

si existen pruebas contundentes de que la productividad es mayor que cuando se utiliza estiércol de pavo. Le han pedido a Sara que realice el estudio y el análisis estadístico para hacerles una recomendación.

Solución

Para comenzar el estudio, Sara especifica un contraste de hipótesis con una hipótesis nula

$$H_0 : \mu_x - \mu_y \leqslant 0$$

frente a una hipótesis alternativa

$$H_1 : \mu_x - \mu_y > 0$$

donde μ_x es la media poblacional de la productividad utilizando estiércol de pavo y μ_y es la media poblacional de la productividad utilizando estiércol de vaca. H_1 indica que el estiércol de pavo aumenta la productividad. Los agricultores no cambiarán de fertilizante a menos que existan pruebas contundentes de que aumenta la productividad. Sara decide antes de recoger los datos que utilizará para este contraste un nivel de significación de $\alpha = 0{,}05$.

Utilizando este diseño, Sara realiza un experimento para contrastar la hipótesis. Utiliza estiércol de vaca en un conjunto de $n_y = 25$ explotaciones agrícolas seleccionadas aleatoriamente. La media muestral de la productividad es $\bar{y} = 100$. Basándose en la experiencia, supone que la varianza de la productividad de estas explotaciones es $\sigma_y^2 = 400$. Utiliza estiércol de pavo en una segunda muestra aleatoria de $n_x = 25$ explotaciones y la media muestral de la productividad es $\bar{x} = 115$. Basándose en algunos estudios publicados, se supone que la varianza de estas explotaciones es $\sigma_x^2 = 625$. Los dos conjuntos de muestras aleatorias son independientes. La regla de decisión es rechazar H_0 en favor de H_1 si

$$\frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}} > z_\alpha$$

Los estadísticos calculados para este problema son

$$n_x = 25 \quad \bar{x} = 115 \quad \sigma_x^2 = 625$$

$$n_y = 25 \quad \bar{y} = 100 \quad \sigma_y^2 = 400$$

$$z = \frac{115 - 100}{\sqrt{\dfrac{625}{25} + \dfrac{400}{25}}} = 2{,}34$$

Comparando el valor calculado de $z = 2{,}34$ con $z_{0{,}05} = 1{,}645$, Sara llega a la conclusión de que se rechaza claramente la hipótesis nula. De hecho, observamos que el *p*-valor de este contraste es 0,0096. Existen, pues, pruebas contundentes de que la productividad es mayor con el estiércol de pavo que con el de vaca.

Dos medias, poblaciones independientes, varianzas desconocidas que se supone que son iguales

En los casos en los que no se conocen las varianzas poblacionales y el tamaño de las muestras es inferior a 100, tenemos que utilizar la distribución t de Student. Hay algunos problemas teóricos cuando se utiliza la distribución t de Student para contrastar las diferencias entre medias muestrales. Sin embargo, estos problemas pueden resolverse utilizando el método siguiente si se puede suponer que las varianzas poblacionales son iguales. Este supuesto es realista en muchos casos en los que comparamos grupos. En el apartado 11.4 presentamos un método para contrastar la igualdad de las varianzas de dos poblaciones normales.

La principal diferencia se encuentra en que este método utiliza un estimador agrupado común de la varianza poblacional igual. Este estimador es

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)}$$

El contraste de hipótesis se realiza utilizando el estadístico t de Student de la diferencia entre dos medias

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}}$$

Obsérvese que la forma de este estadístico es similar a la del estadístico Z, que se utiliza cuando se conocen las varianzas poblacionales. A continuación, se resumen los distintos contrastes en los que se utiliza este método.

Contrastes de la diferencia entre medias poblacionales: varianzas poblacionales desconocidas e iguales

En estos contrastes, se supone que tenemos muestras aleatorias independientes de n_x y n_y observaciones extraídas de poblaciones que siguen una distribución normal que tiene las medias μ_x y μ_y y una varianza común. Se utilizan las varianzas muestrales s_x^2 y s_y^2 para calcular un estimador agrupado de la varianza:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} \qquad \textbf{(11.7)}$$

Hacemos hincapié aquí en que s_p^2 es la media ponderada de las dos varianzas muestrales, s_x^2 y s_y^2.

A continuación, utilizando las medias muestrales observadas $\bar{x}$ e $\bar{y}$, los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \leqslant D_0$$

frente a la alternativa

$$H_1 : \mu_x - \mu_y > D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} > t_{n_x + n_y - 2,\, \alpha} \tag{11.8}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \geqslant D_0$$

frente a la alternativa

$$H_1 : \mu_x - \mu_y < D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} < -t_{n_x + n_y - 2,\, \alpha} \tag{11.9}$$

3. Para contrastar la hipótesis nula

$$H_0 : \mu_x - \mu_y = D_0$$

frente a la hipótesis alternativa bilateral

$$H_1 : \mu_x - \mu_y \neq D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} < -t_{n_x + n_y - 2,\, \alpha/2} \quad \text{o} \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} > t_{n_x + n_y - 2,\, \alpha} \tag{11.10}$$

Aquí, $t_{n_x + n_y - 2,\, \alpha}$ es el número para el que

$$P(t_{n_x + n_y - 2} > t_{n_x + n_y - 2,\, \alpha}) = \alpha$$

Los *p*-valores de todos estos contrastes son la probabilidad de obtener un valor tan extremo como el obtenido, dada la hipótesis nula.

EJEMPLO 11.3. Pautas de ventas al por menor (contraste de hipótesis de las diferencias entre medias)

Una tienda de artículos de deportes se encuentra en un centro comercial de mediano tamaño. Para planificar el volumen de personal, el director le pide que le ayude a averiguar si existen pruebas contundentes de que las ventas son mayores los lunes que los sábados.

Solución

Para responder a esta pregunta, decidimos recoger muestras aleatorias de 25 sábados y 25 lunes de una población de varios años de datos. Las muestras se extraen independientemente. Decidimos contrastar la hipótesis nula

$$H_0 : \mu_M - \mu_S \leqslant 0$$

frente a la hipótesis alternativa

$$H_1 : \mu_M - \mu_S > 0$$

donde los subíndices M y S representan las ventas de los lunes y los sábados. Los estadísticos muestrales son

$$\bar{x}_M = 1.078 \quad s_M = 633 \quad n_M = 25$$
$$\bar{y}_S = 908{,}2 \quad s_S = 469{,}8 \quad n_S = 25$$

La estimación de la varianza agrupada es

$$s_p^2 = \frac{(25-1)(633)^2 + (25-1)(469{,}8)^2}{25+25-2} = 310.700$$

El estadístico del contraste es

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}} = \frac{1.078 - 908{,}2}{\sqrt{\frac{310.700}{25} + \frac{310.700}{25}}} = 1{,}08$$

Utilizando un nivel de significación de $\alpha = 0{,}05$ y 48 grados de libertad, observamos que el valor crítico de t es 1,677. Así pues, llegamos a la conclusión de que no existen pruebas suficientes para rechazar la hipótesis nula y, por lo tanto, no existe razón alguna para concluir que las ventas medias sean mayores los lunes.

EJEMPLO 11.4. Estudio sobre la actividad cerebral (contraste de hipótesis de las diferencias entre medias)

Response to Commercials

En este ejemplo analizamos el efecto de la utilización de diferentes supuestos para la realización de contrastes de las diferencias entre medias poblacionales basados en la t de Student. Recuérdese que en el ejemplo 11.1 realizamos el análisis suponiendo que las observaciones muestrales estaban pareadas. Observamos que existían pruebas para rechazar la hipótesis de que no había ninguna diferencia entre las medias poblacionales y para aceptar la hipótesis de que la media poblacional de la actividad cerebral era mayor en el caso de los anuncios bien recordados. Aquí reconsideramos el ejemplo 11.1 postulando otros supuestos (utilizamos el fichero de datos **Response to Commercials**).

Solución

Primero abandonamos el supuesto de que las observaciones muestrales son datos pareados y están correlacionadas. Suponemos, sin embargo, que las dos varianzas poblacionales son iguales. Estamos contrastando la misma hipótesis que en el ejemplo 11.1. La Figura 11.2 muestra la salida Minitab. El valor de la t de Student calculado es 1,45, el p-valor es 0,082 y los grados de libertad son 18. Por lo tanto, con un nivel de significación de 0,05 no podemos rechazar la hipótesis nula y no hay pruebas contundentes de que exista una diferencia en la actividad cerebral. Sin el supuesto de las muestras pareadas y correlacionadas positivamente, la varianza de la diferencia es demasiado grande para concluir que la diferencia es significativa.

Two-Sample T-Test and CI: X,Y

```
Two-sample T for X vs Y

X       10     108.5      42.5       13
Y       10      85.5      26.5      8.4

Difference = mu (X) - mu (Y)
Estimate for difference:   23.0000
95% lower bound for difference:  -4.4587
T-Test of difference = 0 (vs >): T-Value = 1.45  P-Value = 0.082  DF = 18
Both use Pooled StDev = 35.4079
```

Minitab Instructions
1. Stat
2. Basic statistics
3. 2-sample t
4. Assume equal variances

Figura 11.2. Estudio sobre las ondas cerebrales: muestras independientes, varianzas poblacionales iguales (salida Minitab).

Dos medias, muestras independientes, varianzas poblacionales desconocidas que se supone que no son iguales

Los contrastes de hipótesis de diferencias entre medias poblacionales cuando las varianzas individuales son desconocidas y no son iguales requieren una modificación del cálculo de las varianzas y de los grados de libertad. El cálculo de la varianza muestral de la diferencia entre medias muestrales varía. La determinación de los grados de libertad del valor crítico del estadístico t de Student es muy compleja. La forma de calcularlos se presentó en el apartado 9.1. Las ecuaciones 11.11 a 11.14 resumen los métodos.

Contrastes de la diferencia entre medias poblacionales: varianzas poblacionales desconocidas que se supone que no son iguales

Estos contrastes suponen que tenemos muestras aleatorias independientes de n_x y n_y observaciones procedentes de poblaciones normales que tienen las medias μ_x y μ_y y varianzas desiguales. Se utilizan las varianzas muestrales s_x^2 y s_y^2. Los grados de libertad v del estadístico t de Student vienen dados por

$$v = \frac{\left[\left(\frac{s_x^2}{n_x}\right) + \left(\frac{s_y^2}{n_y}\right)\right]^2}{\left(\frac{s_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{s_y^2}{n_y}\right)^2/(n_y - 1)} \tag{11.11}$$

A continuación, utilizando las medias muestrales observadas $\bar{x}$ e $\bar{y}$, los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \leqslant D_0$$

frente a la hipótesis alternativa

$$H_1 : \mu_x - \mu_y > D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} > t_{v,\alpha} \tag{11.12}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \mu_x - \mu_y = D_0 \quad \text{o} \quad H_0 : \mu_x - \mu_y \geqslant D_0$$

frente a la hipótesis alternativa

$$H_1 : \mu_x - \mu_y < D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}}} < -t_{v,\alpha} \tag{11.13}$$

3. Para contrastar la hipótesis nula

$$H_0 : \mu_x - \mu_y = D_0$$

frente a la hipótesis alternativa bilateral

$$H_1 : \mu_x - \mu_y \neq D_0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}}} < -t_{v,\alpha/2} \quad \text{o} \quad \frac{\bar{x} - \bar{y} - D_0}{\sqrt{\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}}} > t_{v,\alpha/2} \tag{11.14}$$

Aquí, $t_{r,\alpha}$ es el número para el que

$$P(t_r > t_{r,\alpha}) = \alpha$$

El análisis del ejemplo 11.4 se realiza de nuevo sin suponer que las varianzas poblacionales son iguales. La Figura 11.3 muestra la salida Excel. Aquí el único cambio importante es que los grados de libertad son menores, por lo que el p-valor es algo más alto.

Prueba t para dos muestras suponiendo varianzas desiguales		
	Variable 1	*Variable* 2
Media	108,5	85,5
Varianza	1806,72222	700,7222222
Observaciones	10	10
Diferencia hipotética de las medias	0	
Grados de libertad	15	
Estadístico t	1,45248674	
P(T<= t) una cola	0,0834817	
Valor crítico de t (una cola)	1,75305104	
P(T<= t) dos colas	0,1669634	
Valor crítico de t (dos colas)	2,13145086	

Instrucciones de Excel
1. Herramientas
2. Análisis de datos
3. Prueba t para dos muestras suponiendo varianzas desiguales

Figura 11.3. Estudio de las ondas cerebrales: muestras independientes (salida Excel).

EJERCICIOS

Ejercicios básicos

11.1. Le han pedido que averigüe si dos procesos de producción diferentes producen una media diferente de unidades por hora. El proceso 1 tiene una media μ_1 y el 2 tiene una media μ_2. La hipótesis nula y la hipótesis alternativa son

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

Utilizando una muestra aleatoria de 25 observaciones pareadas, las medias muestrales de las poblaciones 1 y 2 son 50 y 60. ¿Puede rechazar la hipótesis nula utilizando una probabilidad de cometer el error de Tipo I $\alpha = 0{,}05$ si

a) La desviación típica muestral de la diferencia es 20?
b) La desviación típica muestral de la diferencia es 30?
c) La desviación típica muestral de la diferencia es 15?
d) La desviación típica muestral de la diferencia es 40?

11.2. Le han pedido que averigüe si dos procesos de producción diferentes producen una media diferente de unidades por hora. El proceso 1 tiene una media μ_1 y el 2 tiene una media μ_2. La hipótesis nula y la hipótesis alternativa son

$$H_0: \mu_1 - \mu_2 \geqslant 0$$
$$H_1: \mu_1 - \mu_2 < 0$$

Utilizando una muestra aleatoria de 25 observaciones pareadas, las medias muestrales de las poblaciones 1 y 2 son 56 y 50. ¿Puede rechazar la hipótesis nula utilizando una probabilidad de cometer el error de Tipo I $\alpha = 0{,}05$ si

a) La desviación típica muestral de la diferencia es 20?
b) La desviación típica muestral de la diferencia es 30?
c) La desviación típica muestral de la diferencia es 15?
d) La desviación típica muestral de la diferencia es 40?

Ejercicios aplicados

11.3. En un estudio en el que se compararon bancos de Alemania y Gran Bretaña, se tomó una muestra de 145 pares de bancos. Cada par contenía un banco de Alemania y uno de Gran Bretaña. Los datos se parearon de manera que los dos miembros fueran lo más parecidos posible en cuanto a tamaño y antigüedad. Se calculó el cociente entre los préstamos totales pendientes y los activos totales de cada uno de los bancos. La diferencia entre las medias muestrales de este cociente (alemana-británica) era 0,0518 y la desviación típica muestral de las diferencias era 0,3055. Contraste la hipótesis nula de que las dos medias poblacionales son iguales frente la hipótesis alternativa bilateral.

11.4. Se ha elaborado un método de selección para medir las actitudes de los directivos hacia las minorías. Una elevada puntuación indica una actitud negativa y una baja puntuación indica una actitud positiva. Se han tomado muestras aleatorias independientes de 151 analistas financieros varones y 108 analistas financieros mujeres. En el caso del primer grupo, la media muestral y la desviación típica muestral de las puntuaciones son 85,8 y 19,13, mientras que en el segundo son 71,5 y 12,2. Contraste la hipótesis nula de que las dos medias poblacionales son iguales frente a la hipótesis alternativa de que la verdadera puntuación media es mayor en el caso de los hombres que en el de las mujeres.

11.5. En una muestra aleatoria de 125 empresarios británicos, el número medio de cambios de empleo es 1,91 y la desviación típica muestral es 1,32. En una muestra aleatoria independiente de 86 directivos británicos, el número medio de cambios de empleo es 0,21 y la desviación típica muestral es 0,53. Contraste la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa de que el número medio de cambios de empleo es mayor en el caso de los empresarios británicos que en el de los directivos británicos.

11.6. Un profesor de ciencia política tiene interés en comparar las características de los estudiantes que votan en las elecciones nacionales y las de los que no votan. En una muestra aleatoria de 114 estudiantes que afirman que han votado en las últimas elecciones presidenciales, observa una media de las calificaciones medias de 2,71 y una desviación típica de 0,64. En una muestra

aleatoria independiente de 123 estudiantes que no han votado, la media de las calificaciones medias es 2,79 y la desviación típica es 0,56. Contraste la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa bilateral.

11.7. Ante las quiebras recientes de grandes empresas, los auditores están cada vez más preocupados por la posibilidad de que existan fraudes. Los auditores pueden averiguar más fácilmente las posibilidades de que existan fraudes si calculan minuciosamente el flujo de caja. Para evaluar esta posibilidad, unas muestras de auditores de nivel medio que trabajan en empresas de auditoría reciben información sobre el flujo de caja de un caso de fraude y se les pide que indiquen la posibilidad de que haya un fraude material en una escala de 0 a 100. Una muestra aleatoria de 36 auditores utiliza la información sobre el flujo de caja. Su valoración media es de 36,21 y la desviación típica muestral es 22,93. En el caso de una muestra aleatoria independiente de 36 auditores que no utilizan la información sobre el flujo de caja, la media muestral y la desviación típica muestral son 47,56 y 27,56, respectivamente. Suponiendo que las dos distribuciones poblacionales son normales y tienen la misma varianza, contraste la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa bilateral.

11.8. Se examinan folletos de ofertas públicas de venta de acciones. En una muestra aleatoria de 70 folletos en los que se revelan las predicciones sobre las ventas, el cociente medio entre la deuda y el capital propio antes de la oferta es 3,97 y la desviación típica muestral es 6,14. En una muestra aleatoria independiente de 51 folletos en los que no se revelan las predicciones sobre las ventas, el cociente medio entre la deuda y el capital propio es 2,86 y la desviación típica muestral es 4,29. Contraste la hipótesis nula de que las medias poblacionales de los cocientes de los que no revelan las predicciones sobre las ventas y los de las que sí las revelan son iguales frente a la hipótesis alternativa bilateral.

11.9. Una editorial tiene interés en saber cómo afectan a las ventas los manuales universitarios que contienen más de 100 ficheros de datos. La editorial planea producir 20 manuales sobre administración de empresas y elige aleatoriamente 10 para introducir en ellos más de 100 ficheros de datos. Los 10 restantes no llevarán más de 100 ficheros de datos. En el caso de los primeros, las ventas son, en promedio, de 9.254 durante el primer año y la desviación típica muestral es 2.107. En el caso de los segundos, las ventas son, en promedio, de 8.167 durante el primer año y la desviación típica muestral es 1.681. Suponiendo que las dos distribuciones poblacionales son normales y tienen la misma varianza, contraste la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa de que la verdadera media es mayor en el caso de los manuales que contienen más de 100 ficheros de datos.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

11.10. El centro de colocación de una universidad quiere averiguar si los licenciados y las licenciadas en economía reciben, en promedio, ofertas salariales diferentes en el primer empleo. Selecciona aleatoriamente ocho pares de licenciados en administración de empresas de tal manera que las calificaciones, los intereses y el origen socioeconómico de cada par sean lo más parecidos posible. El fichero de datos **Salary Pair** contiene la oferta salarial más alta recibida por cada miembro de la muestra al final de la ronda de reclutamiento. Suponiendo que las distribuciones son normales, contraste la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa de que la verdadera media de los hombres es mayor que la de las mujeres.

11.11. Una academia ofrece a los estudiantes cursos de preparación para el examen de admisión en un programa de postgrado. En un experimento para evaluar las virtudes del curso, se eligieron 12 estudiantes y se dividieron en seis pares cuyos miembros tenían parecido expediente académico. Antes de realizar el examen, se eligió aleatoriamente un miembro de cada par para que realizara el curso de preparación y el otro no realizó ningún curso. Las calificaciones obtenidas en el examen se encuentran en el fichero de datos **Student Pair**. Suponiendo que las diferencias entre las calificaciones siguen una distribución normal, contraste al nivel del 5 por ciento la hipótesis nula de que las dos medias poblacionales son iguales frente a la hipótesis alternativa de que la verdadera media es mayor en el caso de los estudiantes que asisten al curso de preparación.

11.2. Contrastes de la diferencia entre dos proporciones poblacionales (grandes muestras)

A continuación, presentamos métodos para comparar dos proporciones poblacionales. Examinamos un modelo aplicable a una muestra aleatoria de n_x observaciones procedentes de una población que tiene una proporción P_x de «éxitos» y una segunda muestra aleatoria independiente de n_y observaciones procedentes de una población que tiene una proporción P_y de «éxitos».

En el Capítulo 6 vimos que, cuando las muestras son grandes, las variables aleatorias que siguen una distribución normal son una buena aproximación de las proporciones, por lo que

$$Z = \frac{(\hat{p}_x - \hat{p}_y) - (P_x - P_y)}{\sqrt{\dfrac{P_x(1 - P_x)}{n_x} + \dfrac{P_y(1 - P_y)}{n_y}}}$$

sigue una distribución normal estándar.

Queremos contrastar la hipótesis de que las proporciones poblacionales P_x y P_y son iguales. Sea P_0 su valor común. Entonces, partiendo de esta hipótesis,

$$Z = \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{P_0(1 - P_0)}{n_x} + \dfrac{P_0(1 - P_0)}{n_y}}}$$

sigue aproximadamente una distribución normal estándar.

Por último, la proporción desconocida P_0 puede estimarse por medio de un estimador agrupado

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

En estos contrastes, la hipótesis nula supone que las proporciones poblacionales son iguales. Si la hipótesis nula es verdadera, entonces puede obtenerse un estimador insesgado y eficiente de P_0 combinando las dos muestras aleatorias y, como consecuencia, se calcula $\hat{p}_0$ utilizando esta ecuación. En ese caso, podemos sustituir la P_0 desconocida por $\hat{p}_0$ para obtener una variable aleatoria que tiene una distribución parecida a la normal estándar, cuando el tamaño de la muestra es grande.

A continuación se resumen los contrastes.

Contraste de la igualdad de dos proporciones poblacionales (grandes muestras)

Tenemos muestras aleatorias independientes de tamaño n_x y n_y que tienen una proporción de éxitos $\hat{p}_x$ y $\hat{p}_y$. Cuando suponemos que las proporciones poblacionales son iguales, una estimación de la proporción común es

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y}$$

Cuando el tamaño de la muestra es grande —$nP_0(1 - P_0) > 9$—, los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: P_x - P_y = 0 \quad \text{o} \quad H_0: P_x - P_y \leqslant 0$$

frente a la hipótesis alternativa

$$H_1: P_x - P_y > 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_\alpha \tag{11.15}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: P_x - P_y = 0 \quad \text{o} \quad H_0: P_x - P_y \geqslant 0$$

frente a la hipótesis alternativa

$$H_1: P_x - P_y < 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < -z_\alpha \tag{11.16}$$

3. Para contrastar la hipótesis nula

$$H_0: P_x - P_y = 0$$

frente a la hipótesis alternativa bilateral

$$H_1: P_x - P_y \neq 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} < -z_{\alpha/2} \quad \text{o}$$

$$\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} > z_{\alpha/2} \tag{11.17}$$

También es posible calcular e interpretar los *p*-valores como la probabilidad de obtener un valor al menos tan extremo como el obtenido, dada la hipótesis nula.

EJEMPLO 11.5. El humor en los anuncios publicados en revistas británicas y estadounidenses (contrastes de hipótesis de diferencias entre proporciones)

Se ha realizado un estudio para averiguar si existe alguna diferencia entre el contenido humorístico de los anuncios de las revistas británicas y las estadounidenses. En una muestra aleatoria independiente de 270 anuncios de revistas estadounidenses, 56 eran humorísticos. En una muestra aleatoria independiente de 203 anuncios de revistas británicas, 52 eran humorísticos. ¿Constituyen estos datos una prueba de que existe una diferencia entre las proporciones de anuncios humorísticos de las revistas británicas y las de las revistas estadounidenses?

Solución

Sean P_x y P_y las proporciones poblacionales de anuncios británicos y estadounidenses humorísticos, respectivamente. La hipótesis nula es

$$H_0 : P_x - P_y = 0$$

y la hipótesis alternativa es

$$H_1 : P_x - P_y \neq 0$$

La regla de decisión es rechazar H_0 en favor de H_1 si

$$\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{P_0(1-P_0)}{n_x} + \dfrac{P_0(1-P_0)}{n_y}}} < -z_{\alpha/2} \quad o \quad > z_{\alpha/2}$$

Los datos de este problema son

$$n_x = 203 \qquad \hat{p}_x = 52/203 = 0{,}256 \qquad n_y = 270 \qquad \hat{p}_y = 56/270 = 0{,}207$$

La estimación de la varianza común P_0 según la hipótesis nula es

$$\hat{p}_0 = \frac{n_x\hat{p}_x + n_y\hat{p}_y}{n_x + n_y} = \frac{(203)(0{,}256) + (270)(0{,}207)}{203 + 270} = 0{,}228$$

El estadístico del contraste es

$$\frac{(\hat{p}_x - \hat{p}_y)}{\sqrt{\dfrac{\hat{p}_0(1-\hat{p}_0)}{n_x} + \dfrac{\hat{p}_0(1-\hat{p}_0)}{n_y}}} = \frac{0{,}256 - 0{,}207}{\sqrt{\dfrac{(0{,}228)(1-0{,}228)}{203} + \dfrac{(0{,}228)(1-0{,}228)}{270}}} = 1{,}26$$

En un contraste de dos colas con $\alpha = 0{,}10$, el valor de $z_{0{,}05}$ es 1,645. Por lo tanto, no es posible rechazar la hipótesis nula, por lo que tenemos pocas pruebas de que exista una diferencia entre los anuncios humorísticos de los dos países.

EJERCICIOS

Ejercicios básicos

11.12. Contraste las hipótesis

$$H_0 : P_x - P_y = 0$$
$$H_1 : P_x - P_y > 0$$

utilizando los siguientes estadísticos de muestras aleatorias:

a) $\hat{p}_x = 0{,}42$, $n_x = 500$; $\hat{p}_y = 0{,}50$, $n_y = 600$
b) $\hat{p}_x = 0{,}60$, $n_x = 500$; $\hat{p}_y = 0{,}64$, $n_y = 600$
c) $\hat{p}_x = 0{,}42$, $n_x = 500$; $\hat{p}_y = 0{,}49$, $n_y = 600$
d) $\hat{p}_x = 0{,}25$, $n_x = 500$; $\hat{p}_y = 0{,}34$, $n_y = 600$
e) $\hat{p}_x = 0{,}39$, $n_x = 500$; $\hat{p}_y = 0{,}42$, $n_y = 600$

Ejercicios aplicados

11.13. Las muestras aleatorias de 900 personas de Estados Unidos y de Gran Bretaña indican que el 60 por ciento de los estadounidenses ve con optimismo el futuro de la economía, mientras que la cifra es del 66 por ciento en el caso de los británicos. ¿Es esta información una prueba contundente de que los británicos ven con más optimismo el futuro de la economía?

11.14. Una muestra aleatoria de 1.556 personas del país A debe responder a la siguiente afirmación: «El aumento del comercio mundial puede aumentar nuestra prosperidad per cápita». El 38,4 por ciento de los miembros de esta muestra está de acuerdo con esta afirmación. Cuando se presenta la misma afirmación a una muestra aleatoria de 1.108 personas del país B, el 52,0 por ciento está de acuerdo. Contraste la hipótesis nula de que las proporciones poblacionales que están de acuerdo con esta afirmación son las mismas en los dos países frente a la hipótesis alternativa de que la proporción que está de acuerdo es mayor en el país B.

11.15. En Estados Unidos, se encuestó a las pequeñas empresas 6 meses después de que fuera posible contratar los servicios telefónicos de larga distancia con otras compañías telefónicas distintas de AT&T. De una muestra aleatoria de 368 pequeñas empresas usuarias de AT&T, 92 declararon que estaban intentando obtener más información sobre sus opciones, al igual que 37 de una muestra aleatoria independiente de 116 usuarias de otras compañías telefónicas. Contraste al nivel de significación del 5 por ciento la hipótesis nula de que las dos proporciones poblacionales son iguales frente a la hipótesis alternativa bilateral.

11.16. Los empleados de una cadena de venta de materiales de construcción a punto de cerrar fueron encuestados para conocer su opinión sobre un plan de compra de la empresa. Algunos se comprometieron a aportar 10.000 $ a este plan, entregando inmediatamente 800 $, mientras que otros declararon que no tenían intención de aportar nada. En una muestra aleatoria de 175 empleados que se comprometieron a aportar dinero, 78 ya habían sido despedidos, mientras que 208 de una muestra aleatoria de 604 que no se comprometieron a aportar nada ya habían sido despedidos. Contraste al nivel del 5 por ciento la hipótesis nula de que las proporciones poblacionales que ya han sido despedidas son iguales en los dos grupos frente a la hipótesis alternativa bilateral.

11.17. En una muestra aleatoria de 381 opciones sobre acciones de alta calidad, 191 tenían una deuda de menos del 30 por ciento. En una muestra aleatoria independiente de 166 opciones sobre acciones de alto riesgo, 145 tenían una deuda de menos del 30 por ciento. Contraste la hipótesis nula de que las dos proporciones poblacionales son iguales frente a la hipótesis alternativa bilateral.

11.18. Se preguntó a muestras aleatorias independientes de consumidores si estaban satisfechos con su sistema informático de dos formas algo distintas. Las respuestas posibles eran las mismas en los dos casos. Cuando se les preguntó hasta qué punto estaban *satisfechos* con su sistema informático, 138 de 240 miembros de la muestra declararon «muy satisfecho». Cuando se les preguntó hasta qué punto estaban *insatisfechos* con su sistema informático, 128 de 240 miembros de la muestra declararon «muy satisfecho». Contraste al nivel de significación del 5 por ciento la hipótesis nula de que las dos proporciones poblacionales son iguales frente a la hipótesis alternativa bilateral evidente.

11.19. En una muestra aleatoria de 1.200 daneses, 480 tenían una actitud positiva hacia los vendedores de automóviles. En una muestra aleatoria de 1.000 franceses, 790 tenían una actitud positiva hacia los vendedores de automóviles. Contraste al nivel del 1 por ciento la hipótesis nula de que las proporciones poblacionales son iguales frente a la hipótesis alternativa de que la proporción de franceses que tienen una actitud positiva hacia los vendedores de automóviles es mayor.

11.3. Contrastes de la varianza de una distribución normal

Además de la necesidad de realizar contrastes basados en la media muestral, hay algunas situaciones en las que queremos saber si la varianza poblacional es un valor específico o un conjunto de valores. En los estudios modernos de control de calidad, esta necesidad es especialmente importante, ya que un proceso que tiene, por ejemplo, una varianza excesivamente grande puede producir muchos artículos defectuosos. Aquí presentamos métodos para contrastar la varianza poblacional σ^2 basándonos en la varianza muestral s_x^2, calculada utilizando una muestra aleatoria de n observaciones extraídas de una población que sigue una distribución normal. La base para realizar contrastes específicos se halla en el hecho de que la variable aleatoria

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

sigue una distribución ji-cuadrado con $(n-1)$ grados de libertad. Si la hipótesis nula es que la varianza poblacional es igual a un valor específico σ_0^2, es decir,

$$H_0 : \sigma^2 = \sigma_0^2$$

entonces, cuando esta hipótesis es verdadera, la variable aleatoria

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$$

sigue una distribución ji-cuadrado con $(n-1)$ grados de libertad. Los contrates de hipótesis se basan en valores calculados de este estadístico. Si la hipótesis alternativa es que la varianza poblacional es mayor que σ_0^2, sospecharíamos de la hipótesis nula si la varianza muestral es muy superior a σ_0^2. Si el valor calculado de χ^2_{n-1} fuera alto, se rechazaría la hipótesis nula. Y a la inversa, se aceptaría la hipótesis alternativa de que la varianza poblacional es inferior a σ_0^2 y se rechazaría la hipótesis nula si el valor de χ^2_{n-1} fuera bajo. En el caso de una hipótesis alternativa bilateral de que la varianza poblacional es diferente de σ_0^2, rechazaríamos la hipótesis nula si el valor fuera excepcionalmente alto o excepcionalmente bajo. Los contrastes basados en una distribución ji-cuadrado son más sensibles al supuesto de la normalidad en la distribución subyacente que los contrastes basados en una distribución normal estándar. Por lo tanto, si la población subyacente se desvía considerablemente de la normal, los niveles de significación calculados utilizando la distribución ji-cuadrado pueden desviarse de los niveles de significación correctos basados en la distribución exacta.

La justificación de la realización de contrastes adecuados sigue la lógica del apartado 11.2 y utiliza la notación de la distribución ji-cuadrado desarrollada en el apartado 9.3. $\chi^2_{v,\alpha}$ representa el número que es superado con una probabilidad α por una variable aleatoria ji-cuadrado con v grados de libertad. Es decir,

$$P(\chi^2_v > \chi^2_{v,\alpha}) = \alpha \quad \text{y/o} \quad P(\chi^2_v < \chi^2_{v,1-\alpha}) = \alpha$$

y en el caso de los contrates de dos colas,

$$P(\chi^2_v > \chi^2_{v,\alpha/2} \quad \text{o} \quad \chi^2_v < \chi^2_{v,1-\alpha/2}) = \alpha$$

Estas probabilidades se muestran en la Figura 9.5 y los distintos contrastes se resumen en las ecuaciones 11.18, 11.19 y 11.20.

También es posible hallar *p*-valores para el contraste ji-cuadrado de varianzas. Del resultado general que acabamos de formular se deduce que el *p*-valor del contraste ji-cuadrado es la probabilidad de obtener un valor al menos tan extremo como el obtenido, dada la hipótesis nula.

Contrastes de la varianza de una población normal

Tenemos una muestra aleatoria de n observaciones procedentes de una población que sigue una distribución normal que tiene una varianza σ^2. Si observamos la varianza muestral s^2, los siguientes contrastes tienen el nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{o} \quad H_0: \sigma^2 \leqslant \sigma_0^2$$

frente a la hipótesis alternativa

$$H_1: \sigma^2 > \sigma_0^2$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{n-1,\alpha} \tag{11.18}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{o} \quad H_0: \sigma^2 \geqslant \sigma_0^2$$

frente a la hipótesis alternativa

$$H_1: \sigma^2 < \sigma_0^2$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha} \tag{11.19}$$

3. Para contrastar la hipótesis nula

$$H_0: \sigma^2 = \sigma_0^2$$

frente a la hipótesis alternativa bilateral

$$H_1: \sigma^2 \neq \sigma_0^2$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{n-1,\alpha/2} \quad \text{o} \quad \frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha/2} \tag{11.20}$$

donde χ^2_{n-1} es una variable aleatoria ji-cuadrado y $P(\chi^2_{n-1} > \chi^2_{n-1,\alpha}) = \alpha$.

El *p*-valor es la probabilidad de obtener un valor al menos tan extremo como el obtenido, dada la hipótesis nula.

EJEMPLO 11.6. Varianza de las impurezas de un producto químico (contrastes de hipótesis de varianzas poblacionales)

El director de control de calidad de Industrias Químicas Asociadas le ha pedido que averigüe si la varianza de las impurezas de sus envíos de fertilizante está dentro de la norma establecida. Esta norma establece que la varianza de los kilos de impurezas de los sacos de 100 kilos no puede ser superior a 4.

Solución

Se obtiene una muestra aleatoria de 20 sacos y se miden los kilos de impurezas de cada saco. Se calcula que la varianza muestral es 6,62. En este problema, contrastamos la hipótesis nula

$$H_0: \sigma^2 \leqslant \sigma_0^2 = 4$$

frente a la hipótesis alternativa

$$H_1: \sigma^2 > 4$$

Basándonos en el supuesto de que la población sigue una distribución normal, la regla de decisión para un contraste de nivel de significación α es rechazar H_0 en favor de H_1 si

$$\frac{(n-1)s_x^2}{\sigma_0^2} > \chi^2_{n-1,\alpha}$$

Para este contraste, con $\alpha = 0{,}05$ y 19 grados de libertad, el valor crítico de la variable ji-cuadrado es 30,14, según la Tabla 7 de la ji-cuadrado del apéndice. Entonces, utilizando los datos del contraste, observamos que

$$\frac{(n-1)s_x^2}{\sigma_0^2} = \frac{(20-1)(6{,}62)}{4} = 31{,}45 > \chi^2_{n-1,\alpha} = 30{,}14$$

Por lo tanto, rechazamos la hipótesis nula y concluimos que la variabilidad de las impurezas es superior a lo que establece la norma. Como consecuencia, recomendamos que se estudie el proceso de producción y se hagan mejoras para reducir la variabilidad de los componentes del producto.

El *p*-valor de este contraste es la probabilidad de obtener un estadístico ji-cuadrado con 19 grados de libertad que sea mayor que el observado 31,45:

$$p\text{-valor} = P\left(\frac{(19)s_x^2}{\sigma_0^2} > \chi^2_{19} = 31{,}45\right) = 0{,}036$$

El *p*-valor de 0,036 se ha calculado utilizando la función de distribución de probabilidad Minitab para la distribución ji-cuadrado.

EJERCICIOS

Ejercicios básicos

11.20. Contraste las hipótesis

$$H_0 : \sigma^2 \leqslant 100$$
$$H_1 : \sigma^2 > 100$$

utilizando los siguientes resultados de una muestra aleatoria.

a) $s^2 = 165$; $n = 25$
b) $s^2 = 165$; $n = 29$
c) $s^2 = 159$; $n = 25$
d) $s^2 = 67$; $n = 38$

Ejercicios aplicados

11.21. Ante la insistencia de un inspector de trabajo, se instala un nuevo mecanismo de seguridad en una cadena de montaje. Tras la instalación, se toma una muestra aleatoria de la producción de 8 días y se obtienen los siguientes resultados sobre el número de componentes acabados producidos:

618 660 638 625 571 598 639 582

A la dirección le preocupa la variabilidad de la producción diaria y considera negativa cualquier varianza superior a 500. Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la varianza poblacional de la producción diaria no es superior a 500.

11.22. El plástico que produce una máquina se revisa periódicamente para ver si fluctúa su grosor. Si la verdadera varianza del grosor es de más de 2,25 milímetros cuadrados, hay motivos para preocuparse por la calidad del producto. Se realizan mediciones del grosor de una muestra aleatoria de 10 rollos de plástico producidos en un turno y se obtienen los siguientes resultados (en milímetros):

226 226 232 227 225
228 225 228 229 230

a) Halle la varianza muestral.
b) Contraste al nivel de significación del 5 por ciento la hipótesis nula de que la varianza poblacional es 2,25 como máximo.

11.23. Una manera de evaluar la eficacia de un profesor ayudante es examinar las calificaciones que obtienen sus estudiantes en el examen final del curso. Evidentemente, es interesante la calificación media. Sin embargo, la varianza también contiene información útil: algunos profesores tienen un estilo que da muy buenos resultados con los estudiantes más capacitados, pero no con los menos capacitados o motivados. Un profesor pone al final de cada cuatrimestre el mismo examen para todos los grupos del curso. La varianza de las calificaciones de este examen normalmente es muy cercana a 300. Un nuevo profesor ayudante tiene una clase de 30 estudiantes, cuyas calificaciones tienen una varianza de 480. Considerando las calificaciones obtenidas por estos estudiantes en el examen como una muestra aleatoria extraída de una población normal, contraste la hipótesis nula de que la varianza poblacional de sus calificaciones es de 300 frente a la hipótesis alternativa bilateral.

11.24. Una empresa produce aparatos eléctricos que se pueden regular con un termostato. La desviación típica de la temperatura a la que se pone en marcha el termostato no debe sobrepasar los 2 °F. En una muestra aleatoria de 20 de estos termostatos, la desviación típica muestral de las temperaturas a las que se pone en marcha es de 2,36 °F. Indicando los supuestos que necesite postular, contraste al nivel del 5 por ciento la hipótesis nula de que la desviación típica poblacional es 2,0 frente a la hipótesis alternativa de que es mayor.

11.25. Un profesor ha decidido introducir un componente mayor de estudio independiente en un curso de microeconomía intermedia para animar a los estudiantes a trabajar por su cuenta y a estudiar más detenidamente la materia. Un colega le advierte de que ese método puede aumentar la variabilidad del rendimiento de los estudiantes. Sin embargo, el profesor le responde que es de esperar que la variabilidad sea menor. Ha observado en sus datos que antes las calificaciones de los estudiantes en el examen final de este curso seguían una distribución normal con una desviación típica de 18,2 puntos. En una clase de 25 estudiantes en que utilizó este nuevo método, la desviación típica de las calificaciones del examen final era de 15,3 puntos. Suponiendo que estos 25 estudiantes pueden considerarse una muestra aleatoria de todos los que podrían tener que seguir el nuevo método, contraste la hipótesis nula de que la desviación típica poblacional es al menos de 18,2 puntos frente a la hipótesis alternativa de que es menor.

11.4. Contrastes de la igualdad de las varianzas entre dos poblaciones distribuidas normalmente

Hay algunas situaciones en las que nos interesa comparar las varianzas de dos poblaciones distribuidas normalmente. Por ejemplo, en el contraste basado en la t de Student del apartado 11.1 hemos supuesto que las varianzas eran iguales y hemos utilizado las dos varianzas muestrales para calcular un estimador agrupado de las varianzas comunes. Veremos que las comparaciones de las varianzas también son importantes métodos inferenciales para el análisis de regresión (véanse los Capítulos 12 y 13) y para el análisis de la varianza (véase el Capítulo 17). En los estudios del control de calidad a menudo se trata de saber qué proceso tiene la menor varianza.

En este apartado presentamos un método para contrastar el supuesto de que las varianzas poblacionales de muestras independientes son iguales. Para realizar esos contrastes, introducimos la distribución de probabilidad F. Comenzamos suponiendo que s_x^2 es la varianza muestral de una muestra aleatoria de n_x observaciones procedentes de una población que sigue una distribución normal que tiene una varianza poblacional σ_x^2, y s_y^2 una varianza muestral de una segunda muestra aleatoria independiente de tamaño n_y procedente de una población normal que tiene una varianza poblacional σ_y^2. En ese caso, la variable aleatoria

$$F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$$

sigue una distribución conocida con el nombre de distribución F. Esta familia de distribuciones, que se utiliza frecuentemente en el análisis estadístico, se identifica por los grados de libertad del numerador y los grados de libertad del denominador. Los grados de libertad del numerador están relacionados con la varianza muestral s_x^2 y son iguales a $(n_x - 1)$. Asimismo, los grados de libertad del denominador están relacionados con la varianza muestral s_y^2 y son iguales a $(n_y - 1)$.

La distribución F es el cociente entre dos variables aleatorias ji-cuadrado, dividida cada una por sus grados de libertad. La distribución ji-cuadrado relaciona la varianza muestral con la varianza poblacional de una población que sigue una distribución normal. Los contrastes de hipótesis que utilizan la distribución F dependen del supuesto de una distribución normal. Las características de la distribución F se resumen a continuación.

La distribución *F*

Tenemos dos muestras aleatorias independientes con n_x y n_y observaciones procedentes de dos poblaciones normales que tienen las varianzas σ_x^2 y σ_y^2. Si las varianzas muestrales son s_x^2 y s_y^2, entonces la variable aleatoria

$$F = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} \qquad \textbf{(11.21)}$$

sigue una distribución F con $(n_x - 1)$ grados de libertad en el numerador y $(n_y - 1)$ grados de libertad en el denominador.

Una distribución F con v_1 grados de libertad en el numerador y v_2 grados de libertad en el denominador se representa de la forma siguiente: F_{v_1, v_2}. $F_{v_1, v_2, \alpha}$ es el número para el que

$$P(F_{v_1, v_2} > F_{v_1, v_2, \alpha}) = \alpha$$

Debemos hacer hincapié en que este contraste es muy sensible al supuesto de la normalidad.

Los puntos de corte de $F_{v_1, v_2, \alpha}$ cuando α es igual a 0,05 y 0,01 se encuentran en la Tabla 9 del apéndice. Por ejemplo, vemos en la tabla que para 10 grados de libertad en el numerador y 20 en el denominador,

$$F_{10,20,0,05} = 2,35 \quad \text{y} \quad F_{10,20,0,01} = 3,37$$

Por lo tanto,

$$P(F_{10,20} > 2,35) = 0,05 \quad \text{y} \quad P(F_{10,20} > 3,37) = 0,01$$

La Figura 11.4 contiene una descripción esquemática de la distribución F correspondiente a este ejemplo.

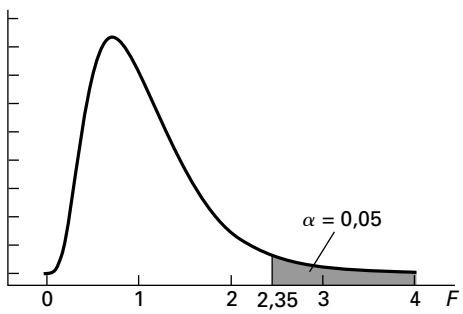


Figura 11.4. Función de densidad de la distribución F con 10 grados de libertad en el numerador y 20 grados de libertad en el denominador.

En las aplicaciones prácticas, normalmente colocamos la varianza muestral mayor en el numerador y la menor en el denominador. Por lo tanto, sólo necesitamos utilizar los puntos de corte superiores para contrastar la hipótesis de la igualdad de las varianzas. Cuando las varianzas poblacionales son iguales, la variable aleatoria F se convierte en

$$F = \frac{s_x^2}{s_y^2}$$

y este cociente entre las varianzas muestrales se convierte en el estadístico del contraste. La idea intuitiva en la que se basa este contraste es bastante sencilla: si una de las varianzas muestrales es muy superior a la otra, debemos concluir que las varianzas poblacionales no son iguales. A continuación, resumimos los contrastes de hipótesis de la igualdad de las varianzas.

Contrastes de la igualdad de las varianzas de dos poblaciones normales

Sean s_x^2 y s_y^2 las varianzas muestrales observadas de muestras aleatorias independientes de tamaño n_x y n_y de poblaciones distribuidas normalmente que tienen las varianzas σ_X^2 y σ_Y^2. Sea s_x^2 la varianza mayor. En ese caso, los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{o} \quad H_0 : \sigma_X^2 \leqslant \sigma_Y^2$$

frente a la hipótesis alternativa

$$H_1 : \sigma_X^2 > \sigma_Y^2$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{s_x^2}{s_y^2} > F_{n_x-1,\,n_y-1,\,\alpha} \tag{11.22}$$

2. Para contrastar la hipótesis nula

$$H_0: \sigma_X^2 = \sigma_Y^2$$

frente a la hipótesis alternativa bilateral

$$H_1: \sigma_X^2 \neq \sigma_Y^2$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{s_x^2}{s_y^2} > F_{n_x-1,\,n_y-1,\,\alpha/2} \tag{11.23}$$

donde s_x^2 es la mayor de las dos varianzas muestrales. Dado que cualquiera de las dos varianzas muestrales podría ser mayor, esta regla se basa en realidad en un contraste de dos colas y, por lo tanto, utilizamos $\alpha/2$ como la probabilidad de la cola superior.

En este caso, $F_{n_x-1,\,n_y-1}$ es el número para el que

$$P(F_{n_x-1,\,n_y-1} > F_{n_x-1,\,n_y-1,\,\alpha}) = \alpha$$

donde $F_{n_x-1,\,n_y-1}$ tiene una distribución F con $(n_x - 1)$ grados de libertad en el numerador y $(n_y - 1)$ grados de libertad en el denominador.

En todos estos contrastes, un *p*-valor es la probabilidad de obtener un valor al menos tan extremo como el obtenido, dada la hipótesis nula. Dada la complejidad de la distribución F, sólo se calculan los valores críticos para unos cuantos casos especiales. Por lo tanto, normalmente los *p*-valores se calculan utilizando un paquete estadístico como Minitab.

EJEMPLO 11.7. Estudio de Inversores Inmediatos sobre las varianzas de los vencimientos (contrastes de hipótesis de dos varianzas)

El personal de investigación de Inversores Inmediatos, sociedad de contratación financiera en línea, tenía interés en averiguar si existe una diferencia entre las varianzas de los vencimientos de los bonos industriales AAA y la de los bonos industriales CCC.

Solución

Para este ejercicio hay que diseñar un estudio que compare las varianzas poblacionales de los vencimientos de los dos tipos de bonos. Contrastaremos la hipótesis nula

$$H_0: \sigma_X^2 = \sigma_Y^2$$

frente a la hipótesis alternativa

$$H_1: \sigma_X^2 \neq \sigma_Y^2$$

donde σ_X^2 es la varianza de los vencimientos de los bonos AAA y σ_Y^2 es la varianza de los vencimientos de los bonos CCC. El nivel de significación del contraste elegido es $\alpha = 0{,}02$.

La regla de decisión es rechazar H_0 en favor de H_1 si

$$\frac{s_x^2}{s_y^2} > F_{n_x-1,\, n_y-1,\, \alpha/2}$$

Obsérvese que cualquiera de las dos varianzas muestrales podría ser mayor y, por lo tanto, estar en el denominador. Así pues, la probabilidad de esta cola superior es $\alpha/2$. En una muestra aleatoria de 17 bonos AAA, la varianza muestral es $s_x^2 = 123{,}35$ y en una muestra aleatoria independiente de 11 bonos CCC, la varianza muestral es $s_y^2 = 8{,}02$. El estadístico del contraste es, pues,

$$\frac{s_x^2}{s_y^2} = \frac{123{,}35}{8{,}02} = 15{,}38$$

Dado un nivel de significación de $\alpha = 0{,}02$, observamos que el valor crítico de F, calculado mediante interpolación en la Tabla 9 del apéndice, es

$$F_{16,\,10,\,0{,}01} = 4{,}53$$

Es evidente que el valor calculado de F (15,38) es superior al valor crítico (4,53), por lo que rechazamos H_0 en favor de H_1. Existen, pues, pruebas contundentes de que las varianzas de los vencimientos de estos dos tipos de bonos son diferentes.

EJERCICIOS

Ejercicios básicos

11.26. Contraste la hipótesis

$$H_0: \sigma_x^2 = \sigma_y^2$$
$$H_1: \sigma_x^2 > \sigma_y^2$$

utilizando los datos siguientes:

a) $s_x^2 = 125$, $n_x = 45$; $s_y^2 = 51$, $n_y = 41$
b) $s_x^2 = 125$, $n_x = 45$; $s_y^2 = 235$, $n_y = 44$
c) $s_x^2 = 134$, $n_x = 48$; $s_y^2 = 51$, $n_y = 41$
d) $s_x^2 = 88$, $n_x = 39$; $s_y^2 = 167$, $n_y = 25$

Ejercicios aplicados

11.27. Se parte de la hipótesis de que cuanto más experto es un grupo de personas que examinan las declaraciones del impuesto sobre la renta, más variables son sus opiniones sobre su exactitud. Se eligieron muestras aleatorias independientes, de 30 personas cada una, de grupos que tenían diferentes niveles de experiencia. El grupo con «poca experiencia» estaba formado por personas que acababan de terminar su primer curso de contabilidad intermedia. Los miembros del grupo de «mucha experiencia» habían terminado los estudios universitarios y trabajaban en empresas auditoras de prestigio. Se pidió a los miembros de las muestras que juzgaran la exactitud de las declaraciones del impuesto sobre la renta. La varianza muestral del grupo con poca experiencia era de 451,770, mientras que la del grupo con mucha experiencia era 1.614,208. Contraste la hipótesis nula de que las dos varianzas poblacionales son iguales frente a la hipótesis alternativa de que la verdadera varianza es mayor en el caso del grupo con mucha experiencia.

11.28. Se parte de la hipótesis de que las ventas totales de una empresa deben variar más en una industria en la que haya competencia de precios que en una que sea un duopolio y en la que haya colusión tácita. En un estudio de la industria de producción de barcos mercantes, se observó que en cuatro años de competencia de precios la varianza de las ventas totales de la empresa A era 114,09. En los siete años siguientes, durante los cuales hubo duopolio y colusión tácita, esta varianza fue 16,08. Suponga que los datos pueden considerarse como una muestra aleatoria inde-

pendiente procedente de dos distribuciones normales. Contraste al nivel del 5 por ciento la hipótesis nula de que las dos varianzas poblacionales son iguales frente a la hipótesis alternativa de que la varianza de las ventas totales es mayor en los años en los que hay competencia de precios.

11.29. En el ejercicio 11.7, hemos partido del supuesto de que las varianzas poblacionales de las valoraciones de la posibilidad de que exista un fraude material de los auditores que utilizaban información sobre el flujo de caja y de los que no la utilizaban eran iguales. Contraste este supuesto frente a la hipótesis alternativa bilateral.

11.30. En el ejercicio 11.9, hemos supuesto que las varianzas poblacionales de las ventas de manuales que contenían más de 100 ficheros de datos y de las ventas de manuales que no contenían más de 100 ficheros eran iguales el primer año. Contraste este supuesto frente a la hipótesis alternativa bilateral.

11.31. Un equipo universitario de investigación estaba estudiando la relación entre la generación de ideas por parte de los grupos con y sin moderador. En una muestra aleatoria de cuatro grupos con moderador, el número medio de ideas generadas por grupo era de 78,0 y la desviación típica era de 24,4. En una muestra aleatoria de cuatro grupos sin moderador, el número medio de ideas generadas era de 63,5 y la desviación típica era de 20,2. Contraste el supuesto de que las dos varianzas poblacionales son iguales frente a la hipótesis alternativa de que la varianza poblacional es mayor en los grupos con moderador.

11.5. Algunas observaciones sobre el contraste de hipótesis

En este capítulo hemos presentado varias aplicaciones importantes de la metodología del contraste de hipótesis. Esta metodología es en un importante sentido fundamental para tomar decisiones y para el análisis cuando hay variabilidad aleatoria, por lo que los métodos pueden aplicarse a muchas decisiones de investigación y de gestión. Son relativamente fáciles de utilizar y algunos procesos informáticos facilitan la realización de los cálculos. Tenemos, pues, un instrumento que es atractivo y bastante fácil de utilizar. Sin embargo, hay algunos sutiles problemas y motivos de preocupación que es necesario examinar para no cometer graves errores.

La hipótesis nula desempeña un papel fundamental en el modelo de contraste de hipótesis. En una investigación, normalmente fijamos el nivel de significación, α, en un bajo valor. A continuación, obtenemos una muestra aleatoria y utilizamos los datos para calcular un estadístico del contraste. Si el estadístico está fuera de la región de aceptación (dependiendo de la dirección del contraste), rechazamos la hipótesis nula y aceptamos la hipótesis alternativa. Cuando rechazamos la hipótesis nula, tenemos pruebas contundentes —una pequeña probabilidad de error— en favor de la hipótesis alternativa. En algunos casos, no podemos rechazar drásticamente las hipótesis nulas falsas simplemente porque sólo tenemos una reducida información muestral o porque el contraste tiene poca potencia. Puede haber importantes casos en los que este resultado es adecuado. Por ejemplo, no cambiaríamos un proceso existente que está funcionando eficazmente a menos que tuviéramos pruebas contundentes de que uno nuevo sería claramente incluso mejor. Sin embargo, en otros casos, el estatus especial de la hipótesis nula no está justificado ni es adecuado. En esos casos, podríamos considerar los costes de cometer tanto errores de Tipo I como errores de Tipo II en un proceso de decisión. También podríamos considerar otra especificación de la hipótesis nula, recordando que el rechazo de la hipótesis nula constituye una prueba contundente a favor de la hipótesis alternativa. Cuando tenemos dos alternativas, podríamos elegir inicialmente cualquiera de las dos como hipótesis nula. En el ejemplo del peso de las cajas de cereales que ponemos al principio del Capítulo 10, la hipótesis nula podría ser o bien que

$$H_0 : \mu \geqslant 16$$

o bien que

$$H_0 : \mu \leqslant 16$$

En el primer caso, el rechazo sería una prueba rotunda de que la media poblacional del peso es inferior a 16. En el segundo caso, el rechazo constituiría una prueba contundente de que la media poblacional del peso es superior a 16. Como hemos indicado, el no rechazar cualquiera de estas dos hipótesis nulas no sería una prueba contundente. También hay métodos para controlar simultáneamente tanto los errores de Tipo I como los de Tipo II (véase, por ejemplo, la referencia bibliográfica 1).

A veces se dispone de abundante información muestral y se rechaza la hipótesis nula incluso cuando las diferencias casi no son importantes. Necesitamos, pues, contrastar la significación estadística con una definición más amplia de significación. Supongamos que se utilizan muestras muy grandes para comparar las rentas familiares medias anuales de dos ciudades. Uno de los resultados podría ser que las medias muestrales se diferencian en 2,67 $ y esa diferencia podría llevarnos a rechazar una hipótesis nula y a concluir, pues, que una de las ciudades tiene una renta familiar media más alta que la otra. Aunque ese resultado podría ser estadísticamente significativo, es evidente que en la práctica no lo es con respecto al consumo o la calidad de vida.

Cuando se especifica una hipótesis nula y una regla de contraste, se definen las condiciones del contraste antes de examinar los datos muestrales generados por un proceso que contiene un componente aleatorio. Por lo tanto, si examinamos los datos antes de definir la hipótesis nula y la hipótesis alternativa, ya no tenemos predeterminada la probabilidad de error y el concepto de «evidencia contundente» resultante del rechazo de la hipótesis nula no es válido. Por ejemplo, si decidimos el nivel de significación de nuestro contraste después de haber visto los *p*-valores, no podemos interpretar nuestros resultados en términos probabilísticos. Supongamos que un economista compara cinco programas de mejora de la renta con respecto a un nivel mínimo básico utilizando un contraste de hipótesis. Después de recoger los datos y de calcular los *p*-valores, decide que la hipótesis nula —una renta no superior al nivel mínimo básico— puede rechazarse en el caso de uno de los cinco programas con un nivel de significación de $\alpha = 0{,}20$. Es evidente que este resultado va en contra del uso adecuado del contraste de hipótesis. Pero hemos visto que esto lo hacen economistas supuestamente profesionales.

Al aumentar la capacidad de los instrumentos de cálculo, hay algunas nuevas formas de violar el principio de especificar la hipótesis nula antes de ver los datos. La reciente popularidad de la «minería de datos» (*data mining*) —la utilización de un programa informático para buscar relaciones entre variables en un gran conjunto de datos— introduce nuevas posibilidades de cometer abusos. La «minería de datos» puede suministrar una descripción de subconjuntos y diferencias en una muestra de datos especialmente grande. Sin embargo, después de ver los resultados de una operación de ese tipo, los analistas pueden tener la tentación de definir contrastes de hipótesis que utilicen muestras aleatorias procedentes del mismo conjunto de datos. Eso viola claramente el principio que establece que hay que definir el contraste de hipótesis antes de ver los datos. Una compañía farmacéutica puede seleccionar un gran número de tratamientos médicos y descubrir que 5 de cada 100 medicamentos producen efectos significativos en el tratamiento de enfermedades para las que no estaban pensados. Ese resultado podría utilizarse legítimamente para identificar posibles temas de investigación para un nuevo estudio de investigación con nuevas muestras aleatorias. Sin embargo, si los datos originales se utilizan entonces para contrastar una hipótesis sobre los beneficios de los 5 medicamentos, tenemos una grave violación de la aplicación correcta del contraste de hipótesis y ninguna de las probabilidades de error es correcta.

Para definir la hipótesis nula y la hipótesis alternativa hay que considerar detenidamente los objetivos del análisis. Por ejemplo, podríamos encontrarnos ante una propuesta para introducir un nuevo proceso de producción. En uno de los casos, el proceso actual podría contener mucho equipo nuevo, trabajadores bien formados y la creencia de que el proceso funciona muy bien. En ese caso, la productividad del proceso actual sería la hipótesis nula y el nuevo proceso sería la hipótesis alternativa. Adoptaríamos el nuevo proceso sólo si existen pruebas contundentes —rechazo de la hipótesis nula con una pequeña α— de que el nuevo tiene una productividad mayor. En el otro caso, el proceso actual podría ser viejo y contener equipo que hay que reponer y algunos trabajadores que necesitan más formación. En ese caso, podríamos utilizar como hipótesis nula la productividad del nuevo proceso. Continuaríamos, pues, manteniendo el viejo proceso sólo si existen pruebas contundentes de que su productividad es mayor.

Cuando trazamos gráficos de control para controlar la calidad de un proceso, como veremos en el Capítulo 20, consideramos que el nivel deseado del proceso es la hipótesis nula y fijamos también un nivel de significación muy bajo: $\alpha < 0{,}01$. Por lo tanto, sólo rechazamos la hipótesis nula cuando hay pruebas muy contundentes de que el proceso ya no funciona bien. Sin embargo, estos contrastes de hipótesis basados en gráficos de control sólo se realizan después de que se han hecho grandes esfuerzos para controlar el proceso y minimizar su variabilidad. Por lo tanto, estamos bastante seguros de que el proceso funciona correctamente y no queremos cambiarlo en respuesta a pequeñas variaciones de los datos muestrales. Pero si encontramos un estadístico basado en los datos muestrales cuyo contraste se sitúa fuera del intervalo de aceptación y, por lo tanto, rechazamos la hipótesis nula, podemos estar bastante seguros de que algo ha ido mal y de que es necesario cambiar el proceso inmediatamente.

Los contrastes presentados en este capítulo se basan en el supuesto de que la distribución subyacente es normal o de que se aplica el teorema del límite central para la distribución de las medias muestrales o las proporciones. Cuando el supuesto de la normalidad ya no se cumple, esas probabilidades de error pueden no ser válidas. Como no podemos estar seguros de que la mayoría de las poblaciones sean exactamente normales, podría preocuparnos seriamente la validez de nuestros contrastes. Muchas investigaciones han demostrado que los contrastes de medias no dependen mucho del supuesto de la normalidad. Se dice que estos contrastes son «robustos» con respecto a la normalidad. Sin embargo, los contrastes de varianzas no lo son. Por lo tanto, hay que tener mayor precaución cuando se utilizan contrastes de hipótesis basados en varianzas.

RESUMEN

En este capítulo hemos continuado presentando nuestra metodología para realizar contrastes de hipótesis clásicos. Basándonos en el Capítulo 10, hemos analizado métodos para comparar medias poblacionales y proporciones poblacionales. Hemos presentado contrastes de hipótesis de las diferencias entre medias poblacionales y entre proporciones poblacionales. También hemos mostrado métodos para contrastar varianzas poblacionales utilizando varianzas muestrales. Por último, hemos presentado métodos para comparar varianzas poblacionales de dos poblaciones diferentes. Hemos examinado, además, las características del entorno en el que se plantea el problema y hemos señalado las aplicaciones adecuadas e inadecuadas del contraste de hipótesis.

La Figura 11.5 muestra un diagrama de flujos para seleccionar el contraste de hipótesis adecuado cuando se comparan medias poblacionales y la 11.6 muestra otro diagrama que indica la forma de seleccionar un contraste de hipótesis adecuado cuando se comparan dos proporciones poblacionales. Los dos constituyen un buen resumen de las distintas opciones para contrastar hipótesis y podrían resultar útiles al lector en su futuro trabajo.

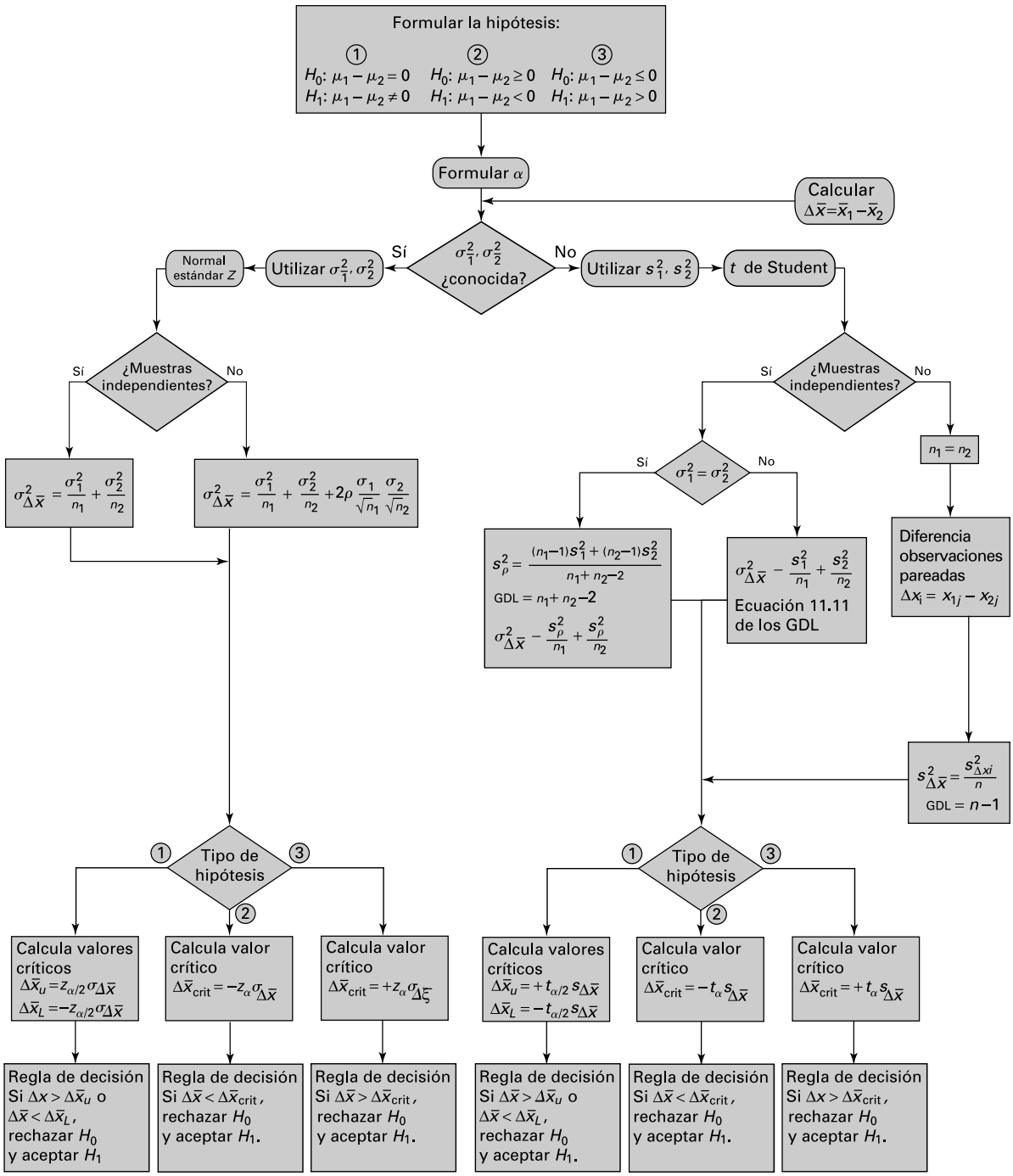


Figura 11.5. Diagrama de flujo para seleccionar el contraste de hipótesis adecuado cuando se comparan dos medias poblacionales.

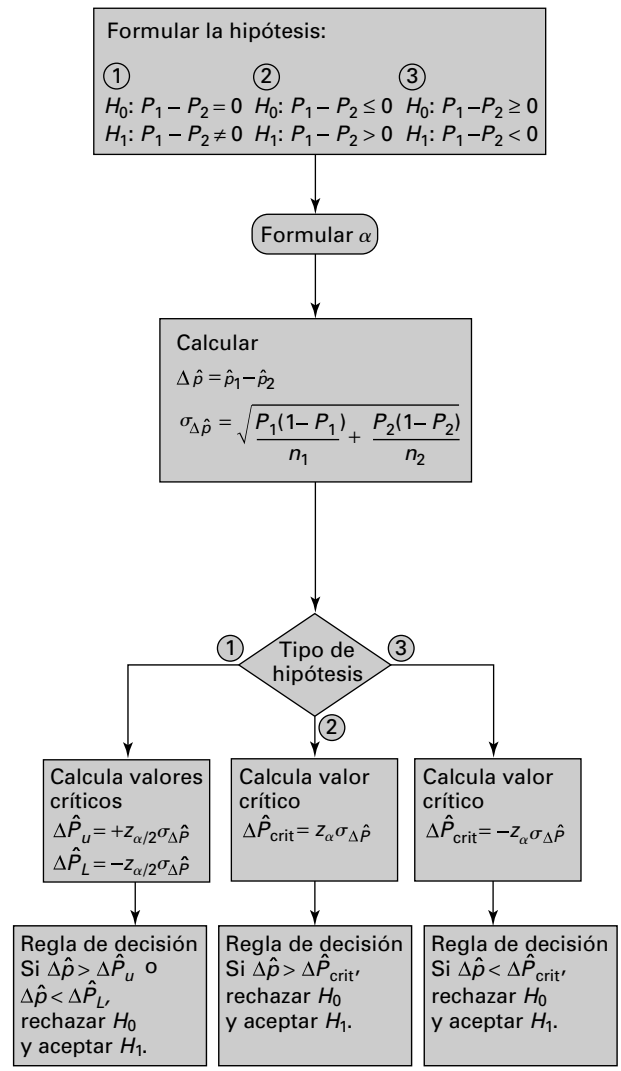


Figura 11.6. Diagrama de flujo para seleccionar el contraste de hipótesis adecuado cuando se comparan dos proporciones poblacionales.

TÉRMINOS CLAVE

contraste de la igualdad de dos proporciones poblacionales (grandes muestras), 408
contrastes de la igualdad de varianzas de dos poblaciones normales, 417
contrastes de la diferencia entre medias poblacionales: muestras independientes (varianzas conocidas), 398
contrastes de la diferencia entre medias poblacionales: datos pareados, 395
contrastes de la diferencia entre medias poblacionales: varianzas poblacionales desconocidas e iguales, 401
contrastes de la diferencia entre medias poblacionales: varianzas poblacionales desconocidas que se supone que no son iguales, 404
contrastes de la varianza de una población normal, 413
distribución F, 416
hipótesis alternativa, 394
hipótesis nula, 394

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

Nota: Si no se indica la probabilidad de cometer un error de Tipo I, seleccione un nivel que sea adecuado para la situación descrita.

11.32. Un estadístico contrasta la hipótesis nula de que la proporción de hombres partidarios de una reforma tributaria propuesta es igual que la de mujeres. Basándose en datos muestrales, se rechaza la hipótesis nula al nivel de significación del 5 por ciento. ¿Implica eso que hay al menos una probabilidad de 0,95 de que la hipótesis nula sea falsa? En caso contrario, formule una afirmación probabilística correcta.

11.33. Un proceso produce cable para la compañía telefónica local. Cuando el proceso está funcionando correctamente, el diámetro del cable sigue una distribución normal de media 1,6 centímetros y desviación típica 0,05 centímetros. En una muestra aleatoria de 16 trozos de cable, los diámetros tenían una media muestral de 1,615 centímetros y una desviación típica muestral de 0,086 centímetros.

a) Suponiendo que la desviación típica poblacional es de 0,05 centímetros, contraste al nivel del 10 por ciento la hipótesis nula de que la media poblacional es de 1,6 centímetros frente a la hipótesis alternativa bilateral. Halle también el nivel de significación más bajo al que puede rechazarse esta hipótesis nula frente a la hipótesis alternativa bilateral.

b) Contraste al nivel del 10 por ciento la hipótesis nula de que la desviación típica poblacional es de 0,05 centímetros frente a la hipótesis alternativa de que es mayor.

11.34. Cuando un proceso de producción funciona normalmente, produce pastillas en las que el peso medio del principio activo es de 5 gramos y la desviación típica es de 0,025 gramos. En una muestra aleatoria de 12 pastillas, se encontraron los siguientes pesos del principio activo:

5,01	4,69	5,03	4,98	4,98	4,95
5,00	5,00	5,03	5,01	5,04	4,95

a) Sin suponer que se conoce la varianza poblacional, contraste la hipótesis nula de que la media poblacional del peso del principio activo por pastilla es de 5 gramos. Utilice una hipótesis alternativa bilateral y un nivel de significación del 5 por ciento. Indique los supuestos que postule.

b) Indicando los supuestos que postule, contraste la hipótesis nula de que la desviación típica poblacional es de 0,025 gramos frente a la hipótesis alternativa de que la desviación típica poblacional es de más de 0,025 gramos. Utilice un nivel de significación del 5 por ciento.

11.35. Una compañía de seguros tiene agentes a comisión. Sostiene que el primer año los agentes perciben una comisión media de 40.000 $ como mínimo y que la desviación típica poblacional no supera los 6.000 $. En una muestra aleatoria de nueve agentes se observa que en lo que se refiere a la comisión percibida el primer año,

$$\sum_{i=1}^{9} x_i = 333 \quad \text{y} \quad \sum_{i=1}^{9} (x_i - \bar{x})^2 = 312$$

expresada en miles de dólares. Puede suponerse que la distribución poblacional es normal.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional es al menos de 40.000 $.

b) Contraste al nivel del 10 por ciento la hipótesis nula de que la desviación típica poblacional es como máximo de 6.000 $.

11.36. En un estudio sobre el índice de rendimiento laboral de antiguos fumadores, una muestra aleatoria de 34 antiguos fumadores tenía un índice medio de 2,21 y una desviación típica muestral de 2,21. En una muestra aleatoria independiente de 86 personas que hacía mucho tiempo que habían dejado de fumar, el índice medio era de 1,47 y la desviación típica muestral era de 1,69. Halle el nivel de significación más bajo al que la hipótesis nula de la igualdad de las dos medias poblacionales puede rechazarse frente a la hipótesis alternativa bilateral.

11.37. Se pide a muestras aleatorias independientes de directivos de empresas y profesores universitarios de economía que valoren en una escala de 1 (totalmente en desacuerdo) a 7 (totalmente de acuerdo) la siguiente afirmación: «Las calificaciones obtenidas en los cursos de economía avanzada son buenos indicadores de la capacidad analítica de los estudiantes». En una muestra de 70 directivos de empresa, la respuesta media es de 4,4 y la desviación típica muestral es de 1,3. En una muestra de 106 profesores de economía, la respuesta media es de 5,3 y la desviación típica muestral es de 1,4.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional de las respuestas de los directivos de empresa es como máximo de 4,0.

b) Contraste al nivel del 5 por ciento la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa de que la media poblacional de las respuestas es mayor en el caso de los profesores de economía que en el de los directivos de empresa.

11.38. En un estudio, se tomaron muestras aleatorias independientes de titulados medios y de titulados superiores en estadística que empezaron trabajando en una gran empresa actuarial y después pasaron a una compañía de seguros. En una muestra de 44 titulados medios, el número medio de meses que tardaron en cambiar de empleo fue de 35,02 y la desviación típica muestral fue de 18,20. En una muestra de 68 titulados superiores, el número medio de meses que tardaron en cambiar de empleo fue de 36,34 y la desviación típica muestral fue de 18,94. Contraste al nivel del 10 por ciento la hipótesis nula de que la media poblacional del número de meses que tardaron los dos grupos en cambiar de empleo es la misma frente a la hipótesis alternativa bilateral.

11.39. Un estudio pretendía evaluar la influencia del tamaño y de las características de los grupos en la generación de conceptos publicitarios. Para evaluar la influencia del tamaño del grupo, se compararon grupos de cuatro y ocho miembros. En una muestra aleatoria de cuatro grupos de 4 miembros, el número medio de conceptos publicitarios generados por grupo fue de 78,0 y la desviación típica muestral fue de 24,4. En una muestra aleatoria independiente de cuatro grupos de 8 miembros, el número medio de conceptos publicitarios generados por grupo fue de 114,7 y la desviación típica muestral fue de 14,6 (en los dos casos, los grupos tenían un moderador). Indicando los supuestos que necesite postular, contraste al nivel del 1 por ciento la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa de que la media es mayor en el caso de los grupos de 8 miembros.

11.40. Se calcula un índice de dificultad de lectura de un texto escrito siguiendo estos pasos:

i. Se halla el número medio de palabras por frase.

ii. Se halla el porcentaje de palabras que tienen cuatro sílabas o más.

iii. El índice es un 40 por ciento de la suma de (i) y (ii).

Una muestra aleatoria de seis anuncios de la revista A tenía los siguientes índices:

15,75 11,55 11,16 9,92 9,23 8,20

Una muestra aleatoria independiente de seis anuncios de la revista B tenía los siguientes índices:

9,17 8,44 6,10 5,78 5,58 5,36

Indicando los supuestos que necesite postular, contraste al nivel del 5 por ciento la hipótesis nula de que la media poblacional de los índices es la misma frente a la hipótesis alternativa de que la verdadera media es mayor en el caso de la revista A que en el de la B.

11.41. En el ejercicio 11.40, los índices de una muestra aleatoria de seis anuncios de la revista C eran los siguientes:

9,50 8,60 8,59 6,50 4,79 4,29

En una muestra aleatoria independiente de seis anuncios de la revista D, los índices eran los siguientes:

10,21 9,66 7,67 5,12 4,88 3,12

Indicando los supuestos que necesite postular, contraste la hipótesis nula de que las medias poblacionales de los índices son iguales frente a una hipótesis alternativa bilateral.

11.42. Se pide a muestras aleatorias independientes de profesores de administración de empresas y de economía que valoren en una escala de 1 (totalmente en desacuerdo) a 4 (totalmente de acuerdo) la siguiente afirmación: «La amenaza y la realidad de las absorciones de empresas que cotizan en bolsa obligan a los consejos de administración y a los directivos a maximizar el valor de las empresas para los accionistas». En una muestra de 202 profesores de administración de empresas, la respuesta media fue de 2,83 y la desviación típica muestral fue de 0,89. En una muestra de 291 profesores de economía, la respuesta media fue de 3,00 y la desviación típica muestral fue de 0,67. Contraste la hipótesis nula de que las medias poblacionales son iguales frente a la hipótesis alternativa de que la media es mayor en el caso de los profesores de economía.

11.43. Se pregunta a muestras aleatorias independientes de pacientes a los que se les han colocado

prótesis de rodilla y de cadera que valoren la calidad del servicio en una escala de 1 (baja) a 7 (alta). En una muestra de 83 pacientes operados de rodilla, la valoración media es de 6,543 y la desviación típica muestral es de 0,649. En una muestra de 54 pacientes operados de cadera, la valoración media es de 6,733 y la desviación típica muestral es de 0,425. Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de estos dos tipos de pacientes son iguales frente a la hipótesis alternativa bilateral.

11.44. En una muestra aleatoria de 148 estudiantes de contabilidad, 75 consideran que tener sentido del humor es una característica muy importante para su carrera. En una muestra aleatoria independiente de 178 estudiantes de economía financiera, 81 piensan lo mismo.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que al menos la mitad de todos los estudiantes de economía financiera consideran que el sentido del humor es muy importante.

b) Contraste al nivel del 5 por ciento la hipótesis nula de que las proporciones poblacionales de los estudiantes de contabilidad y de economía financiera que consideran que el sentido del humor es muy importante son iguales frente a la hipótesis alternativa bilateral.

11.45. En un estudio cuyo objetivo era ver si los beneficios estaban disminuyendo mucho, se tomó una muestra aleatoria de 23 empresas en las que estaban disminuyendo considerablemente y en las que el rendimiento medio de los activos en los tres años anteriores había sido de 0,058 y la desviación típica muestral de 0,055. En una muestra aleatoria independiente de 23 empresas en las que los beneficios no estaban disminuyendo considerablemente, el rendimiento medio había sido de 0,146 y la desviación típica de 0,058 durante ese mismo periodo. Suponga que las dos distribuciones poblacionales son normales y tienen las mismas desviaciones típicas. Contraste al nivel del 5 por ciento la hipótesis nula de que las medias poblacionales de los rendimientos de los activos son iguales frente a la hipótesis alternativa de que la verdadera media es mayor en el caso de las empresas en las que los beneficios no estaban disminuyendo considerablemente.

11.46. En un estudio se extrajeron muestras aleatorias de empleados de restaurantes de comida rápida en los que el empresario da formación. En una muestra de 67 empleados que no habían terminado los estudios secundarios, 11 habían participado en un programa de formación de la empresa. En una muestra aleatoria independiente de 113 empleados que habían terminado los estudios secundarios, pero no habían ido a la universidad, habían participado 27. Contraste al nivel del 1 por ciento la hipótesis nula de que las tasas de participación de los dos grupos son iguales frente a la hipótesis alternativa de que la tasa es mucho más baja en el caso de los que no habían terminado los estudios secundarios.

11.47. En una muestra aleatoria de 69 sociedades de seguros médicos, 47 tenían su propio departamento de relaciones públicas, al igual que 40 de una muestra aleatoria independiente de 69 sociedades de seguros de accidentes. Halle e interprete el *p*-valor de un contraste de la igualdad de las proporciones poblacionales frente a la hipótesis alternativa bilateral.

11.48. En un estudio, se tomaron muestras aleatorias independientes de hombres y mujeres clientes de Centro de Iniciativa Empresarial. Estos clientes estaban considerando la posibilidad de montar una empresa. De 94 hombres clientes, 53 montaron de hecho una empresa, al igual que 47 de 68 mujeres clientes. Halle e interprete el *p*-valor de un contraste de la igualdad de las proporciones poblacionales frente a la hipótesis alternativa de que la proporción de mujeres clientes que montaron realmente una empresa es mayor que la de hombres.

11.49. Se calcula un índice de dificultad de lectura de un texto escrito siguiendo estos pasos:

i. Se halla el número medio de palabras por frase.

ii. Se halla el porcentaje de palabras que tienen cuatro sílabas o más.

iii. El índice es un 40 por ciento de la suma de (i) y (ii).

Una muestra aleatoria de seis anuncios de la revista A tenía los siguientes índices:

15,75 11,55 11,16 9,92 9,23 8,20

Una muestra aleatoria independiente de seis anuncios de la revista B tenía los siguientes índices:

9,17 8,44 6,10 5,78 5,58 5,36

Contraste la hipótesis nula de que la desviación típica poblacional del índice de anuncios de la

revista A es igual que la desviación típica poblacional del índice de anuncios de la revista B frente a la hipótesis alternativa bilateral.

11.50. Se pide a dos analistas financieros que predigan los beneficios por acción que tendrá una muestra aleatoria de 12 empresas el próximo año. Para evaluar la calidad de sus predicciones se utiliza como indicador el error porcentual absoluto de predicción, que se define de la forma siguiente:

$$100 \times \frac{|\text{Efectivos} - \text{Predichos}|}{\text{Efectivos}}$$

Los errores porcentuales absolutos de predicción cometidos se encuentran en el fichero de datos **Analyst Prediction**.

Contraste la hipótesis nula de la igualdad de las varianzas poblacionales de los errores porcentuales absolutos de predicción de los dos analistas financieros.

11.51. Una persona es responsable del desarrollo económico rural en un país que está desarrollándose rápidamente y utilizando el petróleo recién encontrado para desarrollar todo el país. Una de sus responsabilidades es averiguar si existen pruebas de que los nuevos métodos de cultivo del arroz han aumentado la producción por hectárea. Se plantó una muestra aleatoria de 27 arrozales utilizando el viejo método y la media muestral de la producción fue de 60 por hectárea con una varianza muestral de 100. Durante el segundo año, se utilizó el nuevo método en los mismos arrozales y la media muestral de la producción fue de 64 por hectárea con una varianza muestral de 150. La correlación muestral entre los dos arrozales fue de 0,38. Se supone que las varianzas poblacionales son iguales y debe utilizarse ese supuesto para el análisis del problema.

a) Utilice un contraste de hipótesis con una probabilidad de cometer un error de Tipo I = 0,05 para averiguar si hay pruebas contundentes que permitan concluir que el nuevo método aumenta la producción por hectárea e interprete los resultados.

b) Partiendo del supuesto de que las varianzas poblacionales son iguales, construya un intervalo de aceptación al 95 por ciento del cociente entre las varianzas muestrales. ¿Nos llevan las varianzas muestrales observadas a concluir que las varianzas poblacionales son iguales? Explique su respuesta.

11.52. La presidenta de Comercios Planetarios Reunidos (CPR), Susana Perales, le ha pedido ayuda para estudiar el grado de penetración del nuevo teléfono móvil de la empresa en el mercado. Le ha pedido que estudie dos mercados y averigüe si la diferencia entre las cuotas de mercado sigue siendo la misma. Históricamente, en el mercado 1, situado en el oeste de Polonia, CPR ha tenido una cuota de mercado del 30 por ciento. En el mercado 2, situado en el sur de Austria, ha tenido una cuota de mercado del 35 por ciento. Obtiene una muestra aleatoria de clientes potenciales de cada zona. En el mercado 1, 258 de una muestra total de 800 declaran que comprarán el teléfono de CPR. En el mercado 2, 260 de 700 declaran que comprarán el teléfono de CPR.

a) Utilizando una probabilidad de error $\alpha = 0{,}03$, contraste la hipótesis de que las cuotas de mercado son iguales frente a la hipótesis de que no son iguales (mercado 2 − mercado 1).

b) Utilizando una probabilidad de error $\alpha = 0{,}03$, contraste la hipótesis de que las cuotas de mercado son iguales frente a la hipótesis de que la cuota del mercado 2 es mayor.

11.53. En un experimento agrícola van a probar dos variedades de maíz de alto rendimiento y a medir las mejoras de ese rendimiento. El experimento se organiza de tal forma que cada variedad se siembra en 10 pares de parcelas similares. Los datos que se encuentran en el fichero de datos **Corn Yield** son los aumentos porcentuales del rendimiento obtenidos con estas dos variedades. Indicando los supuestos que postule, contraste al nivel del 10 por ciento la hipótesis nula de que las dos medias poblacionales de los aumentos porcentuales del rendimiento son iguales. Utilice la hipótesis alternativa bilateral.

11.54. Usted es el director de producto de la marca 4 de una gran empresa de productos alimenticios. El presidente de la empresa se ha quejado de que una marca rival, llamada marca 2, tiene unas ventas medias mayores. El departamento de datos ha almacenado las cifras más recientes sobre las ventas («saleb2» y «saleb4») y sobre los precios («apriceb2» y «apriceb4») en un fichero llamado **Storet**, que se encuentra en su disco de datos o en el sistema informático local.

a) Basándose en un contraste de hipótesis estadístico, ¿tiene el presidente pruebas contundentes que apoyen su queja? Muestre todo el trabajo y el razonamiento estadísticos.

b) Después de analizar los datos, observa que en la muestra de la marca 2 hay un gran caso atípico de valor 971. Repita el apartado (a) una vez eliminada esta observación extrema. ¿Qué conclusión extrae ahora sobre la queja del presidente?

11.55. Joe Ortega es el director de producto para Helados Ole. Le ha pedido que averigüe si Helados Ole tiene más ventas que Helados Carl, que es un fuerte competidor. El fichero de datos **Ole** contiene datos sobre las ventas y los precios semanales de las marcas rivales del año en tres cadenas de supermercados. Estos datos muestrales representan una muestra aleatoria de todas las ventas de helado de las dos marcas.

a) Diseñe y realice un análisis para averiguar si existen pruebas contundentes que permitan concluir que las ventas medias de Helados Ole son mayores que las de Helados Carl ($\alpha = 0{,}05$). Explique su método y muestre todos los cálculos. Puede incluir una salida Minitab si es adecuado para apoyar su análisis. Explique sus conclusiones.

b) Diseñe y realice un análisis para averiguar si los precios que cobran las dos marcas son diferentes ($\alpha = 0{,}05$). Explique detenidamente su análisis, muestre todos los cálculos e interprete sus resultados.

11.56. María Perlas es responsable de preparar harina mezclada para hacer pan exótico. El proceso consiste en tomar dos tipos diferentes de harina y mezclarlas para lograr un pan de alta calidad. Para uno de los productos, se mezcla harina A y harina B. El paquete de harina A procede de un proceso de empaquetado que tiene un peso medio poblacional de 8 onzas con una varianza poblacional de 0,04. El paquete de harina B tiene un peso medio poblacional de 8 onzas y una varianza poblacional de 0,06. Los pesos de los paquetes tienen una correlación de 0,40. Los paquetes A y B se mezclan para obtener un paquete de 16 onzas de harina exótica especial. Cada 60 minutos se selecciona una muestra aleatoria de cuatro paquetes de harina exótica en el proceso y se calcula el peso medio de los cuatro paquetes. Prepare un intervalo de aceptación del 99 por ciento para un gráfico de control de calidad para las medias muestrales de la muestra de cuatro paquetes. Muestre todos los pasos que sigue y explique su razonamiento. Explique cómo se utilizaría este gráfico de aceptación para garantizar que el peso de los paquetes continúa cumpliendo las normas.

Bibliografía

1. Carlson, W. L. y B. Thorne, *Applied Statistical Methods*, Upper Saddle River, NJ, Prentice Hall, 1997, págs. 539-553.

Capítulo 12

Regresión simple

Esquema del capítulo

Introducción

Hasta ahora hemos centrado la atención en el análisis y la inferencia relacionados con una única variable. En este capítulo extendemos nuestro análisis a las relaciones entre variables. Comenzamos con una breve introducción al análisis de correlación, seguido de la presentación del análisis de regresión simple. Nuestra presentación es paralela a la del Capítulo 3, en el que hicimos hincapié en las relaciones descriptivas, incluido el uso de diagramas de puntos dispersos, coeficientes de correlación y la regresión lineal como instrumentos para describir las relaciones entre variables. Suponemos que el lector está familiarizado con ese capítulo.

En el análisis de los procesos empresariales y económicos se utilizan a menudo las relaciones entre variables. Estas relaciones se expresan en términos matemáticos de la forma siguiente:

$$Y = f(X)$$

donde la función puede adoptar muchas formas lineales y no lineales. En algunos de esos casos, la forma de la relación no se conoce exactamente. Aquí presentamos análisis que se basan en relaciones lineales. En muchos casos, las relaciones lineales constituyen un buen modelo del proceso. En otros casos, nos interesa una parte limitada de una relación no lineal a la que podemos aproximarnos mediante una relación lineal. En el apartado 13.7 mostramos que algunas relaciones no lineales importantes también pueden analizarse utilizando el análisis de regresión. Por lo tanto, los métodos de correlación y de regresión pueden aplicarse a una amplia variedad de problemas.

Las relaciones lineales son muy útiles para muchas aplicaciones empresariales y económicas, como indican los siguientes ejemplos. El presidente de Materiales de Construcción, S.A., fabricante de placas de yeso, cree que la cantidad anual media de placas de yeso vendidas en su región es una función lineal del valor total de los permisos de edificación expedidos durante el año anterior. Un vendedor de cereales quiere saber cómo afecta la producción total al precio por tonelada. Está desarrollando un modelo de predicción que utiliza datos históricos. El departamento de marketing necesita saber cómo afecta el precio de la gasolina a sus ventas totales. Utilizando datos semanales sobre los precios y las ventas, planea desarrollar un modelo lineal que muestre cuánto varían las ventas cuando varía el precio.

Con la aparición de muchos y buenos paquetes estadísticos y hojas de cálculo como Excel, hoy es posible para casi todo el mundo calcular estadísticos de correlación y de regresión. Desgraciadamente, también sabemos que no todo el mundo sabe interpretar y utilizar correctamente estos resultados obtenidos por computador. Aquí el lector aprenderá algunas ideas fundamentales que lo ayudarán a utilizar el análisis de regresión. Comenzaremos examinando el análisis de correlación.

12.1. Análisis de correlación

En este apartado utilizamos los coeficientes de correlación para estudiar las relaciones entre variables. En el Capítulo 3 utilizamos el coeficiente de correlación muestral para describir la relación entre variables indicada en los datos. En el 5 y en el 6 aprendimos lo que era la correlación poblacional. Aquí presentamos métodos inferenciales que utilizan el coeficiente de correlación para estudiar relaciones lineales entre variables.

En principio, dos variables aleatorias pueden estar relacionadas de diversas formas. Es útil postular al comienzo del análisis una forma funcional de su relación. A menudo es razonable suponer, como buena aproximación, que la relación es lineal. Si se examina un par de variables aleatorias, X e Y, entre las que existe una relación lineal, en un diagrama de puntos dispersos las observaciones conjuntas sobre este par de variables tenderán a estar concentradas en torno a una línea recta. Y a la inversa, si no existe una relación lineal, no estarán concentradas en torno a una línea recta. No todas las relaciones que estudiaremos estarán muy concentradas en torno a una línea recta. El diagrama de puntos dispersos de muchas relaciones importantes muestra una tendencia hacia una relación lineal, pero con una considerable desviación con respecto a una línea recta. En los diagramas de puntos dispersos del Capítulo 2 vimos algunos ejemplos.

Las correlaciones tienen muchas aplicaciones en el mundo de la empresa y en la economía. En muchos problemas económicos aplicados, afirmamos que hay una variable independiente o exógena X, cuyos valores son determinados por actividades realizadas fuera del sistema económico examinado y que hay una variable dependiente o endógena Y, cuyo valor depende del valor de X. Si preguntamos si las ventas aumentan cuando bajan los precios, estamos analizando una situación en la que un vendedor ajusta de una forma deliberada e independiente los precios en sentido ascendente o descendente y observa cómo varían las ventas. Supongamos ahora que los precios y las cantidades vendidas son el resultado de equilibrios de la oferta y la demanda como propone el modelo económico básico. En ese caso, podríamos analizar los precios y las cantidades como variables aleatorias y preguntarnos si estas dos variables aleatorias están relacionadas entre sí. El coeficiente de correlación puede utilizarse para averiguar si existe una relación entre variables en cualquiera de estas dos situaciones.

Supongamos que tanto X como Y son determinados simultáneamente por factores que se encuentran fuera del sistema económico analizado. Por lo tanto, suele ser más realista plantear un modelo en el que tanto X como Y sean variables aleatorias. En el Capítulo 5 presentamos el coeficiente de correlación ρ_{xy} como medida de la relación entre dos variables aleatorias, X e Y. En esos casos, utilizamos el coeficiente de correlación poblacional, ρ_{xy}, para indicar la existencia de una relación lineal sin que ello quisiera decir que una de las variables era independiente y la otra dependiente. En las situaciones en las que una de las variables es dependiente lógicamente de otra, el siguiente paso lógico después del análisis de correlación es la utilización del análisis de regresión para desarrollar el modelo lineal. Éste es el tema del siguiente apartado. Aquí presentamos métodos de inferencia estadística que utilizan correlaciones muestrales para averiguar las características de las correlaciones poblacionales.

Contraste de hipótesis de la correlación

El coeficiente de correlación muestral

$$r = \frac{s_{xy}}{s_x s_y}$$

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

es una medida descriptiva útil de la fuerza de la relación lineal en una muestra. También podemos utilizar la correlación para contrastar la hipótesis de que no existe una relación lineal en la población entre un par de variables aleatorias; es decir,

$$H_0 : \rho = 0$$

Esta hipótesis nula de que no existe una relación lineal entre un par de variables aleatorias es muy interesante en algunas aplicaciones. Cuando calculamos la correlación muestral a partir de datos, es probable que el resultado sea diferente de 0 aunque la correlación poblacional sea 0. Nos gustaría, pues, saber en qué medida debe ser diferente de 0 una correlación muestral para contar con una prueba de que la correlación poblacional no es 0.

Podemos demostrar que cuando la hipótesis nula es verdadera y las variables aleatorias siguen una distribución normal conjunta, la variable aleatoria

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

sigue una distribución t de Student con $(n - 2)$ grados de libertad. Las ecuaciones 12.1 a 12.3 muestran los contrastes de hipótesis adecuados.

Contrastes de la correlación poblacional nula

Sea r el coeficiente de correlación muestral, calculado a partir de una muestra aleatoria de n pares de observaciones de una distribución normal conjunta. Los siguientes contrastes de la hipótesis nula

$$H_0 : \rho = 0$$

tienen un valor de significación α:

1. Para contrastar H_0 frente a la hipótesis alternativa

$$H_1 : \rho > 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha} \tag{12.1}$$

2. Para contrastar H_0 frente a la hipótesis alternativa

$$H_1 : \rho < 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2,\alpha} \tag{12.2}$$

3. Para contrastar H_0 frente a la hipótesis alternativa bilateral

$$H_1 : \rho \neq 0$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} < -t_{n-2,\alpha/2} \quad \text{o} \quad \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha/2} \tag{12.3}$$

Aquí, $t_{n-2,\alpha}$ es el número para el que

$$P(t_{n-2} > t_{n-2,\alpha}) = \alpha$$

donde la variable aleatoria t_{n-2} sigue una distribución t de Student con $(n-2)$ grados de libertad.

4. Si introducimos $t_{n-2,\alpha/2} = 2{,}0$ en la ecuación 12.3, podemos demostrar que una «regla práctica» aproximada para contrastar la hipótesis anterior de que la correlación poblacional es 0 es

$$|r| > \frac{2}{\sqrt{n}}$$

EJEMPLO 12.1. Valoración del riesgo político (contraste de hipótesis de la correlación)

Un equipo de investigación estaba intentando averiguar si el riesgo político existente en los países está relacionado con su inflación. En esta investigación, se realizó una encuesta a analistas del riesgo político que permitió elaborar una puntuación media del riesgo político de 49 países (los datos proceden del estudio mencionado en la referencia bibliográfica 2).

Solución

Cuanto más alta es la puntuación, mayor es el riesgo político. La correlación muestral entre la puntuación del riesgo político y la inflación de estos países era de 0,43.

Queremos averiguar si la correlación poblacional, ρ, entre estas medidas es diferente de 0. Concretamente, queremos contrastar

$$H_0: \rho = 0$$

frente a

$$H_1: \rho > 0$$

utilizando la información muestral

$$n = 49 \qquad r = 0{,}43$$

El contraste se basa en el estadístico

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} = \frac{0{,}43\sqrt{(49-2)}}{\sqrt{1-(0{,}43)^2}} = 3{,}265$$

Dado que hay $(n - 2) = 47$ grados de libertad, vemos en la tabla 8 de la t de Student del apéndice que

$$t_{47,\,0{,}005} < 2{,}704$$

Por lo tanto, podemos rechazar la hipótesis nula al nivel de significación del 0,5 por ciento. Tenemos, pues, pruebas contundentes de que existe una relación lineal positiva entre la inflación y la valoración de los expertos del riesgo político de los países. Obsérvese que de este resultado no podemos extraer la conclusión de que una de las variables es la causa de la otra, sólo que están relacionadas.

Antes hemos señalado que la hipótesis nula $H_0: \rho = 0$ puede rechazarse utilizando la regla práctica aproximada $|r| > 2/\sqrt{n}$. Este resultado proporciona un rápido contraste para averiguar si dos variables están relacionadas linealmente cuando se examinan una o más correlaciones muestrales. Así, por ejemplo, en el caso de una muestra de tamaño $n = 25$, el valor absoluto de la correlación muestral tendría que ser superior a $2/\sqrt{25} = 0{,}40$. Pero en el caso de una muestra de tamaño $n = 64$, el valor absoluto de la correlación muestral tendría que ser superior a $2/\sqrt{64} = 0{,}25$ solamente. Se ha observado que este resultado es útil en muchas aplicaciones estadísticas.

EJERCICIOS

Ejercicios básicos

12.1. Dados los pares siguientes de (x, y) observaciones, calcule la correlación muestral.

a) (2, 5), (5, 8), (3, 7), (1, 2), (8, 15).
b) (7, 5), (10, 8), (8, 7), (6, 2), (13, 15).
c) (12, 4), (15, 6), (16, 5), (21, 8), (14, 6).
d) (2, 8), (5, 12), (3, 14), (1, 9), (8, 22).

12.2. Contraste la hipótesis nula

$$H_0: \rho = 0 \quad \text{frente a} \quad H_1: \rho \neq 0$$

dada

a) Una correlación muestral de 0,35 en una muestra aleatoria de tamaño $n = 40$
b) Una correlación muestral de 0,50 en una muestra aleatoria de tamaño $n = 60$

c) Una correlación muestral de 0,62 en una muestra aleatoria de tamaño $n = 45$

d) Una correlación muestral de 0,60 en una muestra aleatoria de tamaño $n = 25$

12.3. El profesor de un curso de estadística puso un examen final y también pidió a los estudiantes que realizaran un proyecto. La tabla adjunta muestra las calificaciones de una muestra aleatoria de 10 estudiantes. Halle la correlación muestral entre las calificaciones del examen y las del proyecto.

Examen	81	62	74	78	93	69	72	83	90	84
Proyecto	76	71	69	76	87	62	80	75	92	79

Ejercicios aplicados

12.4. En el estudio de 49 países analizado en el ejemplo 12.1, la correlación muestral entre la valoración del riesgo político realizada por los expertos y la tasa de mortalidad infantil de estos países era 0,75. Contraste la hipótesis nula de que no existe ninguna correlación entre estas cantidades frente a la hipótesis alternativa de que existe una correlación positiva.

12.5. En una muestra aleatoria de 353 profesores de enseñanza secundaria, se observó que la correlación entre las subidas salariales anuales y las evaluaciones de la docencia era de 0,11. Contraste la hipótesis nula de que estas cantidades no están correlacionadas en la población frente a la hipótesis alternativa de que la correlación poblacional es positiva.

12.6. Se observa que la correlación muestral de 68 pares de rendimientos anuales de acciones ordinarias del país A y del país B es de 0,51. Contraste la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa de que es positiva.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

12.7. La tabla adjunta y el fichero de datos **Dow Jones** muestran las variaciones porcentuales (x_i) del índice Dow-Jones registradas en los cinco primeros días de sesión de cada uno de los años de un periodo de 13 años y las correspondientes variaciones porcentuales (y_i) del índice a lo largo de todo el año.

x	y	x	y
1,5	14,9	5,6	2,3
0,2	−9,2	−1,4	11,9
−0,1	19,6	1,4	27,0
2,8	20,3	1,5	−4,3
2,2	−3,7	4,7	20,3
−1,6	27,7	1,1	4,2
−1,3	22,6		

a) Calcule la correlación muestral.

b) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa bilateral.

12.8. Una universidad distribuye en todos sus cursos un cuestionario de evaluación para que lo rellenen los estudiantes. La tabla adjunta y el fichero de datos **Student Evaluation** muestran tanto la valoración media del profesor (en una escala de 1 a 5) como la calificación media esperada (en una escala de A = 4 a E = 0) de una muestra aleatoria de 12 cursos.

Valoración del profesor	2,8	3,7	4,4	3,6	4,7	3,5	4,1	3,2	4,9	4,2	3,8	3,3
Calificación esperada	2,6	2,9	3,3	3,2	3,1	2,8	2,7	2,4	3,5	3,0	3,4	2,5

a) Halle la correlación muestral entre las valoraciones de los profesores y las calificaciones esperadas.

b) Contraste al nivel de significación del 10 por ciento la hipótesis de que el coeficiente de correlación poblacional es 0 frente a la hipótesis alternativa de que es positivo.

12.9. En un estudio sobre la publicidad, los investigadores querían saber si existía una relación entre el coste per cápita y los ingresos per cápita. Se midieron las siguientes variables en una muestra aleatoria de programas de publicidad:

x_i = coste de la publicidad ÷ n.º de preguntas recibidas

y_i = ingresos generados por las preguntas ÷ n.º de preguntas recibidas

Los datos muestrales se encuentran en el fichero de datos **Advertising Revenue**. Halle la correlación muestral y contraste la hipótesis nula de que la correlación poblacional es 0 frente a la alternativa bilateral.

12.2. Modelo de regresión lineal

Para medir la fuerza de cualquier relación lineal entre un par de variables aleatorias se utilizan coeficientes de correlación. Las variables aleatorias se tratan de una forma totalmente simétrica y da lo mismo que hablemos de «la correlación entre X e Y» que de «la correlación entre Y y X». En el resto de este capítulo, continuamos analizando la relación lineal entre un par de variables, pero desde el punto de vista de la dependencia de una de la otra. Ahora dejamos de tratar las variables aleatorias de una forma simétrica. La idea es que, dado que la variable aleatoria X toma un valor específico, esperamos una respuesta de la variable aleatoria Y. Es decir, el valor que toma X influye en el valor de Y. Podemos pensar que Y depende de X. Las variables dependientes o endógenas —Y— tienen valores que dependen de variables independientes o exógenas —X—, cuyos valores son manipulados o influidos, a su vez, por factores externos a un proceso económico específico.

Los modelos lineales no son tan restrictivos como podría parecer para el análisis empresarial y económico aplicado. En primer lugar, los modelos lineales a menudo constituyen una buena aproximación de una relación en el intervalo examinado. En segundo lugar, en los Capítulos 13 y 14 veremos que algunas funciones no lineales pueden convertirse en funciones lineales implícitas para el análisis de regresión.

En este capítulo realizamos un estudio formal del análisis de regresión y de la correspondiente inferencia estadística en el caso de modelos lineales sencillos. En los Capítulos 2 y 3 introdujimos los instrumentos de los diagramas de puntos dispersos, la correlación y la regresión simple para describir datos. En el 13 aplicaremos estas ideas a los modelos de regresión múltiple que tienen más de una variable de predicción y en el 14 presentamos métodos y aplicaciones avanzados que aumentan nuestra capacidad para analizar problemas empresariales y económicos.

Este análisis comienza con un ejemplo que muestra una aplicación representativa del análisis de regresión y el tipo de resultados que pueden obtenerse.

EJEMPLO 12.2. Predicción sobre las ventas de Northern Household Goods (estimación de un modelo de regresión)

El presidente de Northern Household Goods le ha pedido que desarrolle un modelo que prediga las ventas totales de las nuevas tiendas que se propone abrir. Northern es una cadena de grandes almacenes en rápida expansión y necesita una estrategia racional para averiguar dónde deben abrirse nuevas tiendas. Para realizar este proyecto, necesita estimar una ecuación lineal que prediga las ventas al por menor por hogar en función de la renta disponible del hogar. La empresa ha obtenido datos de una encuesta nacional realizada a los hogares y para desarrollar el modelo se utilizarán las variables de las ventas al por menor (Y) y la renta (X) por hogar.

Solución

Retail Sales

La Figura 12.1 es un diagrama de puntos dispersos que muestra la relación entre las ventas al por menor y la renta disponible de las familias. Los datos efectivos se muestran en la Tabla 12.1 y se encuentran en el fichero de datos llamado **Retail Sales**. Según la teoría económica, las ventas deben aumentar cuando aumenta la renta disponible y el diagrama de puntos dispersos apoya en gran medida esa teoría. El análisis de regresión nos proporciona un modelo lineal que puede utilizarse para calcular las ventas al por

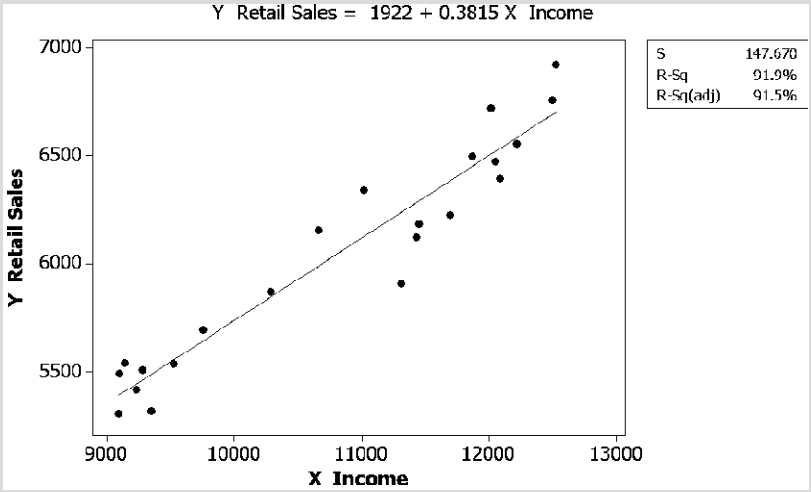


Figura 12.1. Ventas al por menor por hogar en relación con la renta disponible per cápita.

Tabla 12.1. Datos sobre la renta disponible por hogar (*X*) y ventas al por menor por hogar (*Y*).

Año	Renta (*X*)	Ventas al por menor (*Y*)	Año	Renta (*X*)	Ventas al por menor (*Y*)
1	9.098	5.492	12	11.307	5.907
2	9.138	5.540	13	11.432	6.124
3	9.094	5.305	14	11.449	6.186
4	9.282	5.507	15	11.697	6.224
5	9.229	5.418	16	11.871	6.496
6	9.347	5.320	17	12.018	6.718
7	9.525	5.538	18	12.523	6.921
8	9.756	5.692	19	12.053	6.471
9	10.282	5.871	20	12.088	6.394
10	10.662	6.157	21	12.215	6.555
11	11.019	6.342	22	12.494	6.755

menor por hogar correspondientes a varios niveles de renta disponible. La recta del diagrama representa el modelo de regresión simple

$$Y = 1.922{,}39 + 0{,}381517X$$

donde *Y* son las ventas al por menor por hogar y *X* es la renta disponible por hogar. Por lo tanto, la ecuación de regresión nos proporciona, a partir de los datos, el mejor modelo lineal para predecir las ventas correspondientes a una renta disponible dada. Obsérvese que este modelo nos dice que cada aumento de la renta familiar disponible per cápita de 1 $, *X*, va acompañado de un aumento del valor esperado de las ventas al por menor, *Y*, de 0,38 $. Es evidente que el resultado es importante para predecir las ventas al por menor. Por ejemplo, observamos que una renta familiar de 50.000 $ predeciría que las ventas al por menor serán de 20.997 $ (1.922 + 50.000 × 0,3815).

INTERPRETACIÓN

Llegados a este punto, debemos hacer hincapié en que los resultados de la regresión resumen la información que contienen los datos y no «demuestran» que el aumento de la renta sea la «causa» del aumento de las ventas. La teoría económica sugiere que existe una relación causal y estos resultados apoyan esta teoría. Los diagramas de puntos dispersos, las correlaciones y las ecuaciones de regresión no pueden demostrar la existencia de una relación causal, pero pueden aportar pruebas a su favor. Así pues, para extraer conclusiones, necesitamos conjugar la teoría —la experiencia en la administración de empresas y el análisis económico— con un buen análisis estadístico.

Sabemos por nuestros estudios de la economía que la cantidad comprada de bienes, Y, en un mercado específico puede representarse por medio de una función lineal de la renta disponible, X. Si la renta tiene un nivel específico, x_i, los compradores responden comprando la cantidad y_i. En el mundo real, sabemos que hay otros factores que influyen en la cantidad efectiva comprada. Son factores identificables como el precio de los bienes en cuestión, la publicidad y los precios de los bienes rivales. También hay otros factores desconocidos que pueden influir en la cantidad efectiva comprada. En una ecuación lineal simple, representamos el efecto de estos factores, salvo la renta, por medio de un término de error llamado ε.

La Figura 12.2 muestra un ejemplo de un conjunto de observaciones generadas por un modelo lineal subyacente de un proceso. El nivel medio de Y, para todo X, se representa por medio de la ecuación poblacional

$$Y = \beta_0 + \beta_1 X$$

El modelo de regresión lineal permite hallar el valor esperado de la variable aleatoria Y cuando X toma un valor específico. El supuesto de la linealidad implica que esta esperanza puede expresarse de la forma siguiente:

$$E(Y \mid X = x) = \beta_0 + \beta_1 X$$

donde β_0 representa la ordenada en el origen Y de la ecuación y β_1 es la pendiente. El valor observado efectivo de Y para un valor dado de X es igual al valor esperado o media poblacional más un error aleatorio, ε, que tiene una media 0 y una varianza σ^2:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

El término de error aleatorio ε representa la variación de Y que no es estimada por la relación lineal.

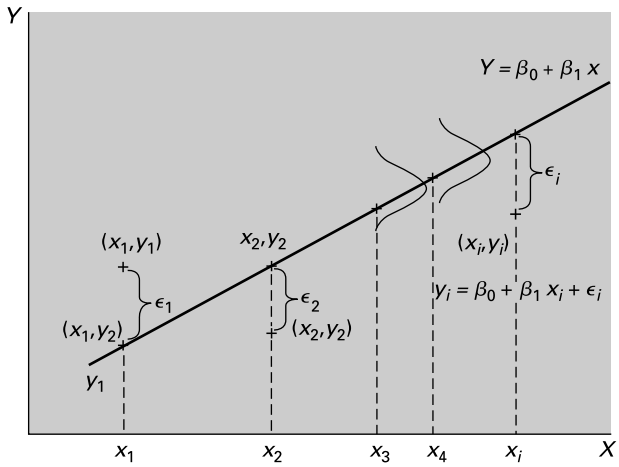


Figura 12.2. Modelo de regresión lineal poblacional.

La regresión por mínimos cuadrados nos proporciona un modelo estimado de la relación lineal entre una variable independiente o exógena y una variable dependiente o endógena. Comenzamos el proceso de formulación de la regresión partiendo de un modelo poblacional en el que X tiene unos valores predeterminados y para todo X hay un valor medio de Y más un término de error aleatorio. Utilizamos la ecuación de regresión estimada —mostrada en la Figura 12.1— para estimar el valor medio de Y para todo valor de X. Los puntos no están alineados siempre en esta recta debido a que existe un término de error aleatorio que tiene una media 0 y una varianza común para todos los valores de X. El error aleatorio representa todos los factores que influyen en Y que no están representados por la relación lineal entre Y y X. Los efectos de estos factores, que se supone que son independientes de X, se comportan como una variable aleatoria cuya media poblacional es 0. Las desviaciones aleatorias ε_i en torno al modelo lineal se muestran en la Figura 12.2 y se combinan con la media de Y_i para todo X_i para obtener el valor observado y_i.

Regresión lineal basada en un modelo poblacional

En la aplicación del análisis de regresión, se representa el proceso estudiado por medio de un modelo poblacional y se calcula un modelo estimado utilizando los datos de que se dispone y realizando una regresión por mínimos cuadrados. El modelo poblacional es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \textbf{(12.4)}$$

donde β_0 y β_1 son los coeficientes del modelo poblacional y ε es un término de error aleatorio. Para todo valor observado, x_i, el modelo poblacional genera un valor observado, y_i. Para realizar la inferencia estadística, como veremos en el apartado 12.4, se supone que ε sigue una distribución normal de media 0 y varianza σ^2. Más adelante, veremos que puede utilizarse el teorema del límite central para abandonar el supuesto de la distribución normal. El modelo de la relación lineal entre Y y X viene definido por los dos coeficientes, β_0 y β_1. La Figura 12.2 lo representa esquemáticamente.

En el modelo de regresión por mínimos cuadrados suponemos que se seleccionan valores de la variable independiente, x_i, y para cada x_i existe una media poblacional de Y. Los valores observados de y_i contienen la media y la desviación aleatoria ε_i. Se observa un conjunto de $n(x_i, y_i)$ puntos y se utiliza para obtener estimaciones de los coeficientes del modelo utilizando el método de mínimos cuadrados. Ampliamos los conceptos de la inferencia clásica presentados en los Capítulos 8 a 11 para hacer inferencias sobre el modelo poblacional subyacente utilizando el modelo de regresión estimado. En el Capítulo 13 veremos cómo pueden considerarse simultáneamente varias variables independientes utilizando la regresión múltiple.

El modelo de regresión estimado y mostrado esquemáticamente en la Figura 12.3 viene dado por la ecuación

$$y_i = b_0 + b_1 x_i + e_i$$

donde b_0 y b_1 son los valores estimados de los coeficientes y e es la diferencia entre el valor predicho de Y en la recta de regresión

$$\hat{y}_i = b_0 + b_1 x_i$$

y el valor observado y_i. La diferencia entre y_i e $\hat{y}_i$ para cada valor de X es el residuo

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (b_0 + b_1 x_i) \end{aligned}$$

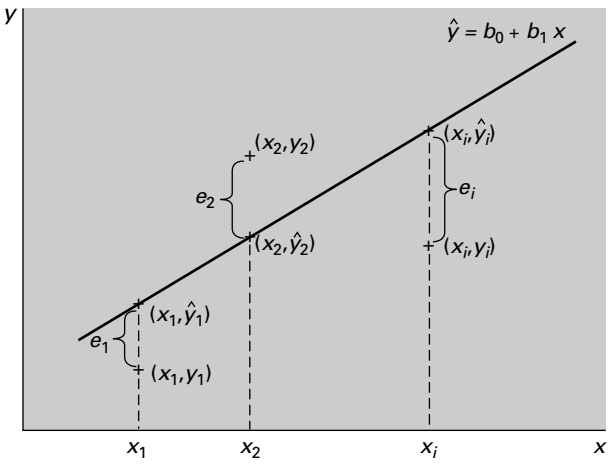


Figura 12.3. Modelo de regresión estimado.

Por lo tanto, para cada valor observado de X hay un valor predicho de Y a partir del modelo estimado y un valor observado. La diferencia entre el valor observado de Y y el predicho es el residuo, e_i. El residuo, e_i, no es el error del modelo, ε, sino la medida combinada del error del modelo y los errores de la estimación de b_0 y b_1 y, a su vez, los errores de la estimación del valor predicho.

Hallamos el modelo de regresión estimado obteniendo estimaciones, b_0 y b_1, de los coeficientes poblacionales utilizando el método llamado análisis de mínimos cuadrados, que presentamos en el apartado 12.3. Empleamos, a su vez, estos coeficientes para obtener los valores predichos de Y para todo valor de X.

Resultados de la regresión lineal

La regresión lineal da dos importantes resultados:

1. Los valores predichos de la variable dependiente o endógena en función de la variable independiente o exógena.
2. La variación marginal estimada de la variable endógena provocada por una variación unitaria de la variable independiente o exógena.

EJERCICIOS

Ejercicios básicos

12.10. Dada la ecuación de regresión

$$Y = 100 + 10X$$

a) ¿Cuál es la variación de Y cuando X varía en $+3$?

b) ¿Cuál es la variación de Y cuando X varía en -4?

c) ¿Cuál es el valor predicho de Y cuando $X = 12$?

d) ¿Cuál es el valor predicho de Y cuando $X = 23$?

e) ¿Demuestra esta ecuación que una variación de X provoca una variación de Y?

12.11. Dada la ecuación de regresión

$$Y = -50 + 12X$$

a) ¿Cuál es la variación de Y cuando X varía en $+3$?

b) ¿Cuál es la variación de Y cuando X varía en -4?

c) ¿Cuál es el valor predicho de Y cuando $X = 12$?

d) ¿Cuál es el valor predicho de Y cuando $X = 23$?

e) ¿Demuestra esta ecuación que una variación de X provoca una variación de Y?

12.12. Dada la ecuación de regresión

$$Y = 43 + 10X$$

a) ¿Cuál es la variación de Y cuando X varía en $+8$?
b) ¿Cuál es la variación de Y cuando X varía en -6?
c) ¿Cuál es el valor predicho de Y cuando $X = 11$?
d) ¿Cuál es el valor predicho de Y cuando $X = 29$?)
e) ¿Demuestra esta ecuación que una variación de X provoca una variación de Y?

12.13. Dada la ecuación de regresión

$$Y = 100 + 21X$$

a) ¿Cuál es la variación de Y cuando X varía en $+5$?
b) ¿Cuál es la variación de Y cuando X varía en -7?
c) ¿Cuál es el valor predicho de Y cuando $X = 14$?
d) ¿Cuál es el valor predicho de Y cuando $X = 27$?
e) ¿Demuestra esta ecuación que una variación de X provoca una variación de Y?

Ejercicios aplicados

12.14. ¿Qué diferencia existe entre un modelo lineal poblacional y un modelo de regresión lineal estimado?

12.15. Explique la diferencia entre el residuo e_i y el error del modelo ε_i.

12.16. Suponga que hemos estimado una ecuación de la regresión de las ventas semanales de «palm pilot» y el precio cobrado durante la semana. Interprete la constante b_0 para el director de la marca.

12.17. Se ha estimado un modelo de regresión de las ventas totales de productos alimenticios con respecto a la renta disponible utilizando datos de pequeñas ciudades aisladas del oeste de Estados Unidos. Elabore una lista de los factores que podrían contribuir al término de error aleatorio.

12.3. Estimadores de coeficientes por el método de mínimos cuadrados

La recta de regresión poblacional es un útil instrumento teórico, pero para las aplicaciones necesitamos estimar el modelo utilizando los datos de que se disponga. Supongamos que tenemos n pares de observaciones, (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Nos gustaría encontrar la línea recta que mejor se ajusta a estos puntos. Para ello, es necesario encontrar estimadores de los coeficientes desconocidos β_0 y β_1 de la recta de regresión poblacional.

Hallamos los estimadores de los coeficientes b_0 y b_1 con ecuaciones obtenidas utilizando el método de mínimos cuadrados. Como mostramos en la Figura 12.3, hay una desviación, e_i, entre el valor observado, y_i, y el valor predicho, $\hat{y}_i$, en la ecuación de regresión estimada para cada valor de X, donde $e_i = y_i - \hat{y}_i$. A continuación, calculamos una función matemática consistente en elevar al cuadrado todos los residuos y sumar las cantidades resultantes. Esta función —cuyo primer miembro se denomina *SCE*— incluye los coeficientes b_0 y b_1. La cantidad *SCE* se denomina *suma de los cuadrados de los errores*. Los estimadores de los coeficientes b_0 y b_1 son los estimadores que minimizan la suma de los cuadrados de los errores.

Método de mínimos cuadrados

El método de mínimos cuadrados obtiene estimaciones de los coeficientes de la ecuación lineal b_0 y b_1 en el modelo

$$\hat{y}_i = b_0 + b_1 x_i \qquad \textbf{(12.5)}$$

minimizando la suma de los cuadrados de los errores e_i:

$$SCE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \qquad \textbf{(12.6)}$$

Los coeficientes b_0 y b_1 se eligen de tal manera que se minimice la cantidad

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \qquad \textbf{(12.7)}$$

Utilizamos el cálculo diferencial para obtener los estimadores de los coeficientes que minimizan la *SCE*. En el apéndice del capítulo se explica cómo se obtienen los estimadores por medio del cálculo.

El estimador del coeficiente resultante es

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{Cov}\,(x,\, y)}{s_x^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x}) x_i} y_i$$

Obsérvese que el numerador del estimador es la covarianza muestral de X e Y y el denominador es la varianza muestral de X. La tercera línea muestra que el coeficiente b_1 es una función lineal de las Y. Dedicamos mucho tiempo al coeficiente de la pendiente porque este resultado es clave para muchas aplicaciones. El coeficiente de la pendiente b_1 es una estimación de la variación que experimenta Y cuando X varía en una unidad. Por ejemplo, si Y es la producción total y X es el número de trabajadores, entonces b_1 es una estimación del aumento marginal de la producción por cada nuevo trabajador. Este tipo de resultados explica por qué la regresión se ha convertido en un instrumento analítico tan importante.

Con algunas manipulaciones algebraicas podemos demostrar que el estimador del coeficiente también es igual a

$$b_1 = r \frac{s_y}{s_x}$$

donde r_{xy} es la correlación muestral y s_y y s_x son las desviaciones típicas muestrales de X e Y. Este resultado es importante porque indica cómo está relacionada directamente la relación estandarizada entre X e Y, la correlación r_{xy}, con el coeficiente de la pendiente.

En el apéndice del capítulo también mostramos que el estimador de la constante es

$$b_0 = \bar{y} - b_1 \bar{x}$$

Sustituyendo b_0 por este valor en la ecuación lineal, tenemos que

$$y = \bar{y} - b_1 \bar{x} + b_1 x$$
$$y - \bar{y} = b_1 (x - \bar{x})$$

En esta ecuación vemos que cuando $x = \bar{x}$, entonces $y = \bar{y}$ y que la ecuación de regresión siempre pasa por el punto $(\bar{x}, \bar{y})$. El valor estimado de la variable dependiente, $\hat{y}_i$, se obtiene utilizando

$$\hat{y}_i = b_0 + b_1 x_i$$

o utilizando

$$\hat{y}_i = \bar{y} + b_1(x_i - \bar{x})$$

Esta última forma pone de relieve que la recta de regresión pasa por las medias de X e Y.

Estimadores de coeficientes por el método de mínimos cuadrados

El estimador del coeficiente de la pendiente es

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r\,\frac{s_y}{s_x}$$

y el estimador de la constante u ordenada en el origen es

$$b_0 = \bar{y} - b_1\bar{x}$$

También señalamos que la recta de regresión siempre pasa por la media $\bar{x}$, $\bar{y}$.

El método de mínimos cuadrados podría utilizarse para calcular estimaciones de los coeficientes b_0 y b_1 utilizando cualquier conjunto de datos pareados. Sin embargo, en la mayoría de las aplicaciones queremos hacer inferencias sobre el modelo poblacional subyacente que forma parte de nuestro problema económico o empresarial. Para hacer inferencias, es necesario que estemos de acuerdo en ciertos supuestos. Dados estos supuestos, puede demostrarse que los estimadores de los coeficientes por mínimos cuadrados son insesgados y tienen una varianza mínima.

Supuestos habituales en los que se basa el modelo de regresión lineal

Para hacer inferencias sobre el modelo lineal poblacional utilizando los coeficientes del modelo estimados se postulan los siguientes supuestos.

1. Las Y son funciones lineales de X más un término de error aleatorio

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

2. Las x son números fijos o son realizaciones de la variable aleatoria X que son independientes de los términos de error, ε_i. En el segundo caso, la inferencia se realiza condicionada a los valores observados de las x.
3. Los términos de error son variables aleatorias que tienen la media 0 y la misma varianza σ^2. El segundo supuesto se llama homocedasticidad o varianza uniforme.

$$E[\varepsilon_i] = 0 \quad \text{y} \quad E[\varepsilon_i^2] = \sigma^2 \quad \text{para } (i = 1, ..., n)$$

4. Los términos de error aleatorio, ε_i, no están correlacionados entre sí, por lo que

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{para todo } i \neq j$$

Generalmente, se considera, con razón, que el segundo de estos supuestos es cierto, aunque en algunos estudios econométricos avanzados es insostenible (el supuesto no se cumple, por ejemplo, cuando no es posible medir x_i con precisión o cuando la regresión forma parte de un sistema de ecuaciones interdependientes). Sin embargo, aquí consideraremos que se satisface este supuesto.

Los supuestos 3 y 4 se refieren a los términos de error, ε_i, de la ecuación de regresión. El término de error esperado es 0 y todos los términos de error tienen la misma varianza. Por lo tanto, no esperamos que las varianzas de los términos de error sean más altas en el caso de algunas observaciones que en el de otras. La Figura 12.2 muestra esta pauta: los errores correspondientes a todos los valores de X proceden de poblaciones que tienen la misma varianza. Por último, se supone que las discrepancias no están correlacionadas entre sí. Así, por ejemplo, la aparición de una gran discrepancia positiva en un punto de observación no nos ayuda a predecir los valores de ninguno de los demás términos de error. Los supuestos 3 y 4 se satisfacen si los términos de error, ε_i, pueden concebirse como una muestra aleatoria procedente de una población que tiene de media 0. En el resto de este capítulo, estos supuestos se cumplen. La posibilidad de abandonar algunos de ellos se examina en el Capítulo 14.

Cálculo por computador del coeficiente de regresión

La extensa aplicación del análisis de regresión ha sido posible gracias a los paquetes estadísticos y a Excel. Como sospechará el lector, los cálculos para obtener estimaciones de los coeficientes de regresión son tediosos. Las ecuaciones de los estimadores y otros importantes cálculos estadísticos están incluidos en los paquetes informáticos y en Excel y se utilizan para estimar los coeficientes de problemas específicos. El programa Excel puede utilizarse para realizar análisis básicos de regresión sin demasiadas dificultades. Pero si se desea utilizar métodos de análisis de regresión aplicado avanzado o un perspicaz análisis gráfico, debe utilizarse un buen paquete estadístico. Dado que nos interesan principalmente las aplicaciones, nuestra tarea más importante es realizar un análisis adecuado de los cálculos de regresión para estas aplicaciones. Este análisis debe realizarse conociendo las ecuaciones de los estimadores y el análisis relacionado con ellas. Sin embargo, no utilizamos estas ecuaciones para calcular realmente las estimaciones u otros estadísticos de la regresión. *Dejamos los cálculos para los computadores*; *nuestra tarea es pensar*, *analizar y hacer recomendaciones*.

La Figura 12.4 muestra una parte de las salidas Minitab y Excel correspondientes al ejemplo de las ventas al por menor. Obsérvese la localización de las estimaciones de la constante, b_0, y el coeficiente de la pendiente, b_1, en la salida informática. Los conceptos restantes de cada línea ayudan a interpretar la calidad de las estimaciones y se explican en apartados posteriores.

En esta regresión, la constante estimada, b_0, es 1.922 y el coeficiente de la pendiente estimado, b_1, es 0,382. Estos valores se calculan utilizando las ecuaciones de los estimadores de los coeficientes antes presentadas. La ecuación estimada puede expresarse de la forma siguiente:

$$\hat{y} = 1.922 + 0{,}382x$$

o, utilizando las medias $\bar{x} = 10.799$ e $\bar{y} = 6.042$, de la forma siguiente:

$$\hat{y} = 6.042 + 0{,}382(x - 10.799)$$

Normalmente, los modelos de regresión sólo deben utilizarse en el rango de los valores observados de X en el que tenemos información sobre la relación porque la relación puede no ser lineal fuera de este rango. La segunda forma del modelo de regresión está centrada en las medias de los datos con una tasa de variación igual a b_1. Utilizando esta forma, centramos la atención en la localización media del modelo de regresión y no en la ordenada

Results for: retail sales.MTW

Regression Analysis: Y Retail Sales versus X Income

```
The regression equation is
Y Retail Sales = 1922 + 0.382 X Income

Predictor       Coef     SE Coef        T        P
Constant      1922.4       274.9     6.99    0.000
X Income     0.38152     0.02529    15.08    0.000

S = 147.670      R-Sq = 91.9%     R-Sq(adj) = 91.5%
```

Coeficientes b_0, b_1

(a)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.958748803					
5	R Square	0.919199267					
6	Adjusted R Square	0.91515923					
7	Standard Error	147.6697181					
8	Observations	22					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	4961434.406	4961434	227.5225	2.17134E-12	
13	Residual	20	436126.9127	21806.35			
14	Total	21	5397561.318				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	1922.392694	274.9493737	6.991806	8.74E-07	1348.858617	2495.92677
18	X Income	0.38151672	0.025293061	15.08305	2.17E-12	0.328756343	0.4342771
19							

Coeficientes b_0, b_1

(b)

Figura 12.4. Análisis de regresión de las ventas al por menor (a) por medio de Minitab y (b) por medio de Excel.

en el origen con el eje de las *Y*. Los usuarios ingenuos del análisis de regresión a veces intentan hacer interpretaciones de la constante b_0, extrayendo ciertas conclusiones sobre la variable dependiente cuando la variable independiente tiene un valor de 0. Consideremos la regresión de las ventas al por menor con respecto a la renta disponible del ejemplo. ¿Afirmaríamos realmente que las ventas al por menor son de 1.922 $ cuando la renta disponible es de 0? En realidad, sencillamente no tenemos datos para afirmar que se vende algo cuando la renta disponible es 0. Éste es otro ejemplo de la importancia de un buen análisis en lugar de interpretaciones tontas. Como analistas profesionales, debemos tener cuidado de no defender resultados que sencillamente no existen.

EJERCICIOS

Ejercicios básicos

12.18. Calcule los coeficientes de una ecuación de regresión por mínimos cuadrados y formule la ecuación, dados los siguientes estadísticos muestrales:

a) $\bar{x} = 50$; $\bar{y} = 100$; $s_x = 25$; $s_y = 75$; $r_{xy} = 0{,}6$; $n = 60$

b) $\bar{x} = 60$; $\bar{y} = 210$; $s_x = 35$; $s_y = 65$; $r_{xy} = 0{,}7$; $n = 60$

c) $\bar{x} = 20$; $\bar{y} = 100$; $s_x = 60$; $s_y = 78$; $r_{xy} = 0{,}75$; $n = 60$

d) $\bar{x} = 10$; $\bar{y} = 50$; $s_x = 100$; $s_y = 75$; $r_{xy} = 0{,}4$; $n = 60$

e) $\bar{x} = 90$; $\bar{y} = 200$; $s_x = 80$; $s_y = 70$; $r_{xy} = 0{,}6$; $n = 60$

Ejercicios aplicados

12.19. Una empresa fija un precio distinto para un sistema de DVD en ocho regiones del país. La ta-

bla adjunta muestra los números de unidades vendidas y los precios correspondientes (en cientos de dólares).

Ventas	420	380	350	400	440	380	450	420
Precio	5,5	6,0	6,5	6,0	5,0	6,5	4,5	5,0

a) Represente estos datos y estime la regresión lineal de las ventas con respecto al precio.
b) ¿Qué efecto sería de esperar que produjera una subida del precio de 100 $ en las ventas?

12.20. Dada una muestra de 20 observaciones mensuales, un analista financiero quiere realizar una regresión de la tasa porcentual de rendimiento (Y) de las acciones ordinarias de una empresa con respecto a la tasa porcentual de rendimiento (X) del índice Standard and Poor's 500. Dispone de la siguiente información:

$$\sum_{i=1}^{20} y_i = 22{,}6 \qquad \sum_{i=1}^{20} x_i = 25{,}4$$

$$\sum_{i=1}^{20} x_i^2 = 145{,}7 \qquad \sum_{i=1}^{20} x_i y_i = 150{,}5$$

a) Estime la regresión lineal de Y con respecto a X.
b) Interprete la pendiente de la recta de regresión muestral.
c) Interprete la ordenada en el origen de la recta de regresión muestral.

12.21. Una empresa realiza un test de aptitud a todos los nuevos representantes de ventas. La dirección tiene interés en saber en qué medida es capaz este test de predecir su éxito final. La tabla adjunta muestra las ventas semanales medias (en miles de dólares) y las puntuaciones obtenidas en el test de aptitud por una muestra aleatoria de ocho representantes.

Ventas semanales	10	12	28	24	18	16	15	12
Puntuación	55	60	85	75	80	85	65	60

a) Estime la regresión lineal de las ventas semanales con respecto a las puntuaciones del test de aptitud.
b) Interprete la pendiente estimada de la recta de regresión.

12.22. Se ha formulado la hipótesis de que el número de botellas de una cerveza importada que se vende cada noche en los restaurantes de una ciudad depende linealmente de los costes medios de las cenas en los restaurantes. Se han obtenido los siguientes resultados de una muestra de $n = 17$ restaurantes que son aproximadamente del mismo tamaño, siendo

y = número de botellas vendidas por noche
x = coste medio, en dólares, de una cena

$$\bar{x} = 25{,}5 \qquad \bar{y} = 16{,}0$$

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = 350 \qquad \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = 180$$

a) Halle la recta de regresión muestral.
b) Interprete la pendiente de la recta de regresión muestral.
c) ¿Es posible dar una interpretación que tenga sentido de la ordenada en el origen de la recta de regresión muestral? Explique su respuesta.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

12.23. Vuelva a los datos del ejercicio 12.7 sobre la variación porcentual (X) del índice Dow-Jones en los cinco primeros días de sesión del año y la variación porcentual (Y) del índice en el conjunto del año.

a) Estime la regresión lineal de Y con respecto a X.
b) Interprete la ordenada en el origen y la pendiente de la recta de regresión muestral.

12.24. El viernes 13 de noviembre de 1989, cayeron vertiginosamente las cotizaciones en la bolsa de Nueva York; el índice Standard and Poor's 500 cayó un 6,1 por ciento ese día. El fichero de datos **New York Stock Exchange Gains and Losses** muestra las *pérdidas* porcentuales (y) que experimentaron los 25 mayores fondos de inversión el 13 de noviembre de 1989. También muestra las *ganancias* porcentuales (x), suponiendo que los dividendos y las ganancias de capital de estos mismos fondos se reinvirtieron en 1989 hasta el 12 de noviembre.

a) Estime la regresión lineal de las pérdidas registradas el 13 de noviembre con respecto a las ganancias obtenidas hasta el 13 de noviembre de 1989.
b) Interprete la pendiente de la recta de regresión muestral.

12.25. Ace Manufacturing está estudiando el absentismo laboral. Los datos del fichero **Employee Absence** se refieren a la variación anual de la tasa total de absentismo y la variación anual de la tasa media de absentismo por enfermedad.

a) Estime la regresión lineal de la variación de la tasa media de absentismo por enfermedad con respecto a la variación de la tasa de absentismo.
b) Interprete la pendiente estimada de la recta de regresión.

12.4. El poder explicativo de una ecuación de regresión lineal

El modelo de regresión estimado que hemos presentado puede concebirse como un intento de explicar los cambios de una variable dependiente Y provocados por los cambios de una variable independiente X. Si sólo tuviéramos observaciones de la variable dependiente, Y, la tendencia central de Y se representaría por medio de la media $\bar{y}$ y la variabilidad total en torno a Y se representaría por medio del numerador del estimador de la varianza muestral, $\Sigma(y_i - \bar{y})^2$. Cuando también tenemos medidas de X, hemos demostrado que la tendencia central de Y ahora puede expresarse en función de X. Esperamos que la ecuación lineal esté más cerca de los valores individuales de Y y que, por lo tanto, la variabilidad en torno a la ecuación lineal sea menor que la variabilidad en torno a la media.

Estamos ya en condiciones de desarrollar medidas que indiquen la eficacia con que la variable X explica la conducta de Y. En nuestro ejemplo de las ventas al por menor mostrado en la Figura 12.1, las ventas al por menor, Y, tienden a aumentar con la renta disponible, X y, por lo tanto, la renta disponible explica algunas de las diferencias entre las ventas al por menor. Sin embargo, los puntos no están todos en la línea, por lo que la explicación no es perfecta. Aquí desarrollamos medidas basadas en la descomposición de la variabilidad, que miden la capacidad de X para explicar Y en una regresión específica.

El análisis de la varianza, ANOVA, para una regresión de mínimos cuadrados se realiza descomponiendo la variabilidad total de Y en un componente explicado y un componente de error. En la Figura 12.5 mostramos que la desviación de un valor de Y con respecto a su media puede descomponerse en la desviación del valor predicho con respecto a la media y la desviación del valor observado con respecto al valor predicho

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

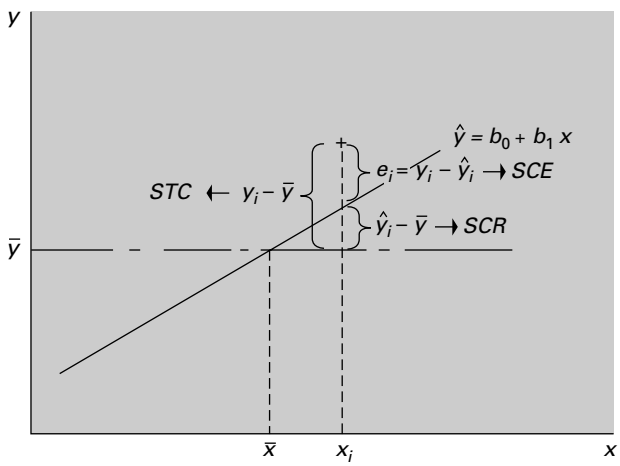


Figura 12.5. Descomposición de la variabilidad.

Elevamos al cuadrado los dos miembros de la ecuación —ya que la suma de las desviaciones en torno a la media es igual a 0— y sumamos el resultado obtenido en los n puntos

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Tal vez algunos lectores se hayan dado cuenta de que la elevación al cuadrado del primer miembro debe incluir el producto de los dos términos además de sus cantidades al cuadrado. Puede demostrarse que el término del producto de los dos términos es igual a 0. Esta ecuación puede expresarse de la forma siguiente:

$$STC = SCR + SCE$$

Aquí vemos que la variabilidad total —*STC*— puede dividirse en un componente —*SCR*— que representa la variabilidad que es explicada por la pendiente de la ecuación de regresión (la media de Y es diferente en distintos niveles de X). El segundo componente —*SCE*— se debe a la desviación aleatoria o sin explicar de los puntos con respecto a la recta de regresión. Esta variabilidad es una indicación de la incertidumbre relacionada con el modelo de regresión. El primer miembro es la *suma total de los cuadrados*:

$$STC = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

La cantidad de variabilidad explicada por la ecuación de regresión es la *suma de los cuadrados de la regresión* y se calcula de la forma siguiente:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Vemos que la variabilidad explicada por la regresión depende directamente de la magnitud del coeficiente b_1 y de la dispersión de los datos de la variable independiente, X. Las desviaciones en torno a la recta de regresión, e_i, que se utilizan para calcular la parte no explicada, o sea, la *suma de los cuadrados de los errores*, pueden definirse utilizando las siguientes formas algebraicas:

$$SCE = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Dado un conjunto de valores observados de las variables dependientes, Y, la *STC* es fija e igual a la variabilidad total de todas las observaciones con respecto a la media. Vemos que en esta descomposición, cuanto más altos son los valores de *SCR* y, por lo tanto, cuanto más bajos son los valores de *SCE*, mejor «se ajusta» o se aproxima la ecuación de regresión a los datos observados. Esta descomposición se muestra gráficamente en la Figura 12.5. En la ecuación de *SCR* vemos que la variabilidad explicada, *SCR*, está relacionada directamente con la dispersión de la variable independiente o X. Por lo tanto, cuando examinamos aplicaciones del análisis de regresión, sabemos que debemos tratar de obtener datos que tengan un gran rango para la variable independiente de manera que el modelo de regresión resultante tenga una variabilidad sin explicar menor.

Análisis de la varianza

La variabilidad total en un análisis de regresión, *STC*, puede descomponerse en un componente explicado por la regresión, *SCR*, y un componente que se debe a un error sin explicar, *SCE*:

$$STC = SCR + SCE \tag{12.8}$$

cuyos componentes se definen de la forma siguiente.

Suma total de los cuadrados:

$$STC = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{12.9}$$

Suma de los cuadrados de los errores:

$$SCE = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{12.10}$$

Suma de los cuadrados de la regresión:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{12.11}$$

Retail Sales

Volvamos con esta información a nuestro ejemplo de las ventas al por menor (ejemplo 12.2) con el fichero de datos **Retail Sales** y veamos cómo utilizamos la descomposición de la variabilidad para averiguar en qué medida explica nuestro modelo el proceso estudiado. La Tabla 12.2 muestra los cálculos detallados de los residuos, e_i; las desviaciones de Y con respecto a la media, y las desviaciones de los valores predichos de Y con respecto a la media. Éstos nos proporcionan los componentes para calcular *SCE*, *STC* y *SCR*. La suma de los cuadrados de las desviaciones de la columna 5 es $SCE = 436.127$. La suma de los cuadrados de las desviaciones de la columna 6 es $STC = 5.397.561$. Por último, la suma de los cuadrados de las desviaciones de la columna 7 es $SCR = 4.961.434$. La Figura 12.6 presenta las salidas Minitab y Excel del análisis de regresión, incluido el análisis de la varianza.

El coeficiente de determinación R^2

Hemos visto que el ajuste de la ecuación de regresión a los datos mejora cuando aumenta *SCR* y disminuye *SCE*. El cociente entre la suma de los cuadrados de la regresión, *SCR*, y la suma total de los cuadrados, *STC*, es una medida descriptiva de la proporción o porcentaje de la variabilidad total que es explicada por el modelo de regresión. Esta medida se llama *coeficiente de determinación* o, en términos más generales, R^2.

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{STC}$$

A menudo se considera que el coeficiente de determinación es el porcentaje de la variabilidad de Y que es explicado por la ecuación de regresión. Antes hemos demostrado que *SCR* aumenta directamente con la dispersión de la variable independiente X:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_i^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Tabla 12.2. Valores efectivos y predichos de las ventas al por menor por hogar y residuos calculados a partir de su regresión lineal con respecto a la renta por hogar.

Año	Renta (X)	Ventas al por menor (Y)	Ventas al por menor predichas	Residuo	Desviación observada con respecto a la media	Desviación predicha con respecto a la media
1	9.098	5.492	5.394	98	−550	−649
2	9.138	5.540	5.409	131	−502	−633
3	9.094	5.305	5.392	−87	−737	−650
4	9.282	5.507	5.464	43	−535	−578
5	9.229	5.418	5.444	−26	−624	−599
6	9.347	5.320	5.489	−169	−722	−554
7	9.525	5.538	5.557	−19	−504	−486
8	9.756	5.692	5.645	47	−350	−397
9	10.282	5.871	5.846	25	−171	−197
10	10.662	6.157	5.991	166	115	−52
11	11.019	6.342	6.127	215	300	84
12	11.307	5.907	6.237	−330	−135	194
13	11.432	6.124	6.284	−160	82	242
14	11.449	6.186	6.291	−105	144	248
15	11.697	6.224	6.385	−161	182	343
16	11.871	6.496	6.452	44	454	409
17	12.018	6.718	6.508	210	676	465
18	12.523	6.921	6.701	220	879	658
19	12.053	6.471	6.521	−50	429	479
20	12.088	6.394	6.535	−141	352	492
21	12.215	6.555	6.583	−28	513	541
22	12.494	6.755	6.689	66	713	647
	Suma de los cuadrados de los valores			436.127	5.397.561	4.961.434

Vemos, pues, que R^2 también aumenta directamente con la dispersión de la variable independiente. Cuando buscamos datos para estimar un modelo de regresión, es importante elegir las observaciones de la variable independiente que abarquen la mayor dispersión posible de X con el fin de obtener un modelo de regresión con el mayor R^2.

Coeficiente de determinación R^2

El coeficiente de determinación de una ecuación de regresión es

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{STC} \qquad \textbf{(12.12)}$$

Esta cantidad varía de 0 a 1 y los valores más altos indican que la regresión es mejor. Las interpretaciones generales de R^2 deben hacerse con cautela, ya que un valor alto puede deberse a que *SCE* es bajo o a que *STC* es alto o ambas cosas a la vez.

R^2 puede variar de 0 a 1, ya que *STC* es fijo y $0 < SCE < STC$. Cuando R^2 es alto, significa que la regresión es mejor, manteniéndose todo lo demás constante. En la salida del análisis de regresión vemos que el R^2 de la regresión de las ventas al por menor es 0,919, o sea, 91,9 por ciento. Normalmente, se considera que R^2 es la *variabilidad porcentual explicada*.

Results for: retail sales.MTW

Regression Analysis: Y Retail Sales versus X Income

```
The regression equation is
Y Retail Sales = 1922 + 0.382 X Income

Predictor       Coef    SE Coef       T       P
Constant      1922.4      274.9    6.99   0.000
X Income     0.38152    0.02529   15.08   0.000

S = 147.670   R-Sq = 91.9%     R-Sq(adj) = 91.5%

Analysis of Variance

Source           DF       SS        MS       F       P
Regression        1  4961434   4961434  227.52   0.000
Residual Error   20   436127     21806
Total            21  5397561

Unusual Observations
                    Y
                Retail
Obs X   Income   Sales      Fit  SE Fit  Residual  St Resid
12       11307  5907.0   6236.2    34.0    -329.2     -2.29R

R denotes an observation with a large standardized residual.
```

s_e, Error típico de la estimación

R^2, Coeficiente de determinación

s_e^2, Varianza del error del modelo

SRC = 4,961,434
SCE = 436,127
STC = 5,397,561

(a)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.958748803					
5	R Square	0.919199267					
6	Adjusted R Square	0.91515923					
7	Standard Error	147.6697181					
8	Observations	22					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	4961434.406	4961434	227.5225	2.17134E-12	
13	Residual	20	436126.9127	21806.35			
14	Total	21	5397561.318				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	1922.392694	274.9493737	6.991806	8.74E-07	1348.858617	2495.92677
18	X Income	0.38151672	0.025293061	15.08385	2.17E-12	0.328756343	0.4342771

s_e, Error típico de la estimación

R^2, Coeficiente de determinación

s_e^2, Varianza del error del modelo

SRC = 4,961,434
SCE = 436,127
STC = 5,397,561

(b)

Figura 12.6. Análisis de regresión de las ventas al por menor con respecto a la renta disponible: (a) salida Minitab; (b) salida Excel.

INTERPRETACIÓN

La segunda forma de la ecuación pone de manifiesto que R^2 depende del cociente entre *SCE* y *STC*. R^2 puede ser alto porque *SCE* es bajo —el objetivo deseado— o porque *STC* es alto o por ambas cosas a la vez. Las interpretaciones generales de R^2 que se aplican a todas las ecuaciones de regresión son peligrosas. Dos modelos de regresión que tengan el mismo conjunto de y_i observadas siempre pueden compararse utilizando el coeficiente de determinación R^2, y el modelo cuyo R^2 sea más alto explica mejor la variable *Y*. Pero las comparaciones generales de R^2 —que afirman que un modelo es bueno porque su R^2 es

superior a un determinado valor— son engañosas. Generalmente, los analistas con experiencia han observado que R^2 es 0,80 o más en los modelos basados en datos de series temporales. En los modelos basados en datos de corte transversal (por ejemplo, ciudades, regiones, empresas), el valor de R^2 oscila entre 0,40 y 0,60 y en los modelos basados en datos de personas individuales a menudo oscila entre 0,10 y 0,20.

Para ilustrar el problema de las interpretaciones generales de R^2, consideremos dos modelos de regresión —cuyos gráficos se muestran en la Figura 12.7—, cada uno de los cuales se basa en un total de 25 observaciones. En ambos modelos, *SCE* es igual a 17,89, por lo

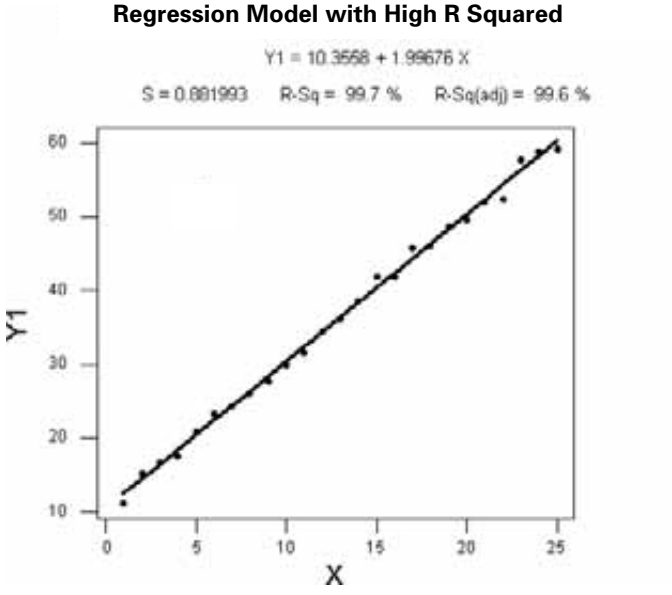


Figura 12.7. Comparación del R^2 de dos modelos de regresión; (a) R^2 alto; (b) R^2 bajo.

que el ajuste de la ecuación de regresión a los puntos de datos es el mismo. Pero en el primer modelo, la suma total de los cuadrados es igual a 5.201,05, mientras que en el segundo es igual a 68,22. Los valores de R^2 de los dos modelos son los siguientes.

Modelo 1:

$$R^2 = 1 - \frac{SCE}{STC} = 1 - \frac{17{,}89}{5.201{,}05} = 0{,}997$$

Modelo 2:

$$R^2 = 1 - \frac{SCE}{STC} = 1 - \frac{17{,}89}{68{,}22} = 0{,}738$$

Dado que *SCE* es igual en ambos modelos y, por lo tanto, la bondad del ajuste es la misma en los dos, no podemos afirmar que el modelo 1 se ajusta mejor a los datos. Sin embargo, en el modelo 1 el valor de R^2 es mucho más alto que en el modelo 2. Como vemos aquí, la interpretación general de R^2 debe hacerse con mucha cautela. Obsérvese que los dos intervalos diferentes del eje de ordenadas de la Figura 12.7 se deben a valores diferentes de *STC*.

También puede establecerse una relación entre el coeficiente de correlación y el R^2, observando que la correlación al cuadrado es igual al coeficiente de determinación. Otra interpretación de la correlación es que es la raíz cuadrada de la variabilidad porcentual explicada.

Correlación y R^2

El coeficiente de determinación, R^2, de la regresión simple es igual al cuadrado del coeficiente de correlación simple:

$$R^2 = r^2 \qquad \textbf{(12.13)}$$

Este resultado establece una importante conexión entre la correlación y el modelo de regresión.

La suma de los cuadrados de los errores puede utilizarse para obtener una estimación de la varianza del error del modelo ε_i. Como veremos, el estimador de la varianza del error del modelo se utiliza para realizar la inferencia estadística en el modelo de regresión. Recuérdese que hemos supuesto que el error poblacional, ε_i, es un error aleatorio que tiene una media 0 y una varianza σ^2. El estimador de σ^2 se calcula de la forma siguiente:

Estimación de la varianza del error del modelo

La cantidad *SCE* es una medida de la suma total de los cuadrados de las desviaciones en torno a la recta de regresión estimada y e_i es el residuo. Un estimador de la varianza del error poblacional del modelo es

$$\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SCE}{n-2} \qquad \textbf{(12.14)}$$

Se divide por $n - 2$ en lugar de $n - 1$ porque el modelo de regresión simple utiliza dos parámetros estimados, b_0 y b_1, en lugar de uno. En el siguiente apartado vemos que este estimador de la varianza es la base de la inferencia estadística en el modelo de regresión.

EJERCICIOS

Ejercicios básicos

12.26. Calcule *SCR*, *SCE*, s_e^2 y el coeficiente de determinación, dados los siguientes estadísticos calculados a partir de una muestra aleatoria de pares de observaciones de *X* e *Y*:

a) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 100.000$; $r^2 = 0{,}50$; $n = 52$

b) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 90.000$; $r^2 = 0{,}70$; $n = 52$

c) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 240$; $r^2 = 0{,}80$; $n = 52$

d) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 200.000$; $r^2 = 0{,}30$; $n = 74$

e) $\sum_{i=1}^{n} (y_i - \bar{y})^2 = 60.000$; $r^2 = 0{,}90$; $n = 40$

Ejercicios aplicados

12.27. Sea la recta de regresión muestral

$$y_i = b_0 + b_1 x_i + e_i = \hat{y}_i + e_i \quad (i = 1, 2, ..., n)$$

y sean $\bar{x}$ e $\bar{y}$ las medias muestrales de las variables independiente y dependiente, respectivamente.

a) Demuestre que

$$e_i = y_i - \bar{y} - b(x_i - \bar{x})$$

b) Utilizando el resultado del apartado (a), demuestre que

$$\sum_{i=1}^{n} e_i = 0$$

c) Utilizando el resultado del apartado (a), demuestre que

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

d) Demuestre que

$$\hat{y}_i - \bar{y} = b_i(x_i - \bar{x})$$

e) Utilizando los resultados de los apartados (c) y (d), demuestre que

$$STC = SCR + SCE$$

f) Utilizando el resultado del apartado (a), demuestre que

$$\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0$$

12.28. Sea

$$R^2 = \frac{SCR}{STC}$$

el coeficiente de determinación de la recta de regresión muestral.

a) Utilizando el apartado (d) del ejercicio 12.27, demuestre que

$$R^2 = b_1^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

b) Utilizando el resultado del apartado (a), demuestre que el coeficiente de determinación es igual al cuadrado de la correlación muestral entre *X* e *Y*.

c) Sea b_1 la pendiente de la regresión por mínimos cuadrados de *Y* con respecto a *X*, b_1^* la pendiente de la regresión por mínimos cuadrados de *X* con respecto a *Y* y *r* la correlación muestral entre *X* e *Y*. Demuestre que

$$b_1 \cdot b_1^* = r^2$$

12.29. Halle e interprete el coeficiente de determinación de la regresión de las ventas del sistema de DVD con respecto al precio, utilizando los datos siguientes.

Ventas	420	380	350	400	440	380	450	420
Precio	5,5	6,0	6,5	6,0	5,0	6,5	4,5	5,0

12.30. Halle e interprete el coeficiente de determinación de la regresión de la variación porcentual del índice Dow-Jones en un año con respecto a la variación porcentual del índice en los cinco primeros días de sesión del año, continuando con el análisis del ejercicio 12.7. Compare su respuesta con la correlación muestral obtenida con estos datos en el ejercicio 12.7. Utilice el fichero de datos **Dow Jones**.

12.31. Basándose en los datos del ejercicio 12.24, halle la proporción de la variabilidad muestral de las pérdidas porcentuales experimentadas por los fondos de inversión el 13 de noviembre de 1989 explicada por su dependencia lineal de las ganancias porcentuales obtenidas en 1989 hasta el 12 de noviembre. Utilice el fichero de datos **New York Stock Exchange Gains and Losses**.

12.32. Vuelva a los datos sobre la tasa de absentismo laboral del ejercicio 12.25. Utilice el fichero de datos **Employee Absence**.

a) Halle los valores predichos, $\hat{y}_i$, y los residuos, e_i, de la regresión por mínimos cuadrados de la variación de la tasa media de absentismo por enfermedad con respecto a la variación de la tasa de desempleo.

b) Halle las sumas de los cuadrados *STC*, *SCR* y *SCE* y verifique que

$$STC = SCR + SCE$$

c) Utilizando los resultados del apartado (a), halle e interprete el coeficiente de determinación.

12.33. Vuelva a los datos sobre las ventas semanales y las puntuaciones obtenidas en un test de aptitud por los representantes de ventas del ejercicio 12.21.

a) Halle los valores predichos, $\hat{y}_i$, y los residuos, e_i, de la regresión por mínimos cuadrados de las ventas semanales con respecto a las puntuaciones del test de aptitud.

b) Halle las sumas de los cuadrados *STC*, *SCR* y *SCE* y verifique que

$$STC = SCR + SCE$$

c) Utilizando los resultados del apartado (a), halle e interprete el coeficiente de determinación.

d) Halle directamente el coeficiente de correlación muestral entre las ventas y las puntuaciones del test de aptitud y verifique que su cuadrado es igual al coeficiente de determinación.

12.34. En un estudio se demostró que en una muestra de 353 profesores universitarios, la correlación entre las subidas salariales anuales y las evaluaciones de la docencia era de 0,11. ¿Cuál sería el coeficiente de determinación de una regresión de las subidas salariales anuales con respecto a las evaluaciones de la docencia en esta muestra? Interprete su resultado.

12.5. Inferencia estadística: contrastes de hipótesis e intervalos de confianza

Una vez desarrollados los estimadores de los coeficientes y un estimador de σ^2, estamos ya en condiciones de hacer inferencias relativas al modelo poblacional. El enfoque básico es paralelo al de los Capítulos 8 a 11. Desarrollamos estimadores de la varianza para los estimadores de los coeficientes, b_0 y b_1, y utilizamos los parámetros y las varianzas estimados para contrastar hipótesis y para calcular intervalos de confianza utilizando la distribución *t* de Student. Las inferencias realizadas a partir del análisis de regresión nos ayudarán a comprender el proceso analizado y a tomar decisiones sobre ese proceso. Suponemos inicialmente que los errores aleatorios del modelo, ε, siguen una distribución normal. Más adelante, sustituiremos este supuesto por el del teorema del límite central. Comenzamos desarrollando estimadores de la varianza y formas útiles de contraste. A continuación, los aplicamos utilizando nuestros datos sobre las ventas al por menor.

En el apartado 12.2 definimos la regresión simple correspondiente al modelo poblacional:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

en la que las x_i tienen valores predeterminados, pero no son variables aleatorias. En los Capítulos 5 y 6 sobre las funciones lineales de variables aleatorias vimos que si ε_i es una variable aleatoria que sigue una distribución normal de varianza σ^2, entonces y_i también sigue una distribución normal que tiene la misma varianza. El segundo miembro es una función lineal de *X*, salvo por la variable aleatoria ε_i. Si sumamos una función de *X* a una

variable aleatoria, no cambiamos la varianza. En el apartado 12.3 observamos que el estimador del coeficiente de la pendiente, b_1, es

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) y_i$$

$$= \sum a_i y_i$$

donde

$$a_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

En este estimador, vemos que b_1 es una función lineal de la variable aleatoria y_i cuya varianza es σ^2. Las y_i son variables aleatorias independientes. Por lo tanto, la varianza de b_1 es una transformación simple de la varianza de Y. Utilizando los resultados del Capítulo 6, la función lineal puede expresarse de la forma siguiente:

$$b_1 = \sum_{i=1}^{n} a_i y_i$$

$$a_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sigma_{b_1}^2 = \sum_{i=1}^{n} a_i^2 \sigma^2$$

$$\sigma_{b_1}^2 = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} \sigma^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Dado que y_i sigue una distribución normal y b_1 es una función lineal de variables normales independientes, esta función lineal implica que b_1 también sigue una distribución normal. De este análisis podemos deducir la varianza poblacional y la varianza muestral.

Distribución en el muestreo del estimador de los coeficientes por mínimos cuadrados

Si se cumplen los supuestos habituales de la estimación por mínimos cuadrados, entonces b_1 es un estimador insesgado de β_1 y tiene una varianza poblacional

$$\sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{(n-1)s_x^2} \tag{12.15}$$

y un estimador insesgado de la varianza muestral

$$s_{b_1}^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_e^2}{(n-1)s_x^2} \tag{12.16}$$

El estimador de la constante de la regresión, b_0, también es una función lineal de la variable aleatoria y_i y, por lo tanto, puede demostrarse que sigue una distribución normal, y su estimador de la varianza puede obtenerse de la forma siguiente:

$$s_{b_0}^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)s_e^2$$

Es importante observar que la varianza del coeficiente de la pendiente, b_1, depende de dos importantes cantidades:

1. La distancia de los puntos con respecto a la recta de regresión medida por s_e^2. Cuando los valores son más altos, la varianza de b_1 es mayor.
2. La desviación total de los valores de X con respecto a la media medida por $(n-1)s_x^2$. Cuanto mayor es la dispersión de los valores de X, menor es la varianza del coeficiente de la pendiente.

Estos dos resultados son muy importantes cuando hay que elegir los datos para realizar un modelo de regresión. Antes hemos señalado que cuanto mayor era la dispersión de la variable independiente, X, mayor era R^2, lo que indicaba que la relación era más estrecha. Ahora vemos que cuanto mayor es la dispersión de la variable independiente —medida por s_x^2—, menor es la varianza del coeficiente estimado de la pendiente, b_1. Por lo tanto, cuanto menores sean los estimadores de la varianza del coeficiente de la pendiente, mejor es el modelo de regresión. También debemos añadir que muchas conclusiones de investigaciones y muchas decisiones de política económica se basan en la variación de Y que se debe a una variación de X, estimada por b_1. Por lo tanto, nos gustaría que la varianza de esta importante variable de decisión, b_1, fuera lo más pequeña posible.

En el análisis de regresión aplicado, nos gustaría saber primero si existe una relación. En el modelo de regresión, vemos que si β_1 es 0, entonces no existe una relación lineal: Y no aumentaría o disminuiría continuamente cuando aumenta X. Para averiguar si existe una relación lineal, podemos contrastar la hipótesis

$$H_0 : \beta_1 = 0$$

frente a

$$H_1 : \beta_1 \neq 0$$

Dado que b_1 sigue una distribución normal, podemos contrastar esta hipótesis utilizando el estadístico t de Student

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

que se distribuye como una t de Student con $n - 2$ grados de libertad. El contraste de hipótesis también puede realizarse con valores de β_1 distintos de 0. Una regla práctica es extraer la conclusión de que existe una relación si el valor absoluto del estadístico t es superior a 2. Este resultado se obtiene exactamente en el caso de un contraste de dos colas con un nivel de significación $\alpha = 0{,}05$ y 60 grados de libertad y constituye una buena aproximación cuando $n > 30$.

Base para la inferencia sobre la pendiente de la regresión poblacional

Sea β_1 la pendiente de la ecuación poblacional y b_1 su estimación por mínimos cuadrados basada en n pares de observaciones muestrales. En ese caso, si se cumplen los supuestos habituales del modelo de regresión y puede suponerse también que los errores, ε_i, siguen una distribución normal, la variable aleatoria

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \qquad \textbf{(12.17)}$$

se distribuye como una t de Student con $(n - 2)$ grados de libertad. Además, el teorema del límite central nos permite concluir que este resultado es aproximadamente válido para una amplia variedad de distribuciones no normales y muestras de un tamaño suficientemente grande, n.

La mayoría de los programas que se emplean para estimar regresiones calculan normalmente la desviación típica de los coeficientes y el estadístico t de Student para $\beta_1 = 0$. La Figura 12.8 muestra las salidas Minitab y Excel correspondientes al ejemplo de las ventas al por menor.

En el caso del modelo de las ventas al por menor, el coeficiente de la pendiente es $b_1 = 0{,}382$ con una desviación típica $s_{b_1} = 0{,}02529$. Para saber si existe relación entre las ventas al por menor, Y, y la renta disponible, X, podemos contrastar la hipótesis

$$H_0 : \beta_1 = 0$$

frente a

$$H_1 : \beta_1 \neq 0$$

En la hipótesis nula, el cociente entre el estimador del coeficiente, b_1, y su desviación típica sigue una distribución t de Student. En el ejemplo de las ventas al por menor, observamos que el estadístico t de Student calculado es

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - 0}{s_{b_1}} = \frac{0{,}38152 - 0}{0{,}02529} = 15{,}08$$

El estadístico t de Student resultante, $t = 15{,}08$, mostrado en la salida del análisis de regresión, constituye una prueba contundente para rechazar la hipótesis nula y concluir que existe una estrecha relación entre las ventas al por menor y la renta disponible. También

Results for: retail sales.MTW

Regression Analysis: Y Retail Sales versus X Income

```
The regression equation is
Y Retail Sales = 1922 + 0.382 X Income

Predictor       Coef    SE Coef       T       P
Constant      1922.4      274.9    6.99   0.000
X Income     0.38152    0.02529   15.08   0.000

S = 147.670  R-Sq = 91.9%    R-Sq(adj) = 91.5%

Analysis of Variance

Source          DF        SS        MS       F       P
Regression       1   4961434   4961434  227.52   0.000
Residual Error  20    436127     21806
Total           21   5397561

Unusual Observations

                  Y
             Retail
Obs X   Income   Sales      Fit   SE Fit   Residual   St Resid
12       11307  5907.0   6236.2    34.0     −329.2     −2.29R

R denotes an observation with a large standardized residual.
```

t_{b_1}, Estadístico t de Student

s_{b_1}, Error típico del coeficiente de la pendiente

s_e, Error típico de la estimación

s_e^2, Varianza del error del modelo

SCR, Suma de los cuadros de la regresión

SCE, Suma de los cuadros de los errores

b_1, Coeficiente de la pendiente

(a)

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.958748803					
5	R Square	0.919199267					
6	Adjusted R Square	0.91515923					
7	Standard Error	147.6697181					
8	Observations	22					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	4961434.406	4961434	227.5225	2.17134E-12	
13	Residual	20	436126.9127	21806.35			
14	Total	21	5397561.318				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	1922.392694	274.9493737	6.991806	8.74E-07	1348.858617	2495.92677
18	X Income	0.38151672	0.025293061	15.08305	2.17E-12	0.328756343	0.4342771
19							

s_e, Error típico de la estimación

SCR, Suma de los cuadros de la regresión

SCE, Suma de los cuadros de los errores

s_e, Varianza del error del modelo

t_{b_1}, Estadístico t de Student

s_{b_1}, Error típico del coeficiente de la pendiente

b_1, Coeficiente de la pendiente

(b)

Figura 12.8. Modelos de ventas al por menor: estimadores de las varianzas de los coeficientes: (a) salida Minitab; (b) salida Excel.

señalamos que el p-valor de b_1 es 0,000, lo que es una prueba alternativa de que β_1 no es igual a 0. Recuérdese que en el Capítulo 10 vimos que el p-valor es el menor nivel de significación al que puede rechazarse la hipótesis nula.

También podrían realizarse contrastes de hipótesis relativos a la constante de la ecuación, b_0, utilizando la desviación típica desarrollada antes y mostrada en la salida Minitab. Sin embargo, como normalmente nos interesan las tasas de variación —medidas por b_1—, los contrastes relativos a la constante generalmente son menos importantes.

Si el tamaño de la muestra es lo suficientemente grande para que se aplique el teorema del límite central, podemos realizar esos contrastes de hipótesis aunque los errores, ε_i, no sigan una distribución normal. La cuestión clave es la distribución de b_1. Si b_1 sigue una distribución normal aproximada, es posible realizar el contraste de hipótesis.

Contrastes de la pendiente de la regresión poblacional

Si los errores de la regresión, ε_i, siguen una distribución normal y se cumplen los supuestos habituales del método de los mínimos cuadrados (o si la distribución de b_1 es aproximadamente normal), los siguientes contrastes tienen un nivel de significación α.

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \beta_1 = \beta_1^* \quad \text{o} \quad H_0: \beta_1 \leqslant \beta_1^*$$

frente a la hipótesis alternativa

$$H_1: \beta_1 > \beta_1^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{b_1 - \beta_1^*}{s_{b_1}} \geqslant t_{n-2,\alpha} \tag{12.18}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0: \beta_1 = \beta_1^* \quad \text{o} \quad H_0: \beta_1 \geqslant \beta_1^*$$

frente a la hipótesis alternativa

$$H_1: \beta_1 < \beta_1^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{b_1 - \beta_1^*}{s_b} \leqslant -t_{n-2,\alpha} \tag{12.19}$$

3. Para contrastar la hipótesis nula

$$H_0: \beta_1 = \beta_1^*$$

frente a la hipótesis alternativa bilateral

$$H_1: \beta_1 \neq \beta_1^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{b_1 - \beta_1^*}{s_{b_1}} \geqslant t_{n-2,\alpha/2} \quad \text{o} \quad \frac{b_1 - \beta_1^*}{s_{b_1}} \leqslant -t_{n-2,\alpha/2} \tag{12.20}$$

Podemos obtener intervalos de confianza para la pendiente β_1 de la ecuación poblacional utilizando los estimadores de los coeficientes y de las varianzas que hemos desarrollado y el razonamiento realizado en el Capítulo 8.

Intervalos de confianza de la pendiente de la regresión poblacional b_1

Si los errores de la regresión, ε_i, siguen una distribución normal y se cumplen los supuestos habituales del análisis de regresión, se obtiene un intervalo de confianza al $100(1-\alpha)\%$ de la pendiente de la recta de regresión poblacional β_1 de la forma siguiente:

$$b_1 - t_{n-2,\alpha/2}s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2}s_{b_1} \qquad \textbf{(12.21)}$$

donde $t_{n-2,\alpha/2}$ es el número para el que

$$P(t_{n-2} > t_{n-2,\alpha/2}) = \alpha/2$$

y la variable aleatoria t_{n-2} sigue una distribución t de Student con $(n-2)$ grados de libertad.

En la salida del análisis de regresión de las ventas al por menor con respecto a la renta disponible de la Figura 12.8, vemos que

$$n = 22 \qquad b_1 = 0{,}3815 \qquad s_b = 0{,}0253$$

Para obtener el intervalo de confianza al 99 por ciento de β_1, tenemos $1-\alpha = 0{,}99$ y $n-2 = 20$ grados de libertad y, por lo tanto, vemos en la tabla 8 del apéndice que

$$t_{n-2,\alpha/2} = t_{20,0{,}005} = 2{,}845$$

Por lo tanto, tenemos el intervalo de confianza al 99 por ciento

$$0{,}3815 - (2{,}845)(0{,}0253) < \beta_1 < 0{,}3815 + (2{,}845)(0{,}0253)$$

o sea

$$0{,}3095 < \beta_1 < 0{,}4535$$

Vemos que el intervalo de confianza al 99 por ciento del aumento esperado de las ventas al por menor por hogar que acompaña a un aumento de la renta disponible por hogar de 1 $ abarca el intervalo de 0,3095 $ a 0,4353 $. La Figura 12.9 muestra los intervalos de confianza al 90, al 95 y al 99 por ciento de la pendiente de la regresión poblacional.

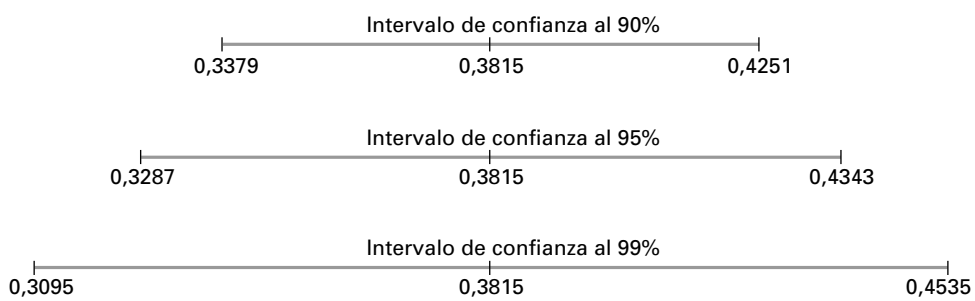


Figura 12.9. Intervalos de confianza de la pendiente de la recta de regresión poblacional de las ventas al por menor a los niveles de confianza del 90, el 95 y el 99 por ciento.

Contraste de hipótesis del coeficiente de la pendiente poblacional utilizando la distribución *F*

Existe otro contraste de la hipótesis de que el coeficiente de la pendiente, β_1, es igual a 0:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Este contraste se basa en la descomposición de la variabilidad que hemos presentado en el apartado 12.4. Este contraste parte del supuesto de que, si la hipótesis nula es verdadera, entonces pueden utilizarse tanto *SCE* como *SCR* para obtener estimadores independientes de la varianza del error del modelo σ^2. Para realizar este contraste, obtenemos dos estimaciones muestrales de la desviación típica poblacional σ, que se denominan términos cuadráticos medios. La suma de los cuadrados de la regresión, *SCR*, tiene un grado de libertad, ya que se refiere al coeficiente de la pendiente, y el cuadrado medio de la regresión, *CMR*, es

$$CMR = \frac{SCR}{1} = SCR$$

Si la hipótesis nula —ausencia de relación— es verdadera, entonces *CMR* es una estimación de la varianza global del modelo, σ^2. También utilizamos la suma de los cuadrados de los errores al igual que antes para hallar el error cuadrático medio, *ECM*:

$$ECM = \frac{SCE}{n-2} = s_e^2$$

En el apartado 11.4 introdujimos la distribución *F*, que era el cociente entre estimaciones muestrales independientes de la varianza, dadas varianzas poblacionales iguales. Puede demostrarse que *CMR* y *ECM* son independientes y que en H_0 ambas son estimaciones de la varianza poblacional, σ^2. Por lo tanto, si H_0 es verdadera, podemos demostrar que el cociente

$$F = \frac{CMR}{ECM} = \frac{SCR}{s_e^2}$$

sigue una distribución *F* con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador. También debe señalarse que el estadístico *F* es igual al cuadrado del estadístico *t* del coeficiente de la pendiente. Esta afirmación puede demostrarse algebraicamente. Aplicando la teoría de la distribución, podemos demostrar que una *t* de Student al cuadrado con $n - 2$ grados de libertad y la *F* con 1 grado de libertad en el numerador y $n - 2$ grados de libertad en el denominador son iguales:

$$F_{\alpha, 1, n-2} = t^2_{\alpha/2, n-2}$$

La Figura 12.8(a) muestra el análisis de varianza de la regresión de las ventas al por menor procedente de la salida Minitab. En nuestro ejemplo de las ventas al por menor, la

suma de los cuadrados de los errores se divide por los 20 grados de libertad para calcular el *ECM*:

$$ECM = \frac{436.127}{20} = 21.806$$

A continuación, se calcula el cociente F, que es como el cociente entre dos cuadrados medios:

$$F = \frac{CMR}{ECM} = \frac{4.961.434}{21.806} = 227{,}52$$

Este cociente F es considerablemente mayor que el valor crítico de $\alpha = 0{,}01$ con 1 grado de libertad en el numerador y 20 grados de libertad en el denominador ($F_{1,\,20,\,0{,}01} = 8{,}10$) según la Tabla 9 del apéndice. La salida Minitab —Figura 12.8(a)— de la regresión de las ventas al por menor muestra que el p-valor de esta F calculada es 0,000, lo que constituye una prueba alternativa para rechazar H_0. Obsérvese también que el estadístico F es igual a t^2, siendo t el estadístico del coeficiente de la pendiente, b_1:

$$F = t^2$$
$$227{,}52 = 15{,}08^2$$

Contraste *F* del coeficiente de regresión simple

Podemos contrastar la hipótesis

$$H_0 : \beta_1 = 0$$

frente a la alternativa

$$H_1 : \beta_1 \neq 0$$

utilizando el estadístico F

$$F = \frac{CMR}{ECM} = \frac{SCR}{s_e^2} \qquad \textbf{(12.22)}$$

La regla de decisión es

$$\text{Rechazar } H_0 \text{ si} \quad F \geqslant F_{1,\,n-2,\,\alpha} \qquad \textbf{(12.23)}$$

También podemos mostrar que el estadístico F es

$$F = t_{b_1}^2 \qquad \textbf{(12.24)}$$

en cualquier análisis de regresión simple.

Este resultado muestra que los contrastes de hipótesis relativos al coeficiente de la pendiente poblacional dan exactamente el mismo resultado cuando se utiliza la t de Student que cuando se utiliza la distribución F. En el Capítulo 13 veremos que la distribución F —cuando se utiliza en un análisis de regresión múltiple— también brinda la oportunidad de contrastar la hipótesis de que varios coeficientes poblacionales de la pendiente son simultáneamente iguales a 0.

EJERCICIOS

Ejercicios básicos

12.35. Dado el modelo de regresión simple

$$Y = \beta_0 + \beta_1 X$$

y los resultados de la regresión siguientes, contraste la hipótesis nula de que el coeficiente de la pendiente es 0 frente a la hipótesis alternativa de que es mayor que cero utilizando la probabilidad de cometer un error de Tipo I igual a 0,05 y halle los intervalos de confianza bilaterales al 95 y al 99 por ciento.

a) Una muestra aleatoria de tamaño $n = 38$ con $b_1 = 5$ y $s_{b_1} = 2{,}1$
b) Una muestra aleatoria de tamaño $n = 46$ con $b_1 = 5{,}2$ y $s_{b_1} = 2{,}1$
c) Una muestra aleatoria de tamaño $n = 38$ con $b_1 = 2{,}7$ y $s_{b_1} = 1{,}87$
d) Una muestra aleatoria de tamaño $n = 29$ con $b_1 = 6{,}7$ y $s_{b_1} = 1{,}8$

12.36. Utilice un modelo de regresión simple para contrastar la hipótesis

$$H_0 : \beta_1 = 0$$

frente a

$$H_1 : \beta_1 \neq 0$$

suponiendo que $\alpha = 0{,}05$, dados los siguientes estadísticos de la regresión:

a) El tamaño de la muestra es 35, $STC = 100.000$ y la correlación entre X e Y es 0,46.
b) El tamaño de la muestra es 61, $STC = 123.000$ y la correlación entre X e Y es 0,65.
c) El tamaño de la muestra es 25, $STC = 128.000$ y la correlación entre X e Y es 0,69.

Ejercicios aplicados

12.37. Considere la regresión lineal de las ventas del sistema DVD con respecto al precio del ejercicio 12.29.

a) Utilice un método de estimación insesgado para hallar una estimación de la varianza de los términos de error en la regresión poblacional.
b) Utilice un método de estimación insesgado para hallar una estimación de la varianza del estimador por mínimos cuadrados de la pendiente de la recta de regresión poblacional.
c) Halle el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.

12.38. Una cadena de comida rápida decidió realizar un experimento para averiguar la influencia de los gastos publicitarios en las ventas. Se introdujeron diferentes cambios relativos en los gastos publicitarios en comparación con el año anterior en ocho regiones del país y se observaron los cambios que experimentaron las ventas como consecuencia. La tabla adjunta muestra los resultados.

Aumento de los gastos publicitarios (%)	0	4	14	10	9	8	6	1
Aumento de las ventas (%)	2,4	7,2	10,3	9,1	10,2	4,1	7,6	3,5

a) Estime por mínimos cuadrados la regresión lineal del aumento de las ventas con respecto al aumento de los gastos publicitarios.
b) Halle el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.

12.39. Un vendedor de bebidas alcohólicas al por mayor tiene interés en averiguar cómo afecta el precio de un whisky escocés a la cantidad vendida. En una muestra aleatoria de datos sobre las ventas de ocho semanas se obtuvieron los resultados de la tabla adjunta sobre el precio, en dólares, y las ventas, en cajas.

Precio	19,2	20,5	19,7	21,3	20,8	19,9	17,8	17,2
Ventas	25,4	14,7	18,6	12,4	11,1	15,7	29,2	35,2

Halle el intervalo de confianza al 95 por ciento de la variación esperada de las ventas provocada por una subida del precio de 1 $.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

12.40. Continúe el análisis del ejercicio 12.30 de la regresión de la variación porcentual del índice Dow-Jones en un año con respecto a la variación porcentual del índice en los cinco primeros días de sesión del año. Utilice el fichero de datos **Dow Jones**.

a) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de los términos de error de la regresión poblacional.

b) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza del estimador por mínimos cuadrados de la pendiente de la recta de regresión poblacional.
c) Halle e interprete el intervalo de confianza al 95 por ciento de la pendiente de la recta de regresión poblacional.
d) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa bilateral.

12.41. Considere el modelo de las pérdidas experimentadas por los fondos de inversión el 13 de noviembre de 1980 del ejercicio 12.24. Utilice el fichero de datos **New York Stock Exchange Gains and Losses**.

a) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza de los términos de error de la regresión poblacional.
b) Utilice un método de estimación insesgado para hallar una estimación puntual de la varianza del estimador por mínimos cuadrados de la pendiente de la recta de regresión poblacional.
c) Halle los intervalos de confianza al 90, al 95 y al 99 por ciento de la pendiente de la recta de regresión poblacional.

12.6. Predicción

Los modelos de regresión pueden utilizarse para hacer predicciones o previsiones sobre la variable dependiente, partiendo de un valor futuro supuesto de la variable independiente. Supongamos que queremos predecir el valor de la variable dependiente, dado que la variable independiente es igual a un valor específico, x_{n+1}, y que la relación lineal entre la variable dependiente y la variable independiente continúa manteniéndose. El valor correspondiente de la variable dependiente será, entonces,

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$$

que, dado x_{n+1}, tiene la esperanza

$$E[y_{n+1} \mid x_{n+1}] = \beta_0 + \beta_1 x_{n+1}$$

Existen dos opciones interesantes:

1. Podemos querer estimar el valor efectivo que se obtendrá con una única observación, y_{n+1}. Esta opción se muestra en la Figura 12.10.
2. Podemos querer estimar el valor esperado condicionado, $E[y_{n+1} \mid x_{n+1}]$, es decir, el valor medio de la variable dependiente cuando la variable independiente es fija e igual a x_{n+1}. Esta opción se muestra en la Figura 12.11.

Dado que los supuestos habituales del análisis de regresión continúan cumpliéndose, se obtiene la misma estimación puntual en las dos opciones. Sustituimos simplemente los β_0 y β_1 desconocidos por sus estimaciones por mínimos cuadrados, b_0 y b_1. Es decir, estimamos $(\beta_0 + \beta_1 x_{n+1})$ por medio de $(b_0 + b_1 x_{n+1})$. Sabemos que el estimador correspondiente es el mejor estimador insesgado lineal de Y, dado X. En la primera opción, nos interesa saber cuál es la mejor predicción de una observación del proceso. Pero en la segunda opción, nos interesa saber cuál es el valor esperado o media a largo plazo del proceso. En ambas opciones, un buen estimador puntual con nuestros supuestos es

$$\hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

ya que no sabemos nada útil sobre la variable aleatoria, ε_{n+1}, salvo que su media es 0. Por lo tanto, sin otra información utilizaremos 0 como estimación puntual.

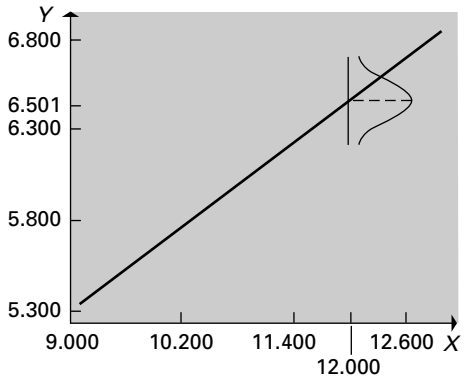


Figura 12.10. Recta de regresión estimada por mínimos cuadrados de las ventas al por menor con respecto a la renta disponible: aplicación a un único valor observado.

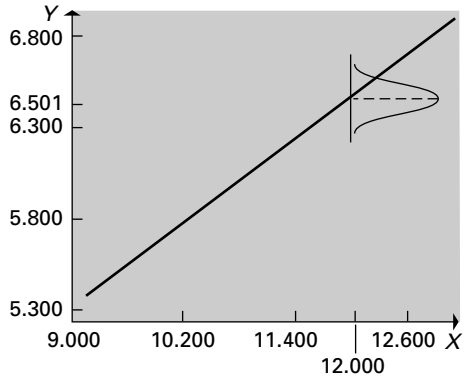


Figura 12.11. Recta de regresión estimada por mínimos cuadrados de las ventas al por menor con respecto a la renta disponible: valor esperado.

Sin embargo, normalmente queremos intervalos, además de estimaciones puntuales, y para eso las dos opciones son diferentes, ya que los estimadores de la varianza de dos cantidades diferentes estimadas son diferentes. Los resultados de estos estimadores diferentes de la varianza llevan a los dos intervalos diferentes. En la primera opción, el intervalo generalmente es un intervalo de predicción porque estamos prediciendo el valor de un único punto. El intervalo de la segunda opción es un intervalo de confianza porque es el intervalo del valor esperado.

Intervalos de confianza de las predicciones e intervalos de predicción

Supongamos que el modelo de regresión poblacional es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, ..., n + 1)$$

que se cumplen los supuestos habituales del análisis de regresión y que los ε_i siguen una distribución normal. Sean b_0 y b_1 las estimaciones por mínimos cuadrados de β_0 y β_1, basadas en (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). En ese caso, puede demostrarse que los intervalos al $100(1 - \alpha)\%$ son los siguientes:

1. Para la predicción del valor efectivo resultante de Y_{n+1}, el intervalo de predicción es

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]} s_e \qquad \textbf{(12.25)}$$

2. Para la predicción de la esperanza condicional $E(Y_{n+1}|x_{n+1})$, el intervalo de confianza es

$$\hat{y}_{n+1} \pm t_{n-2,\alpha/2} \sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]} s_e \qquad \textbf{(12.26)}$$

donde

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{y} \quad \hat{y}_{n+1} = b_0 + b_1 x_{n+1}$$

Retail Sales

EJEMPLO 12.3. Predicción de las ventas al por menor (predicción basada en un modelo de regresión)

Mostramos cómo se calculan los intervalos utilizando el ejemplo 12.2 sobre las ventas al por menor y la renta disponible. Le han pedido que haga una predicción de los valores de las ventas al por menor por hogar cuando la renta disponible por hogar es de 12.000 $: el valor efectivo del año que viene y el valor esperado a largo plazo. También le han pedido que calcule intervalos de predicción e intervalos de confianza para estas predicciones. Utilice el fichero de datos **Retail Sales**.

Solución

Los valores predichos para el próximo año y para el largo plazo son

$$\begin{aligned}\hat{y}_{n+1} &= b_0 + b_1 x_{n+1} \\ &= 1.922 + (0{,}3815)(12.000) = 6.501\end{aligned}$$

Por lo tanto, observamos que las ventas estimadas son de 6.501 $ cuando la renta disponible es de 12.000 $. También observamos que

$$n = 22 \qquad \bar{x} = 10.799 \qquad \sum (x_i - \bar{x})^2 = 34.110.178 \qquad s_e^2 = 21.806$$

Por lo tanto, el error típico de una única observación predicha de Y es

$$\sqrt{\left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]} s_e = \sqrt{\left[1 + \frac{1}{22} + \frac{(12.000 - 10.799)^2}{34.110.178}\right]} \sqrt{21.806} = 154{,}01$$

Asimismo, observamos que el error típico del valor esperado de Y es

$$\sqrt{\left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]} s_e = \sqrt{\left[\frac{1}{22} + \frac{(12.000 - 10.799)^2}{34.110.178}\right]} \sqrt{21.806} = 43{,}76$$

Supongamos que se necesitan intervalos del 95 por ciento para las predicciones suponiendo que $\alpha = 0{,}05$ y

$$t_{n-2,\,\alpha/2} = t_{20,\,0{,}025} = 2{,}086$$

Utilizando estos resultados, observamos que el intervalo de predicción al 95 por ciento para las ventas al por menor del próximo año cuando la renta disponible es de 12.000 $ se calcula de la forma siguiente:

$$\begin{aligned}&6.501 \pm (2{,}086)(154{,}01) \\ &6.501 \pm 321\end{aligned}$$

Por lo tanto, el intervalo de predicción al 95 por ciento para las ventas de un único año en el que la renta es de 12.000 $ va de 6.180 $ a 6.822 $.

En el caso del intervalo de confianza del valor esperado de las ventas al por menor cuando la renta disponible es de 12.000 $, tenemos que

$$\begin{aligned}&6.501 \pm (2{,}086)(43{,}76) \\ &6.501 \pm 91\end{aligned}$$

Por lo tanto, el intervalo de confianza al 95 por ciento del valor esperado va de 6.410 $ a 6.592 $.

INTERPRETACIÓN

Las Figuras 12.10 y 12.11 muestran la distinción entre estos dos problemas de estimación de intervalos. Vemos en ambas figuras la recta de regresión estimada para nuestros datos sobre las ventas al por menor y la renta disponible. También vemos en la Figura 12.10 una función de densidad que representa nuestra incertidumbre sobre el valor que tomarán las ventas al por menor en cualquier año específico en el que la renta disponible sea de 12.000 $. La función de densidad de la Figura 12.11 representa nuestra incertidumbre sobre las ventas al por menor esperadas o medias en los años en los que la renta disponible es de 12.000 $. Naturalmente, tenemos más incertidumbre sobre las ventas de un único año que sobre las ventas medias y eso se refleja en la forma de las dos funciones de densidad. Vemos que ambas están centradas en las ventas al por menor de 6.501 $, pero que la función de densidad de la Figura 12.10 tiene una dispersión mayor. Como consecuencia, el intervalo de predicción de un valor específico es mayor que el intervalo de confianza de las ventas al por menor esperadas.

Podemos extraer algunas conclusiones más estudiando las formas generales de los intervalos de predicción y de confianza. Como hemos visto, cuanto más amplio es el intervalo, mayor es la incertidumbre sobre la predicción puntual. Basándonos en estas fórmulas, hacemos cuatro observaciones:

1. Manteniéndose todo lo demás constante, cuanto mayor es el tamaño de la muestra n, más estrecho es el intervalo de confianza. Vemos, pues, que cuanta más información muestral tengamos, más seguros estaremos de nuestra inferencia.
2. Manteniéndose todo lo demás constante, cuanto mayor es s_e^2, más amplio es el intervalo de confianza. Una vez más, es de esperar, ya que s_e^2 es una estimación de σ^2, la varianza de los errores de la regresión, ε_i. Dado que estos errores

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

   representan la discrepancia entre los valores observados de las variables dependientes y sus esperanzas, dadas las variables independientes, cuanto mayor es la magnitud de esta discrepancia, más imprecisa será nuestra inferencia.
3. Consideremos ahora la cantidad $(\sum_{i=1}^{n} (x_i - \bar{x})^2$. Esta cantidad es simplemente un múltiplo de la varianza muestral de las observaciones de la variable independiente. Cuando la varianza es grande, significa que tenemos información sobre un amplio rango de valores de esta variable, lo que nos permite hacer estimaciones más precisas de la recta de regresión poblacional y, por lo tanto, calcular intervalos de confianza más reducidos.
4. También vemos que cuanto mayores son los valores de la cantidad $(x_{n+1} - \bar{x})^2$, más amplios son los intervalos de confianza de las predicciones. Por lo tanto, los intervalos de confianza son más amplios a medida que nos alejamos de la media de la variable independiente, X. Dado que nuestros datos muestrales están centrados en la media $\bar{x}$, es de esperar que podamos hacer inferencias más definitivas cuando la variable independiente está relativamente cerca de este valor central que cuando está a alguna distancia de él.

INTERPRETACIÓN

No se recomienda extrapolar la ecuación de regresión fuera del rango de los datos utilizados para realizar la estimación. Supongamos que se nos pide que hagamos una predicción de las ventas al por menor por hogar en un año en el que la renta disponible es de 30.000 $. Volviendo a los datos de la Tabla 12.1 y a la recta de regresión de la Figura 12.11, vemos que 30.000 $ se encuentra muy fuera del rango de los datos utilizados para

desarrollar el modelo de regresión. Un analista sin experiencia podría utilizar los métodos antes presentados para hacer una predicción o estimar un intervalo de confianza. En las ecuaciones podemos ver que los intervalos resultantes serían muy amplios y, por lo tanto, la predicción tendría escaso valor. Sin embargo, las predicciones que se realizan fuera del rango de los datos originales plantean un problema más fundamental: no tenemos sencillamente ninguna prueba que indique cómo es la naturaleza de la relación fuera del rango de los datos. No hay ninguna razón en la teoría económica que exija absolutamente que la relación siga siendo lineal con la misma tasa de variación cuando nos salimos del rango de los datos utilizados para estimar los coeficientes del modelo de regresión. Cualquier extrapolación del modelo fuera del rango de los datos para predecir valores debe basarse en otra información o evidencia, además de la que contiene el análisis de regresión basado en los datos de que se dispone. Cuando los analistas intentan hacer este tipo de extrapolación, pueden cometer graves errores.

EJERCICIOS

Ejercicios básicos

12.45. Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 12 + 5x_i$$

y

$$s_e = 9{,}67 \quad \bar{x} = 8 \quad n = 32 \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 500$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que $x = 13$.

12.43. Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 14 + 7x_i$$

y

$$s_e = 7{,}45 \quad \bar{x} = 8 \quad n = 25 \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 300$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que $x = 12$.

12.44. Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 22 + 8x_i$$

y

$$s_e = 3{,}45 \quad \bar{x} = 11 \quad n = 22 \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 400$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que $x = 17$.

12.45. Dado un análisis de regresión simple, suponga que hemos ajustado el siguiente modelo de regresión:

$$\hat{y}_i = 8 + 10x_i$$

y

$$s_e = 11{,}23 \quad \bar{x} = 8 \quad n = 44 \quad \sum_{i=1}^{n} (x_i - \bar{x})^2 = 800$$

Halle el intervalo de confianza al 95 por ciento y el intervalo de predicción al 95 por ciento para el punto en el que $x = 17$.

Ejercicios aplicados

12.46. Se toma una muestra de 25 obreros de una fábrica. Se pide a cada obrero que valore su satisfacción en el trabajo (x) en una escala de 1 a 10. Se averigua también el número de días que estos obreros estuvieron ausentes del trabajo (y) el año pasado. Se estima la recta de regresión muestral por mínimos cuadrados para estos datos.

$$\hat{y} = 12{,}6 - 1{,}2x$$

También se ha observado que

$$\bar{x} = 6{,}0 \quad \sum_{i=1}^{25} (x_i - \bar{x})^2 = 130{,}0 \quad SCE = 80{,}6$$

a) Contraste al nivel de significación del 1 por ciento la hipótesis nula de que la satisfacción en el trabajo no produce un efecto lineal en el absentismo frente a una hipótesis alternativa bilateral adecuada.

b) Un obrero tiene un nivel de satisfacción en el trabajo de 4. Halle un intervalo al 90 por

ciento del número de días que este obrero estaría ausente del trabajo en un año.

12.47. Los médicos tienen interés en saber qué relación existe entre la dosis de un medicamento y el tiempo que necesita un paciente para recuperarse. La tabla adjunta muestra las dosis (en gramos) y el tiempo de recuperación (en horas) de una muestra de cinco pacientes. Estos pacientes tienen parecidas características, salvo la dosis del medicamento administrada.

Dosis	1,2	1,0	1,5	1,2	1,4
Tiempo de recuperación	25	40	10	27	16

a) Estime la regresión lineal del tiempo de recuperación con respecto a la dosis.
b) Halle e interprete el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.
c) ¿Sería útil la regresión muestral obtenida en el apartado (a) para predecir el tiempo de recuperación de un paciente al que se le administran 2,5 gramos de este medicamento? Explique su respuesta.

12.48. En el caso del problema de la tasa de rendimiento de las acciones del ejercicio 12.20, se observó que

$$\sum_{i=1}^{20} y_i^2 = 196{,}2$$

a) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa de que es positiva.
b) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 1 frente a la hipótesis alternativa bilateral.

12.49. Utilizando los datos del ejercicio 12.21, contraste la hipótesis nula de que las ventas semanales de los representantes no están relacionadas linealmente con su puntuación en el test de aptitud frente a la hipótesis alternativa de que existe una relación positiva.

12.50. Vuelva a los datos del ejercicio 12.41. Contraste la hipótesis nula de que las pérdidas que experimentaron los fondos de inversión el viernes 13 de noviembre de 1989 no dependían linealmente de las ganancias obtenidas anteriormente en 1989 frente a la hipótesis alternativa bilateral.

12.51. Sea r la correlación muestral entre un par de variables aleatorias.

a) Demuestre que

$$\frac{1 - r^2}{n - 2} = \frac{s_e^2}{STC}$$

b) Utilizando el resultado del apartado (a), demuestre que

$$\frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{b}{s_e / \sqrt{\sum (x_i - \bar{x})^2}}$$

c) Utilizando el resultado del apartado (b), deduzca que el contraste de la hipótesis nula de la correlación poblacional 0, presentado en el apartado 12.1, es igual que el contraste de la pendiente de la recta de regresión poblacional 0, presentado en el apartado 12.5.

12.52. En el problema del ejercicio 12.22 sobre las ventas de cerveza en los restaurantes se observó que

$$\frac{\sum (y_i - \bar{y})^2}{n - 1} = 250$$

Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa bilateral.

12.53. En una muestra de 74 observaciones mensuales, se estimó la regresión del rendimiento porcentual del oro (y) con respecto a la variación porcentual del índice de precios (x). La recta de regresión muestral, obtenida por mínimos cuadrados, era

$$y = -0{,}003 + 1{,}11x$$

La desviación típica estimada de la pendiente de la recta de regresión poblacional era 2,31. Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa de que la pendiente es positiva.

12.54. Vuelva a los datos del ejercicio 12.39. Contraste al nivel del 5 por ciento la hipótesis nula de que las ventas no dependen linealmente del precio de este whisky escocés frente a la hipótesis alternativa bilateral apropiada.

12.55. Vuelva a los datos del ejercicio 12.29.

a) Halle una estimación puntual del volumen de ventas cuando el precio del sistema DVD es de 480 $ en una región dada.

b) Si el precio del sistema se fija en 480 $, halle intervalos de confianza al 95 por ciento del volumen efectivo de ventas en una región y el número esperado de ventas en esa región.

12.56. Continúe con el análisis del ejercicio 12.7. Si el índice Dow-Jones sube un 1,0 por ciento en los cinco primeros días de sesión de un año, halle intervalos de confianza al 90 por ciento de la variación porcentual *efectiva* y la *esperada* del índice en todo el año. Analice la distinción entre estos intervalos.

12.57. Vuelva a los datos del ejercicio 12.25 (fichero de datos **Employee Absence**). Halle para un año en el que no varía la tasa de desempleo intervalos de confianza al 90 por ciento de la variación *efectiva* de la tasa media de absentismo laboral por enfermedad y de la variación *esperada*.

12.58. Utilice los datos del ejercicio 12.20 para hallar intervalos de confianza al 90 y al 95 por ciento del rendimiento esperado de las acciones de la empresa cuando la tasa de rendimiento del índice Standard and Poor's 500 es del 1 por ciento.

12.59. Un nuevo representante de ventas de la empresa del ejercicio 12.21 obtiene 70 puntos en el test de aptitud. Halle intervalos de confianza al 80 y al 90 por ciento del valor de las ventas semanales que conseguirá.

12.7. Análisis gráfico

Hemos desarrollado los métodos teóricos y analíticos que permiten realizar análisis de regresión y construir modelos lineales. Utilizando contrastes de hipótesis e intervalos de confianza, podemos averiguar la calidad de nuestro modelo e identificar algunas relaciones importantes. Estos métodos inferenciales suponen inicialmente que los errores del modelo siguen una distribución normal. Pero también sabemos que el teorema del límite central nos ayuda a realizar contrastes de hipótesis y a construir intervalos de confianza mientras las distribuciones muestrales de los estimadores de los coeficientes y los valores predichos sean aproximadamente normales. El modelo de regresión también se basa en un conjunto de supuestos. Sin embargo, las aplicaciones del análisis de regresión pueden ser erróneas por muchas razones, incluidos los supuestos que no se satisfacen si los datos no siguen las pautas supuestas.

El ejemplo de la regresión de las ventas al por menor con respecto a la renta disponible —Figura 12.1— tiene un diagrama de puntos dispersos que sigue la pauta supuesta en el análisis de regresión. Sin embargo, esa pauta no siempre se produce cuando se estudian nuevos datos. Una de las mejores formas de detectar posibles problemas en el análisis de regresión simple es realizar diagramas de puntos dispersos y observar la pauta. Aquí examinamos algunos instrumentos analíticos y ejemplos de análisis de regresión que pueden ayudarnos a preparar mejores aplicaciones del análisis de regresión.

En este apartado utilizamos el análisis gráfico para mostrar cómo afectan al análisis de regresión los puntos que tienen valores extremos de X y los puntos que tienen valores de Y que se desvían considerablemente de la ecuación de regresión por mínimos cuadrados. En capítulos posteriores mostramos cómo puede utilizarse el análisis de los residuos para examinar otras desviaciones con respecto a las pautas normales de los datos.

Los puntos extremos son puntos en los que los valores de X se desvían considerablemente de los valores de X de los demás puntos. Volvamos a la ecuación 12.26, que presenta el intervalo de confianza del valor esperado de Y correspondiente a un valor específico

de *X*. Para este intervalo de confianza es fundamental un término llamado normalmente valor de influencia (*leverage*), h_i, de un punto, que se define de la forma siguiente:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Este valor de influencia aumenta la desviación típica del valor esperado cuando los puntos de datos están más lejos de la media de *X* y, por lo tanto, llevan a un intervalo de confianza más amplio. Se considera que un punto *i* es un punto extremo si el valor de *h* de ese punto es muy diferente de los valores de *h* de todos los demás puntos de datos. Vemos en el ejemplo siguiente que el programa Minitab identifica los puntos que tienen un elevado valor de influencia con una *X* si $h_i > 3\ p/n$, donde *p* es el número de predictores, incluida la constante. La mayoría de los paquetes estadísticos buenos permiten identificar estos puntos, pero no así el programa Excel. Utilizando esta opción, es posible identificar los puntos extremos, como muestra el ejemplo 12.4.

Los puntos atípicos son los puntos que se desvían considerablemente en la dirección de *Y* con respecto al valor predicho. Normalmente, estos puntos se identifican calculando el residuo normalizado de la forma siguiente:

$$e_{is} = \frac{e_i}{s_e\sqrt{1 - h_i}}$$

Es decir, el residuo normalizado es el residuo dividido por el error típico del residuo. Obsérvese que en la ecuación anterior los puntos que tienen un elevado valor de influencia —un elevado h_i— tienen un error típico del residuo menor, porque los puntos que tienen un elevado valor de influencia probablemente influyen en la localización de la recta de regresión estimada y, por lo tanto, el valor observado y el esperado de *Y* estarán más cerca. Minitab marca las observaciones que tienen un valor absoluto del residuo normalizado superior a 2,0 con una *R* para indicar que son casos atípicos. También las marcan la mayoría de los buenos paquetes estadísticos, pero no el Excel. Utilizando esta opción, es posible identificar los puntos atípicos, como muestra el ejemplo 12.5.

En los dos ejemplos siguientes, veremos que los puntos extremos y los casos atípicos tienen una gran influencia en la ecuación de regresión estimada en comparación con otras observaciones. En cualquier análisis aplicado, estos puntos inusuales forman parte de los datos que representan el proceso estudiado o no forman parte de ellos. En el primer caso, deben incluirse en el conjunto de datos y en el segundo caso no. El analista debe decidir. Normalmente, para tomar estas decisiones hay que comprender bien el proceso y hacer una buena valoración. En primer lugar, debe examinarse detenidamente cada punto y comprobarse su fuente. Estos puntos inusuales podrían deberse a errores de medición o de recogida de datos y, por lo tanto, se eliminarían o se corregirían. Una investigación más profunda puede revelar circunstancias excepcionales que no se espera que formen parte del proceso habitual y eso indicaría la exclusión de los puntos de datos. Las decisiones sobre qué es un proceso habitual y otras decisiones afines exigen una valoración y un examen detenidos de otra información sobre el proceso estudiado. Un buen analista utiliza los cálculos estadísticos anteriores para identificar las observaciones que deben examinarse más detenidamente, pero no se basa exclusivamente en estas medidas de identificación de las observaciones inusuales para tomar la decisión final.

EJEMPLO 12.4. El efecto de los valores extremos de *X* (análisis mediante un diagrama de puntos dispersos)

Nos interesa saber cómo afectan los valores extremos de *X* a la regresión. En este ejemplo, se analiza el efecto de los puntos que tienen valores de *X* que son muy diferentes de los otros puntos utilizando dos muestras que sólo se diferencian en dos puntos. Estos ejemplos comparativos, aunque son algo excepcionales, se utilizan para poner énfasis en el efecto que producen los puntos extremos en un análisis de regresión.

Solución

La Figura 12.12 es un diagrama de puntos dispersos con una recta de regresión trazada sobre los puntos y la 12.13 es la salida del análisis de regresión calculada con los datos. La pendiente de la recta de regresión es positiva y $R^2 = 0{,}632$. Pero obsérvese que dos puntos extremos parecen determinar la relación de regresión. Examinemos ahora el efecto de un cambio de los dos puntos de datos extremos, mostrado en las Figuras 12.14 y 12.15.

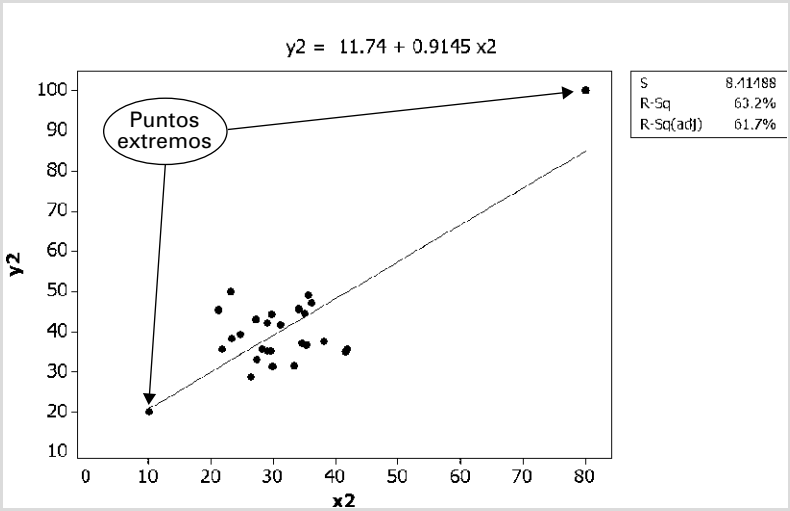


Figura 12.12. Diagrama de puntos dispersos con dos puntos extremos de *X*: pendiente positiva.

Regression Analysis: Y2 versus x2

The regression equation is
Y2 = 11.74 + 0.9145 x2

S = 8.41488 R-Sq = 63.2% R-Sq(adj) = 61.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3034.80	3034.80	42.86	0.000
Error	25	1770.26	70.81		
Total	26	4805.05			

Fitted Line: y2 versus x2

Figura 12.13. Análisis de regresión con dos puntos extremos de *X*: pendiente positiva (salida Minitab).

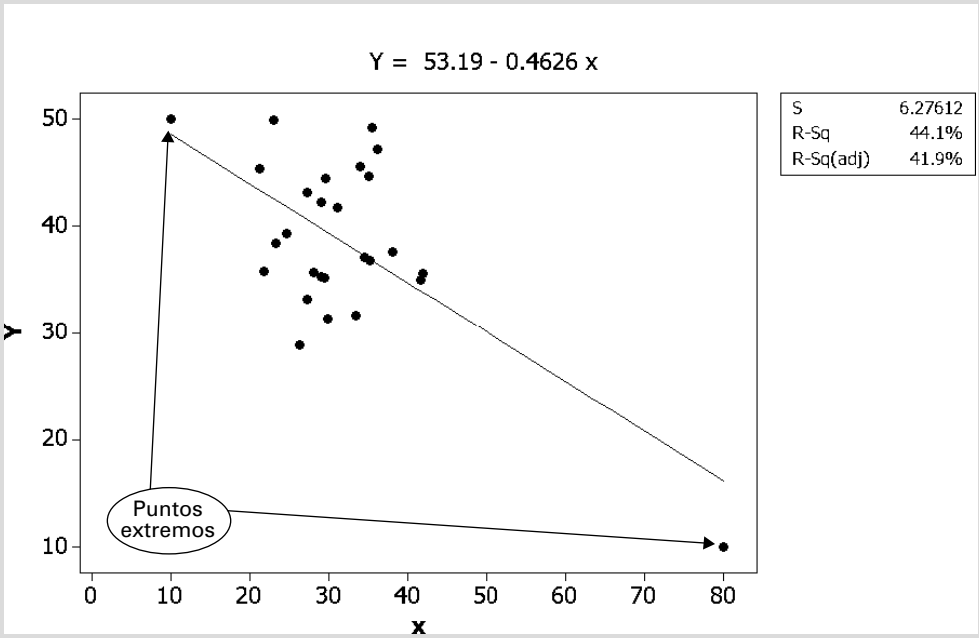


Figura 12.14. Diagrama de puntos dispersos con dos puntos extremos de *X*: pendiente negativa.

Regression Analysis: Y versus X

The regression equation is
Y1 = 53.2 − 0.463 X

Predictor	Coef	SE Coef	T	P
Constant	53.195	3.518	15.12	0.000
X1	−0.4626	0.1042	-4.44	0.000

s = 6.27612 R-Sq = 44.1% R-Sq(adj) = 41.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	776.56	776.56	19.71	0.000
Residual Error	25	984.74	39.39		
Total	26	1761.30			

Unusual Observations

Obs	X	Y	Fit	Se Fit	Residual	St Resid
7	35.5	49.14	36.78	1.27	12.37	2.01R
26	80.0	10.00	16.19	5.17	−6.19	−1.74 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

La observación 26 es un punto extremo con gran influencia

Figura 12.15. Análisis de regresión con dos puntos extremos de *X*: pendiente negativa (salida Minitab).

Como consecuencia del cambio de dos puntos de datos solamente, ahora la relación tiene una pendiente negativa estadísticamente significativa y las predicciones serían muy diferentes. Sin examinar los diagramas de puntos dispersos, no sabríamos por qué la pendiente que se obtiene es positiva o negativa. Podríamos haber pensado que nuestros resultados representaban una situación de regresión normal como la que hemos visto en el diagrama de puntos dispersos de las ventas al por menor. Obsérvese que en la Figura 12.15 la observación 26 se ha denominado observación extrema mediante el símbolo *X*.

Este ejemplo muestra un problema que se plantea habitualmente cuando se utilizan datos históricos. Supongamos que *X* es el número de trabajadores que trabajan en un turno de producción e *Y* es el número de unidades producidas en ese turno. La mayor parte del tiempo la fábrica tiene una plantilla relativamente estable y la producción depende en gran parte de la cantidad de materias primas existentes y de las necesidades de ventas. La producción se ajusta al alza o a la baja en un rango estrecho en respuesta a las demandas y a la plantilla existente, *X*. Por lo tanto, vemos que en la mayoría de los casos el diagrama de puntos dispersos cubre un estrecho rango de la variable *X*. Pero a veces hay una plantilla muy grande o muy pequeña, o el número de trabajadores se ha registrado incorrectamente. Esos días la producción puede ser excepcionalmente grande o pequeña o puede registrarse incorrectamente. Como consecuencia, tenemos puntos extremos que pueden influir mucho en el modelo de regresión. Estos pocos días determinan los resultados de la regresión. Sin los puntos extremos, la regresión indicaría que la relación es pequeña o nula. Si estos puntos extremos representan extensiones de la relación, el modelo estimado es útil. Pero si estos puntos se deben a condiciones excepcionales o a errores de recogida de datos, el modelo estimado es engañoso.

En una aplicación podemos observar que estos puntos extremos son correctos y deben utilizarse para trazar la recta de regresión. Pero el analista tiene que tomar esa decisión sabiendo que ninguno de los demás puntos de datos apoya la existencia de una relación significativa. De hecho, es necesario realizar un estudio detenido para comprender el sistema y el proceso que generaron los datos y para evaluar los datos de los que se dispone.

EJEMPLO 12.5. El efecto de los valores atípicos de la variable *Y* (análisis mediante un diagrama de puntos dispersos)

En este ejemplo consideramos el efecto de los valores atípicos en sentido vertical. Recuérdese que el modelo del análisis de regresión supone que toda la variación se produce en el sentido de las *Y*. Sabemos, pues, que los valores atípicos en el sentido de las *Y* tendrán grandes residuos y éstos residuos darán como resultado una estimación mayor del error del modelo. En este ejemplo, veremos que los efectos pueden ser aún más extremos.

Solución

Para comenzar, observemos el diagrama de puntos dispersos y el análisis de regresión de las Figuras 12.16 y 12.17. En este ejemplo, tenemos una estrecha relación entre las variables *X* e *Y*. El diagrama de puntos dispersos apoya claramente la existencia de una relación lineal, estimándose que $b_1 = 11{,}88$. Además, el R^2 del modelo de regresión es cercano a 1 y el estadístico *t* de Student es muy alto. Es evidente que tenemos pruebas contundentes para apoyar un modelo lineal.

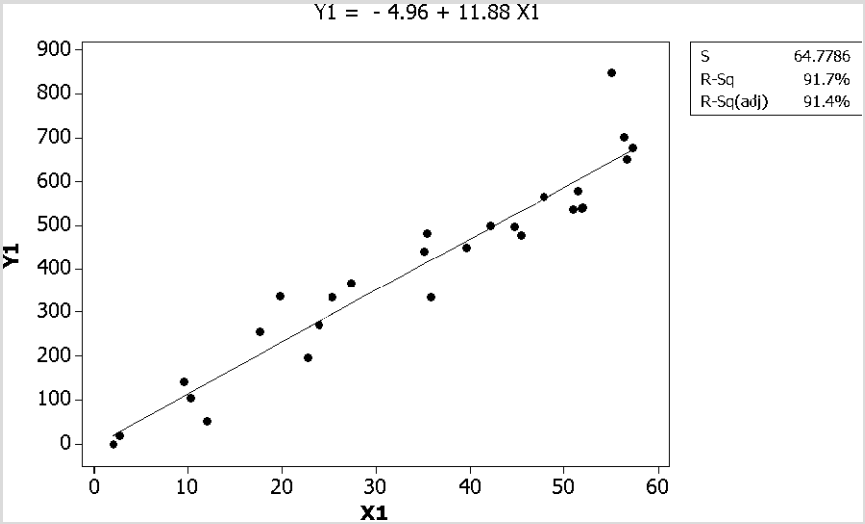


Figura 12.16. Diagrama de puntos dispersos con una pauta prevista.

Regression Analysis: Y1 versus X1

The regression equation is
Y1 = −4.96 + 11.88 X1

s = 64.7786 R-Sq = 91.7% R-Sq(adj) = 91.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1160171	1160171	276.48	0.000
Error	25	104907	4196		
Total	26	1265077			

Fitted Line: Y1 versus X1

Figura 12.17. Regresión con una pauta prevista (salida Minitab).

Veamos ahora cómo afecta un cambio de dos observaciones a los puntos atípicos, como muestra la Figura 12.18, que podría deberse a un error en la recogida de los datos o a la presencia de unas circunstancias muy poco habituales en el proceso estudiado.

La pendiente de la recta de regresión sigue siendo positiva, pero ahora $b_1 = 6{,}40$ y la estimación de la pendiente tiene un error típico mayor, como muestra la Figura 12.19. El intervalo de confianza es mucho más amplio y el valor predicho a partir de la recta de regresión no es tan preciso. Ahora el modelo de regresión correcto no está tan claro. El programa Minitab identifica las observaciones 26 y 27 como observaciones atípicas imprimiendo una R al lado del residuo normalizado. Los residuos normalizados cuyo valor absoluto es superior a 2 se indican en la salida. Si los dos puntos extremos ocurrieron realmente en el funcionamiento normal del proceso, deberíamos incluirlos en

nuestro análisis. Pero el hecho de que se desvíen tanto de la pauta indica que debemos investigar atentamente las situaciones de los datos que generaron esos puntos y estudiar el proceso examinado.

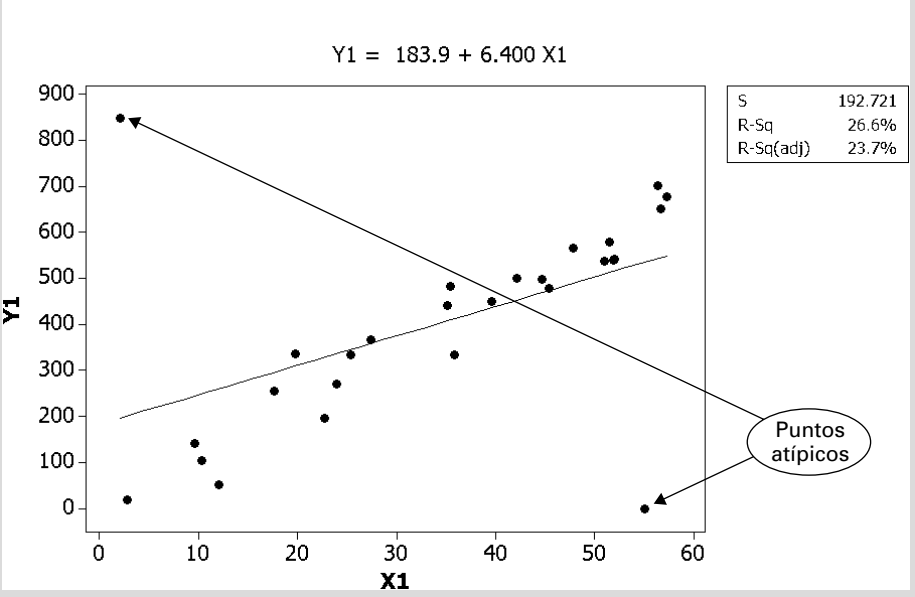


Figura 12.18. Diagrama de puntos dispersos con puntos atípicos de *Y*.

Regression Analysis: Y1 versus X1

The regression equation is
Y1 = 184 + 6.40 X1

Predictor	Coef	SE Coef	T	P
Constant	183.92	82.10	2.24	0.034
X1	6.400	2.126	3.01	0.006

S = 192.721 R-Sq = 26.6% R-Sq(adj) = 23.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	336540	336540	9.06	0.006
Residual Error	25	928537	37141		
Total	26	1265077			

Unusual Observations

Obs	X1	Y1	Fit	Se Fit	Residual	St Resid
26	2.0	850.0	196.7	78.3	653.3	3.71R
27	55.0	0.0	535.9	57.3	−535.9	−2.91R

Casos atípicos marcados con una R

R denotes an observation with a large standardized residual.

Figura 12.19. Regresión con puntos atípicos de *Y* (salida Minitab).

Podríamos proponer otros muchos ejemplos. Podríamos observar que el diagrama de puntos dispersos sugiere la existencia de una relación no lineal y, por lo tanto, sería un modelo mejor para un problema específico. En los Capítulos 13 y 14, veremos cómo puede utilizarse la regresión para analizar relaciones no lineales. Observaremos muchas pautas de datos a medida que examinemos distintas aplicaciones del análisis de regresión. Lo importante es que debemos seguir regularmente los métodos del análisis —incluida la realización de diagramas de puntos dispersos— que puedan suministrar la mayor información posible. Como buen analista, debe **«¡Conocer sus datos!»** En el capítulo siguiente vemos cómo pueden utilizarse también los residuos gráficamente para realizar más contrastes de los modelos de regresión.

EJERCICIOS

Ejercicios básicos

12.60. Frank Anscombe, alto ejecutivo encargado de la investigación, le ha pedido que analice los cuatro modelos lineales siguientes utilizando los datos que contiene el fichero de datos **Anscombe**.

$$Y_1 = \beta_0 + \beta_1 X_1$$
$$Y_2 = \beta_0 + \beta_1 X_1$$
$$Y_3 = \beta_0 + \beta_1 X_1$$
$$Y_4 = \beta_0 + \beta_1 X_1$$

Utilice su paquete informático para estimar una regresión lineal para cada modelo. Trace un diagrama de puntos dispersos de los datos utilizados en cada modelo. Escriba un informe, incluyendo los resultados del análisis de regresión y el gráfico, que compare y contraste los cuatro modelos.

Ejercicio aplicado

12.61. John Foster, presidente de Public Research Inc., le ha pedido ayuda para estudiar el nivel de delincuencia existente en diferentes estados de Estados Unidos antes y después de la realización de elevados gastos federales para reducir la delincuencia. Quiere saber si se puede predecir la tasa de delincuencia en el caso de algunos delitos después de realizados los gastos utilizando la tasa de delincuencia existente antes de realizar los gastos. Le ha pedido que contraste la hipótesis de que la delincuencia existente antes predice la delincuencia posterior en el caso de la tasa total de delincuencia y de las tasas de asesinato, violación y robo. Los datos para su análisis se encuentran en el fichero de datos **Crime Study**. Realice el análisis adecuado y escriba un informe que resuma sus resultados.

RESUMEN

En este capítulo hemos desarrollado los modelos de dos variables o de mínimos cuadrados simples. Nos hemos basado en algunos de los conceptos descriptivos iniciales presentados en el Capítulo 3. El modelo de regresión simple supone que un conjunto de variables exógenas o independientes tiene una relación lineal con el valor esperado de una variable aleatoria endógena o dependiente. Desarrollando estimaciones de los coeficientes de este modelo, podemos comprender mejor los procesos empresariales y económicos y podemos predecir los valores de la variable endógena en función de la variable exógena. En nuestro estudio, hemos desarrollado estimadores de los coeficientes y de las variables dependientes. También hemos desarrollado medidas de la bondad del ajuste de la regresión: análisis de la varianza y de R^2.

Después de ese estudio, hemos presentado métodos de inferencia estadística: contraste de hipótesis e intervalos de confianza de los estimadores de regresión fundamentales. También hemos examinado el análisis de correlación, analizando simplemente la relación entre dos variables. Por último, hemos examinado la importancia de los diagramas de puntos dispersos y el análisis gráfico del desarrollo y el contraste de modelos de regresión.

TÉRMINOS CLAVE

análisis de la varianza, 450
base para la inferencia sobre la pendiente de la regresión poblacional, 459
coeficiente de determinación, R^2, 451
contraste F para el coeficiente de regresión simple, 464
contrastes de la correlación poblacional nula, 433
contrastes de la pendiente de la regresión poblacional, 461
correlación y R^2, 454
distribución en el muestreo del estimador de los coeficientes por mínimos cuadrados, 458
estimación de la varianza del error del modelo, 454
estimadores de los coeficientes, 442
intervalos de confianza de las predicciones, 467
intervalos de confianza de la pendiente de la regresión poblacional b_1, 462
método de mínimos cuadrados, 442
regresión lineal basada en un modelo poblacional, 440
resultados de la regresión lineal, 441
supuestos para los estimadores de los coeficientes por mínimos cuadrados, 442

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

12.62. ¿Qué significa la afirmación de que un par de variables aleatorias están correlacionadas positivamente? Ponga ejemplos de pares de variables aleatorias en los que espera que exista

a) una correlación positiva
b) una correlación negativa
c) una correlación nula

12.63. Una muestra aleatoria de cinco conjuntos de observaciones de un par de variables aleatorias dio los resultados de la tabla adjunta.

X	4	1	0	1	4
Y	−2	−1	0	1	2

a) Halle el coeficiente de correlación muestral.
b) Teniendo en cuenta el hecho de que cada valor de y_i es el cuadrado del valor correspondiente de x_i, comente su respuesta al apartado (a).

12.64. En una muestra aleatoria de 53 tiendas de una cadena de grandes almacenes se observó que la correlación entre las ventas anuales en euros por metro cuadrado de superficie y el alquiler anual en euros por metro cuadrado de superficie era 0,37. Contraste la hipótesis nula de que estas dos cantidades no están correlacionadas en la población frente a la hipótesis alternativa de que la correlación poblacional es positiva.

12.65. En una muestra aleatoria de 526 empresas, se observó que la correlación muestral entre la proporción de directivos que son consejeros y una medida del rendimiento de las acciones de la empresa ajustada para tener en cuenta el riesgo era de 0,1398. Contraste la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa bilateral.

12.66. En una muestra de 66 meses se observó que la correlación entre los rendimientos de los bonos a 10 años de Canadá y de Hong Kong era de 0,293. Contraste la hipótesis nula de que la correlación poblacional es 0 frente a la hipótesis alternativa de que es positiva.

12.67. En una muestra aleatoria de 192 mujeres trabajadoras, se observó una correlación muestral de −0,18 entre la edad y una medida de la disposición a cambiar de empleo. Basándose únicamente en esta información, extraiga todas las conclusiones que pueda sobre la regresión de la disposición a cambiar de empleo con respecto a la edad.

12.68. Basándose en una muestra de n observaciones, (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), se calcula la regresión muestral de y con respecto a x. Demuestre que la recta de regresión muestral pasa por el punto $(x = \bar{x}, y = \bar{y})$, donde $\bar{x}$ e $\bar{y}$ son las medias muestrales.

12.69. Una empresa realiza normalmente un test de aptitud a todo el nuevo personal en formación. Al final del primer año en la empresa, este personal en formación es valorado por sus supervisores inmediatos. En una muestra aleatoria de 12 personas en formación, se obtuvieron los resultados mostrados en el fichero de datos **Employee Test**.

a) Estime la regresión de la valoración realizada por el supervisor con respecto a la puntuación obtenida en el test de aptitud.

b) Interprete la pendiente de la recta de regresión muestral.
c) ¿Es posible dar una interpretación útil a la ordenada en el origen de la recta de regresión muestral?
d) Halle e interprete el coeficiente de determinación de esta regresión.
e) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa unilateral obvia.
f) Halle el intervalo de confianza al 95 por ciento de la valoración que daría el supervisor a una persona en formación que tuviera una puntuación de 70 en el test de aptitud.

12.70. Se ha intentado evaluar la tasa de inflación como predictor del tipo al contado en el mercado de letras del Tesoro alemanas. Partiendo de una muestra de 79 observaciones trimestrales, se obtuvo la regresión lineal estimada

$$\hat{y} = 0{,}0027 + 0{,}7916x$$

donde

y = variación efectiva del tipo al contado
x = variación del tipo al contado predicha por la tasa de inflación

El coeficiente de determinación era 0,097 y la desviación típica estimada del estimador de la pendiente de la recta de regresión poblacional era 0,2759.

a) Interprete la pendiente de la recta de regresión estimada.
b) Interprete el coeficiente de determinación.
c) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa de que la verdadera pendiente es positiva e interprete su resultado.
d) Contraste la hipótesis nula de que la pendiente de la recta de regresión poblacional es 1 frente a la hipótesis alternativa bilateral.

12.71. La tabla muestra las compras por comprador de ocho cosechas de un vino selecto (y) y la valoración del vino realizada por el comprador en un año (x).

x	3,6	3,3	2,8	2,6	2,7	2,9	2,0	2,6
y	24	21	22	22	18	13	9	6

a) Estime la regresión de las compras por comprador con respecto a la valoración realizada por el comprador.
b) Interprete la pendiente de la recta de regresión estimada.
c) Halle e interprete el coeficiente de determinación.
d) Halle e interprete el intervalo de confianza al 90 por ciento de la pendiente de la recta de regresión poblacional.
e) Halle el intervalo de confianza al 90 por ciento de las compras esperadas por comprador de una cosecha a la que el comprador da una valoración de 2,0.

12.72. En una muestra de 306 estudiantes de un curso básico de estadística, se obtuvo la recta de regresión muestral

$$y = 58{,}813 + 0{,}2875x$$

donde

y = calificación final de los estudiantes al terminar el curso
x = calificación en un examen de posición realizado al principio de curso.

El coeficiente de determinación era 0,1158 y la desviación típica estimada del estimador de la pendiente de la recta de regresión poblacional era 0,04566.

a) Interprete la pendiente de la recta de regresión muestral.
b) Interprete el coeficiente de determinación.
c) La información dada permite contrastar la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 de dos formas distintas frente a la hipótesis alternativa de que es positiva. Realice estos contrastes y muestre que llegan a la misma conclusión.

12.73. Basándose en una muestra de 30 observaciones, se estimó el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Las estimaciones por mínimos cuadrados obtenidas fueron

$$b_0 = 10{,}1 \quad \text{y} \quad b_1 = 8{,}4$$

La suma de los cuadrados de la regresión y la suma de los cuadrados de los errores fueron

$$SCR = 128 \quad \text{y} \quad SCE = 286$$

a) Halle e interprete el coeficiente de determinación.
b) Contraste al nivel de significación del 10 por ciento la hipótesis nula de que β_1 es 0 frente a la hipótesis alternativa bilateral.

c) Halle

$$\sum_{i=1}^{30} (x_i - \bar{x})^2$$

12.74. Basándose en una muestra de 25 observaciones, se estimó el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Las estimaciones por mínimos cuadrados obtenidas fueron

$$b_0 = 15{,}6 \quad \text{y} \quad b_1 = 1{,}3$$

La suma total de los cuadrados y la suma de los cuadrados de los errores fueron

$$STC = 268 \quad \text{y} \quad SCE = 204$$

a) Halle e interprete el coeficiente de determinación.
b) Contraste al nivel de significación del 5 por ciento la hipótesis nula de que la pendiente de la recta de regresión poblacional es 0 frente a la hipótesis alternativa bilateral.
c) Halle el intervalo de confianza al 95 por ciento de β_1.

12.75. Un analista cree que el único determinante importante de los rendimientos de los activos (Y) del banco es el cociente entre los préstamos y los depósitos (x). En una muestra aleatoria de 20 bancos se obtuvo la recta de regresión muestral

$$Y = 0{,}97 + 0{,}47x$$

con el coeficiente de determinación de 0,720.

a) Halle la correlación muestral entre los rendimientos de los activos y el cociente entre los préstamos y los depósitos.
b) Contraste la hipótesis nula de que no existe una relación lineal entre los rendimientos y el cociente frente a una hipótesis alternativa bilateral.
c) Halle

$$\frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

12.76. Comente la siguiente afirmación:

Si se estima una regresión del rendimiento por acre del maíz con respecto a la cantidad de fertilizante utilizada empleando las cantidades de fertilizante utilizadas normalmente por los agricultores, la pendiente de la recta de regresión estimada será, desde luego, positiva. Sin embargo, es bien sabido que si se utiliza una cantidad muy grande de fertilizante, el rendimiento del maíz es muy bajo. Por lo tanto, las ecuaciones de regresión no son muy útiles para hacer predicciones.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

12.77. El departamento de economía de una universidad está intentando averiguar si los conocimientos verbales o matemáticos son más importantes para predecir el éxito académico en los estudios de economía. El profesorado del departamento ha decidido utilizar como medida del éxito la calificación media (GPA) obtenida por los licenciados en los cursos de economía. Los conocimientos verbales se miden por medio de las calificaciones obtenidas en dos exámenes estandarizados: el SAT verbal y el ACT de inglés. Los conocimientos matemáticos se miden por medio de las calificaciones obtenidas en el SAT de matemáticas y en el ACT de matemáticas. El fichero de datos llamado **Student GPA**, que se encuentra en su disco de datos, contiene los datos de 112 estudiantes. El nombre de las columnas de las variables se indica al comienzo del fichero de datos. Debe utilizar el paquete estadístico que utilice habitualmente para realizar el análisis de este problema.

a) Represente gráficamente la GPA de economía en relación con cada una de las dos calificaciones de los conocimientos verbales y cada una de las dos calificaciones de los conocimientos matemáticos. ¿Qué variable es el mejor predictor? Observe las pautas poco habituales que haya en los datos.
b) Calcule los coeficientes del modelo lineal y los estadísticos del análisis de regresión para los modelos que predicen la GPA de economía en función de cada calificación en conocimientos verbales y cada calificación en conocimientos matemáticos. Utilizando tanto las medidas matemáticas y verbales del SAT como las medidas de matemáticas e inglés del ACT, averigüe si los conocimientos matemáticos o verbales son el mejor predictor de la GPA de economía.
c) Compare los estadísticos descriptivos —la media, la desviación típica, el cuartil superior y el inferior, el rango— de las variables consideradas predictoras. Observe las diferencias e indique cómo afectan estas diferencias a la capacidad del modelo lineal para realizar predicciones.

12.78. Los responsables de la National Highway Traffic Safety Administration (NHTSA) de Estados Unidos quieren saber si los diferentes tipos de vehículos de un estado tienen relación con la tasa de mortalidad en carretera del esta-

do. Le han pedido que realice varios análisis de regresión para averiguar si el peso medio de los vehículos, el porcentaje de automóviles importados, el porcentaje de camiones ligeros o la antigüedad media de los automóviles están relacionados con las muertes en accidente ocurridas en automóviles y camionetas. Los datos del análisis se encuentran en el fichero de datos llamado **Crash**, que está en su disco de datos. Las descripciones y las localizaciones de las variables se encuentran en el catálogo del fichero de datos del apéndice.

a) Represente gráficamente las muertes en accidente en relación con cada una de las variables potenciales de predicción. Observe la relación y cualquier pauta excepcional en los puntos de datos.

b) Realice un análisis de regresión simple de las muertes totales en accidente con respecto a las variables potenciales de predicción. Indique si alguna de las regresiones muestra una relación significativa y, en caso afirmativo, cuál.

c) Muestre los resultados de su análisis y ordene las variables de predicción según su relación con las muertes totales en accidente.

12.79. El Departamento de Transporte de Estados Unidos desea saber si los estados que tienen un porcentaje mayor de población urbana tienen una tasa más alta de muertes totales en accidente ocurridas en automóviles y camionetas. También quiere saber si existe alguna relación entre la velocidad media a la que se conduce por las carreteras rurales o el porcentaje de carreteras rurales que están asfaltadas y las tasas de muertes en accidente. Los datos de este estudio se encuentran en el fichero de datos **Crash** almacenado en su disco de datos.

a) Represente gráficamente las muertes en accidente en relación con cada una de las variables potenciales de predicción. Observe la relación y cualquier pauta excepcional en los puntos de datos.

b) Realice un análisis de regresión simple de las muertes en accidente con respecto a las variables potenciales de predicción.

c) Muestre los resultados de su análisis y ordene las variables de predicción según su relación con las muertes totales en accidente.

12.80. Un economista desea predecir el valor de mercado de las viviendas de pequeñas ciudades del Medio Oeste ocupadas por sus propietarios. Ha reunido un conjunto de datos de 45 pequeñas ciudades que se refieren a un periodo de dos años y quiere que los utilice como fuente de datos para el análisis. Los datos se encuentran en el fichero **Citydat**, que están en su disco de datos. Quiere que desarrolle dos ecuaciones de predicción: una que utilice el tamaño de la vivienda como predictor y otra que utilice el tipo impositivo como predictor.

a) Represente gráficamente el valor de mercado de las viviendas (hseval) en relación con el tamaño de la vivienda (sizense) y en relación con el tipo impositivo (taxrate). Observe cualquier pauta excepcional en los datos.

b) Realice análisis de regresión para las dos variables de predicción. ¿Qué variable predice mejor el valor de las viviendas?

c) Un promotor industrial de un estado del Medio Oeste ha afirmado que los tipos del impuesto local sobre bienes inmuebles de las pequeñas ciudades debe bajarse porque, en caso contrario, nadie comprará una vivienda en estas ciudades. Basándose en su análisis de este problema, evalúe la afirmación del promotor.

12.81. Stuart Wainwright, vicepresidente de compras para una gran cadena nacional de tiendas de Estados Unidos, le ha pedido que realice un análisis de las ventas al por menor por estados. Quiere saber si el porcentaje de desempleados o la renta personal per cápita están relacionados con las ventas al por menor per cápita. Los datos para realizar este estudio se encuentran en el fichero de datos llamado **Retail**, que está almacenado en su disco de datos y se describe en el catálogo del fichero de datos del apéndice.

a) Trace gráficos y realice análisis de regresión para averiguar las relaciones entre las ventas al por menor per cápita y el porcentaje de desempleados y la renta personal. Calcule intervalos de confianza al 95 por ciento para los coeficientes de la pendiente de cada ecuación de regresión.

b) ¿Cómo afecta una disminución de la renta per cápita de 1.000 $ a las ventas per cápita?

c) ¿Cuál es el intervalo de confianza al 95 por ciento en la ecuación de la renta per cápita de las ventas al por menor correspondientes a la renta media per cápita y a un nivel que esté 1.000 $ por encima de la renta media per cápita?

12.82. Un importante proveedor nacional de materiales de construcción para la construcción de viviendas está preocupado por las ventas totales

del próximo año. Es bien sabido que las ventas de la empresa están relacionadas directamente con la inversión nacional total en vivienda. Algunos banqueros de Nueva York están prediciendo que los tipos de interés subirán alrededor de 2 puntos porcentuales el próximo año. Le han pedido que realice un análisis de regresión para poder predecir el efecto de las variaciones de los tipos de interés en la inversión en vivienda. Los datos de series temporales para realizar este estudio se encuentran en el fichero de datos llamado **Macro2003**, que está almacenado en su disco de datos y se describe en el apéndice del Capítulo 14.

a) Desarrolle dos modelos de regresión para predecir la inversión en vivienda utilizando el tipo de interés preferencial para uno y el tipo de interés de los fondos federales para el otro. Analice los estadísticos de la regresión e indique qué ecuación hace las mejores predicciones.

b) Halle el intervalo de confianza al 95 por ciento del coeficiente de la pendiente en ambas ecuaciones de regresión.

c) Basándose en cada modelo, prediga cómo afecta una subida de los tipos de interés de 2 puntos porcentuales a la inversión en vivienda.

d) Utilizando ambos modelos, calcule intervalos de confianza al 95 por ciento de la variación de la inversión en vivienda provocada por una subida de los tipos de interés de 2 puntos porcentuales.

Apéndice

En este apéndice mostramos cómo se estiman por mínimos cuadrados los parámetros poblacionales de regresión. Queremos hallar los valores b_0 y b_1 tales que la suma de los cuadrados de las discrepancias

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

sea lo más pequeña posible.

En primer lugar, mantenemos constante b_1 y diferenciamos con respecto a b_0, lo que nos da

$$\frac{\partial SCE}{\partial b_0} = 2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$= -2\left(\sum y_i - nb_0 - b_1 \sum x_i\right)$$

Dado que esta derivada debe ser 0 para obtener un mínimo, tenemos que

$$\sum y_i - nb_0 - b_1 \sum x_i = 0$$

Por lo tanto, dividiendo por n resulta que

$$b_0 = \bar{y} - b_1 \bar{x}$$

Introduciendo este resultado de b_0 en la expresión anterior, tenemos que

$$SCE = \sum_{i=1}^{n} [(y_i - \bar{y}) - b_1(x_i - \bar{x})]^2$$

Diferenciando esta expresión con respecto a b_1, obtenemos

$$\frac{\partial SCE}{\partial b_1} = 2 \sum_{i=1}^{n} (x_i - \bar{x})[(y_i - \bar{y}) - b_1(x_i - \bar{x})$$

$$= -2\left(\sum (x_i - \bar{x})(y_i - \bar{y}) - b_1 \sum (x_i - \bar{x})^2\right)$$

Esta derivada debe ser 0 para obtener un mínimo, por lo que tenemos que

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = b_1 \sum (x_i - \bar{x})^2$$

Por lo tanto,

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Bibliografía

1. Dhalla, N. K., «Short-Term Forecasts of Advertising Expenditures», *Journal of Advertising Research*, 19, n.º 1, 1979, págs. 7-14.
2. Mampower, J. L., S. Livingston y T. J. Lee, «Expert Judgments of Political Risk», *Journal of Forecasting*, 6, 1987, págs. 51-65.

Capítulo 13

Regresión múltiple

Esquema del capítulo

Introducción

En el Capítulo 12 presentamos el método de regresión simple para obtener una ecuación lineal que predice una variable dependiente o endógena en función de una única variable independiente o exógena; por ejemplo, el número total de artículos vendidos en función del precio. Sin embargo, en muchas situaciones, varias variables independientes influyen conjuntamente en una variable dependiente. La regresión múltiple nos permite averiguar el efecto simultáneo de varias variables independientes en una variable dependiente utilizando el principio de los mínimos cuadrados.

Existen muchas aplicaciones importantes de la regresión múltiple en el mundo de la empresa y en la economía. Entre estas aplicaciones se encuentran las siguientes:

1. La cantidad vendida de bienes es una función del precio, la renta, la publicidad, el precio de los bienes sustitutivos y otras variables.
2. Existe inversión de capital cuando un empresario cree que puede obtener un beneficio. Por lo tanto, la inversión de capital es una función de variables relacionadas con las posibilidades de obtener beneficios, entre las que se encuentran el tipo de interés, el producto interior bruto, las expectativas de los consumidores, la renta disponible y el nivel tecnológico.
3. El salario es una función de la experiencia, la educación, la edad y el puesto de trabajo.
4. Las grandes empresas del comercio al por menor y la hostelería deciden la localización de los nuevos establecimientos basándose en los ingresos previstos por ventas y/o en la rentabilidad. Utilizando datos de localizaciones anteriores que han tenido éxito y que no lo han tenido, los analistas pueden construir modelos que predicen las ventas o los beneficios de una nueva localización posible.

El análisis económico y empresarial tiene algunas características únicas en comparación con el análisis de otras disciplinas. Los científicos naturales trabajan en un laboratorio en el que es posible controlar muchas variables, pero no todas. En cambio, el laboratorio del economista y del directivo es el mundo y las condiciones no pueden controlarse. Por lo tanto, necesitan instrumentos como la regresión múltiple para estimar el efecto simultáneo de varias variables. La regresión múltiple como «instrumento de laboratorio» es muy importante para el trabajo de los directivos y de los economistas. En este capítulo veremos muchas aplicaciones específicas en los ejemplos y los ejercicios.

Los métodos para ajustar modelos de regresión múltiple se basan en el mismo principio de los mínimos cuadrados que aprendimos en el Capítulo 12 y, por lo tanto, las ideas presentadas en ese capítulo se extenderán directamente a la regresión múltiple. Sin embargo, se introducen algunas complejidades debido a las relaciones entre las distintas variables exógenas. Éstas requieren nuevas ideas que se desarrollan en este capítulo.

13.1. El modelo de regresión múltiple

Nuestro objetivo es aprender a utilizar la regresión múltiple para crear y analizar modelos. Por lo tanto, aprendemos cómo funciona la regresión múltiple y algunas directrices para interpretarla. Comprendiendo perfectamente la regresión múltiple, es posible resolver una amplia variedad de problemas aplicados. Este estudio de los métodos de regresión múltiple es paralelo al de la regresión simple. El primer paso para desarrollar un modelo es la especificación de ese modelo, que consiste en la selección de las variables del modelo y de la forma del modelo. A continuación, se estudia el método de mínimos cuadrados y se analiza la variabilidad para identificar los efectos de cada una de las variables de predicción. Después se estudia la estimación, los intervalos de confianza y el contraste de hipótesis. Se utilizan frecuentemente aplicaciones informáticas para indicar cómo se aplica la teoría a problemas realistas. El estudio de este capítulo será más fácil si se ponen en relación sus ideas con las que presentamos en el Capítulo 12.

Especificación del modelo

Comenzamos con una aplicación que ilustra la importante tarea de la especificación del modelo de regresión. La especificación del modelo consiste en la selección de las variables exógenas y la forma funcional del modelo.

EJEMPLO 13.1. Proceso de producción (especificación del modelo de regresión)

El director de producción de Circuitos Flexibles, S.A., le ha pedido ayuda para estudiar un proceso de producción. Los circuitos flexibles se producen con un rollo continuo de resina flexible que lleva adherida a su superficie una fina película de material conductor hecho de cobre. El cobre se adhiere a la resina pasando la resina por una solución de cobre. El grosor del cobre es fundamental para que los circuitos sean de buena calidad. Depende en parte de la temperatura de la solución de cobre, de la velocidad de la línea de producción, de la densidad de la solución y del grosor de la resina flexible. Para controlar el grosor del cobre adherido a la superficie, el director de producción necesita saber qué efecto produce cada una de estas variables. Le ha pedido ayuda para desarrollar un modelo de regresión múltiple.

Solución

La regresión múltiple puede utilizarse para hacer estimaciones del efecto que produce cada variable en combinación con las demás. El desarrollo del modelo comienza con un análisis detenido del contexto del problema. El primer paso en este ejemplo sería una extensa conversación con los ingenieros responsables del diseño del producto y de la producción, con el fin de comprender detalladamente el proceso del que se pretende desarrollar un modelo. En algunos casos, se estudiaría la literatura existente sobre el proceso. Éste debe ser comprendido y aceptado por todos los interesados antes de poder desarrollar un modelo útil utilizando el análisis de regresión múltiple. En este ejemplo, la variable dependiente, Y, es el grosor del cobre. Las variables independientes son la temperatura de la solución de cobre, X_1; la velocidad de la línea de producción, X_2; la densidad de la solución, X_3, y el grosor de la resina flexible, X_4. Los ingenieros y los científicos que comprendían la tecnología del proceso de recubrimiento identificaron estas variables como posibles predictores del grosor del cobre, Y. Basándose en el estudio del proceso, la especificación del modelo resultante es

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

En el modelo lineal anterior, las β_j' son coeficientes lineales constantes de las X_j que indican el efecto condicionado de cada variable independiente en la determinación de la variable dependiente, Y, en la población. Por lo tanto, las β_j' son parámetros en el modelo de regresión lineal. A continuación, se produciría una serie de lotes para hacer mediciones de distintas combinaciones de las variables independientes y la variable dependiente (véase el análisis del diseño experimental en el apartado 14.2).

EJEMPLO 13.2. Localización de las tiendas (especificación del modelo)

El director de planificación de una gran cadena de comercio al por menor estaba insatisfecho con su experiencia en la apertura de nuevas tiendas. En los cuatro últimos años, el 25 por ciento de las nuevas tiendas no había conseguido las ventas previstas en el periodo de prueba de dos años y se había cerrado con cuantiosas pérdidas económicas. El director quería desarrollar mejores criterios para elegir el emplazamiento de las tiendas y llegó a la conclusión de que debía estudiarse la experiencia histórica de las tiendas que habían tenido éxito y las que habían fracasado.

Solución

Hablando con un consultor, llegó a la conclusión de que podían utilizarse los datos de las tiendas que habían conseguido las ventas que estaban previstas y los datos de las que no las habían conseguido para desarrollar un modelo de regresión múltiple. El consultor sugirió que debía utilizarse como variable dependiente, Y, las ventas del segundo año. Se emplearía un modelo de regresión para predecir las ventas del segundo año en función de varias variables independientes que definen la zona que rodea a la tienda. Sólo se abrirían tiendas en los lugares en los que las ventas predichas superaran un nivel mínimo. El modelo también indicaría cómo afectan varias variables independientes a las ventas.

Tras hablar largo y tendido con personas de la empresa, el consultor recomendó las siguientes variables independientes:

1. X_1 = tamaño de la tienda
2. X_2 = volumen de tráfico de la calle en la que se encuentra la tienda
3. X_3 = apertura de la tienda sola o en un centro comercial
4. X_4 = existencia de una tienda rival a menos de 500 metros
5. X_5 = renta per cápita de la población residente a menos de 8 kilómetros
6. X_6 = número total de personas que residen a menos de 8 kilómetros
7. X_7 = renta per cápita de la población que reside a menos de 15 kilómetros
8. X_8 = número total de personas que residen a menos de 15 kilómetros

Se utilizó la regresión múltiple para estimar los coeficientes del modelo de predicción de las ventas a partir de datos recogidos en todas las tiendas abiertas en los ocho últimos años. En el conjunto de datos había tiendas que seguían abiertas y tiendas que se habían cerrado. Se desarrolló un modelo que podía utilizarse para predecir las ventas del segundo año. Este modelo contenía estimadores, b'_j, de los parámetros del modelo, β'_j. Para aplicar el modelo

$$\hat{y}_i = b_0 + \sum_{j=1}^{8} b_j x_{ji}$$

se hicieron mediciones de las variables independientes de cada nueva localización propuesta y se calcularon las ventas predichas de cada localización. Se utilizó el nivel predicho de ventas, junto con el criterio de los analistas de marketing y de un comité de directores de tiendas de éxito, para elegir el lugar en el que se abrirían tiendas.

En la estrategia para especificar un modelo influyen los objetivos del modelo. Uno de los objetivos es la predicción de una variable dependiente o «de resultado». Entre las aplicaciones se encuentran la predicción de las ventas, de la producción, del consumo total, de la inversión total y otros muchos criterios de los resultados empresariales y económicos. El segundo objetivo es estimar el efecto marginal de cada variable independiente. Los economistas y los directivos necesitan saber cómo cambian las medidas de los resultados cuando varían las variables independientes, X_j, donde $j = 1, ..., K$. Por ejemplo:

1. ¿Cómo varían las ventas como consecuencia de una subida del precio y de los gastos publicitarios?
2. ¿Cómo varía la producción cuando se alteran las cantidades de trabajo y de capital?
3. ¿Disminuye la mortalidad infantil cuando se incrementan los gastos en asistencia sanitaria y en servicios de saneamiento?

Objetivos de la regresión

La regresión múltiple permite obtener dos importantes resultados:

1. Una ecuación lineal estimada que predice la variable dependiente, *Y*, en función de *K* variables independientes observadas, x_j, donde $j = 1, ..., K$.

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki}$$

   donde $i = 1, ..., n$ observaciones.
2. La variación marginal de la variable dependiente, *Y*, provocada por las variaciones de las variables independientes, que se estima por medio de los coeficientes, b_j'. En la regresión múltiple, estos coeficientes dependen de qué otras variables se incluyan en el modelo. El coeficiente b_j indica la variación de *Y*, dada una variación unitaria de x_j, descontando al mismo tiempo el efecto simultáneo de las demás variables independientes.

En algunos problemas, ambos resultados son igual de importantes. Sin embargo, normalmente predomina uno de ellos (por ejemplo, la predicción de las ventas de las tiendas, *Y*, en el ejemplo de la localización de las tiendas).

La variación marginal es más difícil de estimar porque las variables independientes están relacionadas no sólo con las variables dependientes sino también entre sí. Si dos variables independientes o más varían en una relación lineal directa entre sí, es difícil averiguar el efecto que produce cada variable independiente en la variable dependiente.

Examinaremos detalladamente el modelo del ejemplo 13.2. El coeficiente de x_1 —es decir, b_1— indica la variación que experimentan las ventas del segundo año por cada variación unitaria del tamaño de la tienda. El coeficiente de x_5 indica la variación que experimentan las ventas por cada variación unitaria de la renta per cápita de la población que reside a menos de 8 kilómetros, mientras que la de x_7 indica la variación de las ventas por cada variación de la renta per cápita de la población que reside a menos de 15 kilómetros. Es probable, por supuesto, que las variables x_5 y x_7 estén correlacionadas. Por lo tanto, en la medida en que estas variables varíen ambas al mismo tiempo, es difícil averiguar la contribución de cada una de ellas a la variación de los ingresos generados por las ventas de las tiendas. Esta correlación entre variables independientes complica el modelo. Es importante comprender que el modelo predice los ingresos generados por las ventas de las tiendas utilizando la combinación de variables que contiene el modelo. El efecto de una variable de predicción es el efecto que produce esa variable cuando se combina con las demás. Por lo tanto, en general, el coeficiente de una variable no indica el efecto que produce esa variable en todas las condiciones. Estas complejidades se analizarán más detenidamente cuando se desarrolle el modelo de regresión múltiple.

Desarrollo del modelo

Cuando aplicamos la regresión múltiple, construimos un modelo para explicar la variabilidad de la variable dependiente. Para eso queremos incluir las influencias simultáneas e individuales de varias variables independientes. Supongamos, por ejemplo, que queremos desarrollar un modelo que prediga el margen anual de beneficios de las sociedades de ahorro y crédito inmobiliario utilizando los datos recogidos durante un periodo de años. Una especificación inicial del modelo indicaba que el margen anual de beneficios estaba relacionado con los ingresos netos por dólar depositado y el número de oficinas. Se espera que el ingreso neto aumente el margen anual de beneficios y se prevé que el número de oficinas

reducirá el margen anual de beneficios debido al aumento de la competencia. Eso nos llevaría a especificar un modelo de regresión poblacional

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

donde

Y = margen anual de beneficios
X_1 = ingresos anuales netos por dólar depositado
X_2 = número de oficinas existentes ese año

Savings and Loan

La Tabla 13.1 y el fichero de datos **Savings and Loan** contienen 25 observaciones por año de estas variables. Utilizaremos estos datos para desarrollar un modelo lineal que prediga el margen anual de beneficios en función de los ingresos por dólar depositado y del número de oficinas (véase la referencia bibliográfica 4).

Tabla 13.1. Datos de las asociaciones de ahorro y crédito inmobiliario.

Año	Ingresos por dólar	Número de oficinas	Margen de beneficios	Año	Ingresos por dólar	Número de oficinas	Margen de beneficios
1	3,92	7.298	0,75	14	3,78	6.672	0,84
2	3,61	6.855	0,71	15	3,82	6.890	0,79
3	3,32	6.636	0,66	16	3,97	7.115	0,7
4	3,07	6.506	0,61	17	4,07	7.327	0,68
5	3,06	6.450	0,7	18	4,25	7.546	0,72
6	3,11	6.402	0,72	19	4,41	7.931	0,55
7	3,21	6.368	0,77	20	4,49	8.097	0,63
8	3,26	6.340	0,74	21	4,70	8.468	0,56
9	3,42	6.349	0,9	22	4,58	8.717	0,41
10	3,42	6.352	0,82	23	4,69	8.991	0,51
11	3,45	6.361	0,75	24	4,71	9.179	0,47
12	3,58	6.369	0,77	25	4,78	9.318	0,32
13	3,66	6.546	0,78				

Pero antes de poder estimar el modelo, es necesario desarrollar y comprender el método de regresión múltiple. Para comenzar, examinemos el modelo general de regresión múltiple y observemos sus diferencias con el modelo de regresión simple. El modelo de regresión múltiple es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

donde ε_i es el término de error aleatorio que tiene la media 0 y la varianza σ^2, y las β_j' son los coeficientes o efectos marginales de las variables independientes o exógenas, x_j, donde $j = 1, ..., K$, dados los efectos de las demás variables independientes. Las i indican las observaciones, siendo $i = 1, ..., n$. Utilizamos las minúsculas x_{ji} para indicar los valores específicos de la variable X_j en la observación i. Suponemos que las ε_i son independientes de las X_j y entre sí para que las estimaciones de los coeficientes y sus varianzas sean correctas. En el Capítulo 14 explicamos qué ocurre cuando se abandonan estos supuestos.

El modelo muestral estimado es

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_Kx_{Ki} + e_i$$

donde e_i es el residuo o diferencia entre el valor observado de Y y el valor estimado de Y obtenido utilizando los coeficientes estimados, b_j, donde $j = 1, ..., K$. El método de regresión obtiene estimaciones simultáneas, b_j, de los coeficientes del modelo poblacional, β_j, utilizando el método de mínimos cuadrados.

En nuestro ejemplo de las asociaciones de ahorro y crédito inmobiliario, el modelo poblacional para los puntos de datos individuales es

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \varepsilon_i$$

Este modelo reducido con dos variables de predicción solamente brinda la oportunidad de comprender mejor el método de regresión. La función de regresión puede representarse gráficamente en tres dimensiones, como muestra la Figura 13.1. La función de regresión se representa mediante un plano en el que los valores de Y son una función de los valores de la variables independientes X_1 y X_2. Para cada par posible, x_{1i}, x_{2i}, el valor esperado de la variable dependiente, y_i, se encuentra en el plano. La Figura 13.2 ilustra específicamente el ejemplo de las asociaciones de ahorro y crédito inmobiliario. Un aumento de X_1 provoca un aumento del valor esperado de Y, condicionado al efecto de X_2. Asimismo, un aumento de X_2 provoca una disminución del valor esperado de Y, condicionada al efecto de X_1.

Para completar nuestro modelo, añadimos un término de error ε. Este término de error reconoce que no se cumplirá exactamente ninguna relación postulada y que es probable que haya otras variables que también afecten al valor observado de Y. Por lo tanto, cuando aplicamos el modelo, observamos el valor esperado de la variable dependiente, Y —representado por el plano en la Figura 13.2—, más un término de error aleatorio, ε, que representa la parte de Y no incluida en el valor esperado. Como consecuencia, el modelo de datos tiene la forma

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \cdots + \beta_Kx_{Ki} + \varepsilon_i$$

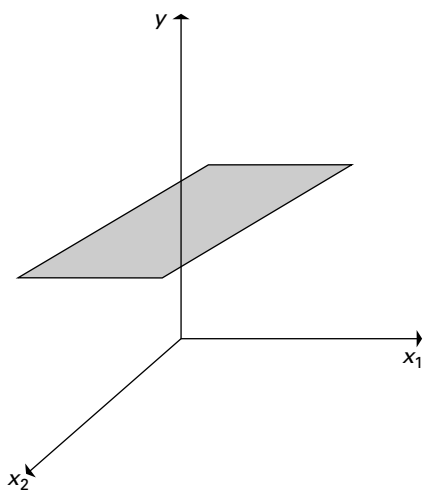


Figura 13.1. El plano es el valor esperado de Y en función de X_1 y X_2.

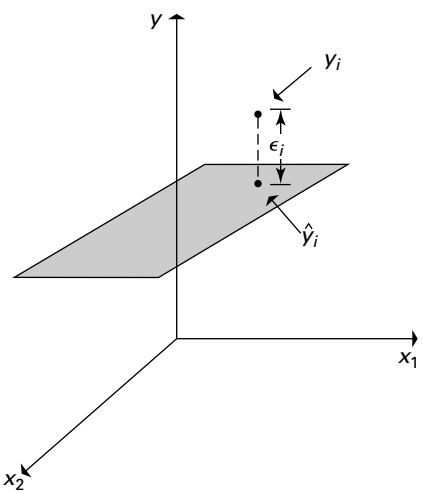


Figura 13.2. Comparación del valor observado y el esperado de Y en función de dos variables independientes.

El modelo de regresión poblacional múltiple

El **modelo de regresión poblacional múltiple** define la relación entre una variable dependiente o endógena, Y, y un conjunto de variables independientes o exógenas, x_j, donde $j = 1, ..., K$. Se supone que las x_{ji} son números fijos; Y es una variable aleatoria definida para cada observación, i, donde $i = 1, ..., n$, y n es el número de observaciones. El modelo se define de la forma siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \qquad \textbf{(13.1)}$$

donde las β_j son coeficientes constantes y las ε son variables aleatorias de 0 y varianza σ^2.

En el ejemplo de las asociaciones de ahorro y crédito inmobiliario, con dos variables independientes, el modelo de regresión poblacional es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Dados valores específicos de los ingresos netos, x_{1i}, y el número de oficinas, x_{2i}, el margen de beneficios observado, y_i, es la suma de dos partes: el valor esperado, $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, y el término de error aleatorio, ε_i. El término de error aleatorio puede concebirse como la combinación de los efectos de otros muchos factores sin identificar que afectan a los márgenes de beneficios. La Figura 13.2 ilustra el modelo; el plano indica el valor esperado de varias combinaciones de las variables independientes y la ε_i es la desviación entre el plano —el valor esperado— y el valor observado de Y —marcado con un punto grande— de un punto de dato específico. En general, los valores observados de Y no se encuentran en el plano sino por encima o por debajo de él, debido a los términos de error positivos o negativos, ε_i.

La regresión simple, presentada en el capítulo anterior, no es más que un caso especial de la regresión múltiple con una única variable de predicción y, por lo tanto, el plano se reduce a una línea. Así pues, la teoría y el análisis que hemos desarrollado para la regresión simple también se aplican a la regresión múltiple. Sin embargo, existen algunas interpretaciones más que desarrollaremos en nuestro estudio de la regresión múltiple. Una de ellas se ilustra en el siguiente análisis de los gráficos tridimensionales.

Gráficos tridimensionales

Tal vez sea más fácil comprender el método de regresión múltiple mediante una imagen gráfica simplificada. Observe el rincón de la habitación en la que está sentado. Las líneas formadas por las dos paredes y el suelo representan los ejes de dos variables independientes, X_1 y X_2. La esquina que forman las dos paredes es el eje de la variable dependiente, Y. Para estimar una recta de regresión, reunimos conjuntos de puntos (x_{1i}, x_{2i} e y_i).

Representemos ahora estos puntos en su habitación utilizando las esquinas de las paredes y el suelo como los tres ejes. Con estos puntos suspendidos en su habitación, buscamos un plano en el espacio que se aproxime a todos ellos. Este plano es la forma geométrica de la ecuación de mínimos cuadrados. Con estos puntos en el espacio, ahora subimos y bajamos un plano y lo hacemos girar en dos direcciones: todos estos movimientos los hacemos simultáneamente hasta que tenemos un plano que está «cerca» de todos los puntos. Recuérdese que en el Capítulo 12 hicimos esto con una línea recta en dos dimensiones para obtener una ecuación

$$\hat{y} = b_0 + b_1 x$$

A continuación, extendemos esa idea a tres dimensiones para obtener una ecuación

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

Este proceso es, por supuesto, más complicado que en el caso de la regresión simple. Pero los problemas reales son complicados y la regresión permite analizar mejor la complejidad de estos problemas. Queremos saber cómo varía Y cuando varía X_1. Pero sabemos que en estas variaciones influye, a su vez, la forma en que varía X_2. Y si X_1 y X_2 siempre varían a la vez, no podemos saber cuánto contribuye cada variable a las variaciones de Y.

Las interpretaciones geométricas de la regresión múltiple son cada vez más complejas a medida que aumenta el número de variables independientes. Sin embargo, la analogía con la regresión simple es extraordinariamente útil. Estimamos los coeficientes minimizando la suma de los cuadrados de las desviaciones de la dimensión Y en torno a una función lineal de las variables independientes. En la regresión simple, la función es una línea recta en un gráfico bidimensional. Con dos variables independientes, la función es un plano en un espacio tridimensional. Cuando consideramos más de dos variables independientes, tenemos varios hiperplanos complejos que son imposibles de visualizar.

EJERCICIOS

Ejercicios básicos

13.1. Dado el modelo lineal estimado

$$\hat{y} = 10 + 3x_1 + 2x_2 + 4x_3$$

a) Calcule $\hat{y}$ cuando $x_1 = 20$, $x_2 = 11$ y $x_3 = 10$.
b) Calcule $\hat{y}$ cuando $x_1 = 15$, $x_2 = 14$ y $x_3 = 20$.
c) Calcule $\hat{y}$ cuando $x_1 = 35$, $x_2 = 19$ y $x_3 = 25$.
d) Calcule $\hat{y}$ cuando $x_1 = 10$, $x_2 = 17$ y $x_3 = 30$.

13.2. Dado el modelo lineal estimado

$$\hat{y} = 10 + 5x_1 + 4x_2 + 2x_3$$

a) Calcule $\hat{y}$ cuando $x_1 = 20$, $x_2 = 11$ y $x_3 = 10$.
b) Calcule $\hat{y}$ cuando $x_1 = 15$, $x_2 = 14$ y $x_3 = 20$.
c) Calcule $\hat{y}$ cuando $x_1 = 35$, $x_2 = 19$ y $x_3 = 25$.
d) Calcule $\hat{y}$ cuando $x_1 = 10$, $x_2 = 17$ y $x_3 = 30$.

13.3. Dado el modelo lineal estimado

$$\hat{y} = 10 + 2x_1 + 12x_2 + 8x_3$$

a) Calcule $\hat{y}$ cuando $x_1 = 20$, $x_2 = 11$ y $x_3 = 10$.
b) Calcule $\hat{y}$ cuando $x_1 = 15$, $x_2 = 24$ y $x_3 = 20$.
c) Calcule $\hat{y}$ cuando $x_1 = 20$, $x_2 = 19$ y $x_3 = 25$.
d) Calcule $\hat{y}$ cuando $x_1 = 10$, $x_2 = 9$ y $x_3 = 30$.

13.4. Dado el modelo lineal estimado

$$\hat{y} = 10 + 2x_1 + 12x_2 + 8x_3$$

a) ¿Cuál es la variación de $\hat{y}$ cuando x_1 aumenta en 4?
b) ¿Cuál es la variación de $\hat{y}$ cuando x_3 aumenta en 1?
c) ¿Cuál es la variación de $\hat{y}$ cuando x_2 aumenta en 2?

13.5. Dado el modelo lineal estimado

$$\hat{y} = 10 - 2x_1 - 14x_2 + 6x_3$$

a) ¿Cuál es la variación de $\hat{y}$ cuando x_1 aumenta en 4?
b) ¿Cuál es la variación de $\hat{y}$ cuando x_3 disminuye en 1?
c) ¿Cuál es la variación de $\hat{y}$ cuando x_2 disminuye en 2?

Ejercicios aplicados

13.6. Una empresa aeronáutica quería predecir el número de horas de trabajo necesario para acabar el diseño de un nuevo avión. Se pensaba que las variables explicativas relevantes eran la velocidad máxima del avión, su peso y el número de piezas que tenía en común con otros modelos construidos por la empresa. Se tomó una muestra de 27 aviones de la empresa y se estimó el siguiente modelo:

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i} + \varepsilon_i$$

donde

y_i = esfuerzo de diseño en millones de horas de trabajo
x_{1i} = velocidad máxima del avión, en kilómetros por hora
x_{2i} = peso del avión, en toneladas

x_{3i} = número porcentual de piezas en común con otros modelos

Los coeficientes de regresión estimados eran

$$b_1 = 0{,}661 \qquad b_2 = 0{,}065 \qquad b_3 = -0{,}018$$

Interprete estas estimaciones.

13.7. En un estudio de la influencia de las instituciones financieras en los tipos de interés de los bonos alemanes, se analizaron datos trimestrales de un periodo de 12 años. El modelo postulado era

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

donde

y_i = variación de los tipos de interés de los bonos en el trimestre
x_{1i} = variación de las compras de bonos por parte de las instituciones financieras en el trimestre
x_{2i} = variación de las ventas de bonos por parte de las instituciones financieras en el trimestre

Los coeficientes de regresión parcial estimados eran

$$b_1 = 0{,}057 \qquad b_2 = -0{,}065$$

Interprete estas estimaciones.

13.8. Se ajustó el siguiente modelo a una muestra de 30 familias para explicar el consumo de leche por familia:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

donde

y_i = consumo de leche, en litros a la semana
x_1 = renta semanal en cientos de dólares
x_2 = tamaño de la familia

Las estimaciones de los parámetros de la regresión por mínimos cuadrados eran

$$b_0 = -0{,}025 \qquad b_1 = 0{,}052 \qquad b_2 = 1{,}14$$

a) Interprete las estimaciones b_1 y b_2.
b) ¿Es posible hacer una interpretación de la estimación b_0 que tenga sentido?

13.9. Se ajustó el siguiente modelo a una muestra de 25 estudiantes utilizando datos obtenidos al final de su primer año de universidad. El objetivo era explicar el aumento de peso de los estudiantes.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \varepsilon_i$$

donde

y_i = aumento de peso en kilos durante el primer año
x_{1i} = número medio de comidas a la semana
x_{2i} = número medio de horas de ejercicio a la semana
x_{3i} = número medio de cervezas consumidas a la semana

Las estimaciones de los parámetros de la regresión por mínimos cuadrados eran

$$b_0 = 7{,}35 \qquad b_1 = 0{,}653$$
$$b_2 = -1{,}345 \qquad b_3 = 0{,}613$$

a) Interprete las estimaciones b_1, b_2 y b_3.
b) ¿Es posible hacer una interpretación de la estimación b_0 que tenga sentido?

13.2. Estimación de coeficientes

Los coeficientes de regresión múltiple se calculan utilizando estimadores obtenidos mediante el método de mínimos cuadrados. Este método de mínimos cuadrados es similar al que presentamos en el Capítulo 12 para la regresión simple. Sin embargo, los estimadores son complicados debido a las relaciones entre las variables independientes X_j que ocurren simultáneamente con las relaciones entre las variables independientes y la variable dependiente. Por ejemplo, si dos variables independientes aumentan o disminuyen al mismo tiempo —correlación positiva o negativa— mientras que al mismo tiempo la variable dependiente aumenta o disminuye, no podemos saber qué variable independiente está relacionada realmente con la variación de la variable dependiente. Como consecuencia, observamos que los coeficientes de regresión estimados son menos fiables si hay estrechas correlaciones entre dos variables independientes o más. Las estimaciones de los coeficientes y sus varianzas siempre se obtienen por computador. Sin embargo, dedicaremos bastantes esfuerzos a estudiar el álgebra y las formas de calcular la regresión por mínimos cuadrados. Estos esfuerzos permitirán comprender el método y averiguar cómo influyen las diferentes pautas de los datos en los resultados. Comenzamos con los supuestos habituales del modelo de regresión múltiple.

Supuestos habituales de la regresión múltiple

El modelo de regresión poblacional múltiple es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

y suponemos que se dispone de n conjuntos de observaciones. Se postulan los siguientes supuestos habituales para el modelo.

1. Las x_{ji} son o bien números fijos, o bien realizaciones de variables aleatorias, X_j, que son independientes de los términos de error, ε_i. En el segundo caso, la inferencia se realiza condicionada a los valores observados de las x_{ji}.
2. El valor esperado de la variable aleatoria Y es una función lineal de las variables independientes X_j.
3. Los términos de error son variables aleatorias cuya media es 0 y que tienen la misma varianza, σ^2. Este último supuesto se denomina homocedasticidad o varianza uniforme.

$$E[\varepsilon_i] = 0 \quad \text{y} \quad E[\varepsilon_i^2] = \sigma^2 \text{ para } (i = 1, ..., n)$$

4. Los términos de error aleatorios, ε_i, no están correlacionados entre sí, por lo que

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{para todo } i = j$$

5. No es posible hallar un conjunto de números que no sean iguales a cero, c_0, c_1, ..., c_K, tal que

$$c_0 + c_1 x_{1i} + c_2 x_{2i} + \cdots + c_K x_{Ki} = 0$$

   Ésta es la propiedad de la ausencia de relación lineal entre las X_j.

Los cuatro primeros supuestos son esencialmente iguales que los que postulamos en el caso de la regresión simple. Sin embargo, el supuesto 5 excluye algunos casos en los que existen relaciones lineales entre las variables de predicción. Supongamos, por ejemplo, que tenemos interés en explicar la variabilidad de las tarifas que se cobran por el envío de maíz. Una variable explicativa evidente sería la distancia a la que se envía el maíz. La distancia podría medirse en diferentes unidades como millas o kilómetros. Pero no tendría sentido utilizar como variables de predicción tanto la distancia en millas como la distancia en kilómetros. Estas dos medidas son funciones lineales una de la otra y no satisfarían el supuesto 5. Además, sería una tontería tratar de evaluar sus efectos independientes. Como veremos, las ecuaciones para calcular las estimaciones de los coeficientes y los programas informáticos no funcionan si no se satisface el supuesto 5. En la mayoría de los casos, la especificación adecuada del modelo evitará que se viole ese supuesto.

Método de mínimos cuadrados

El método de mínimos cuadrados para la regresión múltiple calcula los coeficientes estimados para minimizar la suma de los cuadrados de los residuos. Recuérdese que el residuo es

$$e_i = y_i - \hat{y}_i$$

donde y_i es el valor observado de Y e $\hat{y}_i$ es el valor de Y predicho a partir de la regresión. En términos formales, minimizamos *SCE*:

$$\begin{aligned} SCE &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki}))^2 \end{aligned}$$

Esta minimización consiste en hallar el plano que mejor represente un conjunto de puntos en el espacio, como hemos visto en nuestro análisis de los gráficos tridimensionales. Para realizar el proceso formalmente, utilizamos derivadas parciales para desarrollar un conjunto de ecuaciones normales simultáneas que se resuelve para obtener los estimadores de los coeficientes. Para los que tengan buenos conocimientos de matemáticas, en el apéndice del capítulo presentamos algunos de los detalles del proceso. Sin embargo, se pueden extraer importantes conclusiones dándose cuenta de que queremos encontrar la ecuación que mejor represente los datos observados. Afortunadamente, en las aplicaciones estudiadas en este libro, los complejos cálculos siempre se realizan utilizando un paquete estadístico como Minitab, SAS o SPSS. Nuestro objetivo es comprender cómo se interpretan los resultados de las regresiones y utilizarlos para resolver problemas. Lo haremos examinando algunos de los resultados algebraicos intermedios para ayudar a comprender los efectos que producen distintas pautas de datos en los estimadores de los coeficientes.

Estimación por mínimos cuadrados y regresión muestral múltiple

Comenzamos con una muestra de n observaciones (x_{1i}, x_{2i}, ..., x_{Ki}, y_i, donde $i = 1, ..., n$) medidas para un proceso cuyo modelo de regresión poblacional múltiple es

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Las estimaciones por mínimos cuadrados de los coeficientes $\beta_1, \beta_2, ..., \beta_K$ son los valores $b_0, b_1, ..., b_K$ para los que la suma de los cuadrados de las desviaciones

$$SCE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_K x_{Ki})^2 \qquad \textbf{(13.2)}$$

es la menor posible.

La ecuación resultante

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki} \qquad \textbf{(13.3)}$$

es la regresión muestral múltiple de Y con respecto a $X_1, X_2, ..., X_K$.

Consideremos de nuevo el modelo de regresión con dos variables de predicción solamente.

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

Los estimadores de los coeficientes pueden resolverse utilizando las formas siguientes:

$$b_1 = \frac{s_y(r_{x_1y} - r_{x_1x_2}r_{x_2y})}{s_{x_1}(1 - r^2_{x_1x_2})} \tag{13.4}$$

$$b_2 = \frac{s_y(r_{x_2y} - r_{x_1x_2}r_{x_1y})}{s_{x_2}(1 - r^2_{x_1x_2})} \tag{13.5}$$

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \tag{13.6}$$

donde

r_{x_1y} = correlación muestral entre X_1 y Y
r_{x_2y} = correlación muestral entre X_2 e Y
$r_{x_1x_2}$ = correlación muestral entre X_1 y X_2
s_{x_1} = desviación típica muestral de X_1
s_{x_2} = desviación típica muestral de X_2
s_y = desviación típica muestral de Y

En las ecuaciones de los estimadores de los coeficientes, vemos que la estimación del coeficiente de la pendiente, b_1, no sólo depende de la correlación entre Y y X_1 sino que también la afecta la correlación entre X_1 y X_2 y la correlación entre X_2 e Y. Si la correlación entre X_1 y X_2 es igual a 0, los estimadores de los coeficientes, b_1 y b_2, serán iguales que los estimadores de los coeficientes que se obtendrían en las regresiones simples correspondientes: debemos señalar que esto raras veces ocurre en el análisis empresarial y económico. Y a la inversa, si la correlación entre las variables independientes es igual a 1, los estimadores de los coeficientes serán indefinidos, pero eso se deberá únicamente a que la especificación del modelo es incorrecta y violará el supuesto 5 de la regresión múltiple. Si las variables independientes están correlacionadas perfectamente, ambas experimentan variaciones relativas simultáneas. Vemos que en ese caso no es posible saber qué variable predice la variación de Y. En el ejemplo 13.3 vemos el efecto de las correlaciones entre las variables independientes examinando el problema de las asociaciones de ahorro y crédito inmobiliario, cuyos datos se muestran en la Tabla 13.1.

EJEMPLO 13.3. Márgenes de beneficios de las asociaciones de ahorro y crédito inmobiliario (estimación de los coeficientes de regresión)

El presidente de la confederación de asociaciones de ahorro y crédito inmobiliario le ha pedido que identifique las variables que afectan al margen porcentual de beneficios.

Savings and Loan

Solución

En primer lugar, desarrollamos una especificación del modelo de regresión múltiple que predice los beneficios como una función lineal del porcentaje de ingresos netos por dólar depositado y el número de oficinas. Utilizando los datos de la Tabla 13.1 que se encuentran en el fichero de datos **Savings and Loan**, hemos estimado un modelo de regresión múltiple, que se observa en las salidas Minitab y Excel de la Figura 13.3.

Los coeficientes estimados se identifican en la salida de los programas informáticos. Vemos que cada aumento unitario de los ingresos, X_1, provoca un aumento de los beneficios porcentuales de 0,237 —si la otra variable no varía— y un aumento unitario del

Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices

Predictor          Coef       SE Coef      T      P
Constant        1.56450       0.07940   19.70  0.000
X1 revenue      0.23720       0.05556    4.27  0.000
X2 offices   -0.00024908   0.00003205   -7.77  0.000

S = 0.0533022  R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source            DF       SS        MS      F      P
Regression         2  0.40151   0.20076  70.66  0.000
Residual Error    22  0.06250   0.00284
Total             24  0.46402
```

Coeficientes de regresión b_0, b_1, b_2

RESUMEN						
Estadísticos de la regresión						
Coeficiente de correlación múltiple	0,930212915					
Coeficiente de determinación R	0,865296068					
R ajustado	0,853050256					
Error típico	0,053302217					
Observaciones	25					
ANÁLISIS DE VARIANZA						
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
Regresión	2	0,40151122	0,20075561	70,66057082	2,64962E-10	
Residuos	22	0,06250478	0,002841126			
Total	24	0,464016				
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	1,564496771	0,079395981	19,70498685	1,81733E-15	1,399839407	1,72915414
X1 ingresos	0,237197475	0,055559366	4,269261695	0,000312567	0,121974278	0,35242067
X2 número de oficinas	–0,000249079	3,20485E-05	–7,771949195	9,50879E-08	–0,000315544	–0,00018261

Coeficientes de regresión b_0, b_1, b_2

Figura 13.3. Ecuación de regresión de los beneficios de las asociaciones de ahorro y crédito inmobiliario: (a) salida Minitab; (b) salida Excel.

número de oficinas reduce los beneficios en 0,000249. Consideremos ahora los dos modelos de regresión simple de las Figuras 13.4 y 13.5, en los que hacemos una regresión de Y con respecto a cada variable independiente por separado. Consideremos primero la regresión de Y con respecto a los ingresos, X_1, de la Figura 13.4. En esta regresión simple, el coeficiente de X_1 es $-0{,}169$, que es claramente diferente de $+0{,}237$ de la regresión múltiple. Más abajo, vemos que la correlación entre X_1 y X_2 es 0,941. Esta gran correlación produce un gran efecto en el coeficiente de X_1 en la ecuación de regresión múltiple.

Consideremos a continuación la regresión de Y con respecto a X_2 solamente de la Figura 13.5. En esta regresión simple, el coeficiente de la pendiente del número de ofi-

Regression Analysis: Y profit versus X1 revenue

The regression equation is
Y profit = 1.33 − 0.169 X1 revenue

Predictor	Coef	SE Coef	T	P
Constant	1.3262	0.1386	9.57	0.000
X1 revenue	−0.16913	0.03559	−4.75	0.000

Coeficiente de regresión b_1

S = 0.100891 R-Sq = 49.5% R-Sq(adj) = 47.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.22990	0.22990	22.59	0.000
Residual Error	23	0.23412	0.01018		
Total	24	0.46402			

Figura 13.4. Regresión de los beneficios de las asociaciones de ahorro y crédito inmobiliario con respecto a los ingresos.

Regression Analysis: Y profit versus X2 revenue

The regression equation is
Y profit = 1.55 − 0.000120 X2 offices

Predictor	Coef	SE Coef	T	P
Constant	1.5460	0.1048	14.75	0.000
X2 offices	−0.00012033	0.00001434	−8.39	0.000

Coeficiente de regresión b_2

S = 0.0704917 R-Sq = 75.4% R-Sq(adj) = 74.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.34973	0.34973	70.38	0.000
Residual Error	23	0.11429	0.00497		
Total	24	0.46402			

Figura 13.5. Regresión de los beneficios de las asociaciones de ahorro y crédito inmobiliario con respecto al número de oficinas.

cinas, X_2, es $-0{,}000120$, mientras que en la regresión múltiple es $-0{,}000249$. Este cambio de los coeficientes, aunque no es tan grande como en el caso del coeficiente de X_1, también se debe a la estrecha correlación entre las variables independientes.

Las correlaciones entre las tres variables son

	Y Beneficios	*X*1 Ingresos
X1 Ingresos	−0,704	
X2 Oficinas	−0,868	0,941

INTERPRETACIÓN

Vemos que la correlación entre X_1 y X_2 es 0,941. Por lo tanto, las dos variables tienden a variar a la vez y no es sorprendente que los coeficientes de la regresión múltiple sean diferentes de los coeficientes de la regresión simple. Debemos señalar que los coeficientes de la regresión múltiple son *coeficientes condicionados*; es decir, el coeficiente estimado

b_1 depende de las demás variables incluidas en el modelo. Eso siempre es así en la regresión múltiple, a menos que dos variables independientes tengan una correlación muestral de cero, algo que es muy improbable.

Estas relaciones también pueden estudiarse utilizando un «gráfico matricial» de Minitab, como el que muestra la Figura 13.6. No existen gráficos de este tipo en Excel. Obsérvese que la relación simple entre Y y X_2 es claramente lineal, mientras que la relación simple entre Y y X_1 es algo curvilínea. Esta relación no lineal entre X_1 e Y explica en parte por qué el coeficiente de X_1 de la regresión simple es tan distinto del de la regresión múltiple. Vemos en este ejemplo que las correlaciones entre variables independientes pueden influir considerablemente en los coeficientes estimados. Por lo tanto, si es posible elegir, deben evitarse las variables independientes muy correlacionadas. Pero en muchos casos no es posible elegir. Las estimaciones de los coeficientes de regresión siempre dependen de las demás variables de predicción del modelo. En este ejemplo, los beneficios aumentan en función de los ingresos porcentuales por dólar depositado. Sin embargo, el aumento simultáneo del número de oficinas —que redujo los beneficios— ocultaría el aumento de los beneficios si se utilizara un análisis de regresión simple. Por lo tanto, es muy importante especificar correctamente el modelo, es decir, la elección de las variables de predicción. Para especificar el modelo es necesario comprender el contexto del problema y la teoría.

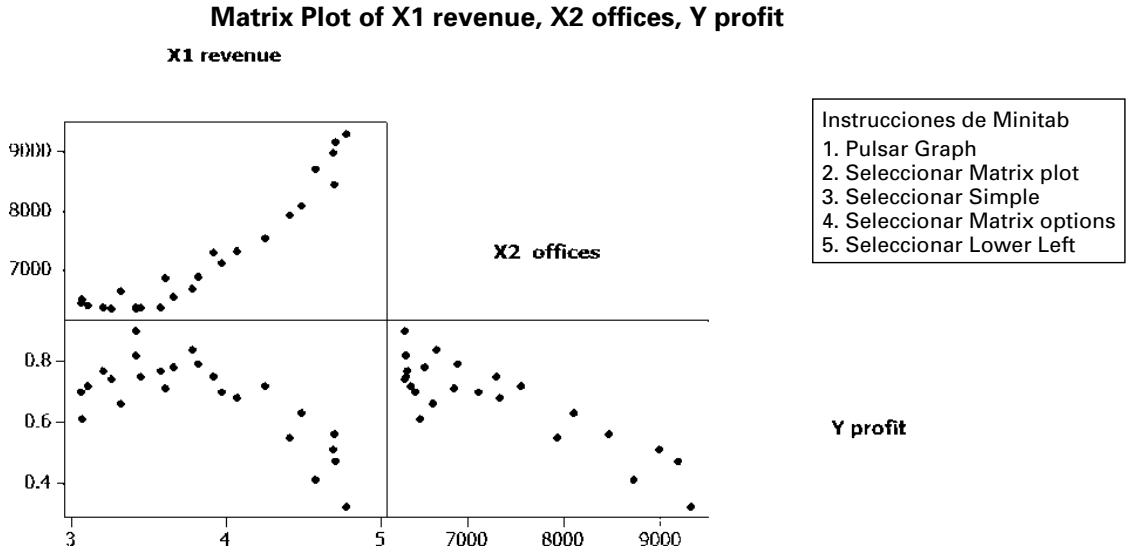


Figura 13.6. Gráficos matriciales de las variables de las asociaciones de ahorro y crédito inmobiliario.

EJERCICIOS

Ejercicios básicos

13.10. Calcule los coeficientes b_1 y b_2 del modelo de regresión

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

dados los siguientes estadísticos sintéticos:

a) $r_{x_1y} = 0{,}60$; $r_{x_2y} = 0{,}70$; $r_{x_1x_2} = 0{,}50$; $s_{x_1} = 200$; $s_{x_2} = 100$; $s_y = 400$

b) $r_{x_1y} = -0{,}60$; $r_{x_2y} = 0{,}70$; $r_{x_1x_2} = -0{,}50$; $s_{x_1} = 200$; $s_{x_2} = 100$; $s_y = 400$

c) $r_{x_1y} = 0{,}40$; $r_{x_2y} = 0{,}450$; $r_{x_1x_2} = 0{,}80$; $s_{x_1} = 200$; $s_{x_2} = 100$; $s_y = 400$

d) $r_{x_1y} = 0{,}60$; $r_{x_2y} = -0{,}50$; $r_{x_1x_2} = -0{,}60$; $s_{x_1} = 200$; $s_{x_2} = 100$; $s_y = 400$

Ejercicios aplicados

13.11. Considere las ecuaciones de regresión lineal estimadas

$$Y = a_0 + a_1X_1$$

$$Y = b_0 + b_1X_1 + b_2X_2$$

a) Muestre detalladamente los estimadores de los coeficientes de a_1 y b_1 cuando la correlación entre X_1 y X_2 es igual a 0.
b) Muestre detalladamente los estimadores de los coeficientes de a_1 y b_1 cuando la correlación entre X_1 y X_2 es igual a 1.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

13.12. Amalgamated Power Inc. le ha pedido que estime una ecuación de regresión para averiguar cómo afectan algunas variables de predicción a la demanda de ventas de electricidad. Realiza una serie de estimaciones de regresión y analiza sus resultados utilizando los datos trimestrales de las ventas de electricidad de los 17 últimos años que se encuentran en el fichero de datos **Power Demand**.

a) Estime una ecuación de regresión utilizando las ventas de electricidad como variable dependiente y el número de clientes y el precio como variables de predicción. Interprete los coeficientes.
b) Estime una ecuación de regresión (ventas de electricidad) utilizando solamente el número de clientes como variable de predicción. Interprete el coeficiente y compare el resultado con el del apartado (a).
c) Estime una ecuación de regresión (ventas de electricidad) utilizando el precio y los grados-días como variables de predicción. Interprete los coeficientes. Compare el coeficiente del precio con el que ha obtenido en el apartado (a).
d) Estime una ecuación de regresión (ventas de electricidad) utilizando la renta y los grados-días como variables de predicción. Interprete los coeficientes.

13.13. Transportation Research Inc. le ha pedido que formule algunas ecuaciones de regresión múltiple para estimar el efecto de algunas variables en el ahorro de combustible. Los datos para este estudio se encuentran en el fichero de datos **Motors** y la variable dependiente son las millas por galón —milpgal— conforme a la certificación del Departamento de Transporte.

a) Formule una ecuación de regresión que utilice la potencia de los vehículos — horsepower— y el peso de los vehículos —weight— como variables independientes. Interprete los coeficientes.
b) Formule una segunda ecuación de regresión que añada el número de cilindros —cylinder— como variable independiente a la ecuación del apartado (a). Interprete los coeficientes.
c) Formule una ecuación de regresión que utilice el número de cilindros y el peso del vehículo como variables independientes. Interprete los coeficientes y compare los resultados con los de los apartados (a) y (b).
d) Formule una ecuación de regresión que utilice la potencia de los vehículos, el peso de los vehículos y el precio como variables de predicción. Interprete los coeficientes.
e) Escriba un breve informe que resuma sus resultados.

13.14. Transportation Research Inc. le ha pedido que formule algunas ecuaciones de regresión múltiple para estimar el efecto de algunas variables en la potencia de los vehículos. Los datos para este estudio se encuentran en el fichero de datos **Motors** y la variable dependiente es la potencia —horsepower— conforme a la certificación del Departamento de Transporte.

a) Formule una ecuación de regresión que utilice el peso de los vehículos —weight— y las pulgadas cúbicas de desplazamiento de los cilindros —displacement— como variables de predicción. Interprete los coeficientes.
b) Formule una ecuación de regresión que utilice el peso de los vehículos, el desplazamiento de los cilindros y el número de cilindros —cylinder— como variables de predicción. Interprete los coeficientes y compare los resultados con los del apartado (a).
c) Formule una ecuación de regresión que utilice el peso de los vehículos, el desplazamiento de los cilindros y las millas por galón —milpgal— como variables de predicción. Interprete los coeficientes y compare los resultados con los del apartado (a).
d) Formule una ecuación de regresión que utilice el peso de los vehículos, el desplazamiento de los cilindros, las millas por galón y el precio como variables de predicción. Interprete los coeficientes y compare los resultados con los del apartado (c).
e) Escriba un breve informe que presente los resultados de su análisis de este problema.

13.3. Poder explicativo de una ecuación de regresión múltiple

La regresión múltiple utiliza variables independientes para explicar la conducta de la variable dependiente. Observamos que la variabilidad de la variable dependiente puede explicarse en parte mediante la función lineal de las variables independientes. En este apartado desarrollamos una medida de la proporción de la variabilidad de la variable dependiente que puede explicarse por medio del modelo de regresión múltiple.

El modelo de regresión estimado a partir de la muestra es

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_Kx_{Ki} + e_i$$

También podríamos expresarlo de la siguiente manera:

$$y_i = \hat{y}_i + e_i$$

donde

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_Kx_{Ki}$$

es el valor predicho de la variable dependiente y el residuo, e_i, es la diferencia entre el valor observado y el predicho. La Tabla 13.2 contiene en las tres primeras columnas estas cantidades correspondientes al ejemplo de las asociaciones de ahorro y crédito inmobiliario.

Tabla 13.2. Valores efectivos, valores predichos y residuos en la regresión de las asociaciones de ahorro y crédito inmobiliario.

y_I	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	$y_i - \bar{y}$	$\hat{y}_i - \bar{y}$
0,75	0,677	0,073	0,076	0,003
0,71	0,713	−0,003	0,036	0,039
0,66	0,699	−0,039	−0,014	0,025
0,61	0,672	−0,062	−0,064	−0,002
0,7	0,684	0,016	0,026	0,010
0,72	0,708	0,012	0,046	0,034
0,77	0,740	0,030	0,096	0,066
0,74	0,759	−0,019	0,066	0,085
0,9	0,794	0,106	0,226	0,120
0,82	0,794	0,026	0,146	0,120
0,75	0,798	−0,048	0,076	0,124
0,77	0,827	−0,057	0,096	0,153
0,78	0,802	−0,022	0,106	0,128
0,84	0,799	0,041	0,166	0,125
0,79	0,754	0,036	0,116	0,080
0,7	0,734	−0,034	0,026	0,060
0,68	0,705	−0,025	0,006	0,031
0,72	0,693	0,027	0,046	0,019
0,55	0,635	−0,085	−0,124	−0,039
0,63	0,613	0,017	−0,044	−0,061
0,56	0,570	−0,010	−0,114	−0,104
0,41	0,480	−0,070	−0,264	−0,194
0,51	0,437	0,073	−0,164	−0,237
0,47	0,395	0,075	−0,204	−0,279
0,32	0,377	−0,057	−0,354	−0,297
Suma de los cuadrados:		0,0625 (*SCE*)	0,4640 (*STC*)	0,4015 (*SCR*)

Restando la media muestral de la variable dependiente de ambos miembros, tenemos que

$$\begin{aligned}(y_i - \bar{y}) &= (\hat{y}_i - \bar{y}) + e_i \\ &= (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\end{aligned}$$

que puede expresarse de la siguiente manera:

$$\begin{array}{l}\text{Desviación observada} \\ \text{con respecto a la media muestral}\end{array} = \begin{array}{l}\text{desviación predicha con} \\ \text{respecto a la media muestral}\end{array} + \text{residuo}$$

A continuación, elevando al cuadrado los dos miembros y sumando con respecto al índice, i, tenemos que

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (\hat{y}_1 - \bar{y})^2 + (y_i - \hat{y}_1)^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2\end{aligned}$$

que es la descomposición de la suma de los cuadrados presentada en el Capítulo 12.

$$STC = SCR + SCE$$

$$\text{Suma total de los cuadrados} = \begin{array}{l}\text{suma de los cuadrados} \\ \text{de la regresión}\end{array} + \begin{array}{l}\text{suma de los cuadrados} \\ \text{de los errores}\end{array}$$

Esta descomposición simplificada se debe a que y e $\hat{y}$ son independientes y, por lo tanto,

$$\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

Descomposición de la suma de los cuadrados y coeficiente de determinación

Comenzamos con el modelo de regresión múltiple ajustado mediante mínimos cuadrados

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki} + e_i = \hat{y}_i + e_i$$

donde las b_j son las estimaciones por mínimos cuadrados de los coeficientes del modelo de regresión poblacional y las e son los residuos del modelo de regresión estimado.

La variabilidad del modelo puede dividirse en los componentes

$$STC = SCR + SCE \tag{13.7}$$

donde estos componentes se definen de la forma siguiente.

Suma total de los cuadrados:

$$STC = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{13.8}$$

$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{13.9}$$

Suma de los cuadrados de los errores:

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{13.10}$$

Suma de los cuadrados de la regresión:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{13.11}$$

Esta descomposición puede interpretarse de la forma siguiente:

Variabilidad muestral total = variabilidad explicada + variabilidad no explicada

El coeficiente de determinación, R^2, de la regresión ajustada es la proporción de la variabilidad muestral total explicada por la regresión

$$R^2 = \frac{SCR}{STC} = 1 - \frac{SCE}{STC} \tag{13.12}$$

y se deduce que

$$0 \leqslant R^2 \leqslant 1$$

La suma de los cuadrados de los errores también se utiliza para calcular la estimación de la varianza de los errores del modelo poblacional, como muestra la ecuación 13.13. Al igual que ocurre en la regresión simple, la varianza de los errores poblacionales se utiliza para la inferencia estadística de la regresión múltiple.

Estimación de la varianza de los errores

Dado el modelo de regresión poblacional múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

y los supuestos habituales de la regresión, sea σ^2 la varianza común del término de error, ε_i. Entonces, una estimación insesgada de esa varianza es

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - K - 1} = \frac{SCE}{n - K - 1} \tag{13.13}$$

donde K es el número de variables independientes en el modelo de regresión. La raíz cuadrada de la varianza, s_e, también se llama **error típico de la estimación**.

Llegados a este punto, también podemos calcular el cuadrado medio de la regresión de la forma siguiente:

$$CMR = \frac{SCR}{K}$$

Utilizamos el *CMR* como medida de la variabilidad explicada ajustada para tener en cuenta el número de variables independientes.

La media muestral de la variable dependiente de los beneficios de las asociaciones de ahorro y crédito inmobiliario es $\bar{y} = 0{,}674$, y hemos utilizado este valor para calcular las dos últimas columnas de la Tabla 13.2. Utilizando los datos de esta tabla y los componentes, podemos demostrar que

$$SCE = 0{,}0625 \qquad STC = 0{,}4640 \qquad R^2 = 0{,}87$$

En estos resultados, vemos que en esta muestra el 87 por ciento de la variabilidad de los beneficios de las asociaciones de ahorro y crédito inmobiliario es explicado por las relaciones lineales con los ingresos netos y el número de oficinas. Obsérvese que también podríamos calcular la suma de los cuadrados de la regresión a partir de la identidad

$$SCR = STC - SCE = 0{,}4640 - 0{,}0625 = 0{,}4015$$

También podemos calcular una estimación de la varianza de los errores σ^2 utilizando la ecuación 13.13:

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - K - 1} = \frac{SCE}{n - K - 1} = \frac{0{,}0625}{25 - 2 - 1} = 0{,}0284$$

La Figura 13.7 presenta la salida Minitab y Excel del análisis de regresión correspondiente al problema de las asociaciones de ahorro y crédito inmobiliario e indica las distintas sumas de los cuadrados calculadas. Los paquetes estadísticos calculan habitualmente estas cantidades; incluimos los detalles de la Tabla 13.2 únicamente para indicar cómo se calculan las sumas de los cuadrados. A partir de ahora, suponemos que las sumas de los cuadrados se calculan mediante un paquete estadístico.

Los componentes de la variabilidad tienen sus correspondientes grados de libertad. La cantidad *STC* tiene $n - 1$ grados de libertad porque se necesita la media de Y para calcularla. El componente *SCR* tiene K grados de libertad porque los coeficientes K se necesitan para calcularla. Por último, el componente *SCE* tiene $n - K - 1$ grados de libertad porque se necesitan los K coeficientes y la media para calcularla. Obsérvese que en la Figura 13.7 se incluyen los grados de libertad (*DF*) correspondientes a cada componente.

INTERPRETACIÓN

Utilizamos el coeficiente de determinación, R^2, habitualmente como estadístico descriptivo para describir la fuerza de la relación lineal entre las variables independientes X y la variable dependiente, Y. Es importante hacer hincapié en que R^2 sólo puede utilizarse para comparar modelos de regresión que tienen el mismo conjunto de observaciones muestrales de y_i, siendo $i = 1, ..., n$. Este resultado se observa en la forma de la ecuación

$$R^2 = 1 - \frac{SCE}{STC}$$

Vemos, pues, que el valor de R^2 puede ser alto bien porque *SCE* es pequeña —lo que indica que los puntos observados están cerca de los puntos predichos—, bien porque *STC* es grande. Hemos visto que *SCE* y s_e^2 indican la cercanía de los puntos observados a los puntos predichos. Cuando dos o más ecuaciones de regresión tienen la misma *STC*, R^2 es una medida comparable de la bondad del ajuste de las ecuaciones.

La utilización de R^2 como medida global de la calidad de una ecuación ajustada puede plantear un problema. Cuando se añaden variables independientes a un modelo de regre-

Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices

Predictor          Coef       SE Coef       T       P
Constant        1.56450       0.07940   19.70   0.000
X1 revenue      0.23720       0.05556    4.27   0.000
X2 offices   -0.00024908    0.00003205   -7.77   0.000

S = 0.0533022   R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source           DF       SS        MS       F       P
Regression        2  0.40151   0.20076   70.66   0.000
Residual Error   22  0.06250   0.00284
Total            24  0.46402

Source        DF      Seq SS
X1 revenus     1     0.22990
X2 offices     1     0.17161
```

Coeficientes b_0, b_1, b_2

Error típico de la estimación s_e

Coeficiente de determinación R^2

$CMR = SCR/K$

Varianza de s_e^2

SCR = 0,40151
SCE = 0,06250
STC = 0,46402

Número de variables independientes (X) = K

Error típico de la estimación s_e

RESUMEN						
Estadísticos de la regresión						
Coeficiente de correlación múltiple	0,930212915					
Coeficiente de determinación R	0,865296068					
R ajustado	0,853050256					
Error típico	0,053302217					
Observaciones	25					
ANÁLISIS DE VARIANZA						
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
Regresión	2	0,40151122	0,20075561	70,66057082	2,64962E-10	
Residuos	22	0,06250478	0,002841126			
Total	24	0,464016				
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	1,564496771	0,079395981	19,70498685	1,81733E-15	1,399839407	1,72915414
X1 ingresos	0,237197475	0,055559366	4,269261695	0,000312567	0,121974278	0,35242067
X2 número de oficinas	–0,000249079	3,20485E-05	–7,771949195	9,50879E-08	–0,000315544	–0,00018261

Coeficiente de determinación R^2

Número de variables independientes (X) = K

$CMR = \frac{SCR}{K}$

Coeficientes b_0, b_1, b_2

SCR = 0,40151
SCE = 0,06250
STC = 0,46402

Figura 13.7. Salida Minitab (a) y salida Excel (b) del análisis de regresión correspondiente al problema de las asociaciones de ahorro y crédito inmobiliario.

sión múltiple —en casi todas las situaciones aplicadas—, la suma explicada de los cuadrados, *SCR*, aumenta aunque la variable independiente adicional no sea una variable de predicción importante. Por lo tanto, podríamos encontrarnos con que R^2 ha aumentado espuriamente después de que se ha añadido una o más variables de predicción poco importantes al modelo de regresión múltiple. En ese caso, el aumento del valor de R^2 sería engañoso. Para evitar este problema, el coeficiente de determinación ajustado puede calcularse como muestra la ecuación 13.14.

Coeficiente de determinación ajustado

El **coeficiente de determinación ajustado**, $\bar{R}^2$, se define de la forma siguiente:

$$\bar{R}^2 = 1 - \frac{SCE/(n - K - 1)}{STC/(n - 1)} \qquad \textbf{(13.14)}$$

Utilizamos esta medida para tener en cuenta el hecho de que las variables independientes irrelevantes provocan una pequeña reducción de la suma de los cuadrados de los errores. Por lo tanto, el $\bar{R}^2$ ajustado permite comparar mejor los modelos de regresión múltiple que tienen diferentes números de variables independientes.

Volviendo a nuestro ejemplo de las asociaciones de ahorro y crédito inmobiliario, vemos que

$$n = 25 \qquad K = 2 \qquad SCE = 0{,}0625 \qquad STC = 0{,}4640$$

y, por lo tanto, el coeficiente ajustado de determinación es

$$\bar{R}^2 = 1 - \frac{0{,}0625/22}{0{,}4640/24} = 0{,}853$$

En este ejemplo, la diferencia entre R^2 y $\bar{R}^2$ no es muy grande. Sin embargo, si el modelo de regresión hubiera contenido algunas variables independientes que no fueran importantes predictores condicionados, la diferencia sería grande. Otra medida de la relación en la regresión múltiple es el coeficiente de correlación múltiple.

Coeficiente de correlación múltiple

El **coeficiente de correlación múltiple** es la correlación entre el valor predicho y el valor observado de la variable dependiente

$$R = r(\hat{y}, y) = \sqrt{R^2} \qquad \textbf{(13.15)}$$

y es igual a la raíz cuadrada del coeficiente múltiple de determinación. Utilizamos *R* como otra medida de la fuerza de la relación entre la variable dependiente y las variables independientes. Por lo tanto, es comparable a la correlación entre *Y* y *X* en la regresión simple.

EJERCICIOS

Ejercicios básicos

13.15. Un análisis de regresión ha producido la siguiente tabla del análisis de la varianza:

Analysis of Variance

Source	DF	SS	MS
Regression	3	4500	
Residual Error	26	500	

a) Calcule s_e y s_e^2.
b) Calcule *STC*.
c) Calcule R^2 y el coeficiente ajustado de determinación.

13.16. Un análisis de regresión ha producido la siguiente tabla del análisis de la varianza:

Analysis of Variance

Source	DF	SS	MS
Regression	2	7000	
Residual Error	29	2500	

a) Calcule s_e y s_e^2.
b) Calcule *STC*.
c) Calcule R^2 y el coeficiente ajustado de determinación.

13.17. Un análisis de regresión ha producido la siguiente tabla del análisis de la varianza:

Analysis of Variance

Source	DF	SS	MS
Regression	4	40000	
Residual Error	45	10000	

a) Calcule s_e y s_e^2.
b) Calcule *STC*.
c) Calcule R^2 y el coeficiente ajustado de determinación.

13.18. Un análisis de regresión ha producido la siguiente tabla del análisis de la varianza:

Analysis of Variance

Source	DF	SS	MS
Regression	5	80000	
Residual Error	200	15000	

a) Calcule s_e y s_e^2.
b) Calcule *STC*.
c) Calcule R^2 y el coeficiente ajustado de determinación.

Ejercicios aplicados

13.19. En el estudio del ejercicio 13.6, en el que las estimaciones por mínimos cuadrados se basaban en 27 conjuntos de observaciones muestrales, la suma total de los cuadrados y la suma de los cuadrados de la regresión eran

$$STC = 3{,}881 \quad \text{y} \quad SCR = 3{,}549$$

a) Halle e interprete el coeficiente de determinación.
b) Halle la suma de los cuadrados de los errores.
c) Halle el coeficiente ajustado de determinación.
d) Halle e interprete el coeficiente de correlación múltiple.

13.20. En el estudio del ejercicio 13.8, en el que las estimaciones por mínimos cuadrados se basaban en 30 conjuntos de observaciones muestrales, la suma total de los cuadrados y la suma de los cuadrados de la regresión eran

$$STC = 162{,}1 \quad \text{y} \quad SCR = 88{,}2$$

a) Halle e interprete el coeficiente de determinación.
b) Halle el coeficiente de determinación ajustado.
c) Halle e interprete el coeficiente de correlación múltiple.

13.21. En el estudio del ejercicio 13.9, se utilizaron 25 observaciones para calcular las estimaciones por mínimos cuadrados. La suma de los cuadrados de la regresión y la suma de los cuadrados de los errores eran

$$SCR = 79{,}2 \quad \text{y} \quad SCE = 45{,}9$$

a) Halle e interprete el coeficiente de determinación.
b) Halle el coeficiente de determinación ajustado.
c) Halle e interprete el coeficiente de correlación múltiple.

13.22. Vuelva a los datos de las asociaciones de ahorro y crédito inmobiliario de la Tabla 13.1.
a) Estime por mínimos cuadrados la regresión del margen de beneficios con respecto al número de oficinas.
b) Estime por mínimos cuadrados la regresión de los ingresos netos con respecto al número de oficinas.
c) Estime por mínimos cuadrados la regresión del margen de beneficios con respecto a los ingresos netos.
d) Estime por mínimos cuadrados la regresión del número de oficinas con respecto a los ingresos netos.

13.4. Intervalos de confianza y contrastes de hipótesis de coeficientes de regresión individuales

En el apartado 13.2 hemos desarrollado y analizado los estimadores puntuales de los parámetros del modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

A continuación, desarrollamos intervalos de confianza y contrastes de hipótesis de los coeficientes de regresión estimados. Estos intervalos de confianza y contrastes de hipótesis dependen de la varianza de los coeficientes y de la distribución de probabilidad de los coeficientes. En el apartado 12.5 mostramos que el coeficiente de regresión simple es una función lineal de la variable dependiente, Y. Los coeficientes de regresión múltiple, b_j, también son funciones lineales de la variable dependiente, Y, pero el álgebra es algo más compleja y no se presentará aquí. En la ecuación de regresión múltiple anterior, vemos que la variable dependiente, Y, es una función lineal de las variables X más el error aleatorio ε. Para un conjunto dado de variables X, la función

$$\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki}$$

es en realidad una constante. También vimos en los Capítulos 5 y 6 que sumando una constante a una variable aleatoria ε se obtiene la variable aleatoria Y que tiene la misma distribución de probabilidad y la misma varianza que la variable aleatoria original ε. Como consecuencia, la variable dependiente, Y, sigue la misma distribución normal y tiene la misma varianza que el término de error, ε. Se deduce, pues, que los coeficientes de regresión, b_j —que son funciones lineales de Y—, también siguen una distribución normal y su varianza puede obtenerse utilizando la relación lineal entre los coeficientes de regresión y la variable dependiente. Este cálculo se realizaría siguiendo los mismos pasos que en el caso de la regresión simple del apartado 12.5, pero el álgebra es más compleja.

Basándonos en la relación lineal entre los coeficientes e Y, sabemos que las estimaciones de los coeficientes siguen una distribución normal si el error del modelo, ε, sigue una distribución normal. Como consecuencia del teorema del límite central, generalmente observamos que las estimaciones de los coeficientes siguen aproximadamente una distribución normal, aunque ε no la siga. Por lo tanto, los contrastes de hipótesis y los intervalos de confianza que desarrollamos no son afectados seriamente por las desviaciones con respecto a la normalidad en la distribución de los términos de error.

Podemos considerar que el término de error, ε, del modelo de regresión poblacional incluye las influencias conjuntas en la variable dependiente de multitud de factores no incluidos en la lista de variables independientes. Estos factores pueden no tener por separado una gran influencia, pero su efecto conjunto puede ser importante. El hecho de que el término de error esté formado por un gran número de componentes cuyos efectos son aleatorios es un argumento intuitivo para suponer que los errores de los coeficientes también siguen una distribución normal.

Como hemos visto antes, los estimadores de los coeficientes, b_j, son funciones lineales de Y, y el valor predicho de Y es una función lineal de los estimadores de los coeficientes de regresión. El computador realiza los cálculos resultantes de las complejas relaciones. Sin embargo, estas relaciones a veces pueden plantear problemas de interpretación, por lo que dedicamos algún tiempo a explicar la forma de calcular las varianzas. Si no compren-

demos cómo se calculan las varianzas, no podremos comprender perfectamente los contrastes de hipótesis y los intervalos de confianza.

La varianza de una estimación de un coeficiente depende del tamaño de la muestra, de la dispersión de las variables *X*, de las correlaciones entre las variables independientes y del término de error del modelo. Por lo tanto, estas correlaciones afectan tanto a los intervalos de confianza como a los contrastes de hipótesis. Antes hemos visto que las correlaciones entre las variables independientes influyen en los estimadores de los coeficientes. Estas correlaciones entre variables independientes también aumentan la varianza de los estimadores de los coeficientes. Una importante conclusión es que la varianza de los estimadores de los coeficientes, además de los estimadores de los coeficientes, depende de todo el conjunto de variables independientes del modelo de regresión.

El análisis anterior de los gráficos tridimensionales hacía hincapié en los complejos efectos que producen varias variables en la varianza de los coeficientes. A medida que son estrechas las relaciones entre las variables independientes, las estimaciones de los coeficientes son más inestables, es decir, tienen una varianza mayor. A continuación, presentamos un análisis más formal de estas complejidades. Para obtener buenas estimaciones de los coeficientes —estimaciones que tengan una baja varianza— debemos buscar un amplio rango para las variables independientes, elegir variables independientes que no estén estrechamente relacionadas entre sí y buscar un modelo que esté cerca de todos los puntos de datos. En la práctica, cuando se realizan estudios estadísticos aplicados en el mundo de la empresa y la economía, a menudo hay que utilizar datos que distan de ser ideales, como los del ejemplo de las asociaciones de ahorro y crédito inmobiliario. Pero conociendo los efectos aquí analizados, podemos contar con elementos para determinar en qué medida son aplicables nuestros modelos.

Para comprender algo el efecto de las correlaciones de variables independientes, examinamos los estimadores de las varianzas a partir del modelo de regresión múltiple estimado con dos variables de predicción:

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

Los estimadores de las varianzas de los coeficientes son

$$s_{b_1}^2 = \frac{s_e^2}{(n-1)s_{x_1}^2(1 - r_{x_1x_2}^2)} \tag{13.16}$$

$$s_{b_2}^2 = \frac{s_e^2}{(n-1)s_{x_2}^2(1 - r_{x_1x_2}^2)} \tag{13.17}$$

y las raíces cuadradas de estos estimadores de las varianzas, s_{b_1} y s_{b_2}, se denominan *errores típicos de los coeficientes*.

La varianza de los estimadores de los coeficientes aumenta directamente con la distancia a la que se encuentran los puntos de la línea, medida por s_e^2, la varianza de los errores estimados. Además, una dispersión mayor de los valores de las variables independientes —medida por $s_{x_1}^2$ o por $s_{x_2}^2$— reduce la varianza de los coeficientes. Recuérdese que estos resultados también se aplican a los estimadores de los coeficientes de regresión simple. También vemos que la varianza de los estimadores de los coeficientes aumenta con los aumentos de la correlación entre las variables independientes del modelo. A medida que aumenta la correlación entre dos variables independientes, es más difícil separar el efecto de cada una de las variables para predecir las variables dependientes. Cuando aumenta el

número de variables independientes en un modelo, las influencias en la varianza de los coeficientes continúan siendo importantes, pero la estructura algebraica se vuelve muy compleja y no se presenta aquí. El efecto de las correlaciones hace que los estimadores de las varianzas de los coeficientes dependan de las demás variables independientes del modelo. Recuérdese que los estimadores efectivos de los coeficientes también dependen de las demás variables independientes del modelo, una vez más debido al efecto de las correlaciones entre las variables independientes.

A continuación, resumimos la base para la inferencia de los coeficientes de la regresión poblacional. Normalmente, nos interesan más los coeficientes de regresión β_j que la constante u ordenada en el origen β_0. Por lo tanto, centraremos la atención en los primeros, señalando que la inferencia sobre la segunda se realiza de una manera parecida.

Base para la inferencia de los parámetros de la regresión poblacional

Sea el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Sean b_0, b_1, ..., b_K las estimaciones por mínimos cuadrados de los parámetros poblacionales y s_{b_0}, s_{b_1}, ..., s_{b_K} las desviaciones típicas estimadas de los estimadores por mínimos cuadrados. Entonces, si se cumplen los supuestos habituales de la regresión y si los términos de error, ε_i, siguen una distribución normal,

$$t_{b_j} = \frac{b_j - \beta_j}{s_{b_j}} \; (j = 1, 2, ..., K) \qquad \textbf{(13.18)}$$

se distribuye como una distribución t de Student con $(n - K - 1)$ grados de libertad.

Intervalos de confianza

Pueden obtenerse intervalos de confianza de los β_j utilizando la ecuación 13.19.

Intervalos de confianza de los coeficientes de regresión

Si los errores de la regresión poblacional, ε_i, siguen una distribución normal y se cumplen los supuestos habituales de la regresión, los intervalos de confianza bilaterales al $100(1 - \alpha)\%$ de los coeficientes de regresión, β_j, son

$$b_j - t_{n-K-1,\,\alpha/2} s_{b_j} < \beta_j < b_j + t_{n-K-1,\,\alpha/2} s_{b_j} \qquad \textbf{(13.19)}$$

donde $t_{n-K-1,\,\alpha/2}$ es el número para el que

$$P(t_{n-K-1} > t_{n-K-1,\,\alpha/2}) = \frac{\alpha}{2}$$

y la variable aleatoria t_{n-K-1} sigue una distribución t de Student con $(n - K - 1)$ grados de libertad.

EJEMPLO 13.4. Desarrollo del modelo de las asociaciones de ahorro y crédito inmobiliario (estimación de intervalos de confianza)

Se nos ha pedido que calculemos intervalos de confianza de los coeficientes del modelo de regresión de las asociaciones de ahorro y crédito inmobiliario presentado en el ejemplo 13.3.

Solución

La Figura 13.8 muestra la salida Minitab del análisis de regresión correspondiente al modelo de regresión de las asociaciones de ahorro y crédito inmobiliario. Los estimado-

Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices

Predictor              Coef      SE Coef        T        P
Constant            1.56450      0.07940    19.70    0.000
X1 revenue          0.23720      0.05556     4.27    0.000
X2 offices       -0.00024908   0.00003205    -7.77    0.000

S = 0.0533022  R-Sq = 86.5%   R-Sq(adj) = 85.3%

Analysis of Variance

Source            DF        SS        MS        F        P
Regression         2   0.40151   0.20076    70.66    0.000
Residual Error    22   0.06250   0.00284
Total             24   0.46402

Source         DF         Seq SS
X1 revenue      1         0.22990
X2 offices      1         0.17161
```

b_1, s_{b_1}, t_{b_1}, t_{b_2}, s_{b_2}, b_2

(a)

Estadísticos de la regresión						
Coeficiente de correlación múltiple	0,930212915					
Coeficiente de determinación R	0,865296068					
R ajustado	0,853050256					
Error típico	0,053302217					
Observaciones	25					
ANÁLISIS DE VARIANZA						
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
Regresión	2	0,40151122	0,20075561	70,66057082	2,64962E-10	
Residuos	22	0,06250478	0.002841126			
Total	24	0,464016				
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	1,564496771	0,079395981	19,70498685	1,81733E-15	1,399839407	1,72915414
X1 ingresos	0,237197475	0,055559366	4,269261695	0,000312567	0,121974278	0,35242067
X2 número de oficinas	–0,000249079	3,20485E-05	–7,771949195	9,50879E-08	–0,000315544	–0,00018261

b_1, b_2, s_{b_1}, s_{b_2}, t_{b_1}, t_{b_2}

(b)

Figura 13.8. Regresión de problema de las asociaciones de ahorro y crédito inmobiliario (salidas Minitab y Excel).

res de los coeficientes y sus desviaciones típicas correspondientes a las variables de predicción de los ingresos, b_1, y el número de oficinas, b_2, son

$$b_1 = 0{,}2372 \qquad s_{b_1} = 0{,}05556; \qquad b_2 = -0{,}000249 \qquad s_{b_2} = 0{,}00003205$$

Vemos, pues, que la desviación típica de la distribución en el muestreo del estimador por mínimos cuadrados de β_1 se estima en 0,05556 y la de β_2 se estima en 0,00003205.

Para obtener intervalos de confianza al 99 por ciento de β_1 y β_2, utilizamos el valor t de Student de la Tabla 8 del apéndice.

$$t_{n-K-1,\,\alpha/2} = t_{22,\,0{,}005} = 2{,}819$$

Basándonos en estos resultados, observamos que el intervalo de confianza al 99 por ciento de β_1 es

$$0{,}237 - (2{,}819)(0{,}05556) < \beta_1 < 0{,}237 + (2{,}819)(0{,}05556)$$

o sea,

$$0{,}080 < \beta_1 < 0{,}394$$

Por lo tanto, el intervalo de confianza al 99 por ciento del aumento esperado del margen de beneficios de las asociaciones de ahorro y crédito inmobiliario provocado por un aumento de los ingresos netos de 1 unidad, dado un número fijo de oficinas, va de 0,080 a 0,394. El intervalo de confianza al 99 por ciento de β_2 es

$$-0{,}000249 - (2{,}819)(0{,}0000320) < \beta_2 - 0{,}000249 + (2{,}819)(0{,}0000320)$$

o sea

$$-0{,}000339 < \beta_2 < -0{,}000159$$

Vemos, pues, que el intervalo de confianza al 99 por ciento de la disminución esperada del margen de beneficios provocada por un aumento de 1.000 oficinas, dado un nivel fijo de ingresos netos, va de 0,159 a 0,339.

Contrastes de hipótesis

Pueden desarrollarse contrastes de hipótesis de los coeficientes de regresión utilizando las estimaciones de las varianzas de los coeficientes. Especialmente interesante es el contraste de hipótesis

$$H_0 : \beta_j = 0$$

que se utiliza frecuentemente para averiguar si una variable independiente específica es importante en un modelo de regresión múltiple.

Contrastes de hipótesis de los coeficientes de regresión

Si los errores de la regresión, ε_i, siguen una distribución normal y se cumplen los supuestos habituales del análisis de regresión, los siguientes contrastes de hipótesis tienen el nivel de significación α:

1. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \beta_j = \beta^* \quad \text{o} \quad H_0 : \beta_j \leqslant \beta^*$$

frente a la hipótesis alternativa

$$H_1 : \beta_j > \beta^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{b_j - \beta^*}{s_{b_j}} > t_{n-K-1,\alpha} \tag{13.20}$$

2. Para contrastar cualquiera de las dos hipótesis nulas

$$H_0 : \beta_j = \beta^* \quad \text{o} \quad H_0 : \beta_j \geqslant \beta^*$$

frente a la hipótesis alternativa

$$H_1 : \beta_j < \beta^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{b_j - \beta^*}{s_{b_j}} < -t_{n-K-1,\alpha} \tag{13.21}$$

3. Para contrastar la hipótesis nula

$$H_0 : \beta_j = \beta^*$$

frente a la hipótesis alternativa bilateral

$$H_1 : \beta_j \neq \beta^*$$

la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{b_j - \beta^*}{s_{b_j}} > t_{n-K-1,\alpha/2} \quad \text{o} \quad \frac{b_j - \beta^*}{s_{b_j}} < -t_{n-K-1,\alpha/2} \tag{13.22}$$

Muchos analistas sostienen que si no podemos rechazar la hipótesis condicionada de que el coeficiente es 0, debemos concluir que la variable no debe incluirse en el modelo de regresión. El estadístico t de Student de este contraste normalmente se calcula en la mayoría de los programas de regresión y se indica al lado de la estimación de la varianza de los coeficientes; además, normalmente se incluye el p-valor del contraste de hipótesis. Éstos se muestran en la salida Minitab de la Figura 13.8(a). Utilizando el estadístico t de Student indicado o el p-valor, podemos saber inmediatamente si una variable de predicción es significativa, dadas las demás variables del modelo de regresión.

Existen claramente otros métodos para decidir si una variable independiente debe incluirse en un modelo de regresión. Vemos que el método de selección anterior no tiene en cuenta el error de Tipo II: el coeficiente poblacional no es igual a 0, pero no rechazamos la hipótesis nula de que es igual a 0. Éste es un problema importante cuando un modelo basado en la teoría económica o en otra teoría y especificado con cuidado incluye ciertas variables independientes. En ese caso, debido a un gran error, ε, y/o a las correlaciones entre variables independientes, no podemos rechazar la hipótesis de que el coeficiente es 0. En este caso, muchos analistas incluirán la variable independiente en el modelo porque creen que debe primar la especificación original del modelo basada en la teoría o la experiencia

económicas. Se trata de una cuestión difícil que exige hacer una buena valoración basándose tanto en los resultados estadísticos como en la teoría económica sobre la relación subyacente analizada.

EJEMPLO 13.5. Desarrollo del modelo de las asociaciones de ahorro y crédito inmobiliario (contrastes de hipótesis de coeficientes)

Se nos ha pedido que averigüemos si los coeficientes del modelo de regresión de las asociaciones de ahorro y crédito inmobiliario son predictores significativos de los beneficios.

Solución

En el contraste de hipótesis para esta cuestión utilizaremos los resultados de la regresión realizada con el programa Minitab mostrados en la Figura 13.8(a). En primer lugar, queremos averiguar si los ingresos totales aumentan significativamente los beneficios dado el efecto del número de oficinas, es decir, descontando la influencia de éste. La hipótesis nula es

$$H_0 : \beta_1 = 0$$

frente a la hipótesis alternativa

$$H_1 : \beta_1 > 0$$

El contraste puede realizarse calculando el estadístico t de Student del coeficiente, dado H_0:

$$t_{b_1} = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{0{,}237 - 0}{0{,}05556} = 4{,}27$$

En la Tabla 8 de la t de Student del apéndice podemos ver que el valor crítico del estadístico t de Student es

$$t_{22,\,0{,}005} = 2{,}819$$

La Figura 13.8(a) también indica que el p-valor del contraste de hipótesis es inferior a 0,005. Basándonos en esta evidencia, rechazamos H_0 y aceptamos H_1 y concluimos que los ingresos totales son un predictor estadísticamente significativo del aumento de los beneficios de las asociaciones de ahorro y crédito inmobiliario, dado que hemos tenido en cuenta el efecto del número de oficinas.

También podemos averiguar si el número total de oficinas reduce significativamente los márgenes de beneficios. La hipótesis nula es

$$H_0 : \beta_2 = 0$$

frente a la hipótesis alternativa

$$H_1 : \beta_2 < 0$$

El contraste puede realizarse calculando el estadístico t de Student del coeficiente, dado H_0:

$$t_{b_2} = \frac{b_2 - \beta_2}{s_{b_2}} = \frac{-0{,}000249 - 0}{0{,}0000320} = -7{,}77$$

En la Tabla 8 del apéndice podemos ver que el valor crítico del estadístico t de Student es

$$t_{22,\,0{,}005} = 2{,}819$$

La Figura 13.8(a) también indica que el p-valor del contraste de hipótesis es inferior a 0,005. Basándonos en esta evidencia, rechazamos H_0 y aceptamos H_1 y concluimos que el número de oficinas es un predictor estadísticamente significativo de la reducción de los beneficios de las asociaciones de ahorro y crédito inmobiliario, dado que hemos tenido en cuenta el efecto de los ingresos totales.

Es importante hacer hincapié en que los dos contrates de hipótesis se basan en el conjunto de variables incluidas en el modelo de regresión. Por ejemplo, si se incluyeran más variables de predicción, estos contrastes ya no serían válidos. Con más variables en el modelo, las estimaciones de los coeficientes y sus desviaciones típicas estimadas serían diferentes y, por lo tanto, también lo sería el estadístico t de Student.

Obsérvese que en la salida Minitab del análisis de regresión mostrada en la Figura 13.8(a), el estadístico t de Student de la hipótesis nula —$H_0 : \beta_j = 0$— es el cociente entre el coeficiente estimado y el error típico del coeficiente estimado, que se encuentra en las dos columnas situadas a la izquierda del estadístico t de Student. También se muestra la probabilidad o p-valor del contraste de hipótesis de dos colas: $H_j : \beta_j \neq 0$. Por lo tanto, cualquier analista puede realizar estos contrastes de hipótesis directamente examinando la salida del análisis de regresión múltiple. El estadístico t de Student y el p-valor se calculan en todos los paquetes estadísticos modernos. La mayoría de los analistas buscan estos resultados habitualmente cuando examinan la salida del análisis de regresión de un programa estadístico.

EJEMPLO 13.6. Factores que afectan al tipo del impuesto sobre bienes inmuebles (análisis de los coeficientes de regresión)

Un ayuntamiento encargó un estudio para averiguar los factores que influyen en los impuestos urbanos sobre los bienes inmuebles de las ciudades de 100.000-200.000 habitantes.

Solución

Utilizando una muestra de 20 ciudades de Estados Unidos, se estimó el siguiente modelo de regresión:

$$\hat{y} = 1{,}79 + \underset{(0{,}000139)}{0{,}000567x_1} + \underset{(0{,}0082)}{0{,}0183x_2} - \underset{(0{,}000446)}{0{,}000191x_3}$$

$$R^2 = 0{,}71 \qquad n = 20$$

donde

y = tipo efectivo del impuesto de bienes inmuebles (impuestos efectivos divididos por el valor de mercado de la base impositiva)

x_1 = número de viviendas por kilómetro cuadrado

x_2 = porcentaje de los ingresos municipales totales representado por las ayudas procedentes de las administraciones de los estados y de la administración federal
x_3 = renta personal per cápita mediana en dólares

Los números entre paréntesis que se encuentran debajo de los coeficientes son los errores típicos de los coeficientes estimados.

La presentación anterior constituye un buen formato para mostrar los resultados de un modelo de regresión. Los resultados indican que las estimaciones condicionadas de los efectos de las tres variables de predicción son las siguientes:

1. Un aumento de una vivienda por kilómetro cuadrado eleva el tipo efectivo del impuesto sobre bienes inmuebles en 0,000567. Obsérvese que los tipos del impuesto sobre bienes inmuebles normalmente se expresan en dólares por cada 1.000 $ de valor catastral de la propiedad. Así, un aumento de 0,000567 indica que los tipos del impuesto sobre bienes inmuebles son 0,567 $ más altos por 1.000 $ de valor catastral de la propiedad.
2. Un aumento de los ingresos municipales totales de un 1 por ciento procedente de las ayudas de las administraciones de los estados y de la administración federal eleva el tipo impositivo efectivo en 0,0183.
3. Un aumento de la renta personal per cápita mediana de 1 $ provoca una disminución esperada del tipo impositivo efectivo de 0,000191.

Hacemos de nuevo hincapié en que estas estimaciones de los coeficientes sólo son válidas en un modelo que incluya las tres variables de predicción anteriores.

Para comprender mejor la exactitud de estos efectos, construiremos intervalos de confianza al 95 por ciento condicionados. En el modelo de regresión estimado, el error tiene $(20 - 3 - 1) = 16$ grados de libertad. Por lo tanto, el estadístico t de Student para calcular los intervalos de confianza es, como se observa en el apéndice, $t_{16,\,0{,}025} = 2{,}12$. El formato del intervalo de confianza es

$$b_j - t_{n-K-1,\,\alpha/2} s_{b_j} < \beta_j < b_j + t_{n-K-1,\,\alpha/2} s_{b_j}$$

Por lo tanto, el coeficiente del número de viviendas por kilómetro cuadrado tiene un intervalo de confianza al 95 por ciento de

$$\begin{gathered} 0{,}000567 - (2{,}12)(0{,}000139) < \beta_1 < 0{,}000567 + (2{,}12)(0{,}000139) \\ 0{,}000272 < \beta_1 < 0{,}000862 \end{gathered}$$

El coeficiente del porcentaje de ingresos representados por las ayudas tiene un intervalo de confianza al 95 por ciento de

$$\begin{gathered} 0{,}0183 - (2{,}12)(0{,}0082) < \beta_2 < 0{,}0183 + (2{,}12)(0{,}0082) \\ 0{,}0009 < \beta_2 < 0{,}0357 \end{gathered}$$

Por último, el coeficiente de la renta personal per cápita mediana tiene un intervalo de confianza al 95 por ciento de

$$\begin{gathered} -0{,}000191 - (2{,}12)(0{,}000446) < \beta_3 < -0{,}000191 + (2{,}12)(0{,}000446) \\ -0{,}001137 < \beta_3 < 0{,}000755 \end{gathered}$$

Una vez más hacemos hincapié en que estos intervalos dependen de que se incluyan las tres variables de predicción en el modelo.

Vemos que el intervalo de confianza al 95 por ciento de β_3 incluye 0 y, por lo tanto, podríamos no rechazar la hipótesis de dos colas de que este coeficiente es 0. Basándonos en este intervalo de confianza, concluimos que X_3 no es una variable de predicción estadísticamente significativa en el modelo de regresión múltiple. Sin embargo, los intervalos de confianza de las otras dos variables no incluyen 0 y, por lo tanto, concluimos que éstas son estadísticamente significativas.

EJEMPLO 13.7. Efectos de los factores fiscales en los precios de la vivienda (estimación de los coeficientes del modelo de regresión)

Citydat

Northern City (Minnesota) tenía interés en saber cómo afectaba la promoción inmobiliaria local al precio de mercado de las viviendas de la ciudad. Northern City es una de las numerosas ciudades no metropolitanas pequeñas del Medio Oeste de Estados Unidos cuya población oscila entre 6.000 y 40.000 habitantes. Uno de los objetivos era averiguar cómo influiría un aumento de la cantidad de locales comerciales en el valor de las viviendas locales. Los datos se encuentran en el fichero de datos **Citydat**.

Solución

Para responder a esta pregunta, se recogieron datos de algunas ciudades y se utilizaron para construir un modelo de regresión que estima el efecto de variables clave en el precio de la vivienda. Para este estudio se obtuvieron las siguientes variables de cada ciudad:

Y (hseval) = precio medio de mercado de las viviendas de la ciudad
X_1 (sizehse) = número medio de habitaciones de las viviendas
X_2 (incom72) = renta media de los hogares
X_3 (taxrate) = tipo impositivo por mil dólares de valor catastral de las viviendas
X_4 (comper) = porcentaje de propiedades inmobiliarias imponibles que son comerciales

La Figura 13.9 muestra los resultados de la regresión múltiple, obtenidos por medio del programa Minitab. El coeficiente del número medio de habitaciones de las viviendas es 7,878 y la desviación típica del coeficiente es 1,809. En este estudio, los valores de las viviendas se expresan en unidades de 1.000 $ y la media de todas las ciudades es de 21.000 $. Así, por ejemplo, si el número medio de habitaciones de las viviendas de una ciudad es mayor en 1,0, el precio medio es mayor en 7.878 $. El estadístico t de Student resultante es 4,35 y el p-valor es 0,000. Por lo tanto, se rechaza la hipótesis condicionada de que este coeficiente es igual a 0. Se obtiene el mismo resultado en el caso de las variables de la renta y del tipo impositivo. La variable «incom72» está expresada en unidades de dólares y, por lo tanto, si la renta media de una ciudad es mayor en 1.000 $, el coeficiente de 0,003666 indica que el precio medio de la vivienda es 3.666 $ mayor. Si el tipo impositivo aumenta un 1 por ciento, el precio medio de la vivienda se reduce en 1.720 $. Vemos que el análisis de regresión lleva a la conclusión de que cada una de estas tres variables es un importante predictor del precio medio de la vivienda de las ciudades incluidas en este estudio. Sin embargo, vemos que el coeficiente del porcentaje de locales comerciales, «comper», es $-10{,}614$ y la desviación típica del coeficiente es 6,491, lo que da un estadístico t de Student igual a $-1{,}64$. Obsérvese que este resultado permite establecer una importante conclusión. El coeficiente tendría un p-valor de

Regression Analysis: hseval versus sizehse, income72, taxrate, Comper

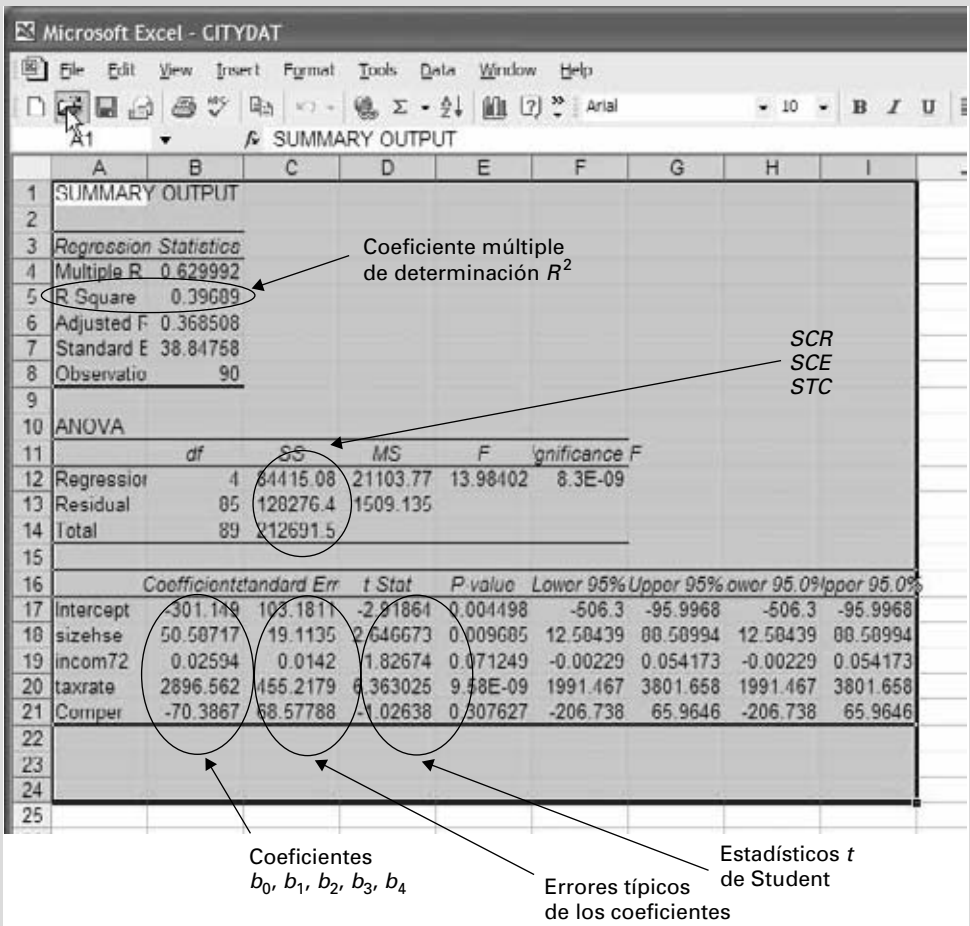
```
The regression equation is
hseval = -28.1 + 7.88 sizehse + 0.000367 incom72 - 172 taxrate -10.6 Comper

Predictor          Coef     SE Coef        T       P
Constant       -28.075       9.766    -2.87   0.005
Sizehse          7.878       1.809     4.35   0.000
incom72       0.003666    0.001344     2.73   0.008
taxrate        -171.80       43.09    -3.99   0.000
Comper         -10.614       6.491    -1.64   0.106


S = 3.67686    R-Sq = 47.4%   R-Sq(adj) = 45.0%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         4   1037.49    259.37   19.19   0.000
Residual Error    85   1149.14     13.52
Total             89   2186.63
```

Figura 13.9. Modelo de regresión del precio de la vivienda (salida Minitab).

una cola de 0,053 o un *p*-valor de dos colas de 0,106. Por lo tanto, parece que reduce algo el precio medio de las viviendas. Dado que se han incluido los efectos del tamaño de las viviendas, la renta y el tipo impositivo en el precio de mercado de las viviendas, vemos que el porcentaje de locales comerciales no eleva los precios de la vivienda. Por lo tanto, este análisis no apoya el argumento de que el valor de mercado de las viviendas aumentará si se construyen más locales comerciales. Esa conclusión sólo es cierta en un modelo que incluya estas cuatro variables de predicción. Obsérvese también que los valores de $R^2 = 47{,}4$ por ciento y s_e (error típico de la regresión) = 3,677 están incluidos en la salida del análisis de regresión.

Los defensores de un aumento de la promoción de locales comerciales también sostenían que el aumento de la cantidad de locales comerciales reduciría los impuestos pagados por las viviendas ocupadas por sus propietarios. Esta tesis se contrastó utilizando los resultados de la regresión de la Figura 13.10 obtenidos con el programa Excel. Se indican los estimadores de los coeficientes y sus errores típicos. Los estadísticos *t* de Student de los coeficientes del tamaño de la vivienda y el tipo impositivo son 2,65 y 6,36, lo cual indica que estas variables son importantes predictores. El estadístico *t* de Student de la renta es 1,83 con un *p*-valor de 0,07 para un contraste de dos colas. Por lo tanto, la renta tiene alguna influencia como predictor, pero su efecto no es tan fuerte como el de las dos variables anteriores. Vemos de nuevo que hay margen para extraer conclusiones sólidas. La hipótesis condicionada de que un aumento de los locales comerciales reduce los impuestos sobre las viviendas ocupadas por sus propietarios puede contrastarse utilizando el estadístico *t* de Student de la variable «comper» en los resultados de la regresión. El estadístico *t* de Student es $-1{,}03$ con un *p*-valor de 0,308. Por lo tanto, la hipótesis de que un aumento de los locales comerciales no reduce los impuestos sobre la vivienda no puede rechazarse. No existen pruebas en este análisis de que los impuestos sobre las viviendas disminuirían si se construyeran más locales comerciales.

Basándose en los análisis de regresión realizados en este estudio, los consultores llegaron a la conclusión de que no existían pruebas de que un aumento de los locales comerciales elevaría el valor de mercado de las viviendas o reduciría los impuestos sobre bienes inmuebles de las viviendas.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	Regression Statistics								
4	Multiple R	0.629992							
5	R Square	0.39689							
6	Adjusted F	0.368508							
7	Standard E	38.84758							
8	Observatio	90							
9									
10	ANOVA								
11		df	SS	MS	F	ignificance F			
12	Regressior	4	84415.08	21103.77	13.98402	8.3E-09			
13	Residual	85	128276.4	1509.135					
14	Total	89	212691.5						
15									
16		Coefficients	tandard Err	t Stat	P value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
17	Intercept	-301.149	103.1811	-2.91864	0.004498	-506.3	-95.9968	-506.3	-95.9968
18	sizehse	50.58717	19.1135	2.646673	0.009685	12.58439	88.58994	12.58439	88.58994
19	incom72	0.02594	0.0142	1.82674	0.071249	-0.00229	0.054173	-0.00229	0.054173
20	taxrate	2896.562	455.2179	6.363025	9.58E-09	1991.467	3801.658	1991.467	3801.658
21	Comper	-70.3867	68.57788	-1.02638	0.307627	-206.738	65.9646	-206.738	65.9646
22									
23									
24									
25									

Figura 13.10. Modelo de regresión de los impuestos sobre las viviendas (salida Excel).

EJERCICIOS

Ejercicios básicos

13.23. Los resultados del análisis de un modelo de regresión son los siguientes:

$$\hat{y} = \underset{}{1{,}50} + \underset{(2,1)}{4{,}8}x_1 + \underset{(3,7)}{6{,}9}x_2 - \underset{(2,8)}{7{,}2}x_3$$

$$R^2 = 0{,}71 \qquad n = 24$$

Los números entre paréntesis situados debajo de las estimaciones de los coeficientes son los errores típicos muestrales de las estimaciones de los coeficientes.

a) Calcule intervalos de confianza al 95 por ciento bilaterales de los tres coeficientes de la pendiente de regresión.

b) Contraste para cada uno de los coeficientes de la pendiente las hipótesis

$$H_0 : \beta_j = 0 \quad \text{frente a} \quad H_1 : \beta_j > 0$$

13.24. Los resultados del análisis de un modelo de regresión son los siguientes:

$$\hat{y} = \underset{}{2{,}50} + \underset{(3,1)}{6{,}8}x_1 + \underset{(3,7)}{6{,}9}x_2 - \underset{(3,2)}{7{,}2}x_3$$

$$R^2 = 0{,}85 \qquad n = 34$$

Los números entre paréntesis situados debajo de las estimaciones de los coeficientes son los errores típicos muestrales de las estimaciones de los coeficientes.

a) Calcule intervalos de confianza al 95 por ciento bilaterales de los tres coeficientes de la pendiente de regresión.
b) Contraste para cada uno de los coeficientes de la pendiente las hipótesis

$$H_0: \beta_j = 0 \quad \text{frente a} \quad H_1: \beta_j > 0$$

13.25. Los resultados del análisis de un modelo de regresión son los siguientes:

$$\hat{y} = \underset{}{-101{,}50} + \underset{(12{,}1)}{34{,}8x_1} + \underset{(23{,}7)}{56{,}9x_2} - \underset{(32{,}8)}{57{,}2x_3}$$

$$R^2 = 0{,}71 \qquad n = 65$$

Los números entre paréntesis situados debajo de las estimaciones de los coeficientes son los errores típicos muestrales de las estimaciones de los coeficientes.

a) Calcule intervalos de confianza al 95 por ciento bilaterales de los tres coeficientes de la pendiente de regresión.
b) Contraste para cada uno de los coeficientes de la pendiente las hipótesis

$$H_0: \beta_j = 0 \quad \text{frente a} \quad H_1: \beta_j > 0$$

13.26. Los resultados del análisis de un modelo de regresión son los siguientes:

$$\hat{y} = \underset{}{-9{,}50} + \underset{(7{,}1)}{17{,}8x_1} + \underset{(13{,}7)}{26{,}9x_2} - \underset{(3{,}8)}{9{,}2x_3}$$

$$R^2 = 0{,}71 \qquad n = 39$$

Los números entre paréntesis situados debajo de las estimaciones de los coeficientes son los errores típicos muestrales de las estimaciones de los coeficientes.

a) Calcule intervalos de confianza al 95 por ciento bilaterales de los tres coeficientes de la pendiente de regresión.
b) Contraste para cada uno de los coeficientes de la pendiente las hipótesis

$$H_0: \beta_j = 0 \quad \text{frente a} \quad H_1: \beta_j > 0$$

Ejercicios aplicados

13.27. En el estudio del ejercicio 13.6, los errores típicos estimados eran

$$s_{b_1} = 0{,}099 \qquad s_{b_2} = 0{,}032 \qquad s_{b_3} = 0{,}002$$

a) Halle intervalos de confianza al 90 y el 95 por ciento de β_1.
b) Halle intervalos de confianza al 95 y el 99 por ciento de β_2.
c) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, el peso del avión no tiene una influencia lineal en su esfuerzo de diseño frente a la hipótesis alternativa bilateral.
d) La suma de los cuadrados de los errores de esta regresión era 0,332. Utilizando los mismos datos, se ajustó una regresión lineal simple del esfuerzo de diseño con respecto al número porcentual de piezas comunes, lo que dio una suma de los cuadrados de los errores de 3,311. Contraste al nivel del 1 por ciento la hipótesis nula de que la velocidad máxima y el peso, considerados conjuntamente, no contribuyen nada en un sentido lineal a la explicación del esfuerzo de diseño, dado que el número porcentual de piezas comunes también se utiliza como variable explicativa.

13.28. En el estudio del ejercicio 13.8, en el que la regresión muestral se basaba en 30 observaciones, los errores típicos estimados eran

$$s_{b_1} = 0{,}023 \qquad s_{b_2} = 0{,}35$$

a) Contraste la hipótesis nula de que, dado el tamaño de la familia, el consumo de leche no depende linealmente de la renta frente a la hipótesis alternativa unilateral adecuada.
b) Halle intervalos de confianza del 90, el 95 y el 99 por ciento de β_2.

13.29. En el estudio de los ejercicios 13.9 y 13.21, en los que la regresión muestral se basaba en 25 observaciones, los errores típicos estimados eran

$$s_{b_1} = 0{,}189 \qquad s_{b_2} = 0{,}565 \qquad s_{b_3} = 0{,}243$$

a) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, las horas de ejercicio no influyen linealmente en el aumento de peso frente a la hipótesis alternativa unilateral adecuada.
b) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, el consumo de cerveza no influye linealmente en el aumento de peso frente a la hipótesis alternativa unilateral adecuada.
c) Halle intervalos de confianza del 90, el 95 y el 99 por ciento de β_1.

13.30. Vuelva a los datos del ejemplo 13.6.

a) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, la renta

personal per cápita mediana no influye en el tipo efectivo del impuesto sobre bienes inmuebles frente a una hipótesis alternativa bilateral.

b) Contraste la hipótesis nula de que las tres variables independientes, consideradas conjuntamente, no influyen linealmente en el tipo efectivo del impuesto sobre bienes inmuebles.

13.31. Vuelva a los datos del ejemplo 13.7 que se encuentran en el fichero de datos **Citydat**.

a) Halle intervalos de confianza al 95 y al 99 por ciento de la variación esperada del precio de mercado de las viviendas provocada por un aumento del número medio de habitaciones de 1 unidad cuando no varían los valores de todas las demás variables independientes.

b) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, la renta media de los hogares no influye en el precio de mercado frente a la hipótesis alternativa de que cuanto mayor es la renta media de los hogares, más alto es el precio de mercado.

13.32. En un estudio de los ingresos generados por las loterías nacionales, se ajustó la siguiente ecuación de regresión de 29 países que tienen loterías:

$$\hat{y} = \underset{}{-31{,}323} + \underset{(0{,}00755)}{0{,}04045}x_1 + \underset{(0{,}3107)}{0{,}8772}x_2 - \underset{(263{,}88)}{365{,}01}x_3 - \underset{(3{,}4520)}{9{,}9298}x_4$$

$$R^2 = 0{,}51$$

donde

y = dólares de ingresos anuales netos per cápita generados por la lotería
x_1 = renta personal media per cápita del país
x_2 = número de hoteles, moteles, hostales y albergues por mil habitantes del país
x_3 = ingresos anuales gastables per cápita generados por las apuestas, las carreras y otros juegos de azar legalizados
x_4 = porcentaje de la frontera nacional que limita con un país o países que tienen una lotería

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete el coeficiente estimado de x_1.

b) Halle e interprete el intervalo de confianza al 95 por ciento del coeficiente de x_2 en la regresión poblacional.

c) Contraste la hipótesis nula de que el coeficiente de x_3 en la regresión poblacional es 0 frente a la hipótesis alternativa de que este coeficiente es negativo. Interprete sus resultados.

13.33. Se realizó un estudio para averiguar si podían utilizarse algunas características para explicar la variabilidad de los precios de los hornos. Se estimó para una muestra de 19 hornos la siguiente regresión:

$$\hat{y} = \underset{}{-68{,}236} + \underset{(0{,}005)}{0{,}0023}x_1 + \underset{(8{,}992)}{19{,}729}x_2 + \underset{(3{,}082)}{7{,}653}x_3$$

$$R^2 = 0{,}84$$

donde

y = precio en dólares
x_1 = potencia del horno en BTU por hora
x_2 = coeficiente de eficiencia energética
x_3 = número de posiciones

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Halle el intervalo de confianza al 95 por ciento de la subida esperada del precio resultante de un aumento de las posiciones cuando los valores de la potencia y el índice de eficiencia energética se mantienen fijos.

b) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, el índice de eficiencia energética de los hornos no afecta a su precio frente a la hipótesis alternativa de que cuanto más alto es el índice de eficiencia energética, más alto es el precio.

13.34. En un estudio de la demanda nigeriana de importaciones se ajustó el siguiente modelo a 19 años de datos:

$$\hat{y} = \underset{}{-58{,}9} + \underset{(0{,}0092)}{0{,}20}x_1 - \underset{(0{,}084)}{0{,}10}x_2 \qquad \bar{R}^2 = 0{,}96$$

donde

y = cantidad de importaciones
x_1 = gastos personales de consumo
x_2 = precio de las importaciones ÷ precios interiores

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Halle el intervalo de confianza al 95 por ciento de β_1.

b) Contraste la hipótesis nula de que $\beta_2 = 0$ frente a la hipótesis alternativa unilateral adecuada.

13.35. En un estudio de las tenencias extranjeras en bancos británicos, se obtuvo la siguiente regresión muestral, basada en 14 observaciones anuales

$$\hat{y} = \underset{}{-3{,}248} + \underset{(0{,}0023)}{0{,}101}x_1 - \underset{(0{,}080)}{0{,}244}x_2 + \underset{(0{,}00925)}{0{,}057}x_3 \quad R^2 = 0{,}93$$

donde

y = proporción de activos a final del año en filiales de bancos británicos en manos de extranjeros en porcentaje de los activos totales
x_1 = variación anual, en miles de millones de libras, de la inversión extranjera directa en Gran Bretaña (excluidos finanzas, seguros y bienes inmuebles)
x_2 = relación precio-beneficios de los bancos
x_3 = índice del valor de cambio de la libra

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Halle el intervalo de confianza al 90 por ciento de β_1 e interprete su resultado.
b) Contraste la hipótesis nula de que β_2 es 0 frente a la hipótesis alternativa de que es negativo e interprete su resultado.
c) Contraste la hipótesis nula de que β_3 es 0 frente a la hipótesis alternativa de que es positivo e interprete su resultado.

13.36. En un estudio de las diferencias entre los niveles de demanda de bomberos por parte de las ciudades, se obtuvo la siguiente regresión muestral, basada en datos de 39 ciudades de Maryland:

$$\hat{y} = \underset{}{-0{,}00232} - \underset{(0{,}00010)}{0{,}00024}x_1 - \underset{(0{,}000018)}{0{,}00002}x_2 + \underset{(0{,}00012)}{0{,}00034}x_3$$
$$+ \underset{(0{,}77954)}{0{,}48122}x_4 + \underset{(0{,}01172)}{0{,}04950}x_5 - \underset{(0{,}00005)}{0{,}00010}x_6 + \underset{(0{,}00306)}{0{,}00645}x_7$$
$$\bar{R}^2 = 0{,}3572$$

donde

y = número de bomberos a tiempo completo per cápita
x_1 = salario base máximo de los bomberos en miles de dólares
x_2 = porcentaje de población
x_3 = renta per cápita estimada en miles de dólares
x_4 = densidad de población
x_5 = cantidad de ayudas intergubernamentales per cápita en miles de dólares
x_6 = número de kilómetros de distancia hasta la capital de la región
x_7 = porcentaje de la población que son varones y tienen entre 12 y 21 años

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Halle e interprete el intervalo de confianza al 99 por ciento de β_5.
b) Contraste la hipótesis nula de que β_4 es 0 frente a la hipótesis alternativa bilateral e interprete su resultado.
c) Contraste la hipótesis nula de que β_7 es 0 frente a la hipótesis alternativa bilateral e interprete su resultado.

13.5. Contrastes de los coeficientes de regresión

En el apartado anterior hemos mostrado cómo puede realizarse un contraste de hipótesis condicionado para averiguar si el coeficiente de una variable específica es significativo en un modelo de regresión. Existen, sin embargo, situaciones en las que nos interesa saber cuál es el efecto de la combinación de varias variables. Por ejemplo, en un modelo que predice la cantidad vendida, podría interesarnos saber cuál es el efecto conjunto tanto del precio del vendedor como del precio del competidor. En otros casos, podría interesarnos saber si la combinación de todas las variables es un útil predictor de la variable dependiente.

Contrastes de todos los coeficientes

En primer lugar, presentamos contrastes de hipótesis para averiguar si los conjuntos de varios coeficientes son todos simultáneamente iguales a 0. Consideremos de nuevo el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Comenzamos examinando la hipótesis nula de que todos los coeficientes son simultáneamente iguales a cero:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

La aceptación de esta hipótesis nos llevaría a concluir que ninguna de las variables de predicción del modelo de regresión es estadísticamente significativa y, por lo tanto, que no suministran ninguna información útil. Si eso ocurriera, tendríamos que volver al proceso de especificación del modelo y desarrollar un nuevo conjunto de variables de predicción. Afortunadamente, en la mayoría de los casos aplicados esta hipótesis se rechaza porque el proceso de especificación normalmente lleva a la identificación de al menos una variable de predicción significativa.

Para contrastar la hipótesis anterior, podemos utilizar la descomposición de la variabilidad desarrollada en el apartado 13.3:

$$STC = SCR + SCE$$

Recuérdese que *SCR* es la cantidad de variabilidad explicada por la regresión y *SCE* es la cantidad de variabilidad no explicada. Recuérdese también que la varianza del modelo de regresión puede estimarse utilizando

$$s_e^2 = \frac{SCE}{(n - K - 1)}$$

Si la hipótesis nula de que todos los coeficientes son iguales a 0 es verdadera, entonces *el cuadrado medio de la regresión*

$$CMR = \frac{SCR}{K}$$

también es una medida del error con K grados de libertad. Como consecuencia, el cociente de

$$F = \frac{SCR/K}{SCE/(n - K - 1)}$$

$$= \frac{CMR}{s_e^2}$$

sigue una distribución F con K grados de libertad en el numerador y $n - K - 1$ grados de libertad en el denominador. Si la hipótesis nula es verdadera, tanto el numerador como el denominador son estimaciones de la varianza poblacional. Como señalamos en el apartado 11.4, el cociente entre las varianzas muestrales independientes de poblaciones que tienen varianzas poblacionales iguales sigue una distribución F si las poblaciones siguen una distribución normal. Se compara el valor calculado de F con el valor crítico de F de la Tabla 9 del apéndice a un nivel de significación α. Si el valor calculado es mayor que el valor crítico de la tabla, rechazamos la hipótesis nula y concluimos que al menos uno de los coeficientes no es igual a 0. Este método de contraste se resume en la ecuación 13.23.

Contraste de todos los parámetros de un modelo de regresión

Consideremos el modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Para contrastar la hipótesis nula

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$$

frente a la hipótesis alternativa

$$H_1 = \text{Al menos un } \beta_j \neq 0$$

a un nivel de significación α, utilizamos la regla de decisión

$$\text{Rechazar } H_0 : \text{ si } \quad \frac{CMR}{s_e^2} > F_{K, n-K-1, \alpha} \tag{13.23}$$

donde $F_{K, n-K-1, \alpha}$ es el valor crítico de F de la Tabla 9 del apéndice para el que

$$P(F_{K, n-K-1} > F_{K, n-K-1, \alpha}) = \alpha$$

La variable aleatoria calculada $F_{K, n-K-1}$ sigue una distribución F con K grados de libertad en el numerador y $(n - K - 1)$ grados de libertad en el denominador.

EJEMPLO 13.8. Modelo de predicción de los precios de la vivienda (contraste simultáneo de coeficientes)

Durante el desarrollo del modelo de predicción de los precios de la vivienda para Northern City, los analistas querían saber si existían pruebas de que la combinación de cuatro variables de predicción no era un predictor significativo del precio de la vivienda. Es decir, querían contrastar la hipótesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

Solución

Citydat

Este método de contraste puede ilustrarse mediante la regresión de los precios de la vivienda de la Figura 13.9 realizada utilizando el fichero de datos **Citydat**. En la tabla del análisis de la varianza, el estadístico F calculado es 19,19 con 4 grados de libertad en el numerador y 85 grados de libertad en el denominador. El cálculo de F es

$$F = \frac{259{,}37}{13{,}52} = 19{,}19$$

Este valor es más alto que el valor crítico de $F = 3{,}6$ para $\alpha = 0{,}01$ de la Tabla 9 del apéndice. Obsérvese, además, que el Minitab —y la mayoría de los paquetes estadísticos— calcula el p-valor, que en este ejemplo es igual a 0,000. Por lo tanto, rechazaríamos la hipótesis de que todos los coeficientes son iguales a cero.

Contraste de un subconjunto de coeficientes de regresión

En los apartados anteriores hemos desarrollado contrastes de hipótesis de parámetros de regresión individuales y de todos los parámetros en conjunto. A continuación, desarrollamos un contraste de hipótesis de un subconjunto de parámetros de regresión, como el ejemplo del conjunto de precios que acabamos de analizar. Utilizamos este contraste para averiguar si el efecto conjunto de varias variables independientes es significativo en un modelo de regresión.

Consideremos un modelo de regresión que contiene las variables independientes X_j y Z_j:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \alpha_1 z_{1i} + \cdots + \alpha_r z_{ri} + \varepsilon_i$$

La hipótesis nula que se contrasta es

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \quad \text{dados} \quad \beta_j \neq 0, \quad j = 1, ..., K$$

Si H_0 es verdadera, las variables Z_j no deben incluirse en el modelo de regresión porque no suministran ninguna información para explicar la conducta de la variable dependiente más que la que suministran las variables X_j. El método para realizar este contraste se resume en la ecuación 13.24 y se analiza detalladamente a continuación.

El contraste se realiza comparando la suma de los cuadrados de los errores, *SCE*, del modelo de regresión completo, que incluye tanto las variables *X* como las variables *Z*, con la *SCE*(*r*) de un modelo restringido que sólo incluye las variables *X*. Primero realizamos una regresión con respecto al modelo de regresión completo anterior y obtenemos la suma de los cuadrados de los errores, *SCE*. A continuación realizamos la regresión restringida, que excluye las variables *Z* (obsérvese que en esta regresión se aplica la restricción de que los coeficientes α_j son iguales a 0):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_j^*$$

A partir de esta regresión obtenemos la suma restringida de los cuadrados de los errores, *SCE*(*r*). A continuación, calculamos el estadístico *F* con *r* grados de libertad en el numerador (*r* es el número de variables eliminadas simultáneamente del modelo restringido) y $n - K - r - 1$ grados de libertad en el denominador (los grados de libertad del error en el modelo que incluye tanto las variables independientes *X* como *Z*). El estadístico *F* es

$$F = \frac{(SCE(r) - SCE)/r}{s_e^2}$$

donde s_e^2 es la varianza estimada del error del modelo completo. Este estadístico sigue una distribución *F* con *r* grados de libertad en el numerador y $n - K - r - 1$ grados de libertad en el denominador. Si el valor de *F* calculado es mayor que el valor crítico de *F*, entonces se rechaza la hipótesis nula y concluimos que las variables *Z* como conjunto deben incluirse en el modelo. Obsérvese que este contraste no implica que las variables *Z* individuales no deban excluirse, por ejemplo, utilizando el contraste *t* de Student antes analizado. Además, el contraste para todas las *Z* no implica que no pueda excluirse un subconjunto de las variables *Z* utilizando este método de contraste con un subconjunto diferente de variables *Z*.

Contraste de un subconjunto de los parámetros de regresión

Dado un modelo de regresión con la descomposición de las variables independientes en los subconjuntos *X* y *Z*,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \alpha_1 z_{1i} + \cdots + \alpha_r z_{ri} + \varepsilon_i$$

Para contrastar la hipótesis nula

$$H_0 : \alpha_1 = \alpha_2 = \cdots\cdots = \alpha_r = 0$$

de que los parámetros de regresión de un subconjunto son simultáneamente iguales a 0, frente a la hipótesis alternativa

$$H_1 : \text{Al menos un } \alpha_j \neq 0 \quad (j = 1, ..., r)$$

comparamos la suma de los cuadrados de los errores del modelo completo con la suma de los cuadrados de los errores del modelo restringido. Primero, hacemos una regresión para el modelo completo, que incluye todas las variables independientes, y obtenemos la suma de los cuadrados de los errores, *SCE*. A continuación, hacemos una regresión restringida, que excluye las variables *Z* cuyos coeficientes son las α: el número de variables excluidas es *r*. A partir de esta regresión obtenemos la suma restringida de los cuadrados de los errores, *SCE*(*r*). A continuación, calculamos el estadístico *F* y aplicamos la regla de decisión para el nivel de significación α:

$$\text{Rechazar } H_0 \text{ si } \quad \frac{(SCE(r) - SCE)/r}{s_e^2} > F_{r, n-K-r-1, \alpha} \qquad \textbf{(13.24)}$$

Comparación de los contrastes *F* y *t*

Si utilizáramos la ecuación 13.24 con $r = 1$, podríamos contrastar la hipótesis de que una única variable, X_j, no mejora la predicción de la variable dependiente, dadas las demás variables independientes del modelo. Por lo tanto, tenemos el contraste de hipótesis

$$H_0 : \beta_j = 0 \,|\, \beta_l \neq 0, j \neq 1$$
$$H_1 : \beta_j \neq 0 \,|\, \beta_l \neq 0, j \neq 1$$

Antes hemos visto que este contraste también podía realizarse utilizando un contraste *t* de Student. Utilizando métodos que no presentamos en este libro, podemos demostrar que los contrastes *F* y *t* correspondientes permiten llegar exactamente a las mismas conclusiones sobre el contraste de hipótesis de una única variable. Además, el estadístico *t* calculado para el coeficiente b_j es igual a la raíz cuadrada del estadístico *F* calculado correspondiente. Es decir,

$$t_{b_j} = \sqrt{F_{x_j}}$$

donde F_{x_j} es el estadístico *F* calculado utilizando la ecuación 13.24 cuando se excluye la variable x_j del modelo y, por lo tanto, $r = 1$. Demostramos este resultado numérico en el ejemplo 13.9.

La teoría estadística de la distribución también demuestra que una variable aleatoria *F* con 1 grado de libertad en el numerador es el cuadrado de una variable aleatoria *t* cuyos grados de libertad son iguales al denominador de la variable aleatoria *F*. Por lo tanto, los contrastes *F* y *t* siempre llevan a las mismas conclusiones sobre el contraste de hipótesis de una única variable independiente en un modelo de regresión múltiple.

EJEMPLO 13.9. Predicción del precio de la vivienda en las pequeñas ciudades (contrastes de hipótesis de subconjuntos de coeficientes)

Los promotores del modelo de predicción del precio de la vivienda del ejemplo 13.8 querían averiguar si el efecto conjunto del tipo impositivo y del porcentaje de locales comerciales contribuye a la predicción después de incluir previamente los efectos del tamaño de la vivienda y de la renta.

Solución

Continuando con el problema de los ejemplos 13.7 y 13.8, tenemos un contraste condicionado de la hipótesis de que dos variables no son predictores significativos, dado que las otras dos son predictores significativos:

$$H_0: \beta_3 = \beta_4 = 0 \,|\, \beta_1,\ \beta_2 \neq 0$$

Este contraste se realiza utilizando el método de la ecuación 13.24. La Figura 13.9 presenta la regresión del modelo completo con las cuatro variables de predicción. En esa regresión, $SCE = 1.149{,}14$. En la Figura 13.11 tenemos la regresión reducida en la que las únicas variables de predicción son el tamaño de la vivienda y la renta. En esa regresión, $SCE = 1.426{,}93$. La hipótesis se contrasta primero calculando el estadístico F cuyo numerador es la suma de los cuadrados de los errores del modelo reducido [$SCE(r)$] menos la SCE del modelo completo.

$$F = \frac{(1.426{,}93 - 1.149{,}14)/2}{13{,}52} = 10{,}27$$

Regression Analysis: hseval versus sizehse, income72

```
The regression equation is
hseval = -42.2 + 91.4 sizehse + 0.000393 incom72



Predictor        Coef    SE Coef        T       P
Constant      -42.208      9.810    -4.30   0.000
Sizehse         9.135      1.940     4.71   0.000
incom72      0.003927   0.001473     2.67   0.009


S = 4.04987    R-Sq = 34.7%   R-Sq(adj) = 33.2%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         2    759.70    379.85   23.16   0.000
Residual Error    87   1426.93     16.40
Total             89   2186.63


Source     DF    Seq SS
sizehse     1    643.12
incom72     1    116.58
```

SCE(*r*)

Figura 13.11. Regresión del precio de la vivienda: modelo reducido (salida Minitab).

El estadístico F tiene 2 grados de libertad —correspondientes a las dos variables contrastadas simultáneamente— en el numerador y 85 grados de libertad en el denominador. Obsérvese que el estimador de la varianza, $s_e^2 = 13{,}52$, se obtiene a partir del modelo completo de la Figura 13.9, en la que el error tiene 85 grados de libertad. Vemos en la Tabla 9 del apéndice que el valor crítico de F con $\alpha = 0{,}01$ y 2 y 85 grados de libertad es aproximadamente 4,9. Como el valor calculado de F es mayor que el valor crítico, rechazamos la hipótesis nula de que el tipo impositivo y el porcentaje de locales comerciales no están en la combinación significativa. El efecto conjunto de estas dos variables sí mejora el modelo que predice el precio de la vivienda. Por lo tanto, el tipo impositivo y el porcentaje de locales comerciales deben incluirse en el modelo.

También hemos calculado esta regresión excluyendo la variable «compr» y hemos observado que la *SCE* resultante era

$$SCE(1) = 1.185{,}29$$

El estadístico F calculado de esta variable era

$$F = \frac{(1.185{,}29 - 1.149{,}14)/1}{13{,}52} = 2{,}674$$

La raíz cuadrada de 2,674 es 1,64, que es el estadístico t calculado para la variable «compr» en la salida del análisis de regresión de la Figura 13.9. Utilizando el estadístico F calculado o el estadístico t calculado, obtendríamos este resultado para las hipótesis de esta variable:

$$H_0 : \beta_{compr} = 0 \,|\, \beta_l \neq 0,\ l \neq compr$$
$$H_1 : \beta_{compr} \neq 0 \,|\, \beta_l \neq 0,\ l \neq compr$$

EJERCICIOS

Ejercicios básicos

13.37. Suponga que ha estimado coeficientes para el siguiente modelo de regresión:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Contraste la hipótesis de que las tres variables de predicción son iguales a 0, dadas las siguientes tablas del análisis de la varianza.

a) Análisis de la varianza

Source	DF	SS	MS
Regression	3	4500	
Residual Error	26	500	

b) Análisis de la varianza

Source	DF	SS	MS
Regression	3	9780	
Residual Error	26	2100	

c) Análisis de la varianza

Source	DF	SS	MS
Regression	3	46000	
Residual Error	26	25000	

d) Análisis de la varianza

Source	DF	SS	MS
Regression	3	87000	
Residual Error	26	48000	

Ejercicios aplicados

13.38. Vuelva al estudio del esfuerzo de diseño de aviones de los ejercicios 13.6 y 13.19.

a) Contraste la hipótesis nula

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

b) Muestre la tabla del análisis de la varianza.

13.39. Para el estudio de la influencia de las instituciones financieras en los precios de las acciones del ejercicio 13.7, se utilizaron 48 observaciones trimestrales y se observó que el coeficiente corregido de determinación era $R^2 = 0{,}463$. Contraste la hipótesis nula.

$$H_0: \beta_1 = \beta_2 = 0$$

13.40. Vuelva al estudio del consumo de leche, descrito en los ejercicios 13.8, 13.20 y 13.28.

a) Contraste la hipótesis nula

$$H_0: \beta_1 = \beta_2 = 0$$

b) Muestre la tabla del análisis de la varianza.

13.41. Vuelva al estudio del aumento de peso, descrito en los ejercicios 13.9, 13.21 y 13.29.

a) Contraste la hipótesis nula

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

b) Muestre la tabla del análisis de la varianza.

13.42. Vuelva al ejercicio 13.32. Contraste la hipótesis nula de que las cuatro variables independientes, consideradas en conjunto, no influyen linealmente en los ingresos generados por las loterías nacionales.

13.43. Vuelva al ejercicio 13.33. Contraste la hipótesis nula de que las tres variables independientes, consideradas en conjunto, no influyen linealmente en el precio de los hornos.

13.44. Vuelva al estudio del ejercicio 13.34. Contraste la hipótesis nula de que los gastos personales de consumo y el precio relativo de las importaciones, considerados en conjunto, no afectan linealmente a la demanda nigeriana de importaciones.

13.45. Vuelva al estudio de los determinantes de la demanda de bomberos en una ciudad analizado en el ejercicio 13.36. Contraste la hipótesis nula

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

e interprete sus resultados.

13.46. Se realiza una regresión de una variable dependiente con respecto a K variables independientes utilizando n conjuntos de observaciones muestrales. *SCE* es la suma de los cuadrados de los errores y R^2 es el coeficiente de determinación de esta regresión estimada. Queremos contrastar la hipótesis nula de que K_1 de estas variables independientes, consideradas en conjunto, no afectan linealmente a la variable dependiente, dado que las demás variables independientes $(K - K_1)$ también se utilizan. Supongamos que se vuelve a estimar la regresión excluyendo las K_1 variables independientes de interés. Sea SCE^* la suma de los cuadrados de los errores y R^{*2} el coeficiente de determinación de esta regresión. Demuestre que el estadístico para contrastar nuestra hipótesis nula, introducido en el apartado 13.5, puede expresarse de la forma siguiente:

$$\frac{(SCE^* - SCE)/K_1}{SCE/(n - K - 1)} = \frac{R^2 - R^{*2}}{1 - R^2} \cdot \frac{n - K - 1}{K_1}$$

13.47. En el estudio de los ejercicios 13.8, 13.20 y 13.28 sobre el consumo de leche, se añadió al modelo de regresión una tercera variable independiente: el número de niños en edad preescolar que había en el hogar. Cuando se estimó este modelo ampliado, se observó que la suma de los cuadrados de los errores era 83,7. Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, el número de niños en edad preescolar que hay en el hogar no afecta linealmente al consumo de leche.

13.48. Suponga que una variable dependiente está relacionada con K variables independientes a través de un modelo de regresión múltiple. Sea R^2 el coeficiente de determinación y $\bar{R}^2$ el coeficiente corregido. Suponga que se utilizan n conjuntos de observaciones para ajustar la regresión.

a) Demuestre que

$$\bar{R}^2 = \frac{(n - 1)R^2 - K}{n - K - 1}$$

b) Demuestre que

$$R^2 = \frac{(n - K - 1)\bar{R}^2 + K}{n - 1}$$

c) Demuestre que el estadístico para contrastar la hipótesis nula de que todos los coeficientes de regresión son 0 puede expresarse de la forma siguiente:

$$\frac{SCR/K}{SCE/(n - K - 1)} = \frac{n - K - 1}{K} \cdot \frac{\bar{R}^2 + A}{1 - \bar{R}^2}$$

donde

$$A = \frac{K}{n - K - 1}$$

13.6. Predicción

Una aplicación importante de los modelos de regresión es predecir los valores de la variable dependiente, dados los valores de las variables independientes. Las predicciones pueden realizarse directamente a partir del modelo de regresión estimado utilizando las estimaciones de los coeficientes de ese modelo, como muestra la ecuación 13.25.

Predicciones a partir de los modelos de regresión múltiple

Dado que se cumple el modelo de regresión poblacional

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i \quad (i = 1, 2, ..., n)$$

y que los supuestos habituales del análisis de regresión son válidos, sean b_0, b_1, ..., b_K las estimaciones por mínimos cuadrados de los coeficientes del modelo, β_j, siendo $j = 1, ..., K$, basados en los puntos de datos x_{1i}, x_{2i}, ..., x_{Ki} ($i = 1, ..., n$). En tal caso, dada una nueva observación de un punto de datos, $x_{1,n+1}$, $x_{2,n+1}$, ..., $x_{K,n+1}$, la mejor predicción lineal insesgada de $\hat{y}_{n+1}$ es

$$\hat{y}_{n+1} = b_0 + b_1 x_{1,n+1} + b_{2,n+1} x_{2,n+1} + \cdots + b_K x_{K,n+1} \qquad \textbf{(13.25)}$$

Es muy arriesgado hacer predicciones que se basan en valores de *X* fuera del rango de los datos utilizados para estimar los coeficientes del modelo, ya que no tenemos pruebas que apoyen el modelo lineal en esos puntos.

Además de querer conocer el valor predicho de Y para un conjunto de x_j, a menudo nos interesa calcular un intervalo de confianza o un intervalo de predicción. Como señalamos en el apartado 12.6, el intervalo de confianza incluye el valor esperado de Y con la probabilidad $1 - \alpha$. En cambio, el intervalo de predicción incluye los valores individuales predichos: los valores esperados de Y más el término de error aleatorio. Para hallar estos intervalos, es necesario calcular estimaciones de las desviaciones típicas del valor esperado de Y y los puntos individuales. Estos cálculos son similares en la forma a los utilizados en la regresión simple, pero las ecuaciones de los estimadores son mucho más complicadas. Las desviaciones típicas de los valores predichos, $s_{\hat{y}}$, son una función del error típico de la estimación, s_e; la desviación típica de las variables de predicción; las correlaciones entre las variables de predicción; y el cuadrado de la distancia entre la media de las variables independientes y las X para la predicción. Esta desviación típica es similar a la desviación típica de las predicciones de la regresión simple del Capítulo 12. Sin embargo, las ecuaciones de la regresión múltiple son muy complejas y no se presentan aquí; lo que hacemos es calcular los valores utilizando el programa Minitab. La mayoría de los paquetes estadísticos buenos calculan las desviaciones típicas del intervalo de predicción y del intervalo de confianza y los correspondientes intervalos. Excel no permite calcular la desviación típica de las variables predichas.

EJEMPLO 13.10. Predicción del margen de beneficios de las asociaciones de ahorro y crédito inmobiliario (predicciones del modelo de regresión)

Le han pedido que haga una predicción del margen de beneficios de las asociaciones de ahorro y crédito inmobiliario para un año en el que el porcentaje de ingresos netos es

Savings and Loan

4,50 y hay 9.000 oficinas, utilizando el modelo de regresión de las asociaciones de ahorro y crédito inmobiliario. Los datos se encuentran en el fichero **Savings and Loan**.

Solución

Utilizando la notación de la ecuación 13.25, tenemos las variables

$$x_{1,n+1} = 4{,}50 \qquad x_{2,n+1} = 9.000$$

Utilizando estos valores, observamos que nuestro predictor puntual del margen de beneficios es

$$\begin{aligned}\hat{y}_{n+1} &= b_0 + b_1 x_{1,n+1} + b_2 x_{2,n+1} \\ &= 1{,}565 + (0{,}237)(4{,}50) - (0{,}000249)(9.000) = 0{,}39\end{aligned}$$

Por lo tanto, en un año en el que el porcentaje de ingresos netos por dólar depositado es 4,50 y el número de oficinas es 9.000, predecimos que el margen porcentual de beneficios de las asociaciones de ahorro y crédito inmobiliario es 0,39.

Los valores predichos, los intervalos de confianza y los intervalos de predicción pueden calcularse directamente por medio del programa Minitab.

La Figura 13.12 muestra la salida del análisis de regresión. Se presenta el valor predicho, $\hat{y} = 0{,}39$ y su desviación típica, 0,0277, junto con el intervalo de confianza y el

Regression Analysis: Y profit versus X1 revenue, X2 offices

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue - 0.000249 X2 offices


Predictor          Coef     SE Coef      T      P
Constant        1.56450     0.07940  19.70  0.000
X1 revenue      0.23720      0.5556   4.27  0.000
X2 offices -0.00024908  0.00003205  -7.77  0.000


S = 0.0533022   R-Sq = 86.5%   R-Sq(adj) = 85.3%


Analysis of Variance

Source         DF       SS       MS      F       P
Regression      2  0.40151  0.20076  70.66  0.000
Residual Error 22  0.06250  0.00284
Total          24  0.46402


Predicted Values for New Observations

New
Obs     Fit   SE Fit          95% CI          95% PI
  1  0.3902   0.0277  (0.3327, 0.4476)  (0.2656, 0.5148)


Values of Predictors for New Observations

New
Obs  X1 revenue  X2 offices
  1        4.50        9000
```

Instrucciones de Minitab
1. STAT > REGRESSION > REGRESSION > OPTIONS
2. Pulsar New Observation Values
3. Seleccionar Fits, Confidence limits, Prediction limits

Valor predicho

Error típico del valor predicho

Intervalo de predicción

Intervalo de confianza

Valores de las variables de predicción

Figura 13.12. Predicciones e intervalos de predicción de la regresión múltiple (salida Minitab).

intervalo de predicción. El intervalo de confianza —CI— es un intervalo del valor esperado de Y en la función lineal definida por los valores de las variables independientes. Este intervalo es una función del error típico del modelo de regresión, la distancia a la que se encuentran los valores de x_j de sus medias muestrales individuales y la correlación entre las variables x_j utilizadas para ajustar el modelo. El intervalo de predicción —PI— es un intervalo para un único valor observado. Por lo tanto, incluye la variabilidad del valor esperado más la variabilidad de un único punto en torno al valor predicho.

EJERCICIOS

Ejercicios básicos

13.49. Dada la ecuación de regresión múltiple estimada

$$\hat{y} = 6 + 5x_1 + 4x_2 + 7x_3 + 8x_4$$

calcular el valor predicho de Y cuando

a) $x_1 = 10$, $x_2 = 23$, $x_3 = 9$, $x_4 = 12$
b) $x_1 = 23$, $x_2 = 18$, $x_3 = 10$, $x_4 = 11$
c) $x_1 = 10$, $x_2 = 23$, $x_3 = 9$, $x_4 = 12$
d) $x_1 = -10$, $x_2 = 13$, $x_3 = -8$, $x_4 = -16$

Ejercicios aplicados

13.50. Utilizando la información del ejercicio 13.9, prediga el aumento de peso de un estudiante de primer año que come una media de 20 comidas a la semana, hace ejercicio durante una media de 10 horas a la semana y consume una media de 6 cervezas a la semana.

13.51. Utilizando la información del ejercicio 13.8, prediga el consumo semanal de leche de una familia de cuatro personas que tiene una renta de 600 $ a la semana.

$$b_0 = 0{,}578$$

13.52. En la regresión del esfuerzo de diseño de aviones del ejercicio 13.6, la ordenada en el origen estimada era 2,0. Prediga el esfuerzo de diseño de un avión que tiene una velocidad máxima de mach 1,0 pesa 7 toneladas y tiene un 50 por ciento de piezas en común con otros modelos.

13.53. Una agencia inmobiliaria afirma que en su ciudad el precio de venta de una vivienda en dólares (y) depende de su tamaño en metros cuadrados de superficie (x_1), el tamaño del solar en metros cuadrados (x_2), el número de dormitorios (x_3) y el número de cuartos de baño (x_4). Basándose en una muestra aleatoria de 20 ventas de viviendas, se obtuvo el siguiente modelo estimado por mínimos cuadrados:

$$\hat{y} = 1.998{,}5 + \underset{(2{,}5543)}{22{,}352x_1} + \underset{(1{,}4492)}{1{,}4686x_2} + \underset{(1820{,}8)}{6.767{,}3x_3} + \underset{(1996{,}2)}{2.701{,}1x_4} \qquad R^2 = 0{,}9843$$

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete en el contexto de este modelo el coeficiente estimado de x_2.
b) Interprete el coeficiente de determinación.
c) Suponiendo que el modelo está especificado correctamente, contraste al nivel del 5 por ciento la hipótesis nula de que, manteniéndose todo lo demás constante, el precio de venta no depende del número de cuartos de baño frente a la hipótesis alternativa unilateral adecuada.
d) Estime el precio de venta de una vivienda de 1.250 metros cuadrados de superficie, un solar de 4.700 metros cuadrados, 3 dormitorios y un cuarto de baño y medio.

13.7. Transformaciones de modelos de regresión no lineales

Hemos visto cómo puede utilizarse el análisis de regresión para estimar relaciones lineales que predicen una variable dependiente en función de una o más variables independientes. Estas aplicaciones son muy importantes. Sin embargo, hay, además, algunas relaciones económicas y empresariales que no son estrictamente lineales. En este apartado desarrolla-

mos métodos para modificar algunos formatos de los modelos no lineales con el fin de poder utilizar los métodos de regresión múltiple para estimar los coeficientes del modelo. Por lo tanto, el objetivo de los apartados 13.7 y 13.8 es ampliar la variedad de problemas que pueden adaptarse a un análisis de regresión. De esta forma vemos que el análisis de regresión tiene aun mayores aplicaciones.

Examinando el algoritmo de mínimos cuadrados, vemos que manipulando con cuidado los modelos no lineales, es posible utilizar los mínimos cuadrados en un conjunto más amplio de problemas aplicados. Los supuestos sobre las variables independientes en la regresión múltiple no son muy restrictivos. Las variables independientes definen puntos en los que medimos una variable aleatoria Y. Suponemos que hay una relación lineal entre los niveles de las variables independientes X_j, donde $j = 1, ..., K$, y el valor esperado de la variable dependiente Y. Podemos aprovechar esta libertad para ampliar el conjunto de modelos que pueden estimarse. Por lo tanto, podemos ir más allá de los modelos lineales en nuestras aplicaciones del análisis de regresión múltiple. En la Figura 13.13 se muestran tres ejemplos:

(a) Las funciones de oferta pueden no ser lineales.
(b) El aumento de la producción total con un aumento del número de trabajadores puede ser cada vez menor a medida que se añaden más trabajadores.
(c) El coste medio por unidad producida a menudo se minimiza en un nivel de producción intermedio.

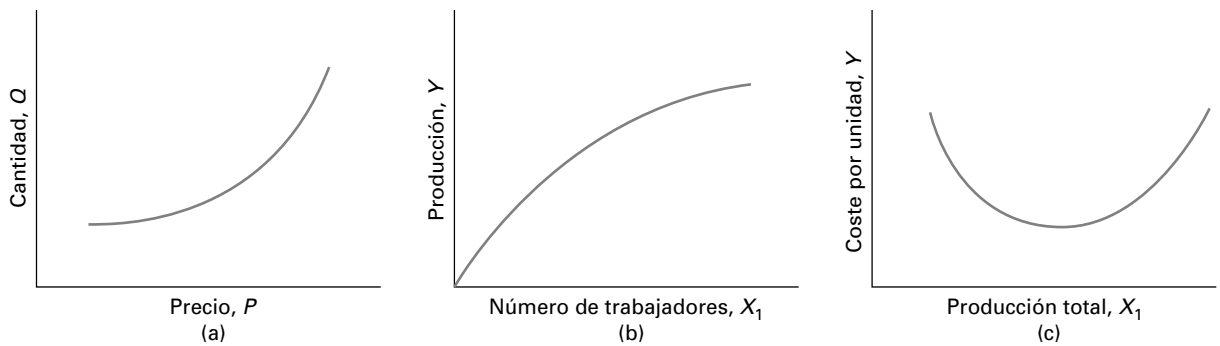


Figura 13.13. Ejemplos de funciones cuadráticas.

Transformaciones de modelos cuadráticos

Hemos dedicado bastante tiempo al desarrollo del análisis de regresión para estimar ecuaciones lineales que representan diversos procesos empresariales y económicos. También hay muchos procesos que pueden representarse mejor mediante ecuaciones no lineales. El ingreso total tiene una relación cuadrática con el precio y el ingreso máximo se obtiene en un nivel intermedio de precios si la función de demanda tiene pendiente negativa. En muchos casos, el coste mínimo de producción por unidad se obtiene en un nivel de producción intermedio y el coste por unidad es decreciente a medida que nos aproximamos al coste mínimo por unidad y después aumenta a partir de ese coste mínimo por unidad. Podemos analizar algunas de estas relaciones económicas y empresariales utilizando un modelo cuadrático:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$$

Para estimar los coeficientes de un modelo cuadrático para aplicaciones de este tipo, podemos transformar o modificar las variables, como muestran las ecuaciones 13.26 y 13.27. De esta forma, un modelo cuadrático no lineal se convierte en un modelo que es lineal en un conjunto modificado de variables.

Transformaciones de modelos cuadráticos

La función cuadrática

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon \qquad \textbf{(13.26)}$$

puede transformarse en un modelo lineal de regresión múltiple definiendo nuevas variables:

$$Z_1 = X_1$$
$$Z_2 = X_1^2$$

y después especificando el modelo

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i \qquad \textbf{(13.27)}$$

que es lineal en las variables transformadas. Las variables cuadráticas transformadas pueden combinarse con otras variables en un modelo de regresión múltiple. Por lo tanto, podemos ajustar una regresión cuadrática múltiple utilizando variables transformadas. El objetivo es encontrar modelos que sean lineales en otras formas matemáticas de una variable.

Transformando las variables, podemos estimar un modelo lineal de regresión múltiple y utilizar los resultados como un modelo no lineal. Los métodos de inferencia para los modelos cuadráticos transformados son los mismos que hemos desarrollado para los modelos lineales. De esta forma, evitamos la confusión que se tendría si se utilizaran unos métodos estadísticos para los modelos lineales y otros para los modelos cuadráticos. Los coeficientes deben combinarse para poder interpretarlos. Así, si tenemos un modelo cuadrático, el efecto de una variable, X, es indicado por los coeficientes tanto de los términos lineales como de los términos cuadráticos. También realizamos un sencillo contraste de hipótesis para averiguar si un modelo cuadrático es una mejora con respecto a un modelo lineal. La variable Z_2 o X_1^2 no es más que una variable adicional cuyo coeficiente puede contrastarse —$H_0: \beta_2 = 0$— utilizando la t de Student condicionada o el estadístico F. Si un modelo cuadrático se ajusta a los datos mejor que un modelo lineal, el coeficiente de la variable cuadrática —$Z_2 = X_1^2$— será significativamente diferente de 0. El método es el mismo si tenemos variables como $Z_3 = X_1^3$ o $Z_4 = X_1^2 X_2$.

EJEMPLO 13.11. Costes de producción (estimación de un modelo cuadrático)

Production Cost

Arnold Sorenson, director de producción de New Frontiers Instruments Inc., tenía interés en estimar la relación matemática entre el número de montajes electrónicos producidos en un turno de 8 horas y el coste medio por montaje. Esta función se utilizaría después para estimar el coste de varios pedidos de producción y averiguar el nivel de producción que minimizaría el coste medio. Los datos se encuentran en el fichero de datos **Production Cost**.

Solución

Arnold recogió datos de nueve turnos durante los cuales el número de montajes osciló entre 100 y 900. También obtuvo en el departamento de contabilidad el coste medio por unidad en que se incurrió durante esos días. Estos datos se presentan en un diagrama de puntos dispersos realizado por medio del programa Excel y mostrado en la Figura 13.14. Sus estudios de economía y su experiencia lo llevaron a sospechar que la función podría ser cuadrática con un coste medio mínimo intermedio. Diseñó su análisis para considerar tanto una función de coste medio de producción lineal como una cuadrática.

La Figura 13.15 es la regresión simple del coste como una función lineal del número de unidades. Vemos que la relación lineal es casi plana, lo que indica que no existe una relación lineal entre el coste medio y el número de unidades producidas. Si Arnold hubiera utilizado simplemente esta relación, habría cometido graves errores en sus métodos de estimación del coste.

La Figura 13.16 presenta la regresión cuadrática que muestra el coste medio por unidad como una función no lineal del número de unidades producidas. Obsérvese que b_2 es diferente de 0 y, por lo tanto, debe incluirse en el modelo. Obsérvese también que el R^2 del modelo cuadrático es 0,962, mientras que en el modelo lineal es 0,174. Utilizando el modelo cuadrático, Arnold ha elaborado un modelo de coste medio mucho más útil.

Número de Unidades	Coste medio por unidad
100	5,11
210	4,42
290	4,07
415	3,52
509	3,33
613	3,44
697	3,77
806	4,07
908	4,28

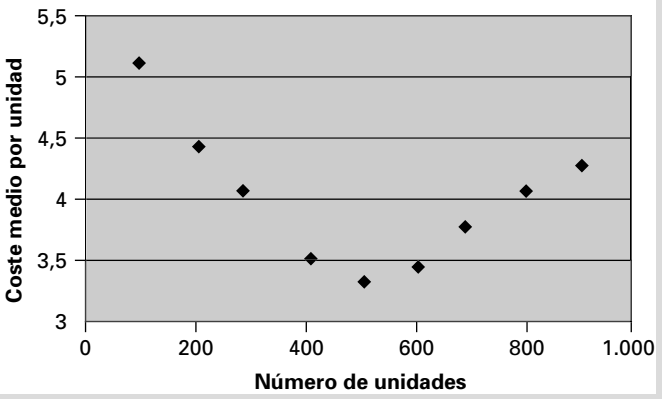


Figura 13.14. Coste medio de producción en función del número de unidades.

Regression Analysis: Mean Cost per Unit versus Number of Units

```
The regression equation is
Mean Cost per Unit = 4.43 - 0.000855 Number of Units


Predictor             Coef      SE Coef       T       P
Constant            4.4330       0.3994   11.10   0.000
Number of Units  -0.0008547    0.0007029   -1.22   0.263


S = 0.547614   R-Sq = 17.4%   R-Sq(adj) = 5.6%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         1    0.4433    0.4433    1.48   0.263
Residual Error     7    2.0992    0.2999
Total              8    2.5425
```

Figura 13.15. Regresión lineal del coste medio en función del número de unidades (salida Minitab).

Regression Analysis: Mean Cost per Unit versus Number of Units, No Units Squared

```
The regression equation is
Mean Cost per Unit = 5.91 - 0.000884 Number of Units + 0.000008
No Units Squared


Predictor              Coef      SE Coef       T       P
Constant             5.9084       0.1614   36.60   0.000
Number of Units   -0.0088415    0.0007344  -12.04   0.000
No Units Squared -0.00000793   0.00000071   11.15   0.000

S = 0.126875   R-Sq = 96.2%   R-Sq(adj) = 94.9%

Analysis of Variance

Source           DF        SS        MS       F       P
Regression        2    2.4459    1.2230   75.97   0.000
Residual Error    6    0.0966    0.0161
Total             8    2.5425
```

Figura 13.16. Modelo cuadrático del coste medio en función del número de unidades (salida Minitab).

Transformaciones logarítmicas

Algunas relaciones económicas pueden analizarse mediante funciones exponenciales. Por ejemplo, si la variación porcentual de la cantidad vendida de bienes varía linealmente en respuesta a las variaciones porcentuales del precio, la función de demanda tendrá una forma exponencial:

$$Q = \beta_0 P^{\beta_1}$$

donde Q es la cantidad demandada y P es el precio por unidad. Las funciones de demanda exponenciales tienen elasticidad constante y, por lo tanto, una variación del precio de un 1 por ciento provoca la misma variación porcentual de la cantidad demandada en todos los niveles de precios. En cambio, los modelos lineales de demanda indican que una variación unitaria de la variable del precio provoca la misma variación de la cantidad demandada en todos los niveles de precios. Los modelos exponenciales de demanda se utilizan mucho en el análisis de la conducta del mercado. Una importante característica de estos modelos es que el coeficiente β_1 es la elasticidad constante, e, de la demanda Q con respecto al precio P:

$$e = \frac{\partial Q/Q}{\partial P/P} = \beta_1$$

Este resultado se desarrolla en la mayoría de los libros de texto de microeconomía. Los coeficientes del modelo exponencial se estiman utilizando transformaciones logarítmicas, como muestra la ecuación 13.29.

La transformación logarítmica supone que el término de error aleatorio multiplica el verdadero valor de Y para obtener el valor observado. Por lo tanto, en el modelo exponencial el error es un porcentaje del verdadero valor y la varianza de la distribución del error aumenta cuando aumenta Y. Si este resultado no es cierto, la transformación logarítmica no es correcta. En ese caso, debe utilizarse una técnica de estimación no lineal mucho más compleja. Estas técnicas están fuera del alcance de este libro.

Transformaciones de modelos exponenciales

Los coeficientes de los modelos exponenciales de la forma

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \varepsilon \tag{13.28}$$

pueden estimarse tomando primero el logaritmo de los dos miembros para obtener una ecuación que es lineal en los logaritmos de las variables:

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \log(\varepsilon) \tag{13.29}$$

Utilizando esta forma, podemos hacer una regresión del logaritmo de Y con respecto a los logaritmos de las dos variables X y obtener estimaciones de los coeficientes β_1 y β_2 directamente del análisis de regresión. Dado que los coeficientes son elasticidades, muchos economistas utilizan esta forma del modelo en la que pueden suponer que las elasticidades son constantes en el rango de los datos. Obsérvese que este método de estimación requiere que los errores aleatorios sean multiplicativos en el modelo exponencial original. Por lo tanto, el término de error, ε, se expresa como un aumento o una disminución porcentual y no como la adición o la sustracción de un error aleatorio, como hemos visto en los modelos lineales de regresión.

Otra importante aplicación de los modelos exponenciales es la función de producción Cobb-Douglas, que tiene la forma

$$Q = \beta_0 L^{\beta_1} K^{\beta_2}$$

donde Q es la cantidad producida, L es la cantidad utilizada de trabajo y K es la cantidad de capital. β_1 y β_2 son las contribuciones relativas de las variaciones del trabajo y de las variaciones del capital a las variaciones de la cantidad producida. En un caso especial, correspondiente a los rendimientos constantes de escala, se plantea la restricción de que la suma de los coeficientes sea igual a 1. En ese caso, β_1 y β_2 son las contribuciones porcentuales del trabajo y el capital al aumento de la productividad.

La estimación de los coeficientes cuando su suma es igual a 1 es un ejemplo de estimación restringida en los modelos de regresión. La ecuación 13.29 es modificada por la restricción

$$\beta_1 + \beta_2 = 1$$

y, por lo tanto, se incluye la sustitución de la forma

$$\beta_2 = 1 - \beta_1$$

y la nueva ecuación de estimación se convierte en

$$\begin{aligned} \log(Y) &= \log(\beta_0) + \beta_1 \log(X_1) + (1 - \beta_1) \log(X_2) + \log(\varepsilon) \\ \log(Y) - \log(X_2) &= \log(\beta_0) + \beta_1 [\log(X_1) - \log(X_2)] + \log(\varepsilon) \\ \log\left(\frac{Y}{X_2}\right) &= \log(\beta_0) + \beta_1 \log\left(\frac{X_1}{X_2}\right) + \log(\varepsilon) \end{aligned} \tag{13.30}$$

Vemos, pues, que el coeficiente β_1 se obtiene haciendo una regresión de $\log(Y/X_2)$ con respecto a $\log(X_1/X_2)$. A continuación, se calcula β_2 restando β_1 de 1,0.

Todos los buenos paquetes estadísticos pueden calcular fácilmente las transformaciones necesarias de los datos para los modelos logarítmicos. En el ejemplo siguiente utilizamos el programa Minitab, pero podrían obtenerse resultados similares utilizando otros muchos paquetes.

EJEMPLO 13.12. Función de producción de Minong Boat Works (estimación del modelo exponencial)

Minong Boat Works comenzó a producir pequeños barcos de pesca a principios de la década de 1970 para los pescadores del norte de Wisconsin. Sus propietarios desarrollaron un método de producción de bajo coste para producir barcos de calidad. Como consecuencia, ha aumentado su demanda con el paso de los años. El método de producción utiliza una terminal de trabajo con un conjunto de plantillas y herramientas eléctricas que pueden ser manejadas por un número variable de trabajadores. El número de terminales (unidades de capital) ha aumentado con el paso de los años de 1 a 20 para satisfacer la demanda de barcos. Al mismo tiempo, la plantilla se ha incrementado de 2 trabajadores al año a 25. Ahora los propietarios están considerando la posibilidad de aumentar sus ventas en otros mercados de Michigan y Minnesota. Por lo tanto, necesitan saber cuánto tienen que aumentar el número de terminales y el número de trabajadores para lograr diversos aumentos del nivel de producción.

Solución

Boat Production

Su hija, licenciada en economía, sugiere que estimen una función de producción Cobb-Douglas restringida utilizando datos de años anteriores. Explica que esta función de producción les permitirá predecir el número de barcos producidos con diferentes niveles de terminales y de trabajadores. Los propietarios están de acuerdo en que ese análisis es una buena idea y le piden que lo realice. Comienza el análisis recogiendo los datos históricos de producción de la empresa, que se encuentran en el fichero de datos **Boat Production**. Para estimar los coeficientes, primero debe transformar la especificación original del modelo en una forma que pueda estimarse mediante una regresión por mínimos cuadrados. El modelo de la función de producción Cobb-Douglas es

$$Y = \beta_0 L^{\beta_1} K^{\beta_2}$$

con la restricción

$$\beta_2 = 1 - \beta_1$$

donde Y es el número de barcos producidos al año, K es el número de terminales (unidades de capital) utilizadas cada año y L es el número de trabajadores utilizados cada año.

La función de producción Cobb-Douglas restringida se transforma en la forma de estimación:

$$\log\left(\frac{Y}{K}\right) = \log(\beta_0) + \beta_2 \log\left(\frac{L}{K}\right)$$

para hacer una estimación por mínimos cuadrados.

La estimación del modelo de regresión se muestra en la Figura 13.17 y la ecuación resultante es:

$$\log\left(\frac{Y}{K}\right) = 3{,}02 + 0{,}845 \log\left(\frac{L}{K}\right) \qquad \textbf{(13.31)}$$

En este resultado, vemos que el coeficiente del modelo estimado, b_1, es 0,845. Por lo tanto, $b_2 = 1 - 0{,}845 = 0{,}155$. Por último, $\log(b_0) = 3{,}02$. Este análisis muestra que el 84,5 por ciento del valor de la producción procede del trabajo y el 15,5 por ciento del

```
The regression equation is
logbotunit = 3.02 + 0.845 logworunit

Predictor        Coef      SE Coef        T        P
Constant      3.02325      0.04387    68.92    0.000
logworun      0.84479      0.09062     9.32    0.000

S = 0.1105     R-Sq = 79.8%    R-Sq(adj) = 78.9%

Analysis of Variance

Source            DF        SS        MS        F        P
Regression         1    1.0618    1.0618    86.90    0.000
Residual Error    22    0.2688    0.0122
Total             23    1.3306
```

Figura 13.17. Análisis de regresión de la función de producción restringida (salida Minitab).

capital. Tras realizar las oportunas transformaciones algebraicas, el modelo de la función de producción es

$$Y = 20{,}49K^{0{,}845}L^{0{,}155} \qquad \textbf{(13.32)}$$

Esta función de producción puede utilizarse para predecir la producción esperada utilizando diversos niveles de capital y de trabajo.

La Figura 13.18 muestra una comparación del número observado de barcos y el número predicho de barcos a partir de la ecuación de regresión transformada. El número predicho de barcos se ha calculado utilizando la ecuación 13.32. Ese análisis también indica que el R^2 de la regresión del número de barcos con respecto al número predicho de barcos es 0,973. Este R^2 puede interpretarse exactamente igual que el R^2 de cualquier modelo de regresión lineal y, por lo tanto, vemos que el número predicho de barcos constituye un buen ajuste de los datos observados sobre la producción de barcos. El R^2 de los datos de la regresión transformada de la Figura 13.17 no puede interpretarse fácilmente como un indicador de la relación entre el número de barcos producidos y las variables independientes del trabajo y el capital, ya que las unidades están expresadas en logaritmos de cocientes.

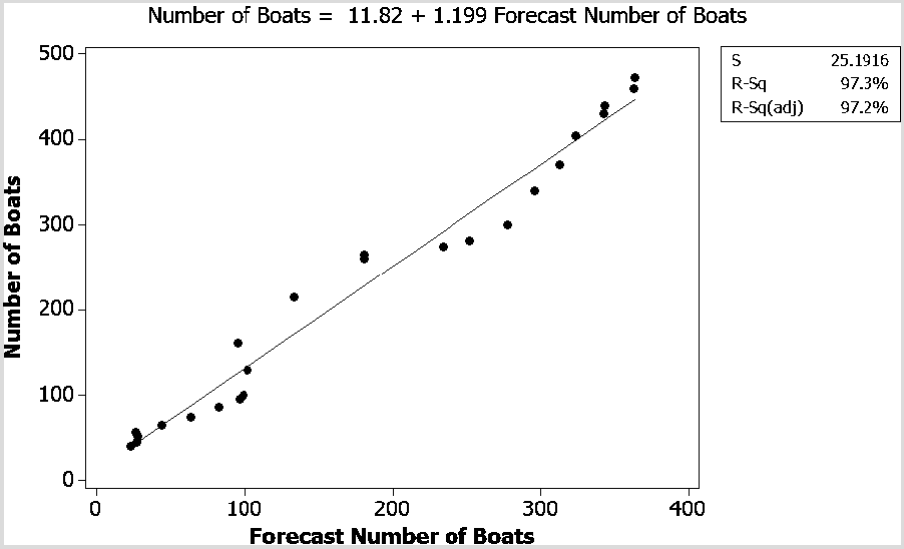


Figura 13.18. Comparación de la producción observada y la predicha.

EJERCICIOS

Ejercicios básicos

13.54. Considere las dos ecuaciones siguientes estimadas utilizando los métodos desarrollados en este apartado.

i. $y_i = 4x^{1,5}$
ii. $y_i = 1 + 2x_i + 2x_1^2$

Calcule los valores de y_i cuando $x_i = 1, 2, 4, 6, 8, 10$.

13.55. Considere las dos ecuaciones siguientes estimadas utilizando los métodos desarrollados en este apartado.

i. $y_i = 4x^{1,8}$
ii. $y_i = 1 + 2x_i + 2x_i^2$

Calcule los valores de y_i cuando $x_i = 1, 2, 4, 6, 8, 10$.

13.56. Considere las dos ecuaciones siguientes estimadas utilizando los métodos desarrollados en este apartado.

i. $y_i = 4x^{1,5}$
ii. $y_i = 1 + 2x_i + 1,7x_i^2$

Calcule los valores de y_i cuando $x_i = 1, 2, 4, 6, 8, 10$.

13.57. Considere las dos ecuaciones siguientes estimadas utilizando los métodos desarrollados en este apartado.

i. $y_i = 3x^{1,2}$
ii. $y_i = 1 + 5x_i + 1,5x_i^2$

Calcule los valores de y_i cuando $x_i = 1, 2, 4, 6, 8, 10$.

Ejercicios aplicados

13.58. Describa un ejemplo extraído de su experiencia en el que un modelo cuadrático sea mejor que un modelo lineal.

13.59. Juan Sánchez, presidente de Estudios de Mercado, S.A., le ha pedido que estime los coeficientes del modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2$$

donde Y son las ventas esperadas de suministros de oficina de un gran distribuidor minorista de suministros de oficina, X_1 es la renta total disponible de los residentes que viven a menos de 5 kilómetros de la tienda y X_2 es el número total de personas empleadas en empresas cuya actividad se basa en la información que se encuentran a menos de 5 kilómetros de la tienda. Según los estudios recientes de una consultora nacional, los coeficientes del modelo deben tener la siguiente restricción:

$$\beta_1 + \beta_2 = 2$$

Describa cómo estimaría los coeficientes del modelo utilizando el método de mínimos cuadrados.

13.60. En un estudio de los determinantes de los gastos de los hogares en viajes de vacaciones, se obtuvieron datos de una muestra de 2.246 hogares (véase la referencia bibliográfica). El modelo estimado era

$$\log y = \underset{}{-4,054} + \underset{(0,0546)}{1,1556 \log x_1} - \underset{(0,0490)}{0,4408 \log x_2}$$

$R^2 = 0,168$

donde

y = gasto en viajes de vacaciones
x_1 = gasto total anual de consumo
x_2 = número de miembros del hogar

Los números entre paréntesis que se encuentran debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete los coeficientes de regresión estimados.
b) Interprete el coeficiente de determinación.
c) Manteniéndose todo lo demás constante, halle el intervalo de confianza al 95 por ciento del aumento porcentual de los gastos en viajes de vacaciones provocado por un aumento del gasto anual total de consumo de un 1 por ciento.
d) Suponiendo que el modelo está especificado correctamente, contraste al nivel de significación del 1 por ciento la hipótesis nula de que, manteniéndose todo lo demás constante, el número de miembros de un hogar no afecta a los gastos en viajes de vacaciones frente a la hipótesis alternativa de que cuanto mayor es el número de miembros del hogar, menor es el gasto en viajes de vacaciones.

13.61. En un estudio, se estimó el siguiente modelo para una muestra de 322 supermercados de grandes zonas metropolitanas (véase la referencia bibliográfica 3):

$$\text{Log } y = \underset{}{2,921} + \underset{(0,077)}{0,680 \log x} \qquad R^2 = 0,19$$

donde

y = tamaño de la tienda
x = renta mediana del distrito postal en el que se encuentra la tienda

Los números entre paréntesis que figuran debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete el coeficiente estimado de $\log x$.
b) Contraste la hipótesis nula de que la renta no influye en el tamaño de la tienda frente a la hipótesis alternativa de que un aumento de la renta tiende a ir acompañado de un aumento del tamaño de la tienda.

13.62. Un economista agrícola cree que la cantidad consumida de carne de vacuno (y) en toneladas al año en Estados Unidos depende de su precio (x_1) en dólares por kilo, del precio de la carne de porcino (x_2) en dólares por kilo, del precio del pollo (x_3) en dólares por kilo y de la renta por hogar (x_4) en miles de dólares. Se ha obtenido la siguiente regresión muestral por mínimos cuadrados utilizando 30 observaciones anuales:

$$\log y = \underset{}{-0{,}024} \underset{(0{,}168)}{-0{,}529 \log x_1} \underset{(0{,}103)}{+0{,}217 \log x_2} \underset{(0{,}106)}{+0{,}193 \log x_3} \underset{(0{,}163)}{+0{,}416 \log x_4} \qquad R^2 = 0{,}683$$

Los números entre paréntesis que se encuentran debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete el coeficiente de $\log x_1$.
b) Interprete el coeficiente de $\log x_2$.
c) Contraste al nivel de significación del 1 por ciento la hipótesis nula de que el coeficiente de $\log x_4$ en la regresión poblacional es 0 frente a la hipótesis alternativa de que es positivo.
d) Contraste la hipótesis nula de que las cuatro variables ($\log x_1$, $\log x_2$, $\log x_3$, $\log x_4$) no tienen, en conjunto, ninguna influencia lineal en $\log y$.
e) Al economista también le preocupa que la creciente concienciación de las consecuencias del consumo frecuente de carne roja para la salud pueda haber influido en la demanda de carne de vacuno. Si eso es así, ¿cómo influiría en su opinión sobre la regresión estimada original?

13.63. Le han pedido que desarrolle una función de producción exponencial —forma Cobb-Douglas— que prediga el número de microprocesadores producidos por un fabricante, Y, en función de las unidades de capital, X_1; las unidades de trabajo, X_2, y el número de informáticos que realizan investigación básica, X_3. Especifique la forma del modelo e indique con cuidado y exhaustivamente cómo estimaría los coeficientes. Hágalo utilizando primero un modelo sin restricciones y a continuación incluyendo la restricción de que los coeficientes de las tres variables deben sumar 1.

13.64. Considere el siguiente modelo no lineal con errores multiplicativos.

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} X_3^{\beta_3} X_4^{\beta_4} \varepsilon$$
$$\beta_1 + \beta_2 = 1$$
$$\beta_3 + \beta_4 = 1$$

a) Muestre cómo obtendría estimaciones de los coeficientes. Deben satisfacerse las restricciones de los coeficientes. Muestre todo lo que hace y explíquelo.
b) ¿Cuál es la elasticidad constante de Y con respecto a X_4?

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

13.65. Angelica Chandra, presidenta de Benefits Research Inc., le ha pedido que estudie la estructura salarial de su empresa. Benefits Research ofrece consultoría y gestión de los programas de seguro médico y de jubilación para los empleados. Sus clientes son grandes y medianas empresas. Primero le pide que desarrolle un modelo de regresión que estime el salario esperado en función de los años de experiencia en la empresa. Debe examinar modelos lineales, cuadráticos y cúbicos y averiguar cuál es más adecuado. Estime modelos de regresión adecuados y escriba un breve informe que recomiende el mejor modelo. Utilice los datos del fichero **Benefits Research**.

13.66. El fichero de datos **German Imports** muestra las importaciones reales alemanas (y), el consumo privado real (x_1) y el tipo de cambio real (x_2) en dólares estadounidenses por marco de un periodo de 31 años. Estime el modelo

$$\log Y_t = \beta_0 + \beta_1 \log x_{1i} + \beta_2 \log x_{2t} + \varepsilon_i$$

y escriba un informe sobre sus resultados.

13.8. Utilización de variables ficticias en modelos de regresión

En el análisis de la regresión múltiple, hemos supuesto hasta ahora que las variables independientes, x_j, existían en un rango y contenían muchos valores diferentes. Sin embargo, en los supuestos de la regresión múltiple la única restricción a la que están sujetas las variables independientes es que son valores fijos. Por lo tanto, podríamos tener una variable independiente que tomara solamente dos valores: $x_j = 0$ y $x_j = 1$. Esta estructura se denomina normalmente *variable ficticia*, y veremos que constituye un valioso instrumento para aplicar la regresión múltiple a situaciones en las que hay variables categóricas. Un importante ejemplo es una función lineal que varía en respuesta a alguna influencia. Consideremos primero una ecuación de regresión simple:

$$Y = \beta_0 + \beta_1 X_1$$

Supongamos ahora que introducimos una variable ficticia, X_2, que toma los valores 0 y 1 y que la ecuación resultante es

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Cuando $X_2 = 0$ en esta ecuación, la constante es β_0, pero cuando $X_2 = 1$, la constante es $\beta_0 + \beta_2$. Vemos, pues, que la variable ficticia desplaza la relación lineal entre Y y X_1 en el valor del coeficiente β_2. De esta forma, podemos representar el efecto de los desplazamientos en nuestra ecuación de regresión. Las variables ficticias también se llaman *variables de indicador*. Comenzamos nuestro análisis con un ejemplo de una importante aplicación.

EJEMPLO 13.13. Análisis de la discriminación salarial (estimación de un modelo utilizando variables ficticias)

Gender and Salary

El presidente de Investors Ltd. quiere averiguar si existe alguna prueba de la presencia de discriminación salarial en los salarios de las mujeres y los hombres analistas financieros. La Figura 13.19 muestra un ejemplo de los salarios anuales de los analistas en relación con sus años de experiencia. Véase el fichero de datos **Gender and Salary**.

Solución

Examinando los datos y el gráfico, vemos dos subconjuntos diferentes de salarios y parece que los salarios de los hombres son uniformemente más altos cualesquiera que sean los años de experiencia.

Este problema puede analizarse estimando un modelo de regresión múltiple del salario, Y, en función de los años de experiencia, X_1, con una segunda variable, X_2, que toma dos valores:

0 Mujeres analistas
1 Hombres analistas

El modelo de regresión múltiple resultante

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

puede analizarse utilizando los métodos que hemos aprendido, señalando que el coeficiente b_1 es una estimación del aumento anual esperado del salario por año de experien-

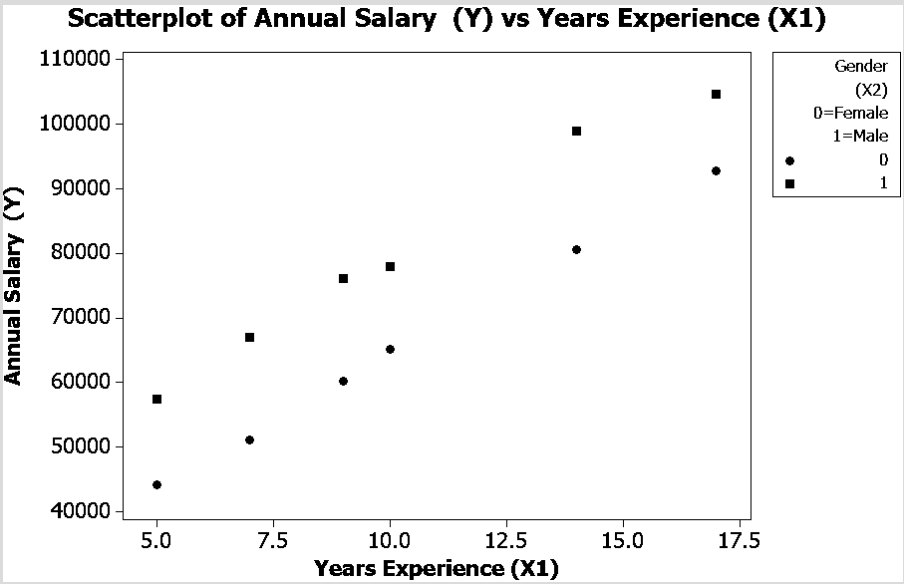


Figura 13.19. Ejemplo de una pauta de datos que indica la existencia de discriminación salarial.

cia y b_2 es el aumento que experimenta el salario medio cuando el analista es un hombre en lugar de una mujer. Si b_2 es positivo, eso indica que los salarios de los hombres son uniformemente más altos.

La Figura 13.20 presenta el análisis de regresión múltiple de Minitab para este problema. En este análisis vemos que el coeficiente de x_1 —gender— tiene un estadístico t de Student igual a 14,88 y un p-valor de 0, lo que nos lleva a rechazar la hipótesis nula de que el coeficiente es igual a 0. Este resultado indica que los salarios de los hombres son significativamente más altos. También vemos que $b_2 = 4.076{,}5$, lo que indica que el valor esperado del aumento anual es 4.076,50 $ y que $b_1 = 14.638{,}7$, lo que indica que los salarios de los hombres son, en promedio, 14.683,70 $ más altos. Este tipo de análisis se ha utilizado con éxito en algunos juicios sobre discriminación salarial, por lo que la mayoría de las empresas realizan análisis parecidos a éste para averiguar si existe alguna prueba de discriminación salarial.

Este tipo de ejemplos tiene numerosas aplicaciones en algunos problemas entre los que se encuentran los siguientes:

1. Es probable que la relación entre el número de unidades vendidas y el precio se desplace si entra un nuevo competidor en el mercado.
2. La relación entre el consumo agregado y la renta disponible agregada puede desplazarse en tiempos de guerra o como consecuencia de algún otro gran acontecimiento nacional.
3. La relación entre la producción total y el número de trabajadores puede desplazarse como consecuencia de la introducción de una nueva tecnología de producción.
4. La función de demanda de un producto puede variar como consecuencia de una nueva campaña publicitaria o de la publicación de una noticia relativa al producto.

Este análisis ha introducido el concepto de regresión utilizando variables ficticias como un método para ampliar nuestra capacidad de análisis. El método se resume a continuación.

```
The regression equation is
Annual Salary (Y) = 23608 + 14684 Gender (X2) 0=Female 1=Male
                          + 4076 Years Experience (X1)

Predictor                       Coef  SE Coef      T       P
Constant                       23608     1434  16.46   0.000
Gender (X2) 0=Female 1=Male  14683.7    987.0  14.88   0.000
Year Experience (X1)          4076.5    121.3  33.61   0.000


S = 1709.48    R-Sq = 99.3%   R-Sq(adj) = 99.2%

Analysis of Variance

Source            DF          SS          MS       F       P
Regression         2   394824096  1974120398  675.53   0.000
Residual Error     9    26300913     2922324
Total             11  3974541710
```

Figura 13.20. Análisis de regresión del ejemplo de la discriminación salarial: salario anual en relación con los años de experiencia y el sexo (salida Minitab).

Análisis de regresión utilizando variables ficticias

La relación entre Y y X_1

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

puede desplazarse en respuesta a un cambio de una determinada condición. El efecto del desplazamiento puede estimarse utilizando una variable ficticia que tiene el valor 0 (no se cumple la condición) y 1 (se cumple la condición). Como muestra la Figura 13.19, todas las observaciones del conjunto superior de puntos de datos tienen la variable ficticia $x_2 = 1$, y las observaciones de los puntos inferiores tienen la variable ficticia $x_2 = 0$. En estos casos, la relación entre Y y X_1 es especificada por el modelo de regresión múltiple

$$\hat{y}_i = b_0 + b_2 x_{2i} + b_1 x_{1i} \qquad \textbf{(13.33)}$$

El coeficiente b_2 representa el desplazamiento de la función entre el conjunto de puntos inferior de la Figura 13.19 y el superior. Las funciones de cada conjunto de puntos son

$$\hat{y} = b_0 + bx_1 \quad \text{cuando } x_2 = 0$$

y

$$\hat{y} = (b_0 + b_2 x_2) + b_1 x_1 \quad \text{cuando } x_2 = 1$$

En la primera función, la constante es b_0, mientras que en la segunda es $b_0 + b_2$. En el Capítulo 14 mostramos cómo pueden utilizarse las variables ficticias para analizar problemas que tienen más de dos categorías discretas.

Esta sencilla especificación del modelo de regresión lineal es un instrumento muy poderoso para resolver los problemas que implican un desplazamiento de la función lineal provocado por factores discretos identificables. Además, la estructura de regresión múltiple es un método directo para realizar un contraste de hipótesis, como hemos hecho en el ejemplo 13.13. El contraste de hipótesis es

$$H_0 : \beta_2 = 0 \,|\, \beta_1 \neq 0$$
$$H_1 : \beta_2 \neq 0 \,|\, \beta_1 \neq 0$$

El rechazo de la hipótesis nula, H_0, lleva a la conclusión de que la constante de los dos subconjuntos de datos es diferente. En el ejemplo 13.13 hemos visto que esta diferencia entre las constantes llevaba a la conclusión de que existía una diferencia significativa entre los salarios masculinos y los femeninos una vez eliminado el efecto de los años de experiencia.

Diferencias entre las pendientes

Podemos utilizar variables ficticias para analizar y contrastar las diferencias entre las pendientes añadiendo una variable de interacción. La Figura 13.21 muestra un ejemplo representativo. Para contrastar tanto las diferencias entre las constantes como las diferencias entre las pendientes, utilizamos un modelo de regresión más complejo.

Gender (X2) 0=Female 1=Male	Experience times Gender	Years Experience (X1)	Annual Salary (Y)
0	0	5	$36,730
0	0	7	$40,650
0	0	9	$46,820
0	0	10	$50,149
0	0	14	$59,679
0	0	17	$67,360
1	5	5	$51,535
1	7	7	$62,289
1	9	9	$72,486
1	10	10	$75,022
1	14	14	$90,079
1	17	17	$105,979

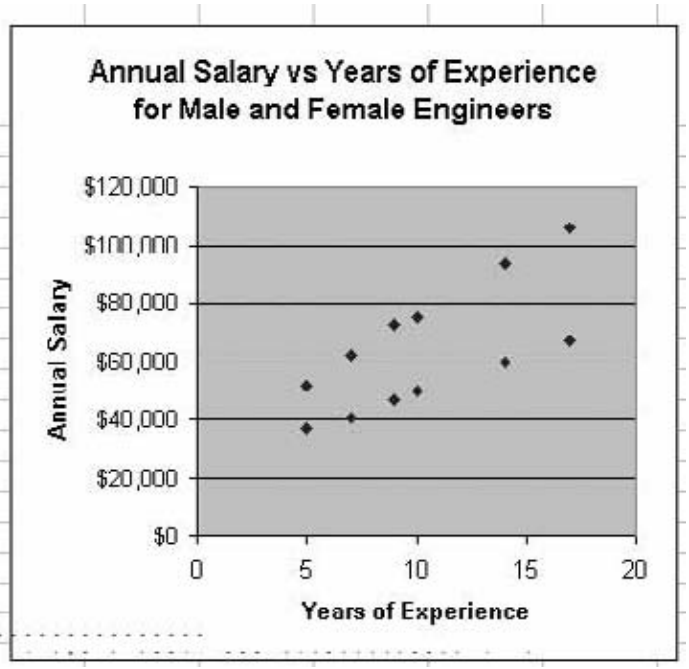


Figura 13.21. Datos salariales anuales de Systems Inc.

Regresión utilizando variables ficticias para contrastar las diferencias entre las pendientes

Para averiguar si existen diferencias significativas entre las pendientes de dos condiciones discretas, hay que expandir nuestro modelo de regresión a una forma más compleja:

$$Y = \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1 \qquad \textbf{(13.34)}$$

Ahora vemos que la pendiente de x_1 contiene dos componentes, β_1 y $\beta_3 X_2$. Cuando X_2 es igual a 0, la pendiente es el β_1 habitual. Sin embargo, cuando X_2 es igual a 1, la pendiente es igual a la suma algebraica de $\beta_1 + \beta_3$. Para estimar el modelo, necesitamos en realidad crear un nuevo conjunto de variables transformadas que sean lineales. Por lo tanto, el modelo utilizado realmente para la estimación es

$$\hat{y}_i = b_0 + b_2 x_{2i} + b_1 x_{1i} + b_3 x_{2i} x_{1i} \qquad \textbf{(13.35)}$$

El modelo de regresión resultante ahora es lineal con tres variables. La nueva variable, x_1x_2, a menudo se llama *variable de interacción*. Obsérvese que cuando la variable ficticia $x_2 = 0$, esta variable tiene un valor de 0, pero cuando $x_2 = 1$, esta variable tiene el valor de X_1. El coeficiente b_3 es una estimación de la diferencia entre el coeficiente de X_1 cuando $x_2 = 1$ y el coeficiente de X_1 cuando $x_2 = 0$. Por lo tanto, puede utilizarse el estadístico t de Student de b_3 para contrastar las hipótesis

$$H_0 : \beta_3 = 0 \,|\, \beta_1 \neq 0,\ \beta_2 \neq 0$$
$$H_1 : \beta_3 \neq 0 \,|\, \beta_1 \neq 0,\ \beta_2 \neq 0$$

Si rechazamos la hipótesis nula, concluimos que existe una diferencia entre las pendientes de los dos subgrupos. En muchos casos, nos interesará tanto la diferencia entre las constantes como la diferencia entre las pendientes y contrastaremos las dos hipótesis presentadas en este apartado.

EJEMPLO 13.14. Modelo de los salarios para Systems Inc. (estimación de un modelo utilizando variables ficticias)

El presidente de Systems Inc. está interesado en saber si las subidas salariales anuales de las ingenieras de la empresa han sigo iguales que las de los ingenieros. Ha habido algunas quejas tanto de los ingenieros como de las ingenieras de que los salarios de éstas no han subido al mismo ritmo que los de aquéllos.

Solución

La Figura 13.21 muestra los datos de la empresa y un diagrama de puntos dispersos. El diagrama sugiere que la pendiente es más alta en el caso del subgrupo superior, que representa a los ingenieros. En la Figura 13.22 presentamos el análisis de regresión múltiple realizado con el programa Excel, que puede utilizarse para contrastar la hipótesis de que las tasas de subida de los dos subgrupos de ingenieros son iguales. En este análisis vemos que la experiencia multiplicada por el sexo tiene un estadístico t de Stu-

Gender and Salary Increase

Análisis de regresión						
Estadísticos de la regresión						
Coeficiente de correlación múltiple	0,9993					
Coeficiente de determinación R	0,9985					
R ajustado	0,9980					
Error típico	936,5446					
Observaciones	12					
ANÁLISIS DE VARIANZA						
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*	
Regresión	3	4773061717	1591020572	1813,92	0,00	
Residuos	8	7016926	877116			
Total	11	4780078643				
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	23459,68	1043,57	22,48	0,00	21053,20	25866,15
Sexo (X2) 0 = Mujer	7053,17	1475,83	4,78	0,00	3649,90	10456,44
Experiencia x sexo	1886,82	132,89	14,20	0,00	1580,37	2193,27
Años de experiencia (X1)	2590,81	93,97	27,57	0,00	2374,11	2807,50

Figura 13.22. Análisis de regresión del salario anual en relación con la experiencia y el sexo (salida Excel).

dent de 14,20 y un *p*-valor de 0. Rechazamos la hipótesis nula de que, a medida que aumenta la experiencia, los salarios de los ingenieros y de las ingenieras han subido al mismo ritmo. Por lo tanto, será importante tomar medidas para abordar la discriminación salarial que es evidente en los datos. Los datos se encuentran en el fichero **Gender and Salary Increase**.

EJERCICIOS

Ejercicios básicos

13.67. ¿Cuál es la constante del modelo cuando la variable ficticia es igual a 1 en las siguientes ecuaciones, donde x_1 es una variable continua y x_2 es una variable ficticia que toma un valor de 0 o 1?

a) $\hat{y} = 4 + 8x_1 + 3x_2$
b) $\hat{y} = 7 + 6x_1 + 5x_2$
c) $\hat{y} = 4 + 8x_1 + 3x_2 + 4x_1x_2$

13.68. ¿Cuál es la constante del modelo y el coeficiente de la pendiente de x_1 cuando la variable ficticia es igual a 1 en las siguientes ecuaciones, donde x_1 es una variable continua y x_2 es una variable ficticia que toma un valor de 0 o 1?

a) $\hat{y} = 4 + 9x_1 + 1{,}78x_2 + 3{,}09x_1x_2$
b) $\hat{y} = -3 + 7x_1 + 4{,}15x_2 + 2{,}51x_1x_2$
c) $\hat{y} = 10 + 5x_1 + 3{,}67x_2 + 3{,}98x_1x_2$

Ejercicios aplicados

13.69. El siguiente modelo se ajustó a las observaciones de 1972-1979 en un intento de explicar la conducta de la fijación de los precios.

$$\hat{y} = \underset{(0{,}029)}{37x_1} + \underset{(0{,}50)}{5{,}22x_2}$$

donde

y = diferencia entre el precio del año actual y el precio del año anterior en dólares por barril
x_1 = diferencia entre el precio al contado en el año actual y el precio al contado en el año anterior
x_2 = variable ficticia que toma el valor 1 en 1974 y 0 en los demás, para representar el efecto específico del embargo del petróleo de ese año

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

Interprete verbal y gráficamente el coeficiente estimado de la variable ficticia.

13.70. Se ha ajustado el siguiente modelo para explicar los precios de venta de los pisos de una muestra de 815 ventas.

$$\hat{y} = -1.264 + \underset{(0{,}91)}{48{,}18x_1} + \underset{(515)}{3.382x_2} - \underset{(488)}{1.859x_3} + \underset{(947)}{3.219x_4} + \underset{(768)}{2.005x_5} \qquad \bar{R}^2 = 0{,}86$$

donde

$\hat{y}$ = precio de venta del piso, en dólares
x_1 = metros cuadrados útiles
x_2 = tamaño del garaje en número de automóviles
x_3 = antigüedad del piso en años
x_4 = variable ficticia que toma el valor 1 si el piso tiene chimenea y 0 en caso contrario
x_5 = variable ficticia que toma el valor 1 si el piso tiene suelos de madera y 0 si tiene suelos de vinilo

a) Interprete el coeficiente estimado de x_4.
b) Interprete el coeficiente estimado de x_5.
c) Halle el intervalo de confianza al 95 por ciento del efecto de una chimenea en el precio de venta, manteniéndose todo lo demás constante.
d) Contraste la hipótesis nula de que el tipo de suelo no afecta al precio de venta frente a la hipótesis alternativa de que, manteniéndose todo lo demás constante, los pisos con suelo de madera tienen un precio de venta más alto que los pisos con suelo de vinilo.

13.71. Se ha ajustado el siguiente modelo a datos sobre 32 compañías de seguros.

$$\hat{y} = 7{,}62 - \underset{(0{,}008)}{0{,}16x_1} + \underset{(0{,}496)}{1{,}23x_2} \qquad R^2 = 0{,}37$$

donde

y = relación precio-beneficios
x_1 = volumen de activos de las compañías de seguros, en miles de millones de dólares
x_2 = variable ficticia que toma el valor 1 en el caso de las compañías regionales y 0 en el de las nacionales.

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete el coeficiente estimado de la variable ficticia.

b) Contraste la hipótesis nula de que el verdadero coeficiente de la variable ficticia es 0 frente a la hipótesis alternativa bilateral.

c) Contraste al nivel del 5 por ciento la hipótesis nula $\beta_1 = \beta_2 = 0$ e interprete su resultado.

13.72. El decano de una facultad de derecho quería evaluar la importancia de factores que podrían ayudar a predecir el éxito en los estudios de postgrado en derecho. Se obtuvieron datos de una muestra aleatoria de 50 estudiantes cuando terminaron sus estudios de postgrado en derecho y se ajustó el siguiente modelo:

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

donde

Y_i = calificación que refleja el rendimiento global de los estudiantes en sus estudios de postgrado en derecho

x_{1i} = calificación media de los estudios de grado

x_{2i} = calificación en el examen de acceso a la universidad

x_{3i} = variable ficticia que toma el valor 1 si las cartas de recomendación del estudiante son excepcionalmente buenas y 0 en caso contrario

Utilice la parte de la salida de la regresión estimada mostrada aquí para escribir un informe que resuma los resultados de este estudio.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	*F* VALUE	*R*-SQUARE
MODEL	3	641.04	213.68	8.48	.356
ERROR	46	1159.66	25.21		
CORRECTED TOTAL	49	1800.70			

PARAMETER	ESTIMATE	T FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	6.512		
X1	3.502	1.45	2.419
X2	0.491	4.59	0.107
X3	10.327	2.45	4.213

13.73. El siguiente modelo se ajustó a datos de 50 estados de Estados Unidos.

$$\hat{y} = \underset{}{13.472} + \underset{(124,3)}{547}x_1 + \underset{(1,858)}{5,48}x_2 + \underset{(208,9)}{493}x_3 + \underset{(234)}{32,7}x_4 + \underset{(2.897)}{5.793}x_5 - \underset{(1.761)}{3.100}x_6 \qquad R^2 = 0,54$$

donde

y = sueldo anual del fiscal general del estado

x_1 = sueldo anual medio de los abogados en miles de dólares

x_2 = número de leyes aprobadas en la legislatura anterior

x_3 = número de actuaciones de los tribunales de los estados que dieron lugar a una anulación de legislación en los 40 años anteriores

x_4 = duración del mandato del fiscal general del estado

x_5 = variable ficticia que toma el valor 1 si los magistrados del tribunal supremo del estado pueden ser cesados por el gobernador, por el consejo del poder judicial o mediante una votación por mayoría del tribunal supremo y 0 en caso contrario

x_6 = variable ficticia que toma el valor 1 si los magistrados del tribunal supremo son designados tras unas elecciones en las que intervienen los partidos políticos y 0 en caso contrario

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete el coeficiente estimado de la variable ficticia x_5.

b) Interprete el coeficiente estimado de la variable ficticia x_6.

c) Contraste al nivel del 5 por ciento la hipótesis nula de que el verdadero coeficiente de la variable ficticia x_5 es 0 frente a la hipótesis alternativa de que es positivo.

d) Contraste al nivel del 5 por ciento la hipótesis nula de que el verdadero coeficiente de la variable ficticia x_6 es 0 frente a la hipótesis alternativa de que es negativo.

e) Halle e interprete un nivel de confianza del 95 por ciento del parámetro β_1.

13.74. Un grupo consultor ofrece cursos de gestión financiera para los ejecutivos. Al final de estos cursos, los participantes deben hacer una valoración global del valor del curso. Se estimó para una muestra de 25 cursos la siguiente regresión por mínimos cuadrados.

$$\hat{y} = \underset{}{42,97} + \underset{(0,29)}{0,38}x_1 + \underset{(0,21)}{0,52}x_2 - \underset{(0,11)}{0,08}x_3 + \underset{(0,359)}{6,21}x_4 \qquad R^2 = 0,569$$

donde

y = valoración media realizada por los participantes en el curso

x_1 = porcentaje del tiempo del curso dedicado a sesiones de discusión en grupo
x_2 = dinero, en dólares, por miembro del curso dedicados a preparar el material del curso
x_3 = dinero, en dólares, por miembro del curso gastado en comida y bebida
x_2 = variable ficticia que toma el valor 1 si interviene en el curso un profesor visitante y 0 en caso contrario.

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete el coeficiente estimado de x_4.
b) Contraste la hipótesis nula de que el verdadero coeficiente de x_4 es 0 frente a la hipótesis alternativa de que es positivo.
c) Interprete el coeficiente de determinación y utilícelo para contrastar la hipótesis nula de que las cuatro variables independientes, consideradas en conjunto, no influyen linealmente en la variable dependiente.
d) Halle e interprete el intervalo de confianza al 95 por ciento de β_2.

13.75. En un estudio, se estimó un modelo de regresión para comparar el rendimiento de los estudiantes que asistían a un curso de estadística para los negocios: un curso normal de 14 semanas o un curso intensivo de 3 semanas. Se estimó el siguiente modelo a partir de las observaciones sobre 350 estudiantes (véase la referencia bibliográfica 5):

$$\hat{y} = -0{,}7052 + \underset{(0{,}4568)}{1{,}4170}x_1 + \underset{(0{,}3287)}{2{,}1624}x_2 + \underset{(0{,}4393)}{0{,}8680}x_3 + \underset{(0{,}3766)}{1{,}0845}x_4 + \underset{(0{,}0628)}{0{,}4694}x_5 + \underset{(0{,}0094)}{0{,}0038}x_6 + \underset{(0{,}0776)}{0{,}0484}x_7$$

$$R^2 = 0{,}344$$

donde

y = calificación obtenida en un examen normalizado sobre los conocimientos de estadística después de asistir al curso
x_1 = variable ficticia que toma el valor 1 si se asistió a un curso de 3 semanas y 0 si se asistió a un curso de 14 semanas
x_2 = calificación media del estudiante
x_3 = variable ficticia que toma el valor 0 o 1, dependiendo de cuál de dos profesores impartiera el curso
x_4 = variable ficticia que toma el valor 1 si el estudiante es varón y 0 si es mujer
x_5 = calificación obtenida en un examen normalizado sobre los conocimientos de matemáticas antes de asistir al curso
x_6 = número de créditos semestrales que había completado el estudiante
x_7 = edad del estudiante

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

Escriba un informe analizando lo que puede aprenderse con esta regresión ajustada.

Se recomienda que los siguientes ejercicios se resuelvan con la ayuda de un computador.

13.76. En un estudio de 27 estudiantes de la Universidad de Illinois se obtuvieron resultados sobre la calificación media (y), el número de horas semanales dedicadas a estudiar (x_1), el número medio de horas dedicadas a estudiar para los exámenes (x_2), el número de horas semanales pasadas en los bares (x_3), el hecho de que los estudiantes tomen notas o subrayen cuando leen los libros de texto ($x_4 = 1$ si sí, 0 si no) y el número medio de créditos realizados por semestre (x_5). Estime la regresión de la calificación media con respecto a las cinco variables independientes y escriba un informe sobre sus resultados. Los datos se encuentran en el fichero de datos **Student Performance** de su disco de datos.

13.77. Le han pedido que desarrolle un modelo para analizar los salarios de una gran empresa. Los datos para desarrollarlo se encuentran en el fichero llamado **Salorg**.

a) Utilizando los datos del fichero, desarrolle un modelo de regresión que prediga el salario en función de las variables que seleccione. Calcule los estadísticos F y t condicionados del coeficiente de cada variable de predicción incluida en el modelo. Muestre todo lo que hace y explíquelo minuciosamente.
b) Contraste la hipótesis de que las mujeres tienen un salario anual más bajo condicionado a las variables de su modelo. La variable «Gender_1F» toma el valor 1 en el caso de las mujeres y 0 en el de los hombres.
c) Contraste la hipótesis de que la tasa de subida salarial de las mujeres ha sido más baja condicionada a las variables del modelo desarrollado en el apartado (b).

13.9. Método de aplicación del análisis de regresión múltiple

Cotton

En este apartado presentamos un extenso caso práctico que indica cómo se realizaría un estudio estadístico. El estudio detenido de este ejemplo puede ayudar a utilizar muchos de los métodos presentados en este capítulo y en los anteriores.

El objetivo de este estudio es desarrollar un modelo de regresión múltiple para predecir las ventas de tejido de algodón. Los datos para el proyecto proceden del fichero de datos **Cotton**, que se encuentra en el disco de datos de este libro. Las variables del fichero de datos son

quarter	Trimestre del año
year	año de observación
cottonq	cantidad de tejido de algodón producida
whoprice	índice de precios al por mayor
impfab	cantidad de tejido importado
expfab	cantidad de tejido exportado

Especificación del modelo

El primer paso para desarrollar el modelo es seleccionar una teoría económica adecuada que sirva de base para el análisis del modelo. Este proceso de identificación de un conjunto de variables de predicción probables y la forma matemática del modelo se conoce con el nombre de *especificación del modelo*. En este caso, la teoría adecuada se basa en la de los modelos económicos de demanda. La teoría económica indica que el precio debe producir un importante efecto: una subida del precio reduce la cantidad demandada. Es probable que también haya otras variables que influyan en la cantidad demandada de algodón. Es de esperar que la cantidad importada de tejido de algodón reduzca la demanda de tejido interior y que la cantidad exportada de tejido de algodón aumente la demanda de tejido interior. En el lenguaje económico, las importaciones y las exportaciones de tejido desplazan la función de demanda. Basándonos en este análisis, nuestra especificación inicial incluye el precio con un coeficiente negativo, el tejido exportado con un coeficiente positivo y el tejido importado con un coeficiente negativo. Se especifica inicialmente que todos los coeficientes tienen efectos lineales. Por lo tanto, el modelo tiene la forma

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

donde x_1 es el precio al por mayor, x_2 es la cantidad de tejido importado y x_3 es la cantidad de tejido exportado.

También existe la posibilidad de que la cantidad demandada varíe con el tiempo, y, por lo tanto, el modelo debe incluir la posibilidad de una variable temporal para reducir la variabilidad no explicada. Para este análisis queremos utilizar una variable que represente el tiempo. Como el tiempo es indicado por una combinación de año y trimestre, utilizamos la transformación

```
Time = Year + 0.25*Quarter
```

para producir una nueva variable del tiempo que sea continuamente creciente.

El paso siguiente en el análisis es hacer una descripción estadística de las variables y de sus relaciones. Excluimos el año y el trimestre de este análisis porque han sido sustituidos por el tiempo y su inclusión sólo introduciría confusión en el análisis. Utilizamos el

programa Minitab para obtener medidas de la tendencia central y de la dispersión y también para comprender algo la pauta de las observaciones. La Figura 13.23 contiene la salida Minitab. El examen de la media, la desviación típica y el mínimo y el máximo indica la región potencial de aplicación del modelo. El modelo de regresión estimado siempre pasa por la media de las variables del modelo. Los valores predichos de la variable dependiente, «cottonq», pueden utilizarse dentro del rango de las variables independientes.

El paso siguiente es examinar las relaciones simples existentes entre las variables utilizando tanto la matriz de correlaciones como la opción de los gráficos matriciales. Éstos deben examinarse conjuntamente para averiguar la fuerza de las relaciones lineales (correlaciones) y para averiguar la forma de las relaciones (gráfico matricial).

La Figura 13.24 contiene la matriz de correlaciones de las variables del estudio elaborada utilizando Minitab. El *p*-valor mostrado con cada correlación indica la probabilidad de que la hipótesis de la correlación 0 entre las dos variables sea verdadera. Utilizando nuestra regla de selección basada en el contraste de hipótesis, podemos concluir que un *p*-valor de menos de 0,05 es una prueba de la existencia de una estrecha relación lineal entre las dos variables. Examinando la primera columna, observamos que existen estrechas relaciones lineales entre «cottonq» y tanto «whoprice» como «time». La variable «expfab» tiene una posible relación simple marginalmente significativa. Una buena regla práctica, mostrada en el apartado 12.1, para examinar los coeficientes de correlación es que el valor

Figura 13.23. Estadísticos descriptivos de las variables del mercado del algodón (salida Minitab).

Results for: Cotton.MtW
Descriptive Statistics: cottonq, whoprice, impfab, expfab, time

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
cottonq	28	0	1779.8	54.9	290.5	1277.0	1535.3	1762.5	2035.0
whoprice	28	0	106.81	1.16	6.11	98.00	100.45	107.40	112.20
impfab	28	0	7.52	1.38	7.33	1.30	2.78	4.85	9.05
expfab	28	0	274.0	20.3	107.7	80.0	190.5	277.1	358.1
time	28	0	69.625	0.389	2.056	66.250	67.813	69.625	71.438

Variable	Maximum
cottonq	2287.0
whoprice	115.80
impfab	27.00
expfab	477.0
time	73.000

Figura 13.24. Correlaciones de las variables del mercado del algodón (salida Minitab).

Correlations: cottonq, whoprice, impfab, expfab, time

	cottonq	whoprice	impfab	expfab
whoprice	−0.950			
	0.000			
impfab	0.291	−0.439		
	0.133	0.019		
expfab	0.370	−0.285	0.181	
	0.052	0.142	0.357	
time	−0.950	0.992	−0.392	−0.238
	0.000	0.000	0.039	0.222

Cell Contents: Pearson correlation
P−Value

absoluto de la correlación debe ser superior a 2 dividido por la raíz cuadrada del tamaño de la muestra, n. En este problema, el valor de selección es $2/\sqrt{28} = 0{,}38$.

La segunda tarea es averiguar si existen estrechas relaciones simples entre los pares de variables de predicción posibles. Vemos una estrechísima correlación entre «time» y «whoprice» y relaciones significativas entre «impfab» y tanto «time» como «whoprice». Estas elevadas correlaciones hacen que la varianza de los estimadores de los coeficientes tanto de «time» como de «whoprice» sea alta si se incluyen ambos como variables de predicción.

También podemos examinar las relaciones entre las variables utilizando los gráficos matriciales mostrados en la Figura 13.25. Los diagramas de puntos dispersos individuales muestran simultáneamente las relaciones entre diferentes variables. Constituyen, pues, un tipo de presentación parecido a una matriz de correlaciones. La ventaja del diagrama de puntos dispersos radica en que incluye todos los puntos de datos. También se puede ver, pues, si existe una relación no lineal simple entre las variables y/o si existe algún agrupamiento extraño de observaciones. Todas las variables, excepto «year» y «quarter», están incluidas en el mismo orden que en la matriz de correlaciones, por lo que hay una comparación directa entre la matriz de correlaciones y los gráficos matriciales.

Obsérvese la correspondencia entre las correlaciones y los diagramas de puntos dispersos. Tanto «whoprice» como «time» tienen estrechas relaciones lineales con «cottonq». Sin embargo, la estrecha relación lineal positiva entre «whoprice» y «time» tendrá una gran influencia en los coeficientes estimados, como se muestra en el apartado 13.2, y en los errores típicos de los coeficientes, como se muestra en el apartado 13.4. No existe ninguna estrecha relación simple entre las variables de predicción potenciales. Ni las importaciones ni las exportaciones están correlacionadas con el precio al por mayor, con el tiempo o entre sí.

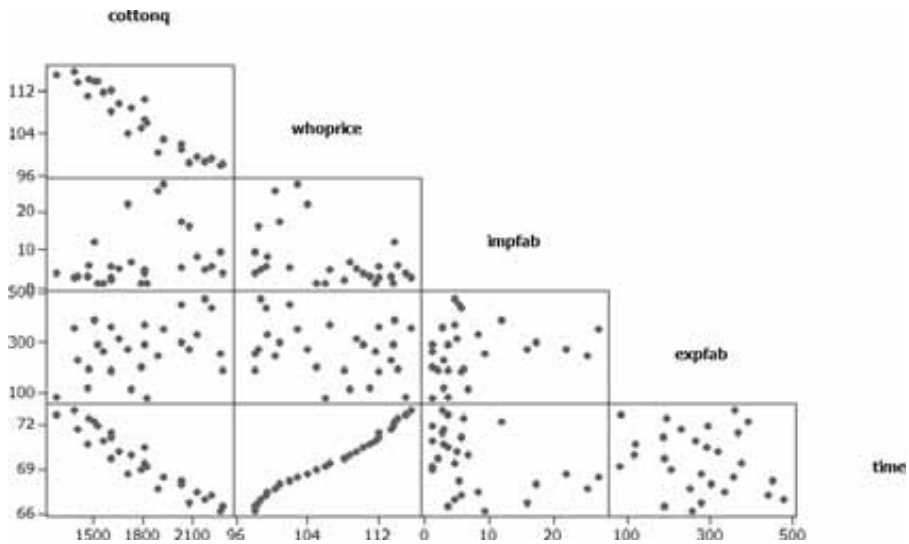


Figura 13.25. Gráficos matriciales de las variables del estudio (salida Minitab).

Regresión múltiple

El paso siguiente consiste en estimar el primer modelo de regresión múltiple. La teoría económica para este análisis sugiere que la cantidad producida de tejido de algodón debe estar relacionada inversamente con el precio y con la cantidad importada de tejido y relacionada directamente con la cantidad exportada de tejido. Además, la estrecha correlación

entre el tiempo y la producción de tejido de algodón indica que la producción disminuyó linealmente con el paso del tiempo, pero que el precio al por mayor también subió linealmente con el paso del tiempo. La estrecha correlación positiva resultante entre el tiempo y el precio al por mayor influye en ambos coeficientes en una ecuación de regresión múltiple. Seleccionamos «cottonq» como variable dependiente y «whoprice», «impfab», «expfab» y «time», por ese orden, como variables independientes. El primer análisis de regresión múltiple se muestra en la Figura 13.26.

El análisis de los estadísticos de la regresión indica que el valor de R^2 es alto y el error típico de la estimación (S) es igual a 78,91, en comparación con la desviación típica de 290,5 (Figura 13.23) de «cottonq», cuando se considera de forma aislada. Las variables «impbaf» y «expfab» son ambas significativas y tienen signos que corresponden a la teoría económica. Los pequeños estadísticos t de Student de «whoprice» y «time» indican que, en realidad, existe un grave problema. Ambas variables no pueden incluirse como predictores porque representan el mismo efecto.

Las reglas para eliminar variables se basan en una combinación tanto de las teorías subyacentes al modelo como de indicadores estadísticos. La regla estadística sería eliminar la variable que tiene el menor t de Student absoluto, es decir, «time». La teoría económica defendería la inclusión de una variable del precio en un modelo para predecir la cantidad producida o la cantidad demandada. Vemos que en este caso ambas reglas llevan a la misma conclusión. No siempre ocurre así, por lo que es muy importante valorar bien los resultados y tener claros los objetivos del modelo.

Figura 13.26. Modelo inicial de regresión múltiple (salida Minitab).

Regression Analysis: cottonq versus whoprice, impfab, expfab, time

```
The regression equation is
cottonq =8876 − 24.3 whoprice − 5.57 impfab + 0.376 expfab − 65.5 time

Predictor    Coef  SE Coef      T      P
Constant     8876     2295   3.87  0.001
whoprice   −24.31    24.45  −0.99  0.331
impfab     −5.565    2.527  −2.20  0.038
expfab     0.3758   0.1595   2.36  0.027
time       −65.51    70.24  −0.99  0.361

S = 78.9141   R−Sq = 93.7%  R−Sq(adj) = 92.6%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       4  2134572  533643  85.69  0.001
Residual Error  23   143231    6227
Total           27  2277803

Source    DF   Seq SS
whoprice   1  2055110
impfab     1    44905
expfab     1    29141
time       1     5417

Unusual Observations

Obs  Whoprice  Cottonq     Fit  SE Fit  Residual  St Resid
18        110   1810.0  1663.3    29.6     146.7     2.00R

R denotes an observation with a large standardized residual.
```

Nota
Esta tabla indica la variabilidad explicada condicionada de cada variable, dado el orden de entrada utilizado para este análisis de regresión.

Es importante formular claramente las razones por las que se seleccionan las variables antes de examinar los resultados. En los modelos económicos de demanda o de oferta como el que examinamos aquí, desearíamos fervientemente seguir la teoría económica e incluir el precio, a menos que los resultados estadísticos fueran muy contrarios a esa decisión previa. Por ejemplo, si el valor absoluto del estadístico *t* de Student del tiempo fuera superior a 2,5 o 3 y el valor absoluto del estadístico *t* de Student del precio al por mayor fuera inferior a 1, habría pruebas contundentes en contra de la teoría de que el precio es una importante variable.

Basándose en este análisis, se estima un segundo modelo de regresión, mostrado en la Figura 13.27, en el que se excluye el tiempo como variable de predicción. Ahora vemos que la variable «whoprice» es muy significativa y que los estadísticos *s* y R^2 son esencialmente iguales que los del primer análisis de regresión (Figura 13.26). Obsérvese también que la suma de los cuadrados de la regresión explicada (*SCR*) y la suma de los cuadrados de los errores residuales (*SCE*) son esencialmente iguales. La desviación típica del coeficiente de «whoprice» ha disminuido de 24,45 a 2,835 y, como consecuencia, la *t* de Student es considerablemente mayor. Como hemos visto en el apartado 13.4, cuando existen correlaciones estrechas entre variables independientes, las varianzas de los estimadores de los coeficientes son mucho mayores. Vemos aquí ese efecto. Obsérvese también que en este modelo de regresión, la estimación del coeficiente del precio al por mayor cambia de $-24{,}31$ a $-46{,}956$. En el apartado 13.2 hemos visto que las correlaciones entre variables de predicción producen un complejo efecto en las estimaciones de los coeficientes, por lo

Figura 13.27. Modelo final del análisis de regresión (salida Minitab).

Regression Analysis: cottonq versus whoprice, impfab, expfab, time

```
The regression equation is


Predictor    Coef  SE Coef      T      P
Constant   6757.0    322.2  20.97  0.000
whoprice  −46.956    2.835 −16.56  0.000
impfab     −6.517    2.306  −2.83  0.009
expfab     0.3190   0.1471   2.17  0.040

S = 78.6998   R−Sq = 93.5%  R-Sq(adj) = 92.7%

Analysis of Variance

Source          DF       SS      MS       F      P
Regression       3  2129156  709719  114.59  0.000
Residual Error  24   148648    6194
Total           27  2277803

Source    DF   Seq SS
whoprice   1  2055110
impfab     1    44905
expfab     1    29141

Unusual Observations

Obs  Whoprice  Cottonq     Fit  SE Fit  Residual  St Resid
18        110   1810.0  1642.0    18.7     168.0     2.20R

R denotes an observation with a large standardized residual.
```

Nota
Estas sucesivas sumas de los cuadrados explicadas condicionadas son iguales que las de la regresión de la Figura 13.26, que incluían el tiempo como variable de predicción.

que no siempre existe una diferencia tan grande. Sin embargo, las correlaciones entre variables independientes siempre aumentan el error típico de los coeficientes. Los errores típicos de los otros dos coeficientes no han cambiado significativamente, debido a que las correlaciones con el tiempo no eran grandes.

El programa Minitab también contiene una lista de observaciones con residuos extremos. Vemos en la observación 18 que el valor observado de «cottonq» es muy superior al valor que predice la ecuación. En este caso, podríamos decidir volver a los datos originales y tratar de averiguar si hay un error en los datos del fichero. Esa investigación también podría ayudar a comprender el proceso estudiado utilizando la regresión múltiple.

Efecto de la eliminación de una variable estadísticamente significativa

En este apartado examinamos el efecto de la eliminación de una variable significativa del modelo de regresión. En la Figura 13.27 hemos visto que «expfab» es una predictor estadísticamente significativo de la cantidad producida de algodón. Sin embargo, el análisis de regresión de la Figura 13.28 ha eliminado «expfab» del modelo de regresión de la Figura 13.27.

Obsérvese que, como consecuencia de la eliminación de «expfab», el error típico de la estimación ha aumentado de 78,70 a 84,33 y R^2 ha disminuido del 93,5 al 92,2 por ciento. Estos resultados indican que el término de error del modelo ahora es mayor y, por lo tanto, ha empeorado la calidad del modelo.

El estadístico F condicionado de «expfab» puede calcularse utilizando las tablas del análisis de la varianza de los modelos de las Figuras 13.27 y 13.28. En la siguiente ecuación, definimos la regresión lineal a partir de la Figura 13.27 como modelo 1 y la regresión de la Figura 13.28, eliminado «expfab», como modelo 2. Utilizando estas convenciones, el estadístico F condicionada de la variable «expfab», X_3, en la hipótesis nula de que su coeficiente es 0, puede calcularse de la forma siguiente:

$$F_{x_3} = \frac{SCR_1 - SCR_2}{s_e^2} = \frac{(2.129.156 - 2.100.015)}{6.194} = 4{,}705$$

Figura 13.28. Análisis de regresión con la eliminación del tejido exportado (salida Minitab).

Regression Analysis: cottonq versus whoprice, impfab, expfab, time

```
The regression equation is
cottonq = 6995 − 48.4 whoprice − 6.20 impfab

Predictor    Coef  SE Coef      T      P
Constant   6994.8    324.6  21.55  0.000
whoprice  −48.388    2.955 −16.38  0.000
impfab     −6.195    2.465  −2.51  0.019

S = 84.3299   R−Sq = 92.2%  R-Sq(adj) = 91.6%

Analysis of Variance

Source          DF       SS      MS       F      P
Regression       2  2100015 1050007  147.65  0.000
Residual Error  25   177788    7112
Total           27  2277803
```

También podemos calcular el estadístico t de Student condicionado de la variable x_3 tomando la raíz cuadrada de la F_{x_3} condicionada:

$$t_{x_3} = \sqrt{4{,}705} = 2{,}169$$

y, naturalmente, vemos que es igual que el estadístico t de Student de la variable «expfab» (x_3) de la Figura 13.27. El contraste F condicionado de una única variable independiente siempre es exactamente igual que el F condicionado, ya que una F con 1 grado de libertad en el numerador es exactamente igual a t^2.

Análisis de los residuos

Después de ajustar el modelo de regresión, es útil examinar los residuos para averiguar cómo se ajusta realmente el modelo a los datos y los supuestos de la regresión. En el apartado 12.7, examinamos el análisis de los casos atípicos y los puntos extremos en la regresión simple. Esas ideas también se aplican directamente a la regresión múltiple y deben formar parte del análisis de los residuos. Recuérdese que los residuos se calculan de la forma siguiente:

$$e_i = y_i - \hat{y}_i$$

Con el programa Minitab o con cualquier otro buen paquete estadístico se puede calcular una variable que contenga los residuos de un análisis de regresión. Se ha hecho para el modelo final de regresión de la Figura 13.27. El primer paso consiste en examinar la pauta de los residuos construyendo un histograma, como el de la Figura 13.29. Vemos que la distribución de los residuos es aproximadamente simétrica. La distribución también parece algo uniforme. Obsérvese que se debe en parte al pequeño tamaño de la muestra utilizada para construir el histograma.

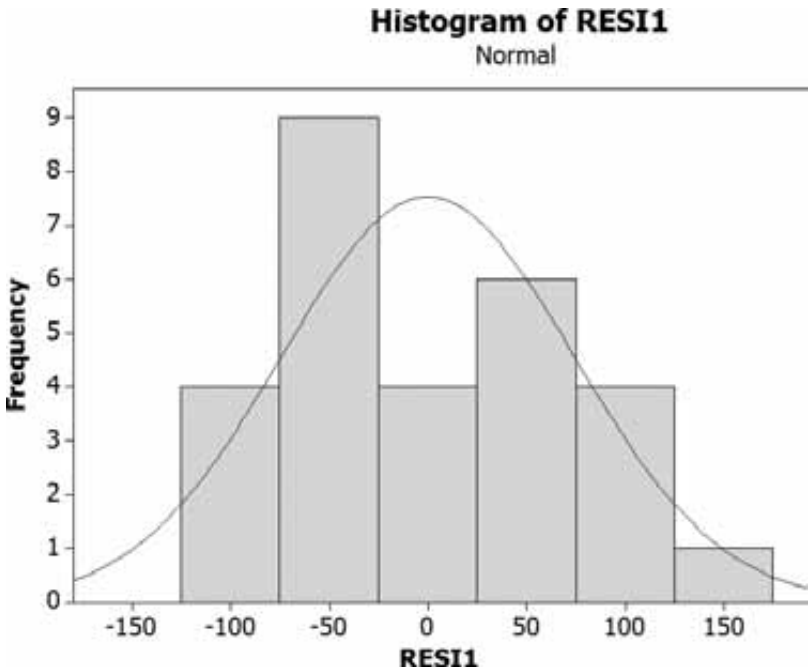


Figura 13.29. Histograma de los residuos del modelo final de regresión.

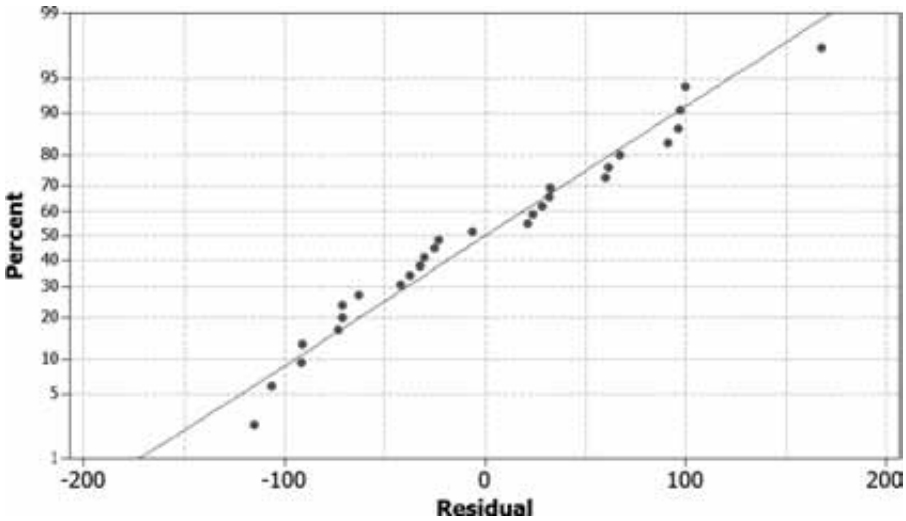


Figura 13.30. Gráfico de probabilidad normal de los residuos del modelo.

La realización de un gráfico de probabilidad normal, como el de la Figura 13.30, es útil para averiguar la pauta de los residuos. El gráfico indica la existencia de una relación lineal aproximada y, por lo tanto, no es posible rechazar el supuesto de que los residuos siguen una distribución normal.

También es bueno representar los residuos en relación con cada una de las variables independientes incluidas en el análisis. Eso permite comprobar que no había unos cuantos puntos de datos excepcionales o una compleja relación no lineal condicionada de una de las variables independientes. Si el modelo se ha especificado y se ha estimado correctamente, esperamos que no exista ninguna pauta de relación entre las variables independientes y los residuos. La Figura 13.31 muestra el gráfico de los residuos en relación con la variable del precio al por mayor. No observamos ninguna pauta excepcional en este gráfico, salvo el elevado caso atípico positivo cuando el precio al por mayor es aproximadamente 110.

En la Figura 13.32 mostramos el gráfico de los residuos en relación con el tejido importado. Una vez más, no vemos ninguna pauta excepcional de los residuos, pero sí observamos que la mayoría de las importaciones están concentradas entre 0 y 10. Por lo tanto,

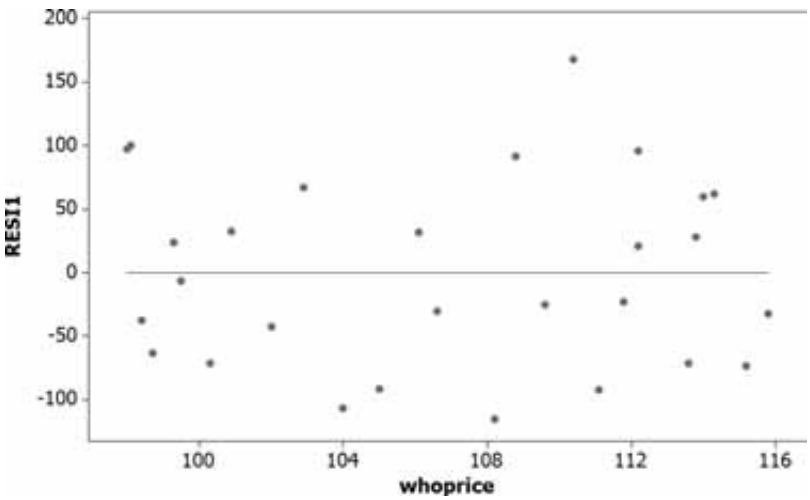


Figura 13.31. Diagrama de puntos dispersos de los residuos en relación con el precio al por mayor.

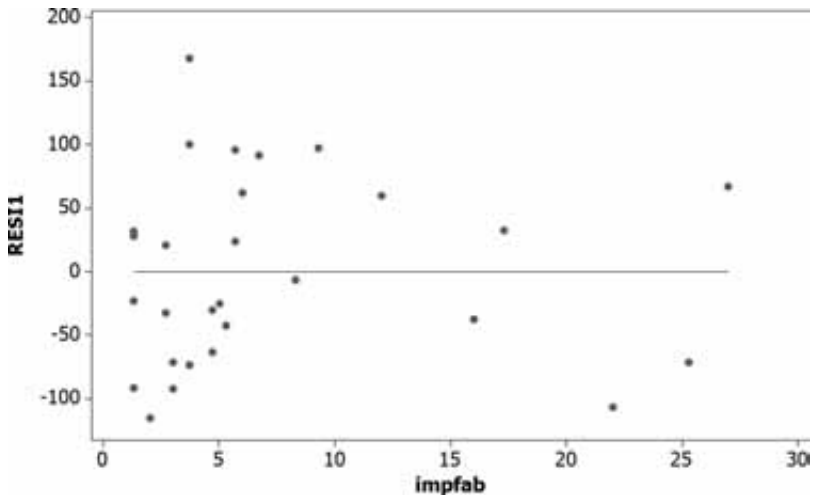


Figura 13.32. Diagrama de puntos dispersos de los residuos en relación con el tejido importado.

los valores más altos del tejido importado podrían producir un gran efecto en el coeficiente de la pendiente de la recta de regresión. Por último, en la Figura 13.33 vemos un gráfico de los residuos en relación con el tejido exportado. De nuevo, la pauta de los residuos no sugiere una alternativa a la relación lineal.

El análisis final de los residuos examina la relación entre los residuos y la variable dependiente. Consideramos un gráfico de los residuos en relación con el valor observado de la variable dependiente en la Figura 13.34 y en relación con el valor predicho de la variable dependiente en la 13.35. Podemos ver en la 13.34 que existe una relación positiva entre los residuos y el valor observado de «cottonq». Hay más residuos negativos en los valores bajos de «cottonq» y más residuos positivos en los valores altos de «cottonq». Es posible demostrar matemáticamente que siempre existe una correlación positiva entre los residuos y los valores observados de la variable dependiente. Por lo tanto, un gráfico de los residuos en relación con el valor observado no suministra ninguna información útil. Sin embargo, siempre se deben representar los residuos en relación con los valores predichos o ajustados de la variable dependiente. De esa forma se averigua si los errores del modelo son estables en el rango de los valores predichos. En este ejemplo, obsérvese que no existe ninguna relación entre los residuos y los valores predichos. Por lo tanto, los errores del modelo son estables en el rango.

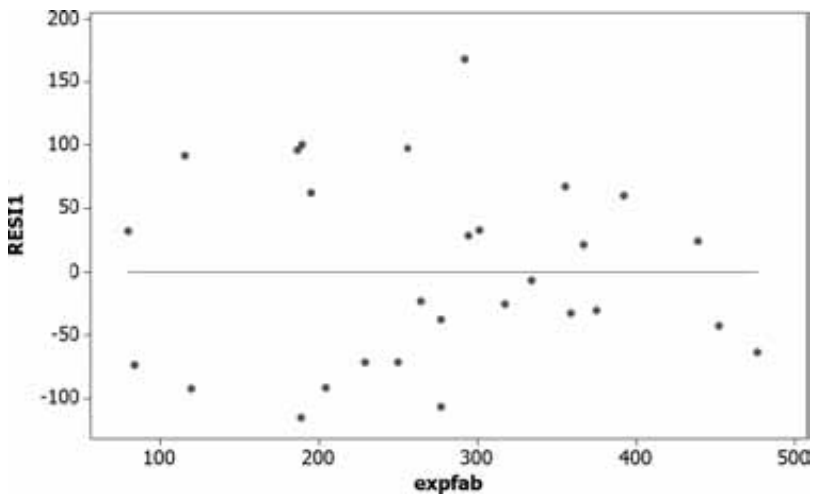


Figura 13.33. Diagrama de puntos dispersos de los residuos en relación con el tejido exportado.

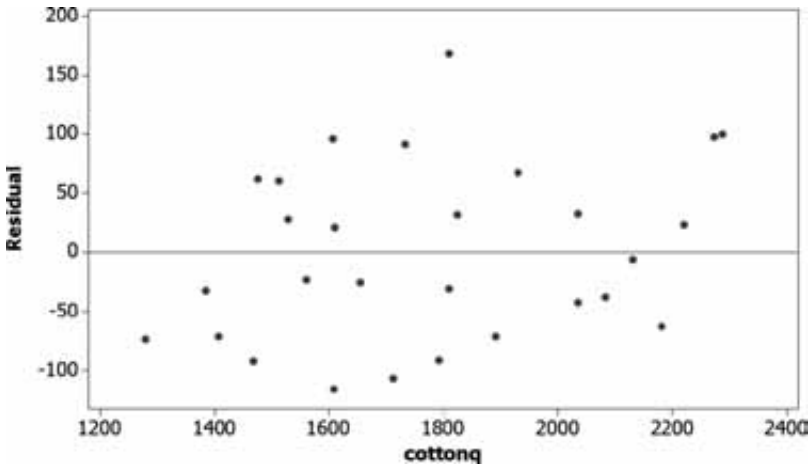


Figura 13.34. Diagrama de puntos dispersos de los residuos en relación con el valor observado del algodón.

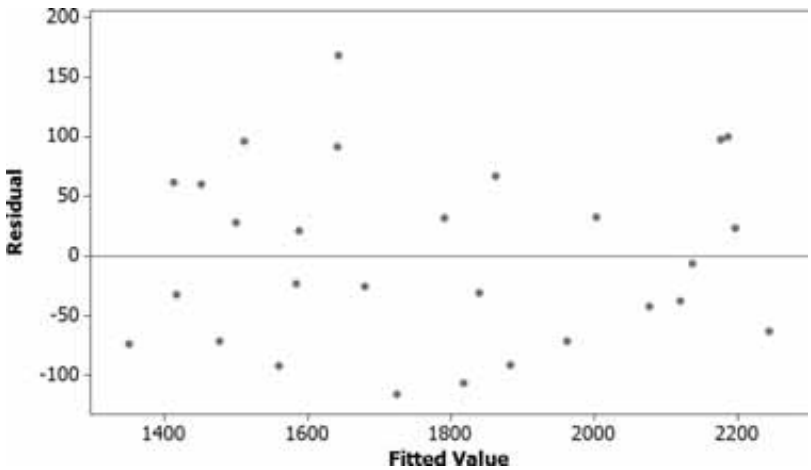


Figura 13.35. Diagrama de puntos dispersos de los residuos en relación con el valor predicho del algodón.

En el Capítulo 14 utilizaremos el análisis de los residuos para identificar dos situaciones del modelo de regresión, la heterocedasticidad y la autocorrelación, que violan el supuesto del análisis de regresión de que la varianza de los errores es la misma en el rango del modelo.

EJERCICIOS

Ejercicios básicos

13.78. Suponga que se incluyen dos variables independientes como variables de predicción en un análisis de regresión múltiple. ¿Cómo cabe esperar que afecte a los coeficientes de la pendiente estimados cuando estas dos variables tienen una correlación igual a

a) 0,78?
b) 0,08?
c) 0,94?
d) 0,33?

13.79. Considere un análisis de regresión con $n = 34$ y cuatro variables independientes posibles. Suponga que una de las variables independientes tiene una correlación de 0,23 con la variable dependiente. ¿Implica eso que esta variable independiente tendrá un estadístico t de Student muy pequeño en el análisis de regresión con las cuatro variables de predicción?

13.80. Considere un análisis de regresión con $n = 47$ y tres variables independientes posibles. Suponga que una de las variables independientes tiene

una correlación de 0,95 con la variable dependiente. ¿Implica eso que esta variable independiente tendrá un estadístico *t* de Student muy grande en el análisis de regresión con las tres variables de predicción?

13.81. Considere un análisis de regresión con $n = 49$ y dos variables independientes posibles. Suponga que una de las variables independientes tiene una correlación de 0,56 con la variable dependiente. ¿Implica eso que esta variable independiente tendrá un estadístico *t* de Student muy pequeño en el análisis de regresión con las dos variables de predicción?

Ejercicios aplicados

13.82. Para averiguar cómo influye en un estado el poder económico de una compañía de seguros de accidentes en su poder político, se desarrolló el siguiente modelo y se ajustó a los datos de los 50 estados de Estados Unidos.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

donde

Y = cociente entre el pago de los impuestos estatales y locales de la empresa, en miles de dólares, y los ingresos fiscales estatales y locales totales en millones de dólares
x_1 = coeficiente de concentración estatal de las compañías de seguros (que mide la concentración de los recursos bancarios)
x_2 = renta per cápita del estado en miles de dólares
x_3 = cociente entre la renta no agrícola y la suma de la renta agrícola y no agrícola
x_4 = cociente entre la renta neta después de impuestos de la compañía de seguros y las reservas de seguro (multiplicado por 1.000)
x_5 = media de las reservas de seguro (dividida por 10.000)

Aquí se muestra parte de la salida informática de la regresión estimada. Realice un informe que resuma los resultados de este estudio.

R-SQUARE = 0.515

Parameter	Estimate	Student's *t* for HO: Parameter = 0	Std. Error of Estimate
Intercept	10.60	2.41	4.40
X1	−0.90	−0.69	1.31
X3	−13.85	−2.83	4.18
X4	0.080	0.50	0.160
X5	0.100	5.00	0.020

13.83. Se pidió a una muestra aleatoria de 93 estudiantes universitarios de primer año de la Universidad de Illinois que valoraran en una escala de 1 (baja) a 10 (alta) su opinión general sobre la vida en la residencia universitaria. También se les pidió que valoraran su nivel de satisfacción con los compañeros, con la planta, con la residencia y con el director de la residencia (se obtuvo información sobre la satisfacción con la habitación, pero ésta se descartó más tarde, porque no suministraba más información para explicar la opinión general). Se estimó el siguiente modelo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

donde

Y = opinión general sobre la residencia
x_1 = satisfacción con los compañeros
x_2 = satisfacción con la planta
x_3 = satisfacción con la residencia
x_4 = satisfacción con el director de la residencia

Utilice la parte de la salida informática de la regresión estimada que se muestra a continuación para realizar un informe que resuma los resultados de este estudio.

DEPENDENT VARIABLE: *Y* OVERALL OPINION

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	*F* VALUE	*R*-SQUARE
MODEL	4	37.016	9.2540	9.958	0.312
ERROR	88	81.780	0.9293		
TOTAL	92	118.79			

PARAMETER	ESTIMATE	STUDENT'S *t* FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	3.950	5.84	0.676
X1	0.106	1.69	0.063
X2	0.122	1.70	0.072
X3	0.092	1.75	0.053
X4	0.169	2.64	0.064

13.84. En un estudio, se ajustó el siguiente modelo a 47 observaciones mensuales en un intento de explicar la diferencia entre los tipos de los certificados de depósito y los tipos del papel comercial:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

donde

Y = tipo de los certificados de depósito menos tipo del papel comercial
x_1 = tipo del papel comercial
x_2 = cociente entre los préstamos y las inversiones y el capital

Utilice la parte de la salida informática de la regresión estimada que se muestra a continuación para escribir un informe que resuma los resultados de este estudio.

R-SQUARE = 0.730

PARAMETER	ESTIMATE	STUDENT'S t FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	−5.559	−4.14	1.343
X1	0.186	5.64	0.033
X2	0.450	2.08	0.216

13.85. Se le ha pedido que desarrolle un modelo de regresión múltiple para predecir el número anual de muertes en carretera en Estados Unidos en función del total de millas recorridas y de la velocidad media. El fichero de datos **Traffic Death Rate** contiene 10 años de datos anuales sobre las tasas de mortalidad por 100 millones de millas-vehículo (y), la distancia total recorrida en miles de millones de millas-vehículo (x_1) y la velocidad media en millas por hora de todos los vehículos (x_2). Calcule la regresión múltiple de y con respecto a x_1 y x_2 y realice un informe que analice sus resultados.

13.86. El fichero de datos **Household Income** contiene datos de los 50 estados de Estados Unidos. Las variables incluidas en el fichero son el porcentaje de mujeres que participan en la población activa (y), la mediana de la renta personal de los hogares (x_1), el número medio de años de estudios de las mujeres (x_2) y la tasa de desempleo de las mujeres (x_3). Calcule la regresión múltiple de y con respecto a x_1, x_2 y x_3 y realice un informe sobre sus resultados.

13.87. Le han pedido que desarrolle un modelo de regresión múltiple que prediga la oferta monetaria real de Alemania en función de la renta y del tipo de interés. El fichero de datos **Real Money** contiene 12 observaciones anuales sobre el dinero real per cápita (y), la renta real per cápita (x_1) y los tipos de interés (x_2) de Alemania. Utilice estos datos para desarrollar un modelo que prediga el dinero real per cápita en función de la renta per cápita y del tipo de interés y realice un informe sobre sus resultados.

13.88. Las Naciones Unidas le han contratado como consultor para ayudar a identificar los factores que predigan el crecimiento de la industria manufacturera de los países en vías de desarrollo. Ha decidido utilizar una regresión múltiple para desarrollar un modelo e identificar las variables importantes que predicen el crecimiento. Ha recogido los datos de 48 países en el fichero de datos **Developing Country**. Las variables incluidas son el crecimiento porcentual de la industria manufacturera (y), el crecimiento agrícola porcentual (x_1), el crecimiento porcentual de las exportaciones (x_2) y la tasa porcentual de inflación (x_3) de 48 países en vías de desarrollo. Desarrolle un modelo de regresión múltiple y escriba un informe sobre sus resultados.

RESUMEN

En este capítulo hemos sentado las bases necesarias para comprender y aplicar los métodos de regresión múltiple. Hemos comenzado analizando detalladamente los supuestos del modelo y las consecuencias de esos supuestos. A partir de ahí, hemos presentado el método de mínimos cuadrados y los métodos para obtener estimaciones de los coeficientes. Con esas bases, hemos desarrollado métodos para averiguar cómo se ajusta el modelo de regresión a los datos observados, lo cual nos ha llevado a desarrollar los métodos clásicos de inferencia para contrastar hipótesis sobre los coeficientes y para construir intervalos de confianza. Eso nos ha llevado a presentar métodos para realizar predicciones de la variable dependiente a partir del modelo e inferencias sobre los valores predichos.

Con estas bases y comprendiendo el modelo básico, hemos pasado a examinar algunas técnicas importantes. Hemos presentado métodos para transformar modelos cuadráticos en funciones lineales. También hemos desarrollado transformaciones para modelos lineales logarítmicos. Por último, hemos comenzado a presentar métodos para utilizar variables ficticias para representar variables de predicción categóricas. El capítulo termina con un extenso modelo de aplicación que muestra cómo realizaría un analista todo el proceso de desarrollo del modelo de regresión. Este proceso comienza con sencillos estadísticos descriptivos, técnicas gráficas y la aplicación de métodos de regresión y termina con un análisis de los residuos para examinar la compatibilidad del modelo con los datos y los supuestos del modelo.

TÉRMINOS CLAVE

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

13.89. El método de mínimos cuadrados se utiliza mucho más a menudo que cualquier otro para estimar los parámetros de un modelo de regresión múltiple. Explique la base de este método de estimación y explique por qué se utiliza tanto.

13.90. Es habitual calcular una tabla del análisis de la varianza junto con una regresión múltiple estimada. Explique detenidamente qué información puede extraerse de esa tabla.

13.91. Indique si cada una de las afirmaciones siguientes es verdadera o falsa.

a) La suma de los cuadrados de los errores debe ser menor que la suma de los cuadrados de la regresión.

b) En lugar de realizar una regresión múltiple, podemos obtener la misma información a partir de regresiones lineales simples de la variable dependiente con respecto a cada variable independiente.

c) El coeficiente de determinación no puede ser negativo.

d) El coeficiente de determinación ajustado no puede ser negativo.

e) El coeficiente de correlación múltiple es la raíz cuadrada del coeficiente de determinación.

13.92. Si se añade una variable independiente más, por irrelevante que sea, a un modelo de regresión múltiple, la suma de los cuadrados de los errores es menor. Explique por qué y analice las consecuencias para la interpretación del coeficiente de determinación.

13.93. Se hace una regresión de una variable dependiente con respecto a dos variables independientes. Es posible que no puedan rechazarse las hipótesis $H_0: \beta_1 = 0$ y $H_0: \beta_2 = 0$ a niveles bajos de significación y, sin embargo, pueda rechazarse la hipótesis $H_0: \beta_1 = \beta_2 = 0$ a un nivel muy bajo de significación. ¿En qué circunstancias podría darse este resultado?

13.94. [*Para hacer este ejercicio es necesario haber leído el apéndice del capítulo*]. Suponga que se estima el modelo de regresión por mínimos cuadrados:

$$Y_1 = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Demuestre que los residuos, e_i, del modelo ajustado suman 0.

13.95. Se realizó un estudio para evaluar la influencia de algunos factores en la creación de nuevas empresas en la industria de chips de computador. Se estimó el siguiente modelo para una muestra de 70 países:

$$\hat{y} = \underset{}{-59{,}31} + \underset{(1{,}156)}{4{,}983}x_1 + \underset{(0{,}210)}{2{,}198}x_2 + \underset{(2{,}063)}{3{,}816}x_3 - \underset{(0{,}330)}{0{,}310}x_4$$

$$- \underset{(3{,}055)}{0{,}886}x_5 + \underset{(1{,}568)}{3{,}215}x_6 + \underset{(0{,}354)}{0{,}085}x_7 \qquad R^2 = 0{,}766$$

donde

y = creación de nuevas empresas en la industria
x_1 = población en millones
x_2 = tamaño de la industria
x_3 = medida de la calidad de vida económica
x_4 = medida de la calidad de vida política

x_5 = medida de la calidad de vida medioambiental
x_6 = medida de la calidad de vida sanitaria y educativa
x_7 = medida de la calidad de vida social

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete los coeficientes de regresión estimados.
b) Interprete el coeficiente de determinación.
c) Halle el intervalo de confianza al 90 por ciento del aumento de la creación de empresas provocado por un aumento de la calidad de vida económica de 1 unidad, manteniéndose todas las demás variables constantes.
d) Contraste al nivel del 5 por ciento la hipótesis nula de que, manteniéndose todo lo demás constante, la calidad de vida medioambiental no influye en la creación de empresas frente a la hipótesis alternativa bilateral.
e) Contraste al nivel del 5 por ciento la hipótesis nula de que, manteniéndose todo lo demás constante, la calidad de vida sanitaria y educativa no influye en la creación de empresas frente a la hipótesis alternativa bilateral.
f) Contraste la hipótesis nula de que estas siete variables independientes, consideradas en conjunto, no influyen en la creación de empresas.

13.96. Una empresa de sondeos realiza habitualmente estudios sobre los hogares por medio de cuestionarios por correo y tiene interés en conocer los factores que influyen en la tasa de respuesta. En un experimento, se enviaron 30 juegos de cuestionarios a posibles encuestados. El modelo de regresión ajustado al conjunto de datos resultantes era

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

donde

Y = porcentaje de respuestas recibidas
x_1 = número de preguntas realizadas
x_2 = longitud del cuestionario en número de palabras

A continuación se muestra una parte de la salida del programa SAS de la regresión estimada.

R-SQUARE = 0.637

PARAMETER	ESTIMATE	STUDENT'S t FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	74.3652		
X1	−1.8345	−2.89	0.6349
X2	−0.0162	−1.78	0.0091

a) Interprete los coeficientes de regresión estimados.
b) Interprete el coeficiente de determinación.
c) Contraste al nivel de significación del 1 por ciento la hipótesis nula de que las dos variables independientes, consideradas en conjunto, no influyen linealmente en la tasa de respuesta.
d) Halle e interprete el intervalo de confianza al 99 por ciento de β_1.
e) Contraste la hipótesis nula

$$H_0 : \beta_2 = 0$$

frente a la hipótesis alternativa

$$H_1 : \beta_2 < 0$$

e interprete sus resultados.

13.97. Una consultora ofrece cursos de gestión financiera para ejecutivos. Al final de estos cursos, se pide a los participantes que hagan una valoración global del valor del curso. Para ver cómo influyen algunos factores en las valoraciones, se ajustó el modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

para 25 cursos, donde

Y = valoración media realizada por los participantes en el curso
x_1 = porcentaje del curso dedicado a realizar sesiones de discusión en grupo
x_2 = cantidad de dinero (en dólares) por asistente al curso dedicado a la preparación del material del curso
x_3 = cantidad de dinero por asistente al curso dedicado a la provisión de material no relacionado con el curso (comida, bebidas, etc.)

A continuación se muestra una parte de la salida del programa SAS de la regresión ajustada.

R-SQUARE = 0.579

PARAMETER	ESTIMATE	STUDENT'S t FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	42.9712		
X1	0.3817	1.89	0.2018
X2	0.5172	2.64	0.1957
X3	0.0753	1.09	0.0693

a) Interprete los coeficientes de regresión estimados.
b) Interprete el coeficiente de determinación.
c) Contraste al nivel de significación del 5 por ciento la hipótesis nula de que las tres variables independientes, consideradas en conjun-

to, no influyen linealmente en la valoración del curso.

d) Halle e interprete el intervalo de confianza al 90 por ciento de β_1.

e) Contraste la hipótesis nula

$$H_0: \beta_2 = 0$$

frente a la hipótesis alternativa

$$H_1: \beta_2 > 0$$

e interprete su resultado.

f) Contraste al nivel del 10 por ciento la hipótesis nula

$$H_0: \beta_3 = 0$$

frente a la hipótesis alternativa

$$H_1: \beta_3 \neq 0$$

e interprete su resultado.

13.98. Al final de las clases, los profesores son evaluados por sus estudiantes en una escala de 1 (malo) a 5 (excelente). También se les pregunta a los estudiantes qué calificación esperan obtener y éstas se codifican de la forma siguiente: A = 4, B = 3, etc. El fichero de datos **Teacher Rating** contiene las evaluaciones de los profesores, las calificaciones medias esperadas y el número de estudiantes de las clases de una muestra aleatoria de 20 clases. Calcule la regresión múltiple de la evaluación con respecto a la calificación esperada y el número de estudiantes y realice un informe sobre sus resultados.

13.99. Sistemas Informáticos Voladores, S.A., quiere saber cómo afectan algunas variables a la eficiencia del trabajo. Basándose en una muestra de 64 observaciones, estimó el siguiente modelo por mínimos cuadrados:

$$\hat{y} = -16{,}528 + 28{,}729x_1 + 0{,}022x_2 - 0{,}023x_3 - 0{,}054x_4 - 0{,}077x_5 + 0{,}411x_6 + 0{,}349x_7 + 0{,}028x_8 \qquad R^2 = 0{,}467$$

donde

y = índice de eficiencia directa del trabajo en la planta de producción
x_1 = cociente entre las horas extraordinarias y las horas ordinarias realizadas por todos los obreros
x_2 = número medio de trabajadores por hora en la planta
x_3 = porcentaje de asalariados que participan en algún programa de calidad de vida laboral
x_4 = número de reclamaciones recibidas por cada 100 trabajadores
x_5 = tasa de acciones disciplinarias
x_6 = tasa de absentismo de los trabajadores por hora
x_7 = actitudes de los trabajadores asalariados, desde baja (insatisfechos) hasta alta, medidas por medio de un cuestionario.
x_8 = porcentaje de trabajadores por hora que hacen al menos una sugerencia en un año al programa de sugerencias de la planta.

También se obtuvo por mínimos cuadrados un modelo ajustado a partir de estos datos:

$$\hat{y} = 9{,}062 - 10{,}944x_1 + 0{,}320x_2 + 0{,}019x_3 \qquad R^2 = 0{,}242$$

Las variables x_4, x_5, x_6, x_7 y x_8 son medidas de los resultados de un sistema de relaciones laborales de la planta. Contraste al nivel del 1 por ciento la hipótesis nula de que no contribuyen a explicar la eficiencia directa del trabajo, dado que también se utilizan x_1, x_2 y x_3.

13.100. Basándose en las calificaciones obtenidas por 107 estudiantes en el primer examen de un curso de estadística para los negocios, se estimó el siguiente modelo por mínimos cuadrados:

$$\hat{y} = \underset{}{2{,}178} + \underset{(0{,}090)}{0{,}469}x_1 + \underset{(0{,}456)}{3{,}369}x_2 + \underset{(1{,}457)}{3{,}054}x_3 \qquad R^2 = 0{,}686$$

donde

y = calificación efectiva del estudiante en el examen
x_1 = calificación esperada por el estudiante en el examen
x_2 = horas semanales dedicadas a estudiar para el curso
x_3 = calificación media del estudiante

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Interprete la estimación de β_1.

b) Halle e interprete el intervalo de confianza al 95 por ciento de β_2.

c) Contraste la hipótesis nula de que β_3 es 0 frente a una hipótesis alternativa bilateral e interprete su resultado.

d) Interprete el coeficiente de determinación.

e) Contraste la hipótesis nula de que

$$\beta_1 = \beta_2 = \beta_3 = 0$$

f) Halle e interprete el coeficiente de correlación múltiple.

g) Prediga la calificación de un estudiante que espera una calificación de 80, estudia 8 horas a la semana y tiene una calificación media de 3,0.

13.101. Basándose en 25 años de datos anuales, se intentó explicar el ahorro en la India. El modelo ajustado era

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

donde

y = variación del tipo real de los depósitos
x_1 = variación de la renta real per cápita
x_2 = variación del tipo de interés real

Las estimaciones de los parámetros por mínimos cuadrados (con los errores típicos entre paréntesis) eran (véase la referencia bibliográfica 1)

$$b_1 = 0{,}0974(0{,}0215) \qquad b_2 = 0{,}374(0{,}209)$$

El coeficiente de determinación corregido era

$$\bar{R}^2 = 0{,}91$$

a) Halle e interprete el intervalo de confianza al 99 por ciento de β_1.
b) Contraste la hipótesis nula de que β_2 es 0 frente a la hipótesis alternativa de que es positivo.
c) Halle el coeficiente de determinación.
d) Contraste la hipótesis nula de que $\beta_1 = \beta_2 = 0$.
e) Halle e interprete el coeficiente de correlación múltiple.

13.102. Basándose en datos de 2.679 jugadores de baloncesto de centros de enseñanza secundaria, se ajustó el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_9 x_{91} + \varepsilon_i$$

donde

y = minutos jugados en la temporada
x_1 = porcentaje de tiros de 2 puntos convertidos
x_2 = porcentaje de tiros libres
x_3 = rebotes por minuto
x_4 = puntos por minuto
x_5 = faltas por minuto
x_6 = robos de balón por minuto
x_7 = tapones por minuto
x_8 = pérdidas de balón por minuto
x_9 = asistencias por minuto

Las estimaciones de los parámetros por mínimos cuadrados (con los errores típicos entre paréntesis) son

$b_0 = 358{,}848\ (44{,}695)$ $\quad b_1 = 0{,}6742\ (0{,}0639)$
$b_2 = 0{,}2855\ (0{,}0388)$ $\quad b_3 = 303{,}81\ (77{,}73)$
$b_4 = 504{,}95\ (43{,}26)$ $\quad b_5 = -3.923{,}5\ (120{,}6)$
$b_6 = 480{,}04\ (224{,}9)$ $\quad b_7 = 1.350{,}3\ (212{,}3)$
$b_8 = -891{,}67\ (180{,}87)$ $\quad b_9 = 722{,}95\ (110{,}98)$

El coeficiente de determinación es

$$R^2 = 0{,}5239$$

a) Halle e interprete el intervalo de confianza al 90 por ciento de β_6.
b) Halle e interprete el intervalo de confianza al 99 por ciento de β_7.
c) Contraste la hipótesis nula de que β_8 es 0 frente a la hipótesis alternativa de que es negativo. Interprete su resultado.
d) Contraste la hipótesis nula de que β_9 es 0 frente a la hipótesis alternativa de que es positivo. Interprete su resultado.
e) Interprete el coeficiente de determinación.
f) Halle e interprete el coeficiente de correlación múltiple.

13.103. Basándose en datos de 63 regiones, se estimó el siguiente modelo por mínimos cuadrados:

$$\hat{y} = 0{,}58 - \underset{(0{,}019)}{0{,}052 x_1} - \underset{(0{,}042)}{0{,}005 x_2} \qquad R^2 = 0{,}17$$

donde

y = tasa de crecimiento del producto interior bruto real
x_1 = renta real per cápita
x_2 = tipo impositivo medio en porcentaje del producto nacional bruto

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

a) Contraste la hipótesis nula de que β_1 es 0 frente a una hipótesis alternativa bilateral. Interprete su resultado.
b) Contraste la hipótesis nula de que β_2 es 0 frente a una hipótesis alternativa bilateral. Interprete su resultado.
c) Interprete el coeficiente de determinación.
d) Halle e interprete el coeficiente de correlación múltiple.

13.104. En un estudio, se ajustó el siguiente modelo de regresión a los datos de 60 golfistas amateurs:

$$\hat{y} = 164.683 + \underset{(100{,}59)}{341{,}10 x_1} + \underset{(167{,}18)}{170{,}02 x_2} + \underset{(305{,}48)}{495{,}19 x_3} - \underset{(90{,}0)}{4{,}23 x_4}$$
$$- \underset{(25.634)}{136.040 x_5} - \underset{(16.240)}{35.549 x_6} + \underset{(106{,}20)}{202{,}52 x_7} \qquad \bar{R}^2 = 0{,}516$$

donde

y = ganancias por torneo en dólares
x_1 = longitud media del golpe
x_2 = porcentaje de veces en que el golpe acaba en la pista
x_3 = porcentaje de veces en que se llega en buena posición al «green» («regulation»)

x_4 = porcentaje de veces en que se consigue el par después de haber caído en zona de arena
x_5 = número medio de «putts» realizados en los «greens» a los que se ha llegado en buena posición
x_6 = número medio de «putts» realizados en los «greens» a los que no se ha llegado en buena posición
x_7 = número de años que lleva jugando el golfista amateur.

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes estimados.

Realice un informe que resuma lo que ha aprendido con estos resultados.

13.105. El Departamento de Economía quiere desarrollar un modelo de regresión múltiple para predecir la calificación media (GPA) de los estudiantes en los cursos de economía. El profesorado del departamento ha reunido datos de 112 licenciados, que contienen las variables GPA de economía, SAT verbal, SAT de matemáticas, ACT de inglés, ACT de ciencias sociales y puesto obtenido en el bachillerato (*rank*). Los datos se encuentran en el fichero de datos llamado **Student GPA** de su disco de datos. El apéndice contiene una descripción de las variables.

a) Utilice las variables SAT y «rank» para averiguar cuál es el mejor modelo de predicción. Elimine las variables independientes que no sean significativas. ¿Cuáles son los coeficientes, su estadístico *t* de Student y el modelo?
b) Utilice las variables ACT y «rank» para averiguar cuál es el mejor modelo de predicción. Elimine las variables independientes que no sean significativas. ¿Cuáles son los coeficientes, su estadístico *t* de Student y el modelo?
c) ¿Qué modelo predice mejor la GPA de economía? Aporte pruebas para apoyar su conclusión.

13.106. El fichero de datos **Salary Model** contiene una variable dependiente y siete variables independientes. Tiene que desarrollar el «mejor» modelo de regresión que prediga *Y* en función de las siete variables independientes. Los datos se encuentran en su disco de datos.

La variable dependiente se llama «*Y*» en el fichero y las variables independientes también tienen su propio nombre. Utilice un análisis de regresión para averiguar qué variables deben estar en el modelo final y para estimar los coeficientes. Muestre el contraste *F* condicionado y el contraste *t* condicionado de cualquier variable eliminada. Analice los residuos del modelo por medio de gráficos. Muestre sus resultados y analice sus conclusiones. Transforme las variables si los residuos indican una relación no lineal. Presente claramente su modelo final, mostrando los coeficientes y los estadísticos *t* de Student de los coeficientes.

13.107. Utilice los datos del fichero **Citydat** para estimar una ecuación de regresión que pueda utilizarse para averiguar el efecto marginal que produce el porcentaje de locales comerciales en el valor de mercado por vivienda ocupada por su propietario. Incluya en su ecuación de regresión múltiple el porcentaje de viviendas ocupadas por su propietario, el porcentaje de locales industriales, el número mediano de habitaciones por vivienda y la renta per cápita como variables de predicción adicionales. Las variables están en su disco de datos y se describen en el apéndice. Indique cuáles son significativas. Su ecuación final debe incluir únicamente las variables significativas. Analice e interprete su modelo final de regresión e indique cómo seleccionaría una ciudad para comprar su vivienda.

13.108. Los responsables de la National Highway Traffic Safety Administration (NHTSA) de Estados Unidos quieren saber si los diferentes tipos de vehículos de un estado tienen relación con la tasa de mortalidad en carretera del estado. Le han pedido que desarrolle varios análisis de regresión múltiple para averiguar si el peso medio de los vehículos, el porcentaje de vehículos importados, el porcentaje de camiones ligeros y la antigüedad media de los automóviles están relacionados con las muertes en accidente ocurridas en automóviles y camionetas. Los datos del análisis se encuentran en el fichero de datos llamado **Crash**, que está en su disco de datos.

a) Prepare una matriz de correlaciones de las muertes en accidente y las variables de predicción. Observe las relaciones simples entre las muertes en accidente y las variables de predicción. Indique además cualquier problema posible de multicolinealidad entre las variables de predicción.

b) Realice un análisis de regresión múltiple de las muertes en accidente con respecto a las variables de predicción posibles. Elimine en el modelo de regresión cualquier variable de predicción no significativa, una de cada vez. Indique su mejor modelo final.

c) Exponga las conclusiones de su análisis y analice la importancia condicionada de las variables desde el punto de vista de su relación con las muertes en accidente.

13.109. El Departamento de Transporte de Estados Unidos quiere saber si los estados que tienen un porcentaje mayor de población urbana tienen una tasa más alta de muertes totales en accidente ocurridas en automóviles y camionetas. También quiere saber si la velocidad media a la que se conduce por las carreteras rurales o el porcentaje de carreteras rurales que está asfaltado están relacionados con las tasas de muertes en accidente, dado el porcentaje de población urbana. Los datos de este estudio se encuentran en el fichero de datos **Crash** almacenado en su disco de datos.

a) Prepare una matriz de correlaciones y estadísticos descriptivos de las muertes en accidente y las variables de predicción posibles. Señale las relaciones y cualquier problema posible de multicolinealidad.

b) Realice un análisis de regresión múltiple de las muertes en accidente con respecto a las variables de predicción posibles. Averigüe cuáles de las variables deben mantenerse en el modelo de regresión porque tienen una relación significativa.

c) Muestre los resultados de su análisis desde el punto de vista de su modelo final de regresión. Indique qué variables son significativas.

13.110. Un economista desea predecir el valor de mercado de las viviendas de pequeñas ciudades del Medio Oeste ocupadas por sus propietarios. Ha reunido un conjunto de datos de 45 pequeñas ciudades que se refieren a un periodo de dos años y quiere que los utilice como fuente de datos para el análisis. Los datos se encuentran en el fichero **Citydat**, que está en su disco de datos. Quiere que desarrolle una ecuación de predicción basada en una regresión múltiple. Las variables de predicción posibles son el tamaño de la vivienda, el tipo impositivo, el porcentaje de locales comerciales, la renta per cápita y el gasto público municipal total.

a) Calcule la matriz de correlaciones y estadísticos descriptivos del valor de mercado de las viviendas y las variables de predicción posibles. Señale los problemas posibles de multicolinealidad. Defina el rango aproximado para su modelo de regresión utilizando la regla siguiente: medias de las variables $\pm$ 2 desviaciones típicas.

b) Realice análisis de regresión múltiple utilizando las variables de predicción. Elimine las variables que no sean significativas. ¿Qué variable, el tamaño de la vivienda o el tipo impositivo, tiene la relación condicionada más estrecha con el valor de las viviendas?

c) Un promotor industrial de un estado del Medio Oeste ha afirmado que los tipos de los impuestos locales sobre bienes inmuebles de las pequeñas ciudades deben bajarse, ya que, de lo contrario, nadie comprará una vivienda en estas ciudades. Basándose en su análisis de este problema, evalúe la afirmación del promotor.

13.111. Stuart Wainwright, vicepresidente de compras para una gran cadena nacional de tiendas de Estados Unidos, le ha pedido que realice un análisis de las ventas al por menor por estados. Quiere saber si el porcentaje de desempleados o la renta personal per cápita están relacionados con las ventas al por menor per cápita. Los datos para realizar este estudio se encuentran en el fichero de datos llamado **Retail**, que está almacenado en su disco de datos.

a) Prepare una matriz de correlaciones, calcule los estadísticos descriptivos y realice un análisis de regresión de las ventas al por menor per cápita con respecto al porcentaje de desempleados y a la renta personal. Calcule intervalos de confianza al 95 por ciento de los coeficientes de la pendiente de cada ecuación de regresión.

b) ¿Cuál es el efecto condicionado de una disminución de la renta per cápita de 1.000 $ en las ventas per cápita?

c) ¿Mejoraría la ecuación de predicción añadiendo la población de los estados como una variable de predicción adicional?

13.112. Un importante proveedor nacional de materiales de construcción para la construcción de viviendas está preocupado por las ventas totales del próximo año. Es bien sabido que las ventas de la empresa están relacionadas directamente con la inversión nacional total en

vivienda. Algunos banqueros de Nueva York están prediciendo que los tipos de interés subirán alrededor de 2 puntos porcentuales el próximo año. Le han pedido que realice un análisis de regresión para poder predecir el efecto de las variaciones de los tipos de interés en la inversión en vivienda. Usted cree que, además del tipo de interés, el PNB, la oferta monetaria, el gasto público y el índice de precios de los bienes acabados podrían ser predictores de la inversión en vivienda, por lo que llega a la conclusión de que necesita dos modelos de regresión múltiple. Uno incluirá el tipo de interés preferencial y otras importantes variables. El otro incluirá el tipo de interés de los fondos federales y otras importantes variables. Los datos de series temporales para realizar este estudio se encuentran en el fichero de datos llamado **Macro2003**, que está almacenado en su disco de datos y se describe en el apéndice del Capítulo 14.

a) Desarrolle dos modelos de regresión para predecir la inversión en vivienda utilizando el tipo de interés preferencial para uno y el tipo de interés de los fondos federales para el otro. Los modelos finales de regresión deben incluir solamente variables de predicción que produzcan un efecto condicionado significativo. Analice los estadísticos de la regresión e indique qué ecuación hace las mejores predicciones.

b) Halle el intervalo de confianza al 95 por ciento del coeficiente de la pendiente del tipo de interés en ambas ecuaciones de regresión.

13.113. La Congressional Budget Office (CBO) de Estados Unidos tiene interés en saber si las tasas de mortalidad infantil de los estados están relacionadas con el nivel de recursos médicos de que dispone cada uno. Los datos para el estudio se encuentran en el fichero de datos llamado **State**, que está almacenado en su disco de datos. La medida de la mortalidad infantil son las muertes de niños de menos de 1 año por cada 100 nacidos vivos. El conjunto de variables de predicción posibles son los médicos por 100.000 habitantes, la renta personal per cápita y los gastos totales de los hospitales (esta variable debe expresarse en magnitudes per cápita dividiendo por la población del estado).

a) Realice un análisis de regresión múltiple y averigüe qué variables de predicción deben incluirse en el modelo de regresión múltiple. Interprete su modelo final de regresión y analice los coeficientes, sus estadísticos *t* de Student, el error típico de la estimación y el R^2.

b) Identifique dos variables más que podrían ser predictores adicionales si se añadieran al modelo de regresión múltiple. Contraste su efecto en un análisis de regresión múltiple e indique si sus sospechas iniciales eran correctas.

13.114. Desarrolle un modelo de regresión múltiple para predecir el salario en función de otras variables independientes utilizando los datos del fichero **Salary Model**, que se encuentra en su disco de datos. Para este problema no utilice los años de experiencia sino la edad como sucedáneo de la experiencia.

a) Describa los pasos seguidos para obtener el modelo final de regresión.

b) Contraste la hipótesis de que la tasa de variación de los salarios femeninos en función de la edad es menor que la tasa de variación de los salarios masculinos en función de la edad. Debe formular su contraste de hipótesis de manera que aporte pruebas contundentes de la existencia de discriminación de las mujeres [*nota*: las mujeres se indican mediante un «1» en la variable «sexo» en la columna 5; el contraste debe realizarse condicionado a las demás variables de predicción significativas del apartado (a)].

13.115. Un grupo de activistas de Peaceful (Montana) está tratando de aumentar el desarrollo de su prístino enclave, que ha sido objeto de algún reconocimiento nacional en el programa de televisión *Four Dirty Old Men*. Sostienen que un aumento del desarrollo comercial e industrial traerá mayor prosperidad e impuestos más bajos a Peaceful. Concretamente, sostienen que un aumento del porcentaje de locales comerciales e industriales reducirá el tipo del impuesto sobre bienes inmuebles y aumentará el valor de mercado de las viviendas ocupadas por sus propietarios.

Le han contratado para analizar sus afirmaciones. Para ello ha obtenido el fichero de datos **Citydat**, que contiene datos de 45 pequeñas ciudades. Con estos datos, primero desarrolla modelos de regresión que predicen el valor medio de las viviendas ocupadas por sus propietarios y el tipo del impuesto sobre bienes inmuebles. A continuación, averigua si y cómo la

adición del porcentaje de locales comerciales y del porcentaje de locales industriales afecta a la variabilidad en estos modelos de regresión. El modelo básico para predecir el valor de mercado de las viviendas (c10) incluye como variables independientes el tamaño de la vivienda (c4), el tipo impositivo (c7), la renta per cápita (c9) y el porcentaje de viviendas ocupadas por sus propietarios (c12). El modelo básico para predecir el tipo impositivo (c7) incluye como variables independientes el valor catastral (c6), los gastos municipales actuales per cápita (c5/c8) y el porcentaje de viviendas ocupadas por sus propietarios (c12).

Averigüe si el porcentaje de locales comerciales (c14) y el porcentaje de locales industriales (c15) mejoran la variabilidad explicada en cada uno de los dos modelos. Realice un contraste F condicionado de cada una de estas variables adicionales. Primero estime el efecto condicionado del porcentaje de locales comerciales por sí solo y, a continuación, el de locales industriales por sí solo. Explique detenidamente los resultados de su análisis. Incluya en su informe una explicación de por qué es importante incluir todas las demás variables en el modelo de regresión en lugar de examinar simplemente el efecto de la relación directa y simple entre el porcentaje de locales comerciales y el de locales industriales en el tipo impositivo y en el valor de mercado de la vivienda.

13.116. Utilice los datos del fichero de datos llamado **Student GPA**, que se encuentra en su disco de datos y se describe en el apéndice, a fin de desarrollar un modelo para predecir la calificación media (GPA) de economía de un estudiante. Comience con las variables «ACT scores», «gender» y «HSpct».

a) Utilice métodos estadísticos adecuados para elegir un subconjunto de variables de predicción estadísticamente significativas. Describa su estrategia y defina minuciosamente su modelo final.

b) Explique cómo podría utilizar la comisión de admisiones de la universidad este modelo para tomar sus decisiones.

13.117. Un economista estimó para una muestra aleatoria de 50 observaciones el modelo de regresión

$$\text{Log}\,\hat{y}_i = \alpha + \beta_1 \log x_{1i} + \beta_2 \log x_{2i} + \beta_3 \log x_{3i} + \beta_4 \log x_{4i} + \varepsilon_i$$

donde

y = ingresos brutos generados por una práctica médica

x_{1i} = número medio de horas trabajadas por los médicos en la práctica

x_{2i} = número de médicos en la práctica

x_{3i} = número de personal sanitario auxiliar (como enfermeras) empleado en la práctica

x_{4i} = número de habitaciones utilizadas en la práctica

Utilice la parte de la salida informática mostrada aquí para realizar un informe sobre estos resultados.

R-SQUARE = 0.927

PARAMETER	ESTIMATE	STUDENT'S t FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	2.347		
LOG X1	0.239	3.27	0.073
LOG X2	0.673	8.31	0.081
LOG X3	0.279	6.64	0.042
LOG X4	0.082	1.61	0.051

Apéndice

1. Obtención de los estimadores por mínimos cuadrados

Los estimadores de los coeficientes de un modelo con dos variables de predicción se obtienen de la forma siguiente:

$$\hat{y}_1 = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

Se minimiza

$$SCE = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i})]^2$$

Aplicando el cálculo diferencial, obtenemos un conjunto de tres ecuaciones normales que pueden resolverse para hallar los estimadores de los coeficientes:

$$\frac{\partial SCE}{\partial b_0} = 0$$

$$2 \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i})](-1) = 0$$

$$\sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_{1i} - b_2 \sum_{i=1}^{n} x_{2i} = 0$$

$$nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} x_{1i}$$

$$\frac{\partial SCE}{\partial b_1} = 0$$

$$2 \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i})](-x_{1i}) = 0$$

$$\sum_{i=1}^{n} x_{1i} y_i - b_0 \sum_{i=1}^{n} x_{1i} - b_1 \sum_{i=1}^{n} x_{1i}^2 - b_2 \sum_{i=1}^{n} x_{1i} x_{2i} = 0$$

$$b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} x_{2i} y_i$$

$$\frac{\partial SCE}{\partial b_2} = 0$$

$$2 \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i})](-x_{2i}) = 0$$

$$\sum_{i=1}^{n} x_{2i} y_1 - b_0 \sum_{i=1}^{n} x_{2i} - b_1 \sum_{i=1}^{n} x_{1i} x_{2i} - b_2 \sum_{i=1}^{n} x_{2i}^2 = 0$$

$$b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i} x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} x_{2i} y_i$$

Como consecuencia de la aplicación del algoritmo de los mínimos cuadrados, tenemos un sistema de tres ecuaciones lineales con tres incógnitas, b_0, b_1 y b_2:

$$nb_0 + b_1 \sum_{i=1}^{n} x_{1i} + b_2 \sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 + b_2 \sum_{i=1}^{n} x_{1i} x_{2i} = \sum_{i=1}^{n} x_{1i} y_i$$

$$b_0 \sum_{i=1}^{n} x_{2i} + b_1 \sum_{i=1}^{n} x_{1i} x_{2i} + b_2 \sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} x_{2i} y_i$$

Se resuelven las ecuaciones normales para obtener los coeficientes deseados calculando primero los distintos cuadrados de X e Y y los términos que incluyen los productos entre ellas.

El término de la ordenada en el origen se estima de la forma siguiente:

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

2. Variabilidad total explicada

El término *SCR* de la variabilidad explicada en la regresión múltiple es más complejo que el término *SCR* calculado en la regresión simple.

En el modelo de regresión con dos variables independientes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

observamos que

$$\begin{aligned} SCR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} [b_0 + b_1 x_{1i} + b_2 x_{2i} - (b_0 + b_1\bar{x}_1 + b_2\bar{x}_2)]^2 \\ &= \sum_{i=1}^{n} [b_1^2(x_{1i} - \bar{x}_1)^2 + b_2^2(x_{2i} - \bar{x}_1)^2 + 2b_1b_2(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)] \\ &= (n-1)(b_1^2 s_{x_1}^2 + b_2^2 s_{x_2}^2 + 2r_{x_1x_2} b_1 b_2 s_{x_1} s_{x_2}) \end{aligned}$$

Vemos que la variabilidad explicada tiene una parte relacionada directamente con cada una de las variables independientes y una parte relacionada con la correlación entre las dos variables.

Bibliografía

1. Ghatak, S. y D. Deadman, «Money, Prices and Stabilization Policies in Some Developing Countries», *Applied Economics*, 21, 1989, págs. 853-865.
2. Hagermann, R. P., «The Determinants of Household Vacation Travel: Some Empirical Evidence», *Applied Economics*, 13, 1981, págs. 225-234.
3. MacDonald, J. M. y P. E. Nelson, «Do the Poor Still Pay More? Food Price Variations in Large Metropolitan Areas», *Journal of Urban Economics*, 30, 1991, págs. 344-359.
4. Spellman, L. J., «Entry and Profitability in a Rate-free Savings and Loan Market», *Quarterly Review of Economics and Business*, 18, n.º 2, 1978, págs. 87-95.
5. Van Scyoc, L. J. y J. Gleason, «Traditional or Intensive Course Lenghts? A Comparison of Outcomes in Economics Learning», *Journal of Economic Education*, 24, 1993, págs. 15-22.

Capítulo 14

Otros temas del análisis de regresión

Esquema del capítulo

Introducción

En los Capítulos 12 y 13 presentamos la regresión simple y la regresión múltiple como instrumentos para estimar los coeficientes de modelos lineales para aplicaciones empresariales y económicas. Ahora comprendemos que el fin de ajustar una ecuación de regresión es utilizar la información sobre las variables independientes para explicar la conducta de las variables dependientes y para hacer predicciones de la variable dependiente. Los coeficientes del modelo también pueden utilizarse para estimar la tasa de variación de la variable dependiente como consecuencia de las variaciones de una variable independiente, siempre y cuando el conjunto específico de otras variables independientes incluidas en el modelo se mantenga fijo. En este capítulo estudiamos un conjunto de especificaciones alternativas. Consideramos, además, situaciones en las que se violan los supuestos básicos del análisis de regresión.

El lector puede seleccionar los temas de este capítulo para complementar su estudio del análisis de regresión. A casi todo el mundo le interesará el análisis de la construcción de modelos del apartado siguiente. El proceso de construcción de modelos es fundamental para todas las aplicaciones del análisis de regresión, por lo que comenzamos con esas ideas. El apartado sobre las variables ficticias y el diseño experimental contiene métodos para extender las aplicaciones de los modelos. Los apartados como el de la heterocedasticidad y las autocorrelaciones indican cómo se aborda la cuestión de las violaciones de los supuestos.

Se desarrollan modelos de regresión en aplicaciones empresariales y económicas para aumentar la comprensión y servir de orientación para tomar decisiones. Para desarrollar estos modelos, es necesario comprender bien el sistema y el proceso estudiados. La teoría estadística sirve de nexo entre el proceso subyacente y los datos observados en ese proceso. Esta relación entre el contexto del problema y un buen análisis estadístico normalmente requiere un equipo interdisciplinar que pueda aportar sus conocimientos sobre todos los aspectos del problema. Los autores piensan por experiencia que estos equipos sólo tendrán éxito cuando todos sus miembros aprendan unos de otros: los expertos en producción deben tener unos conocimientos básicos de los métodos estadísticos y los estadísticos deben comprender el proceso de producción.

14.1. Metodología para la construcción de modelos

Aquí desarrollamos una estrategia general para construir modelos de regresión. Vivimos en un mundo complejo y nadie cree que podamos recoger exactamente las complejidades de la conducta económica y empresarial en una o más ecuaciones. Nuestro objetivo es utilizar un modelo relativamente sencillo que refleje la compleja realidad con la suficiente precisión como para que aporte útiles ideas. El arte de la construcción de modelos reconoce la imposibilidad de representar todos los factores que influyen en una variable dependiente y trata de seleccionar las variables más influyentes. A continuación, es necesario formular un modelo para representar las relaciones entre estos factores. Queremos construir un sencillo modelo que sea fácil de interpretar, pero no tan excesivamente simplificado que no tenga en cuenta las influencias importantes.

El proceso de construcción de modelos estadísticos depende de cada problema. Nuestro enfoque depende de la información de que se dispone sobre la conducta de las cantidades estudiadas y de los datos existentes. En la Figura 14.1 presentamos las distintas fases de la construcción de modelos.

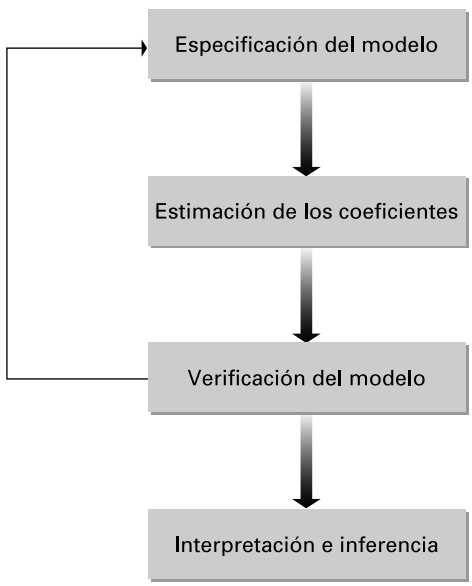


Figura 14.1. Fases de la construcción de modelos estadísticos.

Especificación del modelo

El análisis comienza con el desarrollo de la especificación del modelo. Comprende la selección de la variable dependiente y de las variables independientes y la forma algebraica del modelo. Buscamos una especificación que represente correctamente el sistema y el proceso estudiados. Los ejemplos de los Capítulos 12 y 13 que se refieren a las ventas al por menor, la rentabilidad de las asociaciones de ahorro y crédito inmobiliario y la producción de algodón postulaban todos ellos una relación lineal entre la variable dependiente y las variables independientes. Los modelos lineales a menudo reflejan bien el problema de interés. Pero no siempre es así.

La especificación del modelo comienza con la comprensión de la teoría que constituye el contexto para el modelo. Debemos estudiar detenidamente la literatura existente y enterarnos de qué se sabe sobre la situación de la que tratamos de desarrollar un modelo. Este estudio debe incluir la realización de consultas a los que conocen el contexto, a los que han hecho investigaciones sobre el tema y a los que han desarrollado modelos parecidos. Cuando se trata de estudios aplicados, también debe entrarse en contacto con los profesionales con experiencia que conocen en la práctica el sistema que se pretende estudiar.

La especificación del modelo normalmente exige un profundo estudio del sistema y del proceso que subyace al problema. Cuando tenemos complejos problemas en los que intervienen varios factores, es importante que el equipo interdisciplinario analice minuciosamente todos los aspectos del problema. Puede ser necesario realizar más investigaciones y quizá incluir a otros que tengan ideas importantes. La especificación requiere un estudio y un análisis serios. Éste también es el momento en el que es necesario decidir los datos necesarios para el estudio. En muchos casos, eso puede significar decidir si los datos existentes —o los que podrían obtenerse— serán adecuados para estimar el modelo. Si no sabemos lo que queremos hacer o no comprendemos el contexto del problema, hay sofisticados instrumentos analíticos y analistas competentes que nos darán la mejor respuesta posible. Los analistas sin experiencia a menudo realizan cálculos por computador antes de analizar minuciosamente el problema. Los analistas profesionales saben que con ese enfoque se obtienen resultados inferiores.

Estimación de los coeficientes

Un modelo estadístico, una vez especificado, normalmente tiene algunos coeficientes desconocidos, llamados parámetros. El paso siguiente del ejercicio de construcción de un modelo es emplear los datos de los que se dispone en la estimación de estos coeficientes. Deben realizarse estimaciones puntuales y estimaciones de intervalos para el modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

Desde el punto de vista estadístico, los objetivos del modelo de regresión pueden dividirse en la predicción de la media de la variable dependiente, Y, o la estimación de uno o más de los coeficientes individuales, β_j. En muchos casos, los objetivos no son totalmente independientes, pero estas alternativas identifican importantes opciones.

Si el objetivo es la predicción, queremos un modelo en el que el error típico de la estimación, s_e, sea pequeño. No nos preocupa tanto que las variables independientes estén correlacionadas, porque sabemos que la precisión de la predicción será la misma con una serie de diferentes combinaciones de variables correlacionadas. Sin embargo, necesitamos

saber si las correlaciones entre las variables independientes continuarán cumpliéndose en futuras poblaciones. También necesitamos que las variables independientes tengan una amplia dispersión para que la varianza de la predicción sea pequeña en el rango deseado de la aplicación del modelo.

Si el objetivo es la estimación, la estimación de los coeficientes de la pendiente nos lleva a considerar una variedad mayor de cuestiones. En la desviación típica estimada, s_{b_j}, de los coeficientes de la pendiente influye directamente el error típico del modelo e inversamente la dispersión de las variables independientes y las correlaciones entre las variables independientes, como se observa en el apartado 13.4. La multicolinealidad —las correlaciones entre variables independientes— es una cuestión fundamental, como veremos en el apartado 14.5. También veremos en el apartado 14.4 que cuando no se incluyen importantes variables de predicción, el estimador de los coeficientes de las variables de predicción incluidas en el modelo es un estimador sesgado. Estos dos resultados llevan a un problema estadístico clásico. ¿Incluimos una variable de predicción que está estrechamente correlacionada con las demás para evitar una estimación sesgada de los coeficientes pero aumentamos también considerablemente la varianza del estimador de los coeficientes? ¿O excluimos una variable de predicción correlacionada para reducir la varianza del estimador de los coeficientes pero aumentamos el sesgo? La selección del equilibrio adecuado entre el sesgo del estimador y la varianza a menudo es un problema en la construcción de un modelo aplicado.

Verificación del modelo

Cuando desarrollamos la especificación del modelo, incorporamos ideas sobre la conducta del sistema y el proceso subyacentes. Cuando se trasladan estas ideas a formas algebraicas y cuando se seleccionan datos para estimar el modelo, se realizan algunas simplificaciones y se postulan algunos supuestos. Como algunos pueden resultar insostenibles, es importante comprobar la adecuación del modelo.

Después de estimar una ecuación de regresión, podemos observar que las estimaciones no tienen sentido, dado lo que sabemos del proceso. Supongamos, por ejemplo, que el modelo indica que la demanda de automóviles aumenta cuando suben los precios, lo cual es contrario a la teoría económica básica. Ese resultado puede deberse a que los datos no son adecuados o a que existen algunas correlaciones estrechas entre el precio y otras variables de predicción. Éstas son las razones por las que el signo de los coeficientes puede ser incorrecto. Pero el problema también puede deberse a que el modelo no se ha especificado correctamente. Si no se incluye el conjunto adecuado de variables de predicción, los coeficientes pueden estar sesgados y los signos ser incorrectos. También es necesario verificar los supuestos postulados sobre las variables aleatorias del modelo. Por ejemplo, los supuestos básicos del análisis de regresión establecen que los términos de error tienen todos ellos la misma varianza y no están correlacionados entre sí. En los apartados 14.6 y 14.7 vemos cómo pueden comprobarse estos supuestos utilizando los datos existentes.

Si obtenemos resultados inverosímiles, tenemos que examinar nuestros supuestos, la especificación del modelo y los datos. Eso puede llevarnos a considerar otra especificación del modelo. Así, en la Figura 14.1 lo indicamos con una flecha de retroalimentación en el proceso de construcción de modelos. A medida que adquiramos experiencia en la construcción de modelos y en la resolución de otros difíciles problemas, descubriremos que estos procesos tienden a repetirse y que se vuelve a fases anteriores hasta que se desarrolla un modelo satisfactorio y se soluciona el problema.

Interpretación del modelo e inferencia

Una vez que se ha construido un modelo, puede utilizarse para obtener alguna información sobre el sistema y el proceso estudiados. En el análisis de regresión, puede significar buscar intervalos de confianza para los parámetros del modelo, contrastar hipótesis de interés o predecir los futuros valores de la variable dependiente, dados los valores supuestos de las variables independientes. Es importante reconocer que este tipo de inferencia se basa en el supuesto de que el modelo está especificado y estimado correctamente. Cuanto más graves son los errores de especificación o de estimación, menos fiables son las inferencias realizadas a partir del modelo estimado.

También deberíamos reconocer que algunos resultados de nuestro análisis basado en los datos existentes pueden no estar de acuerdo con lo que se sabía hasta entonces. Cuando eso ocurre, es necesario comparar minuciosamente nuestros resultados con lo que se sabía hasta entonces. Las diferencias pueden deberse a que la especificación del modelo es diferente o incorrecta, a errores de los datos o alguna otra deficiencia. Pero también podríamos descubrir algunos importantes resultados nuevos debido a que la especificación del modelo es mejor o a nuevos datos que representan un cambio del contexto estudiado. En cualquier caso, debemos estar dispuestos a hacer correcciones o a presentar nuestros nuevos resultados de una manera lógica.

14.2. Variables ficticias y diseño experimental

En el apartado 13.8 introdujimos las **variables ficticias** en aplicaciones en las que había modelos de regresión aplicados a dos subconjuntos diferentes de datos. Por ejemplo, vimos cómo podrían utilizarse para averiguar la existencia de discriminación sexual en el ejemplo de los salarios.

En este apartado ampliamos las aplicaciones potenciales de las variables ficticias. En primer lugar, presentamos una aplicación en la que se aplica un modelo de regresión a más de dos subconjuntos de datos. A continuación, mostramos cómo pueden utilizarse las variables ficticias para estimar los efectos estacionales en un modelo de regresión aplicado a datos de series temporales. Por último, mostramos cómo pueden utilizarse las variables ficticias para analizar datos de situaciones experimentales, definidas por variables categóricas que contienen múltiples niveles.

EJEMPLO 14.1. Demanda de productos de lana (análisis del modelo utilizando variables ficticias)

Un analista de marketing para la Asociación de Fabricantes de Productos de Lana tiene interés en estimar la demanda de productos de lana en algunas ciudades en función de la renta total disponible de la ciudad. Se han recogido datos de 30 áreas metropolitanas seleccionadas aleatoriamente. En primer lugar, el analista especifica un modelo de regresión de la relación entre las ventas y la renta disponible:

$$Y = \beta_0 + \beta_1 X_1$$

donde X_1 es la renta disponible anual per cápita de una ciudad e Y son las ventas per cápita de productos de lana en la ciudad. Tras algunas conversaciones más, el analista

se pregunta si los niveles totales de ventas varían de unas regiones geográficas a otras: norte, centro y sur.

Solución

El análisis comienza colocando cada una de las ciudades en una de las tres regiones. La Figura 14.2 es un diagrama de puntos dispersos de las ventas per cápita en relación con la renta disponible. Los datos parecen estar divididos en tres subgrupos que corresponden a las regiones geográficas. Se utilizan dos variables ficticias para identificar cada una de las tres regiones siguientes:

Norte	$x_2 = 0,$	$x_3 = 1$
Centro	$x_2 = 1,$	$x_3 = 0$
Sur	$x_2 = 0,$	$x_3 = 0$

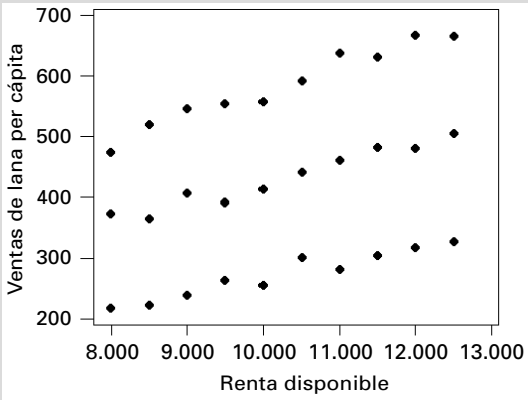


Figura 14.2. Ventas per cápita de lana en relación con la renta disponible per cápita.

En general, pueden identificarse perfectamente k regiones o subconjuntos con $k - 1$ variables ficticias. Si tratamos de utilizar k variables ficticias para representar k subgrupos distintos, obtenemos una relación lineal entre las variables de predicción y es imposible estimar los coeficientes, como se señaló en el apartado 13.2. Eso a veces se denomina «trampa de las variables ficticias».

Los desplazamientos de la constante del modelo podrían estimarse utilizando el modelo

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_1 X_1$$

Aplicando este modelo al norte, se convierte en

$$\begin{aligned} Y &= \beta_0 + \beta_2(0) + \beta_3(1) + \beta_1 X_1 \\ &= (\beta_0 + \beta_3) + \beta_1 X_1 \end{aligned}$$

En la región central, observamos que

$$\begin{aligned} Y &= \beta_0 + \beta_2(1) + \beta_3(0) + \beta_1 X_1 \\ &= (\beta_0 + \beta_2) + \beta_1 X_1 \end{aligned}$$

Por último, en el caso de la región meridional el modelo es

$$\begin{aligned} Y &= \beta_0 + \beta_2(0) + \beta_3(0) + \beta_1 X_1 \\ &= \beta_0 + \beta_1 X_1 \end{aligned}$$

Resumiendo estos resultados, las constantes de las distintas regiones son:

Norte	$\beta_0 + \beta_3$
Centro	$\beta_0 + \beta_2$
Sur	β_0

Esta formulación define el sur como la constante «base»; β_3 y β_2 definen el desplazamiento de la función de las ciudades del norte y el centro, respectivamente. Podrían utilizarse contrastes de hipótesis, utilizando el estadístico *t* de Student de los coeficientes, para averiguar si hay diferencias significativas entre las constantes de las diferentes regiones en comparación, en este caso, con la constante de la región del sur. Podrían obtenerse constantes para más regiones utilizando variables ficticias que continúen esta pauta. Podríamos especificar las variables ficticias de manera que cualquier nivel fuera el nivel base con el que se comparan los demás niveles. En este problema, la especificación del sur como condición base es natural, dados los objetivos del problema.

El modelo en el que se incluyen diferencias entre los coeficientes de la pendiente y las constantes es

$$\begin{aligned} Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + (\beta_1 + \beta_4 X_2 + \beta_3 X_3) X_1 \\ &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_1 X_1 + \beta_4 X_2 X_1 + \beta_5 X_3 X_1 \end{aligned}$$

Aplicando este modelo a la región del norte, vemos que

$$\begin{aligned} Y &= \beta_0 + \beta_2(0) + \beta_3(1) + (\beta_1 + \beta_4(0) + \beta_5(1)) X_1 \\ &= (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1 \end{aligned}$$

En el caso de la región central, el modelo es

$$\begin{aligned} Y &= \beta_0 + \beta_2(1) + \beta_3(0) + (\beta_1 + \beta_4(1) + \beta_5(0)) X_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 \end{aligned}$$

Por último, en el caso de la región del sur

$$\begin{aligned} Y &= \beta_0 + \beta_2(0) + \beta_3(0) + (\beta_1 + \beta_4(0) + \beta_5(0)) X_1 \\ &= \beta_0 + \beta_1 X_1 \end{aligned}$$

El coeficiente de la pendiente de X_1 de las ciudades de diferentes regiones es:

Norte	$\beta_1 + \beta_5$
Centro	$\beta_1 + \beta_4$
Sur	β_1

Una vez más, el sur es la condición base que tiene la pendiente β_1. Pueden utilizarse contrastes de hipótesis para averiguar la significación estadística de las diferencias entre los coeficientes de la pendiente y la condición base, que en este caso es la región del sur. Utilizando este modelo de regresión que contiene variables ficticias, el analista puede estimar la relación entre las ventas y la renta disponible por regiones.

Utilizando la muestra de 30 áreas metropolitanas divididas por igual entre las tres regiones geográficas, se estimó un modelo de regresión múltiple con variables ficticias utilizando Minitab. Los resultados se muestran en la Figura 14.3. A partir del modelo de regresión podemos averiguar las características de las pautas de compra de lana. Pueden utilizarse contrastes de hipótesis condicionados de la forma

$$H_0: \beta_j = 0 \mid \beta_l \neq 0,\ l = 1, ..., K,\ l \neq j$$
$$H_1: \beta_j \neq 0 \mid \beta_l \neq 0,\ l = 1, ..., K,\ l \neq j$$

para averiguar los efectos condicionados de los distintos factores en la demanda de lana. El coeficiente de la variable ficticia X_3, $\beta_3 = 138{,}46$, indica que las personas del norte gastan una media de 138,46 $ más que las del sur. Asimismo, las personas de la región central gastan una media de 96,33 $ más que las del sur. Estos coeficientes son significativos. El coeficiente de la renta disponible es 0,0252, lo que indica que, en el caso de las personas del sur, cada dólar de aumento de la renta per cápita incrementa la compra de productos de lana en 0,025, y este resultado es significativo. En el caso de las personas del norte, cada dólar de aumento de la renta incrementa el gasto en productos de lana en 0,042 (0,0252 + 0,0168) y la diferencia entre los aumentos de la pendiente es significativa. La tasa estimada de aumento de la compra por dólar de aumento de la renta también es mayor en el caso de las personas que viven en la región central que en el de las que viven en la región del sur. Sin embargo, esa diferencia no es significativa. Utilizando estos resultados, las ventas por región pueden predecirse con mayor precisión que con un modelo que combine todas las regiones y sólo utilice la renta per cápita.

```
The regression equation is
Per Capita Wool Sales = 12.7 + 138 North X3 + 96.3 Central X2
           + 0.0252 Disposable Income + 0.0168 NorX3Inc + 0.00608 CentX2Inc


Predictor              Coef        StDev       T       P
Constant              12.73        27.74    0.53   0.600
North X3             138.46        39.22    3.53   0.022
Central X2            96.33        39.22    2.46   0.002
Disposab          0.025231     0.002680    9.42   0.000
NorX3 Inc         0.016839     0.003790    4.44   0.000
CentX2 In         0.006085     0.003790    1.61   0.121

S = 12.17      R-Sq = 99.4%   R-Sq(adj) = 99.2%

Analysis of Variance

Source            DF        SS        MS       F       P
Regression         5    553704    110741  747.71   0.000
Residual Error    24      3555       148
Total             29    557259
```

Figura 14.3. Modelo de regresión múltiple utilizando variables ficticias par estimar el consumo de lana per cápita (salida Minitab).

EJEMPLO 14.2. Predicción de las ventas de productos de lana (variables ficticias estacionales)

Tras acabar el análisis de las ventas regionales, el analista decidió estudiar la relación entre las ventas y la renta disponible utilizando datos de series temporales. Tras realizar algunos análisis, se dio cuenta de que las ventas varían de unos trimestres a otros. Por ejemplo, durante el cuarto trimestre son altas en previsión de los regalos de Navidad y de la bajada de la temperatura. Le ha pedido que lo ayude a realizar el estudio.

Solución

Tras analizar el problema, le recomienda que represente los cuatro trimestres de cada año por medio de tres variables ficticias. De esta forma, puede utilizarse el modelo de regresión múltiple para estimar las diferencias entre las ventas de los diferentes trimestres. Concretamente, le propone una estructura similar a la del modelo de variables ficticias regionales:

Primer trimestre:	$x_2 = 0, x_3 = 0, x_4 = 0$
Segundo trimestre:	$x_2 = 1, x_3 = 0, x_4 = 0$
Tercer trimestre:	$x_2 = 0, x_3 = 1, x_4 = 0$
Cuarto trimestre:	$x_2 = 0, x_3 = 0, x_4 = 1$

Los coeficientes de las variables ficticias son estimaciones de los desplazamientos de la función de consumo de lana entre los trimestres en el modelo de los datos

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_1 X_1$$

donde Y son las ventas totales de productos de lana y X_1 es la renta disponible. Las constantes de los distintos trimestres son:

Primer trimestre:	β_0
Segundo trimestre:	$\beta_0 + \beta_2$
Tercer trimestre:	$\beta_0 + \beta_3$
Cuarto trimestre:	$\beta_0 + \beta_4$

Modelos de diseño experimental

Los métodos de diseño experimental han sido una importante área de investigación y práctica estadísticas durante algunos años. Los primeros estudios se referían a investigaciones agrícolas. Los esfuerzos realizados por estadísticos como R. A. Fisher y O. L. Davies en Inglaterra durante la décaca de 1920 sentaron las bases de la metodología del diseño experimental y de la práctica estadística en general. Los experimentos agrícolas requieren una temporada entera de cultivo para obtener datos. Era, pues, importante desarrollar métodos que pudieran dar respuesta a una serie de cuestiones y conseguir una gran precisión. Además, la mayoría de los experimentos definían la actividad utilizando variables con niveles discretos en lugar de continuos. Los métodos de diseño experimental también se han utilizado mucho para estudiar la conducta humana y para realizar algunos experimentos industriales. El énfasis reciente en la mejora de la calidad y la productividad ha aumentado la actividad en esta área de la estadística con importantes aportaciones de grupos como el Center for Quality and Productivity de la Universidad de Wisconsin.

Diseño experimental

La regresión utilizando variables ficticias puede emplearse como instrumento en los estudios de diseño experimental. Los experimentos tienen una única variable de resultado, que contiene todo el error aleatorio. Cada resultado experimental corresponde a una combinación discreta de las variables experimentales (independientes), X_j.

Existe una importante diferencia de filosofía entre los diseños experimentales y la mayoría de los problemas que hemos examinado. El diseño experimental intenta identificar las causas de las variaciones de la variable dependiente, especificando previamente combinaciones de variables independientes discretas cuyos valores se utilizan para medir la variable dependiente. Un importante objetivo es elegir puntos experimentales, definidos por variables independientes, que constituyan estimadores de las varianzas mínimas. El orden en el que se realizan los experimentos se elige aleatoriamente para evitar sesgos introducidos por variables no incluidas en el experimento.

Los resultados experimentales, Y, corresponden a combinaciones específicas de niveles de las variables de tratamiento y de bloqueo. Una *variable de tratamiento* es una variable cuyo efecto tenemos interés en estimar con una varianza mínima. Por ejemplo, podríamos querer saber cuál de cuatro máquinas de producción es más productiva por hora. En ese caso, el tratamiento son las máquinas de producción representadas por una variable categórica de cuatro niveles, Z_j. Una *variable de bloqueo* representa una variable que forma parte del entorno y, por lo tanto, no puede preseleccionarse el nivel de la variable. Pero queremos incluir el nivel de la variable de bloqueo en nuestro modelo, con el fin de eliminar la variabilidad de la variable de resultado, Y, que está relacionada con los diferentes niveles de las variables de bloqueo. Podemos representar una variable de tratamiento o de bloqueo de K niveles utilizando $K - 1$ variables ficticias. Consideremos un sencillo ejemplo que tiene una variable de tratamiento de cuatro niveles, Z_1, y una variable de bloqueo de tres niveles, Z_2. Estas variables podrían representarse por medio de variables ficticias, como se muestra en la Tabla 14.1. A continuación, utilizando estas variables ficticias, podría estimarse el modelo de diseño experimental mediante el modelo de regresión múltiple

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

Tabla 14.1. Ejemplo de especificación de las variables ficticias para las variables de tratamiento y de bloqueo

Z_1	X_1	X_2	X_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1
Z_2	X_4	X_5	
1	0	0	
2	1	0	
3	0	1	

En este modelo, por ejemplo, el coeficiente β_3 es una estimación de la cantidad en la que la productividad del nivel de tratamiento 4 es mayor que la del nivel de tratamiento 1, para la variable de tratamiento categórica, Z_1. Naturalmente, si β_3 es negativo, sabemos

que el nivel de tratamiento 1 tiene una productividad mayor que el 4. Siguiendo la lógica de la regresión múltiple, sabemos que las variables X_4 y X_5 explican parte de la variabilidad de Y y, por lo tanto, el estimador de la varianza es menor. Este modelo puede expandirse fácilmente para incluir varias variables de tratamiento simultáneamente con algunas otras variables de bloqueo. Además, si hay una variable continua —por ejemplo, la temperatura ambiente— que afecta a la productividad, esa variable también puede añadirse directamente al modelo de regresión. En muchos casos, se replica varias veces el diseño básico para obtener suficientes grados de libertad para el error. Este proceso se muestra en el ejemplo 14.3.

EJEMPLO 14.3. Programa de formación de los trabajadores (especificación del modelo utilizando variables ficticias)

María Cruz es la directora de producción de una gran fábrica de piezas de automóvil. Tiene interés en saber cómo afecta un nuevo programa de formación a la productividad de los trabajadores. Existen muchas investigaciones que apoyan la conclusión de que en la productividad influyen el tipo de máquina y la cantidad de formación que ha recibido el trabajador.

Solución

María define las siguientes variables para el experimento:

Y El número de unidades producidas por turno de 8 horas

Z_1 El tipo de formación

1. Clase tradicional en un aula y presentación de películas
2. Enseñanza interactiva asistida por computador (CAI)

Z_2 Tipo de máquina

1. Máquina de tipo 1
2. Máquina de tipo 2
3. Máquina de tipo 3

Z_3 Nivel de estudios de los trabajadores

1. Nivel de estudios secundarios
2. Al menos un año de estudios postsecundarios

La variable Z_1 se llama *variable de tratamiento* porque el principal objetivo del estudio es evaluar el programa de formación. Las variables Z_2 y Z_3 se llaman *variables de bloqueo* porque se incluyen para ayudar a reducir o bloquear parte de la variabilidad sin explicar. De esta forma se reduce la varianza y el contraste de los principales efectos del tratamiento tiene mayor potencia. La expresión *variable de bloqueo* proviene de los experimentos agrícolas en los que las parcelas se dividían en pequeños bloques, cuyo suelo tenía unas condiciones que variaban de unos a otros. También es posible estimar el efecto de estas variables de bloqueo. Por lo tanto, no se pierde información llamando a ciertas variables «variables de bloqueo» en lugar de «variables de tratamiento».

Las observaciones del diseño experimental se definen previamente utilizando las variables independientes. La Tabla 14.2 contiene una lista de las observaciones, en la que cada observación se designa utilizando los niveles de las variables Z. En este diseño, que se llama diseño factorial completo, hay 12 observaciones, una para cada combina-

Tabla 14.2. Diseño experimental para el estudio de la productividad.

Producción Y	Formación Z_1	Máquina Z_2	Nivel de estudios Z_3
Y_1	1	1	1
Y_2	1	1	2
Y_3	1	2	1
Y_4	1	2	2
Y_5	1	3	1
Y_6	1	3	2
Y_7	2	1	1
Y_8	2	1	2
Y_9	2	2	1
Y_{10}	2	2	2
Y_{11}	2	3	1
Y_{12}	2	3	2

ción de las variables de tratamiento y de bloqueo. Las Y_i observaciones representan las respuestas medidas en cada una de las condiciones experimentales. En los datos, el modelo Y_i contiene el efecto de las variables de tratamiento y de bloqueo más un error aleatorio. En muchos diseños experimentales, esta pauta de 12 observaciones se replica (se repite) para obtener más grados de libertad para el error y estimaciones más bajas de las varianzas de los efectos de las variables de diseño. Este diseño también puede analizarse utilizando los métodos del análisis de la varianza. Sin embargo, aquí mostramos cómo puede realizarse el análisis recurriendo a la regresión basada en variables ficticias.

Los niveles de cada una de las tres variables de diseño —Z_1, Z_2 y Z_3— pueden expresarse como un conjunto de variables ficticias. Definamos las siguientes variables ficticias:

$$z_1 = 1 \rightarrow x_1 = 0$$
$$z_1 = 2 \rightarrow x_1 = 1$$
$$z_2 = 1 \rightarrow x_2 = 0 \ \& \ x_3 = 0$$
$$z_2 = 2 \rightarrow x_2 = 1 \ \& \ x_3 = 0$$
$$z_2 = 3 \rightarrow x_2 = 0 \ \& \ x_3 = 1$$
$$z_3 = 1 \rightarrow x_4 = 0$$
$$z_3 = 2 \rightarrow x_4 = 1$$

Utilizando estas relaciones, el modelo de diseño experimental de la Tabla 14.2, que utiliza las variables Z, puede representarse por medio de variables ficticias, como muestra la Tabla 14.3. Utilizando estas variables ficticias, podemos definir un modelo de regresión múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Los coeficientes de regresión se estiman utilizando las variables especificadas previamente. Los 12 experimentos u observaciones definidos en las Tablas 14.2 y 14.3 son una réplica del diseño experimental. Una réplica contiene todos los experimentos individuales que se incluyen en el diseño experimental. A menudo se realizan varias réplicas del diseño para estimar con mayor precisión los coeficientes y obtener suficientes grados de libertad para estimar la varianza. En el modelo basado en variables ficticias, esti-

Tabla 14.3. Diseño experimental para el estudio de la productividad utilizando variables ficticias.

Productividad Y	X_1	X_2	X_3	X_4
Y_1	0	0	0	0
Y_2	0	0	0	1
Y_3	0	1	0	0
Y_4	0	1	0	1
Y_5	0	0	1	0
Y_6	0	0	1	1
Y_7	1	0	0	0
Y_8	1	0	0	1
Y_9	1	1	0	0
Y_{10}	1	1	0	1
Y_{11}	1	0	1	0
Y_{12}	1	0	1	1

mamos cuatro coeficientes y una constante y quedan $n - 4 - 1$ grados de libertad para estimar la varianza. Con una réplica, $n = 12$ y tenemos 7 grados de libertad para estimar la varianza. Con dos réplicas del diseño, $n = 24$ y tenemos 19 grados de libertad para estimar la varianza, y con tres réplicas tenemos 31 grados de libertad. Normalmente, se necesitan al menos 15 o 20 grados de libertad para obtener estimaciones estables de la varianza. Utilizando las definiciones de las variables ficticias, observamos que los coeficientes de regresión estimados se interpretan de la forma siguiente:

1. b_1 es el aumento de la productividad provocado por el nuevo tipo de formación CAI en comparación con la formación tradicional en el aula.
2. b_2 es el aumento de la productividad provocado por la máquina de tipo 2 en comparación con la de tipo 1.
3. b_3 es el aumento de la productividad provocado por la máquina de tipo 3 en comparación con la de tipo 1.
4. b_4 es el aumento de la productividad provocado por la educación postsecundaria en comparación con la secundaria solamente.

Cualquiera de estos «aumentos» podría ser negativo, lo que implica una disminución.

La importancia de cada uno de estos efectos puede contrastarse utilizando nuestros métodos tradicionales de contraste de hipótesis. Obsérvese que si se pierde o falla una observación experimental, puede seguir utilizándose el mismo modelo de regresión para estimar los coeficientes. Sin embargo, en ese caso tenemos una varianza mayor y, por lo tanto, los contrastes de hipótesis tienen menos potencia.

También es posible añadir al modelo variables continuas u otras variables relacionadas. Supongamos que María sospecha que el número de años de experiencia de los trabajadores y la temperatura ambiente también influyen en la productividad. Se pueden medir estas dos variables continuas para cada experimento y añadir al modelo de regresión basado en variables ficticias. El modelo de regresión se convierte entonces en

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

donde X_5 son los años de experiencia y X_6 es la temperatura ambiente. Si estas últimas son importantes, reducirán la varianza y aumentarán la potencia de los contrastes de hipótesis de los efectos de otras variables.

Otra extensión posible es la inclusión de efectos de interacción. Supongamos que María sospecha que la formación CAI es más beneficiosa para los trabajadores que utilizan la máquina de tipo 3. Para contrastar este efecto, puede incluir una variable de interacción, $X_7 = X_1X_3$. Los valores de X_7 son el producto de las variables X_1 y X_3. Por lo tanto, en la Tabla 14.3 añadiríamos una columna para X_7, que tomaría el valor 1 en el caso de la 11.ª observación y la 12.ª y 0 en el del resto. Si también sospecha que la formación CAI beneficia más a los trabajadores que tienen un nivel de estudios más alto, puede definir otra variable de interacción, $X_8 = X_1X_4$. Esta variable añade otra columna a la Tabla 14.3, que tomaría el valor 1 en el caso de la 8.ª observación, la 10.ª y la 12.ª y 0 en el del resto. Es posible añadir otras variables y términos de interacción. Por lo tanto, el número de opciones con estos diseños experimentales es muy grande.

Con todas estas adiciones, el modelo de regresión es

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8$$

En esta ecuación, hay que estimar ocho coeficientes y una constante y sólo quedan 3 grados de libertad para estimar la varianza si sólo se realiza una réplica del diseño. En las situaciones en las que las mediciones pueden realizarse con precisión y los distintos efectos son grandes, este diseño incluso con una réplica puede suministrar útil información sobre los factores que influyen en la productividad. En la mayoría de los casos, es deseable hacer más de una réplica. Con un número mayor de observaciones, las estimaciones de los coeficientes son mejores y la varianza de los coeficientes es menor. Sin embargo, en una situación industrial es posible que haya que realizar experimentos en toda la fábrica, por lo que pueden ser muy caros. Los analistas tratan de conseguir la máxima información posible en cada conjunto de experimentos.

En este apartado hemos introducido los diseños experimentales y su análisis utilizando variables ficticias. El diseño experimental es una importante área de la estadística aplicada que puede estudiarse en otros muchos cursos y libros. Los programas estadísticos, como el Minitab, normalmente contienen un extenso conjunto de rutinas para desarrollar distintos y sofisticados modelos de diseño experimental. Deben utilizarse únicamente después de conocer sus detalles e interpretaciones específicos. Sin embargo, incluso con la introducción que hemos realizado aquí, el lector tiene un poderoso instrumento para abordar algunos importantes problemas de productividad.

Las aplicaciones del diseño experimental han cobrado una creciente importancia en las operaciones manufactureras y otras operaciones empresariales. Los experimentos para identificar las variables relacionadas con el aumento de la producción y la reducción de los defectos son importantes para mejorar las operaciones de producción. El uso de variables ficticias y de la regresión múltiple para el análisis del diseño experimental amplía los tipos de problemas que pueden abordarse sin aprender más técnicas de análisis. Ésta es una importante ventaja más de los métodos basados en variables ficticias.

EJERCICIOS

Ejercicios básicos

14.1. Formule la especificación de un modelo y defina las variables de un modelo de regresión múltiple para predecir la calificación media obtenida en la universidad en función de la nota media obtenida en el bachillerato y del año de estudios universitarios: primer año, segundo año, tercer año, cuarto año.

14.2. Formule la especificación del modelo y defina las variables de un modelo de regresión múltiple para predecir los salarios en dólares estadounidenses en función de los años de experiencia y del país de empleo (Alemania, Gran Bretaña, Japón, Estados Unidos y Turquía).

14.3. Formule la especificación del modelo y defina las variables de un modelo de regresión múltiple para predecir el coste por unidad producida en función del tipo de fábrica (tecnología clásica, máquinas controladas por computador y manipulación del material controlada por computador) y en función del país (Colombia, Sudáfrica y Japón).

14.4. Un economista quiere estimar una ecuación de regresión que relacione la demanda de un producto (Y) con su precio (X_1) y la renta (X_2). Tiene que basarse en 12 años de datos trimestrales. Sin embargo, se sabe que la demanda de este producto es estacional, es decir, es mayor en unos momentos del año que en otros.

a) Una posibilidad para tener en cuenta la estacionalidad es estimar el modelo

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \beta_5 x_{5t} + \beta_6 x_{6t} + \varepsilon_t$$

donde x_{3t}, x_{4t}, x_{5t} y x_{6t} son valores de las variables ficticias, siendo

$x_{3t} = 1$ en el primer trimestre de cada año, 0 en el resto

$x_{4t} = 1$ en el segundo trimestre de cada año, 0 en el resto

$x_{5t} = 1$ en el tercer trimestre de cada año, 0 en el resto

$x_{6t} = 1$ en el cuarto trimestre, 0 en el resto

Explique por qué este modelo no puede estimarse por mínimos cuadrados.

b) Un modelo que puede estimarse es

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \beta_4 x_{4t} + \beta_5 x_{5t} + \varepsilon_t$$

Interprete los coeficientes de las variables ficticias de este modelo.

Ejercicios aplicados

14.5. Sharon Parsons, presidente de Gourmet Box Mini Pizza, le ha pedido ayuda para desarrollar un modelo que prediga la demanda de la nueva pizza llamada Pizza1. Este producto compite en el mercado con otras tres marcas que llamaremos B2, B3 y B4. Actualmente, los productos son vendidos por tres grandes cadenas de distribución llamadas 1, 2 y 3 para identificarlas. Estas tres cadenas tienen diferentes cuotas de mercado y, por lo tanto, es probable que las ventas de cada distribuidor sean diferentes. El fichero de datos **Market** contiene datos semanales recogidos en las 52 últimas semanas en las tres cadenas de distribución. A continuación, se definen las variables del fichero de datos.

Utilice la regresión múltiple para desarrollar un modelo que prediga la cantidad de Pizza1 vendida a la semana por cada distribuidor. El modelo sólo debe contener variables de predicción importantes.

Distribuidor	Identificador numérico del distribuidor
Weeknum	Número secuencial de la semana en la que se recogieron los datos
Sales Pizza1	Número de unidades de Pizza1 vendidas por el distribuidor durante la semana
Price Pizza1	Precio al por menor de Pizza1 cobrado por el distribuidor durante esa semana
Promotion	Nivel de promoción de la semana: 0 significa Ninguna promoción; 1 significa Anuncios en televisión; 2 significa Exposición en las tiendas; 3 significa Anuncios en la televisión y Exposición en las tiendas
Sales B2	Número de unidades de la marca 2 vendidas por el distribuidor durante la semana
Price B2	Precio al por menor de la marca 2 cobrado por el distribuidor durante la semana
Sales B3	Número de unidades de la marca 3 vendidas por el distribuidor durante la semana
Price B3	Precio al por menor de la marca 3 cobrado por el distribuidor durante la semana
Sales B4	Número de unidades de la marca 4 vendidas por el distribuidor durante la semana
Price B4	Precio al por menor de la marca 4 cobrado por el distribuidor durante la semana

14.6. Le han pedido que desarrolle un modelo de regresión múltiple para predecir las ventas per cápita de cereales de desayuno en las ciudades de más de 100.000 habitantes. En primer lugar, celebra una reunión con los principales directivos de marketing que tienen experiencia en la venta de cereales. En esta reunión, descubre que se es-

pera que en las ventas per cápita influyan el precio de los cereales, el precio de los cereales rivales, la renta media per cápita, el porcentaje de titulados universitarios, la temperatura anual media y la pluviosidad anual media. También se entera de que la relación lineal entre el precio y las ventas per cápita se espera que tenga una pendiente diferente en las ciudades que se encuentran al este del río Misisipi. Se espera que las ventas per cápita sean mayores en las ciudades que tienen una renta per cápita alta y baja que en las ciudades que tienen una renta per cápita intermedia. También se espera que las ventas per cápita sean diferentes en los cuatro sectores siguientes del país: noroeste, sudoeste, noreste y sudeste.

Formule una especificación del modelo cuyos coeficientes puedan estimarse por medio de la regresión múltiple. Defina cada variable completamente e indique la forma matemática del modelo. Analice su especificación, indique qué variables espera que sean estadísticamente significativas y explique las razones por las que lo espera.

14.7. Máximo Márquez, presidente de Piezas Buenas, S.A., le ha pedido que desarrolle un modelo que prediga el número de piezas defectuosas por turno de 8 horas de su fábrica. Cree que existen diferencias entre los tres turnos diarios y entre los cuatro proveedores de materias primas. Además, se piensa que cuanto mayor es la producción y mayor el número de trabajadores, mayor es el número de piezas defectuosas. Máximo visita la fábrica varias veces en los tres turnos para observar las operaciones y dar consejos. Le ha facilitado una lista de los turnos que ha visitado y quiere saber si el número de piezas defectuosas aumenta o disminuye cuando visita la fábrica.

Describa por escrito cómo desarrollaría un modelo para estimar y contrastar los distintos factores que pueden influir en el número de piezas defectuosas producidas por turno. Defina detenidamente cada coeficiente de su modelo y el contraste que utilizaría. Indique cómo recogería los datos y cómo definiría cada variable utilizada en el modelo. Analice las interpretaciones que haría a partir de su especificación del modelo.

14.8. Maderas de Calidad, S.A., lleva 40 años en el sector. Hace muebles de madera de encargo de alta calidad e interiores de armarios y trabajos de madera de interiores de muy buena calidad para viviendas y oficinas caras. La empresa ha tenido mucho éxito debido en gran parte a la elevada cualificación de los artesanos que diseñan y producen sus productos en consulta con sus clientes. Muchos de sus productos han recibido premios nacionales por la calidad de su diseño y el trabajo bien hecho. Cada producto hecho de encargo es producido por un equipo de dos artesanos o más que primero se reúnen con el cliente, realizan un primer diseño, lo revisan con el cliente y después fabrican el producto. Los clientes también pueden reunirse con los artesanos varias veces durante la producción.

Los artesanos tienen una buena formación y han adquirido excelentes cualificaciones en el trabajo de la madera. La mayoría tienen título universitario y se han formado con artesanos cualificados. Los empleados se clasifican en tres niveles: 1. Aprendiz, 2. Profesional y 3. Maestro. Los salarios de los niveles 2 y 3 son más altos y los trabajadores normalmente ascienden conforme adquieren experiencia y cualificación. Actualmente, la empresa tiene una plantilla diversa, en la que hay trabajadores blancos, negros y latinos y tanto hombres como mujeres. Cuando comenzó hace 40 años, todos los trabajadores eran blancos. Hace unos 20 años, comenzó a contratar artesanos negros y latinos, y hace unos 10 años contrató artesanas. Los trabajadores blancos varones tienden a estar sobrerrepresentados en las clasificaciones de los puestos de trabajo más altas debido en parte a que tienen más experiencia. Actualmente, la plantilla tiene un 40 por ciento de hombres blancos, un 30 por ciento de hombres negros y latinos, un 15 por ciento de mujeres blancas y un 15 por ciento de mujeres negras y latinas.

Recientemente, algunos han expresado su preocupación por la discriminación salarial. Concretamente, dicen que las mujeres y los que no son blancos no están recibiendo una remuneración acorde con su experiencia. La dirección de la empresa sostiene que todas las personas cobran en función de los años de experiencia, del nivel de clasificación del puesto de trabajo y de la capacidad personal. Sostiene que no existen diferencias salariales basadas en la raza o el sexo por lo que se refiere al salario base o al incremento por cada año de experiencia.

Explique cómo realizaría un análisis para averiguar si la afirmación de la dirección es cierta. Muestre los detalles de su análisis y razónelos claramente. Indique los datos que deben recogerse y los nombres y las descripciones de las variables que utilizará en el análisis. Indique claramente los contrastes estadísticos que utilizaría

para averiguar cuál es la verdadera situación e indique las reglas de decisión basadas en los contrastes de hipótesis y los resultados de los datos.

14.9. Le han pedido que haga de consultor y de testigo experto en un juicio por discriminación salarial. Un grupo de mujeres latinas y negras ha demandado a su empresa, Distribuidores Reunidos, S.A. Las mujeres, que tienen entre 5 y 25 años de antigüedad en la empresa, alegan que su subida salarial anual media ha sido significativamente menor que la de un grupo de hombres blancos y un grupo de mujeres blancas. Los puestos de trabajo de los tres grupos contienen diversos componentes administrativos, analíticos y directivos. Todos los empleados tenían titulación universitaria de primer ciclo cuando empezaron a trabajar y los años de experiencia son un importante factor para predecir el rendimiento y la productividad de los trabajadores. Le han facilitado el salario mensual actual y el número de años de experiencia de todos los trabajadores de los tres grupos. Además, los datos indican los miembros de los tres grupos que tienen un máster en administración de empresas. Observe que en este problema no realiza ningún análisis de los datos.

a) Desarrolle un modelo y un análisis estadísticos que permitan analizar los datos. Indique los contrastes de hipótesis que pueden utilizarse para aportar pruebas contundentes de la existencia de discriminación salarial si es que existe. La compañía también ha contratado a un estadístico como consultor y testigo experto. Describa su análisis de una forma exhaustiva y clara.

b) Suponga que sus contrastes de hipótesis aportan pruebas contundentes que apoyan la tesis de sus clientes. Resuma brevemente las observaciones clave que hará en su comparecencia en el juicio. Es de esperar que el abogado de la empresa le contrainterrogue con la ayuda de su estadístico, que enseña estadística en una prestigiosa universidad.

14.3. Valores retardados de las variables dependientes como regresores

En este apartado examinamos las variables dependientes retardadas, un importante tema cuando se analizan datos de series temporales, es decir, cuando se realizan mediciones de las cantidades a lo largo del tiempo. Por ejemplo, podemos tener observaciones mensuales, observaciones trimestrales u observaciones anuales. Los economistas normalmente utilizan variables de series temporales como los tipos de interés, medidas de la inflación, la inversión agregada y el consumo agregado para realizar análisis y desarrollar modelos. Especificamos las observaciones de series temporales utilizando el subíndice t para indicar el tiempo en lugar de la i que empleamos para indicar los datos de corte transversal. Por lo tanto, un modelo de regresión múltiple sería

$$y_t = \beta_0 + \beta_t x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t$$

En muchas aplicaciones de series temporales, la variable dependiente en el periodo t a menudo también está relacionada con el valor que tomó esta variable en el periodo anterior, es decir, con y_{t-1}. El valor de la variable dependiente en un periodo anterior se llama *variable dependiente retardada.*

Regresiones que contienen variables dependientes retardadas

Consideremos el siguiente modelo de regresión que relaciona una variable dependiente, Y, con K variables independientes:

$$y_t = \beta_0 + \beta_t x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \gamma y_{t-1} + \varepsilon_t \quad \textbf{(14.1)}$$

donde β_0, β_1, ..., β_K, γ son coeficientes fijos. Si se generan datos con este modelo:

a) Un aumento de la variable independiente X_j de 1 unidad en el periodo t, manteniéndose fijas todas las demás variables independientes, provoca un aumento esperado de la variable dependiente de β_j en el periodo t, $\beta_j\gamma$ en el periodo $(t+1)$, $\beta_j\gamma^2$ en el periodo $(t+2)$, $\beta_j\gamma^3$ en el periodo $(t+3)$, etc. El aumento total esperado en todos los periodos actuales y futuros es

$$\frac{\beta_j}{(1-\gamma)}$$

b) Los coeficientes β_0, β_1, ..., β_K, γ pueden estimarse por mínimos cuadrados como siempre.

c) Pueden calcularse intervalos de confianza y contrastes de hipótesis para los coeficientes de regresión exactamente igual que en el modelo de regresión múltiple ordinario (en rigor, cuando la ecuación de regresión contiene variables dependientes retardadas, estos métodos sólo son aproximadamente válidos. La calidad de la aproximación mejora, manteniéndose todo lo demás constante, cuando aumenta el número de observaciones muestrales).

d) Cuando se utilizan intervalos de confianza y contrastes de hipótesis con datos de series temporales, hay que tener cautela. Existe la posibilidad de que los errores de las ecuaciones, ε_i, ya no sean independientes entre sí. En el apartado 14.7 sobre las autocorrelaciones examinamos esta cuestión. En particular, cuando los errores están correlacionados, las estimaciones de los coeficientes son insesgadas, pero no eficientes. Por lo tanto, los intervalos de confianza y los contrastes de hipótesis ya no son válidos. Los económetras han desarrollado métodos para hacer estimaciones en estas condiciones, que se introducen en el apartado 14.7.

Para ilustrar el cálculo de las estimaciones y de la inferencia basada en la ecuación de regresión ajustada cuando el modelo contiene variables dependientes retardadas, examinamos el extenso ejemplo 14.4 (véase la referencia bibliográfica 1).

EJEMPLO 14.4. Los gastos publicitarios en función de las ventas al por menor (modelo de regresión con variables retardadas)

Un investigador tenía interés en predecir los gastos publicitarios en función de las ventas al por menor, sabiendo que la publicidad del año anterior también había influido.

Solución

Se creía que la publicidad local por hogar dependía de las ventas al por menor por hogar. Además, como los publicistas pueden no querer o no poder ajustar sus planes a los cambios repentinos del nivel de ventas al por menor, se añadió al modelo el valor de los gastos publicitarios locales por hogar del año anterior. Por lo tanto, los gastos publicitarios de este año están relacionados con las ventas al por menor (x_t) de este año y con los gastos publicitarios (y_{t-1}) del año anterior. El modelo que hay que ajustar es, pues,

$$y_t = \beta_0 + \beta_1 x_{1t} + \gamma y_{t-1} + \varepsilon_t$$

donde

y_t = publicidad local por hogar en el año t
x_t = ventas al por menor por hogar en el año t

Advertising Retail

Los datos sobre la publicidad y las ventas al por menor se encuentran en un fichero de datos Minitab llamado **Advertising Retail**. El valor retardado y_{t-1} puede generarse en Minitab utilizando la función retardo (*lag*) en las rutinas de la calculadora y en todos los demás buenos paquetes estadísticos utilizando procedimientos similares. Después de realizar la transformación del retardo, el fichero de datos incluye la variable retardada. La observación 1 de la variable retardada es inexistente, por lo que el conjunto de datos sólo tiene 21 observaciones. Siempre será así cuando se creen variables retardadas. Naturalmente, podríamos tener acceso a datos del año anterior —del año 0 en este ejemplo— y ese valor podría sustituir al valor que faltaba. Ahora ya están listos los datos para realizar una regresión múltiple utilizando los comandos convencionales de Minitab. La Figura 14.4 muestra la salida del análisis de regresión resultante.

```
The regression equation is
Advertising Y(t) = -43.8 + 0.0188 Retail Sales X(t) + 0.479 lag advertising

21 cases used 1 cases contain missing values

Predictor        Coef      SE Coef        T        P
Constant      -43.766        9.843    -4.45    0.000
Retail S     0.018777     0.002855     6.58    0.000
lag adve      0.47906      0.08732     5.49    0.000

S = 3.451      R-Sq = 96.3%   R-Sq(adj) = 95.9%

Analysis of Variance

Source              DF          SS       MS         F         P
Regression           2      5559.1   2779.5    233.43     0.000
Residual Error      18       214.3     11.9
Total               20      5773.4

Source         DF      Seq SS
Retail S        1      5200.7
lag adve        1       358.4

Unusual observations
obs    Retail S   Advertis         Fit        SE Fit   Residual   St Resid
  4        5507    119.220     112.716         1.222      6.504      2.02R
 20        6394    145.370     151.853         1.774     -6.483     -2.19R

R denotes an observation with a large standardized residual
```

Figura 14.4. Gastos publicitarios en función de las ventas al por menor y de los gastos publicitarios retardados (salida Minitab).

La regresión resultante de este problema (con la ausencia de la primera observación) es

$$\hat{y}_t = \underset{}{-43{,}8} + \underset{(0{,}0029)}{0{,}0188}x_t + \underset{(0{,}087)}{0{,}479}y_{t-1}$$

Los números que figuran debajo de los coeficientes de regresión son las desviaciones típicas de los coeficientes. El estadístico t de Student de cada coeficiente es bastante alto y los p-valores resultantes son 0,00, lo que indica que podemos rechazar la hipótesis nula de que los coeficientes son 0. Con 18 grados de libertad para el error, el valor crítico del estadístico t de Student de una hipótesis de dos colas suponiendo que $\alpha = 0{,}05$ es $t = 2{,}101$.

INTERPRETACIÓN

En los modelos de series temporales, el coeficiente de determinación R^2 puede ser algo engañoso. Por ejemplo, el elevado valor de $R^2 = 96{,}3$ por ciento del presente problema no indica necesariamente que exista una estrecha relación entre la publicidad local y las ventas al por menor. Es un hecho empírico perfectamente conocido que los gráficos de muchas series temporales empresariales y económicas muestran una pauta evolutiva bastante uniforme a lo largo del tiempo. Este mero hecho es suficiente para que el coeficiente de determinación tenga un valor alto cuando se incluye una variable dependiente retardada en el modelo de regresión. A efectos prácticos, aconsejamos al lector que preste relativamente poca atención al valor de R^2 en esos modelos.

La regresión estimada para este problema puede interpretarse de la siguiente manera. Supongamos que las ventas al por menor por hogar aumentan 1 $ este año. El efecto esperado en la publicidad local por hogar es un aumento de 0,0188 este año, otro aumento de

$$(0{,}479)(0{,}0188) = 0{,}0090 \ \$$$

el próximo año, otro aumento de

$$(0{,}479)^2\,(0{,}0188) = 0{,}0043 \ \$$$

dentro de dos años, y así sucesivamente. El efecto total en los futuros gastos publicitarios totales por hogar es un aumento esperado de

$$\frac{0{,}0188}{1 - 0{,}479} = 0{,}0361 \ \$$$

Vemos, pues, que el efecto esperado de un aumento de las ventas es un aumento inmediato de los gastos publicitarios, un aumento menor durante el próximo año, un aumento aún menor dentro de dos años, etc. La Figura 14.5 ilustra este efecto geométricamente decreciente de un aumento de las ventas este año en la publicidad de futuros años.

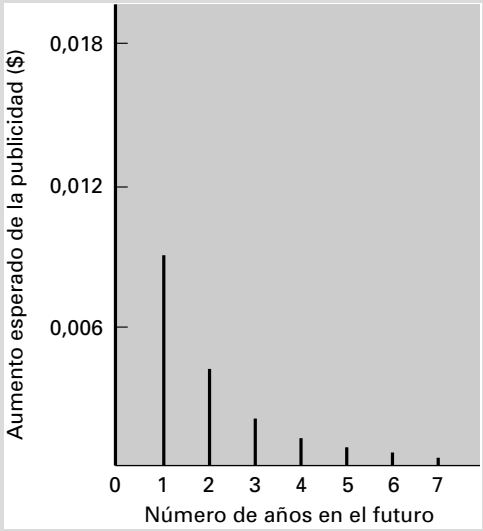


Figura 14.5. Aumentos futuros esperados de la publicidad local por hogar.

EJERCICIOS

Ejercicios básicos

14.10. Considere los siguientes modelos estimados utilizando un análisis de regresión aplicado a datos de series temporales. ¿Qué efecto produce a largo plazo un aumento de x de 1 unidad en el periodo t?

a) $y_t = 10 + 2x_t + 0{,}34y_{t-1}$
b) $y_t = 10 + 2{,}5x_t + 0{,}24y_{t-1}$
c) $y_t = 10 + 2x_t + 0{,}64y_{t-1}$
d) $y_t = 10 + 4{,}3x_t + 0{,}34y_{t-1}$

14.11. Un analista de mercado tiene interés en saber cuál es la cantidad media de dinero que gastan al año los estudiantes universitarios en ropa. Basándose en 25 años de datos anuales, se ha obtenido la siguiente regresión estimada por mínimos cuadrados:

$$y_t = \underset{}{50{,}72} + \underset{(0{,}047)}{0{,}142x_{1t}} + \underset{(0{,}021)}{0{,}027x_{2t}} + \underset{(0{,}136)}{0{,}432y_{t-1}}$$

donde

y = gasto por estudiante, en dólares, en ropa
x_1 = renta disponible por estudiante, en dólares, tras el pago de la matrícula, las tasas y la manutención
x_2 = índice de publicidad sobre ropa destinada al mercado estudiantil

Los números entre paréntesis que se encuentran debajo de los coeficientes son los errores típicos de los coeficientes.

a) Contraste al nivel del 5 por ciento la hipótesis nula de que, manteniéndose todo lo demás constante, la publicidad no afecta a los gastos en ropa en este mercado frente a la hipótesis alternativa unilateral obvia.
b) Halle el intervalo de confianza al 95 por ciento del coeficiente de x_1 de la regresión poblacional.
c) Manteniendo fija la publicidad, ¿cuál sería el efecto esperado con el paso del tiempo de un aumento de la renta disponible por estudiante de 1 $ en el gasto en ropa?

Ejercicios aplicados

14.12. Utilice los datos del fichero **Retail Sales** para estimar el modelo de regresión

$$y_t = \beta_0 + \beta_1 x_t + \gamma y_{t-1} + \varepsilon_t$$

y contraste la hipótesis nula de que $\gamma = 0$, donde

y_t = ventas al por menor por hogar
x_t = renta disponible por hogar

14.13. Utilice el fichero de datos **Money UK**, que contiene observaciones del Reino Unido sobre la cantidad de dinero, en millones de libras (Y); la renta, en millones de libras (X_1); y el tipo de interés de las autoridades locales (X_2). Estime el modelo (véase la referencia bibliográfica 5)

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \gamma y_{t-1} + \varepsilon_t$$

y realice un informe sobre sus resultados.

14.14. El fichero de datos **Pension Funds** contiene datos sobre el rendimiento de mercado (X) de las acciones y el porcentaje (Y) que representan las acciones ordinarias al valor de mercado a finales de año en la cartera de los fondos privados de pensiones. Estime el modelo

$$y_t = \beta_0 + \beta_1 x_t + \gamma y_{t-1} + \varepsilon_t$$

y escriba un informe sobre sus resultados.

14.15. El fichero de datos **Income Canada** muestra observaciones trimestrales sobre la renta (Y) y sobre la oferta monetaria (X) de Canadá. Estime el modelo (véase la referencia bibliográfica 3)

$$y_t = \beta_0 + \beta_1 x_t + \gamma y_{t-1} + \varepsilon_t$$

y realice un informe sobre sus resultados.

14.16. El fichero de datos **Births Australia** muestra observaciones anuales sobre el primer parto de un nacido vivo del matrimonio actual (Y) y el número de primeros matrimonios (de mujeres) registrado en el año anterior (X) en Australia. Estime el modelo (véase la referencia bibliográfica 4)

$$y_t = \beta_0 + \beta_1 x_t + \gamma y_{t-1} + \varepsilon_t$$

y realice un informe sobre sus resultados.

14.17. El fichero de datos **Pinkham Sales** muestra observaciones anuales sobre las ventas unitarias (Y) y sobre los gastos publicitarios (X), ambos en miles de dólares, de Lydia E. Pinkham. Estime el modelo

$$\log y_t = \beta_0 + \beta_1 \log x_t + \gamma \log y_{t-1} + \varepsilon_t$$

y realice un informe sobre sus resultados (véase la referencia bibliográfica 2).

14.18. El fichero de datos **Thailand Consumption** muestra 29 observaciones anuales sobre el consumo privado (Y) y la renta disponible (X) de Tailandia. Ajuste el modelo de regresión

$$\log y_t = \beta_0 + \beta_1 \log x_{1t} + \gamma_2 \log y_{t-1} + \varepsilon_t$$

y realice un informe sobre sus resultados.

14.4. Sesgo de especificación

La especificación de un modelo estadístico que describa correctamente la conducta del mundo real es una tarea delicada y difícil. Sabemos que ningún modelo sencillo puede describir perfectamente la naturaleza de un proceso y los determinantes de sus resultados. El objetivo de la construcción de modelos es descubrir una formulación sencilla que refleje correctamente el proceso subyacente para las cuestiones de interés. Sin embargo, también debemos señalar que hay algunos casos en los que existe una divergencia considerable entre el modelo y la realidad que puede llevar a extraer conclusiones seriamente erróneas.

Hemos visto anteriormente algunas técnicas para especificar un modelo que refleje mejor el proceso. Nuestro uso de variables ficticias en los apartados 13.8 y 14.2 y las transformaciones de modelos no lineales en lineales en el 13.7 son importantes ejemplos. En este apartado examinamos las consecuencias de no incluir importantes variables de predicción en nuestro modelo de regresión.

Para formular un modelo de regresión, un investigador intenta relacionar la variable dependiente de interés con todos sus determinantes importantes. Por lo tanto, si adoptamos un modelo lineal, queremos incluir como variables independientes todas las variables que podrían influir considerablemente en la variable dependiente de interés. Para formular el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

suponemos implícitamente que el conjunto de variables independientes, X_1, X_2, ..., X_K, contiene todas las cantidades que afectan significativamente a la conducta de la variable dependiente, Y. Sabemos que en cualquier problema aplicado real hay otros factores que también afectan a la variable dependiente. La influencia conjunta de estos factores se absorbe dentro del término de error, ε_i. Puede plantearse un grave problema si se omite una variable importante de la lista de variables independientes.

Sesgo provocado por la exclusión de variables de predicción importantes

Cuando se omiten en el modelo variables de predicción importantes, las estimaciones de coeficientes por mínimos cuadrados incluidas en el modelo normalmente están sesgadas y las afirmaciones inferenciales habituales basadas en los contrastes de hipótesis o en los intervalos de confianza pueden ser seriamente engañosas. Además, el error del modelo estimado incluye el efecto de las variables omitidas y, por lo tanto, es mayor. En el raro caso en el que las variables omitidas no están correlacionadas con las variables independientes incluidas en el modelo de regresión, no existe este sesgo en la estimación de los coeficientes.

Examinemos un sencillo ejemplo sobre el mercado al por menor de gasolina. Supongamos que somos propietarios de la estación de servicio A, que vende gasolina, y que la estación de servicio B, que se encuentra a 100 metros de distancia, también vende gasolina. Creemos firmemente que si bajáramos el precio, las ventas unitarias aumentarían y que si lo subiéramos, las ventas unitarias disminuirían. Pero si la estación B subiera y bajara su precio, este precio también influiría en la variación de nuestras ventas unitarias. Por lo tanto, si no tenemos en cuenta el precio de la estación B y sólo consideramos nuestros pre-

cios cuando intentamos predecir las ventas unitarias, normalmente cometeremos graves errores en nuestra estimación de la relación entre nuestro precio y nuestras ventas unitarias. A continuación, mostramos este resultado matemáticamente.

Mostramos cómo se produce el sesgo en la estimación de los coeficientes de regresión mostrando el efecto de la omisión de una variable en un modelo con dos variables independientes:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Supongamos que en esta situación el analista excluye la variable x_2 y estima, en su lugar, el modelo de regresión

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \mu_i$$

Obsérvese que hemos utilizado dos símbolos diferentes para hacer hincapié en el hecho de que los estimadores de los coeficientes serán diferentes. En el modelo de regresión simple, el estimador del coeficiente de x_1 es

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}) y_i}{\sum_{i=1}^{n} (x_{1i} - \bar{x})^2}$$

Sustituyendo el modelo correcto con dos variables de predicción y determinando el valor esperado, observamos que

$$E[\hat{\alpha}_1] = E\left[\frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) y_i}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}\right] = E\left[\frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}\right]$$

Cuando calculamos el valor esperado, observamos que

$$E[\hat{\alpha}_1] = \beta_1 + \beta_2 \left[\frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1) x_{2i}}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}\right]$$

Vemos, pues, que el coeficiente de la variable X_1 está sesgado a menos que la correlación entre X_1 y X_2 sea 0.

Los resultados matemáticos anteriores muestran el sesgo de las estimaciones de los coeficientes que se produce cuando se omite una variable importante. En el Capítulo 13 mostramos matemáticamente y de una forma intuitiva que en las estimaciones de los coeficientes de un modelo de regresión múltiple influyen todas las variables independientes incluidas en el modelo. Por lo tanto, si omitimos una variable independiente importante, los coeficientes estimados del resto de las variables serán diferentes. El ejemplo 14.5 muestra este resultado numéricamente y debe estudiarse atentamente.

EJEMPLO 14.5. Modelo de regresión de las asociaciones de ahorro y crédito inmobiliario con una variable omitida (error de especificación del modelo)

Consideremos el ejemplo de las asociaciones de ahorro y crédito inmobiliario utilizado en el Capítulo 13. En ese ejemplo se hacía una regresión del margen porcentual anual de beneficios (Y) de las asociaciones de ahorro y crédito inmobiliario con respecto a sus ingresos porcentuales netos por dólar depositado (X_1) y el número de oficinas (X_2). En el ejemplo 13.3 estimamos los coeficientes de regresión y observamos que el modelo era

$$\hat{y} = 1{,}565 + \underset{(0{,}0556)}{0{,}237x_1} - \underset{(0{,}0000321)}{0{,}000249x_2} \qquad R^2 = 0{,}865$$

Savings and Loan

Una de las conclusiones de este análisis es que, dado un número fijo de oficinas, un aumento de los ingresos netos por dólar depositado de 1 unidad provoca un aumento esperado del margen de beneficios de 0,237 unidades. ¿Qué ocurriría si hiciéramos una regresión del margen de beneficios únicamente con respecto a los ingresos netos por dólar depositado utilizando los datos almacenados en el fichero **Savings and Loan**?

Solución

Utilizando los datos, hemos hecho una regresión del margen de beneficios (Y) con respecto a los ingresos netos por dólar depositado (X_1) y hemos observado que el modelo era

$$\hat{y} = 1{,}326 - \underset{(0{,}036)}{0{,}169x_1} \qquad R^2 = 0{,}50$$

Comparando los dos modelos ajustados, observamos que una de las consecuencias de omitir X_2 es que la variabilidad porcentual explicada, R^2, disminuye considerablemente.

La omisión produce, sin embargo, un efecto más serio en el coeficiente de los ingresos netos. En el modelo de regresión múltiple, un aumento de los ingresos netos de 1 unidad elevó los beneficios en 0,237, mientras que en el modelo de regresión simple el efecto fue una disminución de 0,169. Este resultado va claramente en contra de la intuición: no es de esperar que un aumento de los ingresos netos reduzca el margen de beneficios. En los dos modelos, rechazaríamos la hipótesis nula de que no existe una relación. Aquí vemos el resultado del estimador sesgado del coeficiente que se obtiene cuando no se incluye una variable importante, X_2, en el modelo. Sin incluir el efecto condicionado del número de oficinas, obtenemos un estimador sesgado.

Este ejemplo ilustra magníficamente la cuestión. Si no se incluye una variable explicativa importante en el modelo de regresión, cualquier conclusión que se extraiga sobre los efectos de otras variables independientes puede ser seriamente engañosa. En este caso, hemos visto que la introducción de otra variable relevante más podría muy bien alterar la conclusión de la existencia de una relación negativa significativa y sustituirla por la conclusión de la existencia de una relación positiva significativa. Observando los datos de la Tabla 13.1, es posible obtener más información. En la segunda parte del periodo, al menos, el margen de beneficios disminuyó y los ingresos netos aumentaron, lo que sugiere la existencia de una relación negativa entre estas variables. Sin embargo, los datos revelan un aumento del número de oficinas durante ese mismo periodo, lo que sugiere la posibilidad

de que este factor fuera la causa de la disminución del margen de beneficios. La única forma legítima de distinguir los efectos de estas dos variables independientes en la variable dependiente es analizarlas conjuntamente en una ecuación de regresión. Este ejemplo muestra la importancia de utilizar el modelo de regresión múltiple en lugar de la ecuación de regresión lineal simple cuando hay más de una variable independiente relevante.

EJERCICIOS

Ejercicios básicos

14.19. Suponga que el verdadero modelo lineal de un proceso era

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

y que ha estimado incorrectamente el modelo

$$Y = \alpha_0 + \alpha_1 X_2$$

Interprete y contraste los coeficientes de X_2 estimados en los dos modelos. Muestre el sesgo que se produce utilizando el segundo modelo.

14.20. Suponga que una relación de regresión viene dada por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Si se estima la regresión lineal simple de Y con respecto a X_1 a partir de una muestra de n observaciones, la estimación resultante de la pendiente β_1 generalmente está sesgada. Sin embargo, en el caso especial en el que la correlación muestral entre X_1 y X_2 es 0, no ocurre así. De hecho, en ese caso la estimación es la misma independientemente de que se incluya o no X_2 en la ecuación de regresión.

a) Explique verbalmente por qué es cierta esta afirmación.

b) Demuestre algebraicamente que esta afirmación es cierta.

Ejercicios aplicados

14.21. Transportation Research Inc. le ha pedido que formule algunas ecuaciones de regresión múltiple para estimar el efecto de algunas variables en el ahorro de combustible. Los datos para realizar este estudio se encuentran en el fichero de datos **Motors** y la variable dependiente está en millas por galón —milpgal— conforme a la certificación del Departamento de Transporte.

a) Formule una ecuación de regresión que utilice la potencia de los vehículos —horsepower— y el peso de éstos —weight— como variables independientes. Interprete los coeficientes.

b) Formule una segunda regresión sesgada que no incluya el peso de los vehículos. ¿Qué conclusiones puede extraer sobre el coeficiente de la potencia?

14.22. Utilice los datos del fichero **Citydat** para estimar una ecuación de regresión que permita averiguar el efecto marginal del porcentaje de locales comerciales en el valor de mercado por vivienda ocupada por su propietario (Hseval). Incluya en su ecuación de regresión múltiple el porcentaje de viviendas ocupadas por sus propietarios (Homper), el porcentaje de locales industriales (Indper), el número mediano de habitaciones por vivienda (sizehse) y la renta per cápita (Incom72) como variables de predicción adicionales. Las variables están incluidas en su disco de datos. Indique qué variables son significativas. Su ecuación final debe incluir solamente las variables significativas. Haga una segunda regresión excluyendo el número mediano de habitaciones por vivienda. Interprete el nuevo coeficiente del porcentaje de locales comerciales que se obtiene en la segunda regresión. Compare los dos coeficientes.

14.5. Multicolinealidad

Si se especifica correctamente un modelo de regresión y se satisfacen los supuestos, las estimaciones por mínimos cuadrados son las mejores que pueden lograrse. No obstante, en algunas circunstancias ¡pueden no ser muy buenas!

Para ilustrarlo supongamos que queremos desarrollar un modelo para predecir las ventas unitarias en función de nuestro precio y del precio del competidor. Imaginemos ahora que estamos en la afortunada posición del científico de laboratorio, que somos capaces de diseñar el experimento para estudiar este problema. El mejor enfoque para seleccionar las observaciones depende algo de los objetivos del análisis, pero hay mejores estrategias.

Existen, sin embargo, opciones que no elegiríamos. Por ejemplo, no elegiríamos los mismos valores de las variables independientes para todas las observaciones. Tampoco seleccionaríamos variables independientes que estén muy correlacionadas. En el apartado 13.2 vimos que sería imposible estimar los coeficientes si las variables independientes estuvieran perfectamente correlacionadas. Y en el 13.4 vimos que la varianza de los estimadores de los coeficientes aumenta a medida que la correlación se aleja de 0. En la Figura 14.6 vemos ejemplos de correlación perfecta entre las variables X_1 y X_2. En estos gráficos vemos que las variaciones de una variable están relacionadas directamente con las variaciones de la otra. Supongamos ahora que estuviéramos intentando utilizar valores de las variables independientes como éstos para estimar los coeficientes del modelo de regresión

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

La inutilidad de esa tarea es evidente. Si X_1 varía al mismo tiempo que X_2, no podemos saber cuál de las variables independientes está relacionada realmente con la variación de Y. Si queremos evaluar los efectos de cada variable independiente por separado, es esencial que no varíen exactamente al unísono en el experimento. Los supuestos habituales del análisis de regresión múltiple excluyen los casos de correlación perfecta entre variables independientes.

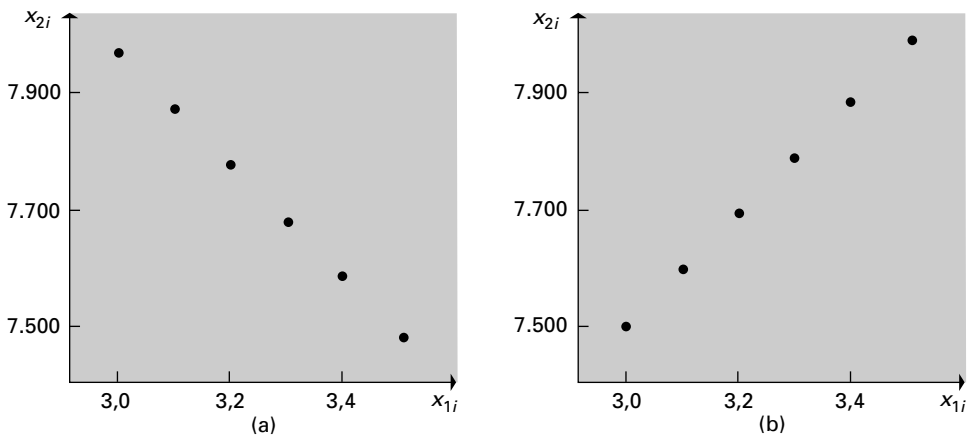


Figura 14.6. Dos diseños con correlación perfecta.

El uso de las variables independientes en la Figura 14.6 sería una mala elección. La 14.7 muestra un caso algo menos extremo. Aquí los puntos del diseño no se encuentran en una única línea recta, pero casi. En esta situación, los resultados suministran alguna información sobre la influencia de cada variable independiente, pero no mucha. Es posible calcular estimaciones por mínimos cuadrados de los coeficientes, pero estas estimaciones tendrían una elevada varianza. Como consecuencia, los coeficientes estimados no serán estadísticamente significativos, incluso aunque las relaciones sean muy estrechas. Este fenómeno se llama **multicolinealidad**. En el Capítulo 13 analizamos extensamente los efectos de las variables independientes correlacionadas.

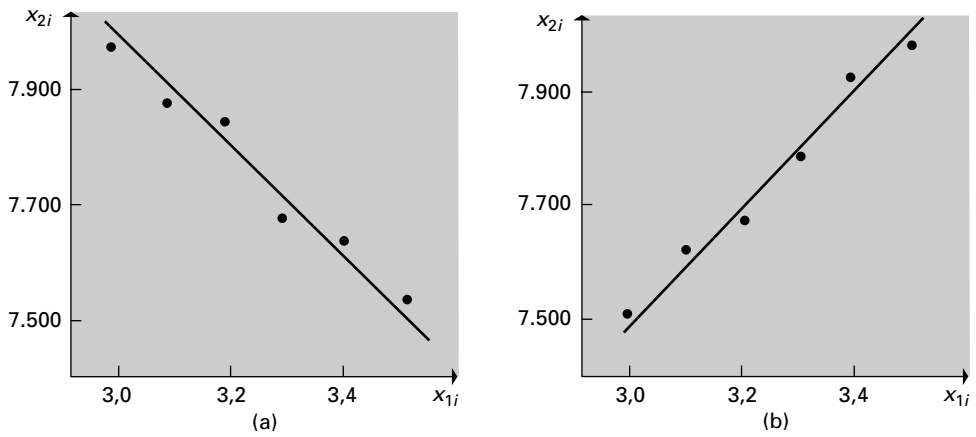


Figura 14.7. Dos diseños con una elevada correlación.

En la inmensa mayoría de los casos prácticos relacionados con el mundo de la empresa y la economía, no podemos controlar la elección de las observaciones de las variables sino que nos vemos obligados a trabajar con el conjunto de datos que el destino nos ha dado. En este contexto, pues, la multicolinealidad es un problema que no se debe a que se hayan elegido mal los datos sino a los datos de que se dispone para hacer el análisis. En el ejemplo de las asociaciones de ahorro y crédito inmobiliario del Capítulo 13, había una elevada correlación entre las variables independientes, pero ésa era la realidad del contexto del problema. En términos más generales, en las ecuaciones de regresión en las que hay varias variables independientes, el problema de multicolinealidad se debe a la existencia de pautas de estrechas intercorrelaciones entre las variables independientes. Quizá el aspecto más frustrante del problema, que puede resumirse en la existencia de datos que no suministran mucha información sobre los parámetros de interés, radique en que normalmente es poco lo que se puede hacer para resolverlo. Sin embargo, aun así es importante ser conscientes del problema y vigilar por si se plantea.

Hay algunos elementos que indican la posibilidad de que haya multicolinealidad. En primer lugar, siempre debe examinarse, por supuesto, una matriz de correlaciones simples de las variables independientes para averiguar si cualquiera de ellas está correlacionada individualmente, como hicimos en el extenso ejemplo del apartado 13.9. Otra indicación de la probable presencia de multicolinealidad es que parezca que un conjunto de variables independientes consideradas como un grupo ejerce una influencia considerable en la variable dependiente y que cuando se examinan por separado, por medio de contrastes de hipótesis, parezca que todas son individualmente insignificantes. En este caso, podría utilizarse una función lineal de las distintas variables para calcular una variable que sustituya a las distintas variables correlacionadas. Otra estrategia es hacer una regresión de las variables individuales independientes con respecto a todas las demás variables independientes del modelo. Eso puede mostrar complejas situaciones de multicolinealidad. Dada la presencia de multicolinealidad, en estas circunstancias sería imprudente extraer la conclusión de que una determinada variable independiente no afecta a la variable dependiente. Es preferible reconocer que el grupo en su conjunto es claramente influyente, pero los datos no son lo suficientemente informativos para poder distinguir con precisión los efectos de cada uno de sus miembros por separado.

Existe otro problema relacionado con éste si se incluyen en un modelo variables de predicción redundantes o irrelevantes. Si estas variables innecesarias están correlacionadas con las demás variables de predicción —y a menudo lo están—, la varianza de las estima-

ciones de los coeficientes de las variables importantes aumentará, como se señala en el apartado 13.4. Como consecuencia, disminuirá la eficiencia global de las estimaciones de los coeficientes. Debe tenerse cuidado de no incluir variables de predicción irrelevantes.

En las situaciones en las que la multicolinealidad es un problema, pueden utilizarse diversos enfoques. En todos ellos, es necesario analizar y valorar atentamente los objetivos del modelo y el entorno del problema que representa. En primer lugar, se puede eliminar una variable independiente que está estrechamente correlacionada con una o más variables independientes. Eso reducirá la varianza de la estimación de los coeficientes, pero, como se muestra en el apartado 14.4, se podría introducir un sesgo en la estimación de los coeficientes si la variable omitida es importante en el modelo. Se podría construir una nueva variable independiente que fuera una función de varias variables independientes estrechamente correlacionadas. Se podría sustituir por una nueva variable independiente que represente la misma influencia, pero no esté correlacionada con otras variables independientes. Ninguno de estos enfoques es siempre la solución perfecta. La multicolinealidad y las variables omitidas del apartado anterior son cuestiones que requieren una buena especificación del modelo basada en una buena valoración, en la experiencia y en la comprensión del contexto del problema.

EJERCICIOS

Ejercicios aplicados

14.23. En el modelo de regresión

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

es posible averiguar en qué medida existe multicolinealidad hallando la correlación entre X_1 y X_2 en la muestra. Explique por qué es así.

14.24. Un economista estima el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Las estimaciones de los parámetros β_1 y β_2 no son muy grandes en comparación con sus errores típicos respectivos. Pero el tamaño del coeficiente de determinación indica la existencia de una relación bastante estrecha entre la variable dependiente y el par de variables independientes. Una vez obtenidos estos resultados, el economista tiene firmes sospechas de la presencia de multicolinealidad. Como lo que más le interesa es saber cómo influye X_1 en la variable dependiente, decide que evitará el problema de multicolinealidad haciendo una regresión de Y con respecto a X_1 solamente. Comente esta estrategia.

14.25. Basándose en los datos de 63 países, se estimó el siguiente modelo por mínimos cuadrados:

$$\hat{y} = 0{,}58 - \underset{(0{,}019)}{0{,}052 x_1} - \underset{(0{,}042)}{0{,}005 x_2} \qquad R^2 = 0{,}17$$

donde

y = tasa de crecimiento del producto interior bruto real
x_1 = renta real per cápita
x_2 = tipo impositivo medio en porcentaje del producto nacional bruto

Los números situados debajo de los coeficientes son los errores típicos de los coeficientes. Una vez eliminada en el modelo la variable independiente X_1, la renta real per cápita, se estimó la regresión de la tasa de crecimiento del producto interior bruto real con respecto a X_2, el tipo impositivo medio, y se obtuvo el modelo ajustado

$$\hat{y} = 0{,}060 - \underset{(0{,}34)}{0{,}074 x_2} \qquad R^2 = 0{,}072$$

Comente este resultado.

14.6. Heterocedasticidad

El método de estimación por mínimos cuadrados y sus métodos inferenciales se basan en los supuestos tradicionales del análisis de regresión. Cuando se cumplen estos supuestos, la regresión por mínimos cuadrados proporciona un poderoso conjunto de instrumentos analí-

ticos. Sin embargo, cuando se viola uno o más de estos supuestos, los coeficientes estimados pueden ser ineficientes y las inferencias realizadas pueden ser engañosas.

En este apartado y en el siguiente, consideramos los problemas que plantean los supuestos relacionados con la distribución de los términos de error ε_i en el modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

Concretamente, hemos supuesto que estos errores tienen una varianza uniforme y no están correlacionados entre sí. En el siguiente apartado, examinamos la posibilidad de que existan errores correlacionados. Aquí analizamos el supuesto de la varianza uniforme.

Existen muchos ejemplos que sugieren la posibilidad de que la varianza no sea uniforme. Consideremos una situación en la que nos interesa conocer los factores que afectan a la producción de una industria. Recogemos datos de varias empresas que contienen medidas de la producción y otras posibles variables de predicción. Si estas empresas son de diferente tamaño, la producción total varía. Es probable, además, que la varianza de la medida de la producción sea mayor en las grandes empresas que en las pequeñas. Eso se debe a la observación de que hay más factores que afectan a los términos de error en una empresa grande que en una pequeña. Por lo tanto, los términos de error serán mayores tanto en los términos positivos como en los negativos.

Se dice que los modelos en los que los términos de error no tienen todos la misma varianza muestran **heterocedasticidad**. Cuando este fenómeno está presente, el método de mínimos cuadrados no es el más eficiente para estimar los coeficientes del modelo de regresión. Además, los métodos habituales para obtener intervalos de confianza y contrastes de hipótesis de estos coeficientes ya no son válidos. Necesitamos, pues, métodos para averiguar si existe heterocedasticidad. La mayoría de los métodos habituales comprueban el supuesto de la varianza constante de los errores frente a alguna alternativa razonable. Podemos observar que la magnitud de la varianza de los errores está relacionada directamente con una de las variables de predicción independientes. Otra posibilidad es que la varianza aumente con el valor esperado de la variable dependiente.

En nuestro modelo de regresión estimado, podemos obtener estimaciones de los valores esperados de la variable dependiente utilizando

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki}$$

Y podemos estimar, a su vez, los términos de error, ε_i, mediante los residuos

$$e_i = y_i - \hat{y}_i$$

A menudo observamos que las técnicas gráficas son útiles para detectar la presencia de heterocedasticidad. En la práctica, trazamos diagramas de puntos dispersos de los residuos en relación con las variables independientes y los valores predichos, $\hat{y}_i$, de la regresión. Consideremos, por ejemplo, la Figura 14.8, que muestra posibles gráficos del residuo, e_i, en relación con la variable independiente X_{1i}. En la parte (a) de la figura, vemos que la magnitud de los errores tiende a aumentar conforme mayores son los valores de X_1, lo que indica que las varianzas de los errores no son constantes. En cambio, la parte (b) de la figura muestra que no existe una relación sistemática entre los errores y X_1. Por lo tanto, en la parte (b) no existen pruebas de que la varianza no sea uniforme.

En el Capítulo 13 desarrollamos un modelo de regresión por mínimos cuadrados para estimar la relación entre el margen de beneficios de las asociaciones de ahorro y crédito

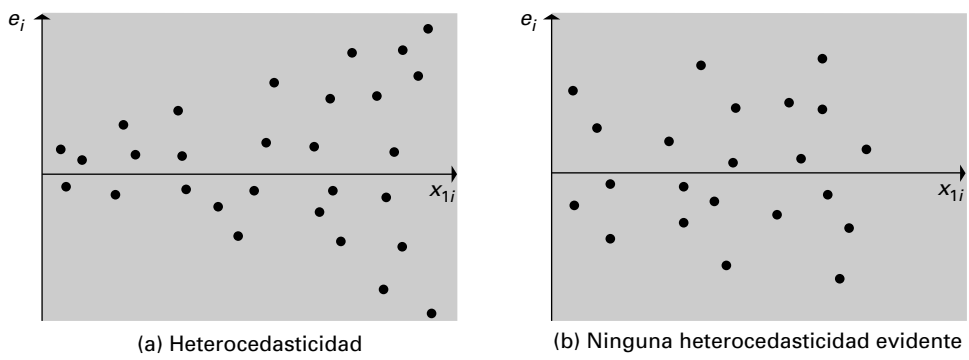


Figura 14.8. Gráficos de los residuos en relación con una variable independiente.

inmobiliario (Y) y los ingresos netos por dólar depositado (X_1) y el número de oficinas (X_2) por medio del modelo

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

Consideremos el modelo de regresión estimado de la Figura 13.3. Calculamos los residuos de todas las observaciones utilizando el método expuesto en el extenso problema del apartado 13.9. En las Figuras 14.9 y 14.10 presentamos diagramas de puntos dispersos de los residuos en relación con los ingresos por dólar depositado y en relación con el número de oficinas. El examen de estos diagramas indica que no parece que exista ninguna relación entre la magnitud de los residuos y cualquiera de las dos variables independientes. La Figura 14.11 presenta un diagrama de puntos dispersos de los residuos en relación con el valor predicho de la variable dependiente. De nuevo, no parece que exista ninguna relación entre el valor predicho de Y y la magnitud de los residuos. Basándonos en el examen de los gráficos de los residuos, no encontramos pruebas de la existencia de heterocedasticidad.

A continuación, examinamos un método más formal para detectar la presencia de heterocedasticidad y para estimar los coeficientes de los modelos de regresión cuando se tienen firmes sospechas de que se viola el supuesto de las varianzas constantes de los errores. Hay muchos tipos de heterocedasticidad que pueden detectarse por medio de diversos métodos. Examinaremos uno de ellos que puede utilizarse para detectar la presencia de heterocedasticidad cuando la varianza del término de error tiene una relación lineal con el valor predicho de la variable dependiente.

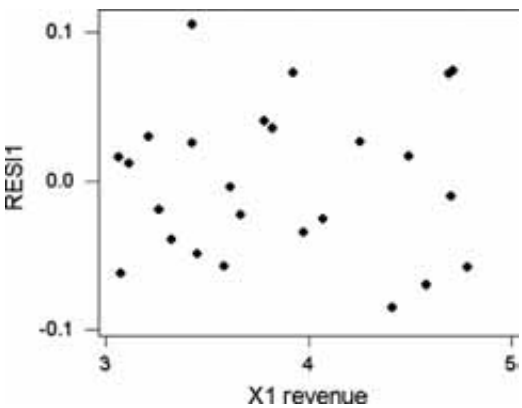


Figura 14.9. Gráfico de los residuos en relación con los ingresos por dólar depositado.

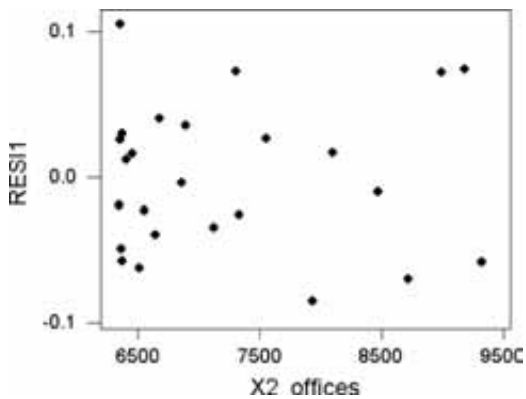


Figura 14.10. Gráfico de los residuos en relación con el número de oficinas.

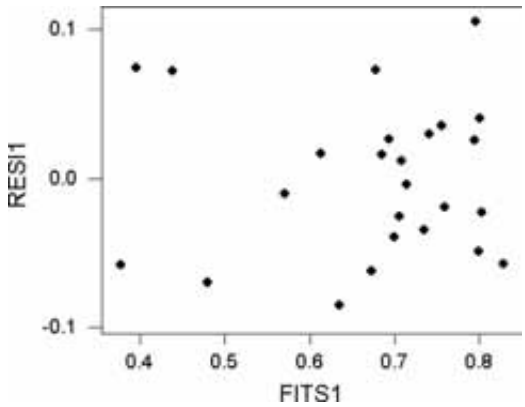


Figura 14.11. Dos diseños con una elevada correlación.

Contraste de la presencia de heterocedasticidad

Consideremos un modelo de regresión

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

que relaciona una variable dependiente con K variables independientes y se basa en n conjuntos de observaciones. Sean b_0, b_1, ..., b_K la estimación por mínimos cuadrados de los coeficientes del modelo, con los valores predichos

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki}$$

y sean los residuos del modelo ajustado

$$e_i = y_i - \hat{y}_i$$

Para contrastar la hipótesis nula de que los términos de error, ε_i, tienen todos ellos la misma varianza frente a la alternativa de que sus varianzas dependen de los valores esperados

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_K x_{Ki}$$

estimamos una regresión simple. En esta regresión, la variable dependiente es la raíz cuadrada de los residuos —es decir, e_i^2— y la variable independiente es el valor predicho, $\hat{y}_i$,

$$e_i^2 = a_0 + a_1 \hat{y}_i \qquad \textbf{(14.2)}$$

Sea R^2 el coeficiente de determinación de esta regresión auxiliar. En ese caso, en un contraste de nivel de significación α, la hipótesis nula se rechaza si nR^2 es mayor que $\chi^2_{1,\alpha}$, donde $\chi^2_{1,\alpha}$ es el valor crítico de la variable aleatoria ji-cuadrado con 1 grado de libertad y una probabilidad de error α.

Pondremos un ejemplo de este contraste utilizando el ejemplo de las asociaciones de ahorro y crédito inmobiliario. La Figura 14.12 muestra un subconjunto de la salida Minitab del análisis de regresión. Se empleó el programa Minitab para calcular los cuadrados de los residuos y se realizó una regresión de los residuos con respecto al valor predicho.

A partir de la regresión de los cuadrados de los residuos con respecto a los valores predichos, obtenemos el modelo estimado

$$e^2 = \underset{(0,00433)}{0,00621 + 0,00550\hat{y}} \qquad R^2 = 0,066$$

Figura 14.12. Regresión de los cuadrados de los residuos con respecto al valor predicho (salida Minitab).

```
The regression equation is
ResSquared = 0.00621 - 0.00550 FITS1

Predictor        Coef      SE Coef        T        P
Constant     0.006211     0.002970     2.09    0.048
FITS1       -0.005503     0.004327    -1.27    0.216

S = 0.002742   R-Sq = 6.6%    R-Sq(adj) = 2.5%

Analysis of Variance

Source            DF           SS           MS      F       P
Regression         1  0.000012158  0.000012158   1.62   0.216
Residual Error    23  0.000172939  0.000007519
Total             24  0.000185097
```

La regresión contiene $n = 25$ observaciones y, por lo tanto, el estadístico del contraste es

$$nR^2 = (25)(0{,}066) = 1{,}65$$

En la Tabla 7 del apéndice observamos que para un contraste al nivel de significación del 10 por ciento

$$\chi^2_{1,0,10} = 2{,}71$$

Por lo tanto, no podemos rechazar la hipótesis nula de que en el modelo de regresión los valores predichos tienen una varianza uniforme. Eso confirma nuestras conclusiones iniciales basadas en el examen de los diagramas de puntos dispersos de los residuos de las Figuras 14.9, 14.10 y 14.11.

Supongamos ahora que hubiéramos rechazado la hipótesis nula de que la varianza era uniforme. En ese caso, el método ordinario de mínimos cuadrados no sería el método de estimación adecuado para el modelo inicial. Existen varias estrategias de estimación dependiendo de cómo sean de poco uniformes los errores. La mayoría de los métodos implican la transformación de las variables del modelo de manera que los términos de error tengan una magnitud uniforme en el rango del modelo. Consideremos el ejemplo en el que la varianza de los términos de error es directamente proporcional al cuadrado del valor esperado de la variable dependiente. En este caso, podríamos expresar aproximadamente el término de error del modelo de la forma siguiente:

$$\varepsilon_i = \hat{y}_i \delta_i$$

donde δ_i es una variable aleatoria que tiene una varianza uniforme en el rango del modelo de regresión. Utilizando este término de error, el modelo de regresión sería

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \hat{y}_i \delta_i$$

En esta aproximación, el término de error aumenta linealmente con el valor esperado, lo cual implica que la varianza aumenta con el cuadrado del valor esperado. Aquí podemos obtener un término de error cuya magnitud es uniforme en el modelo dividiendo cada término de los dos miembros de la ecuación por $\hat{y}_i$. Cuando se parte de esta forma concreta,

se utiliza un sencillo método de dos etapas para estimar los parámetros del modelo de regresión. En la primera etapa, se estima el modelo por mínimos cuadrados de la forma habitual y se registran los valores predichos, $\hat{y}_i$, de la variable dependiente. En la segunda etapa, se estima la ecuación de regresión

$$\frac{y_i}{\hat{y}_i} = \beta_0 \frac{1}{\hat{y}_i} + \beta_1 \frac{x_{1i}}{\hat{y}_i} + \beta_2 \frac{x_{2i}}{\hat{y}_i} + \cdots + \beta_K \frac{x_{Ki}}{\hat{y}_i} + \delta_i$$

con un término de error que satisface los supuestos habituales del análisis de regresión. En este modelo, hacemos una regresión de $y_i/\hat{y}_i$ con respecto a las variables independientes $1/\hat{y}_1$, $x_{1i}/\hat{y}_1$, $x_{2i}/\hat{y}_1$, ..., $x_{Ki}/\hat{y}_1$. Este modelo no incluye una constante y la mayoría de los paquetes estadísticos tienen una opción que calcula estimaciones de los coeficientes excluyendo el término constante. Los coeficientes estimados son las estimaciones de los coeficientes del modelo original. Existen otros muchos métodos en cualquier buen libro de econometría en el apartado dedicado a los «mínimos cuadrados ponderados».

También pueden aparecer errores heterocedásticos si se estima un modelo de regresión lineal en circunstancias en las que es adecuado un modelo logarítmico-lineal. Cuando el proceso es tal que es adecuado un modelo logarítmico-lineal, debemos hacer las transformaciones y estimar un modelo logarítmico-lineal. Tomando logaritmos, disminuye la influencia de las grandes observaciones, sobre todo si éstas se deben al crecimiento porcentual con respecto a momentos anteriores: una pauta de crecimiento exponencial. El modelo resultante a menudo parecerá que está libre de heterocedasticidad. Los modelos logarítmico-lineales a menudo son adecuados cuando los datos estudiados son series temporales de variables económicas, como el consumo, la renta y el dinero, que tienden a crecer exponencialmente con el paso del tiempo.

EJERCICIOS

Ejercicios aplicados

14.26. En el Capítulo 12, se estimó por mínimos cuadrados la regresión de las ventas al por menor por hogar con respecto a la renta disponible por hogar. Los datos se encuentran en la Tabla 12.1 y la 12.2 muestra los residuos y los valores predichos de la variable dependiente.

a) Averigüe gráficamente si existe heterocedasticidad en los errores de regresión.

b) Averigüe si existe heterocedasticidad utilizando un contraste formal.

14.27. Considere un modelo de regresión que utiliza 48 observaciones. Sea e_i los residuos de la regresión ajustada e $\hat{y}_i$ los valores predichos de la variable dependiente dentro del rango de la muestra. La regresión por mínimos cuadrados de e_i^2 con respecto a $\hat{y}_i$ tiene un coeficiente de determinación de 0,032. ¿Qué conclusiones puede extraer de este resultado?

14.28. El fichero de datos **Household Income** contiene datos de 50 estados de Estados Unidos. Las variables incluidas en el fichero son el porcentaje de mujeres que participan en la población activa (y), la mediana de la renta personal de los hogares (X_1), el número medio de años de estudios de las mujeres (X_2) y la tasa de desempleo de las mujeres (X_3).

a) Calcule la regresión múltiple de Y con respecto a X_1, X_2 y X_3.

b) Compruebe gráficamente la presencia de heterocedasticidad en los errores de regresión.

c) Utilice un contraste formal para detectar la presencia de heterocedasticidad.

14.7. Errores autocorrelacionados

En este apartado, vemos qué ocurre con el modelo de regresión si los términos de error están correlacionados entre sí. Hasta ahora hemos supuesto que los errores aleatorios de nuestro modelo son independientes. Sin embargo, en muchos problemas empresariales y económicos utilizamos datos de series temporales. Cuando se analizan datos de series temporales, el término de error representa el efecto de todos los factores, salvo las variables independientes, que influyen en la variable dependiente. En los datos de series temporales, la conducta de muchos de estos factores puede ser bastante parecida en varios periodos de tiempo y el resultado sería una correlación entre los términos de error que están cerca en el tiempo.

Para hacer hincapié en el hecho de que las observaciones son observaciones de series temporales, colocamos el subíndice t y formulamos el modelo de regresión de la siguiente manera:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

En la regresión múltiple, los contrastes de hipótesis y los intervalos de confianza suponen que los errores son independientes. Si no lo son, los errores típicos estimados de los coeficientes están sesgados. Por ejemplo, puede demostrarse que, si existe una correlación positiva entre los términos de error de observaciones de series temporales adyacentes, la estimación del error típico de los coeficientes por mínimos cuadrados es demasiado pequeña. Como consecuencia, el estadístico t de Student calculado para el coeficiente es demasiado grande. Eso puede llevarnos a concluir que algunos coeficientes son significativamente diferentes de 0 —rechazando la hipótesis nula $\beta_j = 0$— cuando, en realidad, no debe rechazarse. Además, los intervalos de confianza estimados serían demasiado estrechos.

Es, pues, fundamental en las regresiones con datos de series temporales contrastar la hipótesis de que los términos de error no están correlacionados entre sí. El hecho de que los errores de primer orden estén correlacionados a lo largo del tiempo se conoce con el nombre de problema de **errores autocorrelacionados**. Cuando estudiamos este problema, es útil tener presente alguna estructura de correlación. Un modelo atractivo es que el error en el periodo t, ε_t, esté estrechamente correlacionado con el error del periodo anterior, ε_{t-1}, pero menos correlacionado con los errores de dos o más periodos anteriores. Definimos

$$\text{Corr}\,(\varepsilon_t,\ \varepsilon_{t-1}) = \rho$$

donde ρ es un coeficiente de correlación y, por lo tanto, su rango es de -1 a $+1$, como vimos en el Capítulo 12. En la mayoría de las aplicaciones, nos interesan sobre todo los valores positivos del coeficiente de correlación. En el caso de los errores que están separados por l periodos, la autocorrelación puede definirse de la siguiente manera:

$$\text{Corr}\,(\varepsilon_t,\ \varepsilon_{t-l}) = \rho^l$$

Como consecuencia, la correlación disminuye rápidamente a medida que aumenta el número de periodos de separación. Vemos, pues, que la correlación entre los errores que están separados en el tiempo es relativamente débil, mientras que la correlación entre los errores que están próximos en el tiempo posiblemente sea bastante estrecha.

Ahora bien, si suponemos que los errores ε_t tienen todos ellos la misma varianza, es posible demostrar que la estructura de autocorrelación corresponde al modelo

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

donde la variable aleatoria u_t tiene una media de 0 y una varianza constante σ^2 y no está autocorrelacionada. Este modelo de conducta autocorrelacionada se denomina modelo autorregresivo de primer orden. Examinando esta ecuación, vemos que el valor que toma el error en el periodo t, ε_t, depende de su valor en el periodo anterior (el grado de dependencia depende del coeficiente de correlación ρ) y de un segundo término aleatorio μ_t. Este modelo se muestra en la Figura 14.13, que contiene gráficos temporales de errores generados por el modelo para valores de $\rho = 0$, 0,3, 0,6 y 0,9. El caso $\rho = 0$ corresponde a la ausencia de autocorrelación de los errores. En la parte (a) de la figura podemos ver que no existe una pauta evidente en la progresión de los errores a lo largo del tiempo. El valor que toma uno no influye en los valores de los demás. A medida que pasamos de una autocorrelación relativamente débil ($\rho = 0{,}3$) a una autocorrelación bastante estrecha ($\rho = 0{,}9$), en las partes (b), (c) y (d), la pauta que muestran los errores a lo largo del tiempo es cada vez menos irregular, de manera que en la parte (d) está bastante claro que es probable que el valor de un error esté relativamente cerca de su vecino inmediato.

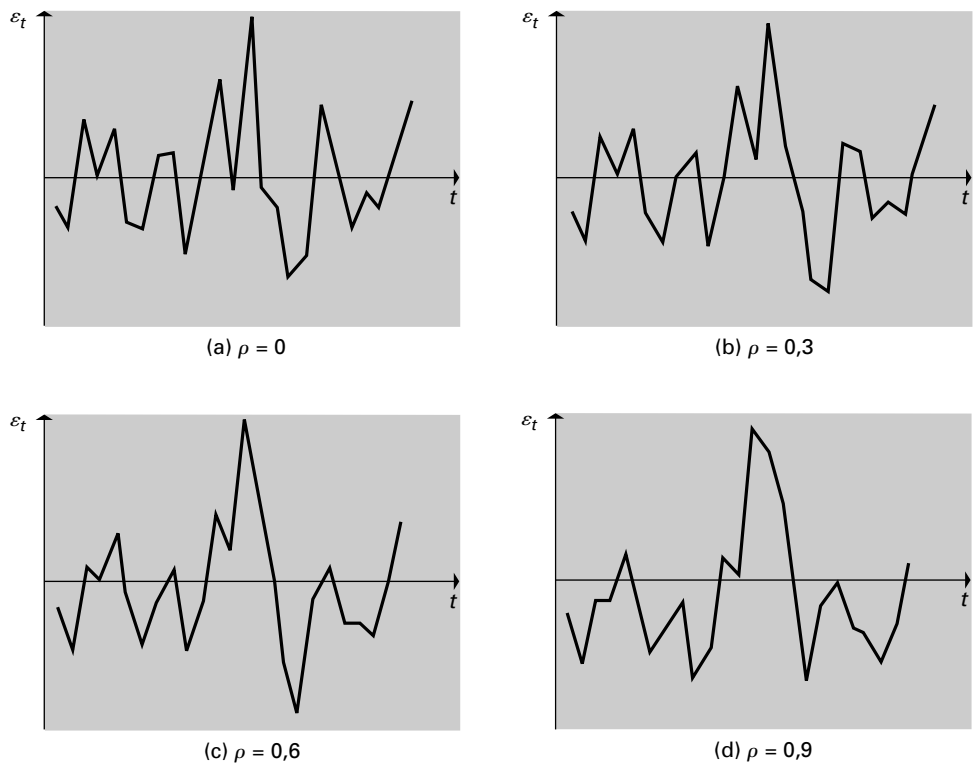


Figura 14.13. Gráficos temporales de los residuos de regresiones cuyos términos de error siguen un proceso autorregresivo de primer orden.

El examen de la Figura 14.13 sugiere que los métodos gráficos pueden ser útiles para detectar la presencia de errores autocorrelacionados. Lo ideal sería poder representar gráficamente los errores del modelo, ε_t, pero éstos son desconocidos, por lo que normalmente examinamos el gráfico de los residuos del modelo de regresión. En concreto, podríamos examinar un gráfico temporal de los residuos como el que muestra la Figura 14.14 en el caso de la regresión de las asociaciones de ahorro y crédito inmobiliario. Este gráfico de series temporales se ha realizado utilizando el programa Minitab.

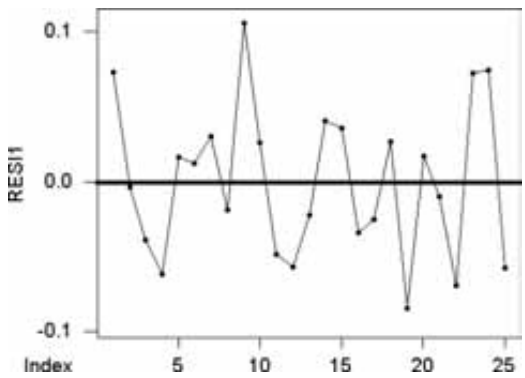


Figura 14.14. Gráfico de series temporales de los residuos de la regresión de las asociaciones de ahorro y crédito inmobiliario.

Examinando el gráfico de series temporales de la Figura 14.14, no vemos ninguna autocorrelación de los residuos sino la pauta irregular de la Figura 14.13(a). Ésta es una prueba en contra de la existencia de autocorrelación. Sin embargo, como el problema es tan importante, es deseable tener un contraste más formal de la hipótesis de que no existe ninguna autocorrelación en los errores de un modelo de regresión.

El contraste que más se utiliza es el **contraste de Durbin-Watson**, basado en los residuos del modelo, e_t. El estadístico del contraste, d, se calcula de la siguiente manera:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

y el método de contraste se describe a continuación.

Podemos demostrar que el estadístico de Durbin-Watson puede expresarse aproximadamente de la forma siguiente:

$$d = 2(1 - r)$$

donde r es la estimación muestral de la correlación poblacional, ρ, entre los errores adyacentes. Si los errores no están autocorrelacionados, entonces r es aproximadamente 0 y d es aproximadamente 2. En cambio, con una correlación positiva los valores de d son bajos y 0 es el límite inferior y con una correlación negativa, los valores de d son altos y 4 es el límite superior. Hay una dificultad teórica cuando se basan los contrastes de los errores autocorrelacionados en el estadístico de Durbin-Watson. El problema estriba en que la distribución muestral efectiva de d, incluso cuando la hipótesis de la ausencia de autocorrelación es verdadera, depende de los valores de las variables independientes. Es evidentemente inviable calcular la distribución correspondiente a todos los conjuntos posibles de valores de las variables independientes. Afortunadamente, se sabe que cualesquiera que sean las variables independientes, la distribución de d se encuentra entre las distribuciones de otras dos variables aleatorias cuyos puntos porcentuales pueden calcularse. La Tabla 12 del apéndice muestra los puntos de corte de estas variables aleatorias en el caso de los contrastes a niveles de significación del 1 y el 5 por ciento. La tabla indica los valores de d_L y d_U correspondientes a diversas combinaciones de n y K. Se rechaza la hipótesis nula de que no existe ninguna autocorrelación frente a la hipótesis alternativa de que existe una autocorrelación positiva si el valor calculado de d es menor que el de d_L. Se acepta la hipótesis nula si el valor de d es mayor que el de d_U y menor que $4 - d_U$, mientras que el

contraste no es concluyente si d se encuentra entre d_L y d_U. Por último, si el estadístico d es mayor que $4 - d_L$, concluiríamos que no existe ninguna autocorrelación negativa. Esta compleja pauta se muestra en la Figura 14.15.

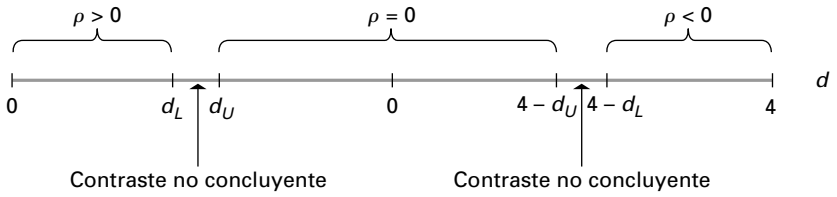


Figura 14.15. Regla de decisión para el contraste de Durbin-Watson.

Contraste de Durbin-Watson

Consideremos el modelo de regresión

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

basado en conjuntos de n observaciones. Nos interesa averiguar si los términos de error están autocorrelacionados y siguen un modelo autorregresivo de primer orden

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

donde u_t no está autocorrelacionado.

El contraste de la hipótesis nula de que no existe autocorrelación

$$H_0 : \rho = 0$$

se basa en el estadístico de Durbin-Watson:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(14.3)}$$

donde los e_t son los residuos cuando la ecuación de regresión se estima por mínimos cuadrados. Cuando la hipótesis alternativa es que existe una autocorrelación positiva de los errores, es decir,

$$H_1 : \rho > 0$$

la regla de decisión es la siguiente:

Rechazar H_0 si $d < d_L$
Aceptar H_0 si $d > d_U$
Contraste no concluyente si $d_L < d < d_U$

donde d_L y d_U corresponden a los valores de n y K y los niveles de significación del 1 y el 5 por ciento que se encuentran en la Tabla 12 del apéndice.

A veces queremos hacer un contraste frente a la hipótesis alternativa de que existe una autocorrelación negativa, es decir,

$$H_1 : \rho < 0$$

En ese caso, la regla de decisión es la siguiente:

Rechazar H_0 si $d > 4 - d_L$
Aceptar H_0 si $d < 4 - d_U$
Contraste no concluyente si $4 - d_L > d > 4 - d_U$

La mayoría de los programas informáticos calculan opcionalmente el estadístico d de Durbin-Watson como parte de la estimación de la regresión. La Figura 14.16 muestra la salida Minitab del ejemplo de las asociaciones de ahorro y crédito inmobiliario con el estadístico d de Durbin-Watson calculado. Éste es igual a 1,95 y en el apéndice vemos que cuando $\alpha = 0{,}01$, $k = 2$ y $n = 25$, los valores críticos son $d_L = 0{,}98$ y $d_U = 1{,}30$. Por lo tanto, $H_0: \rho = 0$ no puede rechazarse, por lo que concluimos que los términos de error no están autocorrelacionados.

Figura 14.16. Cálculo del estadístico de Durbin-Watson d (salida Minitab).

```
The regression equation is
Y profit = 1.56 + 0.237 X1 revenue -0.000249 X2 offices

Predictor        Coef        StDev        T        P
Constant      1.56450      0.07940    19.70    0.000
X1 reven      0.23720      0.05556     4.27    0.000
X2 offit  -0.00024908   0.00003205    -7.77    0.000

S = 0.05330    R-Sq = 86.5%   R-Sq(adj) = 85.3%

 nalysis of Variance

Source            DF         SS          MS        F
Regression         2    0.40151     0.20076    70.66     0.000
Residual Error    22    0.06250     0.00284
Total             24    0.46402

Durbin-Watson statistic = 1.95
```

Estimación de las regresiones con errores autocorrelacionados

Cuando concluimos, basándonos en el contraste de Durbin-Watson, que tenemos errores autocorrelacionados, hay que modificar el método de regresión para eliminar el efecto de estos errores autocorrelacionados. Normalmente, se hace mediante una transformación adecuada de las variables utilizadas en el método de estimación de la regresión. Desarrollamos el método básico en los pasos siguientes. En primer lugar, consideramos un modelo de regresión múltiple con errores autocorrelacionados:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

El mismo modelo de regresión en el periodo $t - 1$:

$$y_{t-1} = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \cdots + \beta_k x_{k,t-1} + \varepsilon_{t-1}$$

Multiplicando los dos miembros de esta ecuación por ρ, la correlación entre los errores adyacentes nos da

$$\rho y_{t-1} = \beta_0 + \beta_1 \rho x_{1,t-1} + \beta_2 \rho x_{2,t-1} + \cdots + \beta_k \rho x_{k,t-1} + \rho \varepsilon_{t-1}$$

A continuación, restamos esta ecuación de la primera para obtener

$$y_t - \rho y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_{1t} - \rho x_{1,t-1}) + \beta_2(x_{2t} - \rho x_{2,t-1}) + \cdots + \beta_k(x_{kt} - \rho_{xk,t-1}) + \mu_t$$

donde

$$u_t = \varepsilon_t - \rho\varepsilon_{t-1}$$

y la variable aleatoria u_t tiene una varianza uniforme y no está autocorrelacionada. Vemos que ahora tenemos un modelo de regresión que relaciona la variable dependiente $(y_t - \rho y_{t-1})$ y las variables independientes $(x_{1t} - \rho x_{1,t-1})$, $(x_{2t} - \rho x_{2,t-1})$, ..., $(x_{kt} - x_{k,t-1})$. Los parámetros de este modelo son exactamente los mismos que los del modelo original, salvo que el término constante es $\beta_0(1 - \rho)$ en lugar de β_0. Más importante es el hecho de que en este modelo los errores no están autocorrelacionados y, por lo tanto, puede utilizarse el método de regresión múltiple por mínimos cuadrados para estimar los coeficientes del modelo. Los métodos inferenciales por mínimos cuadrados para hallar intervalos de confianza y realizar contrastes de hipótesis son adecuados para este modelo transformado.

Basándonos en este análisis, vemos que el problema de los errores autocorrelacionados puede evitarse estimando la regresión por mínimos cuadrados utilizando la variable dependiente $(y_t - \rho y_{t-1})$ y las variables dependientes $(x_{1t} - \rho x_{1,t-1})$, $(x_{2t} - \rho x_{2,t-1})$, ..., $(x_{kt} - \rho x_{k,t-1})$. Desgraciadamente, este enfoque plantea un problema en la práctica porque no conocemos el valor de ρ. En diferentes programas informáticos se utilizan distintos métodos para estimar ρ. Aquí, mostramos un sencillo método en el que utilizamos

$$r = 1 - \frac{d}{2}$$

para estimar ρ.

Estimación de modelos de regresión con errores autocorrelacionados

Supongamos que queremos estimar los coeficientes del modelo de regresión

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

cuando el término de error ε_t está autocorrelacionado.

Podemos estimarlos en dos etapas de la forma siguiente:

1. Estimamos el modelo por mínimos cuadrados, obteniendo el estadístico de Durbin-Watson y, por lo tanto, la estimación

$$r = 1 - \frac{d}{2} \qquad \textbf{(14.4)}$$

del parámetro de autocorrelación.

2. Estimamos por mínimos cuadrados una segunda regresión en la que la variable dependiente es $(y_t - ry_{t-1})$ y las variables independientes son $(x_{1t} - rx_{1,t-1})$, $(x_{2t} - rx_{2,t-1})$, ..., $(x_{kt} - rx_{k,t-1})$.

Los parámetros β_1, β_2, ..., β_k son los coeficientes de regresión estimados en este segundo modelo. Se obtiene una estimación de β_0 dividiendo la constante estimada en el segundo modelo por $(1 - r)$. Los contrastes de hipótesis y los intervalos de confianza de los coeficientes de regresión pueden realizarse utilizando los resultados de la segunda regresión.

EJEMPLO 14.6. Modelo de regresión de series temporales (análisis de regresión con errores correlacionados)

En este ejemplo extenso, mostramos cómo se realiza un análisis de regresión, utilizando el programa Minitab, cuando los errores están autocorrelacionados. En este ejemplo, queremos desarrollar un modelo que prediga el consumo agregado de bienes duraderos en función de la renta disponible y del tipo de interés de los fondos federales.

Solución

Macro2003

Los datos de este proyecto se encuentran en un fichero llamado **Macro2003**. Las variables de este fichero se describen en el apéndice del capítulo. Utilizamos las variables

CDH	Gastos personales de consumo: bienes duraderos (dólares reales de 1996)
YPDH	Renta personal disponible (dólares reales de 1996)
FFED	Tipo efectivo de los fondos federales

El fichero de datos contiene datos trimestrales desde el primer trimestre de 1946 hasta el segundo de 2003, pero queremos estimar el modelo utilizando datos del periodo comprendido entre el primer trimestre de 1980 y el segundo de 2003. Por lo tanto, nuestra primera tarea es obtener un subconjunto de estos datos utilizando el programa Minitab.

A continuación, hacemos la regresión múltiple y mostramos la salida en la Figura 14.17.

Regression Analysis: CDH versus YPDH, FFED

```
The regression equation is
CDH = -654 + 0.224 YPDH + 6.71 FFED

Predictor        Coef      SE Coef        T        P
Constant      -653.52        46.47   -14.06    0.000
YPDH         0.224220     0.006785    33.05    0.000
FFED            6.709        1.893     3.54    0.001

S = 41.4305    R-Sq = 96.3%   R-Sq(adj) = 96.2%

Analysis of Variance

Source            DF          SS          MS         F         P
Regression         2     4139436     2069718   1205.79     0.000
Residual Error    92      157917        1716
Total             94     4297352


Durbin-Watson statistic = 0.284994
```

Figura 14.17. Regresión múltiple para predecir el consumo de bienes duraderos: datos originales (salida Minitab).

El estadístico de Durbin-Watson de este modelo es 0,28, lo que indica que existe una autocorrelación positiva. Por lo tanto, es necesario utilizar transformaciones para obtener variables apropiadas para realizar la regresión. Se calcula un valor estimado de la correlación serial, r, utilizando la relación de la ecuación 14.4:

$$r = 1 - \frac{d}{2} = 1 - \frac{0{,}28}{2} = 0{,}86$$

A continuación, se calculan las variables transformadas en el programa Minitab utilizando el valor estimado $r = 0{,}86$. Como la transformación utiliza un valor retardado de cada variable, perdemos la primera observación del conjunto de datos. Ésa es la razón por la que incluimos el cuarto trimestre de 1979 en el conjunto de datos seleccionados. La Figura 14.18 presenta el modelo de regresión preparado utilizando las variables modificadas.

```
Regression Analysis: cdhadj versus ypdhadj, FFEDadj

The regression equation is
cdhadj = -68.2 + 0.201 ypdhadj - 1.78 FFEDadj

94 cases used, 1 cases contain missing values

Predictor        Coef      SE Coef        T        P
Constant       -68.21        11.13    -6.13    0.000
Ypdhadj       0.20060      0.01318    15.22    0.000
FFEDadj        -1.777        1.886    -0.94    0.349

S = 19.5675    R-Sq = 74.3%   R-Sq(adj) = 73.7%


Analysis of Variance

Source            DF          SS          MS        F          P
Regression         2      100696       50348   131.50      0.000
Residual Error    91       34843         383
Total             93      135538


Durbin-Watson statistic = 2.38972
```

Figura 14.18. Regresión múltiple para predecir el consumo de bienes duraderos: variables transformadas sin autocorrelación (salida Minitab).

La comparación de las salidas de las Figuras 14.17 y 14.18 indica claramente los problemas que plantean los modelos de regresión que tienen errores autocorrelacionados. El primer análisis de regresión es

$$\underset{}{\text{CDH} = -654 + \underset{(0{,}006785)}{0{,}224}\ \text{YPDH} + \underset{(1{,}893)}{6{,}71}\ \text{FFED}}$$

$$R^2 = 0{,}963 \qquad d = 0{,}28$$

Obsérvese que los números que figuran debajo de los coeficientes son los errores estadísticos de los coeficientes.

La primera regresión tiene un estadístico d de Durbin-Watson de 0,28, lo que indica que existe una fuerte autocorrelación positiva. Basándonos en los estadísticos de los coeficientes estimados concluimos que tanto la renta disponible ($b_1 = 0{,}224$) como el tipo de interés de los fondos federales ($b_2 = 6{,}71$) son predictores estadísticamente significativos de los gastos de consumo en bienes duraderos.

Sin embargo, el segundo análisis de regresión —basado en datos del modelo sin errores autocorrelacionados— lleva a una conclusión diferente:

$$\underset{}{\text{CDHadj} = -68{,}2 + \underset{(0{,}01318)}{0{,}201}\ \text{YPDHadj} - \underset{(1{,}886)}{1{,}78}\ \text{FFEDadj}}$$

$$R^2 = 0{,}743 \qquad d = 2{,}39$$

Obsérvese que los nombres de las variables se han modificado para reflejar el hecho de que se han transformado en variables que producirán un modelo que no tendrá autocorrelación. Obsérvese también que el estadístico d de Durbin-Watson es 2,39, lo que indica que no existe autocorrelación. Vemos que el coeficiente estimado de la renta disponible, $b_1 = 0{,}201$, es similar al de la primera regresión y que el error típico del coeficiente es 0,01318. El estadístico t de Student resultante, 15,22, nos lleva a concluir que la renta disponible es un predictor importante del consumo de bienes duraderos. En cambio, el coeficiente del tipo de interés de los fondos federales es $b_2 = -1{,}78$ con un estadístico t de Student de $-0{,}94$. Por lo tanto, no podemos rechazar la hipótesis nula de que el coeficiente del tipo de los fondos federales es 0 y de que debemos eliminar esa variable como predictor en el modelo de regresión.

En este ejemplo, hemos visto que la autocorrelación lleva a extraer una conclusión incorrecta sobre la importancia del tipo de interés de los fondos federales. Sin ajustar los datos para eliminar la correlación, habríamos utilizado el estadístico t de Student del modelo con los datos originales y ese estadístico t de Student de la regresión sin ajustar sobreestima el estadístico t de Student de la regresión ajustada. El estadístico t de Student del coeficiente de la renta disponible de la primera regresión también está sobreestimado. Sin embargo, tras realizar los ajustes pertinentes para obtener el estimador correcto, observamos que el coeficiente sigue siendo considerablemente diferente de 0.

Algunos paquetes estadísticos como Eviews3 y SAS, que están pensados para trabajar con datos de series temporales, tienen rutinas que estiman automáticamente el coeficiente de autocorrelación y realizan los ajustes necesarios para tener en cuenta la autocorrelación. Muchas de estas rutinas tienen rutinas de cálculo iterativas, por lo que generan estimaciones de los coeficientes y de las varianzas del modelo mejores que con la rutina mostrada aquí. Así pues, si el lector tiene acceso a un programa de ese tipo, le resultará más fácil la estimación que con el Minitab o el Excel. En general, esos otros programas informáticos obtienen estimaciones más eficientes de los coeficientes.

Errores autocorrelacionados en los modelos con variables dependientes retardadas

Cuando tenemos un modelo de regresión con variables dependientes retardadas en el segundo miembro y también tenemos errores autocorrelacionados, los métodos habituales de mínimos cuadrados pueden plantear problemas incluso más graves. Además de los problemas habituales que plantea la estimación de los errores de los coeficientes, también sabemos que los estimadores de los coeficientes están sesgados y no son consistentes, debido a que existe una correlación entre el error del modelo y una variable de predicción y eso introduce un sesgo en la estimación de los coeficientes. Desgraciadamente, en esta situación en que hay variables dependientes retardadas, los métodos antes analizados para detectar la presencia de errores autocorrelacionados no son válidos, por lo que presentaremos brevemente un método adecuado.

Consideremos el modelo

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \gamma y_{t-1} + \varepsilon_t$$

Supongamos que se ajusta este modelo a n conjuntos de observaciones muestrales por mínimos cuadrados. Sea d el estadístico de Durbin-Watson habitual con

$$r = 1 - \frac{d}{2}$$

y sea s_c la desviación típica estimada del coeficiente estimado γ de la variable dependiente retardada. Nuestra hipótesis nula es que el parámetro autorregresivo ρ es 0. Un contraste de esta hipótesis, aproximadamente válido en las grandes muestras, se basa en el estadístico h de Durbin:

$$h = r\sqrt{n/(1 - ns_c^2)}$$

En la hipótesis nula, este estadístico tiene una distribución de la que la distribución normal estándar es una buena aproximación cuando las muestras son grandes. Así, por ejemplo, se rechaza la hipótesis nula de que no existe autocorrelación frente a la hipótesis alternativa de que ρ es positivo al nivel de significación del 5 por ciento si el estadístico h es superior a 1,645.

Si el error autorregresivo es

$$u_t = \varepsilon_t - \rho\varepsilon_{t-1}$$

entonces, utilizando una modificación del método antes desarrollado para el ajuste para tener en cuenta la autocorrelación, podemos desarrollar el siguiente modelo:

$$\begin{aligned} y_t = \rho y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_{1t} - \rho x_{1,t-1}) + \beta_2(x_{2t} - \rho x_{2,t-1}) + \cdots \\ + \beta_k(x_{kt} - \rho x_{k,t-1}) + \gamma(y_{t-1} - \rho y_{t-2}) + \delta_t \end{aligned}$$

Uno de los enfoques posibles para estimar los parámetros, que sólo requiere un programa ordinario de estimación por mínimos cuadrados, es introducir, a su vez, en la ecuación anterior los valores posibles de ρ, por ejemplo, 0,1, 0,3, 0,5, 0,7 y 0,9. En ese caso, la regresión de la variable dependiente $(y_t - \rho y_{t-1})$ y las variables independientes $(x_{1t} - \rho x_{1,t-1})$, $(x_{2t} - \rho x_{2,t-1})$, ..., $(x_{kt} - \rho x_{k,t-1})$, $(y_{t-1} - \rho y_{t-2})$ se ajusta por mínimos cuadrados para cada valor posible de ρ. El valor de ρ elegido es aquel con el que la suma resultante de los cuadrados de los errores es menor. La inferencia sobre β_j se basa entonces en la regresión ajustada correspondiente.

EJERCICIOS

Ejercicios básicos

14.29. Suponga que se realiza una regresión con tres variables independientes y 30 observaciones. El estadístico de Durbin-Watson es 0,50. Contraste la hipótesis de que no hay autocorrelación. Calcule una estimación del coeficiente de autocorrelación si los datos indican que hay autocorrelación.

a) Repita con un estadístico Durbin-Watson igual a 0,80.

b) Repita con un estadístico Durbin-Watson igual a 1,10.

c) Repita con un estadístico Durbin-Watson igual a 1,25.

d) Repita con un estadístico Durbin-Watson igual a 1,70.

14.30. Suponga que se realiza una regresión con tres variables independientes y 28 observaciones. El estadístico de Durbin-Watson es 0,50. Contraste la hipótesis de que no hay autocorrela-

ción. Calcule una estimación del coeficiente de autocorrelación si los datos indican que hay autocorrelación.

a) Repita con un estadístico Durbin-Watson igual a 0,80.
b) Repita con un estadístico Durbin-Watson igual a 1,10.
c) Repita con un estadístico Durbin-Watson igual a 1,25.
d) Repita con un estadístico Durbin-Watson igual a 1,70.

Ejercicios aplicados

14.31. En una regresión basada en 30 observaciones anuales, se relacionó la renta agrícola de Estados Unidos con cuatro variables independientes: las exportaciones de cereales, las subvenciones federales, la población y una variable ficticia de los años de mal tiempo. El modelo se ajustó por mínimos cuadrados, lo que dio como resultado un estadístico de Durbin-Watson de 1,29. La regresión de e_i^2 con respecto a $\hat{y}_i$ dio un coeficiente de determinación de 0,043.

a) Realice un contraste de la heteroscedasticidad.
b) Realice un contraste de la existencia de errores autocorrelacionados.

14.32. Considere el modelo de regresión

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + \varepsilon_t$$

Demuestre que si

$$\text{Var}(\varepsilon) = Kx_i^2 \qquad (K > 0)$$

entonces

$$\text{Var}\left(\frac{\varepsilon_i}{x_i}\right) = K$$

Analice la posible relevancia de este resultado en el tratamiento de un tipo de heterocedasticidad.

14.33. Vuelva al ejercicio 14.13. Sea e_i los residuos de la regresión ajustada e $\hat{y}_t$ los valores predichos dentro del rango de la muestra. La regresión por mínimos cuadrados de e_i^2 con respecto a $\hat{y}_i$ tiene un coeficiente de determinación de 0,087. ¿Qué conclusión puede extraer de este resultado?

14.34. Vuelva al ejercicio 14.13 sobre la oferta monetaria del Reino Unido. ¿Qué conclusión puede extraer del estadístico de Durbin-Watson de la regresión ajustada? (Fichero de datos, **Money UK**).

14.35. Vuelva al ejercicio 14.18 sobre el consumo en Tailandia. Contraste la hipótesis nula de que no existen errores autocorrelacionados frente a la alternativa de que existe una autocorrelación positiva (fichero de datos, **Thailand Consumption**).

14.36. Un empresario creía que sus costes de producción unitarios (y) dependían del salario (x_1), de los costes de otros factores (x_2), de los costes generales (x_3) y de los gastos publicitarios (x_4). Se obtuvo una serie de 24 observaciones mensuales y se realizó una estimación por mínimos cuadrados del modelo que dio los siguientes resultados:

$$y_t = \underset{}{0{,}75} + \underset{(0{,}07)}{0{,}24}x_{1t} + \underset{(0{,}12)}{0{,}56}x_{2t} - \underset{(0{,}23)}{0{,}32}x_{3t} + \underset{(0{,}05)}{0{,}23}x_{4t}$$

$$R^2 = 0{,}79 \qquad d = 0{,}85$$

Las cifras entre paréntesis situadas debajo de los coeficientes estimados son sus errores típicos estimados. ¿Qué conclusiones puede extraer de estos resultados?

14.37. El fichero de datos **Advertising Retail** muestra 22 años consecutivos de datos sobre las ventas (y) y la publicidad (x) de una empresa de bienes de consumo.

a) Estime la regresión

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

b) Averigüe si hay errores autocorrelacionados en este modelo.
c) Si es necesario, estime de nuevo el modelo, teniendo en cuenta la posible existencia de errores autocorrelacionados.

14.38. La omisión de una variable independiente importante en un modelo de regresión de series temporales puede provocar la aparición de errores autocorrelacionados. En el ejemplo 14.5, hemos estimado el modelo

$$y_t = \beta_0 + \beta_1 x_{1t} + \varepsilon_t$$

que relaciona el margen de beneficios con los ingresos netos basándose en nuestros datos de las asociaciones de ahorro y crédito inmobiliario. Realice un contraste de Durbin-Watson de los residuos de este modelo. ¿Qué puede inferir de los resultados?

14.39. Vuelva al ejercicio 14.11 sobre el dinero que gastan los estudiantes en ropa. El estadístico de Durbin-Watson del modelo de regresión ajustado es 1,82. Contraste la hipótesis nula de que no hay errores autocorrelacionados frente a la alternativa de que hay una autocorrelación positiva.

RESUMEN

En este capítulo hemos mostrado que la construcción de modelos de regresión consiste en algo más que en los métodos básicos presentados en los Capítulos 12 y 13. En la práctica, la construcción de un buen modelo tiene mucho de arte y exige hacer un detenido análisis. En particular, no deben dejarse de lado importantes variables explicativas. Algunos problemas exigen la utilización de variables ficticias o de variables independientes retardadas. Recuérdese que en el Capítulo 13 mostramos que también pueden utilizarse modelos transformados que incluyan formas cuadráticas y formas logarítmico-lineales.

Como hemos visto, debemos comprobar también, en la medida de lo posible, cualquier supuesto postulado sobre la conducta de los términos de error. Pueden realizarse contrastes de heterocedasticidad y errores autocorrelacionados si se sospecha que existe alguno de los dos problemas. Y si existen, es necesario estimar de nuevo el modelo utilizando métodos adecuados desarrollados en este capítulo y en textos avanzados.

Aquí hemos analizado algunas de las circunstancias posibles en las que es deseable desviarse del análisis de regresión tradicional. Hay otros muchos métodos que se explican en los libros de texto de econometría. Si el lector tiene alguna incertidumbre sobre los supuestos de un método concreto, debe consultar un libro de texto avanzado o a un económetra familiarizado con esos métodos avanzados.

TÉRMINOS CLAVE

contraste de Durbin-Watson, 610
contraste de la presencia de heterocedasticidad, 605
diseño experimental, 584
errores autocorrelacionados, 608
errores autocorrelacionados con variables dependientes retardadas, 616
especificación del modelo, 577
estimación de coeficientes, 577
estimación de modelos de regresión con errores autocorrelacionados, 613
heterocedasticidad, 603
interpretación del modelo e inferencia, 578
multicolinealidad, 600
regresiones que contienen variables dependientes retardadas, 591
sesgo provocado por la exclusión de variables de predicción importantes, 596
variables ficticias, 579
verificación del modelo, 578

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

14.40. Escriba breves informes con ejemplos explicando cómo se utilizan en la especificación de los modelos de regresión de:

a) Las variables ficticias
b) Las variables dependientes retardadas
c) La transformación logarítmica

14.41. Considere el ajuste del modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

donde

Y = ingresos fiscales en porcentaje del producto nacional bruto de un país
X_1 = exportaciones en porcentaje del producto nacional bruto del país
X_2 = renta per cápita del país
X_3 = variable ficticia que toma el valor 1 si el país participa en algún tipo de integración económica y 0 en caso contrario.

Ésta es una forma de tener en cuenta los efectos que produce en los ingresos fiscales la participación en algún tipo de integración económica. Otra posibilidad sería estimar la regresión

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

por separado para los países que participan y no participan en algún tipo de integración económica. Explique en qué se diferencian estos enfoques del problema.

14.42. Analice la siguiente afirmación: «En muchos problemas prácticos de regresión, la multicolinealidad es tan grave que sería mejor realizar regresiones lineales simples independientes de la variable dependiente con respecto a cada variable independiente».

14.43. Explique la naturaleza de cada uno de los siguientes problemas y las dificultades que plantean:

a) La heterocedasticidad
b) Los errores autocorrelacionados

14.44. Se ha ajustado el siguiente modelo a los datos de 90 empresas químicas alemanas:

$$\hat{y} = \underset{}{0{,}819} + \underset{(1{,}79)}{2{,}11}x_1 + \underset{(1{,}94)}{0{,}96}x_2 - \underset{(0{,}144)}{0{,}059}x_3 + \underset{(4{,}08)}{5{,}87}x_4 + \underset{(0{,}00115)}{0{,}00226}x_5 \qquad R^2 = 0{,}410$$

donde los números entre paréntesis son los errores típicos de los coeficientes estimados y

y = precio de la acción
x_1 = beneficios por acción
x_2 = flujo de fondos por acción
x_3 = dividendos por acción
x_4 = valor contable por acción
x_5 = medida del crecimiento

a) Contraste al nivel del 10 por ciento la hipótesis nula de que el coeficiente de x_1 es 0 en la regresión poblacional frente a la hipótesis alternativa de que el verdadero coeficiente es positivo.
b) Contraste al nivel del 10 por ciento la hipótesis nula de que el coeficiente de x_2 es 0 en la regresión poblacional frente a la hipótesis alternativa de que el verdadero coeficiente es positivo.
c) La variable X_2 se ha eliminado del modelo original y se ha estimado la regresión de Y con respecto a (X_1, X_3, X_4, X_5). El coeficiente estimado de X_1 es 2,95 con un error típico de 0,63. ¿Cómo puede conciliarse este resultado con la conclusión del apartado (a)?

14.45. Se ha ajustado el siguiente modelo a los datos de 28 países correspondientes a 1989 para explicar el valor de mercado de su deuda en ese momento:

$$y = 77{,}2 - \underset{(8{,}0)}{9{,}6}x_1 - \underset{(2{,}73)}{17{,}2}x_2 - \underset{(0{,}056)}{0{,}15}x_3 + \underset{(1{,}0)}{2{,}2}x_4 \qquad R^2 = 0{,}84$$

donde

y = precio en el mercado secundario, en dólares, en 1989 de 100 $ de deuda del país
x_1 = 1 si los reguladores bancarios de Estados Unidos han obligado a los bancos de Estados Unidos a amortizar los activos que tienen del país, 0 en caso contrario
x_2 = 1 si el país suspendió el pago de los intereses de la deuda en 1989, 2 si suspendió el pago de los intereses de la deuda antes de 1989 y aún sigue suspendido y 0 en caso contrario
x_3 = cociente entre la deuda y el producto nacional bruto
x_4 = tasa de crecimiento del producto nacional bruto real, 1980-1985

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes.

a) Interprete el coeficiente estimado de x_1.
b) Contraste la hipótesis nula de que, manteniéndose todo lo demás constante, el cociente entre la deuda y el producto nacional bruto no influye linealmente en el valor de mercado de la deuda de un país frente a la alternativa de que cuanto más alto es este cociente, menor es el valor de la deuda.
c) Interprete el coeficiente de determinación.
d) La especificación de la variable ficticia x_2 no es ortodoxa. Una alternativa sería sustituir x_2 por el par de variables (x_5, x_6):

x_5 = 1 si el país suspendió el pago de los intereses de la deuda en 1989, 0 en caso contrario
x_6 = 1 si el país suspendió el pago de los intereses de la deuda antes de 1989 y aún sigue suspendido, 0 en caso contrario

Compare las implicaciones de estas dos especificaciones alternativas.

14.46. Se ha intentado construir un modelo de regresión que explique las calificaciones obtenidas por los estudiantes en los cursos de economía intermedia (véase la referencia bibliográfica 6). El modelo de regresión poblacional suponía que

Y = calificación total de los estudiantes en los cursos de economía intermedia
X_1 = calificación en matemáticas en el examen normalizado SAT
X_2 = calificación en lengua en el examen normalizado SAT
X_3 = calificación obtenida en álgebra en la universidad (A = 4, B = 3, C = 2, D = 1)
X_4 = calificación obtenida en la asignatura de principios de economía de la universidad
X_5 = variable ficticia que toma el valor 1 si el estudiante es mujer y 0 si es hombre
X_6 = variable ficticia que toma el valor 1 si el profesor es hombre y 0 si es mujer
X_7 = variable ficticia que toma el valor 1 si el estudiante y el profesor son del mismo sexo y 0 en caso contrario

Este modelo se ajustó con datos de 262 estudiantes. A continuación, indicamos los estadísticos t_i que son el cociente entre la estimación de

β_i y su error típico estimado correspondiente. Estos cocientes son

$$t_1 = 4{,}69 \quad t_2 = 2{,}89 \quad t_3 = 0{,}46 \quad t_4 = 4{,}90$$
$$t_5 = 0{,}13 \quad t_6 = -1{,}08 \quad t_7 = 0{,}88$$

El objetivo de este estudio era evaluar la influencia del sexo del estudiante y del profesor en el rendimiento. Realice un breve informe esbozando la información que ha obtenido sobre esta cuestión.

14.47. Se ha ajustado la siguiente regresión por mínimos cuadrados a 32 observaciones anuales sobre datos de series temporales:

$$\log y_t = \underset{}{4{,}52} - \underset{(0{,}28)}{0{,}62}\log x_{1t} + \underset{(0{,}38)}{0{,}92}\log x_{2t} + \underset{(0{,}21)}{0{,}61}\log x_{3t} + \underset{(0{,}12)}{0{,}16}\log x_{4t}$$
$$\bar{R}^2 = 0{,}683 \qquad d = 0{,}61$$

donde

y_t = cantidad de trigo exportada por Estados Unidos
x_{1t} = precio del trigo de Estados Unidos en el mercado mundial
x_{2t} = cantidad cultivada de trigo en Estados Unidos
x_{3t} = medida de la renta en los países que importan trigo de Estados Unidos
x_{4t} = precio de la cebada en el mercado mundial

Los números situados debajo de los coeficientes son los errores típicos de los coeficientes.

a) Interprete el coeficiente estimado de $\log x_{1t}$ en el contexto del modelo supuesto.
b) Contraste al nivel del 5 por ciento la hipótesis nula de que, manteniéndose todo lo demás constante, la renta de los países que importan trigo no influye en las exportaciones de trigo de Estados Unidos frente a la hipótesis alternativa de que un aumento de la renta eleva las exportaciones esperadas (no tenga en cuenta de momento el estadístico d de Durbin-Watson).
c) ¿Qué hipótesis nula puede contrastarse por medio del estadístico d? Realice este contraste en el presente problema, utilizando un nivel de significación del 1 por ciento.
d) Dados los resultados obtenidos en el apartado (c), comente sus conclusiones del apartado (b). ¿Cómo contrastaría la hipótesis nula del apartado (b)?

14.48. Se ha ajustado la siguiente regresión por mínimos cuadrados a 30 observaciones anuales sobre datos de series temporales:

$$\log y_t = \underset{}{4{,}31} + \underset{(0{,}17)}{0{,}27}\log x_{1t} + \underset{(0{,}21)}{0{,}53}\log x_{2t} - \underset{(0{,}30)}{0{,}82}\log x_{3t}$$
$$\bar{R}^2 = 0{,}615 \qquad d = 0{,}49$$

donde

y_t = número de quiebras de empresas
x_{1t} = tasa de desempleo
x_{2t} = tipo de interés a corto plazo
x_{3t} = valor de los nuevos pedidos realizados

Los números situados debajo de los coeficientes son los errores típicos de los coeficientes.

a) Interprete el coeficiente estimado de $\log x_{3t}$ en el contexto del modelo supuesto.
b) ¿Qué hipótesis nula puede contrastarse por medio del estadístico d? Realice este contraste en el presente problema utilizando un nivel de significación del 1 por ciento.
c) Dados los resultados del apartado (a), ¿es posible contrastar con la información dada la hipótesis nula de que, manteniéndose todo lo demás constante, los tipos de interés a corto plazo no influyen en las quiebras de empresas?
d) Estime la correlación entre los términos de error adyacentes en el modelo de regresión.

14.49. Un corredor de bolsa tiene interés en saber cuáles son los factores que influyen en la tasa de rendimiento de las acciones ordinarias de los bancos. Se ha estimado por mínimos cuadrados la siguiente regresión con una muestra de 30 bancos:

$$y = \underset{}{2{,}37} + \underset{(0{,}39)}{0{,}84}x_1 + \underset{(0{,}12)}{0{,}15}x_2 - \underset{(0{,}09)}{0{,}13}x_3 + \underset{(1{,}97)}{1{,}67}x_4$$
$$R^2 = 0{,}317$$

donde

y = tasa porcentual de rendimiento de las acciones ordinarias del banco
x_1 = tasa porcentual de crecimiento de los beneficios del banco
x_2 = tasa porcentual de crecimiento de los activos del banco
x_3 = pérdidas por préstamos en porcentaje de los activos del banco
x_4 = 1 si la central del banco está en Nueva York y 0 en caso contrario

Los números situados debajo de los coeficientes son los errores típicos de los coeficientes.

a) Interprete el coeficiente estimado de x_4.
b) Interprete el coeficiente de determinación y utilícelo para contrastar la hipótesis nula de

que las cuatro variables independientes, consideradas en conjunto, no influyen linealmente en la variable dependiente.

c) Sea e_i los residuos de la regresión ajustada e $\hat{y}^1$ los valores predichos de la variable dependiente dentro del rango de la muestra. La regresión de mínimos cuadrados de e_i^2 con respecto a $\hat{y}^1$ generó un coeficiente de determinación de 0,082. ¿Qué conclusiones pueden extraerse de este resultado?

14.50. Un analista de mercado está interesado en saber cuál es la cantidad media de dinero que gastan anualmente los estudiantes en ocio. Se ha estimado por mínimos cuadrados la siguiente regresión con datos anuales de 30 años:

$$y_t = \underset{}{40{,}93} + \underset{(0{,}106)}{0{,}253}x_t + \underset{(0{,}134)}{0{,}546}y_{t-1} \qquad d = 1{,}86$$

donde

y_t = gasto por estudiante, en dólares, en ocio
x_t = renta disponible por estudiante, en dólares, una vez pagada la matrícula, las tasas y la manutención

Los números situados debajo de los coeficientes son los errores típicos de los coeficientes.

a) Halle el intervalo de confianza al 95 por ciento del coeficiente de x_t en la regresión poblacional.

b) ¿Qué efecto es de esperar que produzca a lo largo del tiempo un aumento de la renta disponible por estudiante de 1 $ en el gasto en ocio?

c) Contraste la hipótesis nula de que no existe ninguna autocorrelación en los errores frente a la hipótesis alternativa de que existe una autocorrelación positiva.

14.51. A una empresa local de servicios públicos le gustaría ser capaz de predecir la factura mensual media en electricidad de una vivienda. El estadístico de la empresa ha estimado por mínimos cuadrados el siguiente modelo de regresión:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$$

donde

y = factura mensual media en electricidad, en dólares
x_1 = factura bimestral media en gasolina para automóviles
x_2 = número de habitaciones de la vivienda

El estadístico obtuvo la siguiente salida SAS basándose en una muestra de 25 viviendas:

PARAMETER	ESTIMATE	STUDENT'S *t* FOR HO: PARAMETER = 0	STD. ERROR OF ESTIMATE
INTERCEPT	−10.8030		
X1	−0.0247	−0.956	0.0259
X2	10.9409	18.517	0.5909

a) Interprete, en el contexto del problema, la estimación por mínimos cuadrados de β_2.

b) Contraste la hipótesis nula

$$H_0: \beta_1 = 0$$

frente a la hipótesis alternativa bilateral.

c) El estadístico está preocupado por la posibilidad de que exista multicolinealidad. ¿Qué información se necesita para evaluar la posible gravedad de este problema?

d) Se sugiere que la renta de los hogares es un importante determinante de la cuantía de la factura de electricidad. De ser eso cierto, ¿qué puede decirse sobre la regresión estimada por el estadístico?

e) Dado el modelo ajustado, el estadístico obtiene las facturas predichas de electricidad, $\hat{y}$, y los residuos, e. A continuación, hace una regresión de e^2 con respecto a $\hat{y}$, y observa que la regresión tiene un coeficiente de determinación de 0,0470. Interprete este resultado.

14.52. El fichero de datos **Indonesia Revenue** muestra 15 observaciones anuales de Indonesia sobre los ingresos fiscales totales, salvo los generados por el petróleo (y), la renta nacional (x_1) y el valor añadido por el petróleo en porcentaje del producto interior bruto (x_2). Estime por mínimos cuadrados la regresión

$$\log y_t = \beta_0 + \beta_1 \log x_{1t} + \beta_2 \log x_{2t} + \varepsilon_t$$

Realice un informe que resuma sus resultados, incluido un contraste de la existencia de heterocedasticidad y otro de la existencia de errores autocorrelacionados.

14.53. El fichero de datos **German Income** muestra 22 observaciones anuales de la República Federal de Alemania sobre la variación porcentual de los sueldos y salarios (y), el crecimiento de la productividad (x_1) y la tasa de inflación (x_2) medida por medio del deflactor del producto nacional bruto. Estime por mínimos cuadrados la regresión

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$$

Escriba un informe que resuma sus resultados, incluido un contraste de la existencia de heterocedasticidad y un contraste de la existencia de errores autocorrelacionados.

14.54. El fichero de datos **Japan Imports** muestra 35 observaciones trimestrales de Japón sobre la cantidad de importaciones (y), el cociente entre los precios de las importaciones y los precios interiores (x_1) y el producto nacional bruto real (x_2). Estime por mínimos cuadrados la regresión

$$\log y_t = \beta_0 + \beta_1 \log x_{1t} + \beta_2 \log x_{2t} + \gamma \log Y_{t-1} + \varepsilon_t$$

Realice un informe que resuma sus resultados, incluido un contraste de la existencia de errores autocorrelacionados.

14.55. Se ha realizado un estudio sobre los costes por hora de trabajo de las auditorías realizadas a los bancos por el banco central. Se han obtenido datos sobre 91 auditorías. Algunas han sido realizadas directamente por el banco central y en otras han intervenido auditores externos. Los auditores han calificado la dirección de los bancos de buena, satisfactoria, correcta o insatisfactoria. El modelo estimado es

$$\log y = \underset{}{2{,}41} + \underset{(0{,}0477)}{0{,}3674} \log x_1 + \underset{(0{,}0628)}{0{,}2217} \log x_2 + \underset{(0{,}0287)}{0{,}0803} \log x_3$$
$$- \underset{(0{,}2905)}{0{,}1755} x_4 + \underset{(0{,}1044)}{0{,}2799} x_5 + \underset{(0{,}1657)}{0{,}5634} x_6 - \underset{(0{,}0787)}{0{,}2572} x_7$$
$$R^2 = 0{,}766$$

donde

y = horas de trabajo de los auditores del banco central
x_1 = total de activos del banco
x_2 = número total de oficinas del banco
x_3 = cociente entre los préstamos clasificados como dudosos y los préstamos totales del banco
x_4 = 1 si la valoración de la dirección es «buena» y 0 en caso contrario
x_5 = 1 si la valoración de la dirección es «correcta» y 0 en caso contrario
x_6 = 1 si la valoración de la dirección es «insatisfactoria» y 0 en caso contrario
x_7 = 1 si la auditoría se realizó conjuntamente con auditores externos y 0 en caso contrario

Los números entre paréntesis situados debajo de los coeficientes son los errores típicos de los coeficientes.

14.56. El fichero de datos **Britain Sick Leave** muestra datos de Gran Bretaña sobre el número de días de baja por enfermedad por persona (Y), la tasa de desempleo (X_1), el cociente entre las prestaciones y los ingresos (X_2) y el salario real (X_3). Estime el modelo

$$\log y_t = \beta_0 + \beta_1 \log x_{1t} + \beta_2 \log x_{2t} + \beta_3 \log x_{3t} + \varepsilon_t$$

y realice un informe sobre sus resultados. Incluya en su análisis una comprobación de la posibilidad de que haya errores autocorrelacionados y, si es necesario, una corrección para resolver este problema.

14.57. El Departamento de Comercio de Estados Unidos le ha pedido que desarrolle un modelo de regresión para predecir la inversión trimestral en producción y equipo duradero. Las variables de predicción sugeridas son el PIB, el tipo de interés preferencial, el índice de precios de las mercancías industriales y el gasto público. Los datos de su análisis se encuentran en el fichero de datos **Macro2003**, que está almacenado en su disco de datos y se describe en el diccionario de datos del apéndice de este capítulo. Utilice datos del periodo de tiempo comprendido entre el primer trimestre de 1976 y el segundo de 2003.

a) Estime un modelo de regresión utilizando solamente el tipo de interés para predecir la inversión. Utilice el estadístico de Durbin-Watson para contrastar la existencia de autocorrelación.
b) Halle la mejor ecuación de regresión múltiple para predecir la inversión utilizando las variables de predicción indicadas anteriormente. Utilice el estadístico de Durbin-Watson para contrastar la existencia de autocorrelación.
c) ¿Qué diferencias hay entre los modelos de regresión de los apartados (a) y (b) desde el punto de vista de la bondad del ajuste, la capacidad de predicción, la autocorrelación y la contribución a comprender el problema de inversión?

14.58. Un economista le ha pedido que desarrolle un modelo de regresión para predecir el consumo de servicios en función del PNB y de otras variables importantes. Los datos para hacer el análisis se encuentran en el fichero de datos **Macro2003**, que están almacenados en su disco de datos y se describen en el apéndice del capítulo. Utilice datos del periodo comprendido entre el primer trimestre de 1003 y el cuarto de 2000.

a) Estime un modelo de regresión utilizando solamente el PIB para predecir el consumo de servicios. Contraste la existencia de autocorrelación utilizando el estadístico de Durbin-Watson.
b) Estime un modelo de regresión múltiple utilizando el PNB, el consumo total retardado 1 periodo y el tipo de interés preferencial como predictores adicionales. Contraste la existencia de autocorrelación. ¿Reduce esta

regresión múltiple el problema de la autocorrelación?

14.59. Jack Wong, inversor de Tokio, está considerando la posibilidad de establecer una planta de acero primario en Japón. Tras revisar la propuesta inicial, le preocupa la combinación propuesta de capital y trabajo. Le ha pedido que formule varias funciones de producción utilizando algunos datos históricos de Estados Unidos. El fichero de datos **Metals** contiene 27 observaciones de la producción, medida por el valor añadido, de la cantidad de trabajo y del valor bruto de la planta y equipo de cada fábrica.

a) Utilice una regresión múltiple para estimar una función de producción lineal haciendo una regresión del valor añadido con respecto al trabajo y el capital.

b) Represente gráficamente los residuos en relación con el trabajo y el equipo. Señale las pautas excepcionales que pueda haber.

c) Utilice una regresión múltiple con variables transformadas para estimar una función de producción Cobb-Douglas de la forma

$$y = \beta_0 L^{\beta_1} K^{\beta_2}$$

donde y es el valor añadido, L es la cantidad de trabajo y K es la cantidad de capital.

d) Utilice una regresión múltiple con variables transformadas para estimar una función de producción Cobb-Douglas con rendimientos constantes de escala. Observe que esta función de producción tiene la misma forma que la función estimada del apartado (c), pero tiene la restricción adicional de que $\beta_1 + \beta_2 = 1$. Para desarrollar el modelo de regresión transformado, exprese β_2 en función de β_1 y convierta la expresión a un formato de regresión.

e) Compare las tres funciones de producción utilizando gráficos de los residuos y un error típico de la estimación expresado en la misma escala. Tendrá que convertir los valores predichos de los apartados (c) y (d) (que están en logaritmos) en las unidades originales. A continuación, puede restar los valores predichos de los valores originales de Y para obtener los residuos. Utilice los residuos para calcular errores típicos comparables de la estimación.

14.60. Las autoridades de una pequeña ciudad le han pedido que identifique las variables que influyen en el valor medio de mercado de las viviendas de las ciudades pequeñas del Medio Oeste. El fichero de datos **Citydat** contiene datos de algunas pequeñas ciudades. Las variables de predicción candidatas son el tamaño medio de la vivienda (sizehse), el tipo del impuesto sobre bienes inmuebles (taxrate) (el impuesto dividido por el valor catastral total), los gastos totales en servicios municipales (totexp) y el porcentaje de locales comerciales (comper).

a) Estime el modelo de regresión múltiple utilizando todas las variables de predicción indicadas. Selecciones únicamente las variables estadísticamente significativas para formular su ecuación final.

b) Según un economista, como los datos proceden de ciudades que tienen diferente número de habitantes, es probable que su modelo contenga heterocedasticidad. Sostiene que los precios medios de las viviendas de las ciudades mayores tendrían una varianza menor, ya que el número de viviendas utilizadas para calcular los precios medios de la vivienda sería mayor. Realice un contraste de la existencia de heterocedasticidad.

c) Estime la ecuación de regresión múltiple utilizando mínimos cuadrados ponderados con la población como variable de ponderación. Compare los coeficientes de los modelos de regresión múltiple ponderado y no ponderado.

Apéndice

Diccionario de datos del fichero de datos Macro2003

El fichero de datos contiene datos trimestrales que van del primer trimestre de 1946 al segundo de 2003. Salvo que se indique lo contrario, los datos están expresados en dólares de 1996 utilizando el nuevo índice de precios encadenado. Algunas series no comienzan en 1946, lo cual se indica diciendo que tienen menos de 218 observaciones.

FM2	serie	M	Cantidad de dinero: M2 (desestacionalizada, mm $)
FFED	serie	M	Tipo [efectivo] de los fondos federales (% anual)
FBPR	serie	M	Tipo preferencial de los préstamos bancarios (% anual)
CDH	serie	Q	Gastos personales de consumo: bienes duraderos (TAD —tasa anual desestacionalizada—, mm $ de 1996 encadenados)
CNH	serie	Q	Gastos personales de consumo: bienes no duraderos (TAD, mm $ de 1996 encadenados)
CSH	serie	Q	Gastos personales de consumo: servicios (TAD, mm $ de 1996 encadenados)
CH	serie	Q	Gastos personales de consumo (TAD, mm $ de 1996 encadenados)
Chtot		Q	CDH + CNH + CSH
FNH	serie	Q	Inversión no residencial fija privada (TAD, mm $ de 1996 encadenados)
FRH	serie	Q	Inversión privada fija en viviendas (TAD, mm $ de 1996 encadenados)
VH	serie	Q	Variación de las existencias de las empresas (TAD, mm $ de 1996 encadenados)
IH	serie	Q	Inversión bruta interior privada (TAD, mm $ de 1996 encadenados)
IHTOT		Q	FNH + FRH + VH
XH	serie	Q	Exportaciones de bienes y servicios (TAD, mm $ de 1996 encadenados)
MH	serie	Q	Importaciones de bienes y servicios (TAD, mm $ de 1996 encadenados)
GH	serie	Q	Gasto público de consumo/inversión bruta (TAD, mm $ de 1996 encadenados)
GDPH	serie	Q	Producto interior bruto (TAD, mm $ de 1996 encadenados)
Gdphtot		Q	CHTOT + IHTOT + GH + XH − MH
JGDP	serie	Q	Producto interior bruto: índice de precios encadenado (desestacionalizado, 1996 = 100)
YP	serie	Q	Renta personal (TAD, mm $ de 1996)
YPD	serie	Q	Renta personal disponible (TAD, mm $ de 1996)
YPDH	serie	Q	Renta personal disponible (TAD, mm $ de 1996 encadenados)
YPSV	serie	Q	Ahorro personal (TAD, mm $ de 1996)
YPO	serie	Q	Gasto personal (TAD, mm $ de 1996)

Bibliografía

1. Dhalla, N. K., «Short-Term Forecasts of Advertising Expenditures», *Journal of Advertising Research*, 19, n.º 1, 1979, págs. 7-14.
2. Erikson, G. M., «Using Ridge Regression to Estimate Directly Lagged Effects in Marketing», *Journal of American Statistical Association*, 76, 1981, págs. 766-773.
3. Hsiao, C., «Autoregressive Modeling of Canadian Money and Income Data», *Journal of American Statistical Association*, 74, 1979, págs. 553-560.
4. McDonald, J., «Modeling Demographic Relationships: An Analysis of Forecast Functions for Australian Births», *Journal of the American Statistical Association*, 76, 1981, págs. 782-792.
5. Mills, T. C., «The Functional Form of the UK Demand for Money», *Applied Statistics*, 27, 1978, págs. 52-57.
6. Waldauer, C., V. G. Duggal y M. L. Williams, «Gender Differences in Economic Knowledge: A Further Extension of the Analysis», *Quarterly Review of Economics and Finance*, 32, n.º 4, 1992, págs. 138-143.

Capítulo 15

Estadística no paramétrica

Esquema del capítulo

Introducción

En el Capítulo 2 vimos que los datos se clasifican en numéricos y cualitativos. Los métodos estadísticos que hemos estudiado hasta ahora requieren el uso de datos numéricos. En el caso de esos datos, las medias, las varianzas y las desviaciones típicas tienen sentido. Sin embargo, en el de los datos cualitativos (nominales u ordinales), no pueden aplicarse los métodos paramétricos. En este capítulo introducimos contrastes *no paramétricos* que suelen ser el método necesario para extraer conclusiones inferenciales sobre datos nominales u ordinales. A menudo se obtienen datos de ese tipo en muchos contextos, como los estudios de mercado, las encuestas a empresas y los cuestionarios.

En los Capítulos 10 y 11 introdujimos algunos contrastes de hipótesis que dependían del supuesto de la normalidad de las distribuciones poblacionales. El supuesto de la normalidad a menudo es razonable. Además, en virtud del teorema del límite central, muchos de estos métodos de contraste siguen siendo más o menos válidos cuando las muestras son grandes aunque la distribución poblacional no sea normal. Si embargo, puede darse el caso de que en las aplicaciones prácticas sea insostenible el supuesto de la normalidad. En estas circunstancias, es deseable basar las inferencias en contrastes *no paramétricos* que son válidos en una amplia variedad de distribuciones de la población subyacente. Esos contrastes suelen denominarse contrastes *que no dependen de la distribución*.

En este capítulo describimos algunos de los contrastes no paramétricos que son adecuados para analizar datos nominales, datos ordinales o datos numéricos cuando no puede postularse el supuesto de la normalidad de la distribución de probabilidad de la población. En capítulos posteriores analizamos otros contrastes no paramétricos. No es nuestro objetivo aquí intentar describir toda la amplia variedad de métodos no paramétricos que existen. Nuestra aspiración es más modesta: que el lector se haga una idea de algunos métodos no paramétricos, entre los que se encuentran el contraste de signos, el contraste de Wilcoxon basado en la ordenación de las diferencias, el contraste *U* de Mann-Whitney, el contraste de la suma de puestos de Wilcoxon y el contraste de correlación de orden de Spearman. Éstas son alternativas no paramétricas a los distintos métodos introducidos antes en el libro.

15.1. Contraste de signos e intervalo de confianza

El contraste no paramétrico más sencillo de realizar es el **contraste de signos**. Se utiliza principalmente para contrastar hipótesis sobre la posición central (mediana) de una distribución poblacional o para analizar datos de muestras pareadas. El contraste de signos se emplea en los estudios de mercado para averiguar si los consumidores prefieren uno de dos productos. Dado que los encuestados manifiestan simplemente su preferencia, los datos son nominales y se prestan a métodos no paramétricos.

Contraste de signos de muestras pareadas o enlazadas

Supongamos que se toman muestras pareadas o enlazadas de una población y se descartan las diferencias iguales a 0, por lo que quedan n observaciones. El contraste de signos puede utilizarse para contrastar la hipótesis nula de que la mediana poblacional de las diferencias es 0 (lo que sería cierto, por ejemplo, si las diferencias procedieran de una población cuya distribución fuera simétrica en torno a una media de 0). Sea $+$ una diferencia positiva y $-$ una diferencia negativa. Si la hipótesis nula fuera verdadera, nuestra secuencia de diferencias $+$ y $-$ podría concebirse como una muestra aleatoria extraída de una población en la que las probabilidades de $+$ y $-$ fueran cada una de 0,5. En ese caso, las observaciones constituirían una muestra aleatoria extraída de una población binomial en la que la probabilidad de $+$ sería de 0,5. Por lo tanto, si P representa la verdadera proporción de $+$ que hay en la población (es decir, la verdadera proporción de diferencias positivas), la hipótesis nula es simplemente

$$H_0: P = 0{,}5$$

El contraste de signos se basa entonces en el hecho de que el número de observaciones positivas, S, que hay en la muestra sigue una distribución binominal (donde $P = 0{,}5$ según la hipótesis nula).

Contraste de signos de muestras pareadas

Supongamos que se toman muestras aleatorias pareadas o enlazadas de una población y que se descartan las diferencias iguales a 0, por lo que quedan n observaciones. Calculamos la diferencia para cada par de observaciones y anotamos el signo de esta diferencia. El contraste de signos se utiliza para contrastar

$$H_0: P = 0{,}5$$

donde P es la proporción de observaciones no nulas en la población que son positivas. El estadístico del contraste S para el contraste de signos de muestras pareadas es simplemente

$$S = \text{número de pares que tienen una diferencia positiva}$$

donde S sigue una distribución binomial, donde $P = 0{,}5$ y $n =$ número de diferencias no nulas.

Tras contrastar la hipótesis nula y la hipótesis alternativa y hallar un estadístico del contraste, el paso siguiente es calcular el p-valor y extraer conclusiones basadas en una regla de decisión.

Cálculo del *p*-valor de un contraste de signos

El ***p*-valor de un contraste de signos** se halla utilizando la distribución binomial con n = número de diferencias no nulas, S = número de diferencias positivas y $P = 0{,}5$.

a) En un contraste de la cola superior

$$H_1 : P > 0{,}5 \qquad p\text{-valor} = P(x \geqslant S) \tag{15.1}$$

b) En un contraste de la cola inferior

$$H_1 : P < 0{,}5 \qquad p\text{-valor} = P(x \leqslant S) \tag{15.2}$$

c) En un contraste de dos colas

$$H_1 : P \neq 0{,}5 \qquad 2(p\text{-valor}) \tag{15.3}$$

EJEMPLO 15.1. Preferencia por un producto (contraste de signos)

Un restaurante italiano cercano a un campus universitario está considerando la posibilidad de utilizar una nueva receta para hacer la salsa que echa a las pizzas. Se elige una muestra aleatoria de ocho estudiantes y se pide a cada uno que valore en una escala de 1 a 10 su opinión sobre la salsa original y sobre la salsa propuesta. La Tabla 15.1 muestra las valoraciones obtenidas en la comparación; los números más altos indican que gusta más el producto.

¿Indican los datos una tendencia general a preferir la nueva salsa a la original?

Solución

La Tabla 15.1 también muestra las diferencias de valoración de los estudiantes y los signos de estas diferencias. Así, se asigna un $+$ si se prefiere la salsa original, un $-$ si se prefiere la nueva y 0 si se valoran los dos productos por igual. En este experimento, dos estudiantes prefieren la salsa original y cinco la nueva; uno las valora por igual.

Tabla 15.1. Valoración de la salsa de pizza por parte de los estudiantes.

	Valoración			
Estudiante	Producto original	Producto nuevo	Diferencia (original-nuevo)	Signo de la diferencia
A	6	8	−2	−
B	4	9	−5	−
C	5	4	1	+
D	8	7	1	+
E	3	9	−6	−
F	6	9	−3	−
G	7	7	0	0
H	5	9	−4	−

La hipótesis nula de interés es que en la población en general no hay una tendencia general a preferir un producto al otro. Para evaluar esta hipótesis, comparamos los números que expresan una preferencia por cada producto, descartando los que valoran los

productos por igual. En este ejemplo, los valores del estudiante G se omiten y el tamaño efectivo de la muestra se reduce a $n = 7$. La única información muestral en la que se basa nuestro contraste es que dos de los siete estudiantes prefieren el producto original. Por lo tanto, el estadístico del contraste es $S = 2$.

La hipótesis nula puede concebirse como la hipótesis de que la mediana poblacional de las diferencias es 0. Si la hipótesis nula fuera verdadera, nuestra secuencia de diferencias + y − podría concebirse como una muestra aleatoria extraída de una población en la que las probabilidades de + y − son 0,5 cada una. En ese caso, las observaciones constituirían una muestra aleatoria extraída de una población binomial en la que la probabilidad de + es 0,5. Por lo tanto, si P representa la verdadera proporción de + que hay en la población (es decir, la verdadera proporción de la población que prefiere la salsa original), la hipótesis nula es simplemente

$H_0: P = 0{,}5$ *No hay una tendencia general a preferir uno de los productos al otro*

Se utiliza un contraste de una cola para averiguar si existe una tendencia general a preferir la nueva salsa a la original. La alternativa de interés es que la mayoría de la población prefiere el nuevo producto. Esta alternativa se expresa de la forma siguiente:

$H_1: P < 0{,}5$ *La mayoría prefiere el nuevo producto (o menos del 50% prefiere el producto original)*

A continuación, hallamos la probabilidad de observar en la muestra un resultado tan extremo o más que el que se obtendría si la hipótesis nula fuera, en realidad, verdadera. Este valor es el *p*-valor del contraste. Si representamos por medio de $P(x)$ la probabilidad de observar x «éxitos» (+) en $n = 7$ pruebas binomiales, cada una con una probabilidad de éxito de 0,5, entonces la probabilidad binomial acumulada de observar dos o menos + puede obtenerse utilizando la fórmula binomial, una tabla binomial o un programa informático como Microsoft Excel. El *p*-valor se halla por medio de la ecuación 15.2:

$$\begin{aligned} P\text{-valor} &= P(x \leqslant 2) = P(x = 0) + P(x = 1) + P(x = 2) \\ &= 0{,}0078 + 0{,}0547 + 0{,}1641 = 0{,}2266 \end{aligned}$$

Con un *p*-valor tan grande, no podemos rechazar la hipótesis nula y concluimos que los datos no son suficientes para sugerir que los estudiantes prefieren la nueva salsa. Asimismo, podríamos haber dicho que si adoptamos la regla de decisión «rechazar H_0 si hay dos o menos + en la muestra», entonces la probabilidad de que la hipótesis nula se rechace cuando en realidad es verdadera es 0,2266. Por lo tanto, ese contraste tiene un *p*-valor de 22,66 por ciento. Dado que el *p*-valor es el nivel de significación más bajo al que puede rechazarse la hipótesis nula, en este ejemplo la hipótesis nula puede rechazarse al 22,66 por ciento o más. Es improbable que alguien estuviera dispuesto a aceptar un nivel de significación tan alto. Una vez más, concluimos que los datos no son estadísticamente significativos para recomendar un cambio de salsa. Quizá nuestra decisión se debe a que tenemos un pequeño número de observaciones muestrales.

Para ilustrar un contraste de dos colas, supongamos que queremos averiguar si hay en la población una preferencia general por cualquiera de los dos productos. En ese caso, $H_1: P \neq 0{,}5$ y, por la ecuación 15.3, el *p*-valor $= 2P(x \leqslant 2) = 2(0{,}2266) = 0{,}4532$.

Obsérvese también que

$$p\text{-valor} = P(x \leqslant 2) + P(x \geqslant 5) = P(0) + P(1) + P(2) + P(5) + P(6) + P(7) = 0{,}4532$$

Ese elevado p-valor sugeriría que los datos no son suficientes para pensar que los estudiantes prefieren una de las salsas a la otra. Sólo podríamos rechazar la hipótesis nula y concluir que se prefiere una de las salsas con un nivel de significación del 45,32 por ciento.

Aproximación normal

Como consecuencia del teorema del límite central, puede utilizarse la distribución normal como aproximación de la distribución binomial si la el tamaño de la muestra es grande. Los expertos discrepan sobre la definición exacta de «grande». Sugerimos que la aproximación normal es aceptable si el tamaño de la muestra es de más de 20. Un factor de corrección de continuidad del estadístico del contraste compensa la estimación de datos discretos con una distribución continua y permite aproximarse más al p-valor.

El contraste de signos: aproximación normal (grandes muestras)

Si el número n de observaciones muestrales no nulas es grande, el contraste de signos se basa en la **aproximación normal** de la binomial de media y desviación típica:

Media: $\mu = nP = 0{,}5n$ Desviación típica: $\sigma = \sqrt{nP(1-P)} = \sqrt{0{,}25n} = 0{,}5\sqrt{n}$

El estadístico del contraste es

$$Z = \frac{S^* - \mu}{\sigma} = \frac{S^* - 0{,}5n}{0{,}5\sqrt{n}} \qquad \textbf{(15.4)}$$

donde S^* es el estadístico del contraste corregido para tener en cuenta la continuidad y se define de la forma siguiente:

a) En un contraste de dos colas

$$S^* = S + 0{,}5 \quad \text{si } S < \mu \quad \text{o} \quad S^* = S - 0{,}5 \quad \text{si } S > \mu \qquad \textbf{(15.5)}$$

b) En un contraste de la cola superior

$$S^* = S - 0{,}5 \qquad \textbf{(15.6)}$$

En un contraste de la cola inferior

$$S^* = S + 0{,}5 \qquad \textbf{(15.7)}$$

EJEMPLO 15.2. El helado (contraste de signos: aproximación normal)

Se ha pedido a una muestra aleatoria de 100 niños que comparen dos nuevos sabores de helado: mantequilla de cacahuete y chicle. Cincuenta y seis miembros de la muestra prefieren el helado de mantequilla de cacahuete, 40 el de chicle y 4 no manifiestan ninguna preferencia. Utilice la *aproximación normal* para averiguar si existe una preferencia general por cualquiera de los dos sabores. Compare su resultado con las probabilidades binomiales obtenidas utilizando tanto Excel como Minitab.

Solución

Para contrastar si existe en esta población una preferencia general por uno de los dos sabores, las hipótesis son

$H_0: P = 0{,}5$ *Los niños no tienen ninguna preferencia por ninguno de los dos sabores*
$H_1: P \neq 0{,}5$ *Los niños tienen preferencia por uno de los dos sabores*

Sea P la proporción de la población que prefiere el helado de chicle, por lo que $S = 40$ (P también podría haber sido la proporción de la población que prefiere el helado de mantequilla de cacahuete; en ese caso $S = 56$). Utilizando las ecuaciones 15.4 y 15.5,

$$\mu = nP = 0{,}5n = 0{,}5(96) = 48$$
$$\sigma = 0{,}5\sqrt{96} = 4{,}899$$
$$Z = \frac{S^* - \mu}{\sigma} = \frac{40{,}5 - 48}{4{,}899} = -1{,}53 \quad \text{dado que } 40 < 48,\ S^* = 40{,}5$$

De la distribución normal estándar se deduce que el p-valor aproximado $= 2(0{,}0630) = = 0{,}126$. Por lo tanto, puede rechazarse la hipótesis nula a todos los niveles de significación superiores a 12,6 por ciento. Si no se utiliza ningún factor de corrección de continuidad, el valor Z se convierte en $Z = -1{,}633$, lo que da un p-valor algo menor: 0,1024.

Minitab y Excel (contraste de signos) Dado que el contraste de signos se basa en la distribución de probabilidad binomial, el uso de Minitab o de Excel es sencillo. En la salida Minitab (Figura 15.1A) se observa que el p-valor $= 2(0{,}0626728) = 0{,}1254$ y en la salida Excel (Figura 15.1B) se observa que el p-valor $= P(x \leqslant 40) + P(x \geqslant 56) = = 0{,}0626728 + 0{,}0626728 = 0{,}1253456$. Los dos p-valores son cercanos al p-valor de 0,126 obtenido utilizando las ecuaciones 15.4 y 15.5. Los datos no son suficientes para sugerir que los niños tienen una preferencia general por uno de los sabores o por el otro.

x	P(x <= x)
40,0	0,0626728

Figura 15.1A. Ejemplo del helado: $n = 96$, $P = 0{,}5$, $S = 40$ (salida Minitab).

Figura 15.1B. Ejemplo del helado: $n = 96$, $P = 0{,}5$, $S = 40$ (salida Excel).

Contraste de signos de una mediana poblacional

El contraste de signos también puede utilizarse en el caso de una muestra para contrastar la hipótesis de que la mediana es un valor dado.

EJEMPLO 15.3. Ingresos iniciales de personas recién licenciadas (contraste de signos)

El decano de la facultad de administración de empresas de una universidad querría tener información sobre los ingresos iniciales de las personas recién licenciadas. Éstos son los sueldos iniciales de una muestra aleatoria de 23 licenciados:

29250	29900	28070	31400	31100	29000	33000	50000	28500	31000
34800	42100	33200	36000	65800	34000	29900	32000	31500	29900
32890	36000	35000							

Income

¿Indican los datos que la mediana de los ingresos iniciales es diferente de 35.000 $? Los datos para hacer este problema se encuentran en el fichero de datos **Income**.

Solución

Dado que la distribución de los ingresos a menudo está sesgada, se utilizará el contraste de signos. La hipótesis nula y la hipótesis alternativa son

$$H_0: \text{Mediana} = 35.000\ \$$$
$$H_1: \text{Mediana} \neq 35.000\ \$$$

Aquí contrastamos la hipótesis nula utilizando una distribución binomial en la que $P = 0{,}50$. Primero obtenemos una respuesta aproximada utilizando las ecuaciones 15.4 y 15.5. Obsérvese que hay 17 estudiantes que indicaron que tenían unos ingresos iniciales de más de 35.000 $, 5 que tenían unos ingresos iniciales de menos de 35.000 $ y 1 que tenía unos ingresos iniciales de 35.000 $. El tamaño de la muestra se reduce a $n = 22$ y $S = 17$. Se observa que la media y la desviación típica son

$$\mu = nP = 0{,}5n = 0{,}5(22) = 11$$
$$\sigma = 0{,}5\sqrt{22} = 2{,}345$$

Dado que $S = 17 > \mu = 11$, el estadístico de contraste de la aproximación normal es

$$Z = \frac{16{,}5 - 11}{2{,}345} = 2{,}35$$

Utilizando la tabla de la distribución normal estándar, el p-valor *aproximado* es $2(0{,}0094) = 0{,}0188$. La Figura 15.2 muestra los resultados obtenidos utilizando el programa Excel para resolver este problema:

$$P(X \leqslant 5 \,|\, n = 22,\ P = 0{,}5) = P(X \geqslant 17 \,|\, n = 22,\ P = 0{,}5) = 0{,}0845$$

En este ejemplo, que es de dos colas, el p-valor $= 2(0{,}00845) = 0{,}0169$ (algo menor que el p-valor de 0,0188 obtenido por medio del método de la aproximación normal).

Figura 15.2. Ejemplo de los ingresos iniciales (salida Excel).

La Figura 15.3 muestra la salida Minitab de este ejemplo.

```
Sign test of median = 35000 versus not = 35000

N   Below  Equal  Above      P      Median

23  17       1      5      0.0169   32000
```

Figura 15.3. Ejemplo de los ingresos iniciales (salida Minitab).

Tanto Excel como Minitab calculan el p-valor utilizando las probabilidades binomiales. Si $n > 50$, entonces Minitab calcula el p-valor utilizando la aproximación normal.

Intervalo de confianza de la mediana

Para calcular intervalos de confianza de la mediana basados en el contraste de signos puede utilizarse el programa Minitab. Consideremos los ingresos iniciales que se indican en el ejemplo 15.3 y se encuentran en el fichero de datos **Income**. Obsérvese que en la salida Minitab de la Figura 15.4 se incluyen tres intervalos de confianza. La primera fila indica el nivel de confianza obtenido (0,9069) justo por debajo del deseado (0,95); la tercera indica el nivel de confianza alcanzable (0,9653) justo por encima del deseado (0,95). «El cálculo del primero y el tercer intervalo se realiza con un método parecido al de los signos que se emplea cuando se hace un contraste de hipótesis de la mediana. Primero se ordenan las observaciones. El intervalo que va de la d-ésima observación más pequeña a la d-ésima observación más grande tiene una confianza de $1 - 2P(X < d)$ utilizando la distribución binomial en la que $P = 0{,}5$. Los intervalos que tienen coeficientes de confianza justo por encima y por debajo del deseado son los que se seleccionan. Sólo raras veces puede lograrse la confianza deseada con estos intervalos» (véase la referencia bibliográfica 7). En nuestro ejemplo, el intervalo que va de la octava observación más pequeña a la octava más grande tiene un nivel de confianza de 0,9069, donde X sigue una distribución binomial, siendo $n = 23$ y $P = 0{,}5$. El intervalo intermedio establece que los ingresos medianos se encuentran entre 30.393 $ y 34.442 $ con una confianza del 95 por ciento. Asimismo, el intervalo que va de la séptima observación más pequeña a la séptima más grande tiene un nivel de confianza de 0,9653, donde X sigue una distribución binomial, siendo $n = 23$ y $P = 0{,}5$.

Figura 15.4. Intervalo de confianza del ejemplo de los ingresos iniciales (salida Minitab).

```
Sign confidence interval for median

                                Achieved
                  N   Median    Confidence    Confidence interval   Position
Incomes          23    32000        0.9069    (   31000,   34000)          8
                                    0.9500    (   30393,   34442)        NLI
                                    0.9653    (   29900,   34800)          7
```

Hettmansperger y Sheather desarrollaron el método de interpolación no lineal utilizado por Minitab para hallar el intervalo de confianza intermedio (véase la referencia bibliográfica 3). También se obtienen de una manera parecida otros intervalos de confianza para métodos no paramétricos analizados en este capítulo.

EJERCICIOS

Ejercicios aplicados

15.1. Se pide a una muestra aleatoria de 12 analistas financieros que predigan cuánto subirán en términos porcentuales los precios de las acciones ordinarias de dos empresas el próximo año. La tabla muestra los resultados obtenidos. Utilice el contraste de signos para contrastar la hipótesis nula de que en la población de analistas no hay una preferencia general por la subida del precio de las acciones de una de las empresas o por la subida del precio de las acciones de la otra.

Analista	Acción 1	Acción 2	Analista	Acción 1	Acción 2
A	6,8	7,2	G	9,3	10,1
B	9,8	12,3	H	1,0	2,7
C	2,1	5,3	I	−0,2	1,3
D	6,2	6,8	J	9,6	9,8
E	7,1	7,2	K	12,0	12,0
F	6,5	6,2	L	6,3	8,9

15.2. Una organización ofrece un programa destinado a aumentar el nivel de comprensión de los estudiantes cuando leen trabajos técnicos rápidamente. Se da a cada uno de los miembros de una muestra aleatoria de 10 estudiantes 30 minutos para leer un artículo. A continuación, se realiza un contraste del nivel de comprensión logrado. Este proceso se repite una vez que estos estudiantes terminan el programa. La tabla adjunta muestra los niveles de comprensión obtenidos antes y después de asistir al programa. Utilice el contraste de signos para contrastar la hipótesis nula de que en esta población no hay una mejora general de los niveles de comprensión después de asistir al programa.

Estudiante	Antes	Después	Estudiante	Antes	Después
A	62	69	F	53	61
B	63	72	G	49	63
C	84	80	H	58	59
D	70	70	I	83	87
E	60	69	J	92	98

15.3. Se pregunta a una muestra de 11 encargados de supermercados que tienen una caja rápida si sus clientes tienen una actitud positiva hacia la caja rápida. Siete contestan «sí» y cuatro contestan «no». Contraste la hipótesis nula de que, en la población de encargados, las respuestas se reparten por igual entre «sí» y «no» frente a la hipótesis alternativa bilateral.

15.4. Se ha examinado una muestra de 60 empresas que recompraron franquicias. En estos casos, los rendimientos de las acciones ordinarias en torno a la fecha de anuncio de la recompra fueron positivos 39 veces, negativos 18 y cero 3. Contraste la hipótesis nula de que los rendimientos positivos y los negativos son igual de probables frente a la hipótesis alternativa de que los positivos son más probables (véase la referencia bibliográfica 2).

15.5. En una muestra aleatoria de 130 votantes, 44 eran partidarios de una subida de los impuestos para aumentar los gastos en educación, 68 eran contrarios y 18 no manifestaron su opinión. Contraste la hipótesis nula de que los votantes están repartidos por igual en esta cuestión frente a una hipótesis alternativa bilateral.

15.6. Se ha pedido a una muestra aleatoria de 60 economistas profesionales que predigan si la tasa de inflación será el próximo año más alta, más baja

o más o menos igual que la de este año. Los resultados se muestran en la tabla adjunta. Contraste la hipótesis nula de que los economistas están divididos por igual en esta cuestión.

Predicción	Número
Más alta	20
Más baja	29
Más o menos igual	11

15.2. Contraste de Wilcoxon basado en la ordenación de las diferencias

Uno de los inconvenientes del contraste de signos es que sólo tiene en cuenta una cantidad muy reducida de información, a saber, los signos de las diferencias. Por ejemplo, en la Tabla 15.1 el contraste de signos indica simplemente qué producto se prefiere y *no tiene en cuenta el grado de preferencia*. Cuando el tamaño de la muestra es pequeño, es de esperar, pues, que el contraste no sea muy poderoso. El contraste de Wilcoxon basado en la ordenación de las diferencias es un método para incorporar información sobre la magnitud de las diferencias entre pares enlazados. Sigue siendo un contraste que no depende de la distribución. Al igual que muchos contrastes no paramétricos, se basa en las *ordenaciones*.

El contraste de Wilcoxon en el caso de muestras pareadas

El **contraste de Wilcoxon** puede emplearse cuando se dispone de una muestra aleatoria de pares enlazados de observaciones. Supongamos que la distribución poblacional de las diferencias en estas **muestras pareadas** es simétrica y que queremos contrastar la hipótesis nula de que esta distribución está centrada en 0. Descartando los pares entre los que la diferencia es 0, ordenamos las n diferencias absolutas restantes en sentido ascendente; en caso de empate, el puesto asignado es la media de los puestos que ocupan en la ordenación. Se calculan las sumas de los puestos correspondientes a las diferencias positivas y negativas y la menor de estas sumas es el estadístico de Wilcoxon, T, es decir,

$$T = \min(T_+, T_-) \tag{15.8}$$

donde

T_+ = suma de los puestos correspondientes a diferencias positivas
T_- = suma de los puestos correspondientes a diferencias negativas
n = número de diferencias no nulas

Se rechaza la hipótesis nula si T es menor o igual que el valor de la Tabla 10 del apéndice.

EJEMPLO 15.4. Preferencia por un producto (contraste de Wilcoxon)

Resuelva el ejemplo 15.1 de la valoración de una salsa de pizza utilizando el contraste de Wilcoxon.

Solución

Prescindimos, al igual que en el contraste de signos, de cualquier diferencia de 0, por lo que eliminamos el estudiante G del estudio y el tamaño de la muestra se reduce a $n = 7$. A continuación, ordenamos en sentido ascendente las diferencias absolutas no nulas. Es decir, asignamos un «1» al valor absoluto más bajo. Si dos o más valores son iguales, se les asigna la media de los puestos correspondientes. En nuestro ejemplo, las

dos diferencias absolutas más pequeñas son iguales. Por lo tanto, el puesto que les asignamos es la media de los puestos 1 y 2, es decir, 1,5. Asignamos el 3 al siguiente valor absoluto, y así sucesivamente. Ordenamos todas las diferencias y obtenemos la Tabla 15.2.

Tabla 15.2. Cálculo del estadístico de contraste de Wilcoxon para los datos sobre las preferencias.

Estudiante	Diferencia	Puesto (+)	Puesto (−)
A	−2		3
B	−5		6
C	1	1,5	
D	1	1,5	
E	−6		7
F	−3		4
G	0		
H	−4		5
	Suma de los puestos	**3**	**25**
Estadístico *T* de Wilcoxon = mínimo (3, 25) = 3			

Los puestos de las diferencias positivas y negativas se suman por separado. La menor de estas sumas es el estadístico T de Wilcoxon. En este ejemplo, $T = 3$.

A continuación, suponemos que la distribución poblacional de las diferencias pareadas es simétrica. La hipótesis nula que vamos a contrastar es que el centro de esta distribución es 0. En nuestro ejemplo, pues, suponemos que las diferencias de valoración de los dos productos siguen una distribución simétrica y queremos contrastar si esa distribución está centrada en 0, es decir, si no hay ninguna diferencia entre las valoraciones. Sospecharíamos de la hipótesis nula si la suma de los puestos correspondientes a diferencias positivas fuera muy diferente de la suma de los puestos correspondientes a diferencias negativas. Por lo tanto, se rechazará la hipótesis nula en el caso de los valores bajos del estadístico T.

Los puntos de corte de la distribución de esta variable aleatoria se encuentran en el apéndice y se refieren a los contrastes de que la distribución poblacional de las diferencias pareadas está centrada en algún número mayor que 0 o en algún número menor que 0 frente a la hipótesis alternativa unilateral. Cuando el tamaño de la muestra es n, la tabla muestra el número T_α tal que $P(T < T_\alpha) = \alpha$ correspondiente a distintas probabilidades α. Por ejemplo, si suponemos que $\alpha = 0{,}05$, vemos en la tabla que cuando $n = 7$, $P(T \leqslant 4) = 0{,}05$. Como el estadístico del contraste de Wilcoxon es $T = 3$, se rechaza la hipótesis nula frente a la hipótesis alternativa unilateral al nivel del 5 por ciento. Es probable que, en conjunto, las valoraciones del nuevo producto sean mayores.

Minitab (contraste de Wilcoxon)

Para realizar un contraste de Wilcoxon puede utilizarse el programa Minitab. Consideremos de nuevo las valoraciones que hacen los estudiantes de una salsa de pizza y que se muestran en la Tabla 15.1 (ejemplo 15.1) y se repiten aquí:

Estudiante	A	B	C	D	E	F	G	H
Valoración (original)	6	4	5	8	3	6	7	5
Valoración (nueva)	8	9	4	7	9	9	7	9

En el caso de las muestras pareadas, se introducen los datos de cada par en columnas separadas en una hoja de trabajo de Minitab y se utiliza Calc para obtener las diferencias entre las columnas o se introducen simplemente las diferencias de las dos columnas si se dispone fácilmente de ellas. La Figura 15.5 muestra la salida Minitab del contraste de Wilcoxon. En este caso, el p-valor es 0,038 para un contraste unilateral de la cola inferior. Obsérvese que la información adicional que suministran los paquetes informáticos permite rechazar la hipótesis nula a un nivel de significación mucho más bajo que el que es posible con el contraste de signos. Sabemos que aunque el programa informático suministra información como el p-valor, la interpretación correcta es responsabilidad del lector. Éste es frecuentemente el caso en los estudios que se publican en revistas de investigación (véase la referencia bibliográfica 4). La interpretación correcta es fundamental.

Figura 15.5. Ejemplo de la salsa de pizza (salida Minitab).

Wilcoxon Signed Rank Test for Pizza Sauce Example

```
Test of median = 0.000000 versus median  <  0.000000

                 N for   Wilcoxon           Estimated
            N    Test    Statistic    P      Median
Pizza Sauce 8     7        3.0       0.038  -2.250
```

Aproximación normal

Cuando el número n de diferencias no nulas en la muestra es grande ($n > 20$), la distribución normal constituye una buena aproximación del estadístico de Wilcoxon T en el caso de la hipótesis nula de que las diferencias poblacionales están centradas en 0. Cuando esta hipótesis es verdadera, la media y la varianza de esta distribución se hallan por medio de las ecuaciones siguientes.

Contraste de Wilcoxon: aproximación normal (grandes muestras)

En la hipótesis nula de que las diferencias poblacionales están centradas en 0, el contraste de Wilcoxon tiene una media y una varianza que vienen dadas por

$$E(T) = \mu_T = \frac{n(n+1)}{4} \tag{15.9}$$

y

$$\text{Var}\,(T) = \sigma_T^2 = \frac{n(n+1)(2n+1)}{24} \tag{15.10}$$

Entonces, cuando el tamaño de la muestra, n, es grande, la distribución de la variable aleatoria, Z, es aproximadamente normal estándar donde

$$Z = \frac{T - \mu_T}{\sigma_T} \tag{15.11}$$

Si el número, n, de diferencias no iguales a cero es grande y T es el valor observado del estadístico de Wilcoxon, los siguientes contrastes tienen un nivel de significación α.

1. Si la hipótesis alternativa es unilateral, se rechaza la hipótesis nula si

$$\frac{T - \mu_T}{\sigma_T} < -z_\alpha$$

2. Si la hipótesis alternativa es bilateral, se rechaza la hipótesis nula si

$$\frac{T - \mu_T}{\sigma_T} < -z_{\alpha/2}$$

EJEMPLO 15.5. Métodos de postauditoría (contraste de Wilcoxon)

En un estudio se compararon empresas que tenían sofisticados métodos de postauditoría y empresas que no tenían métodos de ese tipo. Se examinó una muestra de 31 pares de empresas. Se calculó el cociente entre la valoración de mercado y los costes de reposición de los activos de cada una y se utilizó como medida de los resultados de las empresas. En cada uno de los 31 pares, una de las empresas utilizaba un sofisticado método de postauditoría y la otra no. Se calcularon las 31 diferencias entre los cocientes y se ordenaron las diferencias absolutas. La menor de las sumas de los puestos, 189, correspondió a los pares en los que el cociente era mayor en el caso de la empresa que carecía de sofisticados métodos de postauditoría. Contraste la hipótesis nula de que la distribución de las diferencias entre los cocientes está centrada en 0 frente a la hipótesis alternativa de que tiende a ser menor en las empresas que carecen de sofisticados métodos de postauditoría (véase la referencia bibliográfica 8).

Solución

Dada una muestra de $n = 31$ pares, el estadístico de Wilcoxon tiene, según la hipótesis nula, la media

$$\mu_T = \frac{n(n+1)}{4} = \frac{(31)(32)}{4} = 248$$

y la varianza

$$\text{Var}\,(T) = \sigma_T^2 = \frac{n(n+1)(2n+1)}{24} = \frac{(31)(32)(63)}{24} = 2.604$$

por lo que la desviación típica es

$$\sigma_T = 51{,}03$$

El valor observado del estadístico es $T = 189$. Se deduce de las ecuaciones 15.9-15.11 que se rechaza la hipótesis nula frente a la hipótesis alternativa unilateral si

$$Z = \frac{T - \mu_T}{\sigma_T} = \frac{189 - 248}{51{,}03} = \frac{-59}{51{,}03} = -1{,}16 < z_\alpha$$

Suponiendo que $\alpha = 0{,}05$

$$z_\alpha = -1{,}645$$

El resultado del contraste no es suficiente para rechazar la hipótesis nula. Utilizando la distribución normal estándar, la hipótesis nula puede rechazarse a todos los niveles de significación de 12,3 por ciento o más.

EJERCICIOS

Ejercicios aplicados

15.7. Irvine y Rosenfeld (véase la referencia bibliográfica 4) estudiaron «la influencia de la emisión de *Monthly Income Preferred Stock* (MIPS) en los precios de las acciones ordinarias de las empresas emisoras». El fisco ha permitido deducir de los impuestos los dividendos de las MIPS desde el momento en que las introdujo por primera vez Goldman Sachs en 1993. Por lo tanto, «la emisión de MIPS permite a la empresa aumentar su capital social con un coste después de impuestos casi igual al de la deuda a largo plazo». Uno de los aspectos de su estudio es una comparación de algunas características financieras de las empresas que habían emitido MIPS (un total de 185) con las de empresas similares que no habían emitido MIPS antes del 1 de enero de 1999. Las empresas emisoras de MIPS también se dividieron entre empresas industriales que cotizan en bolsa y empresas de servicios públicos, como las telefónicas, las eléctricas, las de gas y las de agua. La tabla adjunta es una lista parcial de algunos resultados de este estudio:

	N de las empresas MIPS	Media de las empresas MIPS	Empresas similares	Contraste de signos	Contraste de la diferencia de ordenaciones
Activos totales (miles de millones)					
Todas las empresas que emiten MIPS	185	26,47	19,42	0,01	0,01
Empresas de servicios públicos	83	10,45	8,65	0,01	0,01
Empresas industriales	102	39,60	28,26	0,01	0,01
Cobertura de intereses					
Todas las empresas que emiten MIPS	164	5,53	7,71	0,04	0,01
Empresas de servicios públicos	83	4,44	5,15	0,01	0,01
Empresas industriales	81	6,63	10,25	0,06	0,01
Cociente entre deuda a corto plazo y activos totales (%)					
Todas las empresas que emiten MIPS	185	23,5	21,4	0,06	0,03
Empresas de servicios públicos	83	32,6	29,3	0,03	0,01
Empresas industriales	102	16,1	14,9	0,19	0,28

Analice los resultados de esta parte del estudio.

15.8. Se pide a una muestra aleatoria de 10 estudiantes que valoren en una cata a ciegas la calidad de dos marcas de cerveza, una nacional y una importada. Las valoraciones se basan en una escala de 1 (mala) a 10 (excelente). La tabla adjunta muestra los resultados. Utilice el contraste de Wilcoxon para contrastar la hipótesis nula de que la distribución de las diferencias pareadas está centrada en 0 frente a la hipótesis alternativa de que la población de todos los estudiantes bebedores de cerveza prefiere la marca importada.

Estudiante	Nacional	Importada	Estudiante	Nacional	Importada
A	2	6	F	4	8
B	3	5	G	3	9
C	7	6	H	4	6
D	8	8	I	5	4
E	7	5	J	6	9

15.9. Dieciséis estudiantes universitarios de primer año se agruparon en ocho pares de tal forma que los dos miembros de cada par fueran lo más parecidos posibles en lo que se refería a su expediente académico —medido por medio de las calificaciones obtenidas en la enseñanza secundaria y en el examen de acceso a la universidad— y a sus orígenes sociales. La principal diferencia existente dentro de los pares era que uno de los estudiantes procedía de la región en la que estaba la universidad y el otro no. Al final del primer año de universidad, se registraron las calificaciones medias obtenidas por estos estudiantes; los resultados se muestran en la tabla. Utilice el contraste de Wilcoxon para analizar los datos. Analice las implicaciones de los resultados del contraste.

Par	De la región	De fuera de la región	Par	De la región	De fuera de la región
A	3,4	2,8	E	3,9	3,7
B	3,0	3,1	F	2,3	2,8
C	2,4	2,7	G	2,6	2,6
D	3,8	3,3	H	3,7	3,3

15.10. En un estudio se pidió a una muestra aleatoria de 40 estudiantes de administración de empresas que acababan de cursar las asignaturas de introducción tanto de estadística como de contabilidad que valoraran el interés de cada una en una escala de 1 (nada interesante) a 10 (muy interesante). Se calcularon las 40 diferencias entre los pares de valoraciones y se ordenaron las diferencias absolutas. La suma menor, que era la de los estudiantes que pensaban que la asignatura de contabilidad era la más interesante, era 281. Contraste la hipótesis nula de que la población de estudiantes de administración de empresas valoraría estos cursos por igual frente a la hipótesis alternativa de que el curso de estadística se considera el más interesante.

15.11. Un consultor tiene interés en saber cómo afecta la introducción de un programa de gestión total de la calidad a la satisfacción de los trabajadores en el trabajo. Se pide a una muestra aleatoria de 30 trabajadores que evalúe el nivel de satisfacción en una escala de 1 (muy insatisfecho) a 10 (muy satisfecho) tres meses antes de que se introduzca el programa. Se pide a los miembros de esta misma muestra que hagan esta evaluación de nuevo tres meses después de la introducción del programa. Se calculan las 30 diferencias entre los pares de valoraciones y se ordenan las diferencias absolutas. La suma menor de todas las sumas de los puestos, que es la de los que están más satisfechos antes de la introducción del programa, es de 160. ¿Qué conclusiones pueden extraerse de este resultado?

15.12. Se toma una muestra aleatoria de 80 propietarios de magnetoscopios. Se pide a cada uno de los miembros de la muestra que valore la cantidad de tiempo que dedica al mes a ver los programas de televisión que ha grabado y a ver las cintas alquiladas. A continuación, se calculan las 80 diferencias entre las cantidades de tiempo y se ordenan sus valores absolutos. La menor de las sumas de los puestos correspondientes a los programas de televisión grabados es de 1.502. Analice las implicaciones de este resultado.

15.3. Contraste *U* de Mann-Whitney

En el Capítulo 11 vimos cómo podía compararse la posición central de dos distribuciones poblacionales cuando se disponía de una muestra aleatoria de datos pareados. En este apartado introducimos un contraste del mismo problema cuando se toman *muestras aleatorias independientes* de las dos poblaciones, el **contraste *U* de Mann-Whitney**. La distribución del estadístico de Mann-Whitney, U, se aproxima a la distribución normal a un ritmo bastante rápido a medida que aumenta el número de observaciones muestrales. La aproximación es adecuada si cada muestra contiene al menos 10 observaciones. Por lo tanto, sólo consideraremos aquí las muestras en las que $n_1 \geqslant 10$ y $n_2 \geqslant 10$. Para contrastar la hipótesis nula de que la posición central de las dos distribuciones poblacionales es igual, suponemos que, aparte de la existencia de cualquier posible diferencia entre las posiciones centrales, las dos distribuciones poblacionales son idénticas.

Estadístico *U* de Mann-Whitney

Supongamos que, aparte de la existencia de posibles diferencias entre las posiciones centrales, las dos distribuciones poblacionales son idénticas. Supongamos que se dispone de n_1 observaciones de la primera población y n_2 observaciones de la segunda. Se juntan las dos muestras y se ordenan las observaciones en sentido ascendente, asignando, en caso de empate, la media de los puestos correspondientes. Sea R_1 la suma de los puestos de las observaciones de la primera población. En ese caso, el **estadístico *U* de Mann-Whitney** se define de la forma siguiente:

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \qquad \textbf{(15.12)}$$

Puede demostrarse entonces que, si la hipótesis nula es verdadera, la variable aleatoria U tiene la media y la varianza definidas en las ecuaciones 15.13 y 15.14.

Contraste *U* de Mann-Whitney: aproximación normal

Suponiendo como hipótesis nula que las posiciones centrales de las dos distribuciones poblacionales son iguales, el estadístico ***U* de Mann-Whitney** tiene la media y la varianza siguientes:

$$E(U) = \mu_U = \frac{n_1 n_2}{2} \tag{15.13}$$

$$\text{Var}\,(U) = \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \tag{15.14}$$

Entonces, cuando las muestras son de gran tamaño (ambas son como mínimo de 10), la distribución normal es una buena aproximación de la distribución de la variable aleatoria

$$Z = \frac{U - \mu_U}{\sigma_U} \tag{15.15}$$

Las reglas de decisión del estadístico del contraste de Mann-Whitney, U, se indican en las ecuaciones 15.16 a 15.18.

Reglas de decisión del contraste *U* de Mann-Whitney

Se supone que las dos distribuciones poblacionales son idénticas, aparte de las diferencias que puedan existir entre sus posiciones centrales. Para contrastar la hipótesis nula de que las dos distribuciones poblacionales tienen la misma posición central, las reglas de decisión para un nivel de significación dado son las siguientes:

1. Si la hipótesis alternativa es la hipótesis de la cola superior unilateral, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{U - \mu_U}{\sigma_U} < -z_\alpha \tag{15.16}$$

2. Si la hipótesis alternativa es la hipótesis de la cola inferior unilateral, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{U - \mu_U}{\sigma_U} > z_\alpha \tag{15.17}$$

3. Si la hipótesis alternativa es la hipótesis bilateral, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \quad \frac{U - \mu_U}{\sigma_U} < -z_{\alpha/2} \quad \text{o} \quad \text{Rechazar } H_0 \text{ si } \quad \frac{U - \mu_U}{\sigma_U} > z_{\alpha/2} \tag{15.18}$$

EJEMPLO 15.6. Horas de estudio (contraste *U* de Mann-Whitney)

Hours

La Tabla 15.3 muestra el número de horas semanales que los estudiantes afirman que dedican a estudiar las asignaturas de introducción a la economía financiera y a la contabilidad. Los datos proceden de muestras aleatorias de 10 estudiantes de economía financiera y 12 de contabilidad.

¿Indican los datos la existencia de una diferencia en el número mediano de horas semanales que dedican los estudiantes a estudiar las asignaturas de introducción a la economía financiera y a la contabilidad? El fichero de datos se llama **Hours**.

Tabla 15.3. Número de horas semanales dedicadas a estudiar las asignaturas de economía financiera y de contabilidad.

Economía financiera	10	6	8	10	12	13	11	9	5	11		
Contabilidad	13	17	14	12	10	9	15	16	11	8	9	7

Solución

Nuestra hipótesis nula es que las posiciones centrales (medianas) de las dos distribuciones poblacionales son idénticas.

H_0: Mediana (1) = Mediana (2) *Los estudiantes dedican la misma cantidad de tiempo a estudiar las asignaturas de economía financiera y de contabilidad*

Se juntan las dos muestras y se ordenan las observaciones en sentido ascendente dando a los empates el mismo tratamiento que antes. Las ordenaciones resultantes se muestran en la Tabla 15.4.

Tabla 15.4. Número de horas semanales dedicadas a estudiar las asignaturas de economía financiera y de contabilidad

Economía financiera	**(Puesto)**	**Contabilidad**	**(Puesto)**
10	(10)	13	(17,5)
6	(2)	17	(22)
8	(4,5)	14	(19)
10	(10)	12	(15,5)
12	(15,5)	10	(10)
13	(17,5)	9	(7)
11	(13)	15	(20)
9	(7)	16	(21)
5	(1)	11	(13)
11	(13)	8	(4,5)
		9	(7)
		7	(3)
Suma de puestos 93,5		**Suma de puestos 159,5**	

Ahora, si la hipótesis nula fuera verdadera, sería de esperar que las ordenaciones medias de las dos muestras fueran muy parecidas. En este ejemplo, el puesto medio de los estudiantes de economía financiera es 9,35, mientras que el de los estudiantes de contabilidad es 13,29. Como ocurre siempre que se contrastan hipótesis, queremos saber cuál es la probabilidad de que hubiera una discrepancia de esta magnitud si la hipótesis nula fuera verdadera.

No es necesario calcular las dos sumas de los puestos, pues si conocemos una, podemos deducir la otra. Por ejemplo, en este caso los puestos deben sumar lo mismo que la suma de los enteros de 1 a 22, es decir, 253. Por lo tanto, cualquier contraste de la hipótesis puede basarse simplemente en una de las sumas de puestos. Si la economía financiera es la primera muestra, entonces

$$n_1 = 10 \qquad n_2 = 12 \qquad R_1 = 93{,}5$$

por lo que el valor observado del estadístico de Mann-Whitney es, de acuerdo con la ecuación 15.12,

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = (10)(12) + \frac{(10)(11)}{2} - 93{,}5 = 81{,}5$$

Utilizando la hipótesis nula de que las posiciones centrales de las dos distribuciones poblacionales son iguales y la ecuación 15.13, la distribución del estadístico tiene una media

$$E(U) = \mu_U = \frac{n_1 n_2}{2} = \frac{(10)(12)}{2} = 60$$

y una varianza

$$\text{Var}\,(U) = \sigma_U^2 = \frac{n_1 n_2(n_1 + n_2 + 1)}{12} = \frac{(10)(12)(23)}{12} = 230$$

Se deduce que

$$\frac{U - \mu_U}{\sigma_U} = \frac{81{,}5 - 60}{\sqrt{230}} = 1{,}42 \quad \text{y} \quad p\text{-valor} = 0{,}1556$$

Por lo tanto, la hipótesis nula puede rechazarse a niveles de significación superiores a 15,56 por ciento. Con el nivel de significación habitual de 0,05, el resultado del contraste no es suficiente para concluir que los estudiantes dedican más tiempo a estudiar una de estas materias que la otra. Podríamos haber utilizado un factor de corrección de continuidad en la aproximación normal. El *p*-valor será de algo más de 0,1556.

Si los estudiantes de contabilidad son la población 1, por lo que $n_1 = 12$ y $R_1 = 159{,}5$, el resultado es el mismo, ya que $z = -1{,}42$. El *p*-valor sigue siendo 0,1556.

Minitab (contraste *U* de Mann-Whitney) Minitab calcula el valor z utilizando un factor de corrección de continuidad. La Figura 15.6 es la salida Minitab del ejemplo 15.6. Obsérvese que el *p*-valor es algo más alto.

Mann–Whitney Test: Finance, Accounting

```
Finance    N =  10     Median =      10.000

Accounting N =  12     Median =      11.500

Point estimate for ETA1-ETA2 is      -2.000
95.6 Percent CI for ETA1-ETA2 is (-5.001,1.000)

W = 93.5

Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.1661
The test is significant at 0.1643 (adjusted for ties)
```

Figura 15.6. Ejemplo del número de horas de estudio (salida Minitab).

EJERCICIOS

Ejercicios aplicados

15.13. En un estudio se compararon empresas que tenían un comité de auditoría con empresas que no lo tenían. Se midió en muestras de empresas de cada tipo el grado de participación de los consejeros en la propiedad por medio del número de acciones que poseía el consejo de administración en porcentaje del número total de acciones emitidas. En la muestra, la participación de los consejeros era, en conjunto, mayor en las empresas que no tenían comité de auditoría. Para contrastar la significación estadística, se calculó el estadístico U de Mann-Whitney. Se observó que $(U - \mu_U)/\sigma_U$ era 2,01 (véase la referencia bibliográfica 1). ¿Qué conclusiones pueden extraerse de este resultado?

15.14. Un analista bursátil elaboró a comienzos del año una lista de acciones para comprar y otra de acciones para vender. En una muestra aleatoria de 10 acciones de la «lista de compra», los rendimientos porcentuales a lo largo del año eran los siguientes:

9,6	5,8	13,8	17,2	11,6
4,2	3,1	11,7	13,9	12,3

En una muestra aleatoria independiente de 10 acciones de la «lista de venta», los rendimientos porcentuales a lo largo del año eran los siguientes:

−2,7	6,2	8,9	11,3	2,1
3,9	−2,4	1,3	7,9	10,2

Utilice el contraste de Mann-Whitney para interpretar estos datos.

15.15. En una muestra aleatoria de 12 titulados en administración de empresas de una universidad privada, los sueldos de partida aceptados después de licenciarse (en miles de dólares) fueron los siguientes:

26,2	29,3	31,3	28,7	27,4	25,1
26,0	27,1	27,5	29,8	32,6	34,6

En una muestra aleatoria independiente de 10 titulados en administración de empresas de una universidad pública, los sueldos de partida aceptados después de licenciarse (en miles de dólares) fueron los siguientes:

25,3	28,2	29,2	27,1	26,8
26,5	30,7	31,3	26,3	24,9

Analice los datos utilizando el contraste de Mann-Whitney y comente los resultados.

15.4. Contraste de la suma de puestos de Wilcoxon

El **contraste de la suma de puestos de Wilcoxon** es parecido al contraste U de Mann-Whitney. Los resultados son los mismos con ambos contrastes. Lo incluimos aquí para completar el análisis, ya que es posible que se prefiera este contraste por su sencillez.

Estadístico *T* de la suma de los puestos de Wilcoxon

Supongamos que se dispone de n_1 observaciones de la primera población y n_2 observaciones de la segunda. Se juntan las dos muestras y se ordenan las observaciones en sentido ascendente, asignando, en caso de empate, la media de los puestos correspondientes. Sea T la suma de los puestos de las observaciones de la primera población (T en el contraste de la suma de puestos de Wilcoxon es igual que R_1 en el contraste U de Mann-Whitney). Suponiendo que la hipótesis nula es verdadera, el estadístico de la suma de puestos de Wilcoxon, T, tiene la media

$$E(T) = \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad \textbf{(15.19)}$$

y la varianza

$$\text{Var}(T) = \sigma_T^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \tag{15.20}$$

Entonces, cuando las muestras son de gran tamaño ($n_1 \geqslant 10$ y $n_2 \geqslant 10$), la distribución normal es una buena aproximación de la distribución de la variable aleatoria

$$Z = \frac{T - \mu_T}{\sigma_T} \tag{15.21}$$

Cuando hay un gran número de empates, la ecuación 15.20 puede no ser correcta (véase la referencia bibliográfica 6).

En el caso de los datos de la Tabla 15.4, $T = R_1 = 93{,}5$ y

$$E(T) = \mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{10(23)}{2} = 115$$

y

$$\text{Var}(T) = \sigma_T^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = 230$$

Obsérvese que la varianza de la distribución muestral del estadístico de la suma de puestos de Wilcoxon, T, es igual que la varianza de la distribución muestral del estadístico de Mann-Whitney, U. Se deduce que

$$\frac{T - \mu_T}{\sigma_T} = \frac{93{,}5 - 115}{\sqrt{230}} = -1{,}42 \quad \text{y} \quad p\text{-valor} = 0{,}1556$$

EJEMPLO 15.7. Beneficios de dos empresas (contraste de la suma de puestos de Wilcoxon)

En un estudio que pretendía comparar los resultados de empresas que revelan las predicciones de la dirección sobre los beneficios con los resultados de las que no las revelan, se tomaron muestras aleatorias de 80 empresas de cada una de las poblaciones. Se midió la variabilidad de la tasa de crecimiento de los beneficios en los 10 periodos anteriores en cada una de las 160 empresas y se ordenaron estas variabilidades. La suma de los puestos de las empresas que no revelan las predicciones de la dirección sobre los beneficios era 7.287 (véase la referencia bibliográfica 5). Contraste la hipótesis nula de que las posiciones centrales de las distribuciones poblacionales de las variabilidades de los beneficios son las mismas en los dos tipos de empresas frente a la hipótesis alternativa bilateral. Demuestre que estos resultados son iguales que los del contraste U de Mann-Whitney y los del contraste de la suma de puestos de Wilcoxon.

Solución

Dado que tenemos que $n_1 = 80$, $n_2 = 80$ y $R_1 = 7.287$, el valor calculado del estadístico de Mann-Whitney es

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = (80)(80) + \frac{(80)(81)}{2} - 7.287 = 2.353$$

Según la hipótesis nula, el estadístico de Mann-Whitney tiene la media

$$\mu_U = \frac{n_1 n_2}{2} = \frac{(80)(80)}{2} = 3.200$$

y la varianza

$$\sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} = \frac{(80)(80)(161)}{12} = 85.867$$

En este caso, tenemos que

$$z = \frac{2.353 - 3.200}{\sqrt{85.867}} = -2{,}89$$

En la Tabla 1 de la distribución normal estándar del apéndice, vemos que el valor de $\alpha/2$ correspondiente a un valor de z de 2,89 es 0,0019, por lo que el p-valor es 0,0038. Por lo tanto, la hipótesis nula puede rechazarse a todos los niveles de más del 0,38 por ciento.

El contraste de la suma de puestos de Wilcoxon utiliza las ecuaciones 15.19 a 15.21. La media de T es

$$E(T) = \frac{n_1(n_1 + n_2 + 1)}{2} = \frac{80(161)}{2} = 6.440$$

La varianza de T es igual que la de U (la ecuación 15.14 es igual que la 15.20). Por lo tanto, por la ecuación 15.21,

$$\frac{T - \mu_T}{\sigma_T} = \frac{7.287 - 6.440}{\sqrt{85.867}} = 2{,}89$$

y, de nuevo, puede rechazarse la hipótesis nula a todos los niveles de más del 0,38 por ciento. Se obtienen los mismos resultados utilizando el contraste U de Mann-Whitney que utilizando el de la suma de los puestos de Wilcoxon. Estos datos constituyen, pues, una prueba contundente en contra de la hipótesis de que las posiciones centrales de las distribuciones de las variabilidades poblacionales de las tasas de crecimiento de los beneficios de las empresas que revelan las predicciones de los beneficios son iguales que las de las empresas que no las revelan.

Ahora bien, si se nos hubieran dado los datos efectivos en lugar de simplemente los puestos en la ordenación, podríamos haber realizado un contraste de la hipótesis nula utilizando los métodos del Capítulo 11. Sin embargo, utilizando el contraste de Mann-Whitney, hemos observado que la hipótesis nula puede rechazarse *sin el supuesto de la normalidad de la población*.

EJERCICIOS

Ejercicios aplicados

15.16. Una empresa entrevista tanto a expertos en marketing como a expertos en economía financiera para cubrir el puesto de dirección general. Un equipo de altos directivos de la empresa realiza una larga entrevista y un largo examen a una muestra aleatoria de 10 expertos en marketing y a una muestra aleatoria independiente de 14 expertos en economía financiera. A continuación, ordena a los candidatos de 1 (el mejor para cubrir el puesto) a 24, como muestra la tabla adjunta. Contraste la hipótesis nula de que, en conjunto, los altos directivos de la empresa no tienen ninguna preferencia por los expertos en marketing o por los expertos en economía financiera frente a la hipótesis alternativa de que prefieren los expertos en economía financiera.

1. economía financiera	13. marketing
2. economía financiera	14. economía financiera
3. marketing	15. economía financiera
4. economía financiera	16. economía financiera
5. economía financiera	17. marketing
6. marketing	18. marketing
7. economía financiera	19. economía financiera
8. marketing	20. economía financiera
9. marketing	21. economía financiera
10. marketing	22. marketing
11. economía financiera	23. marketing
12. economía financiera	24. economía financiera

15.17. Un profesor pidió a una muestra aleatoria de 15 alumnos y a una muestra aleatoria independiente de 15 alumnas que escribieran un ensayo al final de un curso de escritura. A continuación, el profesor ordenó estos ensayos de 1 (el mejor) a 30 (el peor). Ésta es la ordenación.

Alumnos	26	24	15	16	8	29	12	6	18	
	11	13	19	10	28	7				
Alumnas	22	2	17	25	14	21	5	30	3	9
	4	1	27	23	20					

Contraste la hipótesis nula de que en conjunto el orden de los alumnos y el de las alumnas es el mismo frente a la hipótesis alternativa bilateral.

15.18. Un boletín informativo califica los fondos de inversión. Se eligen muestras aleatorias independientes de 10 fondos que tienen la máxima calificación y 10 que tienen la peor calificación. Las cifras siguientes son las tasas porcentuales de rendimiento que obtendrán estos 20 fondos el próximo año.

Mejor calificado	8,1	12,7	13,9	2,3	16,1	5,4	7,3
	9,8	14,3	4,1				
Peor calificado	3,5	14,0	11,1	4,7	6,2	13,3	7,0
	7,3	4,6	10,0				

Contraste la hipótesis nula de que no existe ninguna diferencia entre las posiciones centrales de las distribuciones poblacionales de las tasas de rendimiento frente a la hipótesis alternativa de que los fondos mejor calificados tienden a obtener mayores tasas de rendimiento que los peor calificados.

15.19. Se pregunta a una muestra aleatoria de 50 estudiantes qué sueldo debería estar dispuesta la universidad a pagar para atraer a la persona idónea para entrenar al equipo de fútbol. Se hace la misma pregunta a una muestra aleatoria independiente de 50 profesores. A continuación, se juntan las 100 cifras sobre el sueldo y se ordenan (asignándose 1 al sueldo más bajo). La suma de los puestos de los profesores es 2.024. Contraste la hipótesis nula de que no existe ninguna diferencia entre las posiciones centrales de las distribuciones de los sueldos propuestos por los estudiantes y por los profesores frente a la hipótesis alternativa de que en conjunto los estudiantes propondrían un sueldo más alto para atraer a un entrenador.

15.20. En un estudio se comparó el tiempo (en días) que tardaba una muestra aleatoria de 120 empresas australianas que tienen buenos informes de auditoría en publicar desde finales de año un informe preliminar sobre los beneficios con el que tardaba una muestra aleatoria independiente de 86 empresas cuyos informes no eran buenos. Se juntaron los tiempos que tardaban las 206 empresas y se ordenaron, asignándose al tiempo más corto el puesto 1. La suma de los puestos de las empresas cuya auditoría no era buena era 9.686 (véase la referencia bibliográfica 9). Contraste la hipótesis nula de que las posiciones centrales de las dos distribuciones poblacionales son idénticas frente a la hipótesis alternativa de que las empresas cuya auditoría no era buena tardaban más en publicar un informe preliminar sobre sus beneficios.

15.21. Se comparan los sueldos de partida de licenciados en administración de empresas de dos destacadas facultades de administración de empresas. Se toman muestras aleatorias de 30 estudiantes de cada una y se juntan y ordenan los 60 sueldos de partida. La suma de los puestos de los estudiantes de una de las facultades es 1.243. Contraste la hipótesis nula de que las posiciones centrales de las distribuciones poblacionales son idénticas.

15.5. Correlación de orden de Spearman

El coeficiente de correlación muestral puede verse seriamente afectado por las observaciones extremas. Además, los contrastes basados en él recurren para su validez al supuesto de la normalidad. Puede obtenerse una medida de la correlación en la que no influyen seriamente los valores extremos y en la que pueden basarse contrastes válidos de distribuciones poblacionales muy generales utilizando los puestos en ordenaciones. El contraste resultante será en ese caso no paramétrico.

Correlación de orden de Spearman

Supongamos que se toma una muestra aleatoria (x_1, y_1), ..., (x_n, y_n) de n pares de observaciones. Si las x_i y las y_i se ordenan en sentido ascendente y se calcula la correlación muestral de estos puestos, el coeficiente resultante se llama **coeficiente de correlación de orden de Spearman**. Si no hay empates, una fórmula equivalente para calcular este coeficiente es

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{15.22}$$

donde las d_i son las diferencias entre los puestos de los miembros de los distintos pares.

Los siguientes contrastes de la hipótesis nula H_0 de que no existe ninguna relación en la población tienen un nivel de significación α.

1. Para contrastar la hipótesis nula de que no existe ninguna relación frente a la hipótesis alternativa de que existe una relación positiva, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si} \quad r_s > r_{s,\alpha} \tag{15.23}$$

2. Para contrastar la hipótesis nula de que no existe ninguna relación frente a la hipótesis alternativa de que existe una relación negativa, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si} \quad r_s < -r_{s,\alpha} \tag{15.24}$$

3. Para contrastar la hipótesis nula de que no existe ninguna relación frente a la hipótesis alternativa bilateral de que existe alguna relación, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si} \quad r_s < -r_{s,\alpha/2} \quad \text{o} \quad r_s > r_{s,\alpha/2} \tag{15.25}$$

EJEMPLO 15.8. Promoción de los cruceros (correlación de orden de Spearman)

Para promover los cruceros por el Mediterráneo, supongamos que una empresa de cruceros se anuncia en 17 revistas de viajes. Se invita a los lectores a pedir folletos y literatura. Las dos variables que se quiere relacionar son:

X: coste de la publicidad y la distribución, en miles de dólares
Y: rendimiento de la publicidad

donde este último se define de la forma siguiente:

$$Y = (\text{ingresos estimados de las solicitudes de información} - \\ - \text{coste de la publicidad}) \div \text{coste de la publicidad}$$

La Tabla 15.5 enumera los puestos de estas dos variables de los 17 anuncios de revistas. Calcule el coeficiente de correlación de orden de Spearman y contraste la relación entre las variables.

Tabla 15.5. Cálculos de la correlación de orden del ejemplo de los cruceros.

Revista	Orden (x_i)	Orden (y_i)	D_i = orden (x_i) − orden (y_i)	d_i^2
1	14	2	12	144
2	8	4	4	16
3	1	16	−15	225
4	16	1	15	225
5	17	5	12	144
6	13	6	7	49
7	15	8	7	49
8	2	11	−9	81
9	7	9	−2	4
10	3	13	−10	100
11	6	12	−6	36
12	9	17	−8	64
13	5	3	2	4
14	4	7	−3	9
15	11	14	−3	9
16	12	15	−3	9
17	10	10	0	0
				Suma 1.168

Solución

Dado que no hay empates, utilizamos la ecuación 15.22 y obtenemos

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6(1.168)}{17[(17)^2-1]} = -0{,}431$$

Dado que hay 17 pares de observaciones, los puntos de corte (véase la Tabla 11 del apéndice) en los contrastes al nivel del 10 por ciento y del 5 por ciento son, respectivamente,

$$r_{s,0,05} = 0{,}412 \quad \text{y} \quad r_{s,0,025} = 0{,}49$$

La hipótesis nula de que no existe ninguna relación puede rechazarse frente a la hipótesis alternativa bilateral, según la regla de decisión, al nivel del 10 por ciento, pero no al nivel del 5 por ciento. Nuestras conclusiones no se basan en el supuesto de la normalidad de la población.

Si no hay empates, entonces es sencillo calcular la correlación de orden de Spearman con el programa Minitab o con el Excel. Si hay empate, calculamos la correlación simple (apartado 12.1) entre los puestos.

EJERCICIOS

Ejercicios básicos

15.22. Los estudiantes de un curso de tecnología de comercio electrónico tienen que hacer un examen final escrito y un proyecto para obtener la calificación final. Las calificaciones de una muestra aleatoria de 10 estudiantes tanto en el examen como en el proyecto son

Examen	81	62	74	78	93	69	72	83	90	84
Proyecto	76	71	69	76	87	62	80	75	92	79

a) Halle el coeficiente de correlación de orden de Spearman.
b) Contraste la relación.

15.23. La tabla adjunta muestra el rendimiento porcentual de una muestra aleatoria de 20 fondos de inversión a largo plazo en un periodo de 12 meses y los activos totales (en millones de dólares).

Rendimiento	Activos	Rendimiento	Activos	Rendimiento	Activos
29,3	300	16,0	421	12,9	75
27,6	70	15,5	99	11,3	610
23,7	3.004	15,2	756	9,9	264
22,3	161	15,0	730	7,9	27
22,0	827	14,4	436	6,7	71
19,6	295	14,0	143	3,3	719
17,6	29	13,7	117		

a) Calcule el coeficiente de correlación de orden de Spearman.
b) Realice un contraste no paramétrico de la hipótesis nula de que no existe ninguna relación en la población frente a una hipótesis alternativa bilateral.
c) Analice las ventajas de un contraste no paramétrico de estos datos.

RESUMEN

Los contrastes no paramétricos analizados en este capítulo representan un subconjunto muy pequeño de los métodos no paramétricos que se utilizan actualmente. En capítulos posteriores, encontraremos algunos otros contrastes que no dependen de la distribución.

Es instructivo comparar los contrastes de este capítulo con los de los Capítulos 10 y 11, en los que examinamos el problema de contrastar la igualdad de dos medias poblacionales, *suponiendo que las distribuciones poblacionales son normales*. También puede considerarse que los contrastes desarrollados en este capítulo son contrastes de esta hipótesis nula, pero suponiendo solamente que las dos distribuciones poblacionales tienen la misma forma. Ésta es la principal ventaja de los métodos no paramétricos. Son adecuados con una amplia variedad de supuestos sobre las distribuciones poblacionales subyacentes.

Entre las ventajas de los contrastes no paramétricos se encuentran las siguientes:

1. *Menos supuestos sobre la población*
No es necesario el supuesto de la normalidad. Es-

tos contrastes no paramétricos son adecuados con una amplia variedad de supuestos sobre las distribuciones poblacionales subyacentes.

2. *Los cálculos son más sencillos*
   Los contrastes no paramétricos pueden realizarse bastante deprisa, especialmente el contraste de signos.
3. *Pueden contrastarse datos nominales u ordinales*
   Por ejemplo, si todo lo que se sabe en un estudio de comparación de productos es qué producto se prefiere, puede aplicarse inmediatamente el contraste de signos. En muchas situaciones prácticas, sólo se dispone de datos en forma de ordenaciones, lo que lleva lógicamente a utilizar métodos como el contraste de Wilcoxon o el de Mann-Whitney.
4. *Influyen menos los casos atípicos*
   De la misma forma que en la media pueden influir las observaciones extremas, lo mismo ocurre con las inferencias basadas en los contrates *t* de los Capítulos 10 y 11. En cambio, los contrastes basados en los puestos dan mucho menos peso a los valores muestrales atípicos.

Uno de los inconvenientes de los contrastes no paramétricos es que, con el supuesto de la normalidad de la población, los métodos no paramétricos son *menos poderosos*. En el caso de las poblaciones que siguen una distribución normal, los contrastes paramétricos del Capítulo 11 son más poderosos que los contrastes basados en ordenaciones, ya que estos últimos descartan parte de la información de los datos. Es decir, los contrastes paramétricos tienen más capacidad para detectar los incumplimientos de la hipótesis nula. Sin embargo, al menos en las muestras de moderado tamaño, los contrastes como el de Wilcoxon y el de Mann-Whitney sólo son algo menos poderosos que los contrastes *t* cuando las distribuciones poblacionales son normales. Por esta razón, así como porque pueden aplicarse en muchos más casos, estos contrastes no paramétricos son muy conocidos. Además, cuando la distribución poblacional se aleja mucho de la normal, pueden tener mucho más poder que los contrastes correspondientes basados en la distribución normal. Los programas informáticos también han aumentado el uso de los contrastes no paramétricos.

Dado que los métodos no paramétricos son bastante difíciles de extender a los problemas que implican la construcción de complejos modelos, los métodos tradicionales de los Capítulos 10 y 11, cuyo desarrollo es mucho más sencillo, siguen constituyendo los elementos principales del análisis estadístico.

TÉRMINOS CLAVE

coeficiente de correlación de orden de Spearman, 649
contraste de signos, 628
contraste de la suma de puestos de Wilcoxon, 645
contraste *U* de Mann-Whitney, 641
contraste de Wilcoxon basado en la ordenación de las diferencias, 636

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

15.24. ¿Qué significa que un contraste no sea paramétrico? ¿Cuáles son las ventajas relativas de esos contrastes?

15.25. Ponga un ejemplo realista de un problema estadístico del mundo de la empresa en el que sea preferible un contraste no paramétrico al contraste paramétrico alternativo.

15.26. En una muestra aleatoria de 12 analistas, 7 creen que las ventas de automóviles en Estados Unidos probablemente serán mayores el año que viene que éste, 2 creen que serán mucho menores y los demás prevén que serán más o menos iguales que este año. ¿Qué conclusión podemos extraer de estos datos?

15.27. En una muestra aleatoria de 16 analistas de los tipos de cambio, 8 creen que el yen japonés será una excelente inversión este año, 5 creen que será una mala inversión y 3 no tienen ninguna opinión decidida sobre esta cuestión. ¿Qué conclusiones podemos extraer de estos datos?

15.28. En una muestra aleatoria de 100 estudiantes universitarios, 35 esperan disfrutar de un nivel de vida más alto que el de sus padres, 43 esperan disfrutar de un nivel de vida más bajo y 22 esperan tener el mismo nivel de vida que sus padres. ¿Son estos datos una prueba contundente de que en la población de estudiantes es mayor el número de estudiantes que esperan tener un nivel de vida más bajo que el de sus pa-

dres que el número de estudiantes que esperan tener un nivel de vida más alto?

15.29. En una muestra aleatoria de 120 profesores de administración de empresas, 48 creen que la capacidad de análisis de los estudiantes ha mejorado en la última década, 35 creen que ha empeorado y 37 no ven ningún cambio perceptible. Evalúe la fuerza de la evidencia muestral que sugiere que el número de profesores que creen que la capacidad de análisis ha mejorado es mayor que el número de profesores que creen que ha empeorado.

15.30. Se pide a una muestra aleatoria de 10 analistas de empresas que valoren en una escala de 1 (muy malas) a 10 (muy buenas) las perspectivas de su propia empresa y las de la economía en general en el presente año. Los resultados obtenidos se muestran en la tabla adjunta. Utilizando el contraste de Wilcoxon, analice la proposición de que en conjunto los analistas de empresas son más optimistas sobre las perspectivas de sus propias empresas que sobre las perspectivas de la economía en general.

Analista	Propia empresa	Economía en general	Analista	Propia empresa	Economía en general
1	8	8	6	6	9
2	7	5	7	7	7
3	6	7	8	5	2
4	5	4	9	4	6
5	8	4	10	9	6

15.31. Se construyen nueve pares de perfiles hipotéticos de empleados de empresas que solicitan la admisión en un máster de administración de empresas. Dentro de cada par, los perfiles son idénticos; lo único que varía es que uno de los candidatos es un hombre y el otro es una mujer. En las entrevistas realizadas en el proceso de admisión, se evalúa en una escala de 1 (poca) a 10 (mucha) la idoneidad de los candidatos para el programa. Los resultados se muestran en la tabla adjunta. Analice estos datos utilizando el contraste de Wilcoxon.

Entrevista	Hombre	Mujer	Entrevista	Hombre	Mujer
1	8	8	6	9	9
2	9	10	7	5	3
3	7	5	8	4	5
4	4	7	9	6	2
5	8	8			

Bibliografía

1. Brandbury, M. E., «The Incentives for Voluntary Audit Committee Formation», *Journal of Accounting and Public Policy*, 9, 1990, págs. 19-36.
2. Brickely, F. H. Dark y M. S. Weisbach, «An Agency Perspective on Franchising», *Financial Management*, 20, n.º 1, 1991, págs. 27-35.
3. Hettmansperger, T. P. y S. J. Sheather, «Confidence Intervals Based on Interpolated Order Statistics», *Statistics and Probability Letters*, 4, 1986, págs. 75-79.
4. Irvine, Paul y James Rosenfeld, «Raising Capital Using Monthly Income Preferred Stock: Market Reaction and Implications for Capital Structure Theory», *Financial Management*, 29, verano, 2000, págs. 5-20.
5. Jaggi, B. y P. Grier, «A Comparative Analysis of Forecast Disclosing and Nondisclosing Firms», *Financial Management*, 9, n.º 2, 1980, págs. 38-43.
6. Lehman, E. L., *Nonparametrics*: *Statistical Methods Based on Ranks*, San Francisco, Holden-Day, 1975.
7. *Mintab User's Guide* 2: *Data Analysis and Quality Tools*, State College, PA, Minitab, 1997.
8. Meyers, M. D., L. A. Gordon y M. M. Hamer, «Postauditing Capital Assets and Firm Performance: An Empirical Investigation», *Managerial and Decision Economics*, 12, 1991, págs. 317-327.
9. Whittred, G. P., «Audit Qualification and the Timeliness of Corporate Annual Reports», *Accounting Review*, 55, 1980, págs. 563-577.

Capítulo 16

Contrastes de la bondad del ajuste y tablas de contingencia

Esquema del capítulo

Introducción

En este capítulo analizamos algunos contrastes que se basan en la distribución ji-cuadrado. En primer lugar, examinamos un contraste de la hipótesis de que los datos son generados por una distribución de probabilidad *totalmente especificada*. Los analistas de mercado utilizan a menudo esta técnica para averiguar si los productos son preferidos por igual por los posibles clientes o para averiguar si las cuotas de mercado de diversas marcas de un producto han cambiado en un determinado periodo de tiempo.

A continuación, contrastamos la hipótesis de que los datos son generados por alguna distribución, como la binomial, la distribución de Poisson o la normal sin suponer que se conocen los parámetros de esa distribución. En estas circunstancias, pueden utilizarse los datos de que se dispone para estimar los parámetros poblacionales desconocidos. Cuando se estiman parámetros poblacionales, se utiliza un contraste de la bondad del ajuste.

El contraste de la ji-cuadrado puede extenderse para abordar un problema en el que se toma una muestra de la población y cada uno de sus miembros puede clasificarse de manera inequívoca de acuerdo con un par de atributos. La hipótesis que se contrasta es que no existe ninguna relación en la población entre las posesiones de estos atributos. Los profesionales de las empresas utilizan este método frecuentemente. Para las tablas de contingencia mayores, es cómodo utilizar un programa informático para calcular el estadístico del contraste y el *p*-valor.

16.1. Contrastes de la bondad del ajuste: probabilidades especificadas

Ilustramos el contraste más sencillo de este tipo con un estudio en que se observó una muestra aleatoria de 33 sujetos que compraban una bebida refrescante. De estos sujetos, 8 seleccionaron la marca A, 10 seleccionaron la marca B y el resto seleccionó la marca C. Esta información se muestra en la Tabla 16.1.

Tabla 16.1. Selección de una marca.

Categoría (marca)	A	B	C	Total
Número de sujetos	8	10	15	33

En términos más generales, consideremos una muestra aleatoria de n observaciones que pueden clasificarse en K categorías. Si el número de observaciones que hay en cada categoría es O_1, O_2, ..., O_K, la clasificación es la que muestra la Tabla 16.2.

Tabla 16.2. Clasificación de n observaciones en K categorías.

Categoría	1	2	⋯	K	Total
Número de observaciones	O_1	O_2	⋯	O_K	n

Los datos muestrales se utilizan para contrastar una hipótesis nula que especifica las probabilidades de que una observación pertenezca a cada una de las categorías. En el ejemplo de los 33 sujetos que compran una bebida refrescante, la hipótesis nula (H_0) podría ser que un sujeto elegido aleatoriamente tiene las mismas probabilidades de seleccionar cualquiera de las tres variedades. Esta hipótesis nula especifica, pues, que la probabilidad de que una observación muestral pertenezca a una de las tres categorías es de un tercio. Para contrastar esta hipótesis, es lógico comparar el número *observado* con el que se *esperaría* si la hipótesis nula fuera verdadera. Dado un total de 33 observaciones muestrales, el número esperado de sujetos en cada categoría si se cumple la hipótesis nula sería $(33)(1/3) = 11$. La Tabla 16.3 resume esta información.

Tabla 16.3. Número observado y esperado de compras de tres marcas de bebidas refrescantes.

Categoría (marca)	A	B	C	Total
Número observado de sujetos	8	10	15	33
Probabilidad (según H_0)	1/3	1/3	1/3	1
Número esperado de sujetos (según H_0)	11	11	11	33

En el caso general en el que hay K categorías, supongamos que la hipótesis nula especifica las probabilidades P_1, P_2, ..., P_K de que una observación pertenezca a las categorías. Supongamos que estas posibilidades son mutuamente excluyentes y colectivamente exhaustivas, es decir, cada observación debe pertenecer a una de las categorías y no puede

pertenecer a más de una. En este caso, las probabilidades supuestas deben sumar 1; es decir,

$$P_1 + P_2 + \cdots + P_K = 1$$

Entonces, si hay n observaciones muestrales, el número esperado en cada categoría, según la hipótesis nula, es

$$E_i = nP_i \qquad (i = 1, 2, ..., K)$$

como se muestra en la Tabla 16.4.

Tabla 16.4. Número observado y esperado en el caso de n observaciones y K categorías.

Categoría	**1**	**2**	$\cdots$	K	**Total**
Número observado	O_1	O_2	$\cdots$	O_K	n
Probabilidad (según H_0)	P_1	P_2	$\cdots$	P_K	1
Número esperado de sujetos (según H_0)	$E_1 = nP_1$	$E_2 = nP_2$	$\cdots$	$E_K = nP_K$	n

La hipótesis nula sobre la población especifica las probabilidades de que una observación muestral pertenezca a cada categoría. Las observaciones muestrales se utilizan para contrastar esta hipótesis. Si los valores muestrales observados en cada categoría son muy parecidos a los esperados si la hipótesis nula fuera verdadera, este hecho apoyaría esa hipótesis. En esas circunstancias, los datos constituyen un buen *ajuste* de la distribución de probabilidad que hemos supuesto que sigue la población. Los contrastes de la hipótesis nula se basan en una valoración del grado de ajuste y generalmente se conocen con el nombre de **contrastes de la bondad del ajuste**.

Ahora bien, para contrastar la hipótesis nula, es lógico examinar las magnitudes de las discrepancias entre lo que se observa y lo que se espera. Cuanto mayores son estas discrepancias en valor absoluto, más sospechamos de la hipótesis nula. La variable aleatoria de la Ecuación 16.1 se conoce con el nombre de variable aleatoria ji-cuadrado.

Variable aleatoria ji-cuadrado

Se selecciona una muestra aleatoria de n observaciones, cada una de las cuales puede clasificarse exactamente en una de K categorías. Supongamos que el número observado en cada categoría es O_1, O_2, ..., O_K. Si una hipótesis nula (H_0) especifica las probabilidades P_1, P_2, ..., P_K de que una observación pertenezca a cada una de estas categorías, los números esperados en las categorías, si se cumple H_0, serían

$$E_i = nP_i \qquad (i = 1, 2, ..., K)$$

Si la hipótesis nula es verdadera y el tamaño de la muestra es suficientemente grande para que los valores esperados sean al menos de 5, la variable aleatoria relacionada con

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \qquad \textbf{(16.1)}$$

se aproxima bastante bien a una distribución ji-cuadrado con $(K - 1)$ grados de libertad.

Intuitivamente, el número de grados de libertad se deduce del hecho de que las O_i deben sumar n. Por lo tanto, si se conoce el número de miembros de la muestra, n, así como el número de observaciones que pertenecen a cualquiera $(K - 1)$ de las categorías, también se conoce el número que pertenece a la K-ésima categoría. La hipótesis nula se rechazará cuando el número observado sea muy diferente del esperado, es decir, cuando los valores del estadístico de la ecuación 16.1 sean excepcionalmente altos. A continuación, se muestra el contraste de la bondad del ajuste.

Un contraste de la bondad del ajuste: probabilidades especificadas

Un contraste de la bondad del ajuste, de nivel de significación α, de H_0 frente a la hipótesis alternativa de que las probabilidades especificadas no son correctas se basa en la regla de decisión

$$\text{Rechazar } H_0 \text{ si } \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{K-1,\alpha}$$

donde $\chi^2_{K-1,\alpha}$ es el número tal que

$$P(\chi^2_{K-1} > \chi^2_{K-1,\alpha}) = \alpha$$

y la variable aleatoria χ^2_{K-1} sigue una distribución ji-cuadrado con $(K - 1)$ grados de libertad.

Para ilustrar este contraste, consideremos de nuevo los datos de la Tabla 16.3 sobre la selección de una marca. La hipótesis nula es que las probabilidades de las tres categorías son las mismas. El contraste de esta hipótesis se basa en

$$\chi^2 = \sum_{i-1}^{3} \frac{(O_i - E_i)^2}{11} + \frac{(8 - 11)^2}{11} + \frac{(10 - 11)^2}{11} + \frac{(15 - 11)^2}{11} = 2{,}36$$

Hay $K = 3$ categorías, por lo que los grados de libertad de la distribución ji-cuadrado son $K - 1 = 2$. En la Tabla 7 del apéndice vemos que

$$\chi^2_{2,0,10} = 4{,}61$$

Por lo tanto, según nuestra regla de decisión, la hipótesis nula no puede rechazarse al nivel de significación del 10 por ciento. Estos datos no contienen ninguna prueba contundente en contra de la hipótesis de que un sujeto elegido aleatoriamente tiene las mismas probabilidades de seleccionar cualquiera de las tres marcas de bebidas refrescantes.

EJEMPLO 16.1. Compañía de gas (ji-cuadrado)

Una compañía de gas, basándose en la experiencia, ha llegado a la conclusión de que al final del invierno ha cobrado el 80 por ciento de sus facturas, cobrará el 10 por ciento un mes más tarde, el 6 por ciento 2 meses más tarde y el 4 por ciento más de 2 meses más tarde. Al final de este invierno, la compañía ha comprobado una muestra aleatoria de 400 facturas y ha observado que ha cobrado 287, que cobrará 49 dentro de 1 mes, 30 dentro de 2 meses y 34 dentro de más de 2 meses. ¿Sugieren estos datos que este invierno no está siguiéndose la pauta de años anteriores?

Solución

Según la hipótesis nula de que las proporciones del presente invierno siguen la pauta histórica, las respectivas probabilidades de las cuatro categorías son 0,8, 0,1, 0,06 y 0,04. Según la hipótesis, los números esperados de facturas de cada categoría, en una muestra aleatoria de 400 facturas, serían

$$400(0{,}8) = 320 \qquad 400(0{,}1) = 40 \qquad 400(0{,}06) = 24 \qquad 400(0{,}04) = 16$$

Los números observado y esperado son

Número de meses	0	1	2	Más de 2	Total
Número observado	287	49	30	34	400
Probabilidad (según H_0)	0,80	0,10	0,06	0,04	1
Número esperado (según H_0)	320	40	24	16	400

El contraste de la hipótesis nula (H_0) se basa en

$$\chi^2 = \sum_{i-1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(287 - 320)^2}{320} + \frac{(49 - 40)^2}{40} + \frac{(30 - 24)^2}{24} + \frac{(34 - 16)^2}{16} = 27{,}17$$

Aquí hay $K = 4$ categorías, por lo que hay $K - 1 = 3$ grados de libertad. En la Tabla 7 del apéndice vemos que

$$\chi^2_{3,\,0{,}005} = 12{,}84$$

Dado que 27,178 es mucho mayor que 12,84, la hipótesis nula se rechaza claramente, incluso al nivel de significación del 0,5 por ciento. Estos datos no constituyen, desde luego, una prueba contundente para sospechar que la pauta de cobro de las facturas del gas de este año es diferente de la histórica. El examen de los números de la tabla muestra que este año hay más facturas que se cobrarán más tarde que en años anteriores.

Conviene hacer una advertencia. Las cifras utilizadas para calcular el estadístico del contraste en la ecuación 16.1 deben ser el *número observado* y el *número esperado* en cada categoría. No es correcto, por ejemplo, utilizar los porcentajes de miembros que hay en cada categoría.

EJERCICIOS

Ejercicios aplicados

16.1. Un profesor está pensando utilizar un nuevo libro para el curso de contabilidad financiera y tiene tres posibilidades: *Contabilidad financiera fácil*, *Contabilidad financiera sin lágrimas* y *Contabilidad financiera para obtener un beneficio y por placer*. Se pone en contacto con una muestra aleatoria de 60 estudiantes que ya han asistido al curso y le pide a cada uno que eche una ojeada a los tres libros y le indique cuál es el que más prefiere. La tabla muestra los resultados obtenidos. Contraste la hipótesis nula de que en esta población sus primeras preferencias están distribuidas por igual entre los tres libros.

Libro	Fácil	Sin lágrimas	Beneficio y placer
Número de primeras preferencias	17	25	18

16.2. En un estudio, se seleccionó una muestra aleatoria de 75 fondos de inversión cuyo rendimiento en el periodo 1998-2000 se encontraba en el 20 por ciento más rentable de todos los fondos. Se observó su rendimiento durante los 3 años siguientes. Suponga que en este segundo periodo 13 de los fondos de la muestra se clasificaron en el 20 por ciento más rentable, 20 en el segundo 20 por ciento, 18 en el tercer 20 por ciento, 11 en el cuarto 20 por ciento y el resto en el 20 por ciento inferior. Contraste la hipótesis nula de que un fondo del 20 por ciento más rentable en 1998-2000 seleccionado aleatoriamente tiene las mismas probabilidades de pertenecer a cada una de las cinco categorías posibles de rendimiento en los 3 años siguientes.

16.3. Una compañía de seguros quería averiguar la importancia que tenía el precio en la elección de un hospital de una zona. Pidió a una muestra aleatoria de 450 consumidores que seleccionaran una respuesta entre «ninguna importancia», «es importante» o «mucha importancia». Los números respectivos que seleccionaron estas respuestas fueron 142, 175, 133. Contraste la hipótesis nula de que un consumidor elegido aleatoriamente tiene las mismas probabilidades de seleccionar cada una de estas tres respuestas.

16.4. Los datos de producción indican que el 93 por ciento de los componentes electrónicos que se producen no tiene ningún defecto, el 5 por ciento tiene un defecto y el 2 por ciento tiene más de un defecto. En una muestra aleatoria de 500 componentes producidos en una semana, se observó que 458 no tenían ningún defecto, 30 tenían un defecto y 12 tenían más de un defecto. Contraste al nivel del 5 por ciento la hipótesis nula de que la calidad de la producción de esta semana es conforme a la pauta habitual.

16.5. Una institución benéfica solicita donaciones por teléfono. Se ha observado que el 60 por ciento de todas las personas contactadas por teléfono se niega a hacer una donación, el 30 por ciento pide más información por correo con la promesa de considerar al menos la posibilidad de donar y el 10 por ciento hace inmediatamente una donación por medio de una tarjeta de crédito. En una muestra aleatoria de 100 llamadas realizadas esta semana, 65 se negaron a donar, 31 solicitaron más información por correo y 4 hicieron inmediatamente una donación por medio de una tarjeta de crédito. Contraste al nivel del 10 por ciento la hipótesis nula de que la pauta de resultados de esta semana es similar a la de semanas anteriores.

16.6. El gerente de una universidad ha observado que el 60 por ciento de todos los estudiantes considera que los cursos son muy útiles, el 20 por ciento considera que son algo útiles y el 20 por ciento considera que son inútiles. En una muestra aleatoria de 100 estudiantes que asisten a los cursos de administración de empresas, 68 piensan que el curso en cuestión es muy útil, 68 piensan que es algo útil y 14 que es inútil. Contraste la hipótesis nula de que la distribución poblacional de los cursos de administración de empresas es la misma que la de todos los cursos.

16.7. En un supermercado se venden varios tipos de yogur. El dueño del supermercado sabe, por un estudio anterior sobre los sabores elegidos por los clientes, que el 20 por ciento pidió el sabor A, el 35 por ciento pidió el sabor B, el 18 por ciento pidió el sabor C, el 12 por ciento pidió el sabor D y el resto pidió el sabor E. Ahora el dueño, que piensa que las preferencias de los clientes han cambiado, toma una muestra aleatoria de 80 clientes y observa que 12 prefieren el A, 16 prefieren el B, 30 prefieren el C, 7 prefieren el E y el resto prefiere el D. Averigüe si las preferencias de los clientes han cambiado desde el estudio anterior.

16.8. En una encuesta de mercado reciente, se dieron a probar cinco bebidas refrescantes para averiguar si los clientes preferían alguna de ellas. Se pidió a cada persona que indicara cuál era su bebida favorita. Los resultados fueron los siguientes: bebida A, 20; bebida B, 25; bebida C, 28; bebida D, 15, y bebida E, 27. ¿Existe una preferencia por alguna de estas bebidas refrescantes?

16.9. Un equipo de estudiantes de marketing debía averiguar qué pizza gustaba más a los estudiantes matriculados en su universidad. Hace dos años, se hizo un estudio parecido y se observó que el 40 por ciento de todos los estudiantes de esta universidad prefería la pizza de Bellini, el 25 por ciento prefería la pizza de Anthony, el 20 por ciento prefería la pizza de Ferrara y el resto la pizza de Marie. Para ver si han cambiado las preferencias, se seleccionaron aleatoriamente 180 estudiantes y se les pidió que indicaran sus preferencias en el caso de la pizza. Los resultados fueron los siguientes: 40 seleccionaron la pizza de Ferrara, 32 seleccionaron la de Marie, 80 seleccionaron la de Bellini y el resto seleccionó la de Anthony. ¿Indican los datos que las preferencias han cambiado desde el estudio anterior?

16.10. Se ha pedido a una muestra aleatoria de profesores de estadística que hagan una encuesta con preguntas sobre el contenido del plan de estudios, la integración del uso de computadores y las preferencias por los programas informáticos. De las 250 respuestas, 100 profesores han indicado que prefieren el paquete estadístico M y 80 el programa informático E, mientras que el resto está repartido por igual entre el programa informático S y el P. ¿Indican los datos que los profesores prefieren alguno de estos programas informáticos?

16.2. Contrastes de la bondad del ajuste: parámetros poblacionales desconocidos

En el apartado 16.1 hemos contrastado la hipótesis de que los datos son generados por una distribución de probabilidad *totalmente especificada*. En este contraste, la hipótesis nula especifica la probabilidad de que una observación muestral pertenezca a cualquiera de las categorías. Sin embargo, a menudo hay que contrastar la hipótesis de que los datos son generados por alguna distribución, como la binomial, la distribución de Poisson o la normal, sin suponer que se conocen los parámetros de esa distribución. En estas circunstancias, no puede aplicarse el apartado 16.1, pero pueden utilizarse los datos de que se dispone para estimar los parámetros poblacionales desconocidos. A continuación, formulamos el contraste de la bondad del ajuste que se utiliza cuando se estiman parámetros poblacionales.

Contrastes de la bondad del ajuste cuando se estiman parámetros poblacionales

Supongamos que una hipótesis nula especifica las probabilidades de diferentes categorías que dependen de la estimación (a partir de los datos) de m parámetros poblacionales desconocidos. El **contraste de la bondad del ajuste cuando se estiman parámetros poblacionales** es precisamente el del apartado 16.1, con la salvedad de que el número de grados de libertad de la variable aleatoria ji-cuadrado es

$$\text{Grados de libertad} = (K - m - 1) \tag{16.2}$$

donde K es el número de categorías.

Consideremos un contraste para averiguar si los datos son generados por la distribución de Poisson. Un método para intentar resolver las cuestiones relacionadas con los conflictos sobre la autoría de un texto es contar el número de veces que aparecen determinadas palabras en distintos párrafos del texto y compararlo con los resultados de pasajes cuyo autor se conoce; a menudo puede realizarse esta comparación suponiendo que el número de veces que aparecen determinadas palabras sigue una distribución de Poisson. Un ejemplo de este tipo de investigación es el estudio de *The Federalist Papers* (véase la referencia bibliográfica 10).

EJEMPLO 16.2. Federalist Papers (ji-cuadrado)

En una muestra de 262 párrafos (cada uno de los cuales tenía alrededor de 200 palabras) de *The Federalist Papers* (véase la referencia bibliográfica 10), el número medio de veces que aparecía la palabra *may* era de 0,66. La Tabla 16.5 muestra el número de veces que aparece esta palabra en los 262 párrafos de la muestra. Contraste la hipó-

Tabla 16.5. Número de veces que aparece la palabra *may* en 262 párrafos de *The Federalist Papers*.

Número de apariciones	0	1	2	3 o más
Frecuencia observada	156	63	29	14

tesis nula de que la distribución poblacional de las veces que aparece es una distribución de Poisson, sin suponer que se conoce previamente la media de esta distribución.

Solución

Recuérdese que si la distribución de Poisson es adecuada, la probabilidad de x apariciones es

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

donde λ es el número medio de apariciones. Aunque esta media poblacional es desconocida, puede estimarse por medio de la media muestral de 0,66. En ese caso, sustituyendo λ por 0,66 es posible estimar la probabilidad de cualquier número de apariciones si se cumple la hipótesis nula de que la distribución poblacional es de Poisson. Por ejemplo, la probabilidad de dos apariciones es

$$P(2) = \frac{e^{-0,66}(0,66)^2}{2!}$$

$$= \frac{(0,5169)(0,66)^2}{2} = 0,1126$$

También pueden hallarse las probabilidades de que la palabra no aparezca nunca y de que aparezca una vez, por lo que la probabilidad de que aparezca tres veces o más es

$$P(X \geqslant 3) = 1 - P(0) - P(1) - P(2)$$

Estas probabilidades se muestran en la segunda fila de la Tabla 16.6.

Tabla 16.6. Frecuencia observada y esperada en *The Federalist Papers*.

Número de apariciones	0	1	2	3 o más	Total
Frecuencias observadas	156	63	29	14	262
Probabilidades	0,5169	0,3412	0,1126	0,0293	1
Frecuencias esperadas según H_0	135,4	89,4	29,5	7,7	262

Las frecuencias esperadas si se cumple la hipótesis nula se obtienen entonces, exactamente igual que antes, de la siguiente manera:

$$E_i = nP_i \quad (i = 1, 2, ..., K)$$

Así, por ejemplo, la frecuencia esperada de dos apariciones de la palabra *may* en 262 párrafos del texto es (262) (0,1126) = 29,5. Dado que la propia variable es un número entero, es mejor no redondear estos valores esperados a valores enteros. La fila inferior de la Tabla 16.6 muestra estas frecuencias esperadas. El estadístico del contraste es

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(156 - 135{,}4)^2}{135{,}4} + \frac{(63 - 89{,}4)^2}{89{,}4} + \frac{(29 - 29{,}5)^2}{29{,}5} + \frac{(14 - 7{,}7)^2}{7{,}7} = 16{,}0$$

Dado que hay cuatro categorías y se ha estimado un parámetro, el número aproximado de grados de libertad del contraste es 2. En la Tabla 7 del apéndice vemos que

$$\chi^2_{2,\,0{,}005} = 10{,}60$$

Por lo tanto, la hipótesis nula de que la distribución poblacional es de Poisson puede rechazarse al nivel de significación del 0,5 por ciento. Los datos son una prueba realmente contundente en contra de la hipótesis.

Para resolver el ejemplo 16.2 utilizando el programa Excel, véase el apéndice de este capítulo.

Un contraste de normalidad

La distribución normal desempeña un importante papel en estadística y tanto la validez como algunas propiedades de optimalidad de muchos métodos prácticos dependen del supuesto de que los datos muestrales siguen una distribución normal. En el Capítulo 6 analizamos representaciones gráficas de probabilidades normales para buscar pruebas de la ausencia de normalidad. En el 8 (Figuras 8.2 y 8.9) buscamos visualmente pruebas de la ausencia de normalidad averiguando si los puntos de los gráficos de los distintos cuartiles estaban «cerca» de la línea recta. A continuación, examinamos un contraste del supuesto de la normalidad adaptando el método ji-cuadrado. Este contraste es fácil de realizar y es probablemente más poderoso.

Supongamos que tenemos una muestra $X_1, X_2, ..., X_n$ de n observaciones de una población. Nuestro enfoque se basa en averiguar si estos datos reflejan dos características de la distribución normal. La primera es la simetría en torno a la media. Utilizando la información muestral, el **sesgo** de una población se estima de la siguiente manera:

$$\text{Sesgo} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}$$

donde $\bar{x}$ y s son la media y la desviación típica muestrales, respectivamente. La parte importante de esta expresión es el numerador; el denominador tiene por objeto simplemente la estandarización, de tal forma que las unidades de medición sean irrelevantes. El sesgo será positivo si una distribución está sesgada hacia la derecha, ya que el promedio de los cubos de las diferencias en torno a la media muestral es positivo. El sesgo será negativo en las distribuciones sesgadas hacia la izquierda y 0 en las distribuciones, como la normal, que son simétricas en torno a la media.

Dado que hay diferentes distribuciones simétricas, es necesaria otra característica para distinguir una distribución normal. Para calcular la varianza muestral, se utilizan los cuadrados de las diferencias en torno a la media, mientras que el sesgo se basa en el cubo de las diferencias en torno a la media. El paso lógico siguiente es observar estas diferencias elevadas a la cuarta potencia, lo que da lugar al concepto de **curtosis** muestral:

$$\text{Curtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{ns^4}$$

La curtosis es una medida del peso de las colas de una función de densidad. Se sabe que en el caso de la distribución normal la curtosis poblacional es 3.

El sesgo y la curtosis muestrales pueden calcularse a partir de los datos utilizando estas fórmulas. También se incluyen automáticamente en la salida de la mayoría de los paquetes estadísticos. Sin embargo, en los paquetes estadísticos pueden utilizarse otras fórmulas para hallar estos valores. Un contraste que tiene en cuenta tanto el sesgo como la curtosis es el *estadístico del contraste de la normalidad de Bowman-Shelton*, que se obtiene aplicando la ecuación 6.3.

Contraste de la normalidad de Bowman-Shelton

El **contraste de la normalidad de Bowman-Shelton** se basa en la cercanía del sesgo muestral a 0 y en la cercanía de la curtosis muestral a 3. El estadístico del contraste es

$$B = n\left[\frac{(\text{Sesgo})^2}{6} + \frac{(\text{Curtosis} - 3)^2}{24}\right] \qquad \textbf{(16.3)}$$

Se sabe que cuando el número de observaciones muestrales es muy grande, este estadístico tiene, si se cumple la hipótesis nula de que la distribución poblacional es normal, una distribución ji-cuadrado con 2 grados de libertad. La hipótesis nula se rechaza, por supuesto, cuando los valores del estadístico son altos.

Desgraciadamente, la ji-cuadrado como aproximación de la distribución del estadístico del contraste de Bowman-Shelton, B, sólo es buena cuando la muestra es de gran tamaño. La Tabla 16.7 muestra las diferentes relaciones entre los tamaños muestrales y los niveles de significación del 5 y el 10 por ciento. El método recomendado es calcular el estadístico, B, en la ecuación 16.3 y rechazar la hipótesis nula de la normalidad si el estadístico es superior al valor correspondiente de la Tabla 16.7.

Tabla 16.7. Puntos de significación del estadístico de Bowman-Shelton (véase la referencia bibliográfica 1).

Tamaño muestral n	Significación del 10%	Significación del 5%	Tamaño muestral n	Significación del 10%	Significación del 5%
20	2,13	3,26	200	3,48	4,43
30	2,49	3,71	250	3,54	4,51
40	2,70	3,99	300	3,68	4,60
50	2,90	4,26	400	3,76	4,74
75	3,09	4,27	500	3,91	4,82
100	3,14	4,29	800	4,32	5,46
125	3,31	4,34	∞	4,61	5,99
150	3,43	4,39			

EJEMPLO 16.3. Tasas de rendimiento (contraste de normalidad)

Supongamos que una muestra aleatoria de 300 tasas diarias de rendimiento de un contrato de futuros de cítricos tenía un sesgo de 0,0305 y una curtosis de 3,08. Contraste la hipótesis nula de que la verdadera distribución de estas tasas de rendimiento es normal.

Solución

Hallamos el estadístico de Bowman-Shelton, B:

$$B = 300\left[\frac{(0{,}0305)^2}{6} + \frac{(0{,}08)^2}{24}\right] = 0{,}1265$$

La comparación de este resultado con los puntos de significación de la Tabla 16.7 da, desde luego, pocas razones para pensar que la distribución poblacional no sea normal.

Existen otros muchos contrates de la normalidad, entre los que se encuentran el de Kolmogorov-Smirnov, el de Anderson-Darling y el de Ryan-Joiner. Estos métodos, que no se explican aquí, pueden utilizarse por medio de programas como Minitab.

EJERCICIOS

Ejercicios aplicados

16.11. Durante un periodo de 100 semanas, se observó el número semanal de averías de una máquina y se anotó en la tabla adjunta. Se observó que el número semanal medio de averías era 2,1. Contraste la hipótesis nula de que la distribución poblacional de las averías es de Poisson.

Número de averías	**0**	**1**	**2**	**3**	**4**	**5 o más**
Número de semanas	10	24	32	23	6	5

16.12. En un periodo de 100 minutos, pasó por el puesto de peaje de una autopista un total de 190 vehículos. La tabla adjunta muestra la frecuencia de llegadas por minuto en este periodo. Contraste la hipótesis nula de que la distribución poblacional es de Poisson.

Número de llegadas en minutos	**0**	**1**	**2**	**3**	**4 o más**
Frecuencia observada	10	26	35	24	5

16.13. En un estudio, se pidió a una muestra aleatoria de 50 estudiantes que estimaran la cantidad de dinero que gastaban en libros de texto en un año. Se observó que el sesgo muestral de estas cantidades era 0,83 y la curtosis muestral era 3,98. Contraste al nivel del 10 por ciento la hipótesis nula de que la distribución poblacional de las cantidades gastadas es normal.

16.14. Se tomó una muestra aleatoria de 100 mediciones de la resistencia de los componentes electrónicos producidos en una semana. El sesgo muestral era 0,63 y la curtosis muestral era 3,85. Contraste la hipótesis nula de que la distribución poblacional es normal.

16.15. Utilice el contraste de Bowman-Shelton para averiguar si las cantidades gastadas en tiendas de alimentación por una muestra aleatoria de clientes de Bishop's Supermarket sigue una distribución normal. Utilice el fichero de datos **Bishop**.

16.16. Una muestra aleatoria de 125 saldos de titulares de una tarjeta de crédito indica que el sesgo muestral es 0,55 y la curtosis muestral es 2,77. Contraste la hipótesis nula de que la distribución poblacional es normal.

16.3. Tablas de contingencia

Supongamos que se toma una muestra de una población, cuyos miembros pueden clasificarse de forma inequívoca de acuerdo con un par de atributos, A y B. Debe contrastarse la hipótesis de que no existe ninguna asociación o dependencia en la población entre la posesión del atributo A y la del atributo B. Por ejemplo, una agencia de viajes puede querer saber si hay alguna asociación entre el sexo de los clientes y el método que emplean para hacer una reserva de avión. Una empresa de auditoría puede querer examinar la relación entre la edad de las personas y el tipo de declaración de la renta que hacen. O en un estudio médico, una compañía farmacéutica puede querer saber si el éxito de un medicamento utilizado para controlar el colesterol depende del peso de la persona. Una empresa de estudios de mercado puede averiguar si la elección de los cereales por parte de un cliente depende de alguna manera del color de la caja de cereales. Quizá existe una relación entre la afiliación política y el apoyo a una enmienda que se va a someter a votación en las próximas elecciones.

Supongamos que hay r categorías en A y c categorías en B, por lo que es posible hacer un total de rc cruces de categorías. El número de observaciones muestrales que pertenecen tanto a la i-ésima categoría de A como a la j-ésima categoría de B se representa por medio de O_{ij}, donde $i = 1, 2, ..., r$ y $j = 1, 2, ..., c$. La Tabla 16.8 se llama tabla de contingencia $r \times c$. Por comodidad, hemos añadido en ella los totales de las filas y de las columnas y los representamos, respectivamente, por medio de $R_1, R_2, ..., R_r$ y $C_1, C_2, ..., C_c$.

Tabla 16.8. Clasificación cruzada de n observaciones en una tabla de contingencia $r \times c$.

			Atributo B		
Atributo A	1	2	⋯	c	Total
1	O_{11}	O_{12}	⋯	O_{1c}	R_1
2	O_{21}	O_{22}	⋯	O_{2c}	R_2
⋮	⋮	⋮	⋯	⋮	⋮
r	O_{r1}	O_{r2}	⋯	O_{rc}	R_r
Total	C_1	C_2	⋯	C_c	n

Para contrastar la hipótesis nula de que no existe ninguna asociación entre los atributos A y B, preguntamos cuántas observaciones esperaríamos encontrar en cada cruce de categorías si esa hipótesis fuera verdadera. Esta pregunta tiene sentido cuando los totales de las filas y de las columnas son *fijos*. Consideremos la clasificación conjunta correspondiente a la i-ésima fila y la j-ésima columna de la tabla. Hay un total de C_j observaciones en la j-ésima columna y, suponiendo que no existe ninguna asociación, sería de esperar que cada uno de estos totales de las columnas estuviera distribuido entre las filas en proporción al número total de observaciones de cada i-ésima fila. Por lo tanto, sería de esperar que una proporción R_i/n de estas C_j observaciones estuviera en la i-ésima fila. Por consiguiente, el número esperado estimado de observaciones en cada una de las categorías del cruce es

$$E_{ij} = \frac{R_i C_j}{n} \quad \text{para } (i = 1, 2, ..., r; j = 1, 2, ..., c)$$

donde R_i y C_j son los totales de las filas y de las columnas.

Nuestro contraste de la hipótesis nula de que no existe ninguna asociación se basa en las magnitudes de las diferencias entre los números observados y los que serían de esperar si esa hipótesis fuera verdadera. La variable aleatoria de la ecuación 16.4 es una versión generalizada de la que hemos introducido en el apartado 16.1.

Variable aleatoria ji-cuadrado en el caso de tablas de contingencia

Puede demostrarse que si se cumple la hipótesis nula, la variable aleatoria relacionada con

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad \textbf{(16.4)}$$

se aproxima bastante a una distribución ji-cuadrado con $(r-1)(c-1)$ grados de libertad. La aproximación funciona bien si no más del 20 por ciento de los números esperados estimados E_{ij} es de menos de 5. A veces pueden agregarse clases contiguas para satisfacer este supuesto.

El doble sumatorio de la ecuación 16.4 implica que el sumatorio abarca todas las rc casillas de la tabla. El número de grados de libertad se debe a que los totales de las filas y de las columnas son fijos. Si éstos se conocen y también se conocen las $(r-1)$ $(c-1)$ entradas correspondientes a las $(r-1)$ primeras filas y $(c-1)$ primeras columnas, es posible deducir el resto de las entradas de la tabla. Es evidente que la hipótesis nula de la ausencia de una asociación se rechazará en el caso de que las grandes diferencias absolutas entre los números observados y los esperados sean grandes, es decir, en el caso de los valores altos del estadístico de la ecuación 16.4. A continuación, se resume el método de contraste.

Un contraste de asociación en las tablas de contingencia

Supongamos que se clasifica una muestra de n observaciones según dos atributos en una tabla de contingencia $r \times c$. Sea O_{ij} el número de observaciones que hay en la casilla que está en la i-ésima fila y la j-ésima columna. Si la hipótesis nula es

$$H_0: \text{no existe ninguna asociación entre los dos atributos en la población}$$

el número esperado estimado de observaciones en cada casilla, si se cumple H_0, es

$$E_{ij} = \frac{R_i C_j}{n} \qquad \textbf{(16.5)}$$

donde R_i y C_j son los totales de las filas y de las columnas. **Un contraste de asociación** a un nivel de significación α se basa en la siguiente regla de decisión:

$$\text{Rechazar } H_0 \text{ si } \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)(c-1),\,\alpha}$$

EJEMPLO 16.4. American Traveler Survey (contraste de asociación)

La American Traveler Survey de 1999 realizada por Plog Research Inc. suministra información basada en una muestra aleatoria de 10.536 adultos (de 18 años o más) estadounidenses sobre sus hábitos en los viajes por motivos de negocios y por placer, el uso

de la tecnología y las pautas de gasto en los viajes y una comparación de los hábitos de los que recurren a agencias de viajes con los de los que no recurren a ellas (véase la referencia bibliográfica 6). Supongamos que en un estudio parecido una agencia de viajes tomara una muestra aleatoria de individuos de su mercado para averiguar si existe alguna asociación entre el sexo de los encuestados y los métodos utilizados por ellos para hacer reservas de avión para su último viaje de placer, ya sea nacional o internacional. La Tabla 16.9 muestra los números de observaciones de cada uno de los seis cruces posibles. Por comodidad, también se indican los totales de las filas y de las columnas. Contraste la hipótesis nula de que no existe ninguna asociación entre estos atributos, en este caso, que no existe ninguna asociación entre el sexo de los sujetos y el método utilizado para hacer reservas de avión.

Tabla 16.9. Reservas de avión por sexo y método de reserva.

Método de reserva	Mujeres	Hombres	Total
Agencia de viajes	256	74	330
Internet	41	42	83
Número de teléfono gratuito de la compañía aérea	66	34	100
Total	**363**	**150**	**513**

Solución

La hipótesis nula que se contrasta implica que en la población la proporción de reservas de avión que hace el cliente a través de una agencia de viajes, la que hace por Internet o la que hace llamando al número gratuito de una compañía aérea sería la misma independientemente de que fuera hombre o mujer. Para contrastar la hipótesis nula de que no existe ninguna asociación entre los atributos, nos preguntamos cuántas observaciones *esperaríamos* encontrar en un cruce de categorías si esa hipótesis fuera verdadera.

Por ejemplo, si no existiera ninguna asociación entre el sexo y el método utilizado para hacer una reserva de avión en la Tabla 16.9, esperaríamos, dado que 363 de las 513 reservas fueron realizadas por mujeres, que una proporción de 363/513 de las 330 reservas realizadas a través de una agencia de viajes se debiera a mujeres; es decir,

$$E_{11} = \frac{(330)(363)}{513} = 233{,}5$$

Los demás números esperados se calculan de la misma forma y se muestran en la Tabla 16.10 junto con los números observados correspondientes.

Tabla 16.10. Número observado (y esperado) en cada cruce de categorías.

Método de reserva	Mujeres	Hombres	Total
Agencia de viajes	256 (233,5)	74 (96,5)	330
Internet	41 (58,7)	42 (24,3)	83
Número de teléfono gratuito de la compañía aérea	66 (70,8)	34 (29,2)	100
Total	**363**	**150**	**513**

El contraste de la hipótesis nula de que no existe ninguna asociación se basa en las magnitudes de las diferencias entre los números observados y los que se esperarían si esa hipótesis fuera verdadera. Extendiendo la ecuación 16.1 para incluir cada uno de los seis cruces de categorías, obtenemos el valor del estadístico del contraste de la ji-cuadrado:

$$\chi^2 = \frac{(256 - 233{,}5)^2}{233{,}5} + \frac{(74 - 96{,}5)^2}{96{,}5} + \frac{(41 - 58{,}7)^2}{58{,}7} + \frac{(42 - 24{,}3)^2}{24{,}3}$$
$$+ \frac{(66 - 70{,}8)^2}{70{,}8} + \frac{(34 - 29{,}2)^2}{29{,}2} = 26{,}8$$

Los grados de libertad son $(r - 1)\,(c - 1)$. Aquí, hay $r = 3$ filas y $c = 2$ columnas en la tabla, por lo que el número correcto de grados de libertad es

$$(r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

Vemos en la Tabla 7 del apéndice que

$$\chi^2_{2,\,0{,}005} = 10{,}60$$

Por lo tanto, se rechaza claramente la hipótesis nula de que no existe ninguna asociación, incluso al nivel del 0,5 por ciento. Las pruebas en contra de esta hipótesis son abrumadoras.

Debe señalarse que, al igual que en el caso de los contrastes de la bondad del ajuste de los apartados anteriores, las cifras utilizadas para calcular el estadístico deben ser los *números efectivos* observados y no, por ejemplo, los porcentajes del total.

Aplicaciones informáticas

Las organizaciones profesionales de investigación utilizan diversos programas informáticos para los tipos de métodos analizados en este capítulo. El ejemplo 16.5 ilustra el Minitab en el caso de un estudio sobre una biblioteca universitaria.

EJEMPLO 16.5. Estudio sobre una biblioteca: curso frente a variedad (Minitab)

Library

Un equipo de estudiantes recibió el encargo de realizar una encuesta en su campus universitario. Se les pidió que realizaran un informe sobre su biblioteca: ¿debe ampliarse el horario de apertura? ¿Es fácil localizar los libros en la biblioteca? ¿Existen suficientes bases de datos para investigar? ¿Está al día la tecnología? Los resultados se encuentran en el fichero de datos **Library** (véase la referencia bibliográfica 14). ¿Existe alguna asociación entre el curso en el que se encuentran los estudiantes (1: primer año; 2: segundo año; 3: tercer año; 4: cuarto año) y las respuestas a la pregunta «¿Tiene la biblioteca una variedad suficiente de libros? 1: sí; 2: no».

Solución

En el fichero de datos **Library**, vemos que un total de 355 estudiantes respondió a ambas preguntas. La Figura 16.1 muestra la salida Minitab del cruce de las respuestas. Cada uno de los valores esperados es superior a 5. Si este supuesto no fuera válido, aparecería un mensaje de advertencia en la salida Minitab y podrían agregarse clases contiguas. El bajo *p*-valor indica el rechazo de la hipótesis nula de que no existe ninguna asociación.

Tabulated Statistics: Class Rank, Adequate Variety

```
Rows: Class Rank      Columns: Adequate Variety

          No      Yes      All

1         73       71      144
       54.76    89.24   144.00

2         26       76      102
       38.79    63.21   102.00

3         19       47       66
       25.10    40.90    66.00

4         17       26       43
       16.35    26.65    43.00

All      135      220      355
      135.00   220.00   355.00

Chi-Square = 19.040, DF = 3, P-Value = 0.000
```

Figura 16.1. Curso frente a suficiente variedad (salida Minitab).

Aunque el uso del contraste ji-cuadrado de asociación indique que existe una relación entre dos variables, este método no indica el sentido o el grado de relación.

EJERCICIOS

Ejercicios básicos

16.17. ¿Fomentan los programas de televisión por cable libres de anuncios la ciudadanía en los niños en edad escolar? Véase la referencia bibliográfica 7. Muchos profesores y autoridades creen que el uso de programas de televisión por cable sin anuncios puede aumentar el interés del estudiante por el proceso democrático en los años anteriores a la edad de votar. Otros educadores piensan que la televisión es el enemigo de la educación. Suponga que en un estudio realizado en Texas, se preguntó a una muestra aleatoria de 150 profesores de historia de enseñanza secundaria «¿Le gustaría utilizar programas de televisión por cable sin anuncios en su clase?». La tabla de contingencia adjunta indica las respuestas de los profesores a esta pregunta, así como sus opiniones sobre si esa programación mejora la ciudadanía. ¿Existen pruebas de la presencia de una relación entre las respuestas a estas dos preguntas?

	¿Uso de programas de TV por cable sin anuncios?	
Efecto	**Sí**	**No**
Fomenta la ciudadanía	78	25
No fomenta la ciudadanía	37	10

16.18. Las autoridades universitarias han recogido la siguiente información sobre la calificación media de los estudiantes y la facultad del estudiante.

Facultad	Calificación media < 3,0	Calificación media 3,0 o más
Letras	50	35
Administración de empresas	45	30
Música	15	25

Averigüe si existe alguna relación entre la calificación media y la facultad.

16.19. ¿Debe obligarse a todos los estudiantes universitarios a tener computador portátil? Una escuela de administración de empresas ha encuestado recientemente a sus estudiantes para averiguar su reacción a esta política. Las respuestas se encuentran en la tabla adjunta, junto con la especialidad del estudiante.

	¿Obligar a tener computador portátil?	
Especialidad	**Sí**	**No**
Contabilidad	68	42
Economía financiera	40	15
Dirección de empresas	60	50
Marketing	30	25

¿Indican los datos que existe una asociación entre la especialidad del estudiante y la respuesta a esta pregunta?

16.20. ¿Cómo se enteran los clientes de la existencia de un nuevo producto? Se ha encuestado a una muestra aleatoria de 200 usuarios de un nuevo producto para averiguarlo. También se han recogido otros datos demográficos como la edad. Los encuestados eran 50 personas de menos de 21 años y 90 de entre 21 y 35; el resto tenía más de 35 años. El 60 por ciento de las personas de menos de 21 años había oído hablar del producto a un amigo y el resto había visto un anuncio en la prensa. Un tercio de las personas del grupo de edad 21-35 había visto el anuncio en la prensa. Los otros dos tercios habían oído hablar a un amigo. Sólo el 30 por ciento de las personas de más de 35 años había oído hablar a un amigo, mientras que el resto había visto el anuncio en la prensa. Elabore una tabla de contingencia para las variables edad y forma de enterarse de la existencia del producto. ¿Existe una asociación entre la edad del consumidor y el método por el que se enteró de la existencia del nuevo producto?

16.21. Tras un debate electoral entre dos candidatos, se preguntó a la gente por el sentido de su voto en las siguientes elecciones. ¿Existe alguna asociación entre el sexo del encuestado y la elección del candidato presidencial?

	Sexo	
Preferencia por candidato	**Hombre**	**Mujer**
Candidato A	150	130
Candidato B	100	120

RESUMEN

En este capítulo hemos estudiado algunas de las aplicaciones de la distribución ji-cuadrado. Hemos utilizado contrastes de la bondad del ajuste para contrastar la hipótesis de que los datos son generados por una distribución poblacional totalmente especificada. Esta técnica es utilizada a menudo por los analistas de mercado para averiguar si los clientes prefieren por igual los productos o para averiguar si las cuotas de mercado de varias marcas de un producto han cambiado en un determinado periodo de tiempo.

También hemos utilizado el método de la bondad del ajuste para averiguar si los datos son generados por alguna distribución, como la binomial, la distribución de Poisson o la distribución normal, sin suponer que se conocen los parámetros de esa distribución. Hemos presentado el contraste de normalidad de Bowman-Shelton. También pueden realizarse otros contrastes de normalidad con diversos paquetes estadísticos.

Por último, hemos analizado los contrastes de asociación entre dos variables. En el caso de grandes tablas de contingencia, es cómodo utilizar un paquete estadístico para hallar el estadístico del contraste y el *p*-valor.

TÉRMINOS CLAVE

contraste de asociación, 667
contraste de la bondad del ajuste, 657
contraste de la bondad del ajuste: parámetros estimados, 661
contraste de la bondad del ajuste: probabilidades especificadas, 658
contraste de normalidad de Bowman-Shelton, 664
variable aleatoria ji-cuadrado, 657
variable aleatoria ji-cuadrado: tablas de contingencia, 667

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

16.22. Suponga que se clasificara una muestra aleatoria de empresas que tienen activos insolventes según que los amorticen discrecionalmente y también según que haya o no pruebas de una fusión o adquisición posterior. Utilizando los datos de la tabla adjunta, contraste la hipótesis nula de que no existe ninguna asociación entre estos atributos.

	¿Fusión o adquisición?	
Amortizar	**Sí**	**No**
Sí	32	48
No	25	57

16.23. Un fabricante de un producto tiene tres fábricas en el país. Los defectos de este producto se deben a tres grandes causas, que podemos llamar A, B y C. Durante una semana reciente, se encontró el siguiente número de cada uno de los tres defectos en las tres fábricas:

Fábrica 1	A, 15;	B, 25;	C, 23
Fábrica 2	A, 10;	B, 12;	C, 21
Fábrica 3	A, 32;	B, 28;	C, 44

Basándonos en estas frecuencias, ¿podemos extraer la conclusión de que las pautas de defectos son las mismas en las tres fábricas?

16.24. El departamento de recursos humanos está intentando averiguar si la carrera universitaria de un empleado influye en su rendimiento. Las principales carreras examinadas son administración de empresas, economía, matemáticas y el resto. Las valoraciones del personal son excelente, bueno y medio. Las clasificaciones se basan en los empleados que tienen entre dos y cuatro años de experiencia:

Administración de empresas	excelente, 21;	bueno, 18;	medio, 10
Economía	excelente, 19;	bueno, 15;	medio, 5
Matemáticas	excelente, 10;	bueno, 5;	medio, 5
Resto	excelente, 5;	bueno, 15;	medio, 13

¿Indican estos datos que existe una diferencia entre las valoraciones basadas en la carrera estudiada?

16.25. Se ha pedido a una muestra aleatoria de personas que ocupan tres puestos de trabajo diferentes llamados A, B y C que indiquen sus preferencias por tres marcas de linternas de camping: Big Star, Lone Star y Bright Star. Las preferencias son las siguientes:

Grupo A	Big Star, 54;	Lone Star, 67;	Bright Star, 39
Grupo B	Big Star, 23;	Lone Star, 13;	Bright Star, 44
Grupo C	Big Star, 69;	Lone Star, 53;	Bright Star, 59

¿Indican estos datos que existe una diferencia entre las preferencias de los tres grupos?

16.26. Una universidad tenía interés en saber si los licenciados en historia y en economía seguían programas de doctorado diferentes. Encuestó a una muestra aleatoria de licenciados y observó que un gran número hizo un doctorado de administración de empresas, de derecho y de teología. La tabla adjunta muestra el número de veces que aparecen las personas en las distintas combinaciones. Basándose en estos resultados, ¿existe alguna prueba de que los licenciados en economía y en historia siguen programas de doctorado diferentes?

	Estudios de doctorado		
Licenciado	**Administración de empresas**	**Derecho**	**Teología**
Economía	30	20	10
Historia	6	34	20

16.27. Suponga que ha recogido datos de encuesta sobre el sexo y la compra de un producto. Realice un contraste de la ji-cuadrado para averiguar si

la probabilidad de compra varía de los hombres a las mujeres. Incluya en su respuesta los valores esperados si se cumple la hipótesis nula.

	Sexo	
Decisión	**Hombre**	**Mujer**
Compra	150	150
No compra	50	250

16.28. Sara Sánchez es una avezada directora de campaña electoral. En las elecciones primarias, hay cuatro candidatos. Desea averiguar si las preferencias de los votantes son diferentes en cuatro grandes distritos. La tabla de contingencia muestra el número de veces que aparece cada preferencia por distrito tras realizar una encuesta a una muestra aleatoria. Realice un contraste estadístico adecuado para averiguar si las preferencias por los candidatos están relacionadas con el distrito.

	Preferencia por los candidatos en las elecciones primarias				
Distrito	**A**	**B**	**C**	**D**	**Total**
1	52	34	80	34	200
2	33	15	78	24	150
3	66	54	141	39	300

16.29. Un fabricante de electrodomésticos quería saber si existía relación entre el tamaño de las familias y el tamaño de la lavadora que compraban. Estaba preparando unas directrices para el personal de ventas y quería saber si éste debía hacer recomendaciones específicas a los clientes. Se preguntó a una muestra aleatoria de 300 familias por su tamaño y por el tamaño de la lavadora. En las 40 familias en las que había una o dos personas, 25 tenían una lavadora de 5 kilos, 10 tenían una lavadora de 6 kilos y 5 tenían una lavadora de 7 kilos. En las 140 familias que tenían tres o cuatro personas, 37 tenían una lavadora de 5 kilos, 62 tenían una lavadora de 6 kilos y 41 tenían una lavadora de 7 kilos. En las 120 familias restantes en las que había cinco personas o más, 8 tenían una lavadora de 5 kilos, 53 tenían una lavadora de 6 kilos y 59 tenían una lavadora de 7 kilos. Basándose en estos resultados, ¿qué conclusiones pueden extraerse sobre el tamaño de la familia y el tamaño de la lavadora? Construya una tabla de doble entrada, formule la hipótesis, calcule el estadístico y extraiga su conclusión.

16.30. El departamento de engranajes de una gran empresa produce engranajes de gran calidad. El número que produce un mecánico por hora es 1, 2 o 3, como muestra la tabla. La dirección de la empresa está interesada en saber cómo influye la experiencia de los trabajadores en el número de unidades producidas por hora. La experiencia de los trabajadores se clasifica en tres subgrupos: 1 año o menos, entre 2 y 5 años y más de 5 años. Utilice los datos de la tabla para averiguar si la experiencia y el número de piezas producidas por hora son independientes.

	Unidades producidas por hora			
Experiencia	**1**	**2**	**3**	**Total**
$\leqslant$ 1 año	10	30	10	50
2-5 años	10	20	20	50
$>$ 5 años	10	10	30	50
Total	30	60	60	150

16.31. Ángeles Lara ha estado elaborando un plan para abrir nuevas tiendas dentro de su programa de expansión regional. En una ciudad en la que propone expandirse hay tres posibles localizaciones: norte, este y oeste. Sabe por experiencia que los tres grandes centros de beneficio de sus tiendas son las herramientas, la madera y la pintura. Para seleccionar la localización, son importantes las pautas de demanda de las diferentes partes de la ciudad. Encarga un estudio sobre la ciudad, a partir del cual elabora una tabla de doble entrada de las variables localización residencial y producto comprado. Esta tabla es realizada por el departamento de estudios de mercado utilizando datos procedentes de la muestra aleatoria de hogares de las tres grandes zonas residenciales de la ciudad. Cada zona residencial tiene un prefijo telefónico distinto y se eligen los cuatro últimos dígitos utilizando un generador informático de números aleatorios. ¿Existe una diferencia entre las pautas de demanda de los tres grandes artículos de las diferentes zonas de la ciudad?

	Demanda del producto		
Zona	**Herramientas**	**Madera**	**Pintura**
Este	100	50	50
Norte	50	95	45
Sur	65	70	75

16.32. Una empresa de mensajería está realizando un estudio de sus operaciones de envío de paquetes. En este estudio, ha recogido datos sobre el

tipo de paquete según la fuente de procedencia en un día de operaciones de una oficina del sudeste. Estos datos se muestran en la tabla. Las principales fuentes de procedencia son (1) ciudades pequeñas (ciudades), (2) barrios financieros urbanos (BFU), (3) polígonos industriales (fábricas) y (4) zonas residenciales. Existen tres grandes tipos de paquetes según su tamaño y tarifa. Los sobres urgentes deben pesar 3 kilos o menos y tienen una tarifa fija de 12 $ cualquiera que sea el destino. Los paquetes pequeños pesan entre 4 y 10 kilos y tienen limitaciones sobre su tamaño. Los paquetes grandes pueden pesar entre 11 y 75 kilos y tienen la tarifa más baja por kilo y son los que más tardan en llegar.

	Tamaño del paquete			
Fuente de procedencia	**$\leqslant 3$**	**4-10**	**11-75**	**Total**
Ciudades	40	40	20	100
BFU	119	63	18	200
Fábricas	18	71	111	200
Zonas residenciales	69	64	17	150

a) ¿Existe alguna diferencia entre las pautas de los paquetes procedentes de los diferentes lugares?

b) ¿Qué dos combinaciones tienen la mayor desviación porcentual con respecto a una pauta uniforme?

16.33. Una agencia de viajes tomó una muestra aleatoria de personas de su mercado y le hizo la siguiente pregunta: «¿Reservó su último vuelo a través de una agencia de viajes?». Cruzando las respuestas a esta pregunta con las respuestas al resto del cuestionario, la agencia obtuvo datos como los de la siguiente tabla de contingencia:

	¿Reservó su último vuelo a través de una agencia de viajes?	
Edad	**Sí**	**No**
Menos de 30	15	30
Entre 30 y 39	20	42
Entre 40 y 49	47	42
Entre 50 y 59	36	50
60 o más	45	20

Averigüe si existe relación entre la edad del encuestado y la reserva de su último vuelo a través de una agencia de viajes.

16.34. Cuando en Estados Unidos se aprobó una ley para dar el mismo estatus jurídico a las firmas electrónicas que a las manuales, casi el 60 por ciento de los propietarios de pequeñas empresas pensaba que las firmas digitales no le ayudarían a hacer negocios por Internet (véase la referencia bibliográfica 13). Suponga que se obtienen los siguientes datos en un estudio similar de propietarios de pequeñas empresas clasificadas según el número de años de antigüedad y la opinión del empresario sobre la capacidad de las firmas electrónicas de aumentar el negocio.

	¿Influirán positivamente las firmas digitales en su negocio?		
Antigüedad de la empresa	**Sí**	**No**	**No sabe**
Menos de 5 años	80	68	10
Entre 5 y 10 años	60	90	15
Más de 10 años	72	63	12

¿Existe alguna relación entre la antigüedad de la empresa y la opinión de su propietario sobre la eficacia de las firmas electrónicas?

16.35. La American Society for Quality (ASQ) ofrece a sus miembros instrumentos exclusivos de reclutamiento por Internet. «Sólo los miembros que pretenden contratar profesionales de calidad pueden anunciar sus puestos de trabajo en estos boletines gratuitos y sólo ellos tienen acceso a estos puestos de trabajo por Internet» (véase la referencia bibliográfica 2). Suponga que se toma una muestra aleatoria de empresas y se les pide que indiquen si han recurrido a una empresa de Internet para buscar empleados. También se les hace preguntas sobre la tarifa que se paga por utilizar la página. ¿Existe relación entre el uso de una página de ese tipo y la opinión de los empresarios sobre la tarifa que se paga por utilizarla?

	¿Ha recurrido a una empresa de Internet para buscar empleados?	
Tarifa	**Sí**	**No**
La tarifa es demasiado alta	36	50
La tarifa es más o menos correcta	82	28

16.36. *Business Florida* es la guía oficial del crecimiento y el desarrollo empresarial de Florida. Es publicada anualmente por Enterprise Florida

Inc.; el Florida Economic Development Council, Inc.; y la revista *Florida Trend*. En *Business Florida 2001* (véase la referencia bibliográfica 12), se dan 10 razones para animar a una empresa a seleccionar Florida «para desarrollarse y expandirse». Suponga que en un estudio de seguimiento se encuesta a una muestra aleatoria de empresas situadas en Florida en los tres últimos años. ¿Muestran los datos de la tabla de contingencia adjunta la existencia de alguna relación entre la razón principal del traslado de la empresa a Florida y el tipo de sector?

	Tipo de sector		
Razón principal	**Industria**	**Comercio al por menor**	**Turismo**
Tegnología emergente	53	25	10
Deducciones fiscales	67	36	20
Mano de obra	30	40	33

16.37. ¿Deben los grandes comercios minoristas ofrecer servicios bancarios? Los gigantes del comercio al por menor, como Nordstrom y Federated Department Stores (la empresa matriz de Macy's y Bloomingdale's), comenzaron a ofrecer algunos servicios bancarios a finales de 2000 (véase la referencia bibliográfica 3). Algunos de los incentivos para atraer a los clientes eran la posibilidad de retrasar los pagos, menores comisiones por servicios como las transferencias por cable y la concesión de préstamos para adquirir automóviles o para reformar viviendas. A los bancos pequeños les preocupa su futuro si entran más comercios minoristas en el mundo de la banca. Suponga que una empresa de estudios de mercado ha realizado una encuesta nacional para un comercio minorista que está considerando la posibilidad de ofrecer servicios bancarios a sus clientes. Pide a los encuestados que indiquen el proveedor (banco, comercio minorista, otros) al que recurrirían con mayor probabilidad para ciertos servicios bancarios (suponiendo que la tarifa no influye). ¿Existe alguna relación entre estas dos variables?

	Proveedor		
Servicio	**Banco**	**Comercio minorista**	**Otro**
Cuenta corriente	100	45	10
Cuenta de ahorro	85	25	45
Crédito hipotecario	30	10	80

16.38. Muchos productos de adelgazamiento rápido no son más que ardides publicitarios que atraen a la gente con la esperanza de adelgazar rápidamente. Los grupos de la industria de productos dietéticos, los profesionales sanitarios y las autoridades advierten de que la publicidad engañosa puede llevar a los consumidores a utilizar productos peligrosos (véase la referencia bibliográfica 4). Suponga que se pregunta a una muestra aleatoria de habitantes de una ciudad si han utilizado alguna vez un producto para adelgazar rápidamente. A continuación, se les pregunta si piensan que deben controlarse más estrictamente los anuncios para prohibir la publicidad engañosa de productos de adelgazamiento.

	¿Ha utilizado un producto de adelgazamiento rápido?	
Publicidad	**Sí**	**No**
Es necesario un control más estricto	85	40
No es necesario un control más estricto	25	64

¿Dependen las opiniones de los encuestados sobre los controles de la publicidad de que hayan utilizado o no un producto de adelgazamiento rápido?

16.39. «Nerviosas por la tambaleante bolsa de valores, las empresas en línea han iniciado lo que sin duda será una larga serie de despidos» (véase la referencia bibliográfica 5). Aunque la economía es nueva, parece que en las empresas punto.com está recurriéndose al viejo método de ajuste de plantillas. Estas empresas sostienen que los despidos son necesarios para aumentar los beneficios y ahorrar costes. Suponga que la tabla de contingencia adjunta muestra el número de despidos de tres empresas punto.com y los meses de antigüedad de los empleados despedidos. ¿Existe alguna relación entre estas dos variables?

	Empresa punto.com		
Edad	**A**	**B**	**C**
Menos de 6 meses	23	40	12
Entre 6 meses y 1 año	15	21	12
Más de 1 año	12	9	6

16.40. Algunos estudios de mercado indican el «efecto positivo de la penetración de las marcas blancas en la rentabilidad de las tiendas medida por medio de la cuota de mercado» (véase la referencia bibliográfica 8). Hace dos años, el director de un supermercado local que vende tres marcas nacionales (A, B y C) y una marca blanca (D) de zumo de naranja observó que las marcas A y C se preferían por igual; el 33 por ciento prefería la marca B, y el 27 por ciento prefería la marca blanca D. Ahora el director piensa que han cambiado las preferencias de los clientes y que la preferencia por la marca blanca ha aumentado y quizá contribuyen positivamente al aumento de los beneficios. Los resultados de una muestra aleatoria reciente de compradores indican las siguientes preferencias.

Marca favorita	A	B	C	D (marca blanca)
Número	56	70	28	126

¿Han cambiado las preferencias de los clientes desde el estudio realizado hace 2 años?

16.41. A finales del otoño de 2000, los clientes que querían servicio inalámbrico de Internet podían elegir entre cuatro categorías básicas de equipos: la agenda electrónica Palm y sus sucesoras, que utilizan el sistema operativo de Palm; la agenda Pocket-PC; los teléfonos con acceso a la Web, y los aparatos móviles de lectura del correo electrónico (véase la referencia bibliográfica 11). Analizando los datos adjuntos, procedentes de una encuesta a los usuarios de servicios inalámbricos de Internet, ¿depende la satisfacción del tipo de equipo seleccionado?

	¿Está satisfecho con su compra?	
Tipo de equipo	**Sí**	**No**
Agenda electrónica Palm	128	40
Agenda Pocket-PC	45	15
Teléfonos con acceso a la Web	30	8
Aparatos móviles de lectura del correo electrónico	30	6

16.42. En un estudio exploratorio de mercado, se pidió a los estudiantes de un campus universitario que respondieran a una breve encuesta sobre su biblioteca (véase la referencia bibliográfica 14). Una de las preguntas era si pensaban que debía ampliarse el horario de apertura de la biblioteca.

a) ¿Existe relación entre las respuestas de los estudiantes a esta pregunta y el curso en el que se encuentran? Los datos se encuentran en el fichero de datos **Library**.

b) ¿Qué recomendaciones haría al personal de la biblioteca?

16.43. ¿Puede un estudiante encontrar fácilmente los libros en la biblioteca universitaria? Esta pregunta también se incluyó en la encuesta sobre la biblioteca universitaria (véase la referencia bibliográfica 14).

a) ¿Existe relación entre las respuestas de los estudiantes a esta pregunta y el curso en el que se encuentran? Los datos se encuentran en el fichero de datos **Library**.

b) ¿Qué recomendaciones haría al personal de la biblioteca?

16.44. La Institutional Research Office (IRO) de una importante universidad realiza anualmente encuestas a los estudiantes de primer año, de segundo año y de tercer año para averiguar su nivel de satisfacción con los servicios a los estudiantes, las instalaciones y la política de la universidad. Los estudiantes de último año son encuestados por separado. Suponga que el director de la IRO de una universidad facilita a los administradores, al profesorado y al personal de la universidad análisis de las tendencias, comparaciones y otros datos útiles para mejorar continuamente la universidad.

La encuesta sobre la satisfacción de los estudiantes realizada en la primavera de 2002 entre mediados de marzo y principios de mayo se envió a una muestra aleatoria de 600 estudiantes (200 estudiantes de primer año, 200 de segundo año y 200 de tercer año). Las respuestas recibidas fueron 248, lo que supone una tasa de respuesta del 42,5 por ciento (tras ajustar las cifras para tener en cuenta las encuestas que no pudieron mandarse o no pudieron entregarse por alguna otra razón). Contenía información demográfica sobre la carrera que estaba cursando el estudiante, la edad y el sexo. Suponga que el fichero de datos **IRO** contiene alguna información de la encuesta de 2002 sobre la satisfacción de los estudiantes. Se pidió a los estudiantes que indicaran si estaban muy satisfechos, medio satisfechos o poco satisfechos con el sistema de matrícula por Internet, la librería de la universidad, el servicio de comida, la oficina de atención al estudiante, la planificación financiera de los estudiantes, el programa de estudio-trabajo y algunos otros proveedores de servicios

del campus. Con estos datos pueden investigarse numerosas relaciones. Analice los datos, seleccione y contraste varias relaciones posibles y resuma los resultados que deben presentarse al rector de la universidad. Incluya en su informe un análisis de la relación, si existe, entre la satisfacción del estudiante con el horario de apertura de la biblioteca y el curso en el que se encuentra, su nivel de satisfacción con las tutorías, la existencia de becas de investigación, la matrícula por Internet y los programas internacionales. Puede mejorar su informe por medio de medidas, gráficos y estimaciones.

16.45. Según un estudio reciente sobre el uso de los computadores (véase la referencia bibliográfica 9), «los niños de 2 a 5 años pasaban una media de 27 minutos al día en el computador, mientras que los de 6 a 11 pasaban 49 minutos al día y los de 12 a 17 pasaban una media de 63 minutos al día». Actualmente, la mayoría de las escuelas están conectadas a Internet, pero la forma en que se utilizan estos computadores en clase varía de unos centros a otros. Según Jay Becker, profesor de la Universidad de California en Irvine, «las escuelas en las que había niños pobres tendían más a hacer hincapié en las tareas de procesamiento de textos y en otras sencillas tareas, mientras que en las escuelas en las que había estudiantes de familias más acomodadas se enseñaba a usar el computador para resolver problemas y comprender mejor una asignatura». Suponga que un equipo de pedagogos bajo la dirección de Joy Haugaard realizó una encuesta para contrastar esta hipótesis. El estudio se basó en 225 escuelas tanto de comunidades pobres como de comunidades más acomodadas. La tabla adjunta indica sus respuestas a la pregunta «Por lo que se refiere al uso del computador, ¿es más probable que su escuela haga hincapié en tareas básicas como el procesamiento de textos o en la utilización del computador para resolver problemas?».

	Nivel económico	
Énfasis en el contenido	**Comunidad pobre**	**Comunidad acomodada**
Tareas básicas (procesamiento de textos)	75	40
Dominio del computador (resolución de problemas)	30	80

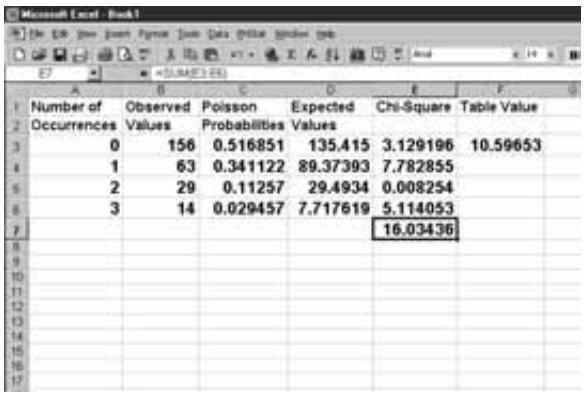
¿Coinciden los datos de este estudio con las conclusiones de Becker?

16.46. En Estados Unidos, la gente puede utilizar distintos métodos para hacer la declaración de la renta. Un método habitual es rellenar el impreso 1040. Algunas personas utilizan otros métodos como la presentación telemática. Otras piden simplemente una prórroga (ampliar la fecha de entrega de la declaración hasta después del 15 de abril). Suponga que en una localidad se realizó un estudio de 200 contribuyentes seleccionados aleatoriamente. Su edad era una importante variable en este estudio. Basándose en la distribución por edades de la población de la región, el estudio incluyó 50 personas de menos de 25 años y 90 de 25-45; el resto tenía más de 45. En el grupo de personas de menos de 25 años, 35 utilizaron un impreso 1040, 8 utilizaron otro método y el resto pidió una prórroga. Dos tercios de las personas de la categoría 25-45 años utilizaron el impreso 1040, 20 utilizaron un método diferente y el resto pidió una prórroga. El 75 por ciento de las personas de más de 45 años utilizó el impreso 1040, 4 pidieron una prórroga y el resto utilizó un método diferente. Averigüe si existe alguna asociación entre la edad de una persona y el método utilizado para hacer la declaración de la renta.

Apéndice

Podemos resolver el ejemplo 16.2 utilizando el programa Excel. Para obtener las probabilidades de Poisson mostradas en la Figura 16.2, hemos utilizado los argumentos de función, como muestra la Figura 16.3 para $x = 0$, para cada una de las ocurrencias ($x = 0$, 1, 2 y 3).

Figura 16.2. Salida Excel para averiguar si la distribución poblacional es de Poisson.



	A	B	C	D	E	F
1	Number of	Observed	Poisson	Expected	Chi-Square	Table Value
2	Occurrences	Values	Probabilities	Values		
3	0	156	0.516851	135.415	3.129196	10.59653
4	1	63	0.341122	89.37393	7.782855	
5	2	29	0.11257	29.4934	0.008254	
6	3	14	0.029457	7.717619	5.114053	
7					16.03436	

Figura 16.3. Probabilidades de Poisson obtenidas con los argumentos de función de Excel.

Bibliografía

1. Bera, A. K. y C. M. Jarque, «An Efficient Large-Sample Test for Normality of Observations and Regression Residuals», *Working Papers in Economics and Econometrics*, 40, Australian National University, 1981.
2. «Career Services Program Updated», *On Q*, American Society for Quality, 15, n.º 4, otoño, 2000.
3. Coolidge, Carrie, «Socks and Bonds», *Forbes*, 3 de julio de 2000, pág. 62.
4. «Dieter Hunger for Gimmicks», *New York Times*, artículo reproducido en *Orlando Sentinel*, 29 de octubre de 2000, p. A11.
5. Godwin, Jennifer, «New Economy, Same Old Downsizing», *Forbes*, 3 de julio de 2000, pág. 60.
6. Jamison, Jane, «Survey Highlights Agents' Strength», *Travel Weekly*, 25 de octubre de 1999, págs. 10-47.
7. Keveney, Bill, «Classroom TV Brings Election to Students: Commercial-free Cable Programs Promote Citizenship», *USA Today*, 30 de octubre de 2000, pág. 4D.
8. Lal, Rajiv y Marcel CorstJensrajiv Lal, «Building Store Loyalty Through Store Brands», *Journal of Marketing Research*, 37, n.º 3, agosto, 2000, pág. 281.
9. Lewin, Tamara, «Children's Computer Use Grows, but Gap Persist, Study Says», *New York Times*, 22 de enero de 2001, pág. A11.

10. Mosteller, F. y D. L. Wallace, *Interference and Disputed Authorship*: *The Federalist* © 1964, Addison-Wesley, Reading, Mass, Tablas 2.3 y 2.4. Premiso de reimpresión.
11. Nadeau, Michael, «Cut the Cord», *Access*: *America's Guide to the Internet*, suplemento especial de *Orlando Sentinel*, 29 de octubre de 2000, págs. 12-14. www.accessmagazine.com.
12. Shepherd, Gary, «10 Reasons Why Your Business Belongs in Florida», *Business Trend's Business Florida 2001*, www.businessflorida.com.
13. «Sign Here Please», *USA Today*, 30 de octubre de 2000, pág. 1B. www.office.com.
14. Thorne, J. Renee *et al*., «University Library Study», artículo inédito. Los datos se encuentran en el fichero de datos **Library**.

Capítulo 17

Análisis de la varianza

Esquema del capítulo

Introducción

En las aplicaciones empresariales modernas del análisis estadístico, hay algunas situaciones que requieren hacer comparaciones de procesos en más de dos niveles. Por ejemplo, al director de Circuitos Integrados S.A. le gustaría saber si cualquiera de cinco procesos para montar componentes aumenta la productividad por hora y reduce el número de componentes defectuosos. Los análisis para responder a estas cuestiones se conocen con el nombre general de diseño experimental. Un importante instrumento para organizar y analizar los datos de este experimento se llama *análisis de la varianza*, que es el tema de este capítulo. El experimento también podría extenderse a un diseño que incluyera la cuestión de cuál de cuatro fuentes de materias primas aumenta más la productividad en combinación con los diferentes métodos de producción. Esta cuestión podría responderse con un análisis de la varianza de dos factores. Por poner otro ejemplo, el presidente de una empresa de cereales tiene interés en comparar las ventas semanales de cuatro marcas diferentes en tres tiendas distintas. Una vez más, tenemos un diseño de un problema que puede analizarse utilizando el análisis de la varianza. En el apartado 14.2 mostramos que también podían utilizarse variables ficticias para analizar problemas de diseño experimental.

17.1. Comparación de las medias de varias poblaciones

En el apartado 11.1 vimos cómo se contrasta la hipótesis de la igualdad de dos medias poblacionales. De hecho, presentamos dos contrastes, que eran adecuados dependiendo del diseño experimental, es decir, del mecanismo empleado para generar las observaciones muestrales. Concretamente, nuestros contrastes partían de observaciones pareadas o de muestras aleatorias independientes. Esta distinción es importante y, para aclararla, nos detendremos a examinar un sencillo ejemplo. Supongamos que nuestro objetivo es comparar el consumo de combustible de dos tipos de automóviles: A y B. Podríamos seleccionar aleatoriamente 10 personas para que recorrieran una determinada distancia con estos automóviles, asignando a cada una un automóvil de cada tipo, de manera que cada una condujera tanto un automóvil A como un automóvil B. Las 20 cifras de consumo de combustible resultantes consistirán en 10 parejas, cada una de las cuales corresponde a un conductor. Éste es el diseño por parejas enlazadas y su atractivo reside en su capacidad para hacer una comparación entre las cantidades de interés (en este caso, el consumo de combustible de los dos tipos de automóvil), teniendo en cuenta al mismo tiempo la posible importancia de otro factor relevante (las diferencias entre los conductores). Así, si se observa la existencia de una diferencia significativa entre el comportamiento de los automóviles A y el de los B, tenemos alguna seguridad de que no se debe a diferencias de conducta de los automovilistas. Otro diseño sería tomar 20 conductores y asignar aleatoriamente 10 a los automóviles A y 10 a los automóviles B (aunque, en realidad, no es necesario hacer el mismo número de pruebas con cada tipo de automóvil). Las 20 cifras de consumo de combustible resultantes constituirían un par de muestras aleatorias independientes de 10 observaciones cada una sobre los automóviles A y B.

En el apartado 11.1 analizamos métodos adecuados para contrastar la hipótesis nula de la igualdad de un par de medias poblacionales en estos dos tipos de diseño. En este capítulo, nuestro objetivo es extender estos métodos al desarrollo de contrastes de la igualdad de la media de varias poblaciones. Supongamos, por ejemplo, que nuestro estudio incluyera un tercer tipo de automóvil, el automóvil C. La hipótesis nula de interés sería en ese caso que la media poblacional del consumo de combustible de los tres tipos de automóviles es igual. Mostramos cómo pueden realizarse contrastes de esas hipótesis, comenzando con el caso en el que se toman muestras aleatorias independientes. En el apartado 17.5 analizamos la extensión del contraste basado en datos pareados.

Supongamos que a 7 de 20 conductores se les asigna un automóvil A, a 7 un automóvil B y a 6 un automóvil C. Utilizando los datos de la Tabla 17.1, calculamos

$$\text{Media muestral de los automóviles A} = \frac{146{,}3}{7} = 20{,}9$$

$$\text{Media muestral de los automóviles B} = \frac{162{,}4}{7} = 23{,}2$$

$$\text{Media muestral de los automóviles C} = \frac{137{,}4}{6} = 22{,}9$$

Naturalmente, estas medias muestrales no son todas iguales. Sin embargo, como siempre, cuando se contrastan hipótesis, interesa saber cuál es la probabilidad de que las diferencias de ese tipo surgieran por casualidad aunque se cumpliera en realidad la hipótesis nula.

Tabla 17.1. Cifras de consumo de combustible de tres muestras aleatorias independientes en kilómetros por litro.

	Automóviles A	Automóviles B	Automóviles C
	22,2	24,6	22,7
	19,9	23,1	21,9
	20,3	22,0	23,2
	21,4	23,5	24,1
	21,2	23,6	22,1
	21,0	22,1	23,4
	20,3	23,5	—
Sumas	**146,3**	**162,4**	**137,4**

Si se llega a la conclusión de que sería muy improbable que surgieran diferencias de ese tipo por casualidad, dudaríamos considerablemente de que la hipótesis nula sea verdadera.

Para aclarar las cuestiones que plantea este análisis, consideremos la Figura 17.1, que representa dos conjuntos hipotéticos de datos. Las medias muestrales de la parte (a) de la figura son exactamente iguales que las de la parte (b). La diferencia fundamental se halla en que en la primera las observaciones están muy concentradas en torno a sus respectivas medias muestrales, mientras que en la segunda la dispersión es mucho mayor. El examen visual de la parte (a) sugiere que los datos proceden, en realidad, de tres poblaciones que tienen diferentes medias. En cambio, observando la parte (b), no nos sorprendería mucho saber que estos datos proceden de una población común.

INTERPRETACIÓN

Esta ilustración sirve para señalar la propia esencia del contraste de la igualdad de las medias poblacionales. El factor crítico es la *variabilidad* de los datos. Si la variabilidad *en torno* a las medias muestrales es pequeña en comparación con la variabilidad *entre* las medias muestrales, como en la Figura 17.1(a), nos inclinamos a dudar de la hipótesis nula de que las medias poblacionales son iguales. Si la variabilidad en torno a las medias muestrales es grande en comparación con la variabilidad entre ellas, como en la Figura 17.1(b), no hay pruebas contundentes para rechazar la hipótesis nula. Si eso es así, parece razonable esperar que el contraste se base en el valor de la varianza. Y así es, en efecto, por lo que la técnica general empleada se conoce con el nombre de análisis de la varianza.

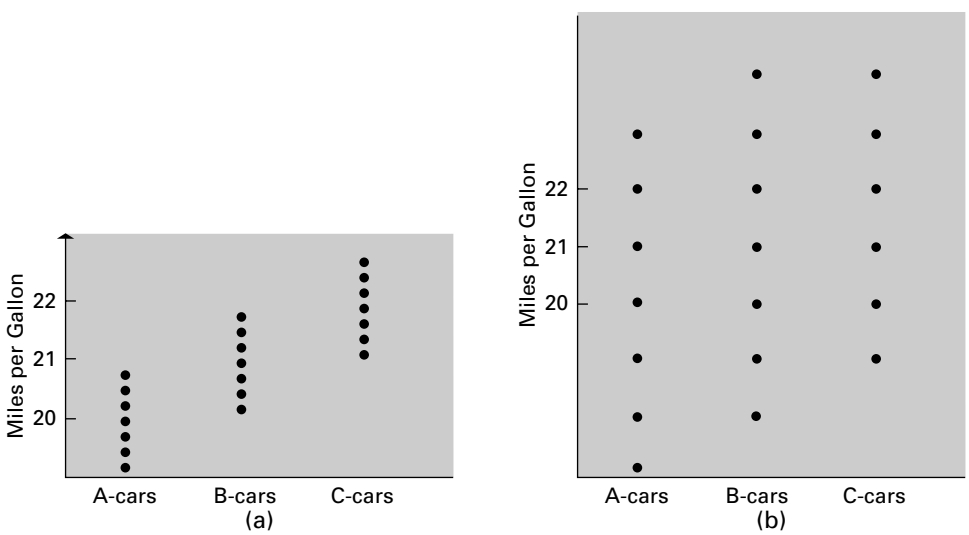


Figura 17.1. Dos conjuntos de datos muestrales sobre el consumo de combustible de tres tipos de automóvil.

17.2. Análisis de la varianza de un factor

El problema presentado en el apartado 17.1 puede tratarse de una forma bastante general. Supongamos que queremos comparar las medias de K poblaciones, *que se supone que tienen todas ellas la misma varianza*. Se toman muestras aleatorias independientes de n_1, n_2, ..., n_K observaciones de estas poblaciones. Utilizamos el símbolo x_{ij} para representar la j-ésima observación de la i-ésima población. Entonces, utilizando el formato de la Tabla 17.1, podemos presentar los datos muestrales como en la Tabla 17.2.

Tabla 17.2. Observaciones muestrales de muestras aleatorias independientes de K poblaciones.

Población			
1	2	⋯	K
x_{11}	x_{21}	⋯	x_{K1}
x_{12}	x_{22}	⋯	x_{K2}
⋮	⋮		⋮
x_{1n}	x_{2n}	⋯	x_{Kn}

El método para contrastar la igualdad de las medias poblacionales en este contexto se denomina análisis de la varianza de un factor, expresión que resultará más clara cuando examinemos otros modelos de análisis de la varianza.

El modelo para un análisis de la varianza de un factor

Supongamos que tenemos muestras aleatorias independientes de n_1, n_2, ..., n_K observaciones de K poblaciones. Si las medias poblacionales son μ_1, μ_2, ..., μ_K, el análisis de la varianza de un factor pretende contrastar la hipótesis nula

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$$
$$H_1: \mu_i \neq \mu_j \quad \text{Para al menos un par } \mu_i, \mu_j$$

En este apartado presentamos un contraste de la hipótesis nula de que las medias de K poblaciones son iguales, dadas muestras aleatorias independientes de esas poblaciones. El primer paso obvio es calcular las medias muestrales de los K grupos de observaciones. Estas medias muestrales se representan por medio de $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_K$. En términos formales,

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i = 1, 2, ..., K)$$

donde n_i representa el número de observaciones muestrales del grupo i. En esta notación, ya hemos observado con los datos de la Tabla 17.1 que

$$\bar{x}_1 = 20{,}9 \qquad \bar{x}_2 = 23{,}2 \qquad \bar{x}_3 = 22{,}9$$

Ahora bien, la hipótesis nula de interés especifica que las K poblaciones tienen una media común. Un paso lógico es, pues, estimar esa media común a partir de las observa-

ciones muestrales. Ésta es simplemente la suma de todos los valores muestrales dividida por su número total. Si n representa el número total de observaciones muestrales, entonces

$$n = \sum_{i=1}^{K} n_i$$

En nuestro ejemplo, $n = 20$. La media global de las observaciones muestrales puede expresarse entonces de la forma siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij}}{n}$$

donde el doble sumatorio indica que sumamos todas las observaciones de cada grupo y las de todos los grupos, es decir, sumamos todas las observaciones existentes. Una expresión equivalente es

$$\bar{x} = \frac{\sum_{i=1}^{K} n_i \bar{x}_i}{n}$$

En el caso de los datos de la Tabla 17.1 sobre el consumo de combustible, la media global es

$$\bar{x} = \frac{(7)(20{,}9) + (7)(23{,}2) + (6)(22{,}9)}{20} = 22{,}305$$

Por lo tanto, si, en realidad, la media poblacional del consumo de combustible de los automóviles A, B y C es igual, estimamos que la media común es de 22,31 kilómetros por litro.

Como indicamos en el apartado 17.1, el contraste de la igualdad de las medias poblacionales se basa en una comparación de dos tipos de variabilidad de los miembros de la muestra. El primero es la variabilidad en torno a las medias muestrales individuales dentro de los K grupos de observaciones. Es cómodo llamarla *variabilidad dentro de los grupos*. En segundo lugar, nos interesa la variabilidad entre las medias de los K grupos. Ésta se llama *variabilidad entre los grupos*. A continuación, buscamos medidas, basadas en los datos muestrales, de estos dos tipos de variabilidad.

Consideremos, en primer lugar, la variabilidad dentro de los grupos. Para medir la variabilidad en el primer grupo, calculamos la suma de los cuadrados de las desviaciones de las observaciones en torno a su media muestral $\bar{x}_1$, es decir,

$$SC_1 = \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2$$

Asimismo, en el caso del segundo grupo, cuya media muestral es $\bar{x}_2$, calculamos

$$SC_2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

y así sucesivamente. Así pues, la variabilidad total que existe dentro de los grupos, denominada *SCD*, es la suma de las sumas de los cuadrados de los K grupos; es decir,

$$SCD = SC_1 + SC_2 + \cdots + SC_K$$

o sea

$$SCD = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

En el caso de los datos sobre el consumo de combustible, tenemos que

$$SC_1 = (22{,}2 - 20{,}9)^2 + (19{,}9 - 20{,}9)^2 + \cdots + (20{,}3 - 20{,}9)^2 = 3{,}76$$
$$SC_2 = (24{,}6 - 23{,}2)^2 + (23{,}1 - 23{,}2)^2 + \cdots + (23{,}5 - 23{,}2)^2 = 4{,}96$$
$$SC_3 = (22{,}7 - 22{,}9)^2 + (21{,}9 - 22{,}9)^2 + \cdots + (23{,}4 - 22{,}9)^2 = 3{,}46$$

La suma de los cuadrados dentro de los grupos es, pues,

$$SCD = SC_1 + SC_2 + SC_3 = 3{,}76 + 4{,}96 + 3{,}46 = 12{,}18$$

A continuación, necesitamos una medida de la variabilidad que existe entre los grupos. Una medida lógica se basa en las diferencias entre las medias individuales de los grupos y la media global. En realidad, al igual que antes, estas diferencias se elevan al cuadrado, por lo que

$$(\bar{x}_1 - \bar{x})^2, (\bar{x}_2 - \bar{x})^2, ..., (\bar{x}_K - \bar{x})^2$$

Para calcular la suma total de los cuadrados entre los grupos, *SCG*, ponderamos cada diferencia al cuadrado por el número de observaciones muestrales del grupo correspondiente (de manera que damos más peso a las diferencias correspondientes a los grupos en los que hay más observaciones), por lo que

$$SCG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2$$

Por lo tanto, en el caso de nuestros datos sobre el consumo de combustible,

$$SCG = (7)(20{,}9 - 22{,}305)^2 + (7)(23{,}2 - 22{,}305)^2 + (6)(22{,}9 - 22{,}305)^2$$
$$= 21{,}55$$

A menudo se calcula otra suma de los cuadrados. Es la suma de los cuadrados de las diferencias de *todas* las observaciones muestrales en torno a su media *global*. Ésta se denomina *suma total de los cuadrados* y se expresa de la forma siguiente:

$$STC = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

De hecho, en el apéndice de este capítulo mostramos que la suma total de los cuadrados es la suma de los cuadrados dentro de los grupos y la suma de los cuadrados entre los grupos; es decir,

$$STC = SCD + SCG$$

Por lo tanto, en el caso de los datos sobre el consumo de combustible, tenemos que

$$STC = 12{,}18 + 21{,}55 = 33{,}73$$

Descomposición de la suma de los cuadrados en el análisis de la varianza de un factor

Supongamos que tenemos muestras aleatorias independientes de n_1, n_2, ..., n_K observaciones de K poblaciones. Sean $\bar{x}_1$, $\bar{x}_2$, ... , $\bar{x}_K$ las medias muestrales de los K grupos y $\bar{x}$ la media muestral global. Definimos las siguientes **sumas de los cuadrados**:

$$\text{Dentro de los grupos:} \quad SCD = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \tag{17.1}$$

$$\text{Entre los grupos:} \quad SCG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2 \tag{17.2}$$

$$\text{Total:} \quad STC = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 \tag{17.3}$$

donde x_{ij} representa la j-ésima observación muestral del i-ésimo grupo.

Entonces,

$$STC = SCD + SCG \tag{17.4}$$

La descomposición de la suma total de los cuadrados en la suma de dos componentes —las sumas de los cuadrados dentro de los grupos y entre los grupos— constituye la base del contraste de la igualdad de las medias poblacionales de los grupos basado en el análisis de la varianza. Podemos considerar que esta descomposición expresa la variabilidad total de todas las observaciones muestrales en torno a su media global como la suma de la variabilidad dentro de los grupos y la variabilidad entre los grupos. La Figura 17.2 lo muestra esquemáticamente.

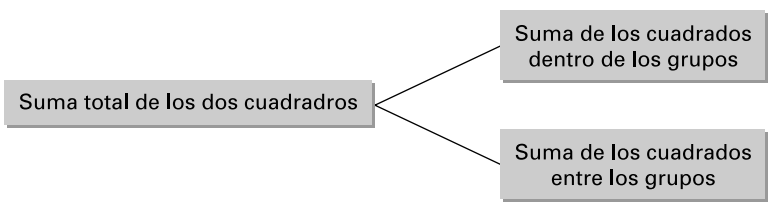


Figura 17.2. Descomposición de la suma de los cuadrados en el análisis de la varianza de un factor.

Nuestro contraste de la igualdad de las medias poblacionales se basa en el supuesto de que las K poblaciones tienen una varianza común. Si la hipótesis nula de que las medias poblacionales son iguales es verdadera, cada una de las sumas de los cuadrados, *SCD* y *SCG*, puede utilizarse como base para estimar la varianza poblacional común. Para obtener estas estimaciones, deben dividirse las sumas de los cuadrados por el número correspondiente de grados de libertad.

En primer lugar, en el apéndice del capítulo mostramos que se obtiene un estimador insesgado de la varianza poblacional si se divide *SCD* por $(n - K)$. La estimación resultante se denomina *media de los cuadrados dentro de los grupos* y se representa por medio de *MCD*, de manera que

$$MCD = \frac{SCD}{n - K}$$

En el caso de nuestros datos, tenemos que

$$MCD = \frac{12{,}18}{20 - 3} = 0{,}7165$$

Si las medias poblacionales son iguales, se obtiene otro estimador insesgado de la varianza poblacional dividiendo *SCG* por $(K - 1)$, que también se muestra en el apéndice del capítulo. La cantidad resultante se llama *media de los cuadrados entre los grupos* y se representa por medio de *MCG*; por lo tanto,

$$MCG = \frac{SCG}{K - 1}$$

En el caso de nuestros datos sobre el consumo de combustible,

$$MCG = \frac{21{,}55}{3 - 1} = 10{,}78$$

Cuando las medias poblacionales *no* son iguales, la media de los cuadrados entre los grupos *no* constituye una estimación insesgada de la varianza poblacional común. El valor esperado de la variable aleatoria correspondiente es mayor que la varianza poblacional común, ya que también contiene información sobre los cuadrados de las diferencias de las verdaderas medias poblacionales.

Si la hipótesis nula fuera verdadera, ahora tendríamos dos estimaciones insesgadas de la misma cantidad, la varianza poblacional común. Sería razonable esperar que estas estimaciones fueran muy parecidas. Cuanto mayor es la diferencia entre estas dos estimaciones, manteniéndose todo lo demás constante, mayor es nuestra sospecha de que la hipótesis nula no es verdadera. El contraste de la hipótesis nula se basa en el cociente entre las medias de los cuadrados (véase el apéndice del capítulo):

$$F = \frac{MCG}{MCD}$$

Si este cociente es cercano a 1, hay pocas razones para dudar de la hipótesis nula de la igualdad de las medias poblacionales. Sin embargo, como ya hemos señalado, si la variabilidad entre los grupos es grande en comparación con la variabilidad dentro de los grupos, sospechamos que la hipótesis nula es falsa. Lo es cuando el cociente *F* tiene un valor muy superior a 1. En ese caso, se rechaza la hipótesis nula.

Cabe deducir un contraste formal del hecho de que si la hipótesis nula de la igualdad de las medias poblacionales es verdadera, la variable aleatoria sigue una distribución *F* (analizada en el apartado 11.4) con $(K - 1)$ grados de libertad en el numerador y $(n - K)$ grados de libertad en el denominador, suponiendo que las distribuciones poblacionales son normales.

Contraste de hipótesis basado en el análisis de la varianza de un factor

Supongamos que tenemos muestras aleatorias independientes de n_1, n_2, ..., n_K observaciones de *K* poblaciones. Sea *n* el tamaño total de la muestra, de manera que

$$n = n_1 + n_2 + \cdots + n_K$$

Definimos las **medias de los cuadrados** de la forma siguiente:

$$\text{Dentro de los grupos:} \quad MCD = \frac{SCD}{n - K} \tag{17.5}$$

$$\text{Entre los grupos:} \quad MCG = \frac{SCG}{K - 1} \tag{17.6}$$

La hipótesis nula que se contrasta es que las K medias poblacionales son iguales; es decir,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$

Postulamos los siguientes supuestos adicionales:

1. Las varianzas poblacionales son iguales.
2. Las distribuciones poblacionales son normales.

La regla de decisión de un contraste de nivel de significación α es:

$$\text{Rechazar } H_0 \text{ si} \quad \frac{MCG}{MCD} > F_{K-1,\,n-K,\,\alpha} \tag{17.7}$$

donde $F_{K-1,\,n-K,\,\alpha}$ es el número para el que

$$P(F_{K-1,\,n-K} > F_{K-1,\,n-K,\,\alpha}) = \alpha$$

y la variable aleatoria $F_{K-1,\,n-K}$ sigue una distribución F con $(K - 1)$ grados de libertad en el numerador y $(n - K)$ grados de libertad en el denominador.

El p-valor de este contraste es el grado más bajo de significación que nos permitiría rechazar la hipótesis nula.

En el caso de los datos sobre el consumo de combustible, observamos que

$$\frac{MCG}{MCD} = \frac{10{,}78}{0{,}7165} = 15{,}04$$

Los grados de libertad del numerador y del denominador son, respectivamente, $(K - 1) = 2$ y $(n - K) = 17$. Por lo tanto, para un contraste al nivel de significación del 1 por ciento, vemos que en la Tabla 9 del apéndice,

$$F_{2,\,17,\,0{,}01} = 6{,}11$$

Por lo tanto, estos datos nos permiten rechazar al nivel de significación del 1 por ciento la hipótesis nula de que la media poblacional del consumo de combustible de los tres tipos de automóvil es igual.

Es muy cómodo resumir los cálculos realizados para hacer este contraste en una **tabla del análisis de la varianza de un factor**. La forma general de la tabla se muestra en la Tabla 17.3. La 17.4 contiene el análisis de la varianza correspondiente a los datos sobre el consumo de combustible. Obsérvese que, en algunas exposiciones, la suma de los cuadrados dentro de los grupos se denomina *suma de los cuadrados de los errores*.

Tabla 17.3. Formato general de la tabla del análisis de la varianza de un factor.

Fuente de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente F
Entre los grupos	SCG	$K-1$	$MCG = \dfrac{SCG}{K-1}$	$\dfrac{MCG}{MCD}$
Dentro de los grupos	SCD	$n-K$	$MCD = \dfrac{SCD}{n-K}$	
Total	STC	$n-1$		

Tabla 17.4. Tabla del análisis de la varianza de un factor correspondiente a los datos sobre el consumo de combustible.

Fuente de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente F
Entre los grupos	21,55	2	10,78	15,04
Dentro de los grupos	12,18	17	0,7165	
Total	33,73	19		

EJEMPLO 17.1. Dificultades para leer los anuncios de las revistas (análisis de la varianza de un factor)

El *índice fog* se utiliza para medir la dificultad para leer un texto escrito: cuanto más alto es el valor del índice, más difícil es el nivel de lectura. Queremos saber si las tres revistas *Scientific American*, *Fortune* y *New Yorker* tienen un índice distinto de dificultad de lectura.

Solución

Se toman muestras aleatorias independientes de 6 anuncios de *Scientific American*, *Fortune* y *New Yorker*, se miden los *índices fog* de los 18 anuncios y se anotan en la Tabla 17.5 (véase la referencia bibliográfica 2).

Tabla 17.5. *Índice fog* de la dificultad de lectura de tres revistas.

Scientific American	*Fortune*	*New Yorker*
15,75	12,63	9,27
11,55	11,46	8,28
11,16	10,77	8,15
9,92	9,93	6,37
9,23	9,87	6,37
8,20	9,42	5,66

Basándose en estos datos, se puede elaborar la tabla del análisis de la varianza utilizando un programa estadístico como Minitab. La Figura 17.3 contiene la salida del análisis de la varianza. Para contrastar la hipótesis nula de que las medias poblacionales de los *índices fog* son iguales, debemos comparar el cociente —$F - F = 6{,}97$— de la tabla

del análisis de la varianza con los valores tabulados de la distribución F con (2, 15) grados de libertad. En la Tabla 9 del apéndice vemos que

$$F_{2,\,15,\,0,01} = 6{,}36$$

INTERPRETACIÓN

Por lo tanto, rechazamos la hipótesis nula de la igualdad de las medias poblacionales de los *índices fog* de las tres revistas al nivel de significación del 1 por ciento. Obsérvese también que el *p*-valor calculado, como se ve en la Figura 17.3, es 0,007. Tenemos pruebas contundentes de que la dificultad de lectura es diferente: el índice más bajo corresponde a *New Yorker*. Obsérvese que la salida Minitab contiene una representación gráfica de las medias de los subgrupos y sus intervalos de confianza. Esta salida contiene una presentación visual de las diferencias entre las medias de los subgrupos, señalando en este caso que *New Yorker* se diferencia mucho de *Scientific American* y *Fortune*.

```
One-way ANOV A: S CIENTIFIC AM ERICAN,  FORTUNE, N EW Y ORKER

Source    DF       SS     MS      F       P
Factor     2    48.53  24.26   6.97   0.007
Error     15    52.22   3.48
Total     17   100.75

S = 1.866    R-Sq = 48.17%     R-Sq(adj) = 41.26%

                                        Individual 95% CIs For Mean Based on
                                        Pooled StDev
Level               N     Mean   StDev  -+--------+--------+--------+----
SCIENTIFIC AMERI    6   10.968   2.647                    (------*------)
FORTUNE             6   10.680   1.202                  (------*-------)
NEW YORKER          6    7.350   1.412   (------*------)
                                         -+--------+--------+--------+----
                                         6.0        8.0       10.0       12.0

Pooled StDev = 1.866
```

Figura 17.3. Análisis de la varianza de un factor de la dificultad de lectura de *Scientific American*, *Fortune* y *New Yorker* (salida Minitab).

Modelo poblacional en el caso del análisis de la varianza de un factor

Es útil observar el modelo del análisis de la varianza de un factor desde una perspectiva diferente. Sea la variable aleatoria X_{ij} la j-ésima observación de la i-ésima población y μ_i la media de esta población. En ese caso, X_{ij} puede concebirse como la suma de dos partes: su media y una variable aleatoria ε_{ij} de media 0. Por lo tanto, podemos escribir

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

Ahora bien, como se toman muestras aleatorias independientes, las variables aleatorias ε_{ij} no están correlacionadas entre sí. Además, dado nuestro supuesto de que las varianzas poblacionales son iguales, se deduce que las ε_{ij} tienen todas ellas las mismas varianzas. Por lo tanto, estas variables aleatorias satisfacen los supuestos habituales (véase el apartado 13.3) impuestos a los términos de error de un modelo de regresión múltiple. Esta ecuación puede

verse como un modelo de regresión con los parámetros desconocidos μ_1, μ_2, ..., μ_K. La hipótesis nula de interés es

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$$

El supuesto añadido de la normalidad facilita el contraste de estos parámetros.

El modelo puede expresarse de una forma algo distinta. Sea μ la media global de las K poblaciones combinadas y G_i la diferencia entre la media poblacional del i-ésimo grupo y esta media global, de manera que

$$G_i = \mu_i - \mu \quad \text{o} \quad \mu_i = \mu + G_i$$

Sustituyendo en la ecuación original, tenemos que

$$X_{ij} = \mu + G_i + \varepsilon_{ij}$$

por lo que una observación está formada por la suma de una media global μ, un término específico del grupo G_i y un error aleatorio ε_{ij}. Entonces, nuestra hipótesis nula es que cada media poblacional μ_i es igual que la media global, o sea

$$H_0: G_1 = G_2 = \cdots = G_K = 0$$

La Figura 17.4 muestra este modelo poblacional y algunos de los supuestos. El consumo efectivo de combustible de cada tipo de automóvil registrado en una prueba cualquiera puede representarse por medio de una variable aleatoria que sigue una distribución normal. Las medias poblacionales del consumo medio de combustible de los automóviles A, B y C, μ_1, μ_2 y μ_3, respectivamente, determinan los centros de estas distribuciones. Según nuestro supuesto, estas distribuciones poblacionales deben tener las mismas varianzas. La Figura 17.4 también muestra la media μ de las tres poblaciones combinadas y las diferencias G_j entre las medias poblacionales individuales y la media global. Por último, en el caso de los automóviles B, hemos marcado con un punto la i-ésima observación muestral. La variable aleatoria ε_{ij} es, pues, la diferencia entre el valor observado y la media de la subpoblación j de la que pro^{M}cede.

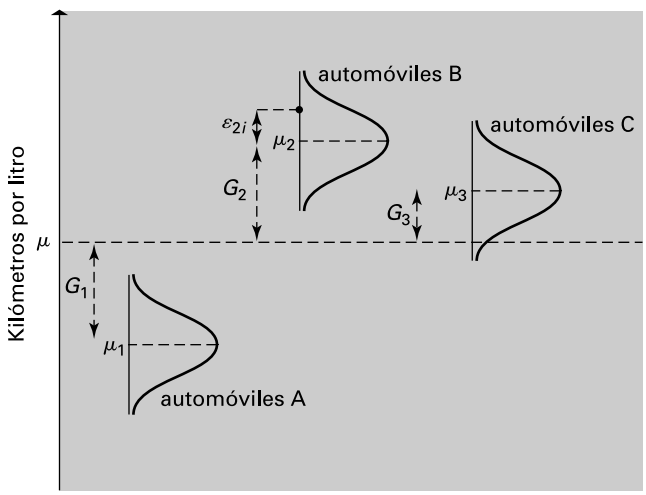


Figura 17.4. Ilustración del modelo poblacional en el caso del análisis de la varianza de un factor.

EJERCICIOS

Ejercicios básicos

17.1. Dada la siguiente tabla del análisis de la varianza:

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre los grupos	1.000	4
Dentro de los grupos	750	15
Total	1.750	19

calcule los cuadrados medios entre los grupos y dentro de los grupos. Calcule el cociente F y contraste la hipótesis de que las medias de los grupos son iguales.

17.2. Dada la siguiente tabla del análisis de la varianza:

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre los grupos	879	3
Dentro de los grupos	798	16
Total	1.677	19

calcule los cuadrados medios entre los grupos y dentro de los grupos. Calcule el cociente F y contraste la hipótesis de que las medias de los grupos son iguales.

17.3. Dada la siguiente tabla del análisis de la varianza:

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre los grupos	1.000	2
Dentro de los grupos	743	15
Total	1.743	17

calcule los cuadrados medios entre los grupos y dentro de los grupos. Calcule el cociente F y contraste la hipótesis de que las medias de los grupos son iguales.

Ejercicios aplicados

17.4. Un fabricante de cereales tiene que elegir entre tres colores para las cajas de cereales: rojo, amarillo y azul. Para averiguar si el color influye en las ventas, se eligen 16 tiendas de tamaño parecido. Se envían cajas rojas a 6 de estas tiendas, cajas amarillas a 5 y cajas azules a las 5 restantes. Después de unos días, se comprueba el número de cajas vendidas en cada tienda. La tabla adjunta muestra los resultados (en decenas de cajas) obtenidos.

Rojo	Amarillo	Azul
43	52	61
52	37	29
59	38	38
76	64	53
61	74	79
81		

a) Calcule la suma de los cuadrados dentro de los grupos, entre los grupos y total.
b) Complete la tabla del análisis de la varianza y contraste la hipótesis nula de que las medias poblacionales de los niveles de ventas de las cajas de los tres colores son iguales.

17.5. Un profesor tiene una clase de 23 estudiantes. Al comienzo de cada cuatrimestre asigna a cada estudiante aleatoriamente a uno de los cuatro profesores ayudantes que tiene: Sánchez, Hervás, Alarcos o Blázquez. Anima a los estudiantes a reunirse con su profesor ayudante para que les explique la materia difícil del curso. Al final del cuatrimestre, se hace un examen. La tabla adjunta muestra las calificaciones obtenidas por los estudiantes que trabajan con estos profesores ayudantes.

Sánchez	Hervás	Alarcos	Blázquez
72	78	80	79
69	93	68	70
84	79	59	61
76	97	75	74
64	88	82	85
	81	68	63

a) Calcule la suma de los cuadrados dentro de los grupos, entre los grupos y total.
b) Complete la tabla del análisis de la varianza y contraste la hipótesis nula de la igualdad de las medias poblacionales de las calificaciones de estos profesores ayudantes.

17.6. Tres proveedores suministran piezas en envíos de 500 unidades. Se han comprobado minuciosamente muestras aleatorias de seis envíos de cada uno de los tres proveedores y se ha anotado el número de piezas que no se ajustan a las normas. La tabla muestra este número.

Proveedor A	Proveedor B	Proveedor C
28	22	33
37	27	29
34	29	39
29	20	33
31	18	37
33	30	38

a) Elabore la tabla del análisis de la varianza de estos datos.
b) Contraste la hipótesis nula de que la igualdad de las medias poblacionales del número de piezas por envío de los tres proveedores no se ajustan a las normas.

17.7. Una empresa está tratando de elegir entre tres tipos de automóvil para su flota: nacionales, japoneses o europeos. Se piden cinco automóviles de cada tipo y, después de recorrer 10.000 kilómetros con ellos, se calcula el coste de explotación por kilómetro de cada uno. Se obtienen los siguientes resultados en centavos por kilómetro.

Nacionales	Japoneses	Europeos
18,0	20,1	19,3
176	17,6	17,4
17,4	16,1	17,1
19,1	17,3	18,6
16,9	17,4	16,1

a) Elabore la tabla del análisis de la varianza de estos datos.
b) Contraste la hipótesis nula de que las medias poblacionales de los costes de explotación medios por kilómetro de los tres tipos de automóviles son iguales.

17.8. Se toman muestras aleatorias de siete estudiantes universitarios de primer año, siete de segundo año y siete de tercero que asisten a una clase de estadística para los negocios. La tabla adjunta muestra las calificaciones obtenidas en el examen final.

Estudiantes de primer año	Estudiantes de segundo año	Estudiantes de tercer año
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87

a) Elabore la tabla del análisis de la varianza de estos datos.
b) Contraste la hipótesis nula de que las medias poblacionales de las calificaciones de los tres grupos son iguales.

17.9. Se pide a muestras de cuatro vendedores de cuatro regiones distintas que predigan los aumentos porcentuales del volumen de ventas de sus territorios en los próximos 12 meses. La tabla adjunta muestra las predicciones.

Oeste	Norte	Sur	Este
6,8	7,2	4,2	9,0
4,2	6,6	4,8	8,0
5,4	5,8	5,8	7,2
5,0	7,0	4,6	7,6

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de las cuatro predicciones del crecimiento de las ventas de las cuatro regiones son iguales.

17.10. Se pide a muestras aleatorias independientes de seis profesores ayudantes, cuatro profesores asociados y cinco profesores titulares que estimen la cantidad de tiempo que dedicaron a sus responsabilidades docentes fuera del aula la semana pasada. La tabla adjunta muestra los resultados en horas.

Ayudante	Asociado	Titular
7	15	11
12	12	7
11	15	6
15	8	9
9		7
14		

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de los tiempos de los tres tipos de profesores son iguales.

17.11. Dos academias ofrecen cursos para prepararse para el examen de acceso a la universidad. Para comprobar la eficacia de sus cursos, se eligen 15 estudiantes. Cinco se asignan aleatoriamente a la academia A, cinco a la B y el resto no asiste a ningún curso. La tabla adjunta muestra las calificaciones obtenidas en el examen, expresadas en porcentajes.

Academia A	Academia B	Academia C
79	74	72
74	69	71
92	87	81
67	81	61
85	64	63

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de las calificaciones de los tres grupos son iguales.

17.12. En el estudio del ejemplo 17.1 se toman muestras aleatorias independientes de seis tipos de anuncios. La tabla adjunta muestra los *índices fog* de estos anuncios. Contraste la hipótesis nula de que las medias poblacionales de los *índices fog* de los tres tipos de anuncios son iguales.

Tipo 1	Tipo 2	Tipo 3
12,89	9,50	10,21
12,69	8,60	9,66
11,15	8,59	7,67
9,52	6,50	5,12
9,12	4,79	4,88
7,04	4,29	3,12

17.13. En el modelo del análisis de la varianza de un factor, expresamos la j-ésima observación del i-ésimo grupo de la forma siguiente:

$$X_{ij} = \mu + G_i + \varepsilon_{ij}$$

donde μ es la media global, G_i es el efecto específico del i-ésimo grupo y ε_{ij} es el error aleatorio de la j-ésima observación del i-ésimo grupo. Considere los datos del ejemplo 17.1.

a) Estime μ.
b) Estime G_i de cada una de las tres revistas.
c) Estime ε_{32}, el término de error correspondiente a la segunda observación (8,28) del *New Yorker*.

17.14. Utilice el modelo del análisis de la varianza de un factor para examinar los datos del ejercicio 17.12.

a) Estime μ.
b) Estime G_i de cada uno de los tres tipos de anuncios.
c) Estime ε_{13}, el término de error correspondiente a la tercera observación (11,15) del primer tipo de anuncio.

17.3. El contraste de Kruskal-Wallis

Como ya hemos señalado, el contraste del análisis de la varianza de un factor del apartado 17.2 generaliza al caso en el que hay varias poblaciones el contraste t utilizado para comparar dos medias poblacionales cuando se dispone de muestras aleatorias independientes. El contraste se basa en el supuesto de que las distribuciones poblacionales subyacentes son normales. En el apartado 15.3 introdujimos el contraste de Mann-Whitney, un contraste no paramétrico que es válido para comparar las posiciones centrales de dos poblaciones basado en muestras aleatorias independientes, incluso cuando las distribuciones poblacionales no son normales. También es posible desarrollar una alternativa no paramétrica al contraste del análisis de la varianza de un factor. Este contraste se conoce con el nombre de **contraste de Kruskal-Wallis** y se emplea cuando un investigador tiene poderosas razones para sospechar que las distribuciones poblacionales subyacentes pueden ser muy diferentes de la normal.

Al igual que la mayoría de los contrastes no paramétricos que ya hemos visto, el contraste de Kruskal-Wallis se basa en los *puestos* ocupados por las observaciones muestrales en las ordenaciones correspondientes. Mostraremos cómo se calcula el estadístico del contraste utilizando los datos sobre el consumo de combustible de la Tabla 17.1. Los valores muestrales se juntan y se ordenan en sentido ascendente, como en la Tabla 17.6, utilizando la media de los puestos en caso de empate.

El contraste se basa en las sumas de los puestos R_1, R_2, ..., R_K de las K muestras. En el ejemplo del consumo de combustible,

$$R_1 = 32 \qquad R_2 = 101{,}5 \qquad R_3 = 76{,}5$$

Tabla 17.6. Cifras de consumo de combustible (en kilómetros por litro) y puestos de tres muestras aleatorias independientes.

Automóviles A	Puesto	Automóviles B	Puesto	Automóviles C	Puesto
22,2	11	24,6	20	22,7	12
19,9	1	23,1	13	21,9	7
20,3	2,5	22,0	8	23,2	14
21,4	6	23,5	16,5	24,1	19
21,2	5	23,6	18	22,1	9,5
21,0	4	22,1	9,5	23,4	15
20,3	2,5	23,5	16,5		
Suma de los puestos	**32**		**101,5**		**76,5**

La hipótesis nula que debe contrastarse es que las tres medias poblacionales son iguales. Sospecharíamos de esa hipótesis si hubiera notables diferencias entre las medias de los puestos de las K muestras. De hecho, nuestro contraste se basa en el estadístico, donde n_i son los tamaños muestrales de los K grupos y n es el número total de observaciones muestrales. Sea W

$$W = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1)$$

La hipótesis nula sería dudosa si el valor de W fuera alto. La base del contraste se deduce del hecho de que, a menos que los tamaños de las muestras sean muy pequeños, la variable aleatoria correspondiente al estadístico del contraste sigue, según la hipótesis nula, una distribución de la que es una buena aproximación la distribución χ^2 con $(K-1)$ grados de libertad.

El contraste de Kruskal-Wallis

Supongamos que tenemos muestras aleatorias independientes de n_1, n_2, ..., n_K observaciones de K poblaciones. Sea

$$n = n_1 + n_2 + \cdots + n_K$$

el número total de observaciones muestrales. Sean R_1, R_2, ..., R_K las sumas de los puestos de las K muestras cuando se juntan las observaciones muestrales y se ordenan en sentido ascendente. El contraste de la hipótesis nula, H_0, de la igualdad de las medias poblacionales se basa en el estadístico

$$W = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1) \qquad \textbf{(17.8)}$$

La regla de decisión de un contraste al nivel de significación a es

$$\text{Rechazar } H_0 \text{ si} \quad W > \chi^2_{K-1,\alpha} \qquad \textbf{(17.9)}$$

donde $\chi^2_{K-1,\alpha}$ es el número que es superado con la probabilidad α por una variable aleatoria χ^2 con $(K-1)$ grados de libertad.

Este método es aproximadamente válido, siempre que la muestra contenga al menos cinco observaciones de cada población.

En el caso de nuestros datos sobre el consumo de combustible, tenemos que

$$W = \frac{12}{(20)(21)}\left[\frac{(32)^2}{7} + \frac{(101{,}5)^2}{7} + \frac{(76{,}5)^2}{6}\right] - (3)(21) = 11{,}10$$

Aquí, tenemos $(K - 1) = 2$ grados de libertad, por lo que en el caso de un contraste al nivel de significación del 0,5 por ciento, vemos en la Tabla 7 del apéndice que

$$\chi^2_{2,\,0{,}005} = 10{,}60$$

Por lo tanto, la hipótesis nula de que las medias poblacionales del consumo de combustible de los tres tipos de automóviles son iguales puede rechazarse incluso al nivel de significación del 0,5 por ciento. Naturalmente, también rechazamos esta hipótesis utilizando el contraste del análisis de la varianza del apartado 17.2. Sin embargo, aquí hemos sido capaces de rechazarlo sin imponer el supuesto de la normalidad de las distribuciones poblacionales.

EJEMPLO 17.2. Importancia de las marcas (contraste de Kruskal-Wallis)

Se ha realizado un estudio para averiguar si las mujeres de diferentes subgrupos profesionales dan diferentes niveles de importancia a las marcas cuando compran bebidas refrescantes.

Solución

Se pidió a muestras aleatorias independientes de 101 empleadas de oficina, 112 administrativas y 96 profesionales que valoraran en una escala de 1 a 7 la importancia que daban a la marca cuando compraban bebidas refrescantes. El valor del estadístico de Kruskal-Wallis de este estudio era 25,22. Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de los tres subgrupos son iguales.

El estadístico del contraste calculado es

$$W = 25{,}22$$

Dado que hay $K = 3$ grupos, tenemos para un contraste al 0,5 por ciento

$$\chi^2_{K-1,\,\alpha} = \chi^2_{2,\,0{,}005} = 10{,}60$$

Por lo tanto, la hipótesis nula de que las medias poblacionales de las valoraciones de los tres subgrupos son iguales se rechaza claramente con los datos de esta muestra, incluso al nivel de significación del 0,5 por ciento. Tenemos pruebas contundentes de que las mujeres de diferentes subgrupos profesionales dan diferentes niveles de importancia a las marcas.

EJERCICIOS

Ejercicios básicos

17.15. Considere un problema con tres subgrupos en el que la suma de los puestos de cada uno de los subgrupos es igual a 45, 98 y 88 y el tamaño de los subgrupos es igual a 6, 6 y 7. Complete el contraste de Kruskal-Wallis y la hipótesis nula de que los puestos de los subgrupos son iguales.

17.16. Considere un problema con cuatro subgrupos en el que la suma de los puestos de cada uno de los subgrupos es igual a 49, 84, 76 y 81 y el ta-

maño de los subgrupos es igual a 4, 6, 7 y 6. Complete el contraste de Kruskal-Wallis y la hipótesis nula de que los puestos de los subgrupos son iguales.

17.17. Considere un problema con cuatro subgrupos en el que la suma de los puestos de cada uno de los subgrupos es igual a 71, 88, 82 y 79 y el tamaño de los subgrupos es igual a 5, 6, 6 y 7. Complete el contraste de Kruskal-Wallis y la hipótesis nula de que los puestos de los subgrupos son iguales.

Ejercicios aplicados

17.18. Basándose en los datos del ejercicio 17.4, utilice el contraste de Kruskal-Wallis de la hipótesis nula de que las medias poblacionales de los niveles de ventas de las cajas de los tres colores son iguales.

17.19. Basándose en los datos del ejercicio 17.5, utilice el contraste de Kruskal-Wallis de la hipótesis nula de que las medias poblacionales de las calificaciones de los estudiantes asignados a los cuatro profesores ayudantes son iguales.

17.20. Basándose en los datos del ejercicio 17.6, realice un contraste de la hipótesis nula de la igualdad de las medias poblacionales del número de piezas por envío de los tres proveedores que no se ajustan a las normas sin suponer que las distribuciones poblacionales son normales.

17.21. Basándose en los datos del ejercicio 17.7, contraste la hipótesis nula de que las medias poblacionales de los costes de explotación por kilómetro de los tres tipos de automóvil son iguales sin suponer que las distribuciones poblacionales son normales.

17.22. Basándose en los datos del ejercicio 17.8, realice un contraste no paramétrico de la hipótesis nula de la igualdad de las medias poblacionales de las calificaciones de los estudiantes de primer año, de segundo año y de tercer año.

17.23. Basándose en los datos del ejercicio 17.9, utilice el método de Kruskal-Wallis para contrastar la hipótesis nula de la igualdad de las medias poblacionales de las predicciones para las cuatro regiones.

17.24. Vuelva al ejercicio 17.10. Sin suponer que las distribuciones poblacionales son normales, contraste la hipótesis nula de que las medias poblacionales del tiempo que dedican los ayudantes, los asociados y los titulares a las responsabilidades docentes fuera del aula son iguales.

17.25. Basándose en los datos del ejercicio 17.11, realice el contraste de Kruskal-Wallis de la hipótesis nula de la igualdad de las medias poblacionales de las calificaciones obtenidas en el examen de acceso a la universidad por los estudiantes que no van a una academia y los que van a la academia A y a la academia B.

17.26. Se pide a muestras aleatorias independientes de 101 estudiantes universitarios de primer año, 112 de segundo año y 96 de tercer año que valoren en una escala de 1 a 7 la importancia que conceden a la marca cuando compran un automóvil. El valor del estadístico de Kruskal-Wallis obtenido es 0,17.

a) ¿Qué hipótesis nula puede contrastarse utilizando esta información?

b) Realice el contraste.

17.4. Análisis de la varianza bifactorial: una observación por celda, bloques aleatorizados

Aunque lo que nos interesa principalmente es el análisis de un aspecto de un experimento, podemos sospechar que hay un segundo factor que influye significativamente en el resultado. En los apartados anteriores de este capítulo hemos analizado un experimento en el que el objetivo era comparar el consumo de combustible de tres tipos de automóviles. Hemos recogido datos de tres muestras aleatorias independientes de pruebas y los hemos analizado por medio de un análisis de la varianza de un factor. Hemos supuesto que la variabilidad de los datos muestrales se debía a dos causas: a la existencia de verdaderas diferencias entre los tres tipos de automóviles y a una variación aleatoria. De hecho, podríamos sospechar que parte de la variabilidad aleatoria observada se debe a las diferencias entre los

hábitos de los conductores. Si fuera posible aislar este último factor, disminuiría la cantidad de variabilidad aleatoria del experimento. Eso permitiría, a su vez, detectar más fácilmente las diferencias de rendimiento entre los automóviles. En otras palabras, diseñando un experimento para tener en cuenta las diferencias entre las características de los conductores, confiamos en conseguir un contrate más poderoso de la hipótesis nula de que las medias poblacionales del consumo de combustible de todos los tipos de automóviles son iguales.

De hecho, es bastante sencillo diseñar un experimento que pueda tener en cuenta la influencia de un segundo factor de este tipo. Supongamos, una vez más, que tenemos tres tipos de automóvil (por ejemplo, automóviles α, automóviles β y automóviles γ) cuyo consumo de combustible queremos comparar. Consideramos un experimento en el que se realizan seis pruebas con cada tipo de automóvil. Si se realizan estas pruebas utilizando seis conductores, cada uno de los cuales conduce un automóvil de los tres tipos, es posible, dado que cada tipo de automóvil será probado por cada conductor, extraer de los resultados información sobre la variabilidad de los conductores, así como información sobre las diferencias entre los tres tipos de automóvil. La variable adicional —en este caso, los conductores— se denomina a veces *variable de bloqueo*. Se dice que este experimento está organizado en *bloques*; en nuestro ejemplo, habría seis bloques, uno por cada conductor.

Este tipo de diseño por bloques puede utilizarse para obtener información sobre dos factores simultáneamente. Supongamos, por ejemplo, que queremos comparar el consumo de combustible de diferentes tipos de automóvil, pero también de diferentes tipos de conductores. En concreto, es posible que nos interese saber cómo influye la edad de los conductores en el consumo de combustible. Para eso, podemos subdividir los conductores en grupos de edad. Podríamos utilizar los seis grupos de edad siguientes (en años):

1. 25 años o menos
2. 26-35
3. 36-45
4. 46-55
5. Más de 65

A continuación, podemos organizar nuestro experimento de tal forma que un automóvil de cada grupo sea conducido por un conductor de cada grupo de edad. De esta forma, además de contrastar la hipótesis de que las medias poblacionales del consumo de combustible de todos los tipos de automóvil son iguales, podemos contrastar la hipótesis de que las medias poblacionales del consumo medio de combustible de todos los grupos de edad son iguales.

De hecho, independientemente de que cada uno de los seis conductores conduzca un automóvil de cada tipo o un conductor de cada una de las seis clases de edad conduzca un automóvil de cada tipo, el método para contrastar la igualdad de las medias poblacionales del consumo de combustible de los tipos de automóviles es el mismo. En este apartado utilizamos el segundo diseño a modo de ilustración.

La Tabla 17.7 contiene los resultados de un experimento realizado con tres tipos de automóvil y conductores de seis grupos de edad. El objetivo principal es comparar los tipos de automóvil y la variable de bloqueo es la edad de los conductores.

Este tipo de diseño se llama **diseño por bloques aleatorizados**. La aleatoriedad se debe a que seleccionamos aleatoriamente un conductor del primer grupo de edad para conducir un automóvil α, un conductor del segundo grupo de edad para conducir un automóvil α, y así sucesivamente. Este procedimiento se repite con cada grupo de edad y con cada tipo de automóvil. Si es posible, las pruebas deben realizarse siguiendo un orden aleatorio, no bloque por bloque.

Tabla 17.7. Observaciones muestrales sobre el consumo de combustible de tres tipos de automóviles conducidos por conductores de seis clases.

	Tipo de automóvil			
Clase de conductores	**Automóviles α**	**Automóviles β**	**Automóviles γ**	**Suma**
1	25,1	23,9	26,0	75,0
2	24,7	23,7	25,4	73,8
3	26,0	24,4	25,8	76,2
4	24,3	23,3	24,4	72,0
5	23,9	23,6	24,2	71,7
6	24,2	24,5	25,4	74,1
Suma	148,2	143,4	151,2	442,8

Supongamos que tenemos K grupos y que hay H bloques. Representaremos por medio de x_{ij} la observación muestral correspondiente al i-ésimo grupo y el j-ésimo bloque. Por lo tanto, los datos muestrales pueden mostrarse como en la Tabla 17.8. Obsérvese que este formato es simplemente una extensión del que utilizamos para realizar el contraste de observaciones pareadas del apartado 11.1, en el que sólo teníamos dos grupos para poder contrastar la igualdad de varias medias poblacionales.

Tabla 17.8. Observación muestral sobre K grupos y H bloques.

	Grupo			
Bloque	**1**	**2**	$\cdots$	***K***
1	x_{11}	x_{21}	$\cdots$	x_{K1}
2	x_{12}	x_{22}	$\cdots$	x_{K2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H	x_{1H}	x_{2H}	$\cdots$	x_{KH}

Para desarrollar un contraste de la hipótesis de que las medias poblacionales de todos los K grupos son iguales, necesitamos las medias muestrales de estos grupos. Para representar la media del i-ésimo grupo, utilizamos la notación $\bar{x}_{i\bullet}$, por lo que

$$\bar{x}_{i\bullet} = \frac{\sum_{j=1}^{H} x_{ij}}{H} \qquad (i = 1, 2, ..., K)$$

Basándonos en la Tabla 17.7, tenemos que

$$\bar{x}_{1\bullet} = \frac{148{,}2}{6} = 24{,}7 \qquad \bar{x}_{2\bullet} = \frac{143{,}4}{6} = 23{,}9 \qquad \bar{x}_{3\bullet} = \frac{151{,}2}{6} = 25{,}2$$

También nos interesan las diferencias entre las medias de los bloques poblacionales. Por lo tanto, necesitamos las medias muestrales de los H bloques. Representamos por medio de $\bar{x}_{\bullet j}$ la media muestral del j-ésimo bloque, por lo que

$$\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{K} x_{ij}}{K} \qquad (j = 1, 2, ..., H)$$

En el caso de los datos sobre el consumo de combustible de la Tabla 17.7, tenemos que

$$\bar{x}_{\bullet 1} = \frac{75{,}0}{3} = 25{,}0 \qquad \bar{x}_{\bullet 2} = \frac{73{,}8}{3} = 24{,}6 \qquad \bar{x}_{\bullet 3} = \frac{76{,}2}{3} = 25{,}4$$

$$\bar{x}_{\bullet 4} = \frac{72{,}0}{3} = 24{,}0 \qquad \bar{x}_{\bullet 5} = \frac{71{,}7}{3} = 23{,}9 \qquad \bar{x}_{\bullet 6} = \frac{74{,}1}{3} = 24{,}7$$

Por último, necesitamos la media global de las observaciones muestrales. Si n representa el número total de observaciones, entonces

$$n = HK$$

y la media muestral de las observaciones es

$$\bar{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} x_{ij}}{n} = \frac{\sum_{i=1}^{K} \bar{x}_{i\bullet}}{K} = \frac{\sum_{j=1}^{H} \bar{x}_{\bullet j}}{H}$$

En el caso de los datos de la Tabla 17.7,

$$\bar{x} = \frac{442{,}8}{18} = 24{,}6$$

Antes de preguntarnos cuál es el contraste adecuado de la hipótesis que nos interesa, es útil examinar el modelo poblacional en el que nos basamos implícitamente. Supongamos que la variable aleatoria X_{ij} corresponde a la observación del i-ésimo grupo y el j-ésimo bloque. Se considera que este valor es la suma de los cuatro componentes siguientes.

1. Una media «global» μ.
2. Un parámetro G_i, que es específico del i-ésimo grupo y que mide la diferencia entre la media de ese grupo y la media global.
3. Un parámetro B_j, que es específico del i-ésimo bloque y que mide la diferencia entre la media de ese bloque y la media global.
4. Una variable aleatoria ε_{ij}, que representa el error experimental, o sea la parte de la observación que no es explicada ni por la media global ni por la pertenencia a los grupos o los bloques.

Podemos escribir, pues,

$$X_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$$

Se supone que el término de error ε_{ij} satisface los supuestos habituales del modelo de regresión múltiple. En concreto, pues, se supone que las varianzas son independientes e iguales.

En tal caso, podemos formular la expresión anterior de la forma siguiente:

$$X_{ij} - \mu = G_i + B_j + \varepsilon_{ij}$$

A continuación, dados los datos muestrales, estimamos la media global μ por medio de la media muestral global $\bar{x}$, por lo que $(x_{ij} - \bar{x})$ es una estimación del primer miembro. La diferencia G_j entre la media poblacional del i-ésimo grupo y la media poblacional global se estima por medio de la correspondiente diferencia entre las medias muestrales, $(\bar{x}_{i\bullet} - \bar{x})$.

Asimismo, B_j se estima por medio de $(\bar{x}_{\bullet j} - \bar{x})$. Por último, restando, estimamos el término de error:

$$(x_{ij} - \bar{x}) - (\bar{x}_{i\bullet} - \bar{x}) - (\bar{x}_{\bullet j} - \bar{x}) = x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x}$$

Tenemos, pues, considerando los miembros muestrales, que

$$(x_{ij} - \bar{x}) = (\bar{x}_{i\bullet} - \bar{x}) - (\bar{x}_{\bullet j} - \bar{x}) + (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})$$

Para ilustrarlo, consideremos el consumo de combustible de un conductor del tercer grupo de edad con un automóvil α. Según la Tabla 17.7,

$$x_{13} = 26{,}0$$

El término del primer miembro es

$$x_{13} - \bar{x} = 26{,}0 - 24{,}6 = 1{,}4$$

El efecto del grupo (automóvil) es

$$\bar{x}_{1\bullet} - \bar{x} = 24{,}7 - 24{,}6 = 0{,}1$$

Obsérvese que este término será el mismo siempre que se conduzca el automóvil α. El efecto del bloque (conductor) es

$$\bar{x}_{\bullet 3} - \bar{x} = 25{,}4 - 24{,}6 = 0{,}8$$

Por último, el término de error es

$$x_{13} - \bar{x}_{1\bullet} - \bar{x}_{\bullet 3} + \bar{x} = 26{,}0 - 24{,}7 - 25{,}4 + 24{,}6 = 0{,}5$$

Por lo tanto, tenemos para esta observación

$$1{,}4 = 0{,}1 + 0{,}8 + 0{,}5$$

Podemos interpretar esta ecuación de la forma siguiente: cuando un conductor del tercer grupo de edad probó el automóvil α, consumió 1,4 kilómetros por litro más que la media de todos los automóviles y los conductores. Se estima que de esta cantidad 0,1 se debe al automóvil, 0,8 al grupo de edad del conductor y el resto, 05, a otros factores, que atribuimos a la variabilidad aleatoria o a un error experimental.

Ahora, si elevamos al cuadrado los dos miembros y sumamos las n observaciones muestrales, puede demostrarse que el resultado es

$$\sum_{i=1}^{K}\sum_{j=1}^{H}(x_{ij} - \bar{x})^2 = H\sum_{i=1}^{K}(\bar{x}_{i\bullet} - \bar{x})^2 + K\sum_{j=1}^{H}(\bar{x}_{\bullet j} - \bar{x})^2 + \sum_{i=1}^{K}\sum_{j=1}^{H}(x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} - \bar{x})$$

En esta ecuación, la variabilidad muestral total de las observaciones en torno a la media global es la suma de las variabilidades que se deben a las diferencias entre los grupos, a las diferencias entre los bloques y al error, respectivamente. Es en la descomposición de estas sumas de los cuadrados en la que se basa el análisis de experimentos de este tipo. El análisis se llama análisis de la varianza bifactorial, ya que los datos se clasifican de dos formas, por grupos y por bloques.

En la Figura 17.5 mostramos esta importante descomposición de la suma de los cuadrados. Obsérvese que, a diferencia de la descomposición del análisis de la varianza de un factor, la descomposición de la suma total de los cuadrados de las observaciones muestra-

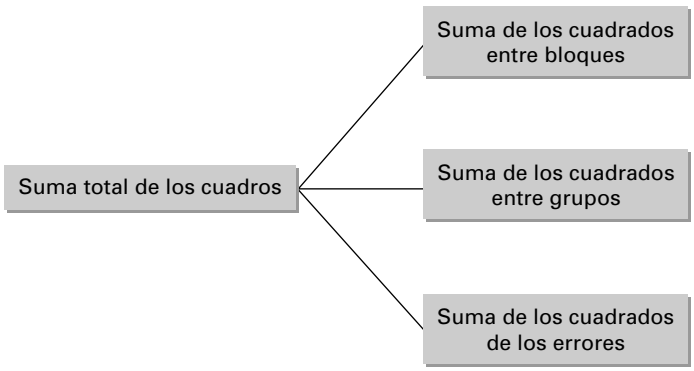


Figura 17.5. Descomposición de la suma de los cuadrados de un análisis de la varianza bifactorial con una observación por celda.

les en torno a su media global se divide aquí en *tres* componentes, que resumimos en las ecuaciones 17.10 a 17.14; el componente extra se debe a nuestra capacidad para extraer de los datos información sobre las diferencias entre los bloques.

En el caso de los datos sobre el consumo de combustible de la Tabla 17.7, tenemos que

$$STC = (25{,}1 - 24{,}6)^2 + (24{,}7 - 24{,}6)^2 + \cdots + (25{,}4 - 24{,}6)^2 = 11{,}88$$
$$SCG = 6[(24{,}7 - 24{,}6)^2 + (23{,}9 - 24{,}6)^2 + (25{,}2 - 24{,}6)^2] = 5{,}16$$
$$SCB = 3[(25{,}0 - 24{,}6)^2 + (24{,}6 - 24{,}6)^2 + \cdots + (24{,}7 - 24{,}6)^2] = 4{,}98$$

por lo que, restando,

$$SCE = STC - SCG - SCB = 11{,}88 - 5{,}16 - 4{,}98 = 1{,}74$$

Descomposición de la suma de los cuadrados del análisis de la varianza bifactorial

Supongamos que tenemos una muestra de observaciones y que x_{ij} es la observación del i-ésimo grupo y el j-ésimo bloque. Supongamos que hay K grupos y H bloques, lo que hace un total de

$$n = KH$$

observaciones. Sean las medias muestrales de los grupos $\bar{x}_{i\bullet}$ ($i = 1, 2, ..., K$), las medias muestrales de los bloques $\bar{x}_{\bullet j}$ ($j = 1, 2, ..., H$) y la media muestral global $\bar{x}$.

Definimos las siguientes sumas de los cuadrados:

Total: $$STC = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ji} - \bar{x})^2 \qquad \textbf{(17.10)}$$

Entre grupos: $$SCG = H \sum_{i=1}^{K} (\bar{x}_{i\bullet} - \bar{x})^2 \qquad \textbf{(17.11)}$$

Entre bloques: $$SCR = K \sum_{j=1}^{H} (\bar{x}_{\bullet i} - \bar{x})^2 \qquad \textbf{(17.12)}$$

Error: $$SCE = \sum_{i=1}^{K} \sum_{j=1}^{H} (x_{ij} - \bar{x}_{i\bullet} - \bar{x}_{\bullet j} + \bar{x})^2 \qquad \textbf{(17.13)}$$

Entonces

$$STC = SCG + SCB + SCE \qquad \textbf{(17.14)}$$

A partir de aquí, los contrastes relacionados con el análisis de la varianza bifactorial se realizan de la misma forma que los contrastes relacionados con el análisis de la varianza de un factor del apartado 17.2. En primer lugar, obtenemos la media de los cuadrados dividiendo cada una de las sumas de los cuadrados por el número correspondiente de grados de libertad. En el caso de la suma total de los cuadrados, los grados de libertad son 1 menos que el número total de observaciones, es decir, $(n - 1)$. En el caso de la suma de los cuadrados entre grupos, los grados de libertad son 1 menos que el número de grupos, o sea $(K - 1)$. Asimismo, en el caso de la suma de los cuadrados entre bloques, el número de grados de libertad es $(H - 1)$. Por lo tanto, restando, los grados de libertad correspondientes a la suma de los cuadrados de los errores son

$$\begin{aligned}(n - 1) - (K - 1) - (H - 1) &= n - K - H + 1 \\ &= KH - K - H + 1 \\ &= (K - 1)(H - 1)\end{aligned}$$

La hipótesis nula de que las medias poblacionales de los grupos son iguales puede contrastarse entonces por medio del cociente entre la media de los cuadrados de los grupos y la media de los cuadrados de los errores, como muestra la ecuación 17.18. A menudo se incluye una variable de bloqueo en el análisis simplemente para reducir la variabilidad debida al error experimental. Sin embargo, a veces también tiene interés la hipótesis de que las medias poblacionales de los bloques son iguales. Esta hipótesis puede contrastarse por medio del cociente entre la media de los cuadrados de los bloques y la media de los cuadrados de los errores de la ecuación 17.19. Al igual que ocurre en el caso del análisis de la varianza de un factor, la comparación proviene de la probabilidad de una cola de la distribución F.

En el caso de los datos sobre el consumo de combustible, la media de los cuadrados es

$$MCG = \frac{SCG}{K - 1} = \frac{5{,}16}{2} = 2{,}58$$

$$MCB = \frac{SCB}{H - 1} = \frac{4{,}98}{5} = 0{,}996$$

$$MCE = \frac{SCE}{(K - 1)(H - 1)} = \frac{1{,}74}{10} = 0{,}174$$

Para contrastar la hipótesis nula de que las medias poblacionales del consumo de combustible de los tres tipos de automóviles son iguales, necesitamos

$$\frac{MCG}{MCE} = \frac{2{,}58}{0{,}174} = 14{,}83$$

En el caso de un contraste al nivel de significación del 1 por ciento, vemos en la Tabla 9 del apéndice que

$$F_{K-1,\,(K-1)(H-1),\,\alpha} = F_{2,\,10,\,0{,}01} = 7{,}56$$

Contrastes de hipótesis en el caso del análisis de la varianza bifactorial

Supongamos que tenemos una observación muestral para cada combinación grupo-bloque en un diseño que contiene K grupos y H bloques:

$$x_{ji} = \mu + G_j + B_i + \varepsilon_{ji}$$

donde G_j es el efecto del grupo y B_i es el efecto del bloque.

Definamos las siguientes **medias de los cuadrados**:

$$\text{Entre grupos:} \quad MCG = \frac{SCG}{K-1} \tag{17.15}$$

$$\text{Entre bloques:} \quad MCB = \frac{SCB}{H-1} \tag{17.16}$$

$$\text{Error:} \quad MCE = \frac{SCE}{(K-1)(H-1)} \tag{17.17}$$

Suponemos que los términos de error ε_{ij} del modelo son independientes entre sí y tienen la misma varianza. Suponemos, además, que estos errores siguen una distribución normal.

La regla de decisión de un contraste al nivel de significación α de la hipótesis nula, H_0, de que las K medias poblacionales de los grupos son iguales es

$$\text{Rechazar } H_0 \text{ si} \quad \frac{MCG}{MCE} > F_{K-1,(K-1)(H-1),\alpha} \tag{17.18}$$

La regla de decisión de un contraste al nivel de significación α de la hipótesis nula, H_0, de que las H medias poblacionales de los bloques son iguales es

$$\text{Rechazar } H_0 \text{ si} \quad \frac{MCB}{MCE} > F_{H-1,(K-1)(H-1),\alpha} \tag{17.19}$$

Aquí, $F_{v_1, v_2, \alpha}$ es el número que es superado con la probabilidad α por una variable aleatoria que sigue una distribución F con v_1 grados de libertad en el numerador y v_2 grados de libertad en el denominador.

Por lo tanto, basándose en estos datos, se rechaza claramente al nivel de significación del 1 por ciento la hipótesis de que las medias poblacionales del consumo de combustible de los tres tipos de automóviles son iguales.

En este ejemplo, la hipótesis nula de la igualdad de las medias poblacionales de los bloques es la hipótesis de que las medias poblacionales del consumo de combustible de todos los grupos de edad son iguales. El contraste se basa en

$$\frac{MCB}{MCE} = \frac{0{,}996}{0{,}174} = 5{,}72$$

En el caso de un contraste al 1 por ciento, vemos en la Tabla 9 del apéndice que

$$F_{H-1,(K-1)(H-1),\alpha} = F_{5,10,0{,}01} = 5{,}64$$

Por lo tanto, la hipótesis nula de la igualdad de las medias poblacionales de los seis grupos de edad también se rechaza al nivel de significación del 1 por ciento.

Una vez más, es muy cómodo resumir los cálculos en una tabla. La Tabla 17.9 muestra la organización general de la **tabla del análisis de la varianza bifactorial** y la Figura 17.6 muestra este análisis de la varianza basado en los datos sobre el consumo de gasolina. El número de grados de libertad depende del número de grupos y de bloques. Las medias de los cuadrados se obtienen dividiendo las sumas de los cuadrados por sus grados de libertad correspondientes. La media de los cuadrados de los errores es el denominador en el cálculo de los dos cocientes F en los que se basa nuestro contraste.

Tabla 17.9. Formato general de la tabla del análisis de la varianza bifactorial.

Fuente de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente F
Entre grupos	SCG	$K-1$	$MCG=\dfrac{SCG}{K-1}$	$\dfrac{MCG}{MCE}$
Entre bloques	SCB	$H-1$	$MCB=\dfrac{SCB}{H-1}$	$\dfrac{MCB}{MCE}$
Error	SCE	$(K-1)(H-1)$	$MCE=\dfrac{SCE}{(K-1)(H-1)}$	
Total	STC	$N-1$		

Figura 17.6. Resultados del análisis de la varianza bifactorial correspondiente al ejemplo 17.3 (salida Minitab).

Two-way ANOVA: Mileage versus Car, Driver

```
Source   DF       SS     MS      F      P
Car       2     5.16  2.580  14.83  0.001
Driver    5     4.98  0.996   5.72  0.009
Error    10     1.74  0.174
Total    17    11.88

S = 0.4171    R-Sq = 85.35%    R-Sq(adj) = 75.10%

                Individual 95% CIs For Mean Based on
                Pooled StDev
Car  Mean   -------+---------+---------+---------+-
1    24.7                (-----*-----)
2    23.9   (-----*-----)
3    25.2                          (-----*-----)
       -------+---------+---------+---------+-
           24.00      24.60      25.20      25.80


                   Individual 95% CIs For Mean Based on
                   Pooled StDev
Driver   Mean   -----+---------+---------+---------+---
1        25.0                  (------*------)
2        24.6            (------*------)
3        25.4                        (------*------)
4        24.0     (------*------)
5        23.9   (------*------)
6        24.7             (------*------)
                -----+---------+---------+---------+---
               23.80      24.50      25.20      25.90
```

EJEMPLO 17.3. Consumo de combustible de los automóviles (análisis de la varianza bifactorial)

Queremos averiguar si existen pruebas contundentes para concluir que hay diferencias entre los niveles de consumo de combustible de diferentes automóviles conducidos por diferentes conductores.

Solución

Los datos sobre el consumo de gasolina de la Tabla 17.7 pueden analizarse utilizando el programa Minitab y la salida mostrada en la Figura 17.6. Esta figura también muestra las distintas sumas de los cuadrados y los cocientes *F*.

EJERCICIOS

Ejercicios básicos

17.27. Considere un análisis de la varianza bifactorial con una observación por celda y bloques aleatorizados con los siguientes resultados:

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre grupos	231	4
Entre bloques	348	5
Error	550	20
Total	1.129	29

Calcule los cuadrados medios y contraste la hipótesis nula de que las medias entre grupos son iguales y las medias entre bloques son iguales.

17.28. Considere un análisis de la varianza bifactorial con una observación por celda y bloques aleatorizados con los siguientes resultados:

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre grupos	380	6
Entre bloques	232	5
Error	387	30
Total	989	41

Calcule los cuadrados medios y contraste la hipótesis nula de que las medias entre grupos son iguales y las medias entre bloques son iguales.

17.29. Considere un análisis de la varianza bifactorial con una observación por celda y bloques aleatorizados con los siguientes resultados:

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre grupos	131	3
Entre bloques	287	6
Error	360	18
Total	778	27

Calcule los cuadrados medios y contraste la hipótesis nula de que las medias entre grupos son iguales y las medias entre bloques son iguales.

Ejercicios aplicados

17.30. Se pide a cuatro analistas financieros que predigan el crecimiento de los beneficios de cinco compañías petroleras el próximo año. La tabla adjunta muestra sus predicciones, expresadas en porcentaje.

a) Elabore la tabla del análisis de la varianza bifactorial.

b) Contraste la hipótesis nula de que las medias poblacionales de las predicciones de los beneficios de todas las compañías son iguales.

	Analista			
Compañía petrolera	A	B	C	D
1	8	12	7	13
2	9	9	8	12
3	12	10	9	10
4	11	10	10	12
5	9	8	10	14

17.31. La tabla adjunta muestra los resultados (en quintales por acre) de un experimento agrícola

destinado a valorar las diferencias de rendimiento de cuatro variedades diferentes de maíz, utilizando tres fertilizantes distintos.

	Variedad			
Fertilizante	**A**	**B**	**C**	**D**
1	86	88	77	84
2	92	91	81	93
3	75	80	83	79

a) Elabore la tabla del análisis de la varianza bifactorial.
b) Contraste la hipótesis nula de que las medias poblacionales del rendimiento de las cuatro variedades de maíz son iguales.
c) Contraste la hipótesis nula de que las medias poblacionales del rendimiento de las tres marcas de fertilizante son iguales.

17.32. Una empresa ha hecho un estudio de mercado de tres nuevos tipos de sopa en algunas tiendas durante un periodo de 1 año. La tabla muestra las ventas (en miles de dólares) de cada una de las tres sopas en cada trimestre del año.

	Sopa		
Trimestre	**A**	**B**	**C**
1	47	57	65
2	63	63	76
3	79	67	54
4	52	50	49

a) Elabore la tabla del análisis de la varianza bifactorial.
b) Contraste la hipótesis nula de que las medias poblacionales de las ventas de los tres tipos de sopas son iguales.

17.33. Un fabricante de una bebida refrescante sin azúcar quiere comparar la influencia en las ventas de las latas de tres colores distintos: rojo, amarillo y azul. Se seleccionan cuatro regiones y se eligen aleatoriamente tres supermercados en cada región para vender en cada uno de ellos latas de uno de los colores. La tabla adjunta muestra las ventas (en decenas de latas) realizadas al final del periodo del experimento.

	Color de la lata		
Región	**Rojo**	**Amarillo**	**Azul**
Este	47	52	60
Sur	56	54	52
Norte	49	63	55
Oeste	41	44	48

a) Elabore la tabla del análisis de la varianza bifactorial.
b) Contraste la hipótesis nula de que las medias poblacionales de las ventas de las latas de los tres colores son iguales.

17.34. Un profesor de economía tiene que elegir entre tres libros de texto. También tiene que elegir entre tres tipos de exámenes: tipo test, redacciones y una mezcla de los dos. Durante el año, da clase a nueve grupos y asigna aleatoriamente a cada grupo una combinación de libro de texto y tipo de examen. Al final del curso obtiene las evaluaciones realizadas por los estudiantes de cada grupo. La tabla adjunta muestra estas evaluaciones.

	Libro de texto		
Examen	**A**	**B**	**C**
Tipo test	4,8	5,3	4,9
Redacción	4,6	5,0	4,3
Mezcla	4,6	5,1	4,8

a) Elabore la tabla del análisis de la varianza bifactorial.
b) Contraste la hipótesis nula de la igualdad de las medias poblacionales de las evaluaciones correspondientes a los tres libros de texto.
c) Contraste la hipótesis nula de la igualdad de las medias poblacionales de las evaluaciones correspondientes a los tres tipos de exámenes.

17.35. Hemos introducido para el análisis de la varianza bifactorial el modelo poblacional

$$X_{ij} - \mu = G_i + \beta_j + \varepsilon_{ij}$$

Basándose en los datos del ejercicio 17.33, obtenga las estimaciones muestrales de cada término del segundo miembro de esta ecuación correspondientes a la combinación región este-lata roja.

17.36. Basándose en los datos del ejercicio 17.34, obtenga las estimaciones muestrales de cada término del segundo miembro de la ecuación utilizada en el ejercicio anterior correspondientes a la combinación libro de texto C-examen tipo test.

17.37. Se pide a cuatro agencias inmobiliarias que valoren 10 viviendas situadas en un determinado barrio. En la tabla se muestran los resultados de las valoraciones, expresadas en miles de dólares.

Fuente de variación	Suma de los cuadrados
Entre agentes	268
Entre viviendas	1.152
Error	2.352

a) Complete la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de estas cuatro agencias son iguales.

17.38. Se evalúan cuatro marcas de fertilizantes. Se utiliza cada marca en seis parcelas de tierra de diferentes tipos. A continuación, se mide el aumento porcentual del rendimiento del maíz en las 24 combinaciones marca-tipo de tierra. La tabla adjunta muestra los resultados obtenidos.

Fuente de variación	Suma de los cuadrados
Entre fertilizantes	135,6
Entre tipos de tierra	81,7
Error	111,3

a) Complete la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales del aumento del rendimiento del maíz son iguales con los cuatro fertilizantes.
c) Contraste la hipótesis nula de que las medias poblacionales del aumento del rendimiento del maíz son iguales en los seis tipos de tierra.

17.39. Se proyectan con carácter experimental tres series de televisión a audiencias de cuatro regiones del país: este, sur, norte y oeste. Basándose en la reacción de la audiencia, se obtiene una puntuación de cada programa (en una escala de 0 a 100). Las sumas de los cuadrados entre los grupos (programas) y entre los bloques (regiones) son

$$SCG = 95{,}2 \quad \text{y} \quad SCB = 69{,}5$$

y la suma de los cuadrados de los errores es

$$SCE = 79{,}3$$

Elabore la tabla del análisis de la varianza y contraste la hipótesis nula de que las medias poblacionales de las puntuaciones de las reacciones de la audiencia a los tres programas son iguales.

17.40. Suponga que en el análisis de la varianza bifactorial con una observación por celda, hay solamente dos grupos. Demuestre que en este caso el cociente F para contrastar la igualdad de las medias poblacionales de los grupos es exactamente el cuadrado del estadístico del contraste analizado en el apartado 11.1 para contrastar la igualdad de las medias poblacionales, dada una muestra de datos pareados. Por lo tanto, deduzca que los dos contrastes son equivalentes en este caso concreto.

17.5. Análisis de la varianza bifactorial: más de una observación por celda

En el análisis de la varianza bifactorial del apartado 17.4, hemos visto que los datos se pueden tabular (como los de las Tablas 17.7 y 17.8) en celdas y que cada celda se refiere a una combinación grupo-bloque. Así, por ejemplo, los resultados obtenidos cuando un conductor del cuarto grupo de edad conduce un automóvil β constituyen una única celda. Una característica del diseño analizado en el apartado 17.4 es que cada celda contiene solamente una observación muestral. Así, por ejemplo, un conductor del cuarto grupo de edad prueba un automóvil β solamente una vez.

En este apartado, consideramos la posibilidad de reproducir el experimento de manera que, por ejemplo, los automóviles β sean conducidos por más de un conductor del cuarto grupo de edad. Los datos resultantes de ese diseño implicarían entonces más de una observación por celda. La extensión de la muestra de esta forma tiene dos grandes ventajas. En primer lugar, cuando se dispone de más datos muestrales, las estimaciones resultantes son más precisas y podemos distinguir mejor las diferencias entre las medias poblacionales. En segundo lugar, un diseño con más de una observación por celda permite aislar otra fuente más de variabilidad: la **interacción** entre los grupos y los bloques. Se producen esas inter-

acciones cuando las diferencias entre los efectos de los grupos no están distribuidas uniformemente entre los bloques. Por ejemplo, los conductores que consiguen unas cifras de consumo de combustible mejores que la media pueden conseguir mejores cifras cuando conducen un automóvil α que cuando conducen un automóvil β. Por lo tanto, estas cifras mejores que la media no están distribuidas de una manera uniforme entre todos los tipos de automóviles sino que son más manifiestas en unos tipos que en otros. Esta posibilidad de interacción puede tenerse en cuenta en un análisis basado en más de una observación por celda.

Para ilustrar el tipo de datos que pueden analizarse, la Tabla 17.10 contiene los resultados del consumo de combustible de conductores de cinco grupos de edad de tres tipos de automóviles: automóviles X, automóviles Y y automóviles Z. Las tres observaciones de cada celda se refieren a pruebas independientes realizadas por conductores de un grupo de edad dado con automóviles de un determinado tipo.

Tabla 17.10. Observaciones muestrales sobre el consumo de combustible de tres tipos de automóviles conducidos por cinco clases de conductores; tres observaciones por celda.

	Tipo de automóvil								
Clase de conductor	**Automóviles X**			**Automóviles Y**			**Automóviles Z**		
1	25,0	25,4	25,2	24,0	24,4	23,9	25,9	25,8	25,4
2	24,8	24,8	24,5	23,5	23,8	23,8	25,2	25,0	25,4
3	26,1	26,3	26,2	24,6	24,9	24,9	25,7	25,9	25,5
4	24,1	24,4	24,4	23,9	24,0	23,8	24,0	23,6	23,5
5	24,0	23,6	24,1	24,4	24,4	24,1	25,1	25,2	25,3

Para representar las observaciones muestrales individuales, necesitamos un subíndice triple, por lo que x_{ijl} representa la l-ésima observación de la ij-ésima celda, es decir, la l-ésima observación de la celda correspondiente al i-ésimo grupo y el j-ésimo bloque. Al igual que antes, K representa el número de grupos y H el número de bloques. L representa el número de observaciones por celda. Por lo tanto, en el ejemplo de la Tabla 17.10, $K = 3$, $H = 5$ y $L = 3$. Esta notación se muestra en la Tabla 17.11.

Tabla 17.11. Observaciones muestrales sobre K grupos y H bloques; L observaciones por celda.

	Grupo			
Bloque	**1**	**2**	$\cdots$	**3**
1	$x_{111}x_{112}\cdots x_{11L}$	$x_{211}x_{212}\cdots x_{21L}$	$\cdots$	$x_{K11}x_{K12}\cdots x_{K1L}$
2	$x_{121}x_{122}\cdots x_{12L}$	$x_{221}x_{222}\cdots x_{22L}$	$\cdots$	$x_{K21}x_{K22}\cdots x_{K2L}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
H	$x_{1H1}x_{1H2}\cdots x_{1HL}$	$x_{2H1}x_{2H2}\cdots x_{2HL}$	$\cdots$	$x_{KH1}x_{KH2}\cdots x_{KHL}$

Basándonos en los resultados de un experimento de este tipo, podemos contrastar tres hipótesis nulas: ninguna diferencia entre las medias de los grupos, ninguna diferencia entre las medias de los bloques y ninguna interacción entre los grupos y los bloques. Para reali-

zar estos contrastes, calculamos de nuevo varias medias muestrales, que se definen y se calculan de la forma siguiente.

1. *Medias de los grupos*
La media de *todas* las observaciones muestrales del i-ésimo grupo se representa por medio de $\bar{x}_{i\bullet\bullet}$, por lo que

$$\bar{x}_{i\bullet\bullet} = \frac{\sum_{j=1}^{H}\sum_{l=1}^{L} x_{ijl}}{HL}$$

Basándonos en la Tabla 17.10, tenemos que

$$\bar{x}_{1\bullet\bullet} = \frac{25{,}0 + 25{,}4 + \cdots + 23{,}6 + 24{,}1}{15} = 24{,}86$$

$$\bar{x}_{2\bullet\bullet} = \frac{24{,}0 + 24{,}4 + \cdots + 24{,}4 + 24{,}1}{15} = 24{,}16$$

$$\bar{x}_{3\bullet\bullet} = \frac{25{,}9 + 25{,}8 + \cdots + 25{,}2 + 25{,}3}{15} = 25{,}10$$

2. *Medias de los bloques*
La media de todas las observaciones muestrales del j-ésimo bloque se representa por medio de $\bar{x}_{\bullet j\bullet}$, por lo que

$$\bar{x}_{\bullet j\bullet} = \frac{\sum_{i=1}^{K}\sum_{l=1}^{L} x_{ijl}}{KL}$$

Basándonos en la Tabla 17.10, tenemos que

$$\bar{x}_{\bullet 1\bullet} = \frac{25{,}0 + 25{,}4 + \cdots + 25{,}8 + 25{,}4}{9} = 25{,}00$$

$$\bar{x}_{\bullet 2\bullet} = \frac{24{,}8 + 24{,}8 + \cdots + 25{,}0 + 25{,}4}{9} = 24{,}53$$

$$\bar{x}_{\bullet 3\bullet} = \frac{26{,}1 + 26{,}3 + \cdots + 25{,}9 + 25{,}5}{9} = 25{,}57$$

$$\bar{x}_{\bullet 4\bullet} = \frac{24{,}1 + 24{,}4 + \cdots + 23{,}6 + 23{,}5}{9} = 23{,}97$$

$$\bar{x}_{\bullet 5\bullet} = \frac{24{,}0 + 23{,}6 + \cdots + 25{,}2 + 25{,}3}{9} = 24{,}47$$

3. *Medias de las celdas*
Para comprobar la posibilidad de que haya interacciones entre los grupos y los bloques, es necesario calcular la media muestral de cada celda. Sea $\bar{x}_{ij\bullet}$ la media muestral de la (ij)-ésima celda. En ese caso,

$$\bar{x}_{ij\bullet} = \frac{\sum_{l=1}^{L} x_{ijl}}{L}$$

Por lo tanto, basándonos en los datos de la Tabla 17.10, tenemos que

$$\bar{x}_{11\bullet} = \frac{25{,}0 + 25{,}4 + 25{,}2}{3} = 25{,}2$$

$$\bar{x}_{12\bullet} = \frac{24{,}8 + 24{,}8 + 24{,}5}{3} = 24{,}7$$

y asimismo,

$$\begin{array}{lllll} & & \bar{x}_{13\bullet} = 26{,}2 & \bar{x}_{14\bullet} = 24{,}3 & \bar{x}_{15\bullet} = 23{,}9 \\ \bar{x}_{21\bullet} = 24{,}1 & \bar{x}_{22\bullet} = 23{,}7 & \bar{x}_{23\bullet} = 24{,}8 & \bar{x}_{24\bullet} = 23{,}9 & \bar{x}_{25\bullet} = 24{,}3 \\ \bar{x}_{31\bullet} = 25{,}7 & \bar{x}_{32\bullet} = 25{,}2 & \bar{x}_{33\bullet} = 25{,}7 & \bar{x}_{34\bullet} = 23{,}7 & \bar{x}_{35\bullet} = 25{,}2 \end{array}$$

4. *Media global*
La media de todas las observaciones muestrales se representa por medio de $\bar{x}$, por lo que

$$\bar{x} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{H} \sum_{l=1}^{L} x_{ijl}}{KHL}$$

En el caso de nuestros datos, esta cantidad se calcula más fácilmente hallando la media de las medias muestrales de los tres grupos, lo que lleva al resultado siguiente:

$$\bar{x} = \frac{24{,}86 + 24{,}16 + 25{,}10}{3} = 24{,}71$$

Ahora bien, para comprender mejor el análisis, es útil basarnos en el modelo poblacional supuesto. Sea X_{jil} la variable aleatoria correspondiente a la l-ésima observación de la ij-ésima celda. En ese caso, el modelo supuesto en nuestro análisis es

$$X_{ijl} = \mu + G_i + B_j + I_{ij} + \varepsilon_{ijl}$$

Los tres primeros términos del segundo miembro son exactamente los mismos que los del modelo en el que no había repeticiones. Representan, al igual que antes, una media global, un factor específico del grupo y un factor específico del bloque. El término siguiente, I_{ij}, representa el efecto de estar en la ji-ésima casilla, dado que ya se tienen en cuenta el efecto global, el efecto del grupo y el efecto del bloque. Si no hubiera ninguna interacción entre los grupos y los bloques, este término sería 0. Su presencia en el modelo nos permite averiguar si hay interacción. Por último, el término de error, ε_{ijl}, es una variable aleatoria que representa el error experimental.

Replanteamos el modelo en forma de desviaciones con respecto a la media:

$$X_{ijl} - \mu = G_i + B_j + I_{ij} + \varepsilon_{ijl}$$

Se demuestra que la suma total de los cuadrados puede descomponerse en la suma de cuatro términos, que representan la variabilidad que se debe a los grupos, a los bloques, a la interacción entre los grupos y los bloques y al error.

En las ecuaciones 17.20 a 17.25 se muestra la descomposición en la que se basan los contrastes sin indicar en detalle cómo se obtienen.

Análisis de la varianza bifactorial: varias observaciones por celda

Supongamos que tenemos una muestra de observaciones sobre K grupos y H bloques y L observaciones por celda. Sea x_{ijl} la l-ésima observación de la celda del i-ésimo grupo y el j-ésimo bloque. Sea $\bar{x}$ la media muestral global, $\bar{x}_{i\bullet\bullet}$ las medias muestrales de los grupos, $\bar{x}_{\bullet j\bullet}$ las medias muestrales de los bloques y $\bar{x}_{ij\bullet}$ las medias muestrales de las celdas.

A continuación, definimos las siguientes sumas de los cuadrados y los grados de libertad correspondientes:

	Suma de los cuadrados	Grados de libertad	
Total:	$STC = \sum_i \sum_j \sum_l (x_{jil} - \bar{x})^2$	$KHL - 1$	**(17.20)**
Entre grupos:	$SCG = HL \sum_{i=1}^{K} (\bar{x}_{i\bullet\bullet} - \bar{x})^2$	$K - 1$	**(17.21)**
Entre bloques:	$SCB = KL \sum_{j=1}^{H} (\bar{x}_{\bullet j\bullet} - \bar{x})^2$	$H - 1$	**(17.22)**
Interacciones:	$SCI = L \sum_{i=1}^{K} \sum_{j=1}^{H} (\bar{x}_{ij\bullet} - \bar{x}_{i\bullet\bullet} + \bar{x})^2$	$(K - 1)(H - 1)$	**(17.23)**
Error:	$SCE = \sum_i \sum_j \sum_l (x_{ijl} - \bar{x}_{ij\bullet})^2$	$KH(L - 1)$	**(17.24)**

Entonces

$$STC = SCG + SCB + SCI + SCE \qquad \textbf{(17.25)}$$

Dividiendo las sumas de los cuadrados de los componentes por sus grados de libertad correspondientes, tenemos las medias de los cuadrados *MCG*, *MCB*, *MCI* y *MCE*.

Los contrastes de las hipótesis de que no hay efectos de los grupos, de los bloques y de la interacción se basan en los respectivos cocientes F:

$$\frac{MCG}{MCE} \qquad \frac{MCB}{MCE} \qquad \frac{MCI}{MCE}$$

Los contrastes se realizan comparando estas cifras con las distribuciones F con los correspondientes grados de libertad del numerador y el denominador. Su validez se basa en el supuesto de que los ε_{ijl} se comportan como una muestra aleatoria extraída de una distribución normal.

La Figura 17.7 muestra que la descomposición de la suma total de los cuadrados de las observaciones muestrales en torno a su media global es la suma de cuatro componentes. Se diferencia de la Figura 17.5 en que, cuando se replica el experimento, ahora podemos aislar la suma de los cuadrados de las interacciones.

Los cálculos pueden resumirse, al igual que antes, en una tabla del análisis de la varianza. La Tabla 17.12 muestra la forma general de la tabla cuando hay L observaciones por celda en un análisis de la varianza de dos factores.

De hecho, existen fórmulas más sencillas para calcular las distintas sumas de los cuadrados. No obstante, los cálculos aritméticos son bastante tediosos y deben realizarse por computador. No entraremos aquí en más detalles sino que nos limitaremos a mostrar en la Figura 17.8 los resultados de los cálculos basados en nuestros datos. En la práctica, los

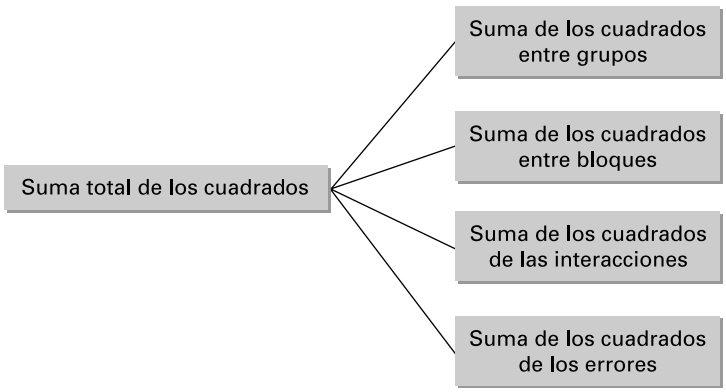


Figura 17.7. Descomposición de la suma de los cuadrados de un análisis de la varianza bifactorial con más de una observación por celda.

Tabla 17.12. Formato general de la tabla del análisis de la varianza bifactorial con L observaciones por celda.

Fuente de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente F
Entre grupos	SCG	$K-1$	$MCG = \dfrac{SCG}{K-1}$	$\dfrac{MCG}{MCE}$
Entre bloques	SCB	$H-1$	$MCB = \dfrac{SCB}{H-1}$	$\dfrac{MCB}{MCE}$
Interacción	SCI	$(K-1)(H-1)$	$MCI = \dfrac{SCI}{(K-1)(H-1)}$	$\dfrac{MCI}{MCE}$
Error	SCE	$KH(L-1)$	$MCE = \dfrac{SCE}{KH(L-1)}$	
Total	STC	$KHL-1$		

cálculos del análisis de la varianza normalmente se realizan utilizando un paquete estadístico como Minitab, por lo que raras veces la complejidad aritmética limita los análisis prácticos.

Los grados de libertad de la Figura 17.8 se deducen del hecho de que en el caso de estos datos tenemos que

$$K = 3 \qquad H = 5 \qquad L = 3$$

Las medias de los cuadrados se obtienen dividiendo las sumas de los cuadrados por los grados de libertad correspondientes. Por último, los cocientes F se obtienen dividiendo, a su vez, cada una de las tres primeras medias de los cuadrados por la media de los cuadrados de los errores.

Utilizando la Figura 17.8, podemos contrastar las tres hipótesis nulas de interés. En primer lugar, contrastamos la hipótesis nula de que no existe ninguna interacción entre los conductores y el tipo de automóvil. Este contraste se basa en el cociente F calculado, 21,35, y el p-valor de 0,000. Dado que los grados de libertad del numerador y del denominador son 8 y 30, respectivamente, vemos en el apéndice que

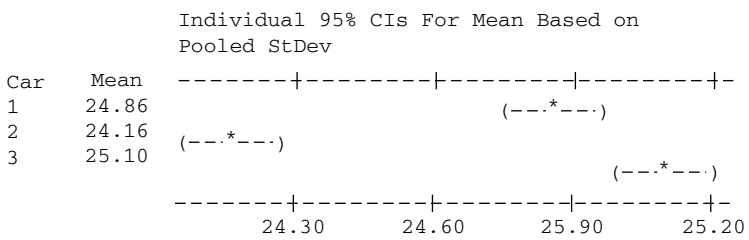
$$F_{8,\,30,\,0{,}01} = 3{,}17$$

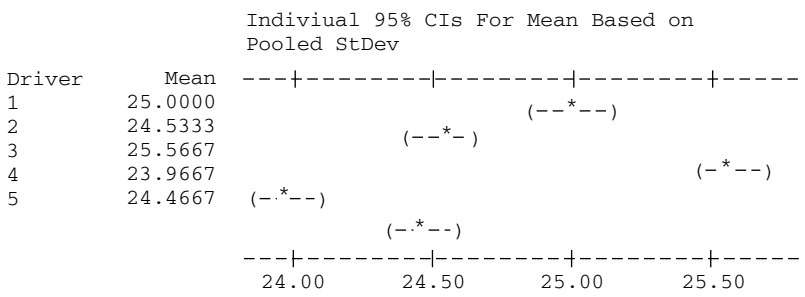
La hipótesis nula de que no existe ninguna interacción entre el tipo de automóvil y el conductor se rechaza claramente al nivel de significación del 1 por ciento.

Figura 17.8. Análisis de la varianza de los datos sobre el consumo de combustible de la Tabla 17.10 (salida Minitab).

Two-way ANOVA: Mileage versus Car, Driver

```
             DF        SS       MS       F       P
Car           2     7.156  3.57800   92.53   0.000
Driver        4    13.148  3.28700   85.01   0.000
Interaction   8     6.604  0.82550   21.35   0.000
Error        30     1.160  0.03867
Total        44    28.068

S = 0.1966     R-Sq = 95.87%     R-Sq(adj) = 93.94%
```





A continuación, contrastamos la hipótesis nula de que las medias poblacionales del consumo de combustible de los automóviles X, Y y Z son iguales. El contraste se basa en el cociente F calculado, 92,53. Vemos en la Tabla 9 del apéndice que en el caso de un contraste al 1 por ciento con 2 y 30 grados de libertad en el numerador y en el denominador, respectivamente,

$$F_{2,\,30,\,0,01} = 5{,}39$$

Por lo tanto, se rechaza abrumadoramente la hipótesis nula de la igualdad de las medias poblacionales del consumo de combustible de los tipos de automóvil al nivel de significación del 1 por ciento.

Por último, contrastamos la hipótesis nula de que las medias poblacionales del consumo de combustible de los cinco grupos de edad de los conductores son iguales. Vemos en la Figura 17.8 que el contraste se basa en el cociente F calculado, 85,01. Por lo tanto, los grados de libertad del numerador y del denominador son 4 y 30, respectivamente, por lo que en un contraste al nivel de significación del 1 por ciento,

$$F_{4,\,30,\,0,01} = 4{,}02$$

La hipótesis nula de la igualdad de las medias poblacionales del consumo de combustible de los grupos de edad de los conductores se rechaza claramente al nivel de significación del 1 por ciento.

Nuestros datos permiten extraer claramente las tres conclusiones siguientes:

1. El consumo medio de combustible de los automóviles X, Y y Z no es el mismo.
2. El consumo medio de combustible de los conductores de todos los grupos de edad no es el mismo.
3. Las diferencias entre los conductores no están distribuidas uniformemente entre los tres tipos de automóviles sino que es probable que un conductor de un grupo de edad obtenga unos resultados relativamente mejores en un tipo de automóvil que en otro en comparación con otros conductores.

En este apartado hemos supuesto hasta ahora que todas las celdas tenían el mismo número de observaciones. Sin embargo, esta restricción no es necesaria y a veces puede ser incómoda para el investigador. De hecho, las fórmulas para calcular las sumas de los cuadrados pueden modificarse para que las celdas no tengan todas ellas el mismo número de observaciones. No nos interesa aquí entrar en los detalles técnicos del cálculo de las sumas de los cuadrados. Generalmente, los investigadores disponen de paquetes estadísticos para realizarlo. Lo que nos interesa es el análisis de los resultados.

EJEMPLO 17.4. Nivel de satisfacción de los trabajadores (análisis de la varianza de dos factores)

Existe un estudio (véase la referencia bibliográfica 1) en el que se comparan los niveles de satisfacción de los trabajadores introvertidos y extrovertidos que realizan tareas estimulantes y no estimulantes. Para realizar este estudio, se utilizaron dos tipos de trabajadores y dos tipos de tareas, lo que nos da cuatro combinaciones. Las medias muestrales de los niveles de satisfacción declarados por los trabajadores de estas cuatro combinaciones fueron:

Trabajador introvertido, tarea no estimulante (16 observaciones):	2,78
Trabajador extrovertido, tarea no estimulante (15 observaciones):	1,85
Trabajador introvertido, tarea estimulante (17 observaciones):	3,87
Trabajador extrovertido, tarea estimulante (19 observaciones):	4,12

La tabla siguiente muestra las sumas de los cuadrados calculadas y los grados de libertad correspondientes. Complete la tabla del análisis de la varianza y analice los resultados de este experimento.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Tarea	62,04	1
Tipo de trabajador	0,06	1
Interacción	1,85	1
Error	23,31	63
Total	87,26	66

Solución

Una vez más, las medias de los cuadrados se obtienen dividiendo las sumas de los cuadrados por sus grados de libertad correspondientes. Los cocientes F se deducen de la división de las medias de los cuadrados de las tareas, de los tipos de trabajadores y de

las interacciones por la media de los cuadrados de los errores. Ahora podemos completar la tabla del análisis de la varianza.

Fuente de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente *F*
Tarea	62,04	1	62,04	167,68
Tipo de trabajador	0,06	1	0,06	0,16
Interacción	1,85	1	1,85	5,00
Error	23,31	63	0,37	
Total	87,26	66		

La tabla del análisis de la varianza puede utilizarse para contrastar tres hipótesis nulas. En el caso de la hipótesis nula de la igualdad de las medias poblacionales de los niveles de satisfacción con los dos tipos de tarea, el cociente *F* calculado es 167,68. Tenemos 1 grado de libertad en el numerador y 63 en el denominador, por lo que vemos en el apéndice que en un contraste al 1 por ciento

$$F_{1,\,63,\,0,01} = 7,07$$

Por lo tanto, se rechaza claramente la hipótesis nula de la igualdad de las medias poblacionales de los niveles de satisfacción con las tareas estimulantes y no estimulantes. Este resultado no es sorprendente. Sería lógico esperar que los trabajadores estuvieran más satisfechos realizando tareas estimulantes que realizando tareas no estimulantes.

A continuación, contrastamos la hipótesis nula de que las medias poblacionales de los niveles de satisfacción de los trabajadores introvertidos y extrovertidos son iguales. En este caso, el cociente *F* calculado es 0,16. Una vez más, los grados de libertad son 1 y 63, por lo que en el caso de un contraste al 5 por ciento,

$$F_{1,\,63,\,0,05} = 4,00$$

La hipótesis nula de la igualdad de los niveles medios de satisfacción de los trabajadores introvertidos y extrovertidos no puede rechazarse al nivel de significación del 5 por ciento.

En muchos estudios, el término de interacción no es en sí mismo muy importante. Se incluye en el análisis principalmente para «absorber» parte de la variabilidad de los datos y poder detectar así más fácilmente las diferencias que pueda haber entre las medias poblacionales. Sin embargo, en este estudio la interacción es muy interesante. La hipótesis nula de que no existe ninguna interacción entre la tarea y el tipo de trabajador en la determinación de los niveles de satisfacción de los trabajadores se contrasta por medio del cociente *F* calculado de 5,00. Una vez más, los grados de libertad del numerador y del denominador son 1 y 63, respectivamente. Por lo tanto, la comparación con los valores calculados de la distribución *F* revela que la hipótesis nula de que no existe ninguna interacción puede rechazarse al nivel del 5 por ciento, pero no al nivel de significación del 1 por ciento.

EJERCICIOS

Ejercicios básicos

17.41. Considere un experimento en el que los factores de tratamiento son A y B y el factor A tiene cuatro niveles y el B tiene tres niveles. La tabla del análisis de la varianza adjunta resume los resultados del experimento.

Calcule las medias de los cuadrados y contraste las hipótesis nulas de que no hay ningún efecto de ninguno de los dos tratamientos y ningún efecto de interacción.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Grupos de tratamiento A	71	3
Grupos de tratamiento B	63	2
Interacción	50	6
Error	280	60
Total	464	71

17.42. Considere un experimento en el que los factores de tratamiento son A y B y el factor A tiene cinco niveles y el B tiene seis niveles. La tabla del análisis de la varianza adjunta resume los resultados del experimento.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Grupos de tratamiento A	86	4
Grupos de tratamiento B	75	5
Interacción	75	20
Error	300	90
Total	536	119

Calcule las medias de los cuadrados y contraste las hipótesis nulas de que no hay ningún efecto de ninguno de los dos tratamientos y ningún efecto de interacción.

17.43. Considere un experimento en el que los factores de tratamiento son A y B y el factor A tiene tres niveles y el B tiene siete niveles. La tabla del análisis de la varianza adjunta resume los resultados del experimento.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Grupos de tratamiento A	37	2
Grupos de tratamiento B	58	6
Interacción	57	12
Error	273	84
Total	425	104

Calcule la media de los cuadrados y contraste las hipótesis nulas de que no hay ningún efecto de ninguno de los dos tratamientos y ningún efecto de interacción.

Ejercicios aplicados

17.44. Suponga que analiza las puntuaciones dadas por los jueces en los saltos de esquí de las olimpiadas de invierno. Suponga que hay 22 participantes y nueve jueces. Cada juez puntúa a cada participante en siete pruebas. Las puntuaciones pueden analizarse, pues, en el marco de un análisis de la varianza de dos factores con 198 celdas participante-juez, siete observaciones por celda. La tabla adjunta muestra las sumas de los cuadrados.

Fuente de variación	Suma de los cuadrados
Entre participantes	364,50
Entre jueces	0,81
Interacción	4,94
Error	1.069,94

a) Complete la tabla del análisis de la varianza.
b) Realice los contrastes F correspondientes e interprete sus resultados.

17.45. Vuelva al ejercicio 17.44. En la competición de patinaje artístico participan doce parejas. Una vez más, hay nueve jueces y se puntúa a los participantes en siete pruebas. Las sumas de los cuadrados entre los grupos (parejas de participantes) y entre los bloques (jueces) son

$$SCG = 60{,}10 \quad \text{y} \quad SCB = 1{,}65$$

mientras que la suma de los cuadrados de las interacciones y de los errores son

$$SCI = 3{,}35 \quad \text{y} \quad SCE = 31{,}61$$

Analice estos resultados e interprete verbalmente las conclusiones.

17.46. Un psicólogo está trabajando con tres tipos de tests de aptitud que pueden hacerse a las personas que solicitan empleo. Una cuestión importante para estructurar los tests es la posibilidad de que exista interacción entre los que solicitan empleo y el tipo de test. Si no hubiera ninguna interacción, sólo sería necesario un tipo de test. Se realizan tres tests de cada tipo (A, B y C) a los miembros de cada uno de los cuatro grupos de solicitantes de empleo. Éstos se distinguen

por las valoraciones de malo, regular, bueno y excelente en las entrevistas preliminares. Las puntuaciones obtenidas se muestran en la tabla adjunta.

	Tipo de test								
Tipo de sujeto	**A**			**B**			**C**		
Malo	65	68	62	69	71	67	75	75	78
Regular	74	79	76	72	69	69	70	69	65
Bueno	64	72	65	68	73	75	78	82	80
Excelente	83	82	84	78	78	75	76	77	75

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que no existe ninguna interacción entre el tipo de sujeto y el tipo de test.

17.47. Se pide a muestras aleatorias de dos estudiantes universitarios de primer año, dos de segundo año, dos de tercer año y dos de cuarto año de cuatro residencias universitarias que valoren en una escala de 1 (mala) a 10 (excelente) la calidad del ambiente de la residencia para estudiar. La tabla muestra los resultados.

	Residencia							
Año	**A**		**B**		**C**		**D**	
Primer año	7	5	8	6	9	8	9	9
Segundo año	6	8	5	5	7	8	8	9
Tercer año	5	4	7	6	6	7	7	8
Cuarto año	7	4	6	8	7	5	6	7

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de las cuatro residencias son iguales.
c) Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de los cuatro tipos de estudiantes son iguales.
d) Contraste la hipótesis nula de que no existe ninguna interacción entre el año de estudios y la valoración de la residencia.

17.48. En algunos experimentos con varias observaciones por celda, el analista está dispuesto a suponer que no existe ninguna interacción entre los grupos y los bloques. Las interacciones que pueda haber se atribuyen a un error aleatorio. Cuando se postula ese supuesto, el análisis se realiza como siempre, con la salvedad de que se suma lo que antes eran las sumas de los cuadrados de las interacciones y de los errores para formar una nueva suma de los cuadrados de los errores. También se suman los grados de libertad correspondientes. Si el supuesto de la ausencia de interacciones es correcto, este enfoque tiene la ventaja de que aumentan los grados de libertad de los errores y, por lo tanto, los contrastes de la igualdad de las medias de los grupos y de los bloques son más poderosos. Para estudiar el ejercicio 17.47 supongamos que ahora postulamos el supuesto de que no hay interacciones entre la valoración de la residencia y el año de estudios del alumno.

a) Explique verbalmente las implicaciones de este supuesto.
b) Dado este supuesto, elabore la nueva tabla del análisis de la varianza.
c) Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de las cuatro residencias son iguales.
d) Contraste la hipótesis nula de que las medias poblacionales de las valoraciones de los cuatro tipos de estudiantes son iguales.

17.49. Vuelva al ejercicio 17.31. Una vez realizado el experimento para comparar el rendimiento medio por acre de cuatro variedades de maíz y tres marcas de fertilizante, un investigador agrario sugirió que podía existir alguna interacción entre la variedad y el fertilizante. Para comprobar esta posibilidad, se realizó otra serie de pruebas, que dieron los rendimientos que se muestran en la tabla.

	Variedad			
Fertilizante	**A**	**B**	**C**	**D**
1	80	88	73	88
2	94	91	79	93
3	81	78	83	83

a) ¿Qué implicaría una interacción entre la variedad y el fertilizante?
b) Combine los datos de los dos conjuntos de pruebas y elabore una tabla del análisis de la varianza.
c) Contraste la hipótesis nula de que las medias poblacionales del rendimiento de las cuatro variedades de maíz son iguales.
d) Contraste la hipótesis nula de que las medias poblacionales del rendimiento de las tres marcas de fertilizante son iguales.
e) Contraste la hipótesis nula de que no existe ninguna interacción entre la variedad de maíz y la marca del fertilizante.

17.50. Vuelva al ejercicio 17.33. Suponga que se añade al estudio una segunda tienda para cada

combinación de región y color de las latas y se obtienen los resultados que muestra la tabla adjunta. Combinando estos resultados con los del ejercicio 17.33, realice los cálculos del análisis de la varianza y analice sus resultados.

	Color de la lata		
Región	**Rojo**	**Amarillo**	**Azul**
Este	45	50	54
Sur	49	51	58
Norte	43	60	50
Oeste	38	49	44

17.51. Una vez realizado el estudio del ejercicio 17.34, el profesor decidió repetirlo un año más tarde. La tabla muestra los resultados obtenidos. Combinando estos resultados con los del ejercicio 17.34, realice los cálculos del análisis de la varianza y analice sus resultados.

	Libro de texto		
Examen	**A**	**B**	**C**
Tipo test	4,7	5,1	4,8
Redacción	4,4	4,6	4,0
Mezcla	4,5	5,3	4,9

RESUMEN

En este capítulo hemos presentado los componentes básicos del método del análisis de la varianza. El análisis de la varianza permite averiguar si uno o más factores cuya dimensión es discreta influyen en la medición de los resultados. Estos procedimientos son fundamentales para el diseño experimental y son utilizados frecuentemente por la industria para saber cuáles son las mejores prácticas para maximizar la productividad y reducir lo más posible los defectos. El análisis de la varianza de un factor es un método para comparar simultáneamente las medias de tres procesos o más. También hemos incluido el contraste de Kruskal-Wallis por ser un útil método no paramétrico para comparar tres o más grupos utilizando datos ordenados. El análisis de la varianza bifactorial considera el efecto que producen dos factores que pueden adoptar varios valores en la medición de los resultados. Podemos considerar el efecto de cada factor por separado y, utilizando celdas con múltiples observaciones, también podemos examinar la interacción entre combinaciones específicas de niveles de los factores. Los métodos del análisis de la varianza son un complemento del análisis de regresión múltiple. También pueden lograrse los mismos objetivos utilizando los procedimientos de las variables ficticias, analizados en el Capítulo 14.

TÉRMINOS CLAVE

análisis de la varianza de un factor, 684
análisis de la varianza bifactorial: una observación por celda, 698
análisis de la varianza bifactorial: varias observaciones por celda, 713
contraste de hipótesis del análisis de la varianza de un factor, 688
contraste de Kruskal-Wallis, 695
contrastes de hipótesis del análisis de la varianza bifactorial, 705
media de los cuadrados, 705
descomposición de la suma de los cuadrados en el análisis de la varianza de un factor, 687
descomposición de la suma de los cuadrados en el análisis de la varianza bifactorial, 703
diseño de bloques aleatorizados, 699
interacción, 709
tabla del análisis de la varianza bifactorial, 706

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

17.52. Distinga detenidamente entre el análisis de la varianza de un factor y el bifactorial. Ponga ejemplos distintos a los que se analizan en el libro y formule problemas empresariales para los que podría ser adecuado cada uno.

17.53. Explique detenidamente qué se entiende por efecto de interacción en el análisis de la varianza bifactorial con más de una observación por celda. Ponga un ejemplo de este efecto en problemas relacionados con el mundo de la empresa.

17.54. Considere un estudio que pretende evaluar la facilidad de lectura de los mensajes de los informes financieros. La eficacia del mensaje escrito se evalúa utilizando un procedimiento tradicional. Se entregan informes financieros a muestras aleatorias independientes de tres grupos: auditores, analistas financieros y responsables de la concesión de préstamos de bancos comerciales en periodo de formación y se anotan las puntuaciones de los miembros de las muestras. La hipótesis nula que se pretende contrastar es que las medias poblacionales de las puntuaciones de los tres grupos son idénticas. Contraste esta hipótesis, dada la información de la tabla adjunta.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre grupos	5.165	2
Dentro de grupos	120.802	1.005
Total	125.967	1.007

17.55. En un experimento realizado para evaluar las ayudas que reciben los profesores universitarios en sus entrevistas con los alumnos graduados a los que supervisan, se asignaron aleatoriamente entrevistadores a uno de los tres tipos de entrevistas: con información sobre entrevistas anteriores, planteando objetivos para la entrevista y grupo de control. En el caso del primer tipo de entrevista, los entrevistadores podían examinar y discutir las reacciones de los estudiantes a entrevistas anteriores. En el caso del segundo tipo, se les animaba a fijar objetivos para la siguiente entrevista. En el caso del grupo de control, las entrevistas se realizaron como siempre sin conocer las entrevistas anteriores y sin fijar objetivos. Una vez terminadas las entrevistas, se valoraron los niveles de satisfacción de los estudiantes con las entrevistas. El nivel medio de satisfacción de las 45 personas del grupo que realizó el primer tipo de entrevista era de 13,98. El de las 49 personas del grupo que realizó el segundo tipo de entrevista era de 17,12, mientras que el de los 41 miembros del grupo de control era de 13,07. El cociente F calculado a partir de los datos era 4,12.

a) Elabore la tabla completa del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de los niveles de satisfacción de los tres tipos de entrevistas son iguales.

17.56. En un estudio se clasificó a 134 abogados en cuatro grupos basándose en la observación y en una entrevista. Se consideró que los 62 abogados del grupo A tenían un elevado nivel de estímulo y de apoyo y un nivel medio de espíritu cívico. Los 52 abogados del grupo B tenían un bajo nivel de estímulo, un nivel medio de apoyo y un elevado nivel de espíritu cívico. El grupo C contenía 7 abogados que tenían un nivel medio de estímulo, un bajo nivel de apoyo y un bajo nivel de espíritu cívico. Los 13 abogados del grupo D tenían un bajo nivel en los tres aspectos. Se compararon los sueldos de estos cuatro grupos. Las medias muestrales eran 7,87 en el caso del grupo A, 7,47 en el del grupo B, 5,14 en el del grupo C y 3,69 en el del grupo D. El cociente F calculado a partir de estos datos era 25,60.

a) Elabore la tabla completa del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de los sueldos de los abogados de estos cuatro grupos eran iguales.

17.57. En un estudio para estimar la influencia del consumo de tabaco en la salud, se clasificaron los empleados en empleados fumadores, empleados que han dejado de fumar recientemente, empleados que dejaron de fumar hace tiempo y empleados que nunca han fumado. Se tomaron muestras de 96, 34, 86 y 206 miembros de estos grupos. Se observó que el número mensual medio de visitas al médico era de 2,15, 2,21, 1,47 y 1,69, respectivamente. El cociente F calculado a partir de estos datos era 2,56.

a) Elabore la tabla completa del análisis de la varianza.
b) Contraste la hipótesis nula de la igualdad de las medias poblacionales de las tasas de riesgo para la salud de los cuatro grupos.

17.58. En un país existen restricciones sobre los anuncios de bebidas alcohólicas. Sin embargo, durante un tiempo, se suprimieron estas restricciones. Se recogieron datos sobre las ventas totales de vino en tres periodos: durante el periodo de restricciones de la publicidad, durante el periodo en el que se eliminaron las restricciones y durante el periodo en que volvieron a establecerse. La tabla adjunta muestra las sumas de los cuadrados y los grados de libertad. Suponiendo que se satisfacen los requisitos habituales del análisis de la varianza —en concreto, que las observaciones muestrales son independientes

entre sí—, contraste la hipótesis nula de la igualdad de las medias poblacionales de las ventas de estos tres periodos de tiempo.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre grupos	11.438,3028	2
Dentro de grupos	109.200,0000	15
Total	120.638,3028	17

17.59. Se toman muestras aleatorias de los precios de venta de las viviendas de cuatro distritos. La tabla adjunta muestra los precios de venta (en miles de dólares). Contraste la hipótesis nula de que las medias poblacionales de los precios de venta de los cuatro distritos son iguales.

Distrito A	Distrito B	Distrito C	Distrito D
73	85	97	61
63	59	86	67
89	84	76	84
75	70	78	67
70	80	76	69

17.60. Basándose en los datos del ejercicio 17.59, utilice el contraste de Kruskal-Wallis para contrastar la hipótesis nula de que las medias poblacionales de los precios de venta de las viviendas de los cuatro distritos son iguales.

17.61. Un estudio pretendía valorar los niveles de satisfacción con los horarios laborales en una escala de 1 (muy insatisfecho) a 7 (muy satisfecho) de los profesores interinos, asociados y ayudantes. El nivel medio de satisfacción de una muestra de 25 interinos era de 6,60; el de una muestra de 24 asociados era de 5,37; el de una muestra de 20 ayudantes era de 5,20. El cociente F calculado a partir de estos datos era 6,62.

a) Elabore la tabla completa del análisis de la varianza.

b) Contraste la hipótesis nula de la igualdad de las medias poblacionales de los niveles de satisfacción de los tres grupos.

17.62. Considere el análisis de la varianza de un factor.

a) Demuestre que la suma de los cuadrados dentro de los grupos puede expresarse de la forma siguiente:

$$SCD = \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ji}^2 - \sum_{i=1}^{K} n_i \bar{x}_i^2$$

b) Demuestre que la suma de los cuadrados entre los grupos puede expresarse de la forma siguiente:

$$SCG = \sum_{i=1}^{K} n_i \bar{x}_i^2 - n\bar{x}^2$$

c) Demuestre que la suma total de los cuadrados puede expresarse de la forma siguiente:

$$STC = \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij}^2 - n\bar{x}^2$$

17.63. Considere el análisis de la varianza de dos factores con una observación por celda.

a) Demuestre que la suma de los cuadrados entre los grupos puede expresarse de la forma siguiente:

$$SCG = H \sum_{i=1}^{K} \bar{x}_{i\bullet}^2 - n\bar{x}^2$$

b) Demuestre que la suma de los cuadrados entre los bloques puede expresarse de la forma siguiente:

$$SCB = K \sum_{j=1}^{H} \bar{x}_{\bullet j}^2 - n\bar{x}^2$$

c) Demuestre que la suma total de los cuadrados puede expresarse de la forma siguiente:

$$STC = \sum_{i=1}^{K} \sum_{j=1}^{H} x_{ji}^2 - n\bar{x}^2$$

d) Demuestre que la suma de los cuadrados de los errores puede expresarse de la forma siguiente:

$$SCE = \sum_{i=1}^{K} \sum_{j=1}^{H} x_{ji}^2 - H \sum_{i=1}^{K} \bar{x}_{i\bullet}^2 - K \sum_{j=1}^{H} \bar{x}_{\bullet j}^2 - n\bar{x}^2$$

17.64. Se ha obtenido de una muestra aleatoria de 125 consumidores información sobre su satisfacción con tres grupos de precios de la cerveza: alto, medio y bajo. La tabla adjunta muestra las sumas de los cuadrados de estas medidas de la satisfacción. Complete la tabla del análisis de la varianza y contraste la hipótesis nula de que los niveles medios de satisfacción con los tres grupos de precios son iguales.

Fuente de variación	Suma de los cuadrados
Entre los consumidores	37.571,5
Entre las marcas	32.987,3
Error	55.710,7

17.65. Se pide a tres agencias inmobiliarias que valoren cinco viviendas de un barrio. La tabla muestra los resultados en miles de dólares. Elabore una tabla del análisis de la varianza y contraste la hipótesis nula de que las valoraciones medias de las tres agencias son iguales.

	Agencia		
Vivienda	**A**	**B**	**C**
1	210	218	226
2	192	190	198
3	183	187	185
4	227	223	237
5	242	240	237

17.66. Los estudiantes se clasifican en función de tres grupos de renta de sus padres y de tres notas posibles en el examen de acceso a la universidad. Se elige aleatoriamente un estudiante de cada una de las nueve combinaciones posibles y se anota la calificación media al final del primer año. La tabla adjunta muestra los resultados.

	Grupo de renta		
Nota de acceso a la universidad	**Alta**	**Moderada**	**Baja**
Muy alta	3,7	3,6	3,6
Alta	3,4	3,5	3,2
Moderada	2,9	2,8	3,0

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de las calificaciones medias del primer año de los estudiantes de los tres grupos de renta son iguales.
c) Contraste la hipótesis nula de que las medias poblacionales de las calificaciones medias del primer año de los estudiantes de los tres grupos de notas de acceso a la universidad son iguales.

17.67. En el modelo del análisis de la varianza bifactorial con una observación por celda, expresamos la observación del i-ésimo grupo y del j-ésimo bloque de la forma siguiente:

$$X_{ij} = \mu + G_i + B_j + \varepsilon_{ij}$$

Vuelva al ejercicio 17.65 y considere la observación sobre la agencia B y la vivienda 1 ($x_{21} = 218$).

a) Estime μ.
b) Estime e interprete G_2.
c) Estime e interprete B_1.
d) Estime ε_{21}.

17.68. Vuelva al ejercicio 17.66 y considere la observación sobre el grupo de renta moderada y una nota alta en el examen de acceso a la universidad ($x_{22} = 3{,}5$).

a) Estime μ.
b) Estime e interprete G_2.
c) Estime e interprete B_1.
d) Estime ε_{21}.

17.69. Considere el análisis de la varianza bifactorial con L observaciones por celda.

a) Demuestre que la suma de los cuadrados entre los grupos puede expresarse de la forma siguiente:

$$SCG = HL \sum_{i=1}^{K} \bar{x}_{i\bullet\bullet}^2 - HKL\bar{x}^2$$

b) Demuestre que la suma de los cuadrados entre los bloques puede expresarse de la forma siguiente:

$$SCB = KL \sum_{j=1}^{H} \bar{x}_{\bullet j\bullet}^2 - HKL\bar{x}^2$$

c) Demuestre que la suma de los cuadrados de los errores puede expresarse de la forma siguiente:

$$SCE = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{l=1}^{L} x_{ijl}^2 - L\sum_{i=1}^{K}\sum_{j=1}^{H} \bar{x}_{ij\bullet}^2$$

d) Demuestre que la suma total de los cuadrados puede expresarse de la forma siguiente:

$$STC = \sum_{i=1}^{K}\sum_{j=1}^{H}\sum_{l=1}^{L} x_{ijl}^2 - HKL\bar{x}^2$$

e) Demuestre que la suma de los cuadrados de las interacciones puede expresarse de la forma siguiente:

$$SCI = L\sum_{i=1}^{K}\sum_{j=1}^{H} \bar{x}_{ij\bullet}^2 - HL\sum_{i=1}^{K} \bar{x}_{i\bullet\bullet}^2 - KL\sum_{j=1}^{H} \bar{x}_{\bullet j\bullet}^2 - HKL\bar{x}^2$$

17.70. Unos agentes de compra reciben información sobre un sistema de telefonía móvil y se les pide que valoren su calidad. La información que reciben es idéntica, salvo por dos factores: el precio y el país de origen. En el caso del precio, hay tres posibilidades: 150 $, 80 $ y ningún precio. En el caso del país de origen, tam-

bién hay tres posibilidades: Estados Unidos, Taiwán y ningún país. Aquí se muestra parte de la tabla del análisis de la varianza de las valoraciones de la calidad realizadas por los agentes de compra. Complete la tabla del análisis de la varianza y realice un análisis completo de estos datos.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre los precios	0,178	2
Entre los países	4,365	2
Interacción	1,262	4
Error	93,330	99

17.71. En el estudio del ejercicio 17.70, también se da información a estudiantes de un máster de administración de empresas. Aquí se muestra parte de la tabla del análisis de la varianza de las valoraciones de la calidad realizadas por los estudiantes. Complete la tabla del análisis de la varianza y realice un análisis completo de estos datos.

Fuente de variación	Suma de los cuadrados	Grados de libertad
Entre los precios	0,042	2
Entre los países	17,319	2
Interacción	2,235	4
Error	70,414	45

17.72. Una vez realizado el estudio del ejercicio 17.66, el investigador decide tomar una segunda muestra aleatoria independiente de un estudiante de cada una de las nueve categorías renta-nota del examen de acceso a la universidad. La tabla adjunta muestra las calificaciones medias obtenidas.

	Grupo de renta		
Nota de acceso a la universidad	Alta	Moderada	Baja
Muy alta	3,9	3,7	3,8
Alta	3,2	3,6	3,4
Moderada	2,7	3,0	2,8

a) Elabore la tabla del análisis de la varianza.
b) Contraste la hipótesis nula de que las medias poblacionales de las calificaciones medias del primer año de los estudiantes de los tres grupos de renta son iguales.
c) Contraste la hipótesis nula de que las medias poblacionales de las calificaciones medias del primer año de los estudiantes de los tres grupos de notas del examen de acceso a la universidad son iguales.
d) Contraste la hipótesis nula de que no existe ninguna interacción entre el grupo de renta y la nota del examen de acceso a la universidad.

17.73. Se realiza un experimento para contrastar los efectos que producen en los rendimientos cinco variedades de maíz y cinco tipos de fertilizante. Se utilizan para cada combinación variedad-fertilizante seis gráficos y se anotan los rendimientos. La tabla muestra los resultados.

Tipo de fertilizante	Variedad de maíz									
	A		B		C		D		E	
1	75	77	74	67	93	90	79	83	72	77
	79	83	73	65	87	82	87	88	79	83
	85	78	79	80	86	88	86	90	78	86
2	80	72	71	69	84	88	77	82	70	75
	76	73	75	62	90	79	84	87	80	80
	70	74	77	63	83	80	82	83	74	81
3	85	87	76	73	88	94	81	86	77	83
	80	79	77	70	89	86	90	90	87	79
	87	80	83	80	89	93	87	88	86	88
4	80	79	74	77	86	87	80	77	79	85
	82	77	69	78	90	85	90	84	88	80
	85	80	74	76	83	88	80	88	87	82
5	75	79	75	80	92	88	82	78	80	87
	86	82	84	80	89	94	85	86	90	83
	79	83	72	77	86	90	82	89	86	83

a) Contraste la hipótesis nula de que los rendimientos medios de las cinco variedades de maíz son iguales.
b) Contraste la hipótesis nula de que los rendimientos medios de las cinco marcas de fertilizante son iguales.
c) Contraste la hipótesis nula de que no existe ninguna interacción entre la variedad y el fertilizante.

Apéndice

1. Suma total de los cuadrados

$$
\begin{aligned}
STC &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 \\
&= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\bar{x})^2 \\
&= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{i=1}^{K}\sum_{j=1}^{n_j}(\bar{x}_i-\bar{x})^2+2\sum_{i=1}^{K}(\bar{x}_i-\bar{x})\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_j) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{i=1}^{K}n_i(\bar{x}_i-\bar{x})^2
\end{aligned}
$$

$$STC = SCD + SCG$$

Nota: $\sum_{j=1}^{n_j}(x_{ij}-\bar{x}_i)=0$

2. Media de los cuadrados dentro de los grupos (*MCD*)

Para cada subgrupo i

$$
\begin{aligned}
\sigma^2 &= E\left[\frac{\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}{n_i}\right] \\
&= E\left[\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i+\bar{x}_i-\mu_i)^2}{n_i}\right] \\
&= E\left[\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n_i}\right]+\frac{\sigma^2}{n_i}
\end{aligned}
$$

$$\frac{(n_i-1)\sigma^2}{n_i} = E\left[\frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n_i}\right]$$

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n_i-1}$$

Sumando los valores de los K subgrupos

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}{n - K} = \frac{SCD}{n - K}$$

$$\hat{\sigma}^2 = MCD$$

3. Media de los cuadrados entre los grupos (*MCG*)

$$\mu_i = \mu \qquad i = 1, ..., K$$

Entonces

$$\sigma^2 = E\left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2}{n - 1}\right]$$

$$= E\left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2}{n - 1}\right]$$

$$= E\left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_j)^2}{n - 1} + \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2}{n - 1}\right]$$

$$= \frac{(n - K)\hat{\sigma}^2}{n - 1} + \frac{\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2}{n - 1}$$

$$\frac{(K - 1)\hat{\sigma}^2}{n - 1} = \frac{\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2}{n - 1}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2}{K - 1}$$

$$\hat{\sigma}^2 = MCG = \frac{SCG}{K - 1}$$

4. Cociente entre las medias de los cuadrados

Si

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$$

es verdadera, entonces *MCG* —con $K - 1$ grados de libertad— es un estimador de σ^2 y

$$\chi^2_{K-1} = \frac{(K-1)MCG}{\sigma^2}$$

Además, *MCD* con $n - K$ grados de libertad es un estimador de σ^2 y, por lo tanto,

$$\chi^2_{n-K} = \frac{(n-K)MCD}{\sigma^2}$$

Por lo tanto,

$$F_{K-1,n-K} = \frac{\dfrac{\chi^2_{K-1}}{K-1}}{\dfrac{\chi^2_{n-K}}{n-K}} = \frac{MCG}{MCD}$$

Bibliografía

1. Kim, J. S., «Relationships of Personality of Perceptual and Behavioral Responses in Stimulating and Nonstimulating Tasks», *Academy of Management Journal*, 23, 1980, págs. 307-319.
2. Shuptrine, F. K. y D. D. McVicker, «Readability Levels of Magazine Advertisements», *Journal of Advertising Research*, 21, n.º 5, 1981, págs. 45-50.

Capítulo 18

Introducción a la calidad

Esquema del capítulo

Introducción

En este capítulo introducimos métodos estadísticos que son bastante sencillos y que distan de ser nuevos. Estos métodos, llamados tradicionalmente *control estadístico de procesos* o *control estadístico de la calidad*, actualmente se incluyen, junto con otras técnicas de mejora de los procesos, en el estudio del control de la gestión y la calidad. Antes las empresas manufactureras aplicaban muchos métodos de mejora de los procesos para controlar los procesos de producción. Pronto se comprendieron los beneficios, y los centros educativos, los servicios sanitarios, los organismos públicos, el poder judicial, el sector del entretenimiento, el turismo, los transportes y otras muchas entidades de servicios, con fines de lucro y sin fines de lucro, adoptaron los principios de la calidad. Las empresas de transporte y las compañías aéreas controlan la llegada puntual de los paquetes y de los aviones; los restaurantes controlan la calidad de los alimentos, el tiempo de preparación y el servicio; los hoteles y los hospitales se preocupan por la satisfacción de los clientes. Las empresas y las organizaciones tienen clientes y los clientes demandan bienes y servicios de calidad. La calidad es esencial en todas las áreas y se aplica a todos los segmentos de la sociedad. Dado que es fundamental en la comunidad en general, la continua mejora de los procesos de producción, de los productos y de los servicios tiene una importancia primordial.

18.1. La importancia de la calidad

¿Qué tienen en común las siguientes empresas?: Nokia Mobile Phones, Europa y África (Finlandia); Inland Revenue, Account Office Cumbernauld (Reino Unido, Escocia); Burton-Apta Refractory Manufacturing Ltd. (Hungría); STMicroelectronics, Inc. (Carrollton, Texas); BI (Minneapolis, Minnesota); The Ritz-Carlton Hotel Company, L.L.C. (Atlanta, Georgia); Sunny Fresh Foods (Monticello, Minnesota); Motorola, Inc. (Schaumburg, Illinois); Texas Nameplate Company, Inc. (Dallas, Texas); Solectron Corporation (Milpitas, California); Xerox Business Services (Rochester, Nueva York); Wainwright Industries, Inc. (St. Peteres, Misuri); y Operations Management International, Inc. (que tiene oficinas en 29 estados de Estados Unidos, Brasil, Canadá, Egipto, Israel, Malasia, Nueva Zelanda, Filipinas y Tailandia)? Algunas son empresas manufactureras; otras son organizaciones de servicios. Algunas son grandes; otras son pequeñas. Algunas son estadounidenses; otras son europeas. Pero el denominador común de todas ellas (y ésta es una lista parcial) es que han recibido un prestigioso premio por su excelencia en la gestión y la continua mejora de la calidad. Por ejemplo, en el foro de la European Foundation for Quality Management (EFQM) celebrado en septiembre de 2000 en Estambul, Nokia Mobile Phones, Inland Revenue y Burton-Apta Refractory se sumaron a la lista de las organizaciones europeas más destacadas tanto del sector público como del sector privado que han recibido el Premio Europeo a la Calidad. La European Foundation ha concedido este premio por la gestión de la calidad desde 1992 (para más información, véase su página web en www.efqm.org). Las otras empresas mencionadas no son más que algunas de las que han recibido el Malcolm Baldrige National Quality Award, que es el principal premio estadounidense a la excelencia y la calidad, que se otorga desde 1988. Existen otros muchos premios a la calidad en los países y en las empresas. La calidad y la mejora continua tienen una importancia internacional.

Los líderes de la calidad

Consideremos la industria manufacturera. Es evidente que el objetivo no es simplemente inspeccionar un producto acabado. Para entonces poco puede hacerse salvo descartarlo o rehacer los artículos defectuosos, lo que supone un despilfarro considerable. Es esencial, por el contrario, controlar el proceso de producción *en cada una de las fases* en las que se produce un producto intermedio que debe satisfacer unas normas verificables. El objetivo es garantizar la calidad en cada fase del proceso de producción, para no perder tiempo y dinero en la producción de productos que no satisfacen las normas de calidad. En la mejora continua de los procesos, pues, se considera que cada fase de producción genera un producto cuya calidad debe evaluarse.

En la industria manufacturera de Estados Unidos, los métodos estadísticos de control de la calidad no se extendieron hasta la década de 1980, una década que fue testigo de una explosión del interés por estas técnicas. Sin embargo, como hemos indicado, los métodos estadísticos de control de la calidad no son, desde luego, nuevos. Tampoco son difíciles de entender o de aplicar. De hecho, los métodos básicos —los que se utilizan más a menudo hoy— no son más que aplicaciones bastante rutinarias de las técnicas estadísticas analizadas en capítulos anteriores de este libro. Aunque los métodos estadísticos de control de la calidad se menospreciaron en Estados Unidos durante muchos años, su desarrollo inicial se debió a un estadounidense, Walter A. Shewhart, quien en la década de 1920 defendió los métodos que subyacen a la metodología que hoy ha logrado una aceptación general.

De hecho, la aplicación general de las ideas de Shewhart en la industria manufacturera privada se logró primero en Japón tras la Segunda Guerra Mundial.

El control de la calidad está en la raíz de la ascensión de Japón hasta convertirse en un líder económico mundial. Su aplicación se debió mucho a la influencia de otro estadístico estadounidense, W. Edwards Deming, antiguo colega de Shewhart. Dos de los conceptos más importantes de Deming (véanse las referencias bibliográficas 2, 3 y 11) son:

1. La calidad es el resultado de un minucioso estudio de todo el proceso de producción y de la intervención directa de la dirección para corregir todos los pequeños problemas que contribuyen a los defectos.
2. Es necesario recoger datos periódicamente y analizarlos mediante métodos estadísticos adecuados para garantizar que el proceso funciona de una manera estable con una varianza mínima. Siempre que se identifican desviaciones de la norma, es necesario corregirlas inmediatamente.

Aparte de los estudios de W. Edward Deming, también contribuyeron al pensamiento moderno sobre la calidad los esfuerzos de Joseph Juran (véanse las referencias bibliográficas 7 y 8), Philip Crosby, Armand V. Feigenbaum, Kaoru Ishikawa y otros muchos. A pesar de las diferencias entre Deming, Juran y Crosby, todos coincidían en que el compromiso de los altos directivos es absolutamente necesario; en que la responsabilidad de la calidad corresponde a la dirección, no a los trabajadores; y en que la mejora es interminable (véase la referencia bibliográfica 5). Sus actividades introdujeron un importante cambio en la filosofía del control de la calidad. Tradicionalmente, se ponía el énfasis en la inspección de los productos finales, bien de todas las unidades, bien de una muestra aleatoria. Mediante esta inspección, se identificaban las unidades defectuosas, se eliminaban o se reparaban o se achatarraban. Según Deming, contrario a este enfoque, «la inspección con el objetivo de encontrar las unidades malas y eliminarlas es demasiado tardía, ineficaz y cara. En primer lugar, no se pueden encontrar las unidades malas, no todas ellas. En segundo lugar, cuesta demasiado. La calidad no se consigue inspeccionando sino mejorando el proceso» (véase la referencia bibliográfica 11). Si sólo se recurre a la inspección final, los trabajadores de las fases anteriores del proceso de producción pueden tener la tentación de tener menos interés en la calidad del producto. Deming insistió en que la calidad era responsabilidad de todos los miembros de la organización y en que la dirección tenía que organizar el proceso para garantizar que los niveles de calidad son siempre altos. El viejo método de la *detección* de los errores debía ser sustituido por el criterio de la *prevención* de los errores (véanse las referencias bibliográficas 2 y 11).

Posteriormente, se prestó mucha atención a la mejora de la calidad en la industria manufacturera japonesa y se desarrollaron y aplicaron algunos refinamientos y modificaciones de los métodos originales. Por ejemplo, Genichi Taguchi, ingeniero japonés, describió el coste de la variación en términos monetarios en lo que a menudo se conoce con el nombre de función de pérdida de Taguchi (véanse las referencias bibliográficas 4, 8 y 10). Su visión de la calidad, basada en implicaciones económicas del incumplimiento de las especificaciones fijadas como objetivo, se refería al diseño del producto. La premisa de Taguchi es «diseñar el producto para lograr una elevada calidad a pesar de la variación que se produzca en la línea de producción» (véase la referencia bibliográfica 5).

A finales de la década de 1970, la industria estadounidense estaba enfrentándose, como nunca hasta entonces, a una feroz competencia extranjera dentro de sus mercados. Las importaciones de bienes manufacturados crecieron espectacularmente, mientras que muchas industrias estadounidenses no lograron un éxito comparable en los mercados exteriores. Las consecuencias fueron profundas. Desde el punto de vista macroeconómico, Estados

Unidos tuvo déficit comerciales durante algún tiempo. Desde el punto de vista microeconómico, industrias enteras entraron en declive, mientras que otras se vieron obligadas a realizar rápidos y, en ocasiones, dolorosos ajustes para hacer frente a la competencia. En estas circunstancias, no es sorprendente que los estadounidenses se fijaran en su competidor más próspero: Japón. Existen, por supuesto, muchas diferencias entre las organizaciones sociales y económicas de Japón y Estados Unidos. El análisis detenido de estas diferencias para intentar explicar el éxito relativo de la industria japonesa nos llevaría muy lejos. Para nuestros fines basta señalar que muchos productos japoneses que pueden adquirirse en el mercado de Estados Unidos llegaron a adquirir una envidiable reputación por su calidad. El reconocimiento en muchas industrias estadounidenses de la necesidad de hacer frente a este reto es lo que explica el rápido crecimiento de la aplicación de métodos estadísticos en Estados Unidos desde la década de 1980.

Algunos de los beneficios de estos métodos deberían ser obvios:

1. *Aumento de la productividad*. Si se detectan en seguida y se corrigen las piezas que no cumplen las normas, puede evitarse mucha pérdida de tiempo y de materiales. La aplicación de un control estadístico satisfactorio de los procesos puede permitir la producción de un volumen mayor con una calidad mejor sin aumentar el coste o el esfuerzo de los trabajadores.
2. *Aumento de las ventas*. La reputación merecida por la calidad del producto es un activo enorme en el competitivo mercado. Esa reputación generalmente es difícil de conseguir, pero en muchas industrias su ausencia puede resultar fatal.
3. *Aumento de los beneficios*. El efecto neto de la reducción de los costes unitarios de producción y del aumento de las ventas se deja sentir, desde luego, en los balances finales de las empresas. Los métodos de control de la calidad están hoy muy extendidos porque son rentables.

Variación

La calidad comienza con un análisis completo del sistema y del proceso utilizados para producir bienes o servicios. Un sistema es un número de componentes que están relacionados entre sí lógica y a veces físicamente con algún fin y un proceso es un conjunto de actividades que operan en un sistema que transforma factores en productos. El objetivo es identificar todos los factores que contribuyen a la producción del producto final y contribuyen así a la calidad del producto. Los problemas que hacen que los productos tengan defectos deben identificarse y corregirse. «En todos los procesos intervienen *factores* a los que el proceso añade *valor* para producir el producto [...] Los indicadores de los resultados miden la producción del proceso y esa información se reintroduce en el proceso para ajustarlo y para eliminar los defectos que impiden lograr lo que quiere el cliente» (véase la referencia bibliográfica 6). Este modelo general de los procesos se muestra en la Figura 18.1 (véase la referencia bibliográfica 6).

En el proceso por el que los factores se transforman en productos, la dirección utiliza la teoría estadística (véase la referencia bibliográfica 9) para controlar y mejorar el proceso. Obsérvese en la Figura 18.1 que la dirección primero debe saber si el factor satisface las normas de calidad o las supera, es decir, el viejo problema de «si lo que entra es basura, lo que sale es basura». Por ejemplo, los fabricantes de automóviles exigen a los proveedores de factores, como juntas de goma para las puertas, que cumplan unas normas especificadas. El programa espacial y los fabricantes de aviones exigen a los proveedores de

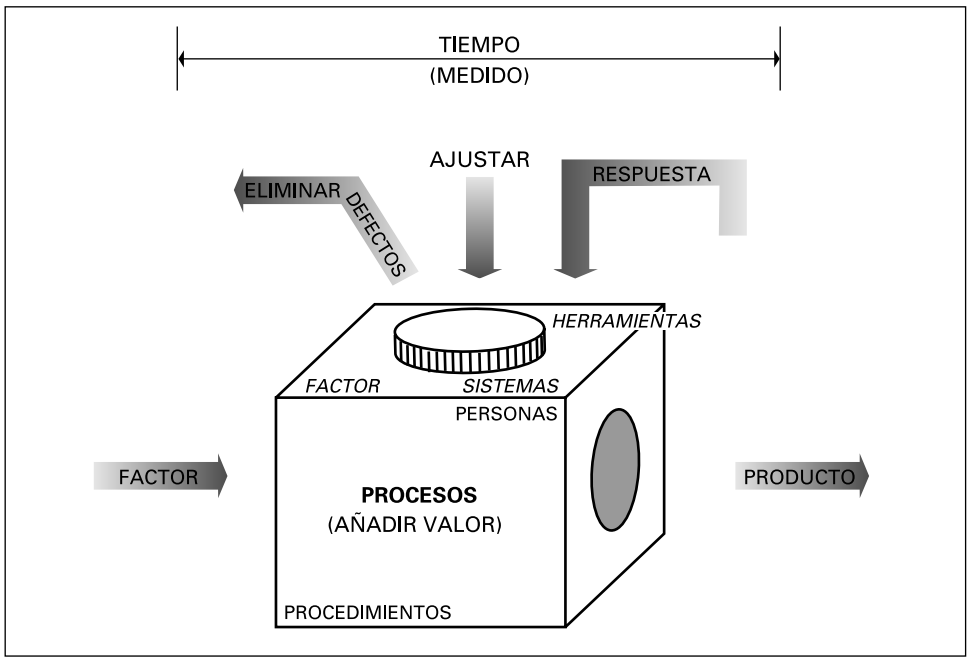


Figura 18.1. Modelo general de los procesos.

piezas, como fuelles metálicos soldados, que certifiquen las normas de calidad de sus productos. En todas las fases del proceso, se utilizan métodos estadísticos para controlar y corregir los defectos. La constante retroalimentación de la información contribuye a la mejora continua del proceso y del producto o del servicio.

Uno de los principios fundamentales del pensamiento estadístico es que *existe variación en todos los procesos* (véase la referencia bibliográfica 9). Es importante comprender la variación para predecir el funcionamiento futuro del proceso. Para comenzar a estudiar los instrumentos estadísticos que se utilizan para mejorar los procesos, primero examinamos dos causas de su variación: las causas comunes y las causas asignables.

Causas comunes y asignables de la variación

Las **causas comunes de la variación** (también llamadas causas aleatorias o incontrolables) son las causas que ocurren aleatoriamente y son inherentes a todos los procesos. El responsable de estas causas es la dirección, no los trabajadores. Las **causas asignables de la variación** (también llamadas causas especiales) son el resultado de fuerzas externas, es decir, de fuerzas ajenas al sistema. Estas causas pueden y deben detectarse y hay que tomar medidas correctoras para eliminarlas del proceso. De lo contrario, aumentará la variación y empeorará la calidad.

En el modelo general de los procesos de la Figura 18.1, se eliminan los defectos que se deben a causas asignables y se realizan los ajustes necesarios en el proceso. Ejemplos de causas comunes son las condiciones de trabajo desagradables (demasiado calor o demasiado frío) y los errores humanos aleatorios. Ejemplos de causas asignables son un lote defectuoso de materias primas, errores de los operadores y errores de ajuste de las máquinas. Las causas asignables deben eliminarse; las causas aleatorias siempre son inherentes a un proceso. Un proceso sólo es estable cuando se eliminan todas las causas asignables.

Proceso estable

Un proceso es **estable** (está controlado) si se eliminan todas las causas asignables; por lo tanto, la variación sólo se debe a causas comunes.

Aunque se ha analizado la importancia de la calidad del producto, aún no está claro dónde entran en escena los métodos estadísticos. Entran en escena porque el método representativo de control de la calidad, una vez que se aplica, debe implicar necesariamente un *muestreo* y un *análisis estadístico*. El objetivo es controlar un proceso de producción operativo. Casi inevitablemente, será inviable medir las características de todos los artículos producidos. Se extraen de vez en cuando muestras relativamente pequeñas de artículos y se realizan mediciones, con el fin de poder representar gráficamente el progreso a lo largo del tiempo y observar e investigar los cambios que hayan podido ocurrir. Lo importante es que la inferencia sobre la conducta del proceso se basa en datos estadísticos. Además, dado que las mediciones de los productos se realizan en la planta —y, en teoría, las valoraciones deben realizarse bastante deprisa—, es deseable que se utilicen métodos relativamente sencillos, como gráficos de control.

Un gráfico de control es un gráfico de la evolución temporal de una característica de un proceso, como la *tendencia central* o la *variación*. Existen varios tipos de gráficos de control y su aplicabilidad depende del tipo de datos de los que se disponga y de las variables que se quiera controlar. En este capítulo, introducimos cuatro gráficos de control que son los que se utilizan más a menudo. Son el gráfico $\bar{X}$ (para las medias), el gráfico s (para la variación), el gráfico p (para la proporción de artículos que no se ajustan a las normas) y el gráfico c (para el número de ocurrencias de un suceso, como las imperfecciones). En el apéndice de este capítulo, se analiza el gráfico R (para los intervalos). Todos los gráficos de control tienen una línea central (LC), un límite de control inferior (LCI) y un límite de control superior (LCS). Las mediciones tomadas a intervalos periódicos se representan en gráficos de control y se examinan en busca de pautas que sugieran la existencia de un posible problema provocado por causas asignables.

La importancia de las ideas estadísticas en el control de la calidad reside en la comprensión de la *variabilidad* y la *aleatoriedad*. El proceso de producción que fabrica artículos *idénticos* no se ha inventado y nunca se inventará. Es inevitable, en la práctica, que haya alguna variabilidad en las características de los artículos. Por lo tanto, para buscar cambios en las características de la producción a lo largo del tiempo, es importante no dejarse engañar por la mera variabilidad aleatoria.

EJERCICIOS

Ejercicios aplicados

18.1. Seleccione 1 de los 50 estados de Estados Unidos que conceden un premio a la calidad. Escriba un breve ensayo indicando el nombre del premio, los criterios utilizados para concederlo, la fecha en que se creó, los tipos de organizaciones que pueden recibirlo y una breve sinopsis de un galardonado recientemente con ese premio. ¿Por qué recibió esa organización ese premio? Incluya bibliografía y direcciones de páginas web.

18.2. Seleccione un premio a la calidad europeo o asiático. Escriba un breve ensayo indicando el nombre del premio, los criterios utilizados para concederlo, la fecha en que se creó, los tipos de organizaciones que pueden recibirlo y una breve sinopsis de un galardonado recientemente con ese premio. ¿Cuáles fueron las iniciativas de calidad de este galardonado que se reconocieron con este premio? Incluya bibliografía y direcciones de páginas web.

18.3. Analice los 14 puntos de Deming para las organizaciones de calidad (véanse las referencias bibliográficas 2 y 11).

18.4. Entre al menos en tres páginas web de la lista mencionada al final de este capítulo. Analice el tipo de información que contienen estas páginas.

18.5. Analice al menos cinco páginas web pertinentes para la calidad, aparte de las de la lista que se encuentra al final de este capítulo.

18.2. Gráficos de control de medias y desviaciones típicas

Consideremos ahora un proceso de producción que genera un producto cuya característica de interés puede medirse en un continuo. Se desea establecer un sistema de control de la calidad para ese proceso. Puede lograrse tomando, a lo largo del tiempo, una secuencia de pequeñas muestras del producto. A menudo se toman muestras de 4 o 5 observaciones y, para establecer un registro razonable del funcionamiento, es deseable tener 20 muestras o más. La frecuencia con que se toman las muestras depende de las características del proceso de producción. A la dirección le interesará tanto el funcionamiento medio de proceso como la variabilidad del funcionamiento. Si hay demasiada variabilidad, están produciéndose muchos artículos que no se ajustan a las normas, aunque el funcionamiento medio sea satisfactorio.

En este apartado, se utilizan las medias y las desviaciones típicas muestrales para controlar el funcionamiento del proceso. Estas cantidades se representan en gráficos de control. Se fijan límites de control para ayudar a comprender las fluctuaciones a lo largo del tiempo de la media muestral y de la desviación típica muestral. Sin embargo, antes de seguir debe señalarse que, aunque es bastante frecuente que se utilice la media, en algunas aplicaciones se utiliza el rango en lugar de la desviación típica para evaluar la variabilidad. El atractivo de esta opción probablemente se halla en que el rango —es decir, la diferencia entre la observación muestral más grande y la más pequeña— se calcula más fácilmente en la planta, donde se realizan ejercicios de control en el trabajo. Sin embargo, es posible que ya no sea así, dada la existencia de calculadoras electrónicas, que calculan automáticamente medias muestrales y desviaciones típicas muestrales, a partir de las observaciones muestrales. La construcción de gráficos de control cuando se utiliza el rango en lugar de la desviación típica se explica detalladamente en el apéndice del capítulo. Los principios de la construcción de gráficos de control y su interpretación son esencialmente los mismos cualquiera que sea la medida de la variabilidad que se emplee, aunque los detalles son algo distintos.

Tres medidas que se utilizan para realizar gráficos de control de medias y desviaciones típicas son la media global, la desviación típica muestral media y la desviación típica del proceso.

Media global, desviación típica muestral media y desviación típica del proceso

Se toma una secuencia de K muestras, cada una de n observaciones, a lo largo del tiempo sobre una característica mensurable del producto de un proceso de producción. Las medias muestrales, representadas por $\bar{x}_i$ para $i = 1, 2, ..., K$, pueden representarse gráficamente en un gráfico $\bar{X}$. La **media de estas medias muestrales** es la **media global** de todas las observaciones muestrales:

$$\bar{\bar{x}} = \sum_{i=1}^{K} \bar{x}_i / K \qquad \textbf{(18.1)}$$

Las desviaciones típicas muestrales, representadas por s_i para $i = 1, 2, ..., K$, pueden representarse gráficamente en un gráfico *s*. La **desviación típica muestral media** es

$$\bar{s} = \sum_{i=1}^{K} s_i/K \qquad \textbf{(18.2)}$$

La **desviación típica del proceso**, σ, es la desviación típica de la población de la que proceden las muestras y debe estimarse a partir de datos muestrales.

Una estimación de la desviación típica del proceso

Para fijar límites de control tanto en el gráfico $\bar{X}$ como en el gráfico *s*, es necesario estimar la desviación típica del proceso, σ. Una de las posibilidades es basar esta estimación en la desviación típica muestral global de todas las observaciones. Sin embargo, en los estudios aplicados del control de la calidad es más habitual basar una estimación de σ en $\bar{s}$, la desviación típica muestral media. Cualquiera que sea la estimación que se utilice, recuérdese que la desviación típica muestral es un estimador sesgado de la desviación típica poblacional. Es deseable intentar corregir este sesgo. De hecho, cuando se sabe que la distribución poblacional es normal, es posible hallar una expresión del valor esperado de la desviación típica muestral. Si la desviación típica muestral s_i se basa en n observaciones, puede demostrarse que

$$E(s_i) = c_4\sigma$$

donde c_4 es un número que puede calcularse como una función del tamaño de la muestra n. Se deduce inmediatamente que

$$E(\bar{s}) = c_4\sigma$$

y, por lo tanto, que una estimación insesgada de la desviación típica del proceso es $\hat{\sigma} = \bar{s}/c_4$. Naturalmente, la distribución poblacional puede no ser exactamente normal. No obstante, se piensa que esta corrección merece la pena y que normalmente reduce el sesgo inherente a la desviación típica muestral como estimador del parámetro poblacional correspondiente.

Estimación de la desviación típica del proceso basada en *s*

Una **estimación de la desviación típica del proceso**, $\hat{\sigma}$, es

$$\hat{\sigma} = \bar{s}/c_4 \qquad \textbf{(18.3)}$$

donde $\bar{s}$ es la desviación típica muestral media; la constante del gráfico de control, c_4, que depende del tamaño de la muestra *n*, puede hallarse en la Tabla 18.1 o en la Tabla 13 del apéndice. Si la distribución de la población es normal, el estimador es insesgado.

La Tabla 18.1 muestra los valores de c_4 correspondientes a tamaños muestrales que van de 2 a 10. También muestra las constantes de otros gráficos de control que se analizarán en este capítulo. La Tabla 13 del apéndice contiene una tabla más completa de constantes. En los estudios prácticos de control de la calidad, se dispone de tablas de las constantes de los gráficos de control que se utilizan habitualmente.

En este capítulo nos referiremos al siguiente ejemplo, basado en el fichero de datos **Signal**.

Tabla 18.1. Constantes de los gráficos de control.

n	c_4	A_3	B_3	B_4
2	0,789	2,66	0	3,27
3	0,886	1,95	0	2,57
4	0,921	1,63	0	2,27
5	0,940	1,43	0	2,09
6	0,952	1,29	0,03	1,97
7	0,959	1,18	0,12	1,88
8	0,965	1,10	0,18	1,82
9	0,969	1,03	0,24	1,76
10	0,973	0,98	0,28	1,72

EJEMPLO 18.1. Señal emitida por un componente electrónico ($\bar{\bar{x}}$, $\bar{s}$, $\hat{\sigma}$)

Signal

La duración, en milisegundos, de una señal emitida por un componente electrónico de una secuencia de 20 muestras, cada una de las cuales tiene cinco observaciones, se muestra en la Tabla 18.2 y se encuentra en el fichero de datos **Signal**. Halle la media global, la desviación típica muestral media y una estimación de la desviación típica del proceso.

Tabla 18.2. Duración, en milisegundos, de la señal emitida por un componente electrónico.

Número muestral						Media muestral	Desviación típica muestral
1	297	296	297	303	298	298,2	2,77
2	301	301	300	304	297	300,6	2,51
3	297	306	296	302	304	301,0	4,36
4	296	302	299	298	309	300,8	5,07
5	305	304	293	309	293	300,8	7,36
6	298	294	303	306	305	301,2	5,07
7	297	304	299	298	306	300,8	3,96
8	292	292	307	295	300	297,2	6,38
9	295	297	307	304	306	301,8	5,45
10	296	297	309	297	305	300,8	5,85
11	299	301	290	298	297	297,0	4,18
12	303	307	296	298	294	299,6	5,32
13	301	292	313	302	307	303,0	7,78
14	299	298	300	301	295	298,6	2,30
15	299	299	306	303	298	301,0	3,39
16	301	303	297	298	304	300,6	3,05
17	300	296	301	300	304	300,2	2,86
18	295	293	300	299	289	295,2	4,49
19	298	298	306	297	295	298,8	4,21
20	296	303	300	304	299	300,4	3,21

Solución

La Tabla 18.2 contiene la media muestral y la desviación típica muestral correspondientes a cada periodo de observación. La media muestral global, que es simplemente la media de las 100 observaciones muestrales, es

$$\bar{\bar{x}} = \frac{(298{,}2 + 300{,}6 + \cdots + 300{,}4)}{20} = 299{,}9$$

La media de las desviaciones típicas muestrales es

$$\bar{s} = \frac{(2{,}77 + 2{,}51 + \cdots + 3{,}21)}{20} = 4{,}48$$

Basándonos en la Tabla 18.1, vemos que con $n = 5$ observaciones,

$$c_4 = 0{,}940$$

Por lo tanto, la desviación típica del proceso es

$$\hat{\sigma} = \frac{\bar{s}}{c_4} = \frac{4{,}48}{0{,}940} = 4{,}77$$

Gráficos de control de medias

La Tabla 18.2 muestra las duraciones medias de las señales de una secuencia de 20 muestras de cinco observaciones cada una a lo largo del tiempo. En los estudios de control de la calidad, para facilitar la interpretación, esa información se representa invariablemente en un gráfico temporal, por ejemplo, en un gráfico $\bar{X}$ o en un gráfico s. En primer lugar, estudiamos los métodos estadísticos para obtener gráficos de control y, a continuación, explicamos cómo se interpretan los gráficos para encontrar indicios de inestabilidad del proceso.

Para la gestión de la producción, es importante buscar indicaciones de empeoramiento de la calidad. Una de las indicaciones posibles de que hay un problema es una media muestral que se desvía considerablemente de su valor «habitual». Por ejemplo, en la Tabla 18.2 la media de la decimoctava muestra es 295,2, que es un valor algo más bajo que los anteriores. ¿Es éste el tipo de resultado que sería razonable esperar como consecuencia de la variabilidad muestral? En el control de la calidad, esta valoración se hace realizando comparaciones con los límites de control trazados en los gráficos de control.

Para fijar los límites de control de los gráficos $\bar{X}$, se supone que el proceso ha venido funcionando a un nivel constante durante todo el periodo de observación y que puede considerarse que todas las observaciones muestrales se han extraído de la misma distribución normal. La media de esa distribución se estima por medio de la media global, $\bar{\bar{x}}$, de todas las observaciones muestrales y la desviación típica se estima por medio de $\hat{\sigma}$ de la ecuación 18.3.

Consideremos ahora una única muestra de cinco observaciones y consideremos que se han extraído de una distribución normal de media $\bar{\bar{x}}$ y desviación típica $\hat{\sigma}$. La distribución muestral de esta media muestral es una distribución normal de media $\bar{\bar{x}}$ y error típico $\hat{\sigma}/\sqrt{n} = \hat{\sigma}/\sqrt{5}$. Este resultado sirve de base para fijar los límites de control.

Cuando existen indicios de que hay un problema, hay que hacer alguna investigación, que puede implicar la interrupción y el estudio exhaustivo del proceso de producción, lo cual puede ser bastante caro. Naturalmente, no es deseable que haya que hacerlo frecuentemente cuando el proceso funciona, en realidad, satisfactoriamente. Para que no aparezcan demasiadas «señales falsas» de este tipo, es habitual en los estudios de control de la calidad fijar unos límites de control equivalentes a tres errores típicos en cualquiera de los lados de la media de la distribución muestral (a veces se llaman límites 3σ). En ese caso, si

la distribución del estadístico muestral —en este caso, la media muestral— es normal, la probabilidad de obtener un valor situado fuera de los límites 3σ es

$$P(Z > 3) + P(Z < -3) = 2(0{,}0014) = 0{,}0028$$

donde Z es una variable aleatoria normal estándar. Por lo tanto, si se fijan límites de esta forma, partiendo de los supuestos postulados, la probabilidad de que aparezca una señal falsa en cualquier muestra es de menos de 3 por 1.000. Desde luego, estos supuestos generalmente no son absolutamente ciertos, por lo que este valor sólo es aproximado. No obstante, debe ser una guía razonable y el uso de límites 3σ es muy frecuente.

Volviendo ahora a la construcción de gráficos de control de medias muestrales, la distribución muestral está centrada en la media global, $\bar{\bar{x}}$, y esta *línea central* se traza en el gráfico. Por lo tanto, si se utilizan límites de tres errores típicos, los límites de control son

$$\bar{\bar{x}} \pm 3\hat{\sigma}/\sqrt{n} = \bar{\bar{x}} \pm 3\bar{s}/(c_4\sqrt{n}) = \bar{\bar{x}} \pm A_3\bar{s} \quad \text{donde } A_3 = \frac{3}{c_4\sqrt{n}}$$

Gráfico $\bar{X}$

El **gráfico $\bar{X}$** es un gráfico temporal de la secuencia de medias muestrales. La **línea central** es

$$LC_{\bar{X}} = \bar{\bar{x}} \tag{18.4}$$

Además, hay límites de control de tres errores típicos. El **límite de control inferior** es

$$LCI_{\bar{X}} = \bar{\bar{x}} - A_3\bar{s} \tag{18.5}$$

y el **límite de control superior** es

$$LCS_{\bar{X}} = \bar{\bar{x}} + A_3\bar{s} \tag{18.6}$$

Algunos valores de A_3 se encuentran en la Tabla 18.1 o en la Tabla 13 del apéndice.

Signal

EJEMPLO 18.2. Gráfico de control de las señales para las medias (gráfico $\bar{X}$)

Construya el gráfico $\bar{X}$ del ejemplo de las señales con el fichero de datos **Signal**.

Solución

Basándose en la Tabla 18.1 y en los cálculos anteriores, con un tamaño muestral de cinco,

$$\bar{\bar{x}} = 299{,}9 \qquad \bar{s} = 4{,}48 \qquad A_3 = 1{,}43$$

La línea central es

$$LC_{\bar{X}} = \bar{\bar{x}} = 299{,}9$$

El límite de control inferior es

$$LCI_{\bar{X}} = \bar{\bar{x}} - A_3\bar{s} = 299{,}9 - (1{,}43)(4{,}48) = 293{,}5)$$

y el límite de control superior es

$$LCS_{\bar{X}} = \bar{\bar{x}} + A_3\bar{s} = 299{,}9 + (1{,}43)(4{,}48) = 306{,}3$$

A continuación, se representa cada una de las medias muestrales, $\bar{x}_i$, en el gráfico $\bar{X}$ de la Figura 18.2.

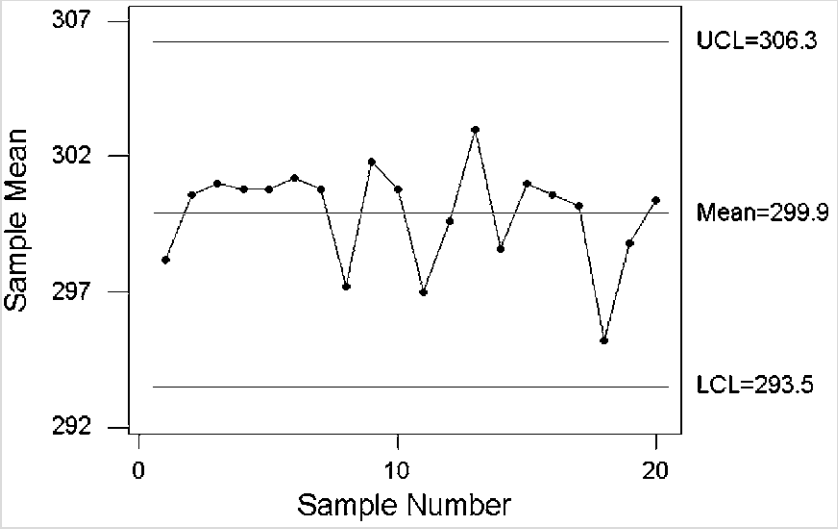


Figura 18.2. Gráfico $\bar{X}$ del ejemplo de las señales.

Gráficos de control de desviaciones típicas

Para evaluar el progreso de la variabilidad de un proceso a lo largo del tiempo, también pueden representarse las desviaciones típicas en un gráfico de control llamado gráfico *s*. La línea central de este gráfico es la desviación típica muestral media, $\bar{s}$, y es habitual fijar límites de tres errores típicos.

Gráfico *s*

El **gráfico *s*** es un gráfico temporal de la secuencia de desviaciones típicas muestrales. La **línea central** de un gráfico *s* es

$$LC_s = \bar{s} \tag{18.7}$$

En el caso de los límites de tres errores típicos, el **límite de control inferior** es

$$LCI_S = B_3\bar{s} \tag{18.8}$$

y el **límite de control superior** es

$$LCS_s = B_4\bar{s} \tag{18.9}$$

Los valores de las constantes B_3 y B_4 del gráfico de control se muestran en la Tabla 18.1.

Cuando el tamaño de la muestra es $n \leqslant 5$, restando tres errores típicos de $\bar{s}$ se obtiene un número negativo. Evidentemente, las desviaciones típicas no pueden ser negativas, por lo que se considera que el límite inferior es 0. En la práctica, raras veces preocupa que haya demasiado poca variabilidad, por lo que el límite inferior normalmente no tiene mucho interés.

EJEMPLO 18.3. Gráfico de control de las señales para las desviaciones típicas (gráfico *s*)

Utilice el fichero de datos **Signal** para trazar el gráfico *s* correspondiente al ejemplo de las señales.

Signal

Solución

Para construir el gráfico *s* correspondiente a los datos del fichero **Signal** (Tabla 18.2), se deduce que

$$\bar{s} = 4{,}48 \qquad B_3 = 0 \qquad B_4 = 2{,}09$$

Por lo tanto, las tres líneas de nuestro gráfico son

$$LC_s = 4{,}48 \qquad LCI_s = 0 \qquad LCS_s = (2{,}09)(4{,}48) = 9{,}36$$

A continuación, se representa cada una de las desviaciones típicas, s_i, en un gráfico de control, en el que $LC_s = 4{,}48$, $LCI_s = 0$ y $LCS_s = 9{,}36$. El gráfico *s* se parece al de la Figura 18.3, que se ha obtenido utilizando el programa Minitab.

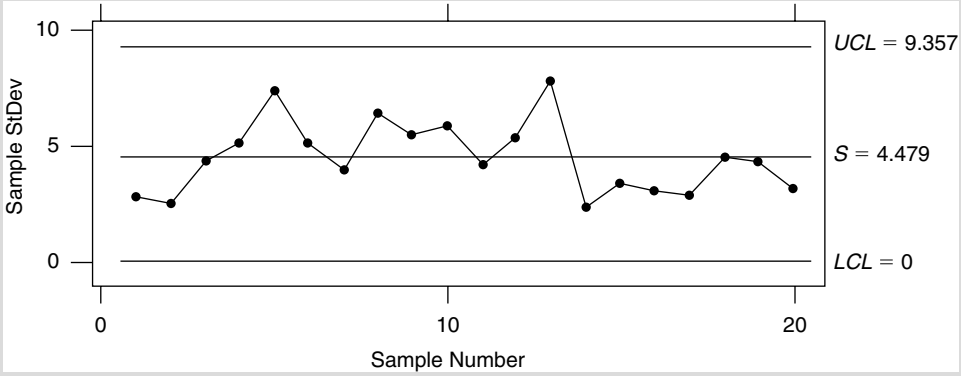


Figura 18.3. Gráfico *s* del ejemplo de las señales.

Interpretación de los gráficos de control

Una vez desarrollados gráficos de control iniciales para controlar el funcionamiento medio y su variabilidad, es necesario profundizar en su análisis y su interpretación. La experiencia y la valoración personal, junto con la comprensión de las pautas de los gráficos de control, permiten hacer mejoras. A continuación, analizamos brevemente algunas cuestiones que podrían plantearse. La principal es una valoración del funcionamiento del proceso en el periodo de observación.

Si un proceso es estable, los puntos de un gráfico de control fluctuarán aleatoriamente entre el límite de control superior y el inferior, por lo que no seguirán una pauta que no sea aleatoria. En esta fase, el analista busca esencialmente una pauta de puntos de datos

distribuidos más o menos aleatoriamente en torno a la línea central y generalmente bien alejados de los límites de control. Desde este punto de vista, las Figuras 18.2 y 18.3 parecen bastante razonables. En esas circunstancias, el proceso estudiado parece estar *bajo control*, lo cual significa que su funcionamiento es bastante estable. El control estadístico de calidad puede concebirse como un medio para averiguar si un proceso está bajo control, como una ayuda para mantenerlo bajo control y como un mecanismo para reducir la variabilidad de la calidad del producto.

Si un proceso no es estable, es posible que los datos no sean fiables o que el proceso que ha generado los datos tenga graves problemas operativos. Esos datos pueden darnos una indicación fiable de lo que cabe esperar cuando el proceso funciona normalmente. Se llama la atención a la dirección sobre las causas asignables que pueden contribuir a la inestabilidad del proceso por medio de diversas pautas de los gráficos de control. Por lo tanto, la interpretación de los gráficos de control comienza por comprender algunas pautas que indican una situación *fuera de control*.

Hay varias pautas de los puntos de datos de un gráfico de control que indican que un proceso puede estar fuera de control. Aquí sólo examinamos tres pruebas para analizar esta posibilidad. Para más pruebas y un estudio más extenso véanse las notas de este capítulo.

Pautas fuera de control

Algunas pautas de los puntos de datos de un gráfico de control indican que el proceso puede estar **fuera de control**. A continuación, mostramos tres:

1. Un valor fuera de los límites de control (un punto que esté alejado más de 3 sigmas de la línea central).
2. Una tendencia de los estadísticos muestrales (seis puntos consecutivos, todos crecientes o decrecientes).
3. Demasiados puntos en uno de los lados de la línea central (nueve puntos consecutivos en el mismo lado de la línea central).

1. **Un valor fuera de los límites de control.** Consideremos la Figura 18.4A. La mayoría de los estadísticos muestrales (en este ejemplo, las medias muestrales) están dentro de los límites de control. Sin embargo, en el caso de la muestra 7 el estadístico está fuera de estos límites; es decir, la media muestral es mayor que la media más 3σ, que es el límite de control superior, 308,5. Con límites de tres errores típicos, esto sería muy poco habitual en un proceso que está bajo control. Es un fenómeno que exige una investigación. El analista tiene que hacer algún trabajo de detective y buscar la causa de este valor extremo. Probablemente se observará que no es una mera variabilidad aleatoria sino que se debe a alguna causa asignable peculiar. Conviene comprobar primero si esa causa asignable es simplemente que el valor de una observación no se ha registrado correctamente.
2. **Una tendencia de los estadísticos muestrales (seis puntos consecutivos, todos crecientes o decrecientes).** La Figura 18.4B es claramente una figura en la que no hay aleatoriedad. Los estadísticos muestrales no parecen estar dispersos aleatoriamente en torno a la línea central. Tienden, por el contrario, a aumentar con el tiempo y los siete últimos son crecientes. Esta variabilidad creciente es un motivo de preocupación, aunque ningún valor muestral esté fuera de los límites de control. Quizá la causa sea que la maquinaria está deteriorándose.
3. **Demasiados puntos en uno de los lados de la línea central (nueve puntos consecutivos en el mismo lado de la línea central).** En la Figura 18.4C, nueve puntos

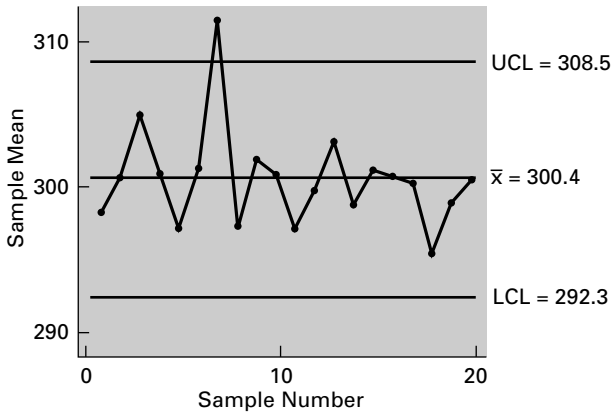


Figura 18.4A. Un valor fuera de los límites de control.

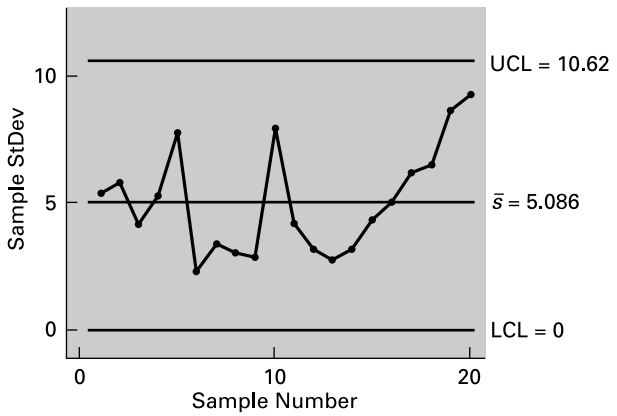


Figura 18.4B. Una tendencia de las desviaciones típicas muestrales.

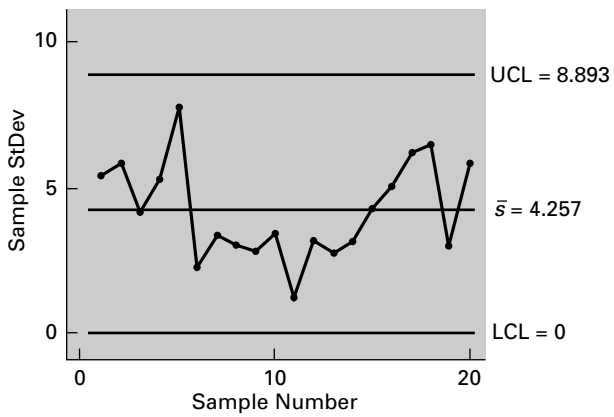


Figura 18.4C. Demasiados puntos en el mismo lado de la línea central.

consecutivos (que representan desviaciones típicas muestrales) se encuentran por debajo de la línea central. La investigación de las causas asignables puede revelar la existencia de problemas.

Sólo es razonable seguir adelante cuando existe alguna seguridad de que un proceso de producción está bajo control. Si observamos de nuevo las Figuras 18.2 y 18.3, no vemos ninguna indicación de que existan pautas fuera de control. Parece que no existe ningún motivo importante de preocupación. Ninguna de las medias muestrales está fuera de los límites de control y, de hecho, la inmensa mayoría de las medias muestrales están dentro de

esos límites. Antes nos hemos preguntado si la media de la muestra 18, 295,2, era un motivo para preocuparse. Aparentemente, no existe ningún motivo de alarma. Asimismo, no parece que esté justificada la preocupación por la variabilidad del proceso, indicada en el gráfico *s* de la Figura 18.3. Las desviaciones típicas muestrales observadas generalmente están muy por debajo del límite de control superior. Parece que aumenta la variabilidad en la parte central del periodo de observación, por lo que quizá merezca la pena buscar una explicación para comprender mejor el proceso de producción. Ninguno de los dos gráficos de control correspondientes al fichero de datos **Signal** sugiere que haya causas asignables en el sistema.

En el siguiente apartado explicamos cómo se averigua si un proceso estable cumple las especificaciones del diseño.

EJERCICIOS

Ejercicios aplicados

18.6. Se ha observado el proceso de producción de un componente y se ha medido la fuerza de la emisión eléctrica de los componentes. Se dispone de resultados de una secuencia de 30 muestras, cada una de las cuales tiene siete observaciones. La media global de las observaciones muestrales es 192,6 y la desviación típica muestral media es 5,42.

a) Utilice un estimador insesgado para estimar la desviación típica del proceso.
b) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
c) Halle la línea central y los límites de control inferior y superior de un gráfico *s*.

18.7. Se toman medidas de la resistencia, en ohmios, de un componente eléctrico. Se obtiene una secuencia de 25 muestras, cada una de las cuales tiene seis observaciones. La media global de las observaciones muestrales es 93,2 y la desviación típica muestral media es 3,67.

a) Utilice un estimador insesgado para estimar la desviación típica del proceso.
b) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
c) Halle la línea central y los límites de control inferior y superior de un gráfico *s*.

18.8. Se pesan muestras de fruta enlatada. Se toma una secuencia de 16 muestras, cada una de las cuales tiene ocho observaciones. La media global de las observaciones muestrales es de 19,86 onzas y la desviación típica muestral media es de 1,23 onzas.

a) Utilice un estimador insesgado para estimar la desviación típica del proceso.
b) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
c) Halle la línea central y los límites de control inferior y superior de un gráfico *s*.

18.9. Recuerde el ejercicio 2.39. Ann Thorne, la directora de operaciones de una fábrica de cremas bronceadoras, quiere asegurarse de que el proceso que se emplea para llenar los botes de un nuevo producto, SunProtector, funciona correctamente. Actualmente, la empresa está comprobando los volúmenes de los botes de 8 onzas (237 ml) de SunProtector. Se hacen mediciones del volumen de los botes de 8 onzas. Se toma una secuencia de 20 muestras de cinco observaciones cada una. La media global de las observaciones muestrales es de 230,5 ml y la desviación típica muestral media es de 1,75 ml. Los volúmenes (en ml) se encuentran en el fichero de datos **Sun**.

a) Utilice un estimador insesgado para estimar la desviación típica del proceso.
b) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
c) Halle la línea central y los límites de control inferior y superior de un gráfico *s*.

18.10. La tabla adjunta muestra las medias y las desviaciones típicas muestrales de una secuencia de 30 muestras de ocho observaciones cada una sobre una característica de la calidad de un producto. El fichero de datos es **Exercise 18-10**.

a) Halle la media global de las observaciones muestrales.
b) Halle la desviación típica muestral media.
c) Utilice un estimador insesgado para estimar la desviación típica del proceso.

d) Halle la línea central y los límites de control inferior y superior para un gráfico $\bar{X}$.
e) Trace el gráfico $\bar{X}$ y analice sus características.
f) Halle la línea central y los límites de control inferior y superior de un gráfico s.
g) Trace el gráfico s y analice sus características.

Muestra	$\bar{x}$	s	Muestra	$\bar{x}$	s
1	148,2	2,26	8	149,2	4,71
2	146,4	4,37	9	153,9	5,82
3	149,9	7,93	10	150,6	4,98
4	152,8	6,79	11	156,0	4,79
5	148,7	5,31	12	150,4	3,92
6	150,6	3,17	13	148,7	8,31
7	151,5	6,15	14	151,1	7,29
15	147,2	3,80	23	151,3	6,20
16	152,9	4,87	24	150,8	7,39
17	150,7	3,88	25	147,2	6,97
18	147,2	8,93	26	141,9	9,68
19	149,4	6,85	27	152,7	4,28
20	154,3	7,29	28	148,6	6,51
21	148,7	6,28	29	150,2	7,29
22	149,7	8,92	30	148,6	4,73

18.11. La tabla adjunta muestra las medias y las desviaciones típicas muestrales de una secuencia de 20 muestras de seis observaciones cada una sobre el peso de las latas de verduras, en onzas. El fichero de datos es **Exercise 18-11**.

Muestra	$\bar{x}$	s	Muestra	$\bar{x}$	s
1	20,2	1,9	11	18,8	2,9
2	18,9	2,7	12	19,3	1,1
3	19,6	1,7	13	19,8	1,3
4	20,8	2,3	14	20,2	1,2
5	19,4	1,2	15	20,7	1,9
6	19,8	2,1	16	19,3	2,2
7	20,9	1,6	17	19,9	3,1
8	21,0	2,3	18	18,8	2,9
9	20,6	1,4	19	19,6	2,2
10	19,1	2,7	20	20,1	1,1

a) Halle la media global de las observaciones muestrales.
b) Halle la desviación típica muestral media.
c) Utilice un estimador insesgado para estimar la desviación típica del proceso.
d) Halle la línea central y los límites de control inferior y superior para un gráfico $\bar{X}$.
e) Trace el gráfico $\bar{X}$ y analice sus características.
f) Halle la línea central y los límites de control inferior y superior de un gráfico s.
g) Trace el gráfico s y analice sus características.

18.3. Capacidad de un proceso

En el apartado 18.2 nos hemos ocupado del uso de gráficos de control, ayudados por límites de control, para averiguar si un proceso está bajo control, es decir, si su funcionamiento es estable. Sin embargo, esta información es insuficiente para saber si el proceso está cumpliendo como es debido las normas para las que se diseñó. Al fin y al cabo, un funcionamiento sistemático podría ser sistemáticamente mediocre o incluso sistemáticamente malo. Antes de seguir con un programa de control de calidad o de mejora de la calidad, es importante averiguar si el proceso de producción funciona de acuerdo con las especificaciones exigidas. Si un proceso está actualmente bajo control, ¿es *capaz* de cumplir estas especificaciones? Esta valoración se hace basándose en los datos generados por un proceso que parece que está bajo control. Por lo tanto, si los datos muestrales contienen observaciones extremas debidas a causas asignables, estos problemas deben corregirse antes de evaluar la *capacidad del proceso*. Más en serio, cuando parece que las cosas han ido mal en el periodo de observación como, por ejemplo, en los casos ilustrados en las Figuras 18.4B y 18.4C, puede que sea necesario que los ingenieros tomen medidas. Sólo cuando se ha establecido un método de control, es posible evaluar la capacidad del proceso.

En este apartado, analizamos un problema frecuente que puede abordarse analizando las medias muestrales y las desviaciones típicas muestrales. Normalmente, la dirección fija un intervalo de valores de alguna característica del proceso productivo, acotado por unos **límites de especificación** inferior y superior. En el caso de la duración de la señal emitida

por un componente electrónico, la dirección puede fijar un intervalo de valores tolerables de 280 a 320 milisegundos para garantizar la calidad del producto. Un proceso capaz de cumplir estas especificaciones es un proceso que probablemente producirá resultados dentro de este intervalo.

En el caso de un proceso que está fuera de control, es lógico basar la evaluación de la capacidad en todas las observaciones muestrales y, en concreto, en estimaciones de la media y la desviación típica del proceso basadas en estas observaciones. En el caso de los datos sobre la señal, las estimaciones son

$$\bar{\bar{x}} = 299{,}9 \qquad \hat{\sigma} = 4{,}77$$

En ese caso, si se supone que la distribución del proceso es normal, alrededor del 99,72 por ciento de toda la producción deberá estar en un margen más/menos tres desviaciones típicas con respecto a la media. Es frecuente, pues, en los estudios de control de la calidad calcular el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$. En nuestro ejemplo,

$$(\bar{\bar{x}} - 3\hat{\sigma},\ \bar{\bar{x}} + 3\hat{\sigma}) = (285{,}6,\ 314{,}2)$$

Éstos son los límites dentro de los cuales el proceso funcionará normalmente. La amplitud de este intervalo

$$6\hat{\sigma} = (6)(4{,}77) = 28{,}6$$

a veces se llama **tolerancia natural** del proceso. Es una medida de la variabilidad de las especificaciones del producto que cabe esperar.

Una vez utilizados los datos muestrales para saber qué puede hacer realmente un proceso de producción, sólo es necesario comparar este resultado con las especificaciones de lo que debe hacer el proceso establecidas por la dirección. Lo que se necesita es que el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ esté, preferiblemente de una manera holgada, entre los límites de especificación inferior y superior. Los datos sobre las señales parecen bastante satisfactorios desde este punto de vista. El intervalo de 285,6 a 314,2 está holgadamente entre 280 y 320 milisegundos. Parece que el proceso es capaz de satisfacer estas especificaciones. Obsérvese que la media muestral global de 299,9 está muy cerca del centro, 300 milisegundos, del intervalo de tolerancia. En esas circunstancias, se dice que el intervalo de funcionamiento está *centrado* en el rango de tolerancia. Normalmente, está centrado y a menudo es deseable que lo esté. Sin embargo, no es necesario para que el proceso sea capaz de satisfacer las normas.

Hay medidas más formales de la capacidad de un proceso y son el índice de capacidad y el índice C_{pk}. En las empresas que se dedican a mejorar procesos, los empleados conocen estas medidas de la capacidad de un proceso y comprenden su importancia.

Medidas de la capacidad de un proceso

Supongamos que la dirección fija unos límites de tolerancia inferior (I) y superior (S) para el funcionamiento de un proceso. La capacidad del proceso se valora por el grado en que $\bar{\bar{x}} \pm 3\hat{\sigma}$ se encuentra dentro de estos límites.

1. **Índice de capacidad (C_p).** Esta medida es adecuada cuando los datos muestrales están **centrados** entre los límites de tolerancia, es decir, $\bar{\bar{x}} \approx (I + S)/2$. El índice es

$$C_p = \frac{S - I}{6\hat{\sigma}} \qquad \textbf{(18.10)}$$

Normalmente se considera que un valor satisfactorio de este índice es un valor de 1,33 como mínimo [eso implica que la **tolerancia natural** del proceso no debe ser más de un 75 por ciento de ($S - I$), la amplitud del intervalo de valores aceptables].

2. **Índice C_{pk}.** Cuando los datos muestrales no están centrados entre los límites de tolerancia, es necesario tener en cuenta el hecho de que el proceso está funcionando más cerca de uno de los límites de tolerancia que del otro. La medida resultante, llamada índice C_{pk}, es

$$C_{pk} = \text{Min}\left[\frac{S - \bar{\bar{x}}}{3\hat{\sigma}}, \frac{\bar{\bar{x}} - I}{3\hat{\sigma}}\right] \qquad \textbf{(18.11)}$$

Una vez más, se considera que es satisfactorio si su valor es de 1,33 como mínimo.

EJEMPLO 18.4. Capacidad del proceso de producción de señales (índices de capacidad)

Signal

Considere de nuevo el caso de la duración de una señal emitida por un componente electrónico y suponga que la dirección fija un intervalo de valores tolerables que va de 280 a 320 milisegundos. Averigüe si el proceso de producción del ejemplo de las señales, que se encuentra en el fichero de datos **Signal** (Tabla 18.2), es capaz de satisfacer las especificaciones. Utilice las medidas de la capacidad de las ecuaciones 18.10 y 18.11.

Solución

En el caso de los datos sobre las señales,

$$\bar{\bar{x}} = 299{,}9 \qquad \hat{\sigma} = 4{,}77 \qquad I = 280 \qquad S = 320$$

Por lo tanto, el índice de capacidad es

$$C_p = \frac{S - I}{6\hat{\sigma}} = \frac{320 - 280}{6(4{,}77)} = 1{,}398$$

El índice C_{pk} es

$$C_{pk} = \text{Min}\left[\frac{S - \bar{\bar{x}}}{3\hat{\sigma}}, \frac{\bar{\bar{x}} - I}{3\hat{\sigma}}\right] = \text{Min}\,(1{,}405,\ 1{,}391) = 1{,}391$$

En este caso concreto, como los datos muestrales están, a todos los efectos, centrados, los dos índices son casi idénticos. Ambos son holgadamente superiores a 1,33, lo que indica que el proceso de producción es capaz de satisfacer las especificaciones.

Las Figuras 18.5 y 18.6 son salidas Minitab que dan los valores tanto de C_p como de C_{pk} del ejemplo 18.4. En la Figura 18.6 vemos, además de los valores de los índices de capacidad, tanto el gráfico $\bar{X}$ como el gráfico s, un gráfico de los 20 últimos subgrupos, el histograma de la capacidad y el gráfico de probabilidad normal.

Una vez evaluada la capacidad del proceso, éste es capaz de satisfacer las especificaciones o las normas o no lo es. Supongamos que observamos que el proceso no es capaz de satisfacer las especificaciones. Este tipo de problema debe comunicarse a la dirección

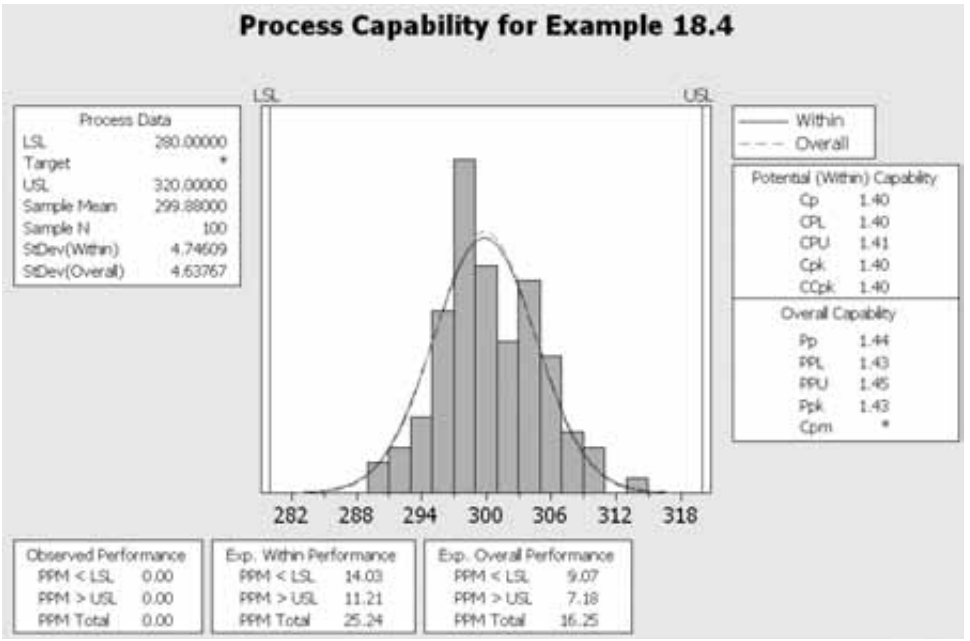


Figura 18.5. «Capability Analysis» (distribución normal) del ejemplo de las señales.

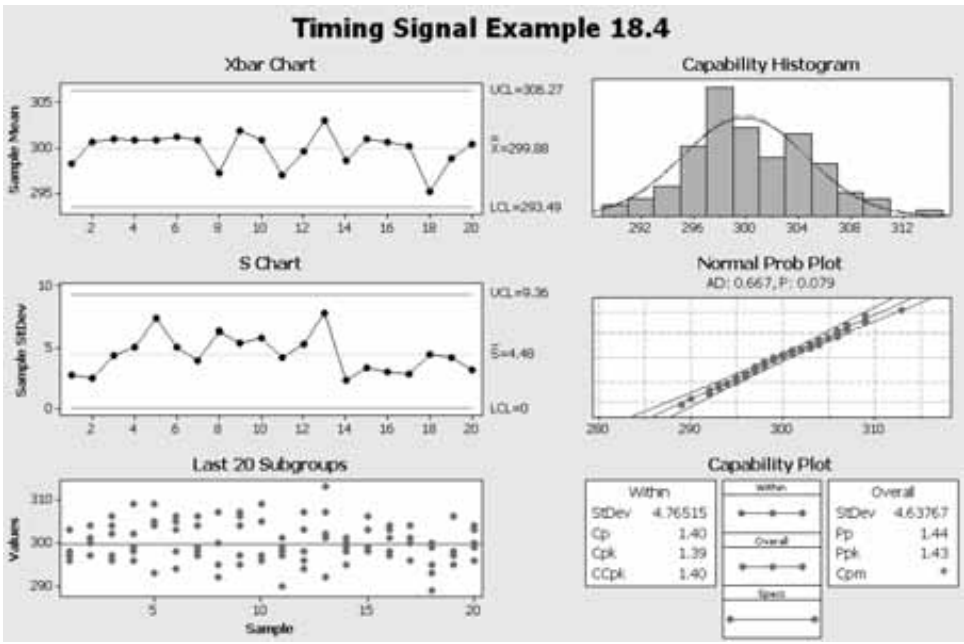


Figura 18.6. «Capability Sixpack» (distribución normal) del ejemplo de las señales.

para que lo analice a fondo y lo corrija. No es un problema para los trabajadores de la planta, que pueden ser capaces de señalar el problema, pero que es improbable que puedan resolverlo. Puede que el equipo de capital no sea adecuado para hacer ese trabajo, posiblemente porque se ha deteriorado. Puede que las normas de funcionamiento que se han fijado sean excesiva e innecesariamente optimistas. Cualquiera que sea la razón, no es muy útil continuar manteniendo el proceso y analizándolo en su estado actual.

El resultado mejor es que se observe que el proceso de producción es capaz de satisfacer las normas de funcionamiento. En ese caso, puede mantenerse el proceso de control de calidad. Se debe controlar periódicamente y se deben trazar gráficos de calidad. De vez en cuando, es deseable calcular de nuevo los límites de control de estos gráficos. También debe comprobarse periódicamente la capacidad del proceso. El control de calidad no es meramente una actividad pasiva. Tampoco es sólo un mecanismo para detectar los problemas, aunque es valioso, desde luego, para ese fin. El objetivo de un ejercicio de control de la calidad es la mejora de la calidad, que puede concebirse como una reducción de la tolerancia natural del proceso. Estas mejoras pueden conseguirse concienciándose más de la importancia de la calidad y de sus fuentes y comprendiéndolas mejor cuando los trabajadores participan en la recogida y la interpretación de datos para los estudios de control de calidad.

EJERCICIOS

Ejercicios aplicados

18.12. Vuelva al ejercicio 18.6. La dirección ha especificado que la fuerza de la emisión eléctrica de los componentes producidos por este proceso debe estar entre 170 y 215.

a) Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.

b) Halle el índice de capacidad C_p y analice el resultado.

c) Halle el índice C_{pk} y analice el resultado.

18.13. Vuelva al ejercicio 18.7. La dirección ha especificado que la resistencia de los componentes producidos por este proceso debe estar entre 85 y 100 ohmios.

a) Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.

b) Halle el índice de capacidad C_p y analice el resultado.

c) Halle el índice C_{pk} y analice el resultado.

18.14. Vuelva al ejercicio 18.8. La dirección ha especificado que el peso de la fruta enlatada debe estar entre 18 y 22 onzas.

a) Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.

b) Halle el índice de capacidad C_p y analice el resultado.

c) Halle el índice C_{pk} y analice el resultado.

18.15. Vuelva al ejercicio 18.10. La dirección ha especificado que los valores de las características de la calidad de este proceso deben estar entre 130 y 170. El fichero de datos es **Exercise 18-10**.

a) Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.

b) Halle el índice de capacidad C_p y analice el resultado.

c) Halle el índice C_{pk} y analice el resultado.

18.16. Vuelva al ejercicio 18.11. La dirección ha especificado que el peso debe estar entre 16 y 24 onzas. Utilice el fichero de datos **Exercise 18-11**.

a) Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.

b) Halle el índice de capacidad C_p y analice el resultado.

c) Halle el índice C_{pk} y analice el resultado.

18.4. Gráfico de control de proporciones

En lugar de analizar datos numéricos que midan alguna característica de un producto, consideremos ahora las situaciones en las que se valoran los productos por separado para ver si se ajustan o no a las especificaciones. Una vez más, se toma una secuencia de muestras a lo largo del tiempo para evaluar la calidad del producto y se representan los resultados en un gráfico de control. Es importante distinguir entre los términos *defecto* y *defectuoso*.

Defecto y defectuoso

«Un **defecto** es una única característica de la calidad de un producto que no se ajusta a las especificaciones. Un producto puede tener varios defectos. El término **defectuoso** se refiere a los productos que tienen uno o más defectos» (véase la referencia bibliográfica 5).

Lo que nos interesa es la proporción de productos de cada muestra que **no se ajustan a las especificaciones**, o sea, que son **defectuosos**. Evidentemente, es deseable que esta proporción sea lo más pequeña posible, por lo que cualquier tendencia ascendente a lo largo del tiempo debe ser motivo de preocupación. Se utiliza el gráfico *p* para controlar la proporción de artículos *defectuosos*. En el siguiente apartado se analiza el gráfico *c*, que se utiliza para controlar los *defectos*.

Una importante diferencia entre el desarrollo de gráficos de control de proporciones y el de gráficos del apartado 18.2 es que se necesitan muestras mucho mayores, ya que un proceso de producción bien desarrollado no va a generar una elevada proporción de productos que no se ajustan a las especificaciones. Por lo tanto, para hacer una evaluación razonable de esta medida de la calidad, es esencial que la muestra sea relativamente grande. En muchas aplicaciones, se recomienda que la muestra tenga entre 50 y 200 artículos, aunque a menudo es necesario que sea mayor. Una regla práctica que suele emplearse es que el número medio de artículos defectuosos por muestra sea, al menos, de cinco o seis. Así, por ejemplo, si se espera que alrededor del 1 por ciento de todos los artículos no se ajuste a las normas, se necesitan muestras de, al menos, 500 o 600 artículos. Una de las consecuencias de la necesidad de que la muestra sea mayor es que puede ser deseable tomar muestras de distinto tamaño. Por ejemplo, puede ser necesario inspeccionar toda la producción de un día o de un turno para tener suficientes observaciones. Normalmente, estos números no permanecen constantes. Aquí centramos la atención por comodidad en el caso en el que las muestras son del mismo tamaño, aunque es bastante sencillo extender el análisis al caso en el que las muestras son de tamaño distinto.

Otra cuestión importante para desarrollar gráficos de control de proporciones de artículos defectuosos es el elemento de subjetividad inherente a la generación de datos. Los artículos son valorados por inspectores y, dado el elemento de subjetividad que implica la valoración, es probable que las valoraciones varíen de unos inspectores a otros, por lo que en los gráficos podría haber variabilidad de más o podría parecer que no hay control. Es importante ser consciente de esta posibilidad cuando se interpretan gráficos de control de proporciones. Cuando los datos van a ser generados por más de un inspector, es necesario ser lo más específico posible al principio en la formulación de los criterios para decidir si un artículo es defectuoso o no.

La ecuación 18.12 permite hallar la media de proporciones muestrales.

Media de proporciones muestrales

Se toma a lo largo del tiempo una secuencia de K muestras, de n observaciones cada una, y se calculan las proporciones de miembros de las muestras que **no se ajustan** a las normas. Estas proporciones muestrales, representadas por $\hat{p}_i$ para $i = 1, 2, ..., K$, pueden representarse en un gráfico *p*. Si las muestras son del mismo tamaño, la **media de las proporciones muestrales** es la **proporción global de artículos defectuosos**. Es decir,

$$\bar{p} = \sum_{i=1}^{K} \frac{\hat{p}_i}{K} \qquad \textbf{(18.12)}$$

Si el proceso ha funcionado correctamente durante todo el periodo de observación, puede considerarse que cada una de las muestras se ha extraído de una población común. La proporción de artículos defectuosos que hay en esa población se estima por medio de la media de las proporciones muestrales, $\bar{p}$. Por lo tanto, recordando nuestro análisis anterior de la distribución muestral de proporciones muestrales, las proporciones muestrales individuales p_i tienen una distribución muestral de media estimada $\bar{p}$ y error típico

$$\hat{\sigma}_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Al igual que en el caso de otras aplicaciones del control de calidad, normalmente en los gráficos de control se fijan límites de tres errores típicos.

Gráfico *p*

El **gráfico *p*** es un gráfico temporal de la secuencia de proporciones muestrales de artículos defectuosos en el que el límite central es $LC_p = \bar{p}$. Los límites de control inferior y superior son

$$LCI_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{y} \quad LCS_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \qquad \textbf{(18.13)}$$

La fórmula del límite de control inferior de la ecuación 18.13 puede dar un valor negativo, que es, por supuesto, un valor imposible para una proporción. En ese caso, el límite de control inferior se fija en 0. En todo caso, la superación del límite inferior normalmente no es motivo de preocupación. Podría significar que el proceso es más fiable. Sin embargo, otra posibilidad podría ser que los inspectores no saben detectar los artículos defectuosos.

EJEMPLO 18.5. Componentes electrónicos defectuosos (gráfico *p*)

Se toman a lo largo del tiempo veinte muestras, de 200 observaciones cada una, de un componente electrónico. El número y la proporción de componentes de cada muestra que no se ajustan a las normas se muestran en la Tabla 18.3 y se encuentran en el fichero de datos **Nonconforming Components**. Construya el gráfico *p* correspondiente a estos datos.

Tabla 18.3. Artículos defectuosos en las muestras de 200 componentes electrónicos.

Muestra	N.º de artículos defectuosos	$\hat{p}$	Muestra	N.º de artículos defectuosos	$\hat{p}$
1	18	0,090	11	19	0,095
2	15	0,075	12	26	0,130
3	23	0,115	13	11	0,055
4	9	0,045	14	28	0,140
5	17	0,085	15	22	0,110
6	29	0,145	16	14	0,070
7	11	0,055	17	25	0,125
8	21	0,105	18	17	0,085
9	25	0,125	19	23	0,115
10	14	0,070	20	18	0,090

Solución

La media de estas proporciones muestrales es

$$\bar{p} = (0{,}090 + 0{,}075 + \cdots + 0{,}090)/20 = 0{,}09625$$

La Figura 18.7 muestra el gráfico p correspondiente a los datos de la Tabla 18.3. La línea central del gráfico es

$$LC_p = \bar{p} = 0{,}09625$$

El límite de control inferior es

$$LCI_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0{,}09625 - 3\sqrt{\frac{(0{,}09625)(0{,}90375)}{200}} = 0{,}09625 - 0{,}06256 = 0{,}03369$$

y el límite de control superior es

$$LCS_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0{,}09625 + 3\sqrt{\frac{(0{,}09625)(0{,}90375)}{200}} = 0{,}09625 + 0{,}06256 = 0{,}15881$$

En la Figura 18.7 puede verse que todas las proporciones muestrales se encuentran entre los límites de control y que la inmensa mayoría se encuentran muy lejos de estos límites. Parece que hay una variabilidad holgadamente alta en la calidad, que podría merecer una investigación en profundidad. Sin embargo, viendo el gráfico sería razonable concluir que el proceso está bajo control. En ese caso, en las condiciones actuales, alrededor del 9,6 por ciento de todos los artículos producidos no se ajusta a las normas.

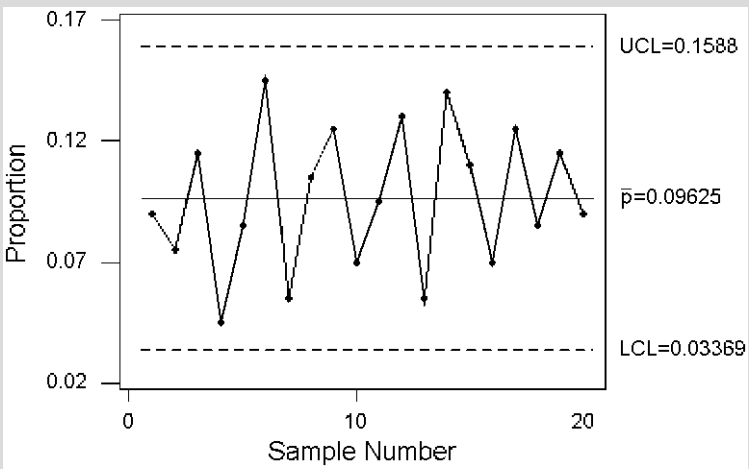


Figura 18.7. Gráfico p de los datos de los componentes defectuosos de la Tabla 18.3.

La interpretación de los gráficos *p* es similar a la de los gráficos del apartado 18.2. Se investigan con mayor profundidad los valores muestrales que se encuentran fuera de los límites de control y, si se encuentran las causas asignables de los valores extremos, se eliminan y se calculan de nuevo los límites de control. Un motivo de especial preocupación sería la aparición de una tendencia ascendente a lo largo del tiempo en un gráfico *p*. Esa tendencia sugeriría que puede estar aumentando la proporción de artículos defectuosos, es decir, que puede estar empeorando la calidad. Una vez que se ha llegado a la conclusión de que el proceso está bajo control, pueden utilizarse los límites para evaluar más datos. Sin embargo, al igual que ocurre con otros gráficos de control, es bueno calcular los límites de control periódicamente para tener en cuenta las mejoras del funcionamiento a medida que avanza el estudio de control de calidad.

Naturalmente, este análisis de los artículos defectuosos puede revelar que están produciéndose demasiados artículos que no se ajustan a las normas. En ese caso, puede ser deseable y posible hacer un análisis más detenido por medio de diagramas de Pareto. En el Capítulo 2 vimos que este tipo de gráfico es esencialmente un gráfico de barras, que aísla las causas por las que hay artículos defectuosos. Se enumeran los distintos problemas de estos artículos y se calcula el número de artículos que hay en cada categoría. Los gráficos de barras pueden organizarse para mostrar el número de productos que tienen diferentes tipos de defectos o los costes totales de corregir estos defectos. Con estos gráficos, la dirección debe ser capaz de hacerse rápidamente una idea de dónde es necesario concentrar los esfuerzos para lograr la máxima reducción de la tasa de productos defectuosos o del coste de rehacer esos productos. De esta manera, el estudio de control de calidad habrá hecho una valiosa contribución a la resolución de los problemas.

EJERCICIOS

Ejercicios aplicados

18.17. En el estudio de componentes de automóviles, se tomaron 30 muestras de 250 observaciones cada una. La media de las proporciones muestrales de piezas defectuosas era 0,056. Halle la línea central y los límites de control inferior y superior del gráfico *p*.

18.18. En el estudio de las piezas de aviones, el fabricante tomó 25 muestras de 500 observaciones cada una. La media de las proporciones muestrales de piezas defectuosas era 0,018. Halle la línea central y los límites de control inferior y superior del gráfico *p*.

18.19. El fichero de datos **Exercise 18-19** muestra las proporciones de artículos defectuosos de una secuencia de 30 muestras de 200 observaciones cada una.

a) Halle la línea central y los límites de control inferior y superior del gráfico *p*.

b) Trace el gráfico *p* y analice sus características.

18.20. El fichero de datos **Exercise 18-20** muestra las proporciones de artículos defectuosos de una secuencia de 20 muestras de 500 observaciones cada una.

a) Halle la línea central y los límites de control inferior y superior del gráfico *p*.

b) Trace el gráfico *p* y analice sus características.

18.21. El fichero de datos **Exercise 18-21** muestra el número de artículos defectuosos de una secuencia de 25 muestras de 250 observaciones cada una.

a) Halle la media de las proporciones muestrales.

b) Halle la línea central y los límites de control inferior y superior del gráfico *p*.

c) Trace el gráfico *p* y analice sus características.

18.5. Gráficos de control del número de ocurrencias

Recuérdese que la distribución de Poisson a menudo es útil para representar el *número de ocurrencias* de un suceso. Una aplicación habitual en el control de la calidad es inspeccionar un producto acabado y contar el número de defectos o imperfecciones de un determinado tipo. Si se inspeccionan artículos a lo largo del tiempo y se *cuenta* el número de imperfecciones de cada uno, esta información puede presentarse en un gráfico de control, llamado *gráfico c*.

He aquí algunas notaciones generales que se utilizan en los gráficos de control del número de ocurrencias.

Número medio muestral de ocurrencias

Se inspecciona a lo largo del tiempo una secuencia de K artículos. Se anota el número de ocurrencias de algún suceso, como una imperfección, en cada artículo. Estos *números de ocurrencias* se representan por medio de c_i para $i = 1, 2, ..., K$. El **número medio muestral de ocurrencias** es

$$\bar{c} = \sum_{i=1}^{K} \frac{c_i}{K} \qquad \textbf{(18.14)}$$

El número medio muestral de ocurrencias, $\bar{c}$, es una estimación de la media poblacional. Además, si la distribución del número de ocurrencias es una distribución de Poisson, la desviación típica de la distribución es la raíz cuadrada de la media:

$$\hat{\sigma}_c = \sqrt{\bar{c}}$$

El gráfico de control del número de ocurrencias puede construirse de la forma habitual.

Gráfico *c*

El **gráfico *c*** es un gráfico temporal del número de ocurrencias de un suceso. La línea central es

$$LC_c = \bar{c} \qquad \textbf{(18.15)}$$

Para los límites de tres errores típicos, el límite de control inferior es

$$\begin{aligned} LCI_c &= \bar{c} - 3\sqrt{\bar{c}} \quad \text{si } \bar{c} > 9 \\ LCI_c &= 0 \qquad\qquad\quad \text{si } \bar{c} \leqslant 9 \end{aligned} \qquad \textbf{(18.16)}$$

y el límite de control superior es

$$LCS_c = \bar{c} + 3\sqrt{\bar{c}} \qquad \textbf{(18.17)}$$

EJEMPLO 18.6. Gráfico *c* de un fabricante de textiles (gráfico *c*)

Un fabricante de textiles produce rollos de tela. Periódicamente inspecciona detenidamente un rollo y anota el número de imperfecciones. La Tabla 18.4 muestra una secuencia de 20 resultados anotados a lo largo del tiempo. En este tipo de situaciones, conviene que sea el mismo inspector el que examine cada pieza. En ese caso, las tendencias aparentes que se observen no se deberán a diferencias de criterio o de experien-

Tabla 18.4. Número de imperfecciones de los rollos de tela.

Rollo de tela	N.º de imperfecciones	Rollo de tela	N.º de imperfecciones	Rollo de tela	N.º de imperfecciones
1	8	8	2	15	1
2	8	9	3	16	7
3	6	10	10	17	9
4	8	11	7	18	11
5	9	12	6	19	9
6	5	13	8	20	6
7	7	14	2		

Cloth

cia de los inspectores. Construya el gráfico c. Los datos se encuentran en el fichero de datos **Cloth**.

Solución

En este ejemplo, el número medio de imperfecciones por rollo de tela es $\bar{c} = (8 + 8 + + \cdots + 6)/20 = 6{,}6$.

Ésta es una estimación natural de la media poblacional del número de imperfecciones por rollo. La desviación típica del número de ocurrencias se estima de la siguiente manera:

$$\sqrt{\bar{c}} = \sqrt{6{,}6} = 2{,}569$$

Dado que $\bar{c} - 3\sqrt{6{,}6}$ sería negativo, el límite de control inferior es $LCI_c = 0$. El límite de control superior es

$$LCS_c = \bar{c} + 3\sqrt{\bar{c}} = 6{,}6 + 3\sqrt{6{,}6} = 14{,}31$$

La Figura 18.8 muestra el gráfico c correspondiente a los datos de la Tabla 18.4.

La inspección de este gráfico c sugiere que no hay motivo alguno de preocupación. Las observaciones se encuentran todas ellas muy por debajo del límite de control superior y no existen pruebas de que esté aumentando el número de imperfecciones con el paso del tiempo. Parece, pues, que el proceso de producción está bajo control.

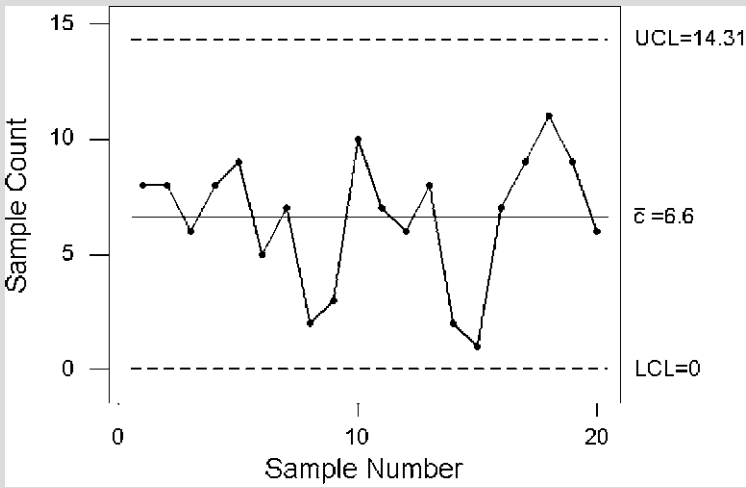


Figura 18.8. Gráfico c del fabricante de textiles.

EJERCICIOS

Ejercicios aplicados

18.22. Un proceso produce rollos de papel recubierto. Se inspecciona en distintos periodos de tiempo una secuencia de 20 rollos y se anota el número de imperfecciones. Los resultados se encuentran en el fichero de datos **Paper**.

a) Halle el número medio muestral de imperfecciones por rollo.
b) Halle la línea central y los límites inferior y superior de un gráfico c.
c) Trace el gráfico c y analice sus características.

18.23. El lector de un periódico lo ha leído detenidamente durante 20 semanas. En la edición de los miércoles ha contado el número de errores tipográficos. Los resultados se encuentran en el fichero de datos **Newspaper**.

a) Halle el número medio muestral de errores de estas 20 ediciones.
b) Halle la línea central y los límites inferior y superior de un gráfico c.
c) Trace el gráfico c y analice sus características.

18.24. Un proceso fabrica bollitos de pasas. Periódicamente se inspecciona uno y se cuenta el número de pasas que contiene. El fichero de datos **Raisins** muestra los resultados de 15 bollitos.

a) Halle el número medio muestral de pasas por bollito.
b) Halle la línea central y los límites inferior y superior de un gráfico c.
c) Trace el gráfico c y analice sus características.

RESUMEN

Los gráficos de control estadístico de este capítulo suministran la información necesaria para hacer un análisis documentado del nivel actual de calidad. Estos métodos no son difíciles de entender y el énfasis en los gráficos hace que la interpretación de los datos sea relativamente sencilla. Eso es importante, ya que permite acceder a la información a una amplia variedad de empleados sin necesidad de que entiendan difíciles conceptos estadísticos. De hecho, la comprensión de la *variabilidad* y de sus causas debería ser de gran ayuda para interpretar de una forma inteligente los datos. Ningún proceso genera productos absolutamente idénticos. Es inevitable que haya alguna *variabilidad natural* atribuible al azar. Un importante elemento del control de calidad es el reconocimiento de pautas en las mediciones que probablemente no se deban a la variabilidad natural sino que sean un indicio de la existencia de alguna causa estructural que debe investigarse.

Este capítulo no es más que una introducción a algunos métodos estadísticos que se emplean en la mejora continua de los procesos. Los gráficos de control no son en modo alguno los únicos instrumentos de los que se dispone. Para un programa de mejora de un proceso son esenciales los gráficos de flujos, los diagramas de Ishikawa (que suelen llamarse diagramas de espina de pescado o de causa-efecto), los diagramas de Pareto (Capítulo 2), los diagramas de puntos dispersos (Capítulo 2) y otras técnicas que quedan fuera del alcance de este libro. Para profundizar en estas cuestiones, véanse las notas y algunas páginas web que se indican al final del capítulo con el fin de obtener información sobre las ideas relativas a la calidad, los premios a la calidad, los seminarios en línea, los libros sobre la calidad y organizaciones como el Deming Institute y el Juran Institute.

En este capítulo hemos utilizado el programa Minitab por su sencillez y precisión. También existen otros paquetes estadísticos.

TÉRMINOS CLAVE

causas asignables de la variación, 733
causas comunes de la variación, 733
defecto, 750
defectuoso (que no se ajusta a las especificaciones), 750
desviación típica del proceso, 736
estimación de la desviación típica del proceso basada en intervalos, 761
estimación de la desviación típica del proceso basada en s, 736
gráfico c, 754
gráfico p, 751
gráfico R, 759
gráfico s, 740
gráfico $\bar{X}$, 739

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

18.25. Un fabricante de tornillos de precisión debe producir tornillos para un automóvil de lujo que tengan una fuerza media de 60.000 libras por pulgada cuadrada (psi). Cada 15 minutos se comprueba la fuerza de cuatro tornillos. El fichero de datos **Bolts** contiene los datos de un periodo de 3 horas. Trace el gráfico $\bar{X}$ y el gráfico s utilizando el programa Minitab.

18.26. Vuelva al fichero de datos **Bolts** del ejercicio 18.25. Halle los siguientes índices de capacidad del proceso utilizando el programa Minitab con $LCI = 58.500$ y $LCS = 61.500$.

a) «Capability Analysis» (distribución normal).
b) «Capability Sixpack» (distribución normal).

18.27. Construya e interprete el gráfico p utilizando el programa Minitab o algún otro para el fichero de datos **Exercise 18-21**.

18.28. Halle el gráfico c utilizando el programa Minitab o algún otro para los ficheros de datos:

a) **Paper** (ejercicio 18.22).
b) **Newspaper** (ejercicio 18.23).
c) **Raisin** (ejercicio 18.24).

18.29. Distinga entre cada uno de los siguientes pares de términos:

a) Un proceso *bajo control* y un proceso *capaz* de funcionar de acuerdo con unas especificaciones.
b) *Variabilidad natural* y *causas asignables*.

18.30. En los estudios de control de calidad es habitual emplear límites de tres errores típicos para trazar los gráficos. Explique la razón y las consecuencias.

18.31. El fichero de datos **Exercise 18-31** contiene las medias y las desviaciones típicas muestrales de una sucesión de 20 muestras de cinco observaciones cada una sobre una característica de la calidad de un producto.

a) Halle la media global de las observaciones muestrales.
b) Halle la desviación típica muestral media.
c) Utilice un estimador insesgado para estimar la desviación típica del proceso.
d) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
e) Trace el gráfico $\bar{X}$ y analice sus características.
f) Halle la línea central y los límites de control inferior y superior de un gráfico s.
g) Trace el gráfico s y analice sus características.
h) La dirección ha especificado que el valor de la característica de la calidad para este proceso debe estar entre 115 y 125.
  i. Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.
  ii. Halle el índice C_p y analice el resultado.
  iii. Halle el índice C_{pk} y analice el resultado.

18.32. El fichero de datos **Exercise 18-32** muestra las medias y las desviaciones típicas muestrales de una sucesión de 25 muestras de ocho observaciones cada una sobre una característica de la calidad de un producto.

a) Halle la media global de las observaciones muestrales.
b) Halle la desviación típica muestral media.
c) Utilice un estimador insesgado para estimar la desviación típica del proceso.
d) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
e) Trace el gráfico $\bar{X}$ y analice sus características.
f) Halle la línea central y los límites de control inferior y superior de un gráfico s.
g) Trace el gráfico s y analice sus características.
h) La dirección ha especificado que el valor de la característica de la calidad para este proceso debe estar entre 325 y 375.
  i. Calcule el intervalo $\bar{\bar{x}} \pm 3\hat{\sigma}$ y comente su resultado.
  ii. Halle el índice C_p y analice el resultado.
  iii. Halle el índice C_{pk} y analice el resultado.
  iv. Utilice el programa Minitab para responder a los apartados (a) a (h).

18.33. El fichero de datos **Exercise 18-33** muestra las proporciones de artículos defectuosos en una

secuencia de 20 muestras de 500 observaciones cada una.

a) Halle la media de las proporciones muestrales.
b) Halle la línea central y los límites de control inferior y superior de un gráfico *p*.
c) Trace el gráfico *p* y analice sus características.
d) Utilice el programa Minitab para trazar el gráfico *p*.

18.34. Unos grandes almacenes han registrado el número de reclamaciones que han presentado los clientes en un periodo de 18 semanas. Los resultados se encuentran en el fichero de datos **Complaints**.

a) Halle el número semanal medio muestral de reclamaciones.
b) Halle la línea central y los límites de control inferior y superior de un gráfico *c*.
c) Trace el gráfico *c* y analice sus características.
d) Utilice el programa Minitab para trazar el gráfico *c*.

18.35. El fichero de datos **Exercise 18-35** muestra las observaciones muestrales de una secuencia de 16 muestras de cuatro observaciones cada una sobre una característica de la calidad de un producto.

a) Halle las 16 medias muestrales y desviaciones típicas muestrales.
b) Halle la media global de las observaciones muestrales.
c) Halle la desviación típica muestral media.
d) Utilice un estimador insesgado para estimar la desviación típica del proceso.
e) Halle la línea central y los límites de control inferior y superior de un gráfico $\bar{X}$.
f) Utilice el programa Minitab o algún otro para trazar el gráfico $\bar{X}$ y el gráfico *s*.
g) Halle la línea central y los límites de control inferior y superior de un gráfico *s*.
h) Trace el gráfico *s* y analice sus características.

18.36. Averigüe si es más probable que cada una de las causas siguientes sea una causa común o una causa asignable:

a) Mala iluminación
b) Elevado grado de humedad
c) Sustitución de un operario
d) Ajuste incorrecto de la máquina
e) Anotación incorrecta de los datos

18.37. Un producto de consumo que ha prosperado en los últimos años es el agua mineral embotellada. Jon Thorne es el director general de una empresa que vende agua mineral embotellada. Ha pedido un informe del proceso mediante el cual se llenan las botellas de 24 onzas (710 ml) para asegurarse de que están llenándose correctamente. Para comprobar si el poceso debe ajustarse, Emma Astrom, que lo controla, toma muestras aleatorias y pesa cinco botellas cada 15 minutos durante un periodo de 5 horas. Los datos se encuentran en el fichero de datos **Bottles**.

a) Trace el gráfico $\bar{X}$ y el gráfico *s* de este problema.
b) Busque las causas asignables y averigüe si el proceso es estable.
c) Si el límite de especificación inferior es 685 ml y el superior es 730 ml, halle la capacidad del proceso.

18.38. Prairie Flower Cereal Inc. es un productor pequeño, pero en expansión, de cereales de desayuno que sólo deben calentarse para comerlos. Gordon Thorson, próspero agricultor que cultiva cereales, creó la empresa en 1910 (véase la referencia bibliográfica 1). Se le ha pedido que compruebe el proceso de empaquetado de cajas de cereales de trigo azucarados de 18 onzas (510 gramos). En el proceso de empaquetado se utilizan dos máquinas. Se toman aleatoriamente veinte muestras de cinco cajas cada una y se pesan. Los datos se encuentran en el fichero **Sugar Coated Wheat**. Los límites de especificación inferior y superior se han establecido en 500 y 525 gramos, respectivamente.

a) Averigüe si el proceso de empaquetado de la máquina 1 está bajo control.
b) Averigüe si el proceso de empaquetado de la máquina 2 está bajo control.
c) ¿Es la máquina 1 capaz de cumplir los límites de especificación?
d) ¿Es la máquina 2 capaz de cumplir los límites de especificación?
e) ¿Qué recomendaciones haría a Prairie Flower Cereal Inc. sobre el proceso de empaquetado de cereales de trigo azucarados?

18.39. Otro producto empaquetado por Prairie Flower Cereal Inc. es el de cereales con canela y manzana. Para comprobar el proceso de empaquetado de cajas de este cereal de 40 onzas (1.134 gramos), se toman aleatoriamente 23 muestras de seis cajas cada una y se pesan.

Los límites de especificación inferior y superior se han fijado en 1.120 y 1.150 gramos, respectivamente. Los datos se encuentran en el fichero de datos **Granola**.

a) ¿Es estable el proceso de empaquetado?
b) Si es estable, averigüe la capacidad del proceso para satisfacer las especificaciones dadas.

18.40. Al Fiedler, director de planta de LDS Vacuum Products, que se encuentra en Altamonte Springs (Florida), aplica la teoría estadística en su centro de trabajo. LDS, importante proveedor de los fabricantes de automóviles, quiere estar seguro de que la tasa de incidencia de fugas (en centímetros cúbicos por segundo) de los enfriadores del aceite de la transmisión (TOC) satisface los límites de especificación establecidos. Se comprueban muestras aleatorias de enfriadores y se registran las tasas de incidencia de fugas en el fichero de datos **TOC**. Compruebe si el proceso es estable. El tamaño de los subgrupos es de cinco.

Apéndice

Antes de que existieran los programas informáticos, para examinar la variabilidad de los procesos se utilizaban más a menudo gráficos R de los intervalos que gráficos s de las desviaciones típicas, ya que para los trabajadores de la planta era más fácil calcular la diferencia entre el mayor valor muestral y el menor que calcular las desviaciones típicas muestrales. Si el gráfico R mostraba que el proceso era estable, se examinaba el gráfico $\bar{X}$ basado en intervalos muestrales. Aquí analizamos el gráfico R para completar el estudio.

1. Gráfico *R*

Las ecuaciones 18.18 y 18.19 son la línea central y los límites de control del gráfico R.

Gráfico *R*

El **gráfico *R*** es un gráfico temporal de la secuencia de intervalos que tiene la línea central

$$LC_R = \bar{R} \tag{18.18}$$

y los límites de control

$$LCI_R = D_3\bar{R} \qquad LCS_R = D_4\bar{R} \tag{18.19}$$

donde las constantes D_3 y D_4 se indican en la Tabla 13 del apéndice. La Tabla 18.5 contiene algunas constantes del gráfico de control.

Tabla 18.5. Algunas constantes de los gráficos de control.

n	d_2	A_2	D_3	D_4
2	1,128	1,88	0	3,27
3	1,693	1,02	0	2,57
4	2,059	0,73	0	2,28
5	2,326	0,58	0	2,11
6	2,534	0,48	0	2,00
7	2,704	0,42	0,08	1,92
8	2,847	0,37	0,14	1,86
9	2,970	0,34	0,18	1,82
10	3,078	0,31	0,22	1,78

A continuación hallamos el gráfico *R* correspondiente al fichero de datos **Signal** (ejemplo 18.1). Los intervalos muestrales se indican en la Tabla 18.6.

Tabla 18.6. Intervalos muestrales correspondientes al ejemplo de las señales.

Muestra	*R*	Muestra	*R*	Muestra	*R*	Muestra	*R*
1	7	6	12	11	11	16	7
2	7	7	9	12	13	17	8
3	10	8	15	13	21	18	11
4	13	9	12	14	6	19	11
5	16	10	13	15	8	20	8

El intervalo muestral medio es

$$\bar{R} = (7 + 7 + \cdots + 8)/20 = 10{,}9$$

Utilizando las ecuaciones 18.18 y 18.19, la línea central del gráfico *R* es

$$LC_R = \bar{R} = 10{,}9$$

Si el tamaño de los subgrupos es $n = 5$, los límites de control son

$$LCI_R = D_3\bar{R} = 0(10{,}9) = 0 \quad \text{y} \quad LCS_R = D_4\bar{R} = (2{,}11)(10{,}9) = 23$$

Trazamos los intervalos muestrales en el gráfico *R* o utilizando el programa Minitab y obtenemos la Figura 18.9.

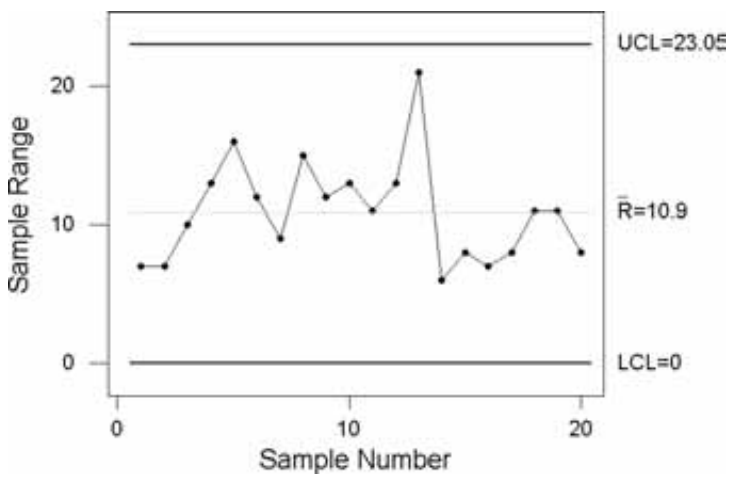


Figura 18.9. Gráfico de intervalos del ejemplo de las señales.

2. Gráfico $\bar{X}$ y gráfico *R*

Dado que la inspección del gráfico *R* no indica que haya motivo alguno para preocuparse, ahora desarrollamos el gráfico $\bar{X}$ basado en intervalos.

Gráfico $\bar{X}$ basado en intervalos

El **gráfico $\bar{X}$ basado en intervalos** es un gráfico temporal de la secuencia de medias que tiene la línea central

$$LC_{\bar{X}} = \bar{\bar{x}} \tag{18.20}$$

y los límites de control

$$LCI_{\bar{X}} = \bar{\bar{x}} - A_2\bar{R} \quad \text{y} \quad LCS_{\bar{X}} = \bar{\bar{x}} + A_2\bar{R} \tag{18.21}$$

donde A_2 se encuentra en la Tabla 13 del apéndice. La Tabla 18.5 muestra algunos valores de A_2. Puede demostrarse que

$$A_2 = \frac{3}{d_2\sqrt{n}} \tag{18.22}$$

En el ejemplo 18.1 observamos que la media global era $\bar{\bar{x}} = 299{,}9$ y en la Tabla 13 del apéndice o en la 18.5 vemos que con $n = 5$, el valor de la constante A_2 es 0,58. Los límites de control son, pues,

$$LCI_{\bar{X}} = \bar{\bar{x}} - A_2\bar{R} = 299{,}9 - (0{,}58)(10{,}9) = 293{,}6$$

y

$$LCS_{\bar{X}} = \bar{\bar{x}} + A_2\bar{R} = 299{,}9 + (0{,}58)(10{,}9) = 306{,}2$$

Ahora utilizamos los datos de **Signal** para obtener tanto el gráfico $\bar{X}$ como el R juntos (véase la Figura 18.10).

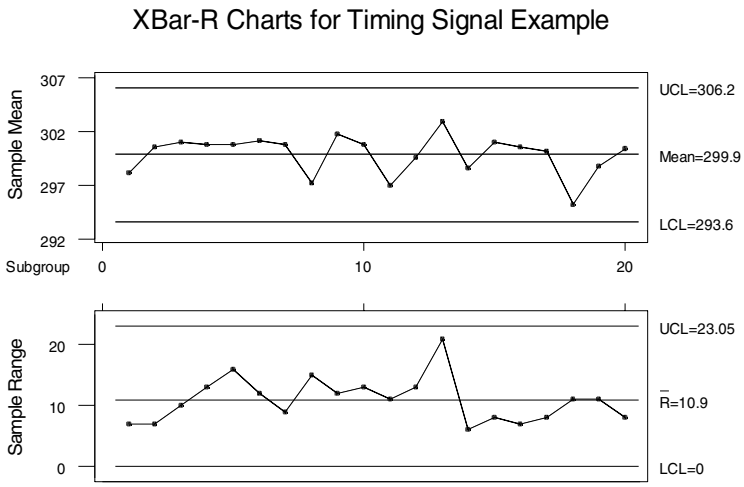


Figura 18.10. Gráfico $\bar{X}$ y gráfico R del ejemplo de las señales.

Estimación de la desviación típica del proceso basada en intervalos

La ecuación 18.23 es una **estimación de la desviación típica del proceso basada en intervalos**:

$$\hat{\sigma} = \bar{R}/d_2 \tag{18.23}$$

donde d_2 se encuentra en la Tabla 13 del apéndice.

En el ejemplo de las señales, la estimación de la desviación típica del proceso es, pues,

$$\hat{\sigma} = \bar{R}/d_2 = 10{,}9/2{,}326 = 4{,}69$$

Ahora puede calcularse la capacidad del proceso, C_p y C_{pk}, por medio de las ecuaciones 18.10 y 18.11, utilizando esta estimación de la desviación típica del proceso.

Bibliografía

1. Carlson, William L., *Cases in Managerial Data Analysis*, San Francisco, Duxbury Press, 1997.
2. Deming, W. Edwards, *Out of the Crisis*, Cambridge, MIT Center for Advanced Engineering Study, 1986.
3. Deming, W. Edwards, *The New Economics for Business*, *Industry*, *and Government*, Cambridge, MIT Center for Advanced Engineering Study, 1993.
4. Evans, James R., *Production/Operations Management*: *Quality*, *Performance*, *and Value*, Minneapolis/St. Paul, MN, West Publishing Company, 1997, 5.ª ed.
5. Evans, James R. y William M. Lindsay, *The Management and Control of Quality*, Cincinnati, OH, Southwestern College Publishing, 2005, 6.ª ed.
6. Fiedler, Alfred W., *LDS Vacuum Products Study*: *Delphi Leak Detector #1*, Altamonte Springs, FL, LDS Vacuum Products, 2000.
7. Juran, Joseph M., *Juran on Quality by Design*, Nueva York, Free Press, 1995, revision.
8. Juran, Joseph M. y A. Blanton Godfrey, *Juran's Quality Handbook*, Nueva York, McGraw-Hill, 1999, 5.ª ed.
9. «Statistical Thinking», ASQ Statistics Division Newsletter, Special Publication, 1996.
10. Taguchi, Genichi, *Introduction to Quality Engineering*, Tokio, Asian Productivity Organization, 1986.
11. Walton, Mary, *The Deming Management Method*, Nueva York, Putnam Publishing Group, 1986.
12. Zimmerman, Steven M. y Marjorie L. Icenogle, *Statistical Quality Control Using Excel*, Milwaukee, WI, ASQ Quality Press, 2002, 2.ª ed.

Algunas direcciones actuales de páginas web de interés

Dirección	Organización
www.asq.org	American Society for Quality (ASQ)
www.deming.org	W. Edwards Deming Institute
www.efqm.org	European Foundation for Quality Management
www.nist.gov	National Institute of Standards and Technology
www.juran.com	Juran Institute
www.nokia.com	Nokia
www.philipcrosby.com	Philip Crosby Associates II, Inc.
www.qualitypress.asq.org	ASQ On-Line Bookstore

Capítulo 19

Análisis de series temporales y predicción

Esquema del capítulo

Introducción

En este capítulo presentamos métodos para analizar conjuntos de datos que contienen mediciones de varias variables a lo largo del tiempo. Ejemplos de datos de series temporales son las ventas mensuales de un producto y los tipos de interés, los beneficios empresariales trimestrales y el consumo agregado y las cotizaciones al cierre de la bolsa.

Serie temporal

Una serie temporal es un conjunto de mediciones, ordenadas en el tiempo, sobre una cantidad de interés. En una serie temporal, la secuencia de observaciones es importante, a diferencia de lo que ocurre en los datos de corte transversal, en el que la secuencia de observaciones no es importante.

Los datos de series temporales normalmente poseen características especiales —relacionadas con la secuencia de observaciones— que exigen el desarrollo de métodos de análisis estadístico especiales. Casi todos los métodos de análisis de datos y de inferencia que hemos desarrollado se basan en el supuesto de que las muestras son

aleatorias, en concreto, de que los errores de las observaciones son independientes. El supuesto de la independencia raras veces es realista en el caso de los datos de series temporales. Consideremos, por ejemplo, una serie de ventas mensuales de un producto manufacturado y observemos las razones posibles por las que no son independientes. Si el mes pasado las ventas fueron superiores a la media, es razonable esperar que continúen siendo altas, ya que no es probable que cambie bruscamente la situación de la economía y de las empresas. Por lo tanto, es de esperar que las ventas de meses contiguos sean similares. También observamos que las ventas de muchos productos tienen una pauta estacional: los pantalones cortos y los bañadores se venden más en primavera y a principios del verano que en invierno. Muchas tiendas minoristas venden más en el cuarto trimestre debido a las compras de regalos de Navidad. Éstos y otros muchos ejemplos demuestran la ausencia de independencia.

La ausencia de independencia entre las observaciones de series temporales plantea serios problemas si se utilizan con datos de series temporales los métodos estadísticos convencionales, que suponen que las observaciones son independientes. Ya vimos el problema en el apartado 14.7 cuando analizamos las dificultades que se plantean si se utilizan métodos convencionales de regresión cuando los errores están correlacionados. El supuesto de la independencia es fundamental; también pueden plantearse otros problemas serios si se utilizan métodos convencionales cuando las observaciones son dependientes. En este capítulo, centramos la atención en los métodos de análisis de series temporales que se utilizan cuando hay una única serie temporal.

Hemos analizado el aspecto negativo de los tipos de pautas de dependencia que es probable que aparezcan en los datos de series temporales. Estos problemas son reales y requieren métodos especiales. Sin embargo, esta dependencia también puede explotarse para realizar predicciones de los futuros valores de los datos de series temporales cuya varianza es menor. Por ejemplo, si hay una correlación entre errores de meses contiguos en una serie de ventas al por menor, esa correlación puede utilizarse para hacer una predicción de las ventas del próximo mes mejor que una predicción basada en una muestra aleatoria. Presentaremos métodos basados en el supuesto de que las pautas anteriores de relación entre mediciones de una serie temporal se mantendrán en el futuro y pueden utilizarse para hacer predicciones, lo cual es como afirmar que podemos aprender en realidad del estudio de la historia.

En el primer apartado desarrollamos números índice, que se utilizan en algunos estudios económicos. Los métodos de análisis de series temporales que se presentan en los apartados posteriores no requieren el conocimiento de los números índice. Se incluyen aquí para hacer una presentación completa de los temas relacionados con el análisis de series temporales.

19.1. Números índice

Nuestro análisis comienza con el desarrollo de números índice. Consideremos, a modo de introducción, la siguiente pregunta: ¿qué variaciones ha experimentado el precio de los automóviles fabricados en Estados Unidos en los 10 últimos años? Ni que decir tiene que ha subido, pero ¿cómo puede describirse cuantitativamente esta subida? A primera vista, no parece que sea muy difícil responder a esta pregunta. El primer paso sería recoger información sobre el precio de estos automóviles en cada uno de los 10 últimos años y representarlo en un gráfico temporal.

Sin embargo, el análisis detenido del problema podría plantear algunas preguntas. En primer lugar, observamos que los automóviles no son homogéneos, por lo que es necesario definir con más precisión el tipo de automóvil. Existe claramente una amplia variedad de

precios y de calidades y la variación del precio medio de todos los automóviles vendidos podría deberse meramente a un cambio de la pauta de compra: ¿se venden automóviles de precio más alto? En este caso, el precio medio subiría, porque tenemos automóviles de precio más alto. Otros cambios de la combinación de mercado podrían provocar otras variaciones de la media. La Tabla 19.1 muestra un sencillo ejemplo hipotético de un mercado en el que sólo hay automóviles de precio bajo y automóviles de precio alto. Obsérvese que el precio medio baja, pero que esta bajada se debe a que en la mezcla hay más automóviles de precio bajo y menos de precio alto. Esta forma de comparar el precio de los automóviles de dos años diferentes no es especialmente útil.

Tabla 19.1. Datos hipotéticos sobre los precios y las ventas de automóviles.

	Automóviles pequeños		Automóviles de lujo		Todos los automóviles
Año	Precio (miles de dólares)	Número vendido (miles)	Precio (miles de dólares)	Número vendido (miles)	Precio medio (miles de dólares)
1	10	5	30	15	25,0
2	11	15	33	5	16,5

Otra solución es calcular el precio medio considerando un único automóvil de cada tipo, como en la Tabla 19.2. Este método también tiene problemas, porque tenemos un mercado en el que los automóviles pequeños son considerablemente más populares que los de lujo. El precio de los primeros es el mismo en los dos años, mientras que el de los segundos se duplica. Como consecuencia, la media calculada considerando un único automóvil de cada tipo es mucho más alta en el segundo año. Pero esta media no refleja exactamente la situación, ya que da el mismo peso a los dos tipos de automóvil cuando, en realidad, los automóviles pequeños se compran mucho más a menudo.

Tabla 19.2. Datos hipotéticos sobre los precios y las ventas de automóviles: igual ponderación.

	Automóviles pequeños		Automóviles de lujo		Todos los automóviles
Año	Precio (miles de dólares)	Número vendido (miles)	Precio (miles de dólares)	Número vendido (miles)	Precio medio de cada tipo de automóvil (miles de dólares)
1	10	100	24	1	17
2	10	100	48	1	29

Estos ejemplos demuestran que, para hacernos una idea fiable de la pauta general de los precios a lo largo del tiempo, hay que tener en cuenta las cantidades compradas en cada periodo. Veremos cómo pueden calcularse medias ponderadas adecuadas.

Se plantea el mismo problema si los compradores compran más automóviles con más extras el segundo año que el primero. En ese caso, compran implícitamente automóviles de mayor calidad que en el primer año. Podríamos examinar solamente los precios de los automóviles sin extras para hacer una comparación válida.

Las mejoras tecnológicas plantean otra dificultad. No es sorprendente observar que los automóviles actuales consumen menos gasolina y duran más que los que se fabricaban ha-

ce 20 o 30 años. Por lo tanto, los cambios de la calidad pueden influir mucho en las subidas de los precios. Es muy importante tenerlos en cuenta cuando se hacen comparaciones de precios, pero las técnicas para analizar su influencia quedan fuera del alcance de este libro.

Hemos puesto ejemplos de un único producto para ilustrar el problema, pero esas comparaciones normalmente sólo tienen interés para las personas relacionadas directamente con la compraventa de ese producto. Nos dedicaremos, pues, a comparar las variaciones de los precios de unos productos con las variaciones de los precios de otros.

El problema de números índice que examinamos a continuación tiene por objeto comparar las variaciones de los precios de un grupo de mercancías. Por ejemplo, el precio de las acciones de empresas que cotizan en bolsa varía en un mes. Nos gustaría desarrollar una medida de la variación agregada de los precios. Los números índice pretenden resolver esos problemas.

Índice de precios de un único artículo

Comenzamos nuestro análisis de los números índice con un sencillo caso. La Figura 19.1 es una hoja de cálculo Excel que muestra el cálculo de un índice de precios de las acciones de Ford Motor Company en un periodo de 12 semanas. La segunda columna contiene el precio efectivo de las acciones. Es algo difícil interpretar estos números, pero esta tarea puede simplificarse calculando un índice de precios utilizando el precio de la primera semana como periodo base. En la tercera columna, vemos el índice de precios calculado. Así, el índice de precios de la segunda semana es

$$100\left(\frac{19{,}875}{20{,}25}\right) = 98{,}1$$

basándose en el precio de la segunda semana de 19,875. Los porcentajes calculados de esta forma se llaman *números índice del precio*. La elección del periodo base es arbitraria. Podríamos haber elegido cualquier otra semana como base y haber expresado todos los precios en porcentaje del precio de esa semana.

La ventaja de utilizar aquí números índice reside en que es más fácil interpretar los números. Por ejemplo, en la Figura 19.1 vemos inmediatamente que el precio de las acciones de Ford Motor Company fue un 13,6 por ciento más alto en la semana 12 que en la 1.

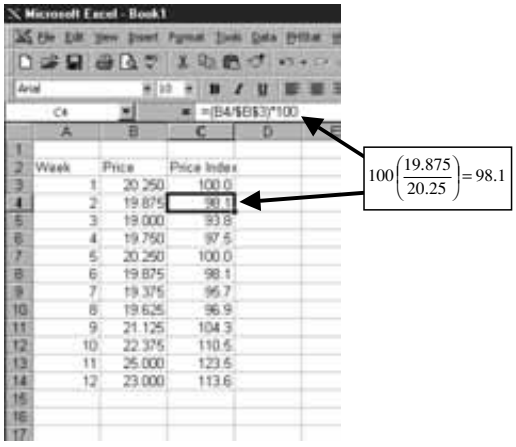


Figura 19.1. Precios e índice de precios de las acciones de Ford Motor Company en 12 semanas.

Cálculo de índices de precios de un único artículo

Supongamos que tenemos una serie de observaciones a lo largo del tiempo del precio de un único artículo. Para construir un índice de precios, elegimos como base un periodo de tiempo y expresamos el precio de cada periodo en porcentaje del precio del periodo base. Por lo tanto, si p_0 representa el precio del periodo base y p_1 el precio del segundo periodo, el índice de precios del segundo periodo es

$$100\left(\frac{p_1}{p_0}\right)$$

Índice de precios agregado no ponderado

A continuación, vemos cómo se representan las variaciones de los precios agregados de un grupo de artículos. La Figura 19.2 es una hoja de cálculo Excel que muestra los precios pagados a los agricultores estadounidenses, en dólares, por quintal por el trigo, el maíz y la soja en 10 años. La tabla también muestra una manera de lograr un índice de precios agregado de estos cultivos. Calculamos el precio medio de cada año y utilizamos esa media para construir un índice de la media, utilizando el primer año como base.

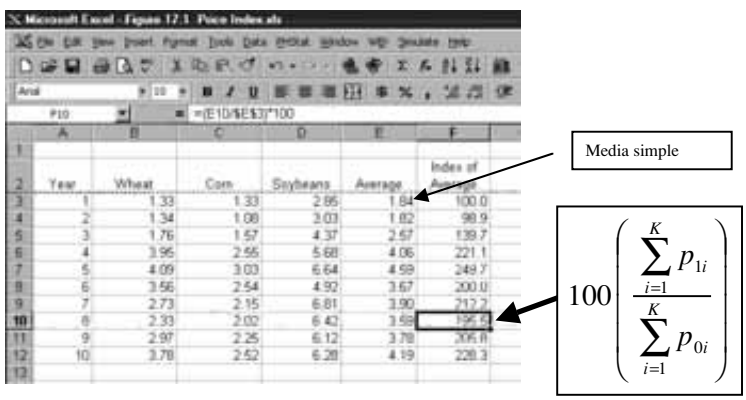


Figura 19.2. Precios por quintal de tres cultivos en 10 años: índice de precios agregado no ponderado.

Es fácil calcular el índice de precios agregado no ponderado, como muestra la Figura 19.2. Expresa el precio medio de cada año en porcentaje del precio medio del año base. Sin embargo, no tiene en cuenta las diferencias entre las cantidades cultivadas de estos productos. La fórmula de la Figura 19.2 indica la división de las sumas de los precios. Eso es, por supuesto, lo mismo que dividir por las medias de estos precios. Estas medias serían el resultado de dividir las sumas del numerador y del denominador por 3.

Un índice de precios no ponderado

Supongamos que tenemos una serie de observaciones en el tiempo sobre los precios de un grupo de K artículos. Se elige como base un periodo de tiempo.

El **índice de precios agregado no ponderado** se obtiene calculando el precio medio de estos artículos en cada periodo de tiempo y calculando a continuación un índice de estos precios medios. Es decir, el precio medio de cada periodo se expresa en porcentaje del precio medio del periodo base. Sea p_{0i} el precio del i-ésimo artículo en el periodo base y p_{1i} el precio

de este artículo en el segundo periodo. El índice agregado no ponderado de precios de este segundo periodo es

$$100\left(\frac{\sum_{i=1}^{K} p_{1i}}{\sum_{i=1}^{K} p_{0i}}\right)$$

Índice de precios agregado ponderado

En general, nos gustaría ponderar los precios por alguna medida de la cantidad vendida. Una posibilidad es utilizar las cantidades medias de algunos de los periodos en cuestión o de todos. En muchos casos, es caro obtener cantidades, por lo que los índices se basan en cantidades de un único periodo. Cuando estas cantidades proceden del periodo base, el índice resultante se llama *índice de precios de Laspeyres*.

El índice de Laspeyres compara, en efecto, el coste total de comprar las cantidades del periodo base en el periodo base con el coste total de comprar estas mismas cantidades en otros periodos. Para ilustrarlo, consideremos los datos de la Figura 19.2 sobre los precios de los cultivos con la información adicional de que la producción en el año 1 fue de 1.352 millones de quintales de trigo, de 4.152 millones de quintales de maíz y de 1.127 millones de quintales de soja. Por lo tanto, el coste, en millones de dólares, de la producción total del año 1 fue

$$(1.352)(1{,}33) + (4.152)(1{,}33) + (1.127)(2{,}85) = 10.532$$

En el año 2, a los precios vigentes entonces, el coste total de comprar las cantidades del año base habría sido

$$(1.352)(1{,}34) + (4.152)(1{,}08) + (1.127)(3{,}03) = 9.711$$

El índice de precios de Laspeyres del año 2 es, pues,

$$100\left(\frac{9.711}{10.532}\right) = 92{,}2$$

La Figura 19.3 muestra el índice completo correspondiente a estos datos calculado de esta forma.

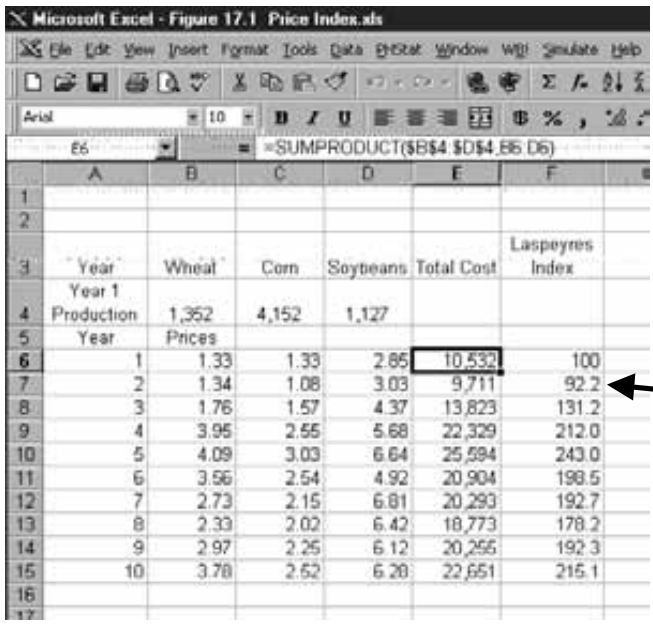
El índice de precios de Laspeyres

Supongamos que tenemos un grupo de K mercancías de las cuales se dispone de información sobre los precios que tenían en un periodo de tiempo. Se selecciona un periodo como base del índice. El **índice de precios de Laspeyres** en cualquier periodo es el coste total de comprar las cantidades comerciadas en el periodo base a los precios del periodo de interés, en porcentaje del coste total de comprar estas mismas cantidades en el periodo base.

Sea p_{0i} el precio y q_{0i} la cantidad comprada del i-ésimo artículo en el periodo base. Si p_{1i} es el precio del i-ésimo artículo en el segundo periodo, el índice de precios de Laspeyres del periodo es

$$100\left(\frac{\sum_{i=1}^{K} q_{0i}p_{1i}}{\sum_{i=1}^{K} q_{0i}p_{0i}}\right)$$

Figura 19.3.
Índice de precios de Laspeyres de tres cultivos.



Microsoft Excel - Figure 17.1 Price Index.xls

File Edit View Insert Format Tools Data StatAdd-Ins Window Help Simulate Help

Arial 10

E6 =SUMPRODUCT($B$4:$D$4,B6:D6)

	A	B	C	D	E	F
1						
2						
3	Year	Wheat	Corn	Soybeans	Total Cost	Laspeyres Index
4	Year 1 Production	1,352	4,152	1,127		
5	Year	Prices				
6	1	1.33	1.33	2.85	10,532	100
7	2	1.34	1.08	3.03	9,711	92.2
8	3	1.76	1.57	4.37	13,823	131.2
9	4	3.95	2.55	5.68	22,329	212.0
10	5	4.09	3.03	6.64	25,594	243.0
11	6	3.56	2.54	4.92	20,904	198.5
12	7	2.73	2.15	6.81	20,293	192.7
13	8	2.33	2.02	6.42	18,773	178.2
14	9	2.97	2.25	6.12	20,256	192.3
15	10	3.78	2.62	6.28	22,651	215.1
16						

$$100\left(\frac{9{,}711}{10{,}532}\right) = 92.2$$

Es útil comparar la fórmula del índice de precios de Laspeyres con la del índice de precios agregado no ponderado. La diferencia es que, cuando se calcula el índice de Laspeyres, el precio de cada artículo se pondera por la cantidad comerciada en el periodo base.

Vemos que el índice de precios de Laspeyres utiliza únicamente la información sobre la cantidad del periodo base. Eso es valioso cuando es difícil obtener esa información de cada periodo. Podría ser un inconveniente si las cantidades del periodo base no fueran representativas de la serie temporal examinada. Por lo tanto, el índice de precios podría quedarse anticuado. Este problema puede resolverse calculando un índice de precios de Laspeyres móvil, en el que el periodo base se cambia de vez en cuando obteniendo información sobre la cantidad de los nuevos periodos base. Muchos de los índices de precios oficiales que se publican, como el índice de precios de consumo, se calculan esencialmente de esta forma.

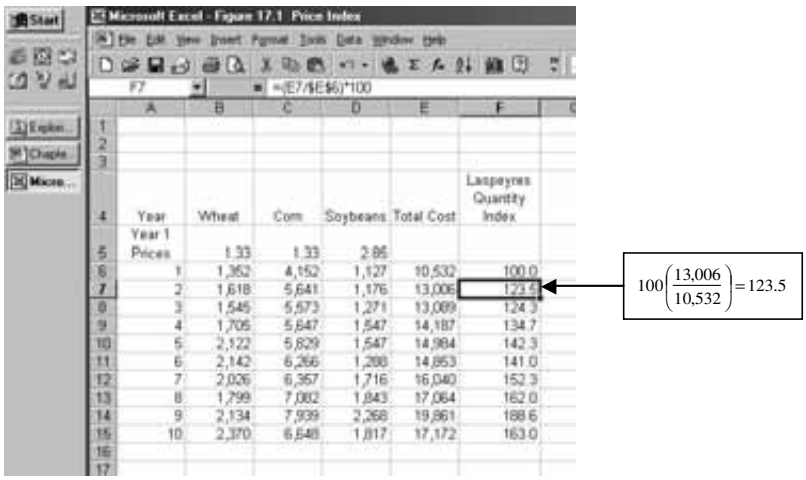
Índice de cantidades agregado ponderado

Los índices de precios constituyen una representación de la evolución de los precios agregados de un grupo de mercancías. También podríamos querer una representación de la evolución de las cantidades totales comerciadas. De nuevo, es probable que cualquier enfoque razonable de este problema dé como resultado un índice de cantidades ponderado, ya que probablemente querríamos dar más peso a un cambio de la cantidad comprada de un artículo muy caro que a un cambio de la misma cantidad comprada de un artículo barato. Un método para lograrlo es el *índice de cantidades de Laspeyres*, que ilustramos con las cantidades producidas de trigo, maíz y soja de la Figura 19.4.

El índice de cantidades de Laspeyres pondera las cantidades por los precios del periodo base. Las ponderaciones de los precios son 1,33, 1,33 y 2,85 en el caso del trigo, el maíz y la soja, lo que da como resultado un valor total en el año 1 de 10.532 millones de dólares. Para obtener un índice de cantidades del año 2, lo comparamos con el valor total de la producción del año 2, si hubieran estado vigentes los precios del año 1; es decir,

$$(1.618)(1{,}33) + (5.641)(1{,}33) + (1.176)(2{,}85) = 13.006$$

Figura 19.4. Producción, en millones de quintales, e índice de cantidades.



El índice de cantidades de Laspeyres del año 2 es, pues,

$$100\left(\frac{13.006}{10.532}\right) = 123{,}5$$

La Figura 19.4 muestra las cantidades producidas y el índice de cantidades de un periodo de 10 años.

El índice de cantidades de Laspeyres

Tenemos datos sobre la cantidad de un conjunto de artículos recogidos durante un conjunto de K años. Se selecciona un periodo como periodo base. El **índice de cantidades de Laspeyres** en cualquier periodo es el coste total de las cantidades comerciadas en ese periodo, basado en los precios del periodo base y expresado en porcentaje del coste total de las cantidades del periodo base.

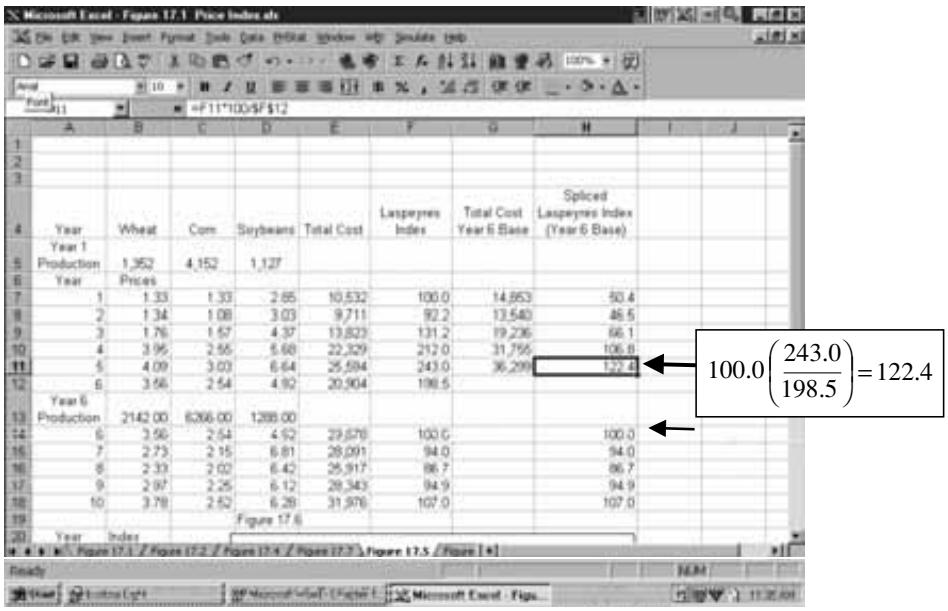
Sean q_{0i} y p_{0i} la cantidad y el precio del i-ésimo artículo en el periodo base y q_{1i} la cantidad de ese artículo en el periodo de interés. El índice de cantidades de Laspeyres de ese periodo es, pues,

$$100\left(\frac{\sum_{i=1}^{K} q_{1i}p_{0i}}{\sum_{i=1}^{K} q_{0i}p_{0i}}\right)$$

Cambio del periodo base

Las series oficiales de números índice se actualizan cambiando el periodo base por uno más reciente. En estos casos, normalmente se calcula el valor del índice original en el periodo que ahora se toma como base. Obsérvese a modo de ilustración el cálculo de la columna F de la Figura 19.5, que muestra los índices de precios del trigo, el maíz y la soja. La columna F muestra el índice de precios de los cultivos de los años 1 a 6, utilizando el año 1 como base comenzando por la fila 14 de la columna F. La columna H indica el índice de precios de Laspeyres de los años 6 a 10, utilizando el año 6 como base. Estos índices se representan en la Figura 19.6, en la que es evidente la discontinuidad en el año 6.

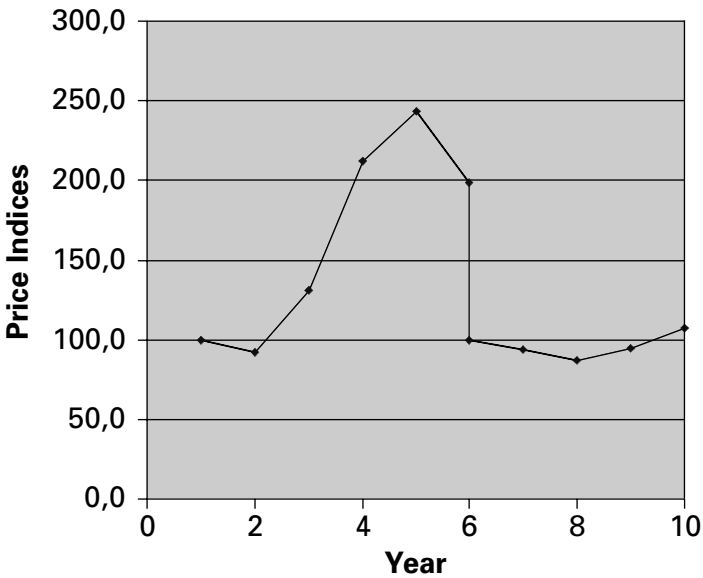
Figura 19.5. Índice de precios agregado de Laspeyres utilizando diferentes años base.



Examinando la Figura 19.6, es difícil comprender claramente las pautas de precios de todo el periodo. Por lo tanto, preferiríamos examinar un **índice de precios enlazado** que tuviera el año 6 como año base. En el índice original basado en el año 1, el índice del año 6 era 198,5 como se ve en la Figura 19.5. Para transformar el índice del año 6 basado en el año 1 en un índice del año 6 tomando como base el año 6, dividimos por 198,5 y multiplicamos por 100. También podemos convertir todos los demás índices cuya base es el año 1 a una base del año 6 dividiendo por 198,5 y multiplicando por 100. Por ejemplo, el nuevo índice del año 5 es

$$100{,}0\left(\frac{243{,}0}{198{,}5}\right) = 122{,}4$$

Figura 19.6. Gráfico temporal del índice de precios agregado de Laspeyres con los años 1-6 (año base 1) y los años 6-10 (año base 6).



La Figura 19.7 representa el índice enlazado que se obtiene utilizando como base el año 6. Este gráfico es una representación más clara de la pauta de variación de los precios en el periodo de 10 años.

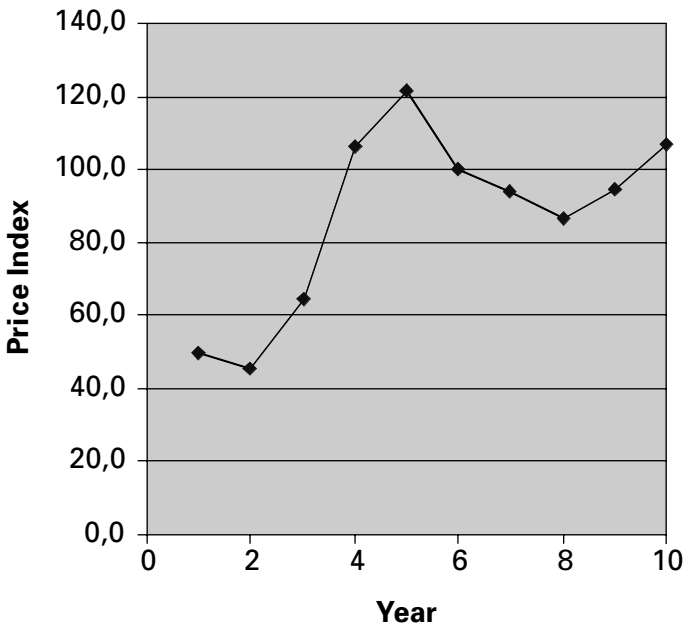


Figura 19.7. Índice de precios agregado de Laspeyres enlazado del trigo, el maíz y la soja (año 6 = 100).

EJERCICIOS

Ejercicios básicos

19.1. Suponga que está analizando un mercado y encuentra un índice de precios de Laspeyres que se calculó utilizando el año 2000 como periodo base. Interprete los resultados suponiendo que el índice de 2003 es:

a) 134,5
b) 97,4
c) 101,7

19.2. Vuelva a la Figura 19.4. Calcule el índice de cantidades de Laspeyres revisado de los años 1 a 6 suponiendo que los precios del año 1 son 1,45 (trigo), 1,21 (maíz) y 2,98 (soja).

19.3. Las universidades tienen muchos costes, entre los cuales se encuentran los costes de la energía, los libros, el laboratorio y demás equipo, el material de oficina y la mano de obra. Suponga que le piden que muestre cómo han variado los niveles de precios a los que se enfrenta su universidad en los 10 últimos años. ¿Qué dificultades esperaría encontrarse y cómo intentaría resolverlas?

Ejercicios aplicados

Nota: los ejercicios 19.4 a 19.7 deben realizarse mediante el programa Excel.

19.4. La tabla adjunta muestra el precio por acción del Banco de Nueva York, Inc., de 12 semanas.

Semana	Precio	Semana	Precio	Semana	Precio
1	35	5	35	9	34 6/8
2	35 7/8	6	34 7/8	10	35 2/8
3	34 6/8	7	35	11	38 6/8
4	34 3/8	8	34 6/8	12	37 1/8

a) Calcule un índice de precios utilizando la semana 1 como periodo base.
b) Calcule un índice de precios utilizando la semana 4 como periodo base.

19.5. Un restaurante ofrece tres platos especiales: bistec, pescado y pollo. La tabla adjunta muestra sus precios medios (en dólares) en los 12 meses del año pasado.

Mes	Bistec	Pescado	Pollo
Enero	7,12	6,45	5,39
Febrero	7,41	6,40	5,21
Marzo	7,45	6,25	5,25
Abril	7,70	6,60	5,40
Mayo	7,72	6,70	5,45
Junio	7,75	6,85	5,60
Julio	8,10	6,90	5,54
Agosto	8,15	6,84	5,70
Septiembre	8,20	6,96	5,72
Octubre	8,30	7,10	5,69
Noviembre	8,45	7,10	5,85
Diciembre	8,65	7,14	6,21

La tabla adjunta muestra el número mensual de pedidos de estos platos especiales. Tome enero como base.

Mes	Bistec	Pescado	Pollo
Enero	123	169	243
Febrero	110	160	251
Marzo	115	181	265
Abril	101	152	231
Mayo	118	140	263
Junio	100	128	237
Julio	92	129	221
Agosto	87	130	204
Septiembre	123	164	293
Octubre	131	169	301
Noviembre	136	176	327
Diciembre	149	193	351

a) Halle el índice de precios agregado no ponderado.
b) Halle el índice de precios de Laspeyres.
c) Halle el índice de cantidades de Laspeyres.

19.6. La tabla adjunta muestra los salarios por hora de tres tipos de empleados de una pequeña empresa en 6 años.

Año	Obreros	Administrativos	Supervisores
1	10,60	8,40	16,40
2	11,10	8,70	19,50
3	11,80	9,10	19,90
4	11,90	9,20	18,80
5	12,30	9,60	19,00
6	12,50	9,70	19,30

Tome el año 1 como base. Ese año había 72 obreros, 23 administrativos y 10 supervisores.

a) Halle el índice de salarios por hora no ponderado.
b) Halle el índice de salarios por hora de Laspeyres.

19.7. La tabla adjunta muestra un índice de precios de un grupo de mercancías en 6 años. Calcule un índice enlazado utilizando el año 4 como base.

Año	1	2	3	4	5	6
Año base 1	100	108,4	114,3	120,2		
Año base 2				100	103,5	107,8

19.8. Explique por qué es útil desarrollar un índice de precios de un grupo de productos, por ejemplo, un índice de precios de la energía. ¿Cuáles son las ventajas de un índice de precios *ponderado*?

19.2. Un contraste no paramétrico de aleatoriedad

Para analizar datos de series temporales, hay que realizar en primer lugar un contraste de aleatoriedad de las series temporales. Presentamos el *contraste de rachas*, que es un contraste no paramétrico especialmente fácil de realizar.

Para mostrar el contraste, examinaremos primero una serie de 16 observaciones diarias sobre un índice del volumen de acciones negociadas en la bolsa. Los datos se muestran en la Tabla 19.3 y se representan en la Figura 19.8. En esta figura, se ha trazado una línea en la mediana. La mediana de un número par de observaciones es la media del par central cuando las observaciones se ordenan en sentido ascendente. En este caso, es

$$\text{Mediana} = \frac{107 + 108}{2} = 107{,}5$$

Si esta serie fuera aleatoria, el volumen negociado en un día sería independiente del volumen negociado en cualquier otro día. En concreto, un día de un elevado volumen de

Tabla 19.3. Índice del volumen de acciones negociado.

Día	Volumen	Día	Volumen	Día	Volumen	Día	Volumen
1	98	5	113	9	114	13	109
2	93	6	111	10	107	14	108
3	82	7	104	11	111	15	128
4	103	8	103	12	109	16	92

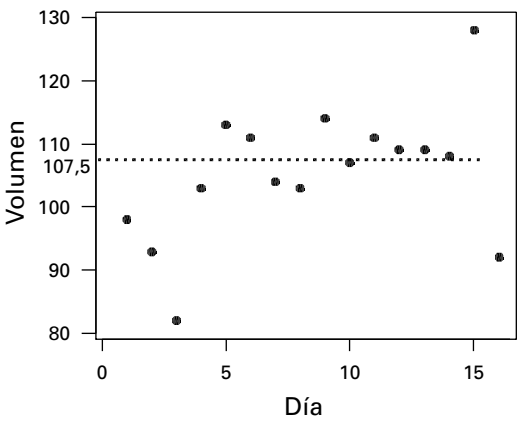


Figura 19.8. Índice del volumen de acciones negociado según el día.

contrataciones no tendría más probabilidades que cualquier otro día de ir seguido de otro día de un elevado volumen de contrataciones. El contraste de rachas que presentamos aquí divide las observaciones en un subgrupo situado por encima de la mediana y un subgrupo situado por debajo de la mediana, como muestra la Figura 19.8; la mediana es 107,5. Si + representa las observaciones situadas por encima de la mediana y − las observaciones situadas por debajo de la mediana, observamos la siguiente pauta a lo largo de los días consecutivos:

$$- - - - + + - - + - + + + + + -$$

Esta secuencia está formada por una racha de cuatro «−», seguida de una racha de dos «+», una racha de dos «−», una racha de un «+», una racha de un «−», una racha de cinco «+» y, finalmente, una racha de un «−». En total, hay, pues, $R = 7$ rachas.

Si, como cabría sospechar aquí, existe una relación positiva entre las observaciones contiguas en el tiempo, sería de esperar que hubiera relativamente pocas rachas. En nuestro ejemplo, nos preguntamos qué probabilidad hay de observar siete rachas o menos si la serie es realmente aleatoria. Para eso es necesario saber cuál es la distribución del número de rachas cuando la hipótesis nula de la aleatoriedad es verdadera. La Tabla 14 del apéndice muestra los valores tabulados de la distribución acumulada. En esa tabla vemos que, cuando $n = 16$ observaciones, la probabilidad según la hipótesis nula de encontrar 7 rachas o menos es 0,214. Por lo tanto, la hipótesis nula de la aleatoriedad sólo puede rechazarse frente a la alternativa de una relación positiva entre las observaciones contiguas al nivel de significación del 21,4 por ciento. Éste no es suficientemente pequeño para que sea razonable rechazar la hipótesis nula ni suficientemente grande para apoyar firmemente la hipótesis nula. No hemos encontrado simplemente pruebas contundentes para rechazarla. Los contrastes de aleatoriedad basados en muestras pequeñas como ésta tienen poca potencia.

El contraste de rachas

Supongamos que tenemos una serie temporal de *n* observaciones. Representemos las observaciones situadas por encima de la media con el signo «+» y las observaciones situadas por debajo de la media con el signo «−». Utilicemos estos signos para definir la secuencia de observaciones de la serie. Sea *R* el número de rachas que hay en la secuencia. La hipótesis nula es que la serie es un conjunto de variables aleatorias. La Tabla 14 del apéndice indica el nivel de significación más bajo al que puede rechazarse esta hipótesis nula frente a la alternativa de una relación positiva entre las observaciones contiguas, como una función de *R* y *n*.

Si la alternativa es una hipótesis bilateral sobre la ausencia de aleatoriedad, el nivel de significación debe duplicarse si es de menos de 0,5. Si el nivel de significación α de la tabla es superior a 0,5, el nivel de significación adecuado para el contraste frente a la alternativa bilateral es $2(1 - \alpha)$.

En el caso de las series temporales en las que $n > 20$, la distribución normal es una buena aproximación de la distribución del número de rachas según la hipótesis nula. Puede demostrarse que según la hipótesis nula

$$Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}}$$

sigue una distribución normal estándar. Este resultado es un contraste de aleatoriedad.

El contraste de rachas: grandes muestras

Dado que tenemos una serie temporal de *n* observaciones y $n > 20$, el número de rachas, *R*, es el número de secuencias que se encuentran por encima o por debajo de la mediana. Queremos contrastar la hipótesis nula

$$H_0: \text{la serie es aleatoria}$$

Los siguientes contrastes tienen un nivel de significación α.

1. Si la hipótesis alternativa es una relación positiva entre las observaciones contiguas, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} < -z_\alpha \qquad \textbf{(19.1)}$$

2. Si la hipótesis alternativa es una hipótesis bilateral de ausencia de aleatoriedad, la regla de decisión es

$$\text{Rechazar } H_0 \text{ si } \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} < -z_{\alpha/2} \quad \text{o} \quad \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} > z_{\alpha/2} \qquad \textbf{(19.2)}$$

EJEMPLO 19.1. Análisis de los datos sobre las ventas (contraste de rachas)

Le han pedido que averigüe si los 30 años de ventas anuales siguen una pauta aleatoria de una observación a la siguiente en una serie temporal.

Pinkham Sales Data

Solución

Los datos para realizar este estudio se encuentran en un fichero de datos llamado **Pinkham Sales Data** y en el disco de datos. La Figura 19.9 es un gráfico de series temporales de los datos en el que se ha trazado la mediana. El examen de este gráfico sugiere que las observaciones no son independientes, ya que parece que siguen una pauta. Los estadísticos del contraste de rachas pueden calcularse utilizando el programa Minitab u otro paquete estadístico. Realizando un análisis por computador u observando la Figura 19.9, vemos que la serie tiene ocho rachas y que la hipótesis nula de una serie temporal aleatoria se rechaza con un p-valor = 0,0030.

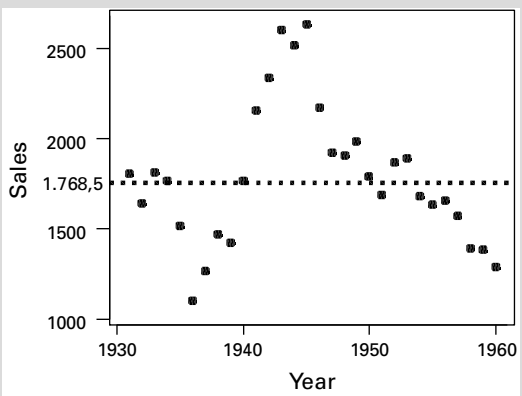


Figura 19.9. Datos sobre las ventas de Lydia Pinkham a lo largo del tiempo.

También podríamos utilizar el número de rachas y el estadístico del contraste para calcular el valor de Z del contraste:

$$Z = \frac{R - \frac{n}{2} - 1}{\sqrt{\frac{n^2 - 2n}{4(n-1)}}} = \frac{8 - 15 - 1}{\sqrt{\frac{900 - 60}{116}}} = -2{,}97$$

y en la Tabla 1 del apéndice vemos que el p-valor resultante de un contraste de dos colas es 0,0030. Vemos, pues, que las pruebas a favor de la hipótesis de que la serie no es aleatoria son abrumadoras.

EJERCICIOS

Ejercicios básicos

19.9. Una serie temporal contiene 18 observaciones. ¿Cuál es la probabilidad de que el número de rachas sea
a) inferior a 5?
b) superior a 11?
c) inferior a a 8?

19.10. Una serie temporal contiene 50 observaciones. ¿Cuál es la probabilidad de que el número de rachas sea
a) inferior a 14?
b) inferior a 17?
c) superior a 38?

19.11. Una serie temporal contiene 100 observaciones. ¿Cuál es la probabilidad de que el número de rachas sea
a) inferior a 25?
b) inferior a 41?
c) superior a 90?

Ejercicios aplicados

19.12. El fichero de datos **Exchange Rate** muestra un índice del valor del dólar estadounidense frente a las monedas de sus socios comerciales durante 12 meses consecutivos. Utilice el contraste de rachas para hacer un contraste de aleatoriedad de esta serie.

19.13. El fichero de datos **Inventory Sales** muestra el cociente entre las existencias y las ventas de la industria y el comercio de Estados Unidos en un periodo de 12 años. Realice un contraste de aleatoriedad de esta serie utilizando el contraste de rachas.

19.14. El fichero de datos **Stock Market Index** muestra los rendimientos anuales de un índice bursátil durante 14 años. Realice un contraste de aleatoriedad utilizando el contraste de rachas.

19.15. El fichero de datos **Gold Price** muestra el precio del oro (en dólares) vigente a finales de año de 14 años consecutivos. Utilice el contraste de rachas para realizar un contraste de aleatoriedad de esta serie.

19.3. Componentes de una serie temporal

En los apartados 19.3 a 19.5 presentamos algunos métodos descriptivos para analizar datos de series temporales. La serie de interés se representa por medio de X_1, X_2, ..., X_n y en el periodo t el valor de la serie es X_t.

Un modelo convencional de la conducta de las series temporales identifica varios componentes de la serie. Tradicionalmente, en la mayoría de las series temporales se representan cuatro componentes al menos en parte:

1. El componente tendencial
2. El componente estacional
3. El componente cíclico
4. El componente irregular

Macro2000

Muchas series temporales muestran una tendencia a aumentar o a disminuir a un ritmo bastante continuo durante largos periodos de tiempo, lo que indica la existencia de un componente tendencial. Por ejemplo, los indicadores de la riqueza nacional, como el producto interior bruto, normalmente crecen con el paso del tiempo. Las tendencias a menudo se mantienen y, en ese caso, este componente es importante para hacer predicciones. La Figura 19.10 muestra la serie temporal del producto interior bruto trimestral de más de 50 años procedente del fichero de datos **Macro2000** que se encuentra en el disco de datos. Esta pauta muestra claramente una fuerte tendencia ascendente que es mayor en unos periodos que en otros. Este gráfico temporal revela un notable componente tenden-

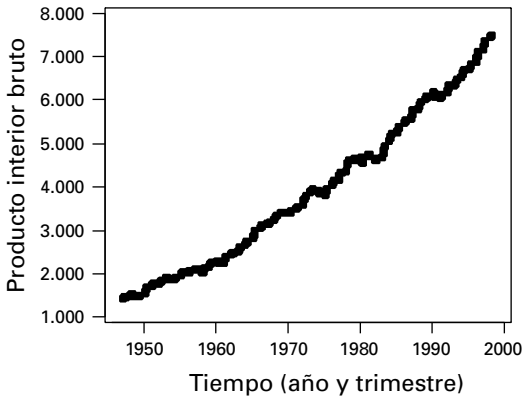


Figura 19.10. Evolución del producto interior bruto a lo largo del tiempo que indica la existencia de una tendencia.

cial que es importante para el análisis inicial y que normalmente va seguido de análisis más sofisticados, como mostramos en futuros apartados.

Otro importante componente es la pauta estacional. La Figura 19.11 muestra los beneficios trimestrales por acción de una empresa. Los beneficios del cuarto trimestre son considerablemente más altos y los del segundo trimestre son algo más altos que los de los demás periodos. Obsérvese que esta pauta continúa repitiéndose en el ciclo de cuatro trimestres que representa cada año. Además del componente estacional, también hay una notable tendencia ascendente en los beneficios por acción. Nuestro tratamiento de la estacionalidad depende de nuestros objetivos. Por ejemplo, si es importante predecir cada trimestre de la forma más precisa posible, incluimos un componente de estacionalidad en nuestro modelo. En el apartado 14.2, por ejemplo, mostramos que pueden utilizarse variables ficticias para estimar un componente de estacionalidad en una serie temporal. Por lo tanto, si prevemos que la pauta de estacionalidad continuará, debemos incluir la estimación del componente de estacionalidad en nuestro modelo de predicción.

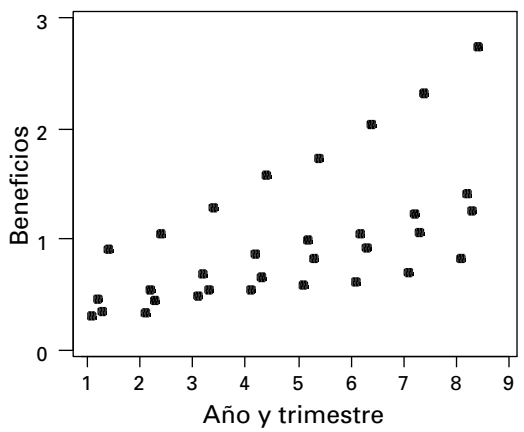


Figura 19.11. Beneficios trimestrales por acción de una empresa que indican la existencia de un componente estacional.

Para algunos otros fines, la estacionalidad puede ser una molestia. En muchas aplicaciones, el analista requiere una valoración de las variaciones globales de una serie temporal, que no esté contaminada por la influencia de factores estacionales. Supongamos, por ejemplo, que acabamos de recibir las cifras más recientes de los beneficios del cuarto trimestre de la empresa de la Figura 19.11. Ya sabemos que éstas serán probablemente mucho más altas que las del trimestre anterior. Lo que nos gustaría hacer es averiguar qué

parte de este aumento de los beneficios se debe a factores puramente estacionales y qué parte representa un verdadero crecimiento subyacente. En otras palabras, nos gustaría producir una serie temporal libre de la influencia estacional. Se dice que una serie de ese tipo está desestacionalizada. En el apartado 19.5 nos extenderemos algo más sobre el ajuste estacional.

Las pautas estacionales en una serie temporal constituyen una forma de conducta oscilatoria regular. Además, muchas series temporales empresariales y económicas muestran pautas oscilatorias o cíclicas que no están relacionadas con la conducta estacional. Por ejemplo, muchas series económicas siguen pautas cíclicas ascendentes y descendentes. En la Figura 19.9 vemos una pauta cíclica en los datos sobre las ventas de Lydia Pinkham. Observamos una disminución de las ventas hasta un mínimo en 1936, seguida de un aumento hasta un máximo a mediados de los años 40 y, a partir de entonces, una disminución continua. Esta pauta es una serie temporal cíclica frecuente y podemos describir la conducta histórica por medio de los movimientos cíclicos. Sin embargo, no estamos sugiriendo que en esas pautas históricas exista suficiente regularidad para poder hacer una predicción fiable de los futuros máximos y mínimos. De hecho, los datos de los que se dispone inducen a pensar que no es así.

Hemos analizado tres fuentes de variabilidad en una serie temporal. Si pudiéramos caracterizar las series temporales principalmente por medio del componente tendencial, el estacional y el cíclico, las series variarían de una manera uniforme con el paso del tiempo y podríamos hacer predicciones utilizando estos componentes. Sin embargo, los datos efectivos no se comportan de esa forma. La serie muestra, además de los principales componentes, componentes irregulares, inducidos por multitud de factores que influyen en la conducta de cualquier serie real y que muestran pautas que parecen impredecibles basándose en la experiencia anterior. Puede considerarse que estas pautas son similares al término de error aleatorio de un modelo de regresión. En todos los ejemplos de componentes que hemos representado hasta ahora, podemos ver claramente el componente irregular añadido a los componentes estructurales.

Análisis de los componentes de las series temporales

Una serie temporal puede describirse mediante modelos basados en los siguientes componentes:

- T_t Componente tendencial
- S_t Componente estacional
- C_t Componente cíclico
- I_t Componente irregular

Utilizando estos componentes, podemos decir que una serie temporal es la suma de sus componentes:

$$X_t = T_t + S_t + C_t + I_t$$

En otras circunstancias, también podríamos decir que una serie temporal es el producto de sus componentes, representado a menudo como un modelo de suma logarítmica:

$$X_t = T_t S_t C_t I_t$$

No tenemos que limitarnos a estas dos formas estructurales. Por ejemplo, en algunos casos podríamos tener una combinación de formas aditivas y multiplicativas.

Una gran parte de los primeros análisis de series temporales trataban de aislar los componentes de una serie, lo que permitía expresar en cualquier momento del tiempo el valor de la serie en función de los componentes. Este enfoque, en el que a menudo se utilizaban medias móviles, que analizamos en los dos apartados siguientes, se ha sustituido en gran parte por enfoques más modernos. Una excepción es el problema de la desestacionalización, que requiere la extracción del componente estacional de la serie y que analizamos en el apartado 19.5.

El enfoque más moderno del análisis de series temporales implica la construcción de un modelo formal, en el que están presentes, explícita o implícitamente, varios componentes, para describir la conducta de una serie de datos. Cuando se construyen modelos, hay dos formas posibles de tratar los componentes de una serie. Una es considerarlos fijos a lo largo del tiempo, de tal manera que una tendencia podría representarse por medio de una línea recta. Este enfoque a menudo es útil para analizar datos físicos, pero dista de ser adecuado en las aplicaciones empresariales y económicas, en las que la experiencia sugiere que cualquier regularidad aparentemente fija es con demasiada frecuencia ilusoria cuando se examina detenidamente. Para ilustrarlo, supongamos que examinamos solamente los datos de Lydia Pinkham correspondientes a los años 1936-1943. Vemos en la Figura 19.9 que en este periodo parece que hay una tendencia ascendente fija y continua. Sin embargo, si esta «tendencia» se hubiera proyectado hacia delante unos cuantos años a partir de 1943, las predicciones resultantes de las futuras ventas habrían sido muy inexactas. Sólo mirando el gráfico de los años siguientes vemos lo inadecuado que habría sido un modelo de tendencia fija.

Cuando se trata de datos empresariales y económicos, es preferible tratar de otra forma los componentes regulares de una serie temporal. En lugar de considerar que son fijos permanentemente, suele ser más sensato pensar que evolucionan continuamente con el tiempo. Por lo tanto, no necesitamos estipular pautas tendenciales o estacionales fijas sino que podemos tener en cuenta la posibilidad de que estos componentes cambien con el tiempo. Examinaremos este tipo de modelos después de haber analizado las medias móviles.

EJERCICIOS

Ejercicios aplicados

19.16. El fichero de datos **Housing Starts** muestra las viviendas iniciadas por mil habitantes en Estados Unidos en un periodo de 24 años.

a) Utilice la variante del contraste de rachas con grandes muestras para realizar un contraste de aleatoriedad de esta serie.

b) Trace un gráfico temporal de esta serie y comente los componentes de la serie que revela este gráfico.

19.17. El fichero de datos **Earnings per Share** muestra los beneficios por acción obtenidos por una empresa en un periodo de 28 años.

a) Utilice la variante del contraste de rachas con grandes muestras para realizar un contraste de aleatoriedad de esta serie.

b) Trace un gráfico temporal de esta serie y comente los componentes de la serie que revela este gráfico.

19.4. Medias móviles

El componente irregular de algunas series temporales puede ser tan grande que oculte las regularidades subyacentes y dificulte la interpretación visual del gráfico temporal. En estas circunstancias, el gráfico real parecerá bastante irregular y es posible que queramos suavi-

zarlo para tener una imagen más clara. Podemos reducir este problema utilizando una media móvil.

Podemos suavizar el gráfico utilizando el método de las medias móviles, que se basa en la idea de que cualquier gran componente irregular en cualquier momento del tiempo ejercerá un efecto menor si promediamos el punto con sus vecinos inmediatos. El método más sencillo que podemos utilizar es una media móvil centrada simple de $(2m + 1)$ puntos. Es decir, sustituimos cada observación x_t por la media de sí misma y sus vecinas, de manera que

$$x_t^* = \frac{1}{2m+1} \sum_{j=-m}^{m} x_{t+j}$$
$$= \frac{x_{t-m} + x_{t-m+1} + \cdots + x_t + \cdots + x_{t+m-1} + x_{t+m}}{2m+1}$$

Por ejemplo, si fijamos m en 2, la media móvil de 5 puntos es

$$x_t^* = \frac{x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2}}{5}$$

Dado que la primera observación es x_1, la primera media móvil sería

$$x_3^* = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$

Ésta es la media de las cinco primeras observaciones. En el caso de los datos sobre las ventas de Lydia Pinkham del ejemplo 19.1, tenemos que en 1933

$$x_3^* = \frac{1.806 + 1.644 + 1.814 + 1.770 + 1.518}{5} = 1.710{,}4$$

Asimismo, x_4^* es la media de la segunda a la sexta observación, y así sucesivamente. La Tabla 19.4 muestra la serie original y la serie suavizada. Obsérvese que en el caso de las medias móviles centradas perdemos la primera y la última m observaciones. Por lo tanto, aunque la serie original va de 1931 a 1960, la serie suavizada va de 1933 a 1958.

Medias móviles centradas simples de $(2m + 1)$ puntos

Sean x_1, x_2, x_3, ..., x_n observaciones de una serie temporal de interés. Puede obtenerse una serie suavizada utilizando una media móvil centrada simple de $(2m + 1)$ puntos.

$$x_t^* = \frac{1}{2m+1} \sum_{j=-m}^{m} x_{t+j} \qquad (t = m + 1, m + 2, ..., n - m) \qquad \textbf{(19.3)}$$

Las medias móviles pueden hallarse utilizando el programa Minitab, como muestra la Figura 19.12. Vemos tanto la serie original como la serie suavizada —la serie de medias móviles de 5 puntos— representadas en relación con el tiempo. Como puede observarse, la serie de medias móviles es de hecho más suave que la serie original. Por lo tanto, la serie de medias móviles ha eliminado el componente irregular subyacente de la serie para mostrar mejor los componentes estructurales.

Tabla 19.4. Ventas anuales de Lydia Pinkham con la media móvil centrada simple de 5 puntos.

Año	Ventas	Media1	Año	Ventas	Media1
1931	1.806*		1946	2.177	2.232,4
1932	1.644*		1.947	1.920	2.125,6
1933	1.814	1.710,4	1948	1.910	1.955,6
1934	1.770	1.569,8	1949	1.984	1.858
1935	1.518	1.494,2	1950	1.787	1.847,2
1936	1.103	1.426	1951	1.689	1.844,4
1937	1.266	1.356,6	1952	1.866	1.784,4
1938	1.473	1.406,4	1953	1.896	1.753,6
1939	1.423	1.618	1954	1.684	1.747,2
1940	1.767	1.832	1955	1.633	1.687,8
1941	2.161	2.057,8	1956	1.657	1.586,6
1942	2.336	2.276,8	1957	1.569	1.527,2
1943	2.602	2.450,8	1958	1.390	1.458,4
1944	2.518	2.454	1959	1.387*	
1945	2.637	2.370,8	1960	1.289*	

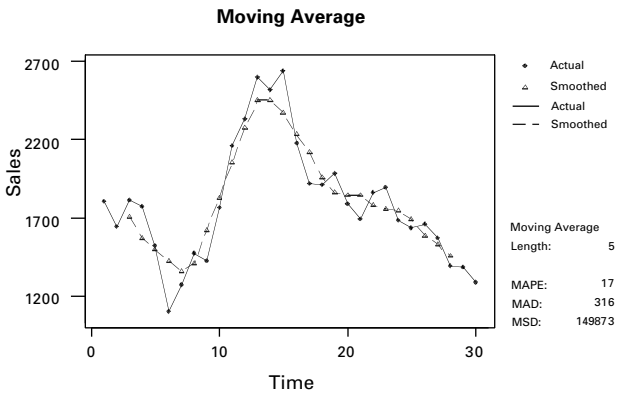


Figura 19.12. Media móvil centrada simple de 5 puntos de los datos sobre las ventas de Lydia Pinkham.

El tipo de media móvil que analizamos en este apartado no es más que uno de los muchos que podrían utilizarse. A menudo se considera deseable utilizar una media ponderada, en la que se da la mayor parte del peso a la observación central y el peso de otros valores disminuye conforme están más lejos de la observación central. Por ejemplo, podríamos utilizar una media ponderada como

$$x_t^* = \frac{x_{t-2} + 2x_{t-1} + 4x_t + 2x_{t+1} + x_{t+2}}{10}$$

En todo caso, el objetivo al utilizar medias móviles es la eliminación del componente irregular con el fin de tener una imagen más clara de las irregularidades subyacentes en una serie temporal. La técnica quizá sea más valiosa con fines descriptivos, en la elaboración de gráficos como el de la Figura 19.12.

Extracción del componente estacional por medio de medias móviles

A continuación, presentamos un método para utilizar medias móviles con el fin de extraer los componentes estacionales de las series empresariales y económicas. Los componentes estacionales pueden ser molestos y el analista puede querer eliminarlos de la serie para apreciar mejor la conducta de otros componentes. Recuérdese también que en el apartado 14.2 mostramos que pueden utilizarse variables ficticias para estimar y controlar los efectos estacionales.

Consideremos una serie temporal trimestral que tiene un componente estacional. Nuestra estrategia para eliminar la estacionalidad es calcular medias móviles de cuatro puntos para reunir los valores estacionales en una única media móvil estacional. Por ejemplo, utilizando los datos de la Tabla 19.5 sobre los beneficios por acción, el primer miembro de la serie sería

$$\frac{0{,}300 + 0{,}460 + 0{,}345 + 0{,}910}{4} = 0{,}50375$$

y el segundo miembro sería

$$\frac{0{,}460 + 0{,}345 + 0{,}910 + 0{,}330}{4} = 0{,}51125$$

La Tabla 19.5 muestra la serie completa.

Esta nueva serie de medias móviles debería estar libre de estacionalidad, pero aún hay un problema. La localización en el tiempo de los miembros de la serie de medias móviles no corresponde exactamente a la de los miembros de la serie original. El primer término es la media de las cuatro primeras observaciones y, por lo tanto, podríamos considerar que está centrado entre la segunda observación y la tercera:

$$x^*_{2,5} = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

Asimismo, el segundo término podría expresarse de la forma siguiente:

$$x^*_{3,5} = \frac{x_2 + x_3 + x_4 + x_5}{4}$$

Este problema puede superarse centrando nuestra serie de medias móviles de 4 puntos, lo cual puede hacerse calculando las medias de pares contiguos, que en el caso del primer valor es

$$x^*_3 = \frac{x^*_{2,5} + x^*_{3,5}}{2} = \frac{0{,}50375 + 0{,}51125}{2} = 0{,}5075$$

Este valor es la media móvil centrada correspondiente a la tercera observación de la serie original. El resto de la serie de medias móviles centradas está en la primera columna de la Tabla 19.5. Obsérvese de nuevo que con este método se pierden dos observaciones de cada extremo de la serie.

La Figura 19.13 representa la serie de medias móviles centradas, junto con la serie original. Es evidente que se ha eliminado el componente estacional. Además, como hemos

Tabla 19.5. Beneficios efectivos por acción de una empresa y media móvil centrada de 4 puntos.

Trimestre del año	Beneficios	Medias móviles de 4 puntos	Medias móviles centradas de 4 puntos
1,1	0,3	*	*
1,2	0,46	*	*
1,3	0,345	0,50375	0,5075
1,4	0,91	0,51125	0,5219
2,1	0,33	0,53250	0,5444
2,2	0,545	0,55625	0,5725
2,3	0,44	0,58875	0,6094
2,4	1,04	0,63000	0,6469
3,1	0,495	0,66375	0,6769
3,2	0,68	0,69000	0,7206
3,3	0,545	0,75125	0,7581
3,4	1,285	0,76500	0,7888
4,1	0,55	0,81250	0,8269
4,2	0,87	0,84125	0,8781
4,3	0,66	0,91500	0,9200
4,4	1,58	0,92500	0,9400
5,1	0,59	0,95500	0,9763
5,2	0,99	0,99750	1,0163
5,3	0,83	1,03500	1,0375
5,4	1,73	1,04000	1,0475
6,1	0,61	1,05500	1,0663
6,2	1,05	1,07750	1,1163
6,3	0,92	1,15500	1,1663
6,4	2,04	1,17750	1,2000
7,1	0,7	1,22250	1,2400
7,2	1,23	1,25750	1,2925
7,3	1,06	1,32750	1,3425
7,4	2,32	1,35750	1,3800
8,1	0,82	1,40250	1,4263
8,2	1,41	1,45000	1,5013
8,3	1,25	1,55250	*
8,4	2,73	*	*

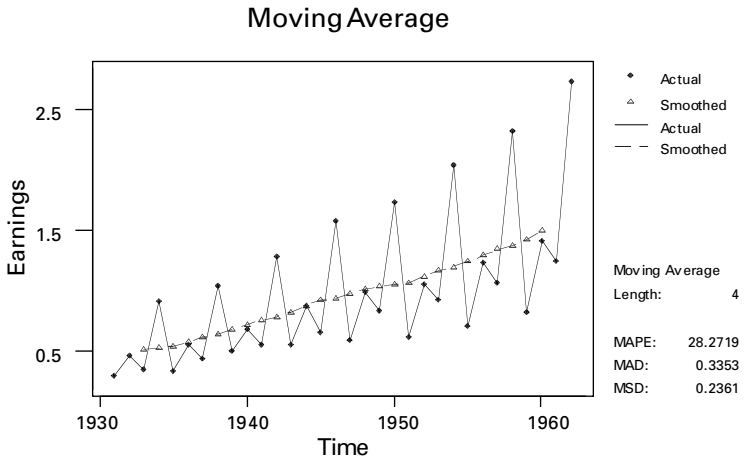


Figura 19.13. Media móvil centrada de 4 puntos y serie original de los beneficios por acción de una empresa.

utilizado medias móviles, también se ha suavizado el componente irregular. La imagen resultante nos permite, pues, juzgar las regularidades no estacionales de los datos. Vemos que en la serie suavizada domina una tendencia ascendente. Un examen más detenido muestra un crecimiento continuo de los beneficios en la primera parte de la serie, una parte central de crecimiento bastante más lento y una reanudación en la última parte del periodo de una pauta similar a la primera.

Método de desestacionalización mediante medias móviles simples

Sea x_t ($t = 1, 2, ..., n$) una serie temporal estacional del periodo s ($s = 4$ en el caso de los datos trimestrales y $s = 12$ meses en el caso de los datos mensuales). Se obtiene una serie de medias móviles centradas de s puntos, x_t^*, siguiendo estos dos pasos, en los que se supone que s es par:

1. Calcular las medias móviles de s puntos:

$$x^*_{t+0,5} = \frac{\sum_{j=-(s/2)+1}^{s/2} x_{t+j}}{s} \qquad \left(t = \frac{s}{2}, \frac{s}{2}+1, ..., n-\frac{s}{2}\right) \tag{19.4}$$

2. Calcular las medias móviles centradas de s puntos:

$$x^*_t = \frac{x^*_{t-0,5} + x^*_{t+0,5}}{2} \qquad \left(t = \frac{s}{2}+1, \frac{s}{2}+2, ..., n-\frac{s}{2}\right) \tag{19.5}$$

Hemos visto que la serie de medias móviles centradas de s puntos pueden ser útiles para comprender la estructura de una serie temporal. Como está libre en gran medida de la estacionalidad y se ha suavizado el componente irregular, es adecuada para identificar un componente tendencial o cíclico. Esta serie de medias móviles también constituye la base de muchos métodos prácticos de desestacionalización. El método específico depende de una serie de factores, entre los que se encuentran el grado de estabilidad que se supone que tiene la pauta estacional y si la estacionalidad se considera aditiva o multiplicativa. En el segundo caso, a menudo tomamos logaritmos de los datos.

A continuación, analizamos un método de desestacionalización que se basa en el supuesto implícito de que la pauta estacional es estable a lo largo del tiempo. El método se conoce con el nombre de *método del índice estacional*. Suponemos que en cualquier mes o trimestre, en cada año, el efecto de la estacionalidad es un aumento o una reducción de la serie en el mismo porcentaje.

Ilustraremos el método del índice estacional utilizando los datos sobre los beneficios de la empresa. La serie desestacionalizada se calcula en la Tabla 19.6. Las dos primeras columnas contienen la serie original y la media móvil centrada de 4 puntos. Para evaluar la influencia de la estacionalidad, expresamos la serie original en porcentaje de la serie de medias móviles centradas de 4 puntos. Así, por ejemplo, en el caso del tercer trimestre del año 1, tenemos que

$$100\left(\frac{x_3}{x_3^*}\right) = 100\left(\frac{0,345}{0,5075}\right) = 67,98$$

Estos porcentajes también se encuentran en la Tabla 19.7, en la que se muestra el cálculo del índice estacional. Para evaluar el efecto de la estacionalidad en el primer trimestre, observamos la mediana de los siete porcentajes de ese trimestre. Éste es el cuarto valor cuan-

Tabla 19.6. Ajuste estacional de los beneficios por acción de una empresa mediante el método del índice estacional.

Trimestre del año	X_t	x_t^*	$100\left(\frac{x_t}{x_t^*}\right)$	Índice estacional	Serie ajustada
1,1	0,300*			61,06	0,4913
1,2	0,460*			96,15	0,4784
1,3	0,345	0,5075	67,98	72,95	0,4729
1,4	0,910	0,5219	174,37	169,84	0,5358
2,1	0,330	0,5444	60,62	61,06	0,5405
2,2	0,545	0,5725	95,20	96,15	0,5668
2,3	0,440	0,6094	72,20	72,95	0,6032
2,4	1,040	0,6469	160,77	169,84	0,6123
3,1	0,495	0,6769	73,13	61,06	0,8107
3,2	0,680	0,7206	94,37	96,15	0,7072
3,3	0,545	0,7581	71,89	72,95	0,7471
3,4	1,285	0,7888	162,91	169,84	0,7566
4,1	0,550	0,8269	66,51	61,06	0,9008
4,2	0,870	0,8781	99,08	96,15	0,9048
4,3	0,660	0,9200	71,74	72,95	0,9047
4,4	1,580	0,9400	168,09	169,84	0,9303
5,1	0,590	0,9763	60,43	61,06	0,9663
5,2	0,990	1,0163	97,41	96,15	1,0296
5,3	0,830	1,0375	80,00	72,95	1,1378
5,4	1,730	1,0475	165,16	169,84	1,0186
6,1	0,610	1,0663	57,21	61,06	0,9990
6,2	1,050	1,1163	94,06	96,15	1,0920
6,3	0,920	1,1663	78,88	72,95	1,2611
6,4	2,040	1,2000	170,00	169,84	1,2011
7,1	0,700	1,2400	56,45	61,06	1,1464
7,2	1,230	1,2925	95,16	96,15	1,2793
7,3	1,060	1,3425	78,96	72,95	1,4531
7,4	2,320	1,3800	168,12	169,84	1,3660
8,1	0,820	1,4263	57,49	61,06	1,3429
8,2	1,410	1,5013	93,92	96,15	1,4665
8,3	1,250*			72,95	1,7135
8,4	2,730*			169,84	1,6074

Tabla 19.7. Cálculo del índice estacional de los datos sobre los beneficios por acción de la empresa.

	Trimestre				
Año	1	2	3	4	Sumas
1			67,98	174,36	
2	60,62	95,20	72,20	160,77	
3	73,13	94,37	71,89	162,91	
4	66,51	99,08	71,74	168,09	
5	60,43	97,41	80,00	165,16	
6	57,21	94,06	78,88	170,00	
7	56,45	95,16	78,96	168,12	
8	57,49	93,92			
Mediana	60,43	95,16	72,20	168,09	395,88
Índice estacional	61,06	96,15	72,95	169,84	400

do se ordenan en sentido ascendente, es decir, 60,43. También hallamos la mediana de x_t en porcentaje de x_t^* para cada uno de los demás trimestres.

Para calcular los índices estacionales, también ajustamos los índices de manera que su media sea 100. Vemos en la Tabla 19.7 que las cuatro medianas sólo suman 395,88. Podemos calcular los índices finales —que tienen una media de 100— multiplicando cada mediana por (400/395,88). En el caso del primer trimestre tenemos que

$$\text{Índice estacional} = 60{,}43\left(\frac{400}{395{,}88}\right) = 61{,}06$$

Esta cifra estima que la estacionalidad reduce los beneficios del primer trimestre a un 61,06 por ciento de los que se habrían obtenido en ausencia de factores estacionales.

Los índices estacionales de la última fila de la Tabla 19.7 se encuentran en la quinta columna de la 19.6. Obsérvese que se utiliza el mismo índice para cualquier trimestre de cada año. Por último, obtenemos nuestro valor desestacionalizado:

$$\text{Valor ajustado} = 100\left(\frac{\text{Valor original}}{\text{Índice estacional}}\right)$$

Por ejemplo, en el caso del tercer trimestre del año 1, el valor desestacionalizado es

$$100\left(\frac{0{,}345}{72{,}95}\right) = 0{,}4729$$

La serie desestacionalizada completa que se obtiene de esta forma se muestra en la última columna de la Tabla 19.6 y se representa en la Figura 19.14. Obsérvese que parece que sigue quedando una cierta estacionalidad en la última parte del periodo, lo cual induce a pensar que podría ser deseable un enfoque más elaborado, que tuviera en cuenta los cambios de las pautas estacionales.

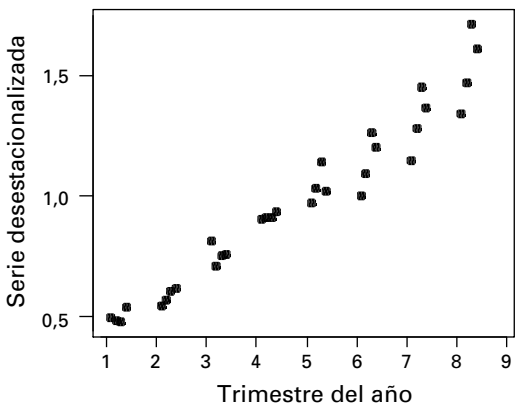


Figura 19.14. Beneficios ajustados estacionalmente por cada acción de una empresa.

El método del índice estacional aquí presentado es una sencilla solución al problema de los índices. Muchas series temporales importantes —como el producto interior bruto y sus componentes, el empleo y el desempleo, los precios y los salarios— tienen un fuerte componente estacional. Generalmente, los organismos oficiales publican datos sobre esas cantidades tanto desestacionalizados con sin desestacionalizar. Los métodos oficiales de ajuste, aunque son más complejos que el que hemos descrito aquí, normalmente se basan en me-

dias móviles. El método de desestacionalización que se utiliza más a menudo en las publicaciones oficiales de Estados Unidos es el método del Censo X-11. Se diferencia del método del índice estacional en que tiene en cuenta el posible cambio de la pauta estacional a lo largo del tiempo. Puede demostrarse que en su versión aditiva X-11 estima de una manera bastante aproximada el componente estacional de una serie temporal mensual por medio de

$$S_t = \frac{z_{t-36} + 2z_{t-24} + 3z_{t-12} + 3z_t + 3z_{t+12} + 2z_{t+24} + z_{t+36}}{15}$$

donde

$$Z_t = Z_t - X_t^*$$

siendo x_t el valor original de la serie en el periodo t y x_t^* la media móvil centrada de 12 puntos. Naturalmente, si se utiliza ese método, es necesario dar un tratamiento especial a los valores que se encuentran al final de la serie, ya que la expresión del factor estacional implica valores de la serie temporal que aún no han ocurrido. Una forma posible de lograrlo es sustituir los valores futuros desconocidos de la media móvil por predicciones basadas en los datos de los que se dispone.

EJERCICIOS

Ejercicios aplicados

19.18. El fichero de datos **Quarterly Earnings 19.18** muestra las ventas trimestrales realizadas por una empresa en un periodo de 6 años.

a) Trace un gráfico temporal de esta serie y analice sus características.

b) Utilice el método del índice estacional para desestacionalizar esta serie. Represente gráficamente la serie desestacionalizada y analice sus características.

19.19. El fichero de datos **Quarterly Sales** muestra las ventas trimestrales realizadas por una empresa en un periodo de 6 años.

a) Trace un gráfico temporal de esta serie y analice sus características.

b) Utilice el método del índice estacional para desestacionalizar esta serie. Represente gráficamente la serie desestacionalizada y analice sus características.

19.20. Calcule una serie de medias móviles centradas simples de 3 puntos de los datos sobre el precio del oro del ejercicio 19.15. Represente la serie suavizada y analice el gráfico resultante.

19.21. Calcule una serie de medias móviles centradas simples de 5 puntos de los datos sobre la construcción de viviendas del ejercicio 19.16. Trace un gráfico temporal de la serie suavizada y comente sus resultados.

19.22. Calcule una serie de medias móviles centradas simples de 7 puntos de los datos sobre los beneficios de la empresa del ejercicio 19.17. Basándose en un gráfico temporal de la serie suavizada, ¿qué puede decirse de sus componentes regulares?

19.23. Sea

$$x_t^* = \frac{1}{2m+1} \sum_{j=-m}^{m} x_{t+j}$$

una media móvil centrada simple de $(2m + 1)$ puntos. Demuestre que

$$x_{t+1}^* = x_t^* \frac{x_{t+m+1} - x_{t-m}}{2m+1}$$

¿Cómo podría utilizarse este resultado en el cálculo eficiente de la serie de medias móviles centradas?

19.24. El fichero de datos **Quarterly Earnings 19.24** muestra los beneficios por acción obtenidos por una empresa en un periodo de 7 años.

a) Trace un gráfico temporal de estos datos. ¿Sugiere su gráfico la presencia de un fuerte componente estacional en esta serie de beneficios?

b) Utilizando el método del índice estacional, obtenga una serie de beneficios desestacionalizada. Represente gráficamente esta serie y comente su conducta.

19.25. **a)** Demuestre que la serie de medias móviles centradas de s puntos del apartado 19.4 puede expresarse de la forma siguiente:

$$x_t^* = \frac{x_{t-(s/2)} + 2(x_{t-(s/2)+1} + \cdots + x_{t+(s/2)-1} - x_{t+(s/2)}}{2s}$$

b) Demuestre que

$$x_{t+1}^* = x_t^* + \frac{x_{t+(s/2)+1} + x_{t+(s/2)} - x_{t-(s/2)+1} - x_{t-(s/2)}}{2s}$$

Analice las ventajas de esta fórmula, desde el punto de vista del cálculo, para desestacionalizar series temporales mensuales.

19.26. El fichero de datos **Monthly Sales** muestra las ventas mensuales de un producto en un periodo de 3 años. Utilice el método del índice estacional para obtener una serie desestacionalizada.

19.5. Suavización exponencial

A continuación analizamos algunos métodos para utilizar los valores actuales y pasados de una serie temporal para predecir sus valores futuros. Este problema, fácil de formular, puede ser muy difícil de resolver satisfactoriamente. Generalmente, se utiliza una amplia variedad de métodos de predicción y la elección final de uno de ellos depende en gran medida del problema, de los recursos y de los objetivos del analista y de la naturaleza de los datos de los que dispone.

Nuestro objetivo es utilizar las observaciones existentes, x_1, x_2, ..., x_t, sobre una serie para predecir los valores futuros desconocidos x_{t+1}, x_{t+2}, ... La predicción tiene una importancia fundamental en el mundo de la empresa como base racional para tomar decisiones. Por ejemplo, la predicción de las ventas mensuales de un producto es la base de la política de control de las existencias. Las predicciones sobre los futuros beneficios se utilizan cuando se toman decisiones de inversión.

En este apartado, introducimos un método de predicción que se conoce con el nombre de **suavización exponencial simple** que da buen resultado en algunas aplicaciones. Constituye, además, la base de algunos métodos de predicción más complejos. La suavización exponencial es adecuada cuando la serie no es estacional y no tiene una tendencia ascendente o descendente sistemática.

En ausencia de tendencia y de estacionalidad, el objetivo es estimar el nivel actual de la serie temporal y utilizar esta estimación para predecir los futuros valores. Nuestra posición es que nos encontramos en el periodo t, estamos observando retrospectivamente la serie de observaciones x_t, x_{t-1}, x_{t-2}, ..., y queremos tener una idea del nivel actual de la serie. Para empezar, consideramos dos posibilidades extremas. En primer lugar, podríamos utilizar simplemente la observación más reciente para predecir todas las futuras observaciones. En algunos casos, como en el de los precios de los mercados especulativos, es posible que sea lo mejor que podemos hacer, pero el resultado no tiene mucho éxito. Sin embargo, en muchas series que tienen componentes irregulares, probablemente querríamos utilizar algunas observaciones anteriores de la serie. Eso identificaría las pautas que pudieran existir en la serie temporal y evitaría utilizar solamente una fluctuación aleatoria como base de nuestra predicción.

En el extremo opuesto, podríamos utilizar la media de todos los valores pasados como estimación del nivel actual. Basta una breve reflexión para pensar que a menudo eso no sería útil, ya que todos los valores pasados se tratarían por igual. Así, por ejemplo, si intentáramos predecir las futuras ventas mediante este procedimiento, daríamos la misma importancia a las ventas de hace muchos años que a las ventas recientes. Parece razonable que la experiencia más reciente influya más en nuestra predicción.

La suavización exponencial simple es una solución intermedia entre estos extremos; hace una predicción basada en una media ponderada de los valores actuales y de los pasados. Cuando se calcula esta media, se da más peso a la observación más reciente, bastante menos al valor inmediatamente anterior, menos al valor anterior, y así sucesivamente. Estimamos el nivel del periodo actual t de la siguiente manera:

$$\hat{x}_t = (1 - \alpha)x_t + \alpha(1 - \alpha)x_{t-1} + \alpha^2(1 - \alpha)x_{t-2} + \cdots$$

donde α es un número comprendido entre 0 y 1. Por ejemplo, suponiendo que $\alpha = 0{,}5$, la predicción de las futuras observaciones es

$$\hat{x}_t = 0{,}5x_t + 0{,}25x_{t-1} + 0{,}125x_{t-2} + \cdots$$

por lo que en el cálculo de las predicciones se aplica a las observaciones actuales y pasadas una media ponderada con unos pesos cada vez menores.

En este modelo, vemos que la predicción de la serie en cualquier periodo t se estima de la siguiente manera:

$$\hat{x}_t = (1 - \alpha)x_t + \alpha(1 - \alpha)x_{t-1} + \alpha^2(1 - \alpha)x_{t-2} + \cdots$$

y, asimismo, el nivel del periodo anterior $(t - 1)$ se estimaría de la forma siguiente:

$$\hat{x}_{t-1} = (1 - \alpha)x_{t-1} + \alpha(1 - \alpha)x_{t-2} + \alpha^2(1 - \alpha)x_{t-3} + \cdots$$

Multiplicando por α, tenemos que

$$\alpha\hat{x}_{t-1} = \alpha(1 - \alpha)x_{t-1} + \alpha^2(1 - \alpha)x_{t-2} + \alpha^3(1 - \alpha)x_{t-3} + \cdots$$

Por lo tanto, restando estas dos ecuaciones, tenemos que

$$\hat{x}_t - \alpha\hat{x}_{t-1} = (1 - \alpha)x_t$$

Y mediante una sencilla manipulación, tenemos la ecuación para calcular la predicción basada en la suavización exponencial simple:

$$\hat{x}_t = \alpha\hat{x}_{t-1} + (1 - \alpha)x_t \qquad \text{para } 0 < \alpha < 1$$

Esta expresión es un útil algoritmo recursivo para calcular predicciones. El valor predicho, $\hat{x}_t$, del periodo t es una media ponderada de la predicción del periodo anterior $\hat{x}_{t-1}$ y la última observación x_t. Las ponderaciones dadas a cada uno dependen de la elección de α, que es la constante de suavización. Obsérvese que un elevado valor de α da más peso a $\hat{x}_{t-1}$, que se basa en la historia pasada de la serie, y un peso menor a x_t, que representa los datos más recientes.

Podemos ilustrar el método utilizando los datos sobre las ventas de Lydia Pinkham suponiendo que el valor de $\alpha = 0{,}4$. El proceso comienza fijando el primer elemento de la serie

$$\hat{x}_1 = x_1 = 1.806$$

El segundo valor de la predicción sería

$$\begin{aligned}\hat{x}_2 &= 0{,}4\hat{x}_1 + 0{,}6x_2 \\ &= (0{,}4)(1.806) + (0{,}6)(1.644) = 1.708{,}8\end{aligned}$$

Y este proceso continúa con toda la serie de manera que

$$\begin{aligned}\hat{x}_3 &= 0{,}4\hat{x}_2 + 0{,}6x_3 \\ &= (0{,}4)(1.708{,}8) + (0{,}6)(1.814) = 1.771{,}9\end{aligned}$$

Predicción por medio de una suavización exponencial simple

Sea x_1, x_2, ..., x_n un conjunto de observaciones de una serie temporal no estacional sin ninguna tendencia ascendente o descendente sistemática. El **método de suavización exponencial simple** para hacer predicciones es el siguiente:

1. Se obtiene la serie suavizada $\hat{x}_t$:

$$\hat{x}_1 = x_1$$

$$\hat{x}_t = \alpha\hat{x}_{t-1} + (1 - \alpha)x_t \qquad (0 < \alpha < 1;\ t = 2, 3, ..., n) \qquad \textbf{(19.6)}$$

   donde α es una constante de suavización cuyo valor se fija entre 0 y 1.
2. A partir del periodo n, se obtienen predicciones de los futuros valores, x_{n+h}, de la serie de la siguiente manera:

$$\hat{x}_{n+h} = \hat{x}_n \qquad (h = 1, 2, 3, ...)$$

Hasta ahora apenas nos hemos referido a la elección de la constante de suavización, α, en las aplicaciones prácticas. En las aplicaciones, esta elección puede basarse en razones subjetivas u objetivas. Una posibilidad es basarse en la experiencia o en el criterio personal. Por ejemplo, un analista que quiera predecir la demanda de un producto puede haber trabajado muchas veces con datos sobre líneas de producto similares y puede basarse en esa experiencia para seleccionar el valor de α. La inspección visual de un gráfico de los datos de los que se dispone también puede ser útil para elegir el valor de la constante de suavización. Si la serie parece que contiene un componente irregular considerable, no queremos dar demasiado peso únicamente a la observación más reciente, ya que podría no indicar qué esperamos en el futuro. Eso sugiere que debemos elegir un valor relativamente alto para la constante de suavización. Pero si la serie es bastante suave, daríamos un valor más bajo a α para dar más peso a la observación más reciente.

Un enfoque más objetivo es probar con diferentes valores y ver cuál ha conseguido predecir mejor los movimientos históricos de la serie temporal. Por ejemplo, podríamos calcular la serie suavizada con los valores de α de 0,2, 0,4, 0,6 y 0,8 y elegir el valor que predice mejor la serie histórica. Calcularíamos el error de cada predicción:

$$e_t = x_t - \hat{x}_{t-1}$$

Una posibilidad es calcular, para cada valor de α utilizado, la suma de los cuadrados de los errores:

$$SC = \sum_{t=2}^{n} e_t^2 = \sum_{t=2}^{n} (x_t - \hat{x}_{t-1})^2$$

El valor de α que minimiza la suma de los cuadrados de los errores es el que se utilizará para hacer futuras predicciones. La suavización exponencial simple puede realizarse utilizando el programa Minitab. La Figura 19.15 muestra un gráfico de la serie original y de la serie suavizada utilizando un valor de $\alpha = 0{,}1$, que se ha elegido probando diferentes valores y hallando el que producía un ajuste satisfactorio. El indicador MSD de la Figura 19.15 es la suma de los cuadrados de los errores dividida por el número de observaciones.

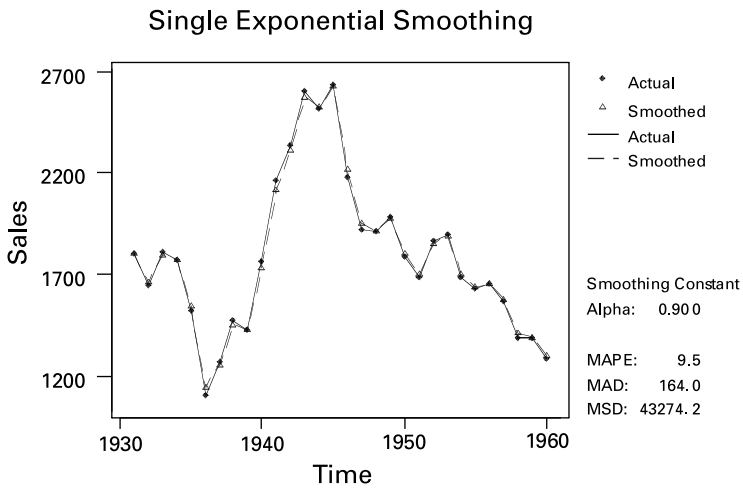


Figura 19.15. Datos sobre las ventas de Lydia Pinkham: valores originales y valores suavizados siguiendo el método exponencial simple.

Cualquiera que sea el valor de la constante de suavización que se utilice, la ecuación 19.6 puede considerarse un mecanismo de actualización. En el periodo $(t - 1)$, el nivel de la serie se estima por medio de $\hat{x}_{t-1}$. En el siguiente periodo, se utiliza la nueva observación x_t para actualizar esta estimación, por lo que la nueva estimación del nivel es una media ponderada de la estimación anterior y la nueva observación.

Modelo de predicción por medio de la suavización exponencial con el método Holt-Winters

Muchos métodos de predicción que se utilizan en el mundo de la empresa se basan en extensiones de la suavización exponencial simple. La suavización exponencial por medio del método de Holt-Winters tiene en cuenta la tendencia y posiblemente también la estacionalidad de una serie temporal.

Consideremos, en primer lugar, una serie temporal no estacional. Queremos estimar no sólo el nivel actual de la serie sino también la tendencia, que es la diferencia entre el nivel actual y el nivel anterior.

Representamos el valor observado por medio de x_t y la estimación del nivel por medio de $\hat{x}_t$. La estimación de la tendencia se representa por medio de T_t. El principio en el que se basa la estimación de estas dos cantidades es igual que el del algoritmo de la suavización exponencial simple. Las dos ecuaciones de estimación son

$$\hat{x}_t = \alpha(\hat{x}_{t-1} + T_{t-1}) + (1 - \alpha)x_t \qquad (0 < \alpha < 1)$$
$$T_t = \beta T_{t-1} + (1 - \beta)(\hat{x}_t - \hat{x}_{t-1}) \qquad (0 < \beta < 1)$$

donde α y β son constantes de suavización cuyos valores se fijan entre 0 y 1.

El método de Holt-Winters, comparable a la suavización exponencial simple, utiliza estas ecuaciones para actualizar las estimaciones anteriores utilizando una nueva observación. La estimación del nivel, $\hat{x}_{t-1}$, realizada en el periodo $(t - 1)$, tomada junto con la estimación de la tendencia, T_{t-1}, sugiere un nivel $(\hat{x}_{t-1} + T_{t-1})$ en el periodo t. Esta estimación se modifica, a la luz de la nueva observación, x_t, para obtener una estimación actualizada del nivel, $\hat{x}_t$, utilizando la ecuación dada.

Asimismo, se estima la tendencia en el periodo $(t - 1)$ como T_{t-1}. Sin embargo, una vez que se dispone de la nueva observación, x_t, la estimación de la tendencia es la diferen-

cia entre las dos estimaciones más recientes del nivel. La tendencia estimada en el periodo t es, pues, la media ponderada indicada.

Comenzamos los cálculos estableciendo que

$$T_2 = x_2 - x_1 \quad \text{y} \quad \hat{x}_2 = x_2$$

A continuación, aplicamos las ecuaciones anteriores, para $t = 3, 4, ..., n$. Mostramos estos cálculos en el ejemplo 19.2. A continuación, resumimos todo el procedimiento.

Predicción con el método de Holt-Winters: series no estacionales

Sea $x_1, x_2, ..., x_n$ un conjunto de observaciones sobre una serie temporal no estacional. El **método de Holt-Winters** para realizar predicciones consiste en lo siguiente.

1. Se obtienen estimaciones del nivel $\hat{x}_t$ y de la tendencia T_t de la forma siguiente:

$$\begin{aligned} &\hat{x}_t = x_2 \qquad T_2 = x_2 - x_1 \\ &\hat{x}_t = \alpha(\hat{x}_{t-1} + T_{t-1}) + (1 - \alpha)x_t \qquad (0 < \alpha < 1;\ t = 3, 4, ..., n) \\ &T_t = \beta T_{t-1} + (1 - \beta)(\hat{x}_t - \hat{x}_{t-1}) \qquad (0 < \beta < 1;\ t = 3, 4, ..., n) \end{aligned} \tag{19.7}$$

   donde α y β son constantes de suavización cuyos valores se fijan entre 0 y 1.

2. A partir del periodo n, se obtienen predicciones de los futuros valores, x_{n+h}, de la serie por medio de

$$\hat{x}_{n+h} = \hat{x}_n + hT_n \tag{19.8}$$

   donde h es el número de periodos futuros.

EJEMPLO 19.2. Predicción del crédito al consumo (suavización exponencial con el método Holt-Winters)

Se le ha pedido que haga una predicción del crédito al consumo pendiente utilizando el método de suavización exponencial de Holt-Winters.

Solución

Los cálculos siguientes se basan en los datos sobre el crédito al consumo de la Tabla 19.8, que también contiene los cálculos del método de Holt-Winters.

Las estimaciones iniciales del nivel y de la tendencia del año 2 son

$$\hat{x}_2 = x_2 = 155$$

y

$$T_2 = x_2 - x_1 = 155 - 133 = 22$$

Esta aplicación de la suavización utiliza los valores de $\alpha = 0{,}3$ y $\beta = 0{,}4$ y las ecuaciones

$$\begin{aligned} \hat{x}_t &= 0{,}3(\hat{x}_{t-1} + T_{t-1}) + 0{,}7x_t \\ T_t &= 0{,}4T_{t-1} + 0{,}6(\hat{x}_t - \hat{x}_{t-1}) \end{aligned}$$

Tabla 19.8. Cálculos del crédito al consumo pendiente basados en el método de Holt-Winters ($\alpha = 0{,}3$, $\beta = 0{,}4$) y realizados a partir de la salida Minitab.

t	x_t	$\hat{x}_t$	T_t
1	133		
2	155	155	22
3	165	169	17
4	171	175	11
5	194	192	14
6	231	223	25
7	274	266	36
8	312	309	40
9	313	324	25
10	333	338	18
11	343	347	13

Para $t = 3$,

$$\begin{aligned}\hat{x}_3 &= 0{,}3(\hat{x}_2 + T_2) + 0{,}7x_3 \\ &= (0{,}3)(155 + 22) + (0{,}7)(165) \\ &= 168{,}6\end{aligned}$$

y, además,

$$\begin{aligned}T_3 &= 0{,}4T_2 + 0{,}6(\hat{x}_3 - \hat{x}_2) \\ &= (0{,}4)(22) + (0{,}6)(168{,}6 - 155) \\ &= 16{,}96\end{aligned}$$

Para $t = 4$,

$$\begin{aligned}\hat{x}_4 &= 0{,}3(\hat{x}_3 + T_3) + 0{,}7x_4 \\ &= (0{,}3)(168{,}6 + 16{,}96) + (0{,}7)(171) \\ &= 175{,}4\end{aligned}$$

y, además,

$$\begin{aligned}T_4 &= 0{,}4T_3 + 0{,}6(\hat{x}_4 - \hat{x}_3) \\ &= (0{,}4)(16{,}96) + (0{,}6)(175{,}4 - 168{,}6) \\ &= 10{,}86\end{aligned}$$

Los cálculos restantes se hacen de la misma forma, fijando $t = 5, 6, ..., 11$. La Tabla 19.8 muestra los resultados de estos cálculos.

Utilicemos ahora estas estimaciones del nivel y de la tendencia para predecir las futuras observaciones. Dada una serie $x_1, x_2, ..., x_n$, las estimaciones más recientes del nivel y de la tendencia son $\hat{x}_t$ y T_n, respectivamente. En la realización de predicciones se supone que esta tendencia más reciente se prolongará a partir del nivel más reciente. Por lo tanto, hacemos una predicción utilizando la relación

$$\hat{x}_{n+1} = \hat{x}_n + T_n$$

y para el periodo siguiente

$$\hat{x}_{n+2} = \hat{x}_n + 2T_n$$

y, en general, para h periodos venideros

$$\hat{x}_{n+h} = \hat{x}_n + hT_n$$

En la Tabla 19.8 vemos que las estimaciones más recientes del nivel y de la tendencia son

$$\hat{x}_{11} = 347 \qquad T_{11} = 13$$

Las predicciones para los tres periodos siguientes son

$$\hat{x}_{12} = 347 + 13 = 360$$
$$\hat{x}_{13} = 347 + (2)(13) = 373$$
$$\hat{x}_{14} = 347 + (3)(13) = 386$$

El método de Holt-Winters puede calcularse utilizando el programa Minitab y la Figura 19.16 muestra el gráfico de series temporales y las predicciones. El método del Minitab es algo distinto del que acabamos de describir. En primer lugar, las entradas del nivel y de la tendencia son

$$\text{Nivel} = 1 - \alpha$$
$$\text{Tendencia} = 1 - \beta$$

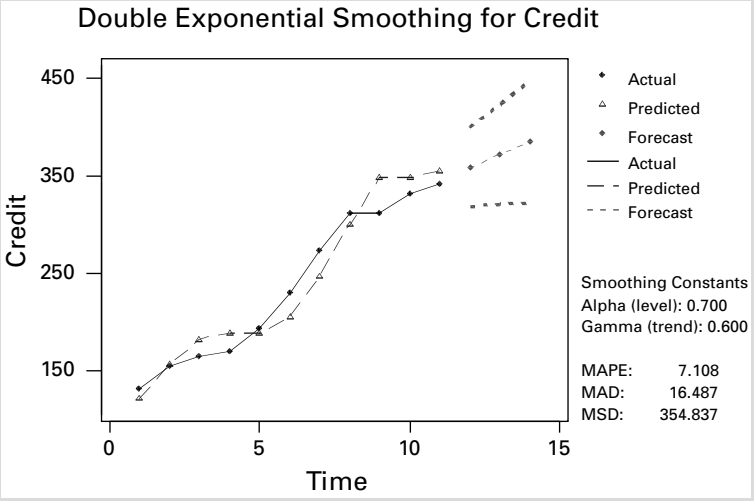


Figura 19.16. Crédito al consumo pendiente observado y predicho.

Además, el Minitab calcula una estimación para el primer periodo utilizando el siguiente método:

1. El Minitab ajusta un modelo de regresión lineal a datos de series temporales (variable y) en relación con el tiempo (variable x).
2. La constante de esta regresión es la estimación inicial del componente del nivel; el coeficiente de la pendiente es la estimación inicial del componente tendencial.

Como consecuencia, los valores calculados con el programa Minitab, que se muestran en la Tabla 19.9, son algo distintos de los que figuran en la 19.8. El método del Minitab generalmente hace predicciones algo mejores que el método más simplificado que hemos mostrado. Si el lector utiliza otros paquetes estadísticos, compruebe los algoritmos específicos utilizados para asegurarse de que comprende lo que calcula. Normalmente, puede hacerse pulsando la opción Ayuda.

Tabla 19.9. Cálculos del crédito al consumo pendiente ($\alpha = 0{,}3$, $\beta = 0{,}4$) y realizados con el programa Minitab.

Periodo	Crédito al consumo observado	Valor esperado del nivel	Tendencia	Predicciones
1	133	130	28	
2	155	156	27	
3	165	170	19	
4	171	177	12	
5	194	192	14	
6	231	224	24	
7	274	266	35	
8	312	309	40	
9	313	324	25	
10	333	338	18	
11	343	347	13	
12				360
13				373
14				385

Predicción de series temporales estacionales

A continuación, examinamos una extensión del método de Holt-Winters que tiene en cuenta la estacionalidad. En la mayoría de los problemas prácticos, el factor estacional se considera multiplicativo, por lo que, por ejemplo, cuando se analizan cifras de ventas mensuales, se puede considerar que las ventas de enero son una proporción de las ventas mensuales medias. Se supone, al igual que antes, que el componente tendencial es aditivo.

Al igual que en el caso no estacional, utilizamos los símbolos x_t, $\hat{x}_t$ y T_t para representar el valor observado y las estimaciones del nivel y de la tendencia, respectivamente, del periodo t. El factor estacional es F_t, por lo que si la serie temporal contiene s periodos al año, el factor estacional del periodo correspondiente del año anterior es F_{t-s}.

En el modelo de Holt-Winters, las estimaciones del nivel, de la tendencia y del factor estacional se actualizan por medio de las tres ecuaciones siguientes:

$$\hat{x}_t = \alpha(\hat{x}_{t-1} + T_{t-1}) + (1 - \alpha)\frac{x_t}{F_{t-s}} \qquad (0 < \alpha < 1)$$

$$T_t = \beta T_{t-1} + (1 - \beta)(\hat{x}_t - \hat{x}_{t-1}) \qquad (0 < \beta < 1)$$

$$F_t = \gamma F_{t-s} + (1 - \gamma)\frac{x_t}{\hat{x}_t} \qquad (0 < \gamma < 1)$$

donde α, β y γ son constantes de suavización cuyos valores están comprendidos entre 0 y 1.

El término $(\hat{x}_{t-1} + T_{t-1})$ es una estimación del nivel del periodo t calculada en el periodo anterior $t - 1$. Esta estimación se actualiza cuando se dispone de x_t. Pero también eliminamos la influencia de la estacionalidad deflactándola por la estimación más reciente, F_{t-s}, del factor estacional de ese periodo. La ecuación de actualización de la tendencia, T_t, es la misma que antes.

Por último, el factor estacional, F_t, se estima utilizando la tercera ecuación. La estimación más reciente del factor, que es la del año anterior, es F_{t-s}. Sin embargo, dividiendo la nueva observación, x_t, por la estimación del nivel, $\hat{x}_t$, se obtiene un factor estacional $x_t/\hat{x}_t$. La nueva estimación del factor estacional es una media ponderada de estas dos cantidades.

Predicción con el método de Holt-Winters: series estacionales

Sean x_1, x_2, ..., x_n un conjunto de observaciones sobre una serie temporal estacional del periodo s (siendo $s = 4$ en el caso de los datos trimestrales y $s = 12$ en el de los datos mensuales). El **método de Holt-Winters** para realizar predicciones utiliza un conjunto de estimaciones recursivas a partir de la serie histórica. Estas estimaciones utilizan una constante del nivel, α; una constante de la tendencia, β, y una constante estacional multiplicativa, γ. Las estimaciones recursivas se basan en las siguientes ecuaciones:

$$\begin{aligned} \hat{x}_t &= \alpha(\hat{x}_{t-1} + T_{t-1}) + (1 - \alpha)\frac{x_t}{F_{t-s}} \quad (0 < \alpha < 1) \\ T_t &= \beta T_{t-1} + (1 - \beta)(\hat{x}_t - \hat{x}_{t-1}) \quad (0 < \beta < 1) \\ F_t &= \gamma F_{t-s} + (1 - \gamma)\frac{x_t}{\hat{x}_t} \quad (0 < \gamma < 1) \end{aligned} \tag{19.9}$$

donde $\hat{x}_t$ es el nivel suavizado de la serie, T_t es la tendencia suavizada de la serie y F_t es el ajuste estacional suavizado de la serie. Los detalles del cálculo son tediosos y lo mejor es hacerlo por computador. Hemos mostrado el algoritmo que utiliza el programa Minitab, pero numerosos paquetes estadísticos de calidad emplean métodos parecidos. Estos métodos pueden diferir en la forma en que abordan la generación de constantes para los periodos iniciales de una serie temporal observada y, por lo tanto, debe consultarse la documentación del programa para averiguar cuál es exactamente el programa utilizado. Minitab utiliza un método de regresión mediante variables ficticias para obtener estimaciones de los periodos iniciales.

Una vez que el método inicial genera las constantes del nivel, la tendencia y la estacionalidad a partir de una serie histórica, podemos utilizar los resultados para predecir los futuros valores de h periodos futuros a partir de la última observación, x_n, de la serie histórica. La ecuación de predicción es

$$\hat{x}_{t+h} = (\hat{x}_t + hT_t)F_{t+h-s} \tag{19.10}$$

Observamos que el factor estacional, F, es el generado para el periodo de tiempo estacional más reciente.

El método que hemos desarrollado aquí puede aplicarse utilizando el procedimiento del Minitab llamado «Winters method». Concretamente, el método aquí descrito utiliza la opción «multiplicative». El método Winters emplea un componente del nivel, un componente tendencial y un componente estacional de cada periodo. Utiliza tres ponderaciones o parámetros de suavización para actualizar los componentes de cada periodo. Los valores iniciales del componente del nivel y del componente tendencial se obtienen a partir de una regresión lineal con respecto al tiempo. Los valores iniciales del componente estacional se obtienen a partir de una regresión mediante variables ficticias utilizando datos desestacio-

nalizados. Las ecuaciones de suavización del método de Winters para el modelo multiplicativo son las antes utilizadas.

Este método se mostrará utilizando los beneficios por acción de una empresa en el programa Minitab. En la Figura 19.17 se muestra un gráfico de los valores observados y ajustados, junto con predicciones para los cuatro periodos siguientes. Se realizan predicciones utilizando las estimaciones más recientes de la tendencia y del nivel y se ajustan para tener en cuenta el factor estacional. Dada una estación que contiene s periodos de tiempo, la predicción para un periodo en el futuro sería

$$\hat{x}_{t+1} = (\hat{x}_n + T_t)F_{t+1-s}$$

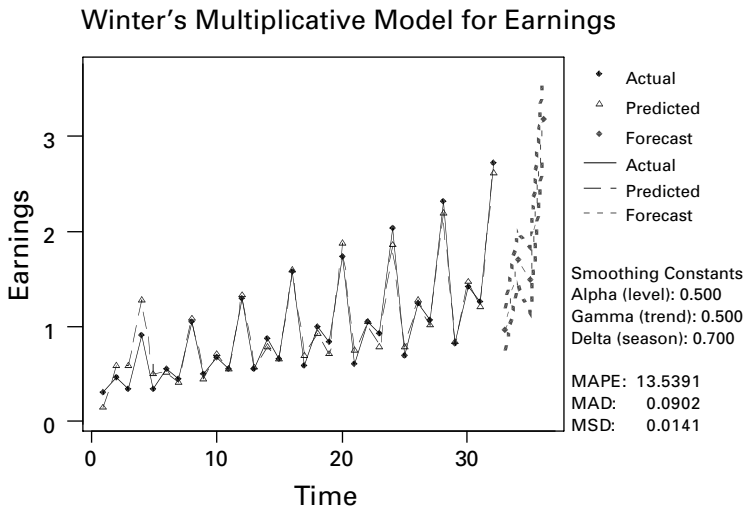


Figura 19.17. Historia y predicción de los beneficios de una empresa utilizando el método de Holt-Winters.

Los datos de nuestro ejemplo contienen 32 periodos de tiempo y un factor estacional $s = 4$, lo que indica que son datos trimestrales. Por lo tanto, para predecir la siguiente observación después del final de la serie, utilizamos la expresión

$$\hat{x}_{33} = (\hat{x}_{32} + T_{32})F_{29}$$

Esta predicción es para el primer trimestre; por lo tanto, utilizamos el factor estacional del primer trimestre más reciente y es F_{29}. En general, si estamos prediciendo h periodos en el futuro, realizamos la predicción de la siguiente manera:

$$\hat{x}_{n+h} = (\hat{x}_n + hT_n)F_{n+h-s}$$

La predicción utiliza una constante del nivel, $\alpha = 0{,}5$, una constante de la tendencia, $\beta = 0{,}5$ y una constante estacional, $\gamma = 0{,}3$.

Por último, en la Tabla 19.10 mostramos los resultados detallados del cálculo de los factores de la tendencia, del nivel y el factor estacional de cada periodo.

Las predicciones efectivas realizadas por medio del método de Holt-Winters dependen de los valores específicos elegidos para las constantes de suavización. Al igual que en nuestro análisis anterior de la suavización exponencial, esta elección podría basarse en cri-

Tabla 19.10. Resultados de la aplicación del método de suavización de Holt-Winters en Minitab.

Trimestre del año	Beneficios de la empresa	Valor suavizado	Estimación del nivel	Estimación de la tendencia	Estimación estacional	Predicción
1,1	0,300	0,043	0,387	0,242	0,713	
1,2	0,460	0,360	0,562	0,208	0,851	
1,3	0,345	0,433	0,609	0,128	0,628	
1,4	0,910	1,055	0,631	0,075	1,529	
2,1	0,330	0,450	0,584	0,014	0,609	
2,2	0,545	0,498	0,619	0,024	0,872	
2,3	0,440	0,389	0,672	0,039	0,646	
2,4	1,040	1,028	0,696	0,031	1,505	
3,1	0,495	0,424	0,770	0,053	0,633	
3,2	0,680	0,671	0,801	0,042	0,856	
3,3	0,545	0,518	0,843	0,042	0,646	
3,4	1,285	1,269	0,869	0,034	1,486	
4,1	0,550	0,550	0,886	0,025	0,624	
4,2	0,870	0,758	0,964	0,052	0,888	
4,3	0,660	0,623	1,019	0,053	0,648	
4,4	1,580	1,514	1,067	0,051	1,482	
5,1	0,590	0,666	1,032	0,008	0,588	
5,2	0,990	0,916	1,077	0,026	0,910	
5,3	0,830	0,697	1,193	0,071	0,681	
5,4	1,730	1,767	1,215	0,047	1,441	
6,1	0,610	0,714	1,150	−0,009	0,548	
6,2	1,050	1,047	1,147	−0,006	0,914	
6,3	0,920	0,782	1,246	0,046	0,721	
6,4	2,040	1,795	1,354	0,077	1,487	
7,1	0,700	0,741	1,355	0,039	0,526	
7,2	1,230	1,238	1,370	0,027	0,902	
7,3	1,060	0,988	1,433	0,045	0,734	
7,4	2,320	2,131	1,519	0,066	1,515	
8,1	0,820	0,799	1,572	0,059	0,523	
8,2	1,410	1,419	1,597	0,042	0,889	
8,3	1,250	1,172	1,671	0,058	0,744	
8,4	2,730	2,531	1,765	0,076	1,537	
9,1						0,963
9,2						1,705
9,3						1,48
9,4						3,18

terios subjetivos u objetivos. La experiencia del analista en el análisis de conjuntos de datos similares podría ayudarlo a dar valores adecuados a las constantes de suavización. También podría probar diferentes conjuntos de valores posibles con los datos históricos de que dispone y hacer las predicciones utilizando el conjunto de valores que dieran las mejores predicciones de esos datos. Esta estrategia es fácil de poner en práctica utilizando un paquete estadístico, como muestra el ejemplo que hemos presentado con el programa Minitab.

EJERCICIOS

Ejercicios aplicados

19.27. Basándose en los datos del ejercicio 19.13, utilice el método de la suavización exponencial simple para hacer predicciones del cociente entre las existencias y las ventas de los 4 próximos años. Utilice una constante de suavización de $\alpha = 0{,}4$. Represente gráficamente la serie temporal y las predicciones.

19.28. Utilice el método de la suavización exponencial simple con una constante de suavización de $\alpha = 0{,}3$ para predecir el precio que tendrá el oro en los 5 próximos años, basándose en los datos del ejercicio 19.15.

19.29. Utilizando los datos del ejercicio 19.16, utilice el método de la suavización exponencial simple con una constante de suavización $\alpha = 0{,}5$ para predecir la construcción de viviendas de los 3 próximos años.

19.30. El fichero de datos **Earnings per Share 19.30** muestra los beneficios por acción que obtendrá una empresa en un periodo de 18 años.

a) Utilizando las constantes de suavización $\alpha = 0{,}2$, 0,4, 0,6 y 0,8, realice predicciones basándose en la suavización exponencial simple.

b) ¿Cuál de las predicciones elegiría?

19.31. **a)** Si las predicciones se basan en una suavización exponencial simple y $\hat{x}_t$ representa el valor suavizado de la serie en el periodo t, demuestre que el error cometido en la predicción de x_t, realizada en el periodo $(t-1)$, puede expresarse de la forma siguiente:

$$e_t = x_t - \hat{x}_{t-1}$$

b) Por lo tanto, demuestre que podemos escribir $\hat{x}_t = x_t - \alpha e_t$, donde vemos que se utiliza la observación más reciente y el error de predicción más reciente para calcular la predicción siguiente.

19.32. Suponga que en el método de la suavización exponencial simple la constante de suavización α se fija en un valor igual a 1. ¿Qué predicciones se obtendrán?

19.33. Comente la siguiente afirmación: «Sabemos que todas las series temporales empresariales y económicas muestran variabilidad a lo largo del tiempo. Sin embargo, si se utiliza el método de la suavización exponencial simple, se obtienen las mismas predicciones de todos los futuros valores de las series temporales. Dado que sabemos que todos los futuros valores no serán iguales, eso es absurdo».

19.34. El fichero de datos **Industrial Production Canada** muestra un índice de producción industrial de Canadá correspondiente a un periodo de 15 años. Utilice el método de Holt-Winters con las constantes de suavización $\alpha = 0{,}3$ y $\beta = 0{,}5$ para hacer predicciones para los 5 próximos años.

19.35. El fichero de datos **Hourly Earnings** muestra los ingresos por hora de la industria manufacturera de Estados Unidos correspondientes a un periodo de 24 meses. Utilice el método de Holt-Winters con las constantes de suavización $\alpha = 0{,}3$ y $\beta = 0{,}4$ para hacer predicciones para los 3 próximos meses.

19.36. El fichero de datos **Food Prices** muestra un índice de los precios de los alimentos desestacionalizado de Estados Unidos correspondiente a un periodo de 14 meses. Utilice el método de Holt-Winters, con las constantes de suavización $\alpha = 0{,}5$ y $\beta = 0{,}5$, para hacer predicciones para los 3 próximos meses.

19.37. El fichero de datos **Profit Margins** muestra los márgenes porcentuales de beneficios de una empresa correspondientes a un periodo de 11 años. Realice predicciones para los 2 próximos años utilizando el método de Holt-Winters con las constantes de suavización $\alpha = 0{,}6$ y $\beta = 0{,}6$.

19.38. Utilice el método estacional de Holt-Winters para realizar predicciones de las ventas para dentro de ocho trimestres, basándose en los datos del ejercicio 19.18. Emplee las constantes de suavización $\alpha = 0{,}6$, $\beta = 0{,}5$ y $\gamma = 0{,}4$. Represente gráficamente los datos y las predicciones.

19.39. Utilice el método estacional de Holt-Winters para hacer predicciones de las ventas para dentro de ocho trimestres, basándose en los datos del ejercicio 19.19. Emplee las constantes de suavización $\alpha = 0{,}5$, $\beta = 0{,}4$ y $\gamma = 0{,}3$. Represente gráficamente los datos y las predicciones.

19.6. Modelos autorregresivos

En este apartado presentamos otro enfoque para hacer predicciones de series temporales. Este enfoque implica la utilización de los datos de los que se dispone para estimar parámetros de un modelo del proceso que podría haber generado la serie temporal. En este apartado examinamos un método muy utilizado, los *modelos autorregresivos*, que se basa en el enfoque de la construcción de modelos.

En el apartado 14.3 introdujimos el uso de variables dependientes retardadas en los modelos de regresión múltiple y ese enfoque es la base de los modelos que analizamos aquí. La idea es esencialmente considerar las series temporales como series de variables aleatorias. A efectos prácticos, a menudo podríamos estar dispuestos a suponer que estas variables aleatorias tienen todas ellas las mismas medias y las mismas varianzas. Sin embargo, no podemos suponer que son independientes entre sí. Ciertamente, si consideramos una serie de ventas de un producto, es muy probable que las ventas de periodos contiguos estén relacionadas entre sí. Las pautas de correlación como las que hay entre periodos contiguos a veces se conocen con el nombre de *autocorrelación*.

En principio, es posible cualquier número de pautas de autocorrelación. Sin embargo, unas son considerablemente más probables que otras. Se plantea una posibilidad especialmente atractiva cuando se examina una correlación bastante estrecha entre observaciones contiguas en el tiempo, una correlación menos estrecha entre observaciones separadas por dos periodos, una correlación más débil aún entre los valores separados por tres periodos, etc. Surge una sencilla pauta de autocorrelación de este tipo cuando la correlación entre valores contiguos es algún número —por ejemplo, ϕ_1— que entre valores separados por dos periodos es ϕ_1^2, que entre valores separados por tres periodos es ϕ_1^3, y así sucesivamente. Por lo tanto, si x_t representa el valor de la serie en el periodo t, tenemos en este modelo de autocorrelación que

$$\text{Corr}\,(x_t, x_{t-j}) = \phi_1^j \qquad (j = 1, 2, 3, \ldots)$$

Esta estructura de autocorrelación da lugar a un modelo de series temporales de la forma

$$x_t = \gamma + \phi_1 x_{t-1} + \varepsilon_t$$

donde γ y ϕ_1 son parámetros fijos y las variables aleatorias ε_t tienen una media de 0 y una varianza fija para todo t y no están correlacionadas entre sí. El fin del parámetro γ es tener en cuenta la posibilidad de que la serie x_t tenga alguna media distinta de 0. Por lo demás, éste es el modelo que utilizamos en el apartado 14.7 para representar la autocorrelación de los términos de error de una ecuación de regresión. Se llama *modelo autorregresivo de primer orden*.

El modelo autorregresivo de primer orden expresa el valor actual, x_t, de una serie en el valor anterior, x_{t-1}, y una variable aleatoria no autocorrelacionada, ε_t. Dado que la variable aleatoria ε_t no está autocorrelacionada, es impredecible. En el caso de las series generadas por el modelo autorregresivo de primer orden, las predicciones de los futuros valores sólo dependen del valor más reciente de la serie. Sin embargo, en muchas aplicaciones querríamos utilizar más de una observación como base para hacer predicciones. Una extensión obvia del modelo sería hacer depender el valor actual de la serie de las dos observaciones más recientes. Por lo tanto, podríamos utilizar un modelo

$$x_t = \gamma + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$$

donde γ, ϕ_1 y ϕ_2 son parámetros fijos. Este modelo se llama *modelo autorregresivo de segundo orden*.

En términos más generales, dado un entero positivo cualquiera p, el valor actual de la serie puede hacerse dependiente (linealmente) de los p valores anteriores por medio del modelo autorregresivo de orden p:

$$x_t = \gamma + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$$

donde γ, ϕ_1 y ϕ_2, ..., ϕ_p son parámetros fijos. Esta ecuación describe el modelo autorregresivo general. En el resto de este apartado, consideramos el ajuste de esos modelos y su uso para predecir los valores futuros.

Supongamos que tenemos una serie de observaciones x_1, x_2, ..., x_n. Queremos utilizarlas para estimar los parámetros desconocidos γ, ϕ_1, ϕ_2, ..., ϕ_p para los que la suma de los cuadrados de las diferencias son

$$SC = \sum_{t=p+1}^{n} (x_t - \gamma - \phi_1 x_{t-1} - \phi_2 x_{t-2} - \cdots - \phi_p x_{t-p})^2$$

sea la menor posible. Por lo tanto, la estimación puede realizarse utilizando un programa de regresión múltiple. Mostramos este método en el ejemplo 19.3 utilizando los datos sobre las ventas de Lydia Pinkham.

Modelos autorregresivos y su estimación

Sea x_t ($t = 1, 2, ..., n$) una serie temporal. Un modelo que puede utilizarse a menudo eficazmente para representar esa serie es el modelo autorregresivo de orden p:

$$x_t = \gamma + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t \qquad \textbf{(19.11)}$$

donde γ, ϕ_1, ϕ_2, ..., ϕ_p son parámetros fijos y las ε_t son variables aleatorias que tienen una media de 0 y una varianza constante y que no están correlacionadas entre sí.

Los parámetros del modelo autorregresivo se estiman por medio de un algoritmo de mínimos cuadrados, tal que los valores de γ, ϕ_1, ϕ_2, ..., ϕ_p minimizan la suma de los cuadrados siguiente:

$$SC = \sum_{t=p+1}^{n} (X_t - \gamma - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p})^2 \qquad \textbf{(19.12)}$$

Pinkham Sales Data

EJEMPLO 19.3. Predicción de los datos sobre las ventas (modelo autorregresivo)

Se le ha pedido que desarrolle un modelo autorregresivo para predecir los datos sobre las ventas de Lydia Pinkham (véase el fichero de datos **Pinkham Sales Data**).

Solución

Para utilizar un modelo autorregresivo que permita hacer predicciones de los futuros valores, es necesario fijar un valor para p, el orden de la autorregresión. Debemos elegir un valor de p lo suficientemente alto para tener en cuenta toda la conducta importante de autocorrelación de la serie. Pero tampoco queremos que p sea tan grande que incluyamos parámetros irrelevantes y que la estimación de los parámetros importantes sea

ineficiente como consecuencia. En general, se prefieren los modelos «parsimónicos» —sencillos, pero suficientes para lograr el objetivo— para hacer buenas predicciones de datos de series temporales.

Una posibilidad es fijar el valor de p arbitrariamente, quizá basándose en la experiencia anterior con conjuntos de datos similares. Otro enfoque es fijar un orden máximo, K, de la autorregresión y estimar, a su vez, modelos de orden $p = K, K-1, K-2, ...$ Se contrasta para cada valor de p la hipótesis nula de que el último parámetro de la autorregresión, ϕ_p, del modelo es 0 frente a la alternativa bilateral. El procedimiento concluye cuando hallamos un valor de p para el que esta hipótesis nula no se rechaza. Nuestro objetivo es, pues, contrastar la hipótesis nula

$$H_0: \phi_p = 0$$

frente a la alternativa

$$H_1: \phi_p \neq 0$$

En el Capítulo 12 presentamos métodos para contrastar la hipótesis nula, H_0. Sabemos básicamente que el cociente entre la estimación del coeficiente y la desviación típica del coeficiente sigue una distribución t de Student. La salida Minitab del análisis de regresión —y la salida del análisis de regresión de cualquier paquete estadístico— incluye ese cálculo de la t de Student y, además, la probabilidad de que la hipótesis nula sea verdadera —el p-valor de la hipótesis nula— dada la t de Student calculada.

Predicción a partir de modelos autorregresivos estimados

Supongamos que tenemos las observaciones x_1, x_2, ..., x_t de una serie temporal y que se ha ajustado un modelo autorregresivo de orden p a estos datos. Expresamos el modelo estimado de la siguiente manera:

$$x_t = \hat{\gamma} + \hat{\phi}_1 x_{t-1} + \hat{\phi}_2 x_{t-2} + \cdots + \hat{\phi}_p x_{t-p} + \varepsilon_t \qquad \textbf{(19.13)}$$

Partiendo del periodo n, hacemos predicciones de los futuros valores de la serie de la siguiente manera:

$$\hat{x}_{t+h} = \hat{\gamma} + \hat{\phi}_1 \hat{x}_{t+h-1} + \hat{\phi}_2 \hat{x}_{t+h-2} + \cdots + \hat{\phi}_p \hat{x}_{t+h-p} \quad (h = 1, 2, ...) \qquad \textbf{(19.14)}$$

donde para $j > 0$, $\hat{x}_{n+j}$ es la predicción de x_{t+j} partiendo del periodo n, y para $j \leqslant 0$, $\hat{x}_{t+j}$ es simplemente el valor observado de x_{t+j}.

La Figura 19.18 muestra copias abreviadas de la salida Minitab del análisis de regresión para modelos autorregresivos utilizando los datos sobre las ventas de Lydia Pinkham y suponiendo que $p = 1, 2, 3, 4$.

Aplicaremos este método a los datos sobre las ventas de Pinkham utilizando un nivel de significación del 10 por ciento para nuestros contrastes. Basándonos en los resultados de la Figura 19.18, comenzamos con la regresión suponiendo que $p = 4$. Observamos que el coeficiente de x_{t-4} tiene un estadístico t de Student de $-1{,}39$ y un p-valor de 0,180. Por lo tanto, no podemos rechazar la hipótesis nula de que el coeficiente es 0 y pasamos a la regresión suponiendo que $p = 3$. En este caso, vemos que el coeficiente de X_{t-3} tiene un

Figura 19.18. Modelos autorregresivos para los datos sobre las ventas de Lydia Pinkham (salida Minitab).

Regression with p = 1

```
Sales = 193 + 0.883 Salelag1
29 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant        193.3       189.0       1.02    0.316
Salelag1       0.8831      0.1024       8.62    0.000
```

S = 207.0 R-Sq = 73.4% R-Sq(adj) = 72.4%

Regression with p = 2

```
Sales = 314 + 1.18 Salelag1 - 0.358 Salelag2
28 cases used 2 cases contain missing values

Predictor        Coef       StDev          T        P
Constant        313.7       192.5       1.63    0.116
Salelag1       1.1801      0.1870       6.31    0.000
Salelag2      -0.3578      0.1914      -1.87    0.073
```

S = 199.6 R-Sq = 76.9% R-Sq(adj) = 75.1%

Regression with p = 3

```
Sales = 322 + 1.19 Salelag1 - 0.317 Salelag2 - 0.057 Salslag3
27 cases used 3 cases contain missing values

Predictor        Coef       StDev          T        P
Constant        322.3       215.7       1.49    0.149
Salelag1       1.1881      0.2065       5.75    0.000
Salelag2      -0.3168      0.3081      -1.03    0.315
Salslag3      -0.0574      0.2098      -0.27    0.787
```

S = 203.0 R-Sq = 78.1% R-Sq(adj) = 75.2%

Regression with p = 4

```
Sales = 446 + 1.19 Salelag1 - 0.439 Salelag2 + 0.286 Salslag3 - 0.291
Salelag4
26 cases used 4 cases contain missing values
Predictor      Coef     StDev      T      P
Constant        446.2       232.8       1.92    0.069
Salelag1       1.1937      0.2108       5.66    0.000
Salelag2      -0.4391      0.3238      -1.36    0.190
Salslag3       0.2859      0.3174       0.90    0.378
Salelag4      -0.2914      0.2101      -1.39    0.180
S = 202.6       R-Sq = 80.1%     R-Sq(adj) = 76.3%
```

estadístico t de Student igual a $-0{,}27$ y un p-valor de 0,787. Una vez más, no podemos rechazar la hipótesis nula de que este coeficiente es 0. En el caso del modelo de regresión en el que se supone que $p = 2$, vemos que el coeficiente de x_{t-2} tiene un estadístico t de Student de $-1{,}87$ y un p-valor de 0,073. Por lo tanto, podemos rechazar la hipótesis nula de que el coeficiente de x_{t-2} es 0. El modelo elegido es el modelo con dos valores retardados, $p = 2$. La ecuación final es

$$\hat{x}_t = 313{,}7 + 1{,}1801x_{t-1} - 0{,}3578x_{t-2}$$

Ahora que tenemos el modelo, queremos aplicarlo para hacer predicciones con los datos sobre las ventas de Lydia Pinkham. Comenzamos señalando que los dos últimos valores de la serie de datos son

$$x_{29} = 1.387 \quad \text{y} \quad x_{30} = 1.289$$

Ahora podemos predecir el siguiente valor x_{31}:

$$\begin{aligned}\hat{x}_{31} &= 313{,}68 + 1{,}180x_{30} - 0{,}358x_{29} \\ &= 313{,}68 + (1{,}180)(1.289) - (0{,}358)(1.387) = 1.338{,}2\end{aligned}$$

Reconocemos que el valor predicho del término de error, ε_t, es 0. Ahora podemos predecir el siguiente valor de la serie siguiendo el mismo procedimiento, con la salvedad de que ahora debemos utilizar el valor predicho de x_{31}, es decir, $\hat{x}_t$:

$$\begin{aligned}\hat{x}_{32} &= 313{,}68 + 1{,}180\hat{x}_{31} - 0{,}358x_{30} \\ &= 313{,}68 + (1{,}180)(1.338{,}2) - (0{,}358)(1.289) = 1.431{,}29\end{aligned}$$

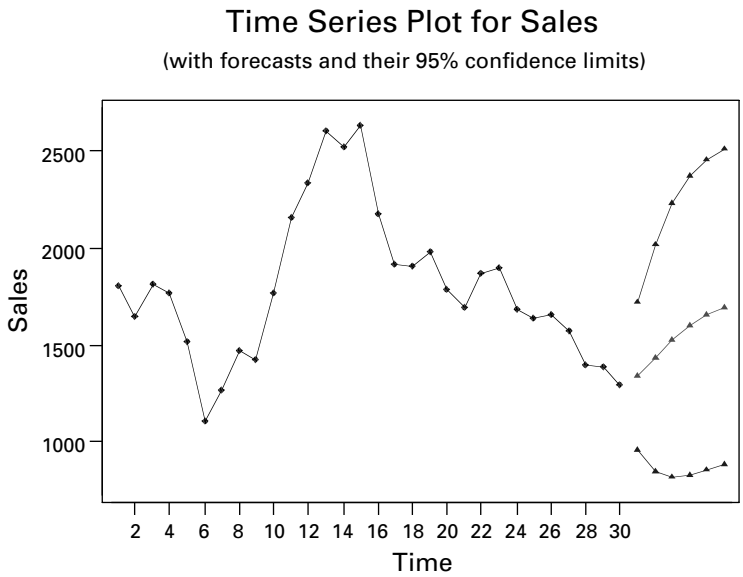
Estos cálculos pueden realizarse directamente mediante el programa Minitab —o mediante cualquier otro buen paquete estadístico— y los resultados se muestran en la Figura 19.19.

Podemos continuar con este proceso y hacer predicciones para tantos periodos futuros como queramos. La serie temporal de ventas y las predicciones para seis periodos se muestran en la Figura 19.20.

Figura 19.19. Valores predichos a partir de un modelo autorregresivo para los datos sobre las ventas de Pinkham (salida Minitab).

```
Sales = 314 + 1.18 Salelag1 - 0.358 Salelag2

28 cases used 2 cases contain missing values

Predictor        Coef       StDev          T        P
Constant        313.7       192.5       1.63    0.116
Salelag1       1.1801      0.1870       6.31    0.000
Salelag2      -0.3578      0.1914      -1.87    0.073

S = 199.6       R-Sq = 76.9%     R-Sq(adj) = 75.1%


Predicted Values

     Fit  StDev Fit         95.0% CI             95.0% PI
  1338.6       63.5   (  1207.7,  1469.4)  (   907.1,  1770.1)
```

Figura 19.20. Ventas de Lydia Pinkham y predicciones basadas en el ajuste de un modelo autorregresivo de segundo orden.



EJERCICIOS

Ejercicios aplicados

19.40. Basándose en los datos de la Tabla 19.10, estime un modelo autorregresivo de primer orden para calcular el índice del volumen de acciones negociadas. Utilice el modelo ajustado para hacer predicciones para los 4 próximos días.

19.41. El fichero de datos **Trading Volume** muestra el volumen de transacciones (en cientos de miles) de acciones de una empresa realizadas en un periodo de 12 meses. Estime con estos datos un modelo autorregresivo de primer orden y utilice el modelo ajustado para hacer predicciones del volumen para las 3 próximas semanas.

19.42. Basándose en el fichero de datos **Housing Starts** del ejercicio 19.16, estime modelos autorregresivos de órdenes 1 a 4. Utilice el método de este apartado para contrastar la hipótesis de que el orden de la autorregresión es $(p-1)$ frente a la alternativa de que es p, con un nivel de significación del 10 por ciento. Seleccione uno de estos modelos y haga predicciones de la construcción de viviendas para los 5 próximos años. Trace un gráfico temporal que muestre las observaciones originales junto con las predicciones. ¿Serían diferentes las predicciones si se utilizara un nivel de significación del 5 por ciento para los contrastes del orden autorregresivo?

19.43. Basándose en el fichero de datos **Earnings per Share** del ejercicio 19.17 sobre los beneficios por acción de una empresa, ajuste modelos autorregresivos de órdenes 1 a 4. Utilice el método de este apartado para contrastar la hipótesis de que el orden de la autorregresión es $(p-1)$ frente a la alternativa de que el verdadero orden es p, con un nivel de significación del 10 por ciento. Seleccione uno de estos modelos y haga predicciones de los beneficios por acción para los 5 próximos años. Trace un gráfico que muestre las observaciones originales junto con las predicciones. ¿Serían diferentes los resultados si se utilizara un nivel de significación del 5 por ciento para los contrastes?

19.44. Vuelva al fichero de datos **Earnings per Share 19.30** del ejercicio 19.30 sobre los beneficios por acción de una empresa. Ajuste modelos autorregresivos de órdenes 1, 2 y 3. Utilice el método del apartado 19.6 para contrastar la hipótesis de que el orden de la autorregresión es $(p-1)$ frente a la alternativa de que es p, con un nivel de significación del 10 por ciento y seleccione un valor para el orden autorregresivo. Utilice el modelo seleccionado para hacer predicciones de los beneficios por acción para dentro de 4 años. Trace un gráfico temporal de las observaciones y las predicciones. ¿Serían diferentes los resultados si se utilizara un nivel de significación del 5 por ciento para los contrastes?

19.45. En la Figura 19.18, se muestran modelos autorregresivos ajustados de órdenes 1 a 4 para datos sobre las ventas anuales. A continuación, seleccionamos un modelo contrastando la hipótesis nula de una autorregresión de orden $(p-1)$ frente a la alternativa de una autorregresión de orden p al nivel de significación del 10 por ciento. Repita este procedimiento, pero haga un contraste al nivel de significación del 5 por ciento.

a) ¿Qué modelo autorregresivo se selecciona ahora?

b) Realice predicciones de las ventas para los 3 próximos años basándose en este modelo seleccionado.

19.46. Se ha observado que las ventas anuales de un producto podrían muy bien describirse por medio de un modelo autorregresivo de tercer orden. El modelo estimado es

$$X_t = 202 + 1{,}10X_{t-1} - 0{,}48X_{t-2} + 0{,}17X_{t-3} + \varepsilon_t$$

En 1993, 1994 y 1995, las ventas fueron de 867, 923 y 951, respectivamente. Calcule las predicciones de las ventas para los años 1996 a 1998.

19.47. En el caso de muchas series temporales, especialmente en el de los precios de los mercados especulativos, se ha observado que el modelo del *paseo aleatorio* representa satisfactoriamente los datos efectivos. Este modelo es

$$x_t = x_{t-1} + \varepsilon_t$$

Demuestre que, si este modelo es adecuado, las predicciones de X_{n+h}, partiendo del periodo n, vienen dadas por

$$\hat{x}_{n+h} = x_n \qquad (h = 1, 2, 3, \ldots)$$

19.48. Vuelva al fichero de datos **Hourly Earnings** del ejercicio 19.35, que muestra los beneficios de 24 meses. Sean x_t $(t = 1, 2, \ldots, 24)$ las

observaciones. A continuación, construya la serie de primeras diferencias:

$$z_t = x_t - x_{t-1} \qquad (t = 2, 3, ..., 24)$$

Ajuste modelos autorregresivos de órdenes 1 a 4 a la serie Z_t. Utilizando el método de este apartado para contrastar la hipótesis de que el orden autorregresivo es $(p - 1)$ frente a la alternativa de orden p, con un nivel de significación del 10 por ciento, seleccione uno de estos modelos. Utilizando el modelo seleccionado, realice predicciones para Z_t, donde $t = 25$, 26 y 27. Realice predicciones de los beneficios para los 3 próximos meses.

19.7. Modelos autorregresivos integrados de medias móviles

En este apartado introducimos brevemente un método para hacer predicciones de datos de series temporales que se utiliza mucho en las aplicaciones empresariales. Los modelos que analizamos incluyen como caso especial los modelos autorregresivos que hemos estudiado en el apartado 19.6.

En un libro clásico, George Box y Gwilyn Jenkins introdujeron una metodología lo suficientemente versátil para que un usuario moderadamente hábil obtenga buenos resultados en una amplia variedad de problemas de predicción que se plantean en la práctica (véase la referencia bibliográfica 1). La esencia del método de Box-Jenkins es el examen de una amplia clase de modelos a partir de los cuales pueden realizarse predicciones, junto con una metodología para elegir, en función de las características de los datos de los que se dispone, un modelo adecuado para cualquier problema de predicción.

La clase general de modelos es la clase de modelos autorregresivos integrados de medias móviles (ARIMA). Son extensiones bastante naturales de los modelos autorregresivos del apartado 19.6. Además, la suavización exponencial simple y los predictores de Holt-Winters pueden obtenerse a partir de miembros específicos de esta clase general, al igual que otros muchos algoritmos que se utilizan frecuentemente para hacer predicciones. Los modelos y las técnicas de análisis de series temporales de Box-Jenkins pueden generalizarse para tener en cuenta la estacionalidad y también para analizar series temporales relacionadas, por lo que es posible predecir los futuros valores de una serie a partir de información no sólo sobre su propio pasado sino también sobre el pasado de otras series relevantes. Esta última posibilidad permite adoptar un enfoque para realizar predicciones que generaliza los métodos de regresión analizados en los Capítulos 12 a 14.

No es posible en el espacio de que disponemos analizar exhaustivamente la metodología de Box-Jenkins (para una introducción a esta metodología, véase la referencia bibliográfica 3). Consta, esencialmente, de tres fases:

1. Basándose en estadísticos sintéticos que son fáciles de calcular a partir de los datos de que se dispone, el analista selecciona un modelo específico de la clase general. No se trata simplemente de seguir automáticamente una serie de reglas sino que hace falta un cierto grado de criterio personal y de experiencia. Sin embargo, el analista no se compromete para siempre a seguir el modelo elegido en esta fase sino que puede abandonarlo en favor de otro en una fase posterior si parece deseable.
2. El modelo específico elegido tiene casi invariablemente algunos coeficientes desconocidos. Éstos deben estimarse a partir de los datos de los que se dispone utilizando técnicas estadísticas eficientes, como mínimos cuadrados.
3. Por último, hay que averiguar si el modelo estimado es una representación adecuada de los datos de series temporales de los que se dispone. Cualquier indicio de

que no lo es en esta fase puede sugerir alguna especificación alternativa y el proceso de selección del modelo, de estimación de los coeficientes y de comprobación del modelo se repite hasta que se encuentra uno satisfactorio.

El enfoque de Box-Jenkins para hacer predicciones tiene la gran ventaja de la flexibilidad: existe una amplia variedad de predictores y la elección entre ellos se basa en los datos. Además, cuando se ha comparado este enfoque con otros métodos, utilizando series temporales económicas y empresariales efectivas, normalmente se ha observado que funciona muy bien. Por lo tanto, puede decirse que ha superado la prueba de fuego: ¡en la práctica, funciona!

Para concluir este breve análisis, obsérvese que existen programas informáticos para realizar análisis de series temporales ajustando a los datos modelos ARIMA, incluido un conjunto de procedimientos del programa Minitab. Sin embargo, el método tiene un inconveniente en comparación con otros más sencillos analizados en apartados anteriores de este capítulo. Como hay flexibilidad para elegir un modelo adecuado de la clase general, el enfoque de Box-Jenkins es más caro que los métodos que imponen una única estructura del modelo a todas las series temporales porque debe ser utilizado por personas cualificadas.

RESUMEN

Este capítulo es una introducción al análisis de los datos de series temporales. Hemos presentado, en primer lugar, los números índice como medida estandarizada de las variaciones a lo largo del tiempo. En el resto del capítulo, hemos mostrado algunos útiles métodos para predecir datos de series temporales.

Los números índice constituyen una base coherente a lo largo del tiempo para representar precios, cantidades y otras medidas importantes. Los números índice simples son una medida del cambio con respecto a un periodo de tiempo fijo. Los números índice ponderados, como el índice de Laspeyres, parten de proporciones de bienes constantes e indican cómo influyen las variaciones de los precios de cada bien en el precio agregado de la cesta de mercado.

Hemos comenzado la predicción de datos de series temporales con un análisis de los principales componentes de las series temporales: tendencial, cíclico, estacional e irregular. A continuación, hemos presentado una serie de instrumentos aplicados que han demostrado ser eficaces para hacer predicciones. Hemos mostrado algunas versiones de los modelos de medias móviles ponderadas y los modelos exponenciales. Hemos visto cómo pueden utilizarse algunas variantes de estos métodos para controlar y estimar el efecto de los principales componentes.

Hemos introducido los modelos autorregresivos para ilustrar el enfoque estocástico de las predicciones de datos de series temporales. En ese enfoque, estimamos parámetros de un modelo que podrían haber generado la serie temporal. Un enfoque consiste en utilizar modelos autorregresivos en los que se plantea que una medida en el periodo t es una función lineal de las observaciones pasadas más un término de error aleatorio. El desarrollo del modelo implica la especificación del modelo, la estimación y a continuación la realización de un contraste para averiguar la eficacia del modelo para hacer predicciones. Por último, hemos ofrecido una visión panorámica de los modelos integrados autorregresivos de medias móviles, que son la base de una amplia variedad de especificaciones de modelos, dependiendo de la estructura que se crea que tiene el proceso.

TÉRMINOS CLAVE

análisis de los componentes de las series temporales, 779
cálculo de índices de precios de un único artículo, 767
cambio del periodo base, 770
contraste de rachas, 775
contraste de rachas: grandes muestras, 775
índice de cantidades agregado ponderado, 769
índice de cantidades de Laspeyres, 770
índice de precios agregado ponderado, 768
índice de precios enlazado, 771
índice de precios de Laspeyres, 768
índice de precios no ponderado, 767

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

19.49. Vuelva al ejercicio 19.35 y al fichero de datos **Hourly Earnings**, que muestra los ingresos mensuales por hora de la industria manufacturera.

a) Calcule un índice con el mes 1 como base.
b) Calcule un índice con el mes 5 como base.

19.50. Una biblioteca compra libros y revistas. La tabla adjunta y el fichero de datos **Library Purchases** muestran los precios medios (en dólares) pagados por cada uno y las cantidades compradas en un periodo de 6 años. Utilice el año 1 como base.

	Libros		Revistas	
Año	Precio	Cantidad	Precio	Cantidad
1	20,4	694	30,1	155
2	22,3	723	33,4	159
3	23,3	687	36,0	160
4	24,6	731	39,8	163
5	27,0	742	45,7	160
6	29,2	748	50,7	155

a) Halle el índice de precios agregado no ponderado.
b) Halle el índice de precios de Laspeyres.
c) Halle el índice de cantidades de Laspeyres.

19.51. Explique la afirmación de que puede considerarse que una serie temporal está formada por varios componentes. Ponga ejemplos de series temporales empresariales y económicas en las que es de esperar que sean importantes determinados componentes.

19.52. En muchas aplicaciones empresariales, las predicciones de los futuros valores de las series temporales, como las ventas y los beneficios, se hacen exclusivamente con información pasada sobre la serie temporal en cuestión. ¿Qué características de la conducta de las series temporales se explota en la producción de esas predicciones?

19.53. Una persona encargada del control de las existencias solicita predicciones mensuales de las ventas de varios productos para los 6 próximos meses. Esta persona tiene datos sobre las ventas mensuales de cada uno de estos productos de los 4 últimos años. Decide utilizar como predicciones para cada uno de los 6 próximos meses las ventas mensuales medias de los 4 últimos años. ¿Cree que es una buena estrategia? Explique su respuesta.

19.54. ¿Qué se entiende por ajuste estacional de una serie temporal? Explique por qué los organismos oficiales realizan muchos esfuerzos para desestacionalizar las series temporales económicas.

19.55. El fichero de datos **US Industrial Production** muestra un índice de producción industrial de Estados Unidos de 14 años.

a) Realice un contraste de aleatoriedad de esta serie utilizando el contraste de rachas.
b) Trace un gráfico temporal de estos datos y analice las características que revela el gráfico.
c) Calcule la serie de medias móviles centradas simples de 3 puntos. Represente gráficamente esta serie suavizada y analice su conducta.

19.56. El fichero de datos **Product Sales** muestra 24 observaciones anuales sobre las ventas de un producto.

a) Utilice la variante del contraste de rachas para grandes muestras para hacer un contraste de aleatoriedad de esta serie.
b) Trace un gráfico temporal de los datos y analice las características de la serie mostrada en este gráfico.
c) Calcule la serie de medias móviles centradas simples de 5 puntos. Represente gráficamente esta serie suavizada y analice su conducta.

19.57. El fichero de datos **Quarterly Earnings 19.57** muestra los beneficios trimestrales por acción de una empresa en 7 años.

a) Represente gráficamente estos datos. ¿Sugiere este gráfico la presencia de un fuerte componente estacional?

b) Utilice el método del índice estacional para obtener una serie desestacionalizada.

19.58. El fichero de datos **Price Index** muestra 15 valores mensuales del índice de precios de una mercancía.

a) Calcule la serie de medias móviles centradas simples de 3 puntos.

b) Trace un gráfico temporal de la serie suavizada y comente sus características.

19.59. Vuelva al ejercicio 19.56 y al fichero de datos **Product Sales**. Utilice la suavización exponencial simple con una constante de suavización $\alpha = 0{,}5$ para hacer predicciones de las ventas para los 3 próximos años.

19.60. Vuelva al ejercicio 19.58 y al fichero de datos **Price Index**. Utilice el método de Holt-Winters con las constantes de suavización $\alpha = 0{,}3$ y $\beta = 0{,}4$ para hacer predicciones del índice de precios para los 4 próximos meses.

19.61. Vuelva al ejercicio 19.57 y al fichero de datos **Quarterly Earnings 19.57**. Utilice el método estacional de Holt-Winters con las constantes de suavización $\alpha = 0{,}4$, $\beta = 0{,}4$ y $\gamma = 0{,}2$ para hacer predicciones de esta serie de beneficios por acción para los cuatro próximos trimestres.

19.62. Basándose en el fichero de datos **Product Sales** del ejercicio 19.59, estime modelos autorregresivos de órdenes 1 a 4 para las ventas del producto. Utilizando el método del apartado 19.6 para contrastar la hipótesis de que el orden autorregresivo es $(p - 1)$ frente a la alternativa de que el orden es p, con un nivel de significación del 10 por ciento, elija uno de estos modelos. Haga predicciones para los 3 próximos años a partir del modelo elegido.

Bibliografía

1. Box, G. E. P. y G. M. Jenkins, *Time Series Analysis*, *Forecasting*, *and Control*, San Francisco, Holden-Day, 1970.
2. Granger, C. W. y P. Newbold, *Forecasting Economic Time Series*, Orlando, Fl, Academic Press, 1986, 2.ª ed.
3. Newbold, P. y T. Bos, *Introductory Business Forecasting*, Cincinnati, OH, South-Western, 1994, 2.ª ed.

Capítulo 20

Otros temas relacionados con el muestreo

Esquema del capítulo

Introducción

Una gran parte de la inferencia estadística se refiere a problemas en los que se hacen afirmaciones sobre una población basándose en información procedente de una muestra. Hasta ahora hemos tratado de una manera bastante superficial dos importantes temas. En primer lugar, apenas nos hemos referido a la forma en que se seleccionan realmente los miembros de la muestra. En segundo lugar, hemos supuesto en general que el número de miembros de la población es muy grande en comparación con el número de miembros de la muestra. En este capítulo examinamos el problema del investigador que quiere descubrir algo sobre una población que no es necesariamente grande. El investigador pretende reunir información únicamente sobre un subconjunto de la población y necesita orientación para saber cómo debe reunirla.

20.1. Pasos básicos de un estudio realizado por muestreo

Los analistas de mercado a menudo estudian las poblaciones humanas para obtener información sobre sus preferencias por un producto. Los auditores normalmente seleccionan una muestra de facturas pendientes de cobro de una empresa. Se hacen inferencias sobre la población correspondiente basándose en estas muestras. Los directores de personal requieren información sobre las actitudes de los empleados hacia los nuevos métodos de producción propuestos y les resulta útil tomar una muestra de la plantilla. Naturalmente, el uso de métodos de muestreo está muy extendido y va más allá del campo de la empresa. Tal vez los ejemplos más conocidos sean las encuestas que se hacen periódicamente sobre las preferencias de los votantes antes de las elecciones. La información recogida tiene interés no sólo para el público en general sino también para los asesores de los candidatos que tratan de averiguar dónde deben concentrar más los esfuerzos. Esas encuestas a los votantes han aumentado tanto que se recaba la opinión de los votantes sobre todos los aspectos de la política y los encuestadores profesionales se han convertido en una importante figura en el séquito del político.

Antes de preguntar cómo debe tomarse una muestra de una población, tal vez se pregunte el lector por qué hay que tomar una muestra. La alternativa es intentar obtener información de todos los miembros de la población. En ese caso, hablaríamos de *censo* y no de muestra. Hay varias razones por las que a menudo se prefiere una muestra a un censo. En primer lugar, en muchas aplicaciones sería enormemente caro tomar un censo completo, a menudo prohibitivo. En segundo lugar, muchas veces es necesario disponer de información bastante deprisa; un censo completo, incluso aunque sea económicamente viable, puede tardar tanto en realizarse que el valor de los resultados puede disminuir seriamente. Otra razón para tomar una muestra es que con los métodos estadísticos modernos generalmente es posible obtener resultados con el grado deseado de precisión por medio del muestreo. El tiempo y el dinero necesarios para obtener números cuya precisión aparente es mayor que la que necesita el investigador podrían dedicarse mejor a otras cosas. Además, si se toma una muestra relativamente pequeña, los beneficios que se obtendrían haciendo un esfuerzo mayor para conseguir información precisa de los miembros de la muestra podrían muy bien ser mayores que los beneficios de obtener información de un grupo mayor que puede ser menos fiable debido a las limitaciones de tiempo y de costes. En cuarto lugar, algunos muestreos son destructivos y los sujetos contrastados se destruyen en el estudio. Así sucedería si se tratara de contrastar la duración de las bombillas, la duración de una determinada marca de neumáticos o la resistencia de los tubos de vidrio a las roturas. Estos factores —coste, tiempo, precisión y carácter destructivo— considerados en conjunto llevan a preferir en muchas ocasiones las muestras a los censos.

Supongamos ahora que se necesita información sobre una población y que se ha decidido tomar una muestra. Es cómodo considerar que un estudio realizado por muestreo consta de los seis pasos siguientes, cada uno destinado a dar una respuesta a una pregunta. La Figura 20.1 muestra estos pasos.

1. Primer paso: ¿qué información se necesita?
2. Segundo paso: ¿cuál es la población relevante y existe un listado de esa población?
3. Tercer paso: ¿cómo deben seleccionarse los miembros de la muestra?
4. Cuarto paso: ¿cómo debe obtenerse información de los miembros de la muestra?
5. Quinto paso: ¿cómo debe utilizarse la información muestral para hacer inferencias sobre la población?
6. Sexto paso: ¿qué conclusiones pueden extraerse sobre la población?

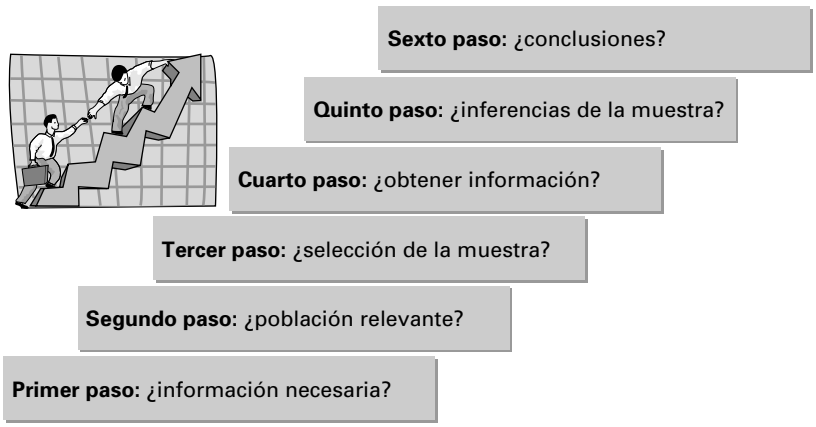


Figura 20.1.
Pasos en un estudio realizado por muestreo.

Se analiza cada uno de esos pasos en relación con un problema de un estudio de mercado. Supongamos que un editor pretende publicar un nuevo libro de texto de estadística y quiere información sobre la situación actual del mercado. La información valiosa podría ser el número de estudiantes matriculados en los cursos de estadística para los negocios, la penetración de los textos existentes en el mercado y las opiniones de los profesores sobre los temas que son más importantes para sus cursos. Supongamos que el editor quiere recoger datos de una muestra de campus universitarios.

1. ¿Qué información se necesita?

La respuesta a esta pregunta es tanto el motivo como el punto de partida para realizar el estudio. Si la información necesaria ya existe o es imposible de obtener, no tiene sentido realizar el estudio. Por muy sencilla que parezca la pregunta, a menudo es necesario lograr un equilibrio bastante delicado en esta fase. El investigador puede estar pensando en un único tema o puede haber varios temas de interés. Pero dado que va a realizarse el estudio, con todos sus costes, normalmente merece la pena preguntarse si puede obtenerse en el estudio más información potencialmente útil con un gasto adicional mínimo. En el caso del editor del libro de estadística para los negocios, las preguntas más útiles se refieren al tamaño del mercado, a la situación de los competidores y a los temas que los profesores consideran más importantes. Dado que hay que entrar en contacto con los miembros de la muestra para recabar esta información, puede merecer la pena hacer algunas preguntas más. Éstas pueden ser si el curso es de un cuatrimestre o de dos, si es optativo u obligatorio, el departamento del profesor, el método para adoptar el libro y el tiempo que lleva utilizándose el libro actual. Una vez elegido ese camino, se puede tener la tentación de dejar que la lista de preguntas aumente espectacularmente, ya que eso generalmente no incrementa mucho el coste del estudio. Sin embargo, puede *tener* un problema. Es más probable que los encuestados cooperen en un estudio en el que se hacen relativamente pocas preguntas, ya que se les quita poco tiempo. Es importante, pues, para el investigador buscar el equilibrio, es decir, hacer preguntas sobre cuestiones centrales (pues, si se descubre una omisión importante, puede ser demasiado caro repetir todo el ejercicio) y conseguir que el número de preguntas sea tolerable para los encuestados.

2. ¿Cuál es la población relevante y existe un listado de esa población?

Parece bastante trivial señalar que para hacer inferencias sobre una población, ésa es la población que debe muestrearse. No obstante, a menudo se han extraído dudosas conclusiones tras un análisis, por lo demás absolutamente respetable, de los datos de encuesta precisamente porque no se ha tenido en cuenta este punto elemental. Muchas publicaciones piden la opinión de sus lectores sobre determinadas cuestiones. Sin embargo, sería peligroso generalizar sus respuestas a la población en general. La población estudiada en este caso es simplemente la de lectores de la publicación y es probable que estos lectores no sean representativos del público en general. En muchos estudios prácticos, la población *real* de interés puede ser imposible de definir. Por ejemplo, una organización que intenta predecir el resultado de unas elecciones presidenciales sólo está interesada realmente en la población que votará. Aunque ésta es la población relevante, sus miembros no son fáciles de distinguir. Una posibilidad es, por supuesto, preguntar a un miembro de una muestra si tiene intención de votar. Si embargo, es bien sabido que la proporción que responde afirmativamente a una pregunta de ese tipo es mayor que la proporción que acaba votando. Otra posibilidad es preguntar si el encuestado votó en las elecciones anteriores, pero esta pregunta también dista de ser totalmente satisfactoria.

Es probable que el editor del libro de texto considere que la población relevante son todos los profesores (o quizá todas las universidades) que imparten cursos de estadística para los negocios. La población es bastante fácil de identificar y, como consecuencia de actividades de marketing anteriores, el editor tendrá casi con toda seguridad un listado bastante preciso de sus miembros.

3. ¿Cómo deben seleccionarse los miembros de la muestra?

Una gran parte del resto de este capítulo se dedica a responder a esta pregunta. En pocas palabras, no existe una única forma de conseguir el «mejor» sistema de muestreo. La elección correcta depende generalmente del problema en cuestión y de los recursos del investigador. Ya hemos introducido anteriormente el concepto de *muestreo aleatorio simple*, en el que todos los miembros de una población tienen la misma probabilidad de ser elegidos para la muestra. De hecho, todos los instrumentos para analizar los datos que hemos introducido hasta ahora se basaban en el supuesto de que la muestra se elegía de esta forma. Existen, sin embargo, muchas circunstancias en las que podría preferirse otro sistema de muestreo. Supongamos que a nuestro editor le interesan las diferencias entre el tratamiento que se da a la estadística empresarial en las escuelas universitarias de grado medio y el que se le da en las facultades de grado superior. Sería importante que la muestra contuviera suficientes centros de cada tipo para poder extraer conclusiones fiables sobre ambos. Sin embargo, el muestreo aleatorio simple no garantiza en modo alguno que se logre ese objetivo. Por ejemplo, es absolutamente posible que la muestra elegida contenga una preponderancia de facultades. Para evitar esta posibilidad, pueden extraerse muestras aleatorias simples de las respectivas poblaciones de los dos tipos. Éste es un ejemplo de *muestreo estratificado*, que se analiza más detalladamente en el apartado 20.4. Otra cuestión que hay que decidir en esta fase es el número de miembros de la muestra. En este caso, la elección depende esencialmente del grado de precisión necesario y de los costes que implica. Esta cuestión se aborda en el apartado 20.5.

4. ¿Cómo debe obtenerse información de los miembros de la muestra?

Esta pregunta es extraordinariamente importante y ha sido objeto de muchas investigaciones. En términos generales, plantea dos importantes cuestiones. En primer lugar, el investigador quiere obtener respuestas de la mayor proporción posible de los miembros de la muestra. Si el número que no responde es alto, será difícil estar seguro de que los que han respondido son representativos de la población en general. Por ejemplo, los profesores que no facilitan información al editor del libro de texto pueden estar más dedicados a la investigación, a la consultoría o a otras actividades y sus preferencias sobre los libros pueden muy bien ser diferentes de las de sus colegas. Recuérdese que el número de preguntas formuladas en una encuesta puede afectar a la tasa de respuesta. También influye la forma en que se contacta con los miembros de la muestra. A menudo los cuestionarios se envían por correo a las personas seleccionadas para la muestra y a menudo ocurre que la proporción que responde es decepcionantemente baja. Muchos investigadores intentan mejorar la tasa de respuesta adjuntando una carta en la que explican los fines del estudio y solicitan ayuda educadamente. La garantía del anonimato también puede ser valiosa. La inclusión de un sobre con el franqueo pagado para devolver el cuestionario generalmente merece la pena; también puede prometerse algún pequeño incentivo monetario o regalo. No obstante, habrá casi inevitablemente una proporción de personas que no respondan y es una buena práctica instituir un estudio de seguimiento para tratar de obtener más información sobre ellas. Es probable que los métodos de contacto más caros, como las llamadas telefónicas o las visitas de los entrevistadores a las casas, logren un nivel de respuesta más alto. Sin embargo, esos métodos pueden ser caros en tiempo y dinero y la decisión de cómo recoger información debe depender de los recursos del investigador y del grado en que se piense que la falta de respuesta puede ser un problema serio.

El editor del libro de texto puede decidir enviar cuestionarios por correo a los miembros de la muestra. Sería barato, por lo que podría extraerse una muestra inicial relativamente grande. La esperanza es que la proporción de personas que no responden no sea demasiado alta y que las respuestas obtenidas sean razonablemente representativas. Si se teme que la falta de respuesta introduzca un sesgo considerable si se envía un cuestionario por correo, se podría tomar una muestra inicial más pequeña y hacer un esfuerzo mayor para contactar con sus miembros. Una estrategia viable es pedir a los representantes de la empresa, que visitan periódicamente los campus, que realicen entrevistas con miembros de la muestra en su siguiente visita. Ese método debería garantizar una tasa de respuesta bastante alta. Su principal dificultad estriba en el tiempo necesario para realizar todas las entrevistas más que en el coste adicional, que sería bastante bajo.

El segundo punto es obtener respuestas que sean lo más exactas y sinceras posible. No sirve de nada hacer un sofisticado análisis estadístico de información que no es fiable. Formular las preguntas, ya sea para enviarlas por correo o para que las realice un encuestador, de tal forma que se consigan respuestas sinceras y exactas es todo un arte. Es importante que las preguntas se formulen de la manera más clara e inequívoca posible, de modo que los sujetos entiendan lo que se les pregunta. También se sabe perfectamente que la formulación de las preguntas o el tono del entrevistador pueden inducir a los encuestados a dar determinadas respuestas. Los entrevistadores no deben dar en modo alguno la impresión de que tienen firmes ideas sobre el tema en cuestión o de que quieren una respuesta concreta. También es importante no predisponer a los encuestados: las preguntas deben formularse de la forma más neutral posible. Por poner un ejemplo extremo, consideremos los dos métodos siguientes para preguntar esencialmente lo mismo:

a) ¿Qué tres temas considera más importantes en su curso de estadística para los negocios?
b) ¿Está de acuerdo en que los métodos modernos de gestión de la calidad, debido a su enorme importancia en el mundo de la empresa, ahora deben considerarse uno de los más importantes en cualquier curso de estadística para los negocios?

Naturalmente, nadie que tenga interés en tener una idea precisa de las opiniones de los profesores haría la segunda pregunta. Sin embargo, se ha observado que formulaciones que tienen un sesgo mucho menos claro que el de ésta influyen significativamente en las respuestas de los sujetos.

5. ¿Cómo debe utilizarse la información de la muestra para hacer inferencias sobre la población?

Hemos dedicado la mayor parte de este libro a dar respuesta justamente a esta pregunta. En los apartados posteriores de este capítulo, analizamos métodos de inferencia de diseños de muestreo específicos. El objetivo principal del presente apartado es señalar la importancia de otros aspectos de un estudio por muestreo.

6. ¿Qué conclusiones pueden extraerse sobre la población?

Por último, cerramos el círculo y preguntamos qué puede decirse sobre la población estudiada como consecuencia de una investigación estadística. ¿Ha dado el estudio claras respuestas a las preguntas que lo motivaron? ¿Han surgido otras cuestiones importantes en el curso del estudio? En esta fase, el investigador tiene la tarea de resumir y presentar la información recogida. Para eso pueden ser necesarias estimaciones puntuales o por intervalos, así como tablas o gráficos que resuman los principales resultados. ¿Cuál es la mejor estimación del número de estudiantes matriculados en los cursos de estadística para los negocios y pueden estimarse intervalos de confianza en torno a esta estimación? ¿Cuáles son los libros de texto más populares en este momento? ¿Qué temas consideran más importantes los profesores? ¿Existen diferencias significativas entre los mercados de las escuelas universitarias y las facultades? En esta fase, la tarea es informar sobre los resultados del estudio y decidir cómo proceder. Puede que el análisis sugiera la conveniencia de recoger más información.

A menudo surgen importantes cuestiones imprevistas durante el curso del estudio que inducen al investigador a estudiar en mayor profundidad la población. Ésta es la razón por la que nuestro editor hace una pregunta abierta como la siguiente: «Nuestra empresa está considerando la posibilidad de introducir en el mercado un nuevo libro de texto de economía. ¿Hay alguna característica que le gustaría que tuviera ese libro?». Supongamos, además, que cuando se devuelven los cuestionarios, un número considerable menciona la posibilidad de que se venda simultáneamente una gran base de datos que contenga datos sobre problemas reales del mundo de la empresa. Analizando estos datos, los estudiantes podrían adquirir experiencia práctica en temas del curso. Antes de incurrir en el coste de producir este programa informático, al editor podría merecerle la pena tomar otra muestra para evaluar las probabilidades de éxito de este proyecto.

EJERCICIOS

Ejercicios básicos

20.1. Suponga que quiere realizar un estudio para conocer las opiniones de los estudiantes de administración de empresas de su campus sobre la necesidad de que la asignatura de estadística sea obligatoria. Analice los pasos que seguiría para realizar este estudio, los problemas que esperaría encontrar y las técnicas que podría utilizar para resolver los problemas.

20.2. Las autoridades universitarias tienen interés en conocer las opiniones de los estudiantes sobre algunos servicios universitarios (como la matrícula, los comedores o el servicio médico). Le han pedido que haga una encuesta. Sugiera cómo seguiría los seis pasos de un estudio de muestreo.

20.3. El director de una tienda de ropa situada en el campus está considerando la posibilidad de introducir algunos artículos más de marca y quiere evaluar la demanda de estos artículos por parte de los estudiantes. Se le ha encargado que diseñe una encuesta para obtener esta información. Explique detalladamente lo que haría.

20.4. Una empresa de servicios financieros está considerando la posibilidad de introducir tres nuevos tipos de fondos de inversión. Se cree que, al menos inicialmente, la mayor parte del apoyo probablemente provendría de sus clientes actuales. A la empresa le gustaría evaluar el grado de interés que tienen estos clientes en los nuevos productos propuestos y preferiblemente conocer también las características relevantes de las personas más interesadas. Le han encargado un estudio con un presupuesto limitado. ¿Qué haría?

20.5. A los ejecutivos de una compañía de seguros, conscientes de que han aumentado significativamente algunos tipos de primas de seguro en los últimos años, les preocupa la imagen pública de su sector y la posibilidad de que tenga repercusiones políticas. Se ha decidido lanzar una campaña de relaciones públicas para informar al público sobre las causas de los incrementos de los costes. Sin embargo, existe mucha incertidumbre sobre los temas que más preocupan a la gente y sobre el grado en que se comprenden los factores que subyacen a las subidas de los precios. Explique cómo podría organizar un estudio para obtener información relevante. Siga los pasos básicos de un plan de muestreo.

20.2. Errores de muestreo y errores ajenos al muestreo

Cuando se toma una muestra de una población, no es posible saber cuál es *exactamente* el valor de cualquier parámetro poblacional, como la media o la proporción. Cualquier estimación puntual tendrá inevitablemente un error. Recuérdese que una de las fuentes de error, llamado **error de muestreo**, se debe a que sólo se dispone de información sobre un subconjunto de todos los miembros de la población. Dados ciertos supuestos, la teoría estadística nos permite caracterizar la naturaleza del error de muestreo y hacer afirmaciones probabilísticas bien definidas sobre los parámetros poblacionales, como los intervalos de confianza analizados en los Capítulos 8 y 9. En apartados posteriores de este capítulo, analizamos métodos de inferencia estadística para varios sistemas importantes de muestreo. Sin embargo, es importante reconocer primero otra fuente posible de error, que no puede analizarse de una forma tan exacta o clara.

En los análisis prácticos, puede haber errores que no tengan que ver con el tipo de sistema de muestreo utilizado. De hecho, esos errores podrían cometerse también si se tomara un censo completo de la población. Son **errores ajenos al muestreo**. En cualquier encuesta, existe la posibilidad de que haya en algunos lugares un error ajeno al muestreo. He aquí algunos ejemplos:

1. **La población de la que se hace realmente el muestreo no es la relevante.** En 1936, ocurrió un conocido caso de este tipo, cuando la revista *Literary Digest* pre-

dijo con seguridad que Alfred Landon ganaría las elecciones frente a Franklin Roosevelt. Sin embargo, Roosevelt ganó por un amplio margen. Este error de predicción se debió a que los miembros de la muestra de *Digest* se habían tomado de las guías de teléfono y de otros listados, como las listas de suscriptores a revistas y los registros de automóviles. En estas fuentes, estaban claramente subrepresentados los pobres, que eran predominantemente demócratas. Para hacer una inferencia sobre una población (en este caso, sobre el electorado estadounidense), es importante hacer una muestra de esa población y no de algún subgrupo, por muy cómodo que parezca esto último.

2. **Los sujetos de la encuesta pueden dar una respuesta inexacta o falsa.** Eso podría ocurrir porque las preguntas se formulan de una manera difícil de entender o de una forma que parece que una respuesta es más agradable o más deseable. Además, muchas preguntas que uno querría hacer son tan delicadas que sería imprudente esperar que todas las respuestas fueran sinceras. Supongamos, por ejemplo, que un jefe de planta quiere evaluar las pérdidas anuales de la empresa que se deben a robos de los empleados. En principio, se podría seleccionar una muestra aleatoria de empleados y preguntar a sus miembros «¿qué ha robado en esta planta en los 12 últimos meses?». ¡Ésta no es, desde luego, la forma más fiable de conseguir la información necesaria!
3. **Falta de respuesta a las preguntas de la encuesta.** Los sujetos de una encuesta pueden no responder a ninguna pregunta o pueden no responder a algunas. Si ocurre en muchos casos, puede haber más errores de muestreo o errores ajenos al muestreo. El error de muestreo se debe a que el tamaño de la muestra logrado será menor que el pretendido. El error ajeno al muestreo puede deberse a que la población de la muestra no es la población que interesa. Los resultados obtenidos pueden considerarse una muestra aleatoria *de la población que está dispuesta a responder*. Estas personas pueden ser diferentes en importantes aspectos de la población en general. En ese caso, habrá un sesgo en las estimaciones resultantes.

No existe ningún método general para identificar y analizar los errores ajenos al muestreo, pero éstos pueden ser importantes. El investigador debe tener cuidado en cuestiones como la identificación de la población relevante, el diseño del cuestionario y la falta de respuesta para reducir lo más posible su importancia. En el resto de este capítulo, suponemos que se tiene ese cuidado, por lo que en nuestro análisis centramos la atención en el tratamiento de los errores de muestreo.

EJERCICIOS

Ejercicios básicos

20.6. Vuelva al estudio del ejercicio 20.2.

a) Dentro del sistema de muestreo que ha diseñado, ¿ve la posibilidad de que haya errores ajenos al muestreo? En caso afirmativo, ¿qué medidas tomaría para reducir lo más posible su magnitud?

b) ¿Es probable que la falta de respuesta sea una cuestión grave en este estudio? En caso afirmativo, ¿qué podría hacerse para resolverla?

20.7. Vuelva al estudio del ejercicio 20.3.

a) Analice las causas probables de los errores ajenos al muestreo e indique cómo podrían reducirse lo más posible.

b) ¿Es de esperar que la falta de respuesta sea un problema grave para realizar este estudio? En caso afirmativo, ¿cómo podría paliarse el problema?

20.8. En el caso del estudio del ejercicio 20.5, analice la posibilidad de que haya errores ajenos al

muestreo y falta de respuesta. Indique qué haría para reducir lo más posible estos problemas.

20.9. Un método para hacer frente a un tipo de falta de respuesta es el *método del recuerdo*. Se realiza una encuesta a los hogares en la que los entrevistadores llaman el jueves por la tarde. Se vuelve a llamar el jueves siguiente a los hogares en los que no hay nadie en casa. Este proceso puede continuar hasta que se logra hablar el jueves siguiente con los hogares con los que no se pudo hablar los dos jueves anteriores. ¿Cuál podría ser el valor de la información obtenida de esta forma?

20.3. Muestreo aleatorio simple

En el resto de este capítulo, analizamos problemas en los que se extrae una muestra de n individuos u objetos de una población que contiene un total de N miembros. En las aplicaciones prácticas, se han utilizado muchos sistemas para seleccionar esas muestras. Nuestros análisis centrarán en gran parte la atención en los métodos de *muestreo probabilístico*, que son métodos en los que se utiliza algún mecanismo en el que interviene el *azar* para decidir los miembros de la muestra y se sabe cuál es la probabilidad de obtener una determinada muestra. Hacemos de nuevo hincapié en el concepto de muestreo aleatorio simple y en la forma en que se toma una muestra aleatoria simple de una población finita, debido a su importancia.

Muestreo aleatorio simple

Supongamos que tenemos que seleccionar una muestra de n objetos de una población de N objetos. Un método de **muestreo aleatorio simple** es aquel en el que todos los miembros de una población tienen la misma probabilidad de ser elegidos para la muestra.

Supongamos que nuestra población está formada por 1.000 individuos, numerados del 1 al 1.000 y que se necesita una muestra aleatoria simple de 100 miembros de la población. El programa Minitab puede generar fácilmente una muestra aleatoria simple. Por ejemplo, una lista parcial de los 100 números aleatorios que generamos con Minitab incluye las personas que tienen los números

457 229 843 460 918 311

Sólo consideraremos el *muestreo sin repetición*, en el que se excluye cualquier número que ya ha salido y el proceso continúa hasta que se obtienen 100 números *diferentes*. No analizamos aquí la alternativa, el *muestreo con repetición*, que permite incluir un individuo en la muestra más de una vez.

El muestreo sistemático es un método de muestreo estadístico que se utiliza a menudo como alternativa al muestreo aleatorio.

Muestreo sistemático

Supongamos que la lista de la población se ordena de una forma que no tiene ninguna relación con el tema de interés. El **muestreo sistemático** implica la selección de todo j-ésimo sujeto de la población, donde j es el cociente entre el tamaño de la población N y el tamaño que se desea que tenga la muestra, n; es decir, $j = N/n$. Se selecciona aleatoriamente un número del 1 al j para obtener el primer sujeto que va a incluirse en la muestra sistemática.

Supongamos que se desea que el tamaño de la muestra sea de 100 y que la población está formada por 5.000 nombres en orden alfabético. En ese caso, $j = 50$. Seleccionamos aleatoriamente un número del 1 al 50. Si el número es el 20, seleccionamos ese número y los sucesivos números obtenidos sumando 50 al número inicial; de esa manera, se obtiene una muestra sistemática formada por los elementos que llevan los números 20, 70, 120, 170, etc. hasta que se seleccionan los 100 sujetos. Una muestra sistemática se analiza de la misma forma que una muestra aleatoria simple, ya que, en relación con el tema investigado, la lista de la población ya está en orden aleatorio. El peligro está en que exista alguna relación sutil e inesperada entre el orden de la población y el tema estudiado. En ese caso, habría un sesgo si se empleara un muestreo sistemático. Las muestras sistemáticas constituyen una buena representación de la población si la población no experimenta ninguna variación cíclica.

Análisis de los resultados de un muestreo aleatorio simple

En este apartado se amplían las estimaciones del intervalo de confianza desarrolladas en el Capítulo 8. Sin embargo, aquí se analizan los casos en los que el número de miembros de la muestra no es una proporción insignificante del número de miembros de la población. Por lo tanto, se utiliza el **factor de corrección en el caso de una población finita**, $(N - n)/N$. Se supondrá que la muestra es lo suficientemente grande para poder recurrir al teorema del límite central.

Estimación de la media poblacional, muestra aleatoria simple

Sean x_1, x_2, ..., x_n los valores observados en una muestra aleatoria simple de tamaño n, tomada de una población de N miembros que tiene una media μ.

1. La media muestral es un estimador insesgado de la media poblacional, μ. La estimación puntual es

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

2. Un método de estimación insesgada de la varianza de la media muestral genera la estimación puntual

$$\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} \times \frac{N - m}{N} \qquad \textbf{(20.1)}$$

3. Siempre que el tamaño de la muestra es grande, los intervalos de confianza al $100(1 - \alpha)\%$ de la media poblacional son

$$\bar{x} - z_{\alpha/2}\hat{\sigma}_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2}\hat{\sigma}_{\bar{x}} \qquad \textbf{(20.2)}$$

EJEMPLO 20.1. Créditos hipotecarios (intervalo de confianza)

En una ciudad, se solicitaron 1.118 créditos hipotecarios el año pasado. Una muestra aleatoria de 60 de estos créditos era de una cuantía media de 87.300 $ y tenía una desviación típica de 19.200 $. Estime la cantidad media de todos los créditos hipotecarios solicitados en esta ciudad el año pasado y halle el intervalo de confianza al 95 por ciento.

Solución

Sea μ la media poblacional. Se sabe que

$$N = 1.118 \qquad n = 60 \qquad \bar{x} = 87.300\ \$ \qquad s = 19.200$$

Para obtener estimaciones de intervalos, utilizamos la ecuación 20.1:

$$\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} \times \frac{(N-n)}{N} = \frac{(19.200)^2}{60} \times \frac{1.058}{1.118} = 5.814.268$$

y tomamos la raíz cuadrada para hallar el error típico estimado,

$$\hat{\sigma}_{\bar{x}} = 2.411$$

Por lo tanto, el intervalo de confianza al 95 por ciento de la cantidad media de todas las hipotecas solicitadas en esta ciudad el año pasado es

$$87.300\ \$ - (1{,}96)(2.411) < \mu < 87.300\ \$ + (1{,}96)(2.411)$$

o sea

$$82.574\ \$ < \mu < 92.026\ \$$$

Es decir, el intervalo va de 82.574 $ a 92.026 $.

A menudo, lo que interesa es el total poblacional en lugar de la media. Por ejemplo, el editor de un libro de texto de estadística para los negocios querrá una estimación del número total de estudiantes que asisten a los cursos de estadística para los negocios en todo el país. Es fácil hacer una inferencia sobre el total poblacional. Los resultados relevantes se deducen del hecho de que en nuestra notación, el total poblacional $= N\mu$.

Estimación del total poblacional, muestra aleatoria simple

Supongamos que se selecciona una muestra aleatoria simple de tamaño n de una población de tamaño N y que la cantidad que se quiere estimar es el total poblacional $N\mu$.

1. Un método de estimación insesgada del total poblacional $N\mu$ genera la estimación puntual $N\bar{x}$.
2. Un método de estimación insesgada de la varianza de nuestro estimador del total poblacional genera la estimación puntual:

$$N^2\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} N(N-n) \qquad \textbf{(20.3)}$$

3. Siempre que el tamaño de la muestra es grande, se obtiene un intervalo de confianza al $100(1-\alpha)\%$ del total poblacional de la forma siguiente:

$$N\bar{x} - z_{\alpha/2}N\hat{\sigma}_{\bar{x}} < N\mu < N\bar{x} + z_{\alpha/2}N\hat{\sigma}_{\bar{x}} \qquad \textbf{(20.4)}$$

EJEMPLO 20.2. Número de matriculados en los cursos de estadística para los negocios (intervalo de confianza)

Supongamos que hay 1.395 universidades en un país. En una muestra aleatoria simple de 400 universidades, se observa que la media muestral del número de matriculados el año pasado en los cursos de estadística para los negocios era de 320,8 estudiantes y que la desviación típica muestral era de 149,7 estudiantes. Estime el número total de estudiantes matriculados en estos cursos durante el año y halle el intervalo de confianza al 99 por ciento.

Solución

Si la media poblacional es μ, para estimar $N\mu$ se utilizan los datos siguientes:

$$N = 1.395 \qquad n = 400 \qquad \bar{x} = 320{,}8 \qquad s = 149{,}7$$

Nuestra estimación puntual del total es

$$N\bar{x} = (1.395)(320{,}8) = 447.516$$

Se estima que hay un total de 447.516 alumnos matriculados en los cursos. Para obtener estimaciones de intervalos, se utiliza la ecuación 20.3 para calcular la varianza del estimador:

$$N^2\hat{\sigma}_{\bar{x}}^2 = \frac{s^2}{n} N(N - n) = \frac{(149{,}7)^2}{400} (1.395)(995) = 77.764{,}413$$

Tomando la raíz cuadrada, tenemos que

$$N\hat{\sigma}_{\bar{x}} = 8.818{,}4$$

Por lo tanto, el intervalo de confianza al 99 por ciento del total poblacional se obtiene aplicando la ecuación 20.4, siendo $z_{\alpha/2} = 2{,}58$:

$$N\bar{x} - Z_{\alpha/2}N\hat{\sigma}_{\bar{x}} < N\mu < N\bar{x} + Z_{\alpha/2}N\hat{\sigma}_{\bar{x}}$$

o sea

$$447.516 - (2{,}58)(8.818{,}4) < N\mu < 447.516 + (2{,}58)(8.818{,}4)$$

o sea

$$447.516 \pm 22.751$$

$$424.765 < N\mu < 470.267$$

Por lo tanto, nuestro intervalo va de 424.765 a 470.267 estudiantes.

Consideremos, por último, el caso en el que hay que estimar la *proporción p* de individuos de la población que poseen una característica específica. La inferencia sobre esta proporción debe basarse en la distribución hipergeométrica cuando el número de miembros de la muestra no es muy pequeño en comparación con el número de miembros de la población. Supongamos, de nuevo, que el tamaño de la muestra es lo suficientemente grande para poder invocar el teorema del límite central.

Estimación de la proporción poblacional, muestra aleatoria simple

Sea $\hat{p}$ la proporción que posee una determinada característica en una muestra aleatoria de n observaciones de una población que tiene una proporción, P, que posee esa característica.

1. La proporción muestral, $\hat{p}$, es un estimador insesgado de la proporción poblacional, P.
2. Un método de estimación insesgada de la varianza de nuestro estimador de la proporción poblacional genera la estimación puntual

$$\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1} \times \frac{(N-n)}{N} \qquad \textbf{(20.5)}$$

3. Siempre que el tamaño de la muestra es grande, los intervalos de confianza del $100(1-\alpha)\%$ de la proporción poblacional son

$$\hat{p} - z_{\alpha/2}\hat{\sigma}_{\hat{p}} < P < \hat{p} + z_{\alpha/2}\hat{\sigma}_{\hat{p}} \qquad \textbf{(20.6)}$$

EJEMPLO 20.3. Cursos anuales de estadística para los negocios (intervalo de confianza)

Se ha observado en una muestra aleatoria simple de 400 universidades de las 1.395 que hay en nuestra población que el curso de estadística para los negocios era un curso anual en 141 de las universidades de la muestra. Estime la proporción de todas las universidades en la que el curso es anual y halle el intervalo de confianza al 90 por ciento.

Solución

Dados

$$N = 1.395 \qquad n = 400 \qquad \hat{p} = \frac{141}{400} = 0{,}3525$$

nuestra estimación puntual de la proporción poblacional, P, es simplemente $\hat{p} = 0{,}3525$. Es decir, el curso es anual en alrededor del 35,25 por ciento de todas las universidades. Para calcular estimaciones de intervalos, la varianza de nuestra estimación se halla mediante la ecuación 20.5:

$$\hat{\sigma}^2_{\hat{p}} = \frac{\hat{p}(1-\hat{p})}{n-1} \times \frac{(N-n)}{N} = \frac{(0{,}3525)(0{,}6475)}{399} \times \frac{995}{1.395} = 0{,}0004080$$

por lo que

$$\hat{\sigma}_{\hat{p}} = 0{,}0202$$

En el caso de un intervalo de confianza al 90 por ciento, $z_{\alpha/2} = z_{0{,}05} = 1{,}645$. El intervalo de confianza al 90 por ciento se halla por medio de la ecuación 20.6:

$$\hat{p} - z_{\alpha/2}\hat{\sigma}_p < P < \hat{p} + z_{\alpha/2}\hat{\sigma}_{\hat{p}}$$

o sea

$$0{,}3525 - (1{,}645)(0{,}0202) < P < 0{,}3525 + (1{,}645)(0{,}0202)$$

o sea

$$0{,}3193 < P < 0{,}3857$$

Por lo tanto, el intervalo de confianza al 90 por ciento del porcentaje de todas las universidades en las que el curso de estadística para los negocios es anual va del 31,93 al 38,57 por ciento.

EJERCICIOS

Ejercicios aplicados

20.10. Consulte un periódico económico para obtener un listado de todas las acciones que cotizan en bolsa. Utilice el programa Minitab para obtener una muestra aleatoria simple de 20 acciones. Halle la subida porcentual media que experimentó el precio de las acciones de esta muestra la semana pasada.

20.11. Obtenga en su periódico local un listado de todos los anuncios de viviendas en venta en su ciudad. Utilice el programa Minitab para obtener una muestra aleatoria simple de 15 anuncios y halle la media muestral de los precios anunciados.

20.12. Un campus tiene 12.723 estudiantes. Quiere una muestra aleatoria de 100 de un listado completo de estos estudiantes. Explique cómo utilizaría el programa Minitab para obtener esa muestra aleatoria.

20.13. Tome una muestra aleatoria de 50 páginas de este libro y estime la proporción de todas las páginas que contienen cifras.

20.14. Una empresa tiene 189 contables. En una muestra aleatoria de 50 de ellos, el número medio de horas extraordinarias trabajadas en una semana fue de 9,7 y la desviación típica muestral fue de 6,2 horas. Halle el intervalo de confianza al 95 por ciento del número medio de horas extraordinarias trabajadas por cada contable en esta empresa esa semana.

20.15. Un auditor, examinando un total de 820 facturas pendientes de cobro de una empresa, tomó una muestra aleatoria de 60. La media muestral era de 127,43 $ y la desviación típica muestral era de 43,27 $.

a) Halle una estimación de la media poblacional utilizando un método de estimación insesgada.
b) Halle una estimación de la varianza de la media muestral utilizando un método de estimación insesgada.
c) Halle el intervalo de confianza al 90 por ciento de la media poblacional.
d) Un estadístico obtuvo un intervalo de confianza de la media poblacional que iba de 117,43 $ a 137,43 $. ¿Cuál es el contenido probabilístico de este intervalo?

20.16. Un día una organización de consumidores recibió 125 llamadas. Se observó que en una muestra aleatoria de 40 llamadas, el tiempo medio dedicado a dar la información solicitada era de 7,28 minutos y la desviación típica muestral era de 5,32 minutos. Halle el intervalo de confianza al 99 por ciento del tiempo medio por llamada.

20.17. Indique si es verdadera o falsa cada una de las afirmaciones siguientes:

a) Dado un número de miembros de una población y dada una varianza muestral, cuanto mayor es el número de miembros de la muestra, mayor es el intervalo de confianza al 95 por ciento de la media poblacional.
b) Dado un número de miembros de una población y dado un número de miembros de la muestra, cuanto mayor es la varianza muestral, mayor es el intervalo de confianza al 95 por ciento de la media poblacional.
c) Dado un número de miembros de una muestra y dada una varianza muestral, cuanto mayor es el número de miembros de la población, mayor es el intervalo de confianza al 95 por ciento de la media poblacional.
d) Dado un número de miembros de una población, dado un número de miembros de la muestra y dada una varianza muestral, un intervalo de confianza al 95 por ciento de la media poblacional es mayor que un intervalo de confianza al 90 por ciento de la media poblacional.

20.18. Demuestre que nuestra estimación de la varianza de la media muestral puede expresarse de la forma siguiente:

$$\hat{\sigma}_{\bar{x}}^2 = s^2\left(\frac{1}{n} - \frac{1}{N}\right)$$

20.19. Basándose en los datos del ejercicio 20.14, halle el intervalo de confianza al 99 por ciento del número total de horas extraordinarias trabajadas por los contables en la empresa durante la semana de interés.

20.20. Basándose en los datos del ejercicio 20.15, halle el intervalo de confianza al 95 por ciento de la cuantía total de estas 820 facturas pendientes de cobro.

20.21. Basándose en los datos del ejercicio 20.16, halle el intervalo de confianza al 90 por ciento de la cantidad total de tiempo dedicado a responder a estas 125 llamadas.

20.22. Un alto directivo, responsable de un grupo de 120 ejecutivos, está interesado en saber cuánto tiempo dedican en total cada semana estas personas a reuniones internas. Se pide a una muestra aleatoria de 35 ejecutivos que anoten diariamente sus actividades la próxima semana. Cuando se analizan los resultados, se observa que estos miembros de esta muestra dedican un total de 143 horas a reuniones internas. La desviación típica muestral es de 3,1 horas. Halle el intervalo de confianza al 90 por ciento del número total de horas dedicadas a reuniones internas por los 120 ejecutivos durante la semana.

20.23. Una muestra aleatoria simple de 400 universidades de un total de 1.395 contenía 39 que utilizaban el libro de texto *Estadística difícil y aburrida*. Halle el intervalo de confianza al 95 por ciento de la proporción de universidades que utilizaban este libro.

20.24. El decano de una escuela de administración de empresas está considerando la posibilidad de proponer un cambio de los requisitos para obtener el título. Actualmente, los estudiantes tienen que cursar una asignatura de ciencias elegida de una lista de asignaturas posibles. La propuesta es que se sustituya por una asignatura de ecología. La escuela tiene 420 estudiantes. En una muestra aleatoria de 100 estudiantes, 56 han declarado que son contrarios a esta propuesta. Halle el intervalo de confianza al 90 por ciento de la proporción de todos los estudiantes que se oponen al cambio de los requisitos.

20.25. En una residencia universitaria, 257 de los residentes son estudiantes de primer año. En una muestra aleatoria de 120 de ellos, 37 declaran que tienen mucho interés en vivir en la residencia el próximo año. Halle el intervalo de confianza al 95 por ciento de la proporción de estudiantes de primer año de esta residencia que tienen mucho interés en vivir en ella el próximo año.

20.26. Una clase tiene 420 estudiantes. El examen final es optativo: si se hace, la nota puede subir, pero nunca bajar. En una muestra aleatoria de 80 estudiantes, 31 declararon que harían el examen final. Halle el intervalo de confianza al 90 por ciento del número total de estudiantes de esta clase que tienen intención de hacer el examen final.

20.4. Muestreo estratificado

Supongamos que decidimos investigar las opiniones de los estudiantes de nuestro campus universitario sobre algún tema delicado y que puede ser difícil formular las preguntas. Es probable que queramos hacer varias preguntas a cada miembro de la muestra y, dada la limitación de recursos, sólo es posible tomar una muestra bastante pequeña. Probablemente elegiríamos una muestra aleatoria simple, por ejemplo, de 100 estudiantes de una lista de todos los estudiantes del campus. Supongamos, sin embargo, que tras examinar más detenidamente los expedientes de los miembros de la muestra, observamos que sólo dos estudian administración de empresas, aunque la proporción poblacional de estudiantes de administración de empresas es mucho mayor. Nuestro problema en esta fase es doble. En primer lugar, podemos muy bien tener interés en comparar las opiniones de los estudiantes de administración de empresas con las del resto de la población de estudiantes. Eso es difícilmente viable, dada su mínima representación en nuestra muestra. En segundo lugar, podemos sospechar que las opiniones de los estudiantes de administración de empresas sobre esta cuestión serán diferentes de las de sus compañeros. Si fuera así, nos preocupará la fiabilidad de la inferencia basada en una muestra en la que este grupo está seriamente subrepresentado.

Tal vez podríamos consolarnos pensando que, como hemos tomado una muestra aleatoria, cualquier estimador obtenido de la forma habitual será insesgado, por lo que la inferencia resultante, en el sentido estadístico, será estrictamente válida. Sin embargo, basta una breve reflexión para convencernos de que apenas sirve de consuelo. Lo que significa que el estimador es insesgado es que si se repite el método de muestreo muchas veces y se

calcula el estimador, su media será igual al valor poblacional correspondiente. Pero en realidad *no* vamos a repetir el método de muestreo muchas veces. Tenemos que basar nuestras conclusiones en una *única muestra*, y el hecho de que los estudiantes de administración de empresas pudieran haber estado sobrerrepresentados en otras muestras que hubiéramos podido tomar, lo que a largo plazo habría compensado, no es muy útil.

Existe una segunda y tentadora posibilidad que es preferible en muchos sentidos a la de utilizar la muestra original. Podríamos descartar simplemente la muestra original y tomar otra. Si la constitución de la muestra lograda en el segundo intento parece más representativa de la población en general, puede muy bien que sea mejor trabajar con ella. Ahora la dificultad estriba en que el método de muestreo que hemos adoptado —se muestrea la población hasta que se logra una muestra que nos gusta— es muy difícil de formalizar, por lo que los resultados de la muestra son muy difíciles de analizar con alguna validez estadística. Ya no es un muestreo aleatorio simple, por lo que los métodos del apartado 20.3 no son estrictamente válidos.

Afortunadamente, existe un tercer sistema de muestreo para no tener este tipo de problema. Si se sospecha al principio que algunas características identificables de los miembros de la población están relacionadas con el tema de investigación o si algunos subgrupos de la población tienen un interés especial para el investigador, no es necesario (y probablemente no es deseable) conformarse con el muestreo aleatorio simple para seleccionar a los miembros de la muestra. En lugar de eso, se puede dividir la población en subgrupos o *estratos* y tomar una muestra aleatoria simple de cada estrato. El único requisito es que sea posible identificar que cada miembro de la población pertenece a un estrato y sólo a uno.

Muestreo aleatorio estratificado

Supongamos que una población de N individuos puede subdividirse en K grupos mutuamente excluyentes y colectivamente exhaustivos o estratos. Un **muestreo aleatorio estratificado** es la selección de muestras aleatorias simples independientes de cada estrato de la población. Si los K estratos de la población contienen N_1, N_2, ..., N_K miembros, entonces

$$N_1 + N_2 + \cdots + N_K = N$$

No es necesario tomar el mismo número de miembros de la muestra de cada estrato. Sea el número de la muestra n_1, n_2, ..., n_K. En ese caso, el número total de miembros de la muestra es

$$n_1 + n_2 + \cdots + n_K = n$$

La población de estudiantes cuyas ideas se quieren conocer podría dividirse en dos estratos: estudiantes de administración de empresas y resto. También es posible hacer una estratificación menos sencilla. Supongamos que, en algún otro tema, creemos que el sexo y el curso del estudiante (cuarto curso, tercer curso, segundo curso o primer curso) pueden ser relevantes. En ese caso, para satisfacer el requisito de que los estratos sean mutuamente excluyentes y colectivamente exhaustivos, se necesitan ocho estratos: mujeres de cuarto curso, hombres de cuarto curso, etc.

Más adelante en este apartado, nos preguntamos cómo se reparte el esfuerzo de muestreo entre los estratos. Una atractiva posibilidad, empleada a menudo en la práctica, es la *asignación proporcional*: la proporción de miembros de la muestra perteneciente a cualquier estrato es igual que la proporción de miembros de la población perteneciente a ese estrato.

Análisis de los resultados de un muestreo aleatorio estratificado

El análisis de los resultados de una muestra aleatoria estratificada es relativamente sencillo. Sean μ_1, μ_2, ..., μ_K las medias poblacionales de los K estratos y $\bar{x}_1$, $\bar{x}_2$, ..., $\bar{x}_K$ las medias muestrales correspondientes. Consideremos un estrato, por ejemplo, el i-ésimo estrato. Dado que se ha tomado una muestra aleatoria simple en este estrato, la media muestral del estrato es un estimador insesgado de la media poblacional μ_j. Utilizando un método de estimación insesgada de la varianza de la media muestral del estrato, la estimación puntual es

$$\sigma^2_{\bar{x}_j} = \frac{s_j^2}{n_j} \times \frac{(N_j - n_j)}{N_j}$$

donde s_j^2 es la varianza muestral del j-ésimo estrato. Es posible, pues, hacer una inferencia sobre los estratos individuales de la misma forma que en el apartado 20.3.

Generalmente, tienen interés las inferencias sobre la media poblacional μ del conjunto de la población, que es

$$\mu = \frac{N_1\mu_1 + N_2\mu_2 + \cdots + N_K\mu_K}{N} = \frac{1}{N}\sum_{j=1}^{K} N_j\mu_j$$

Una estimación puntual natural es

$$\bar{x}_{st} = \frac{1}{N}\sum_{j=1}^{K} N_j\bar{x}_j$$

Un estimador insesgado de la varianza del estimador de μ se deduce del hecho de que las muestras de cada estrato son independientes entre sí y la estimación puntual es

$$\hat{\sigma}^2_{\bar{x}_{st}} = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2\hat{\sigma}^2_{\bar{x}_j}$$

Las inferencias sobre la media del conjunto de la población pueden basarse en estos resultados.

Estimación de la media poblacional, muestra aleatoria estratificada

Supongamos que se toman muestras aleatorias de n_j individuos de estratos que contienen N_j individuos ($j = 1, 2, ..., K$). Sea

$$\sum_{j=1}^{K} N_j = N \quad \text{y} \quad \sum_{j=1}^{K} n_j = n$$

Sean las medias y las varianzas muestrales de los estratos $\bar{x}_j$ y s_j^2 ($j = 1, 2, ..., K$) y la media del conjunto de la población μ.

1. Un método de estimación insesgada de la media del conjunto de la población μ genera la estimación puntual

$$\bar{x}_{st} = \frac{1}{N}\sum_{j=1}^{K} N_j\bar{x}_j \qquad \textbf{(20.7)}$$

2. Un método de estimación insesgada de la varianza de nuestro estimador de la media del conjunto de la población genera la estimación puntual

$$\hat{\sigma}^2_{\bar{x}_{st}} = \frac{1}{N^2} \sum_{j=1}^{K} N_j^2 \hat{\sigma}^2_{\bar{x}_j} \tag{20.8}$$

donde

$$\hat{\sigma}^2_{\bar{x}_j} = \frac{s_j^2}{n_j} \times \frac{(N_j - n_j)}{N_j} \tag{20.9}$$

3. Siempre que el tamaño de la muestra es grande, se obtienen **intervalos de confianza** al 100(1 − α)% **de la media poblacional de muestras aleatorias estratificadas** de la forma siguiente:

$$\bar{x}_{st} - z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}} < \mu < \bar{x}_{st} + z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}} \tag{20.10}$$

EJEMPLO 20.4. Cadena de restaurantes (estimación)

Una cadena de restaurantes tiene 60 en Illinois, 50 en Indiana y 45 en Ohio. La dirección está considerando la posibilidad de añadir un nuevo plato a su menú. Para averiguar cuál es la demanda probable de este plato, se introduce en el menú de muestras aleatorias de 20 restaurantes de Illinois, 10 de Indiana y 9 de Ohio. Utilizando los subíndices 1, 2 y 3 para representar Illinois, Indiana y Ohio, respectivamente, las medias y las desviaciones típicas muestrales del número de pedidos de este plato por restaurante en los tres estados en una semana es

$$\bar{x}_1 = 21{,}2 \qquad s_1 = 12{,}8$$
$$\bar{x}_2 = 13{,}3 \qquad s_2 = 11{,}4$$
$$\bar{x}_3 = 26{,}1 \qquad s_3 = 9{,}2$$

Estime el número medio de pedidos semanales por restaurante, μ, en todos los restaurantes de esta cadena.

Solución

Se sabe que

$$N_1 = 60 \qquad N_2 = 50 \qquad N_3 = 45 \qquad N = 155$$
$$n_1 = 12 \qquad n_2 = 10 \qquad n_3 = 9 \qquad n = 31$$

Nuestra estimación de la media poblacional es

$$\bar{x}_{st} = \frac{1}{N} \sum_{j=1}^{K} N_j \bar{x}_j = \frac{(60)(21{,}2) + (50)(13{,}3) + (45)(26{,}1)}{155} = 20{,}1$$

Por lo tanto, el número medio estimado de pedidos semanales por restaurante es 20,1.

El paso siguiente es calcular las cantidades

$$\hat{\sigma}^2_{\bar{x}_1} = \frac{s_1^2}{n_1} \times \frac{(N_1 - n_1)}{N_1} = \frac{(12{,}8)^2}{12} \times \frac{48}{60} = 10{,}923$$

$$\hat{\sigma}^2_{\bar{x}_2} = \frac{s_2^2}{n_2} \times \frac{(N_2 - n_2)}{N_2} = \frac{(11{,}4)^2}{10} \times \frac{40}{50} = 10{,}397$$

$$\hat{\sigma}^2_{\bar{x}_3} = \frac{s_3^2}{n_3} \times \frac{(N_3 - n_3)}{N_3} = \frac{(9{,}2)^2}{9} \times \frac{36}{45} = 7{,}524$$

Estas cantidades, junto con las medias muestrales de cada estrato, pueden utilizarse para calcular intervalos de confianza de las medias poblacionales de los tres estratos, exactamente como en el ejemplo 20.1 (aunque en este caso el tamaño de la muestra es demasiado pequeño por comodidad). Centramos la atención en la media del conjunto de la población. Para obtener intervalos de confianza para esta cantidad,

$$\hat{\sigma}^2_{\bar{x}_{st}} = \frac{1}{N^2} \sum_{j=1}^{K} N_j^2 \hat{\sigma}^2_{\bar{x}_j}$$

$$= \frac{(60)^2(10{,}923) + (50)^2(10{,}397) + (45)^2(7{,}524)}{(155)^2} = 3{,}353$$

y, tomando la raíz cuadrada,

$$\hat{\sigma}_{\bar{x}_{st}} = 1{,}83$$

Por lo tanto, el intervalo de confianza al 95 por ciento del número medio de pedidos por restaurante realizados en una semana es

$$20{,}1 - (1{,}96)(1{,}83) < \mu < 20{,}1 + (1{,}96)(1{,}83)$$

o sea

$$16{,}5 < \mu < 23{,}7$$

El intervalo de confianza al 95 por ciento va de 16,5 a 23,7 pedidos por restaurante.

Dado que el total poblacional es el producto de la media poblacional y el número de miembros de la población, estos métodos pueden modificarse fácilmente para poder estimarlo.

Estimación del total poblacional, muestra aleatoria estratificada

Supongamos que se toman muestras aleatorias de n_j individuos de estratos que contienen N_j individuos ($j = 1, 2, ..., K$) y que la cantidad que quiere estimarse es el total poblacional, $N\mu$.

1. Un método de estimación insesgada de $N\mu$ genera la estimación puntual

$$N\bar{x}_{st} = \sum_{j=1}^{K} N_j \bar{x}_j \qquad \textbf{(20.11)}$$

2. Un método de estimación insesgada de la varianza de nuestro estimador del total poblacional genera la estimación

$$N^2\hat{\sigma}^2_{\bar{x}_{st}} = \sum_{j=1}^{K} N_j^2 \hat{\sigma}^2_{\bar{x}_j} \qquad \textbf{(20.12)}$$

3. Siempre que el tamaño de la muestra es grande, se obtienen **intervalos de confianza** al 100(1 − α)% **del total poblacional de muestras aleatorias estratificadas** de la forma siguiente:

$$N\bar{x}_{st} - z_{\alpha/2}N\hat{\sigma}_{\bar{x}_{st}} < N\mu < N\bar{x}_{st} + z_{\alpha/2}N\hat{\sigma}_{\bar{x}_{st}} \qquad \textbf{(20.13)}$$

EJEMPLO 20.5. Número anual total de matriculados en estadística para los negocios (estimación)

De las 1.395 universidades que hay en un país, 364 son escuelas universitarias, en las que la duración de los estudios es de 2 años, y 1.031 son facultades, en las que la duración de los estudios es de 4 años. Se toma una muestra aleatoria de 40 escuelas universitarias y una muestra aleatoria simple independiente de 60 facultades. La tabla adjunta muestra las medias muestrales y las desviaciones típicas muestrales del número de estudiantes matriculados el año pasado en la asignatura de estadística para los negocios. Estime el número total anual de matriculados en esa asignatura.

	Escuelas universitarias	**Facultades**
Media	154,3	411,8
Desviación típica	87,3	219,9

Solución

Se sabe que

$$N_1 = 364 \quad n_1 = 40 \quad \bar{x}_1 = 154{,}3 \quad s_1 = 87{,}3$$
$$\mathrm{N}_2 = 1.031 \quad \mathrm{n}_2 = 60 \quad \bar{\mathrm{x}}_2 = 411{,}8 \quad \mathrm{s}_2 = 219{,}9$$

Nuestra estimación del total poblacional es

$$N\bar{x}_{st} = \sum_{j=1}^{K} N_j\bar{x}_j = (364)(154{,}3) + (1.031)(411{,}8) = 480.731$$

A continuación,

$$\hat{\sigma}^2_{\bar{x}_1} = \frac{s_1^2}{n_1} \times \frac{(N_1 - n_1)}{N_1} = \frac{(87{,}3)^2}{40} \times \frac{324}{364} = 169{,}59$$

$$\hat{\sigma}^2_{\bar{x}_2} = \frac{s_2^2}{n_2} \times \frac{(N_2 - n_2)}{N_2} = \frac{(219{,}9)^2}{60} \times \frac{971}{1.031} = 759{,}03$$

Por último,

$$N^2\hat{\sigma}^2_{\bar{x}_{st}} = \sum_{j=1}^{K} N_j^2\hat{\sigma}^2_{\bar{x}_{st}} = (364)^2(169{,}59) + (1.031)^2(759{,}03) = 820.289.284$$

y, tomando la raíz cuadrada,

$$N\hat{\sigma}_{\bar{x}_{st}}^2 = 28.797$$

En el caso del intervalo de confianza al 95 por ciento,

$$z_{\alpha/2} = z_{0,025} = 1{,}96$$

El intervalo al 95 por ciento que buscamos es, pues,

$$480.731 - (1{,}96)(28.797) < N\mu < 480.731 + (1{,}96)(28.797)$$

o sea

$$424.289 < N\mu < 537.173$$

Por lo tanto, nuestro intervalo de confianza al 95 por ciento va de 424.289 a 537.173 estudiantes matriculados.

Consideremos ahora el problema de estimar una proporción poblacional basándonos en una muestra aleatoria estratificada. Sean P_1, P_2, ..., P_K las proporciones poblacionales de los K estratos y $\hat{p}_1$, $\hat{p}_2$, ..., $\hat{p}_K$ las proporciones muestrales correspondientes. Si P representa la proporción de la población total, su estimación se basa en el hecho de que

$$P = \frac{N_1P_1 + N_2P_2 + \cdots + N_KP_K}{N} = \frac{1}{N}\sum_{j=1}^{K} N_jP_j$$

A continuación, se muestran los métodos para estimar la proporción poblacional a partir de una muestra aleatoria estratificada.

Estimación de la proporción poblacional, muestra aleatoria estratificada

Supongamos que se toman muestras aleatorias de n_j individuos de estratos que contienen N_j individuos ($j = 1, 2, ..., K$). Sea P_j la proporción poblacional y $\hat{p}_j$ la proporción muestral en el j-ésimo estrato de los que poseen una determinada característica. Si P es la proporción de la población total:

1. Un método de estimación insesgada de P genera

$$\hat{p}_{st} = \frac{1}{N}\sum_{j=1}^{K} N_j\hat{p}_j \qquad \textbf{(20.14)}$$

2. Un método de estimación insesgada de la varianza de nuestro estimador de la proporción de la población total es

$$\hat{\sigma}_{\hat{p}_{st}}^2 = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2\hat{\sigma}_{\hat{p}_j}^2 \qquad \textbf{(20.15)}$$

donde

$$\hat{\sigma}_{\hat{p}_j}^2 = \frac{\hat{p}_j(1-\hat{p}_j)}{n_j - 1} \times \frac{(N_j - n_j)}{N_j} \qquad \textbf{(20.16)}$$

es la estimación de la varianza de la proporción muestral del j-ésimo estrato.

3. Siempre que el tamaño de la muestra es grande, se obtienen **intervalos de confianza** al 100(1 − α)% **de la proporción poblacional de muestras aleatorias estratificadas** de la forma siguiente:

$$\hat{p}_{st} - z_{\alpha/2}\hat{\sigma}_{\hat{p}_{st}} < P < \hat{p}_{st} + z_{\alpha/2}\hat{\sigma}_{\hat{p}_{st}} \qquad \textbf{(20.17)}$$

EJEMPLO 20.6. Estadística impartida en los departamentos de economía (estimación)

Supongamos que en el estudio del ejemplo 20.5 observamos que la asignatura de estadística para los negocios se imparte en el departamento de economía de 7 escuelas universitarias y de 13 facultades de la muestra. Estime la proporción de todas las universidades en las que se imparte esta asignatura en el departamento de economía.

Solución

Se sabe que

$$N_1 = 364 \qquad n_1 = 40 \qquad \hat{p}_1 = \frac{7}{40} = 0{,}175$$

$$N_2 = 1.031 \qquad n_2 = 60 \qquad \hat{p}_2 = \frac{13}{60} = 0{,}217$$

Nuestra estimación de la proporción poblacional es

$$\hat{p}_{st} = \frac{1}{N}\sum_{j=1}^{K} N_j\hat{p}_j = \frac{(364)(0{,}175) + (1.031)(0{,}217)}{1.395} = 0{,}206$$

Por lo tanto, se estima que en el 20,6 por ciento de todas las escuelas universitarias el departamento de economía imparte la asignatura.

A continuación,

$$\hat{\sigma}^2_{\hat{p}_1} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1 - 1} \times \frac{(N_1 - n_1)}{N_1} = \frac{(0{,}175)(0{,}825)}{39} \times \frac{324}{364} = 0{,}003295$$

$$\hat{\sigma}^2_{\hat{p}_2} = \frac{\hat{p}_2(1-\hat{p}_2)}{n_2 - 1} \times \frac{(N_2 - n_2)}{N_2} = \frac{(0{,}217)(0{,}783)}{59} \times \frac{971}{1.031} = 0{,}002712$$

Estos valores, junto con las proporciones muestrales de cada estrato, pueden utilizarse para calcular intervalos de confianza de las proporciones de la población de los dos estratos, exactamente como en el ejemplo 20.3. Aquí centramos la atención en la estimación por intervalos de la proporción de la población total, para la que

$$\hat{\sigma}^2_{\hat{p}_{st}} = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2\hat{\sigma}^2_{\hat{p}_j} = \frac{(364)^2(0{,}003295) + (1.031)^2(0{,}002712)}{(1.395)^2} = 0{,}001706$$

por lo que, tomando la raíz cuadrada, tenemos que

$$\hat{\sigma}_{\hat{p}_{st}} = 0{,}0413$$

En el caso del intervalo de confianza al 90 por ciento,

$$z_{\alpha/2} = z_{0,05} = 1,645$$

y el intervalo de confianza al 90 por ciento de la proporción poblacional de una muestra aleatoria estratificada es

$$(0,206) - (1,645)(0,0413) < P < (0,206) + (1,645)(0,0413)$$
$$0,138 < \mathrm{P} < 0,274$$

Este intervalo va del 13,8 al 27,4 por ciento de todas las universidades.

Afijación del esfuerzo muestral a los distintos estratos

Queda por analizar la cuestión del reparto del esfuerzo muestral entre los estratos. Suponiendo que se selecciona un total de n miembros, ¿cuántas de estas observaciones muestrales deben asignarse a cada estrato? En realidad, el estudio en cuestión puede tener muchos objetivos, lo cual significa que no existe una clara respuesta. No obstante, es posible especificar unos criterios de elección que el investigador debe tener presentes. Si se sabe poco o nada de antemano sobre la población y si no hay ningún requisito para la producción de información acerca de estratos poco poblados, es lógico elegir una *afijación proporcional*.

Afijación proporcional: tamaño de la muestra

La proporción de miembros de la muestra que hay en un estrato es igual que la proporción de miembros de la población que hay en ese estrato. Por lo tanto, considerando el j-ésimo estrato,

$$\frac{n_j}{n} = \frac{N_j}{N} \tag{20.18}$$

por lo que el **tamaño de la muestra del *j*-ésimo estrato utilizando la afijación proporcional** es

$$n_j = \frac{N_j}{N} \times n \tag{20.19}$$

Este mecanismo de afijación intuitivamente razonable se emplea frecuentemente y permite, por lo general, realizar un análisis satisfactorio. Obsérvese que en el ejemplo 20.4 utilizamos la afijación proporcional. Dividimos un total de $N = 155$ restaurantes en tres estratos (Illinois, Indiana y Ohio). Seleccionamos una muestra de $n = 31$, siendo

$$n_1 = \frac{60}{155} \times 31 = 12 \qquad n_2 = \frac{50}{155} \times 31 = 10 \qquad n_3 = \frac{45}{155} \times 31 = 9$$

A veces la utilización estricta de la afijación proporcional produce relativamente pocas observaciones en los estratos que le interesan especialmente al investigador. En ese caso, la inferencia sobre los parámetros poblacionales de estos estratos podría ser bastante imprecisa. En estas circunstancias, puede ser preferible afijar más observaciones a esos estratos que las que dicta la afijación proporcional. En los ejemplos 20.5 y 20.6, 364 de las

1.395 universidades son escuelas universitarias y se toma una muestra de 100 observaciones. Si se hubiera utilizado la afijación proporcional, el número de escuelas incluidas en la muestra habría sido

$$n_1 = \frac{N_1}{N} \times n = \frac{364}{1.395} \times 100 = 26$$

Dado que al editor le interesaba especialmente obtener información sobre este mercado, se pensó que no sería adecuada una muestra de 26 observaciones solamente. Por esta razón, 40 de las 100 observaciones muestrales se afijaron a este estrato.

Si el único objetivo de un estudio es estimar con la mayor precisión posible un parámetro relativo al conjunto de la población, como la media, el total o la proporción, y si se tiene bastante información sobre la población, es posible establecer una *afijación óptima*.

Afijación óptima: tamaño de la muestra del *j*-ésimo estrato, media o total del conjunto de la población

Si lo que se necesita es estimar una media o un total del conjunto de una población y si las varianzas poblacionales de los estratos individuales se representan por medio de σ_j^2, puede demostrarse que los estimadores más precisos se obtienen con la afijación óptima. El **tamaño de la muestra del *j*-ésimo estrato utilizando la afijación óptima** es

$$n_j = \frac{N_j \sigma_j}{\sum_{i=1}^{K} N_i \sigma_i} \times n \qquad \textbf{(20.20)}$$

Esta fórmula es razonable intuitivamente. Comparada con la afijación proporcional, asigna relativamente más esfuerzo muestral a los estratos en los que la varianza poblacional es mayor. Es decir, se necesita una muestra de mayor tamaño donde la variabilidad poblacional es mayor. Así, en el ejemplo 20.4, en el que hemos utilizado la afijación proporcional, si las diferencias observadas en las desviaciones típicas muestrales reflejaran correctamente las diferencias que existen en las cantidades poblacionales, habría sido preferible tomar menos observaciones en el tercer estrato y más en el primero.

El uso de la ecuación 20.20 plantea inmediatamente una objeción. Requiere conocer las desviaciones típicas poblacionales, σ_j, mientras que antes de que se tome la muestra, a menudo ni siquiera se dispone de estimaciones de estos valores que merezcan la pena. Esta cuestión se analiza en el último apartado del capítulo.

A continuación, se examina el tamaño de la muestra necesario en la afijación óptima correspondiente a una proporción poblacional.

Afijación óptima: tamaño de la muestra del *j*-ésimo estrato, proporción poblacional

Para estimar la proporción de la población total, se obtienen estimadores con la menor varianza posible por medio de una afijación óptima. El **tamaño de la muestra del *j*-ésimo estrato de la proporción poblacional utilizando la afijación óptima** es

$$n_j = \frac{N_j \sqrt{P_j(1 - P_j)}}{\sum_{i=1}^{K} N_i \sqrt{P_i(1 - P_i)}} \times n \qquad \textbf{(20.21)}$$

Esta fórmula, en comparación con la afijación proporcional, asigna más observaciones muestrales a los estratos en los que las verdaderas proporciones poblacionales son más cercanas a 0,5, pues si una proporción es cercana a 0 o a 1, puede saberse con bastante seguridad con una muestra relativamente pequeña. La dificultad que plantea el uso de la ecuación 20.21 estriba en que implica las proporciones desconocidas P_j para ($j = 1, 2, ..., K$), que son las propias cantidades que el estudio pretende estimar.

No obstante, a veces la información anterior sobre la población puede permitir hacerse al menos una idea aproximada de qué estratos tienen proporciones más cercanas a 0,5. En el ejemplo 20.6, las proporciones muestrales sugieren que el número de escuelas universitarias que hay en la muestra debería haber sido menor que el número resultante de la afijación proporcional. Se llega a la misma conclusión en este estudio cuando se comparan las desviaciones típicas muestrales del ejemplo 20.5 con la ecuación 20.20. A pesar de eso, se decidió incluir en la nuestra *más* escuelas universitarias en lugar de menos. La razón era que en este estudio el editor quería tener información fiable tanto sobre el mercado de escuelas universitarias como sobre el de facultades.

Esta ilustración es un ejemplo de una importante cuestión. Aunque la división del esfuerzo muestral que sugieren las ecuaciones 20.20 y 20.21 a menudo se denomina *afijación óptima*, sólo es óptima con respecto al estricto criterio de la estimación eficiente de los parámetros correspondientes al conjunto de la población. A menudo, los estudios tienen objetivos más amplios que ése, en cuyo caso puede muy bien ser razonable no utilizar la afijación óptima.

EJERCICIOS

Ejercicios aplicados

20.27. Una pequeña ciudad contiene un total de 1.800 hogares. La ciudad está dividida en tres distritos, que contienen 820, 540 y 440 hogares, respectivamente. Una muestra aleatoria estratificada de 300 hogares contiene 120, 90 y 90 hogares, respectivamente, de estos tres distritos. Se pide a los miembros de la muestra que estimen su factura total de electricidad consumida en los meses de invierno. Las respectivas medias muestrales son 290 $, 352 $ y 427 $ y las respectivas desviaciones típicas muestrales son 47 $, 61 $ y 93 $.

a) Utilice un método de estimación insesgada para estimar la factura media de electricidad consumida en los meses de invierno por todos los hogares de esta ciudad.

b) Utilice un método de estimación insesgada para estimar la varianza del estimador del apartado (a).

c) Halle el intervalo de confianza al 95 por ciento de la media poblacional de las facturas de electricidad consumida en invierno por los hogares de esta ciudad.

20.28. Una universidad tiene 152 profesores ayudantes, 127 titulares y 208 catedráticos. Las autoridades universitarias están investigando la cantidad de tiempo que dedican estos profesores a reuniones en un cuatrimestre. Se pide a muestras aleatorias de 40 profesores ayudantes, 40 titulares y 50 catedráticos que lleven la cuenta del tiempo que dedican a reuniones en un cuatrimestre. Las medias muestrales son 27,6 horas en el caso de los profesores ayudantes, 39,2 en el de los titulares y 43,3 en el de los catedráticos. Las desviaciones típicas muestrales son 7,1 horas en el caso de los profesores ayudantes, 9,9 en el de los titulares y 12,3 en el de los catedráticos.

a) Halle un intervalo de confianza al 90 por ciento del tiempo medio dedicado a reuniones por los catedráticos de esta universidad en un cuatrimestre.

b) Utilice un método de estimación insesgada para estimar el tiempo medio dedicado a reuniones por todos los profesores de esta universidad en un cuatrimestre.

c) Halle intervalos de confianza del 90 y el 95 por ciento del tiempo medio dedicado a reu-

niones por todos los profesores de esta universidad en un cuatrimestre.

20.29. Una empresa de autobuses está planificando una nueva ruta para dar servicio a cuatro barrios. Se toman muestras aleatorias de hogares de cada barrio y se pide a los miembros de las muestras que valoren en una escala de 1 (totalmente en contra) a 5 (totalmente a favor) su reacción al servicio propuesto. La tabla adjunta muestra los resultados.

	Barrio 1	Barrio 2	Barrio 3	Barrio 4
N_i	240	190	350	280
n_i	40	40	40	40
$\bar{x}_i$	2,5	3,6	3,9	2,8
s_i	0,8	0,9	1,2	0,7

a) Halle un intervalo de confianza al 90 por ciento de la reacción media de los hogares de la barrio 1.
b) Utilice un método de estimación insesgada para estimar la reacción media de todos los hogares a la nueva ruta.
c) Halle intervalos de confianza al 90 y al 95 por ciento de la reacción media de todos los hogares a la nueva ruta.

20.30. En una muestra aleatoria estratificada de estudiantes de una pequeña universidad, se pide a los miembros de la muestra que valoren en una escala de 1 (pocas) a 5 (muchas) las oportunidades para realizar actividades extracurriculares. La tabla adjunta muestra los resultados.

	Estudiantes de primer y segundo año	Estudiantes de tercer y cuarto año
N_i	632	529
n_i	50	50
$\bar{x}_i$	3,12	3,37
s_i	1,04	0,86

a) Halle el intervalo de confianza al 95 por ciento de la valoración media que harían todos los estudiantes de primer y segundo año de este campus.
b) Halle el intervalo de confianza al 95 por ciento de la valoración media que harían todos los estudiantes de tercer y cuarto año de este campus.
c) Halle el intervalo de confianza al 95 por ciento de la valoración media que harían todos los estudiantes de este campus.

20.31. Vuelva al ejercicio 20.28.
a) Halle el intervalo de confianza al 90 por ciento de la cantidad total de tiempo dedicada a reuniones por todos los profesores catedráticos de esta universidad en un cuatrimestre.
b) Halle el intervalo de confianza al 90 por ciento de la cantidad total de tiempo dedicada a reuniones por todos los profesores de esta universidad en un cuatrimestre.

20.32. Una empresa tiene tres divisiones y los auditores están intentado estimar la cantidad total de facturas pendientes de cobro de la empresa. Se toman muestras aleatorias de estas facturas en cada una de las tres divisiones y se obtienen los resultados que muestra la tabla.

	División 1	División 2	División 3
N_i	120	150	180
n_i	40	45	50
$\bar{x}_i$	237 $	198 $	131 $
s_i	93 $	64 $	47 $

a) Utilice un método de estimación insesgada para hallar una estimación puntual del valor total de todas las facturas pendientes de cobro de esta empresa.
b) Halle el intervalo de confianza al 95 por ciento del valor total de todas las facturas pendientes de cobro de esta empresa.

20.33. De las 1.395 universidades que hay en un país, 364 son escuelas universitarias. En una muestra aleatoria de 40 escuelas universitarias, se observa que en 10 de ellas se utiliza el libro de texto *La estadística puede ser divertida*. En otra muestra aleatoria de 60 facultades, se utiliza este libro de texto en 8 de ellas.
a) Estime la proporción de todas las universidades que utilizan este libro de texto empleando un método de estimación insesgada.
b) Halle el intervalo de confianza al 95 por ciento de la proporción de todas las escuelas universitarias que utilizan este libro de texto.

20.34. Una consultora ha desarrollado un curso breve sobre métodos modernos de predicción para ejecutivos de empresa. Al primer curso han asistido 150 ejecutivos. Con la información suministrada por ellos, se ha llegado a la conclusión de que las cualificaciones técnicas de 100 asistentes al curso eran más que suficientes para seguir la materia, mientras que las de los 50 restantes no lo eran. Después de terminar el

curso, se han enviado cuestionarios a muestras aleatorias independientes de 25 personas de cada uno de estos grupos para obtener información con el fin de mejorar la presentación de los cursos posteriores. Seis del grupo más cualificado y 14 del grupo menos cualificado han indicado que creen que el curso es demasiado teórico.

a) Estime la proporción de todos los asistentes al curso que tienen esta opinión utilizando un método de estimación insesgada.

b) Halle intervalos de confianza al 90 por ciento y al 95 por ciento de esta proporción poblacional.

20.35. Una universidad tiene 152 profesores ayudantes, 127 titulares y 208 catedráticos. Un periodista del periódico estudiantil tiene interés en saber si los profesores están realmente en su despacho a las horas indicadas. Decide investigar muestras de 40 profesores ayudantes, 40 titulares y 50 catedráticos. Envía estudiantes voluntarios a los despachos de los miembros de la muestra durante las horas indicadas. Se observa que 31 de los profesores ayudantes, 29 de los titulares y 34 de los catedráticos están realmente en su despacho a esas horas.

a) Utilice un método de estimación insesgada para hallar una estimación puntual de la proporción de todos los profesores que están en su despacho a las horas indicadas.

b) Halle el intervalo de confianza al 90 por ciento y al 95 por ciento de la proporción de todos los profesores que están en su despacho a las horas indicadas.

20.36. Vuelva al ejercicio 20.28. Si se toma una muestra total de 130 profesores, averigüe cuántos son catedráticos utilizando cada uno de los sistemas siguientes:

a) Afijación proporcional.

b) Afijación óptima, suponiendo que las desviaciones típicas poblacionales de los estratos son iguales que los valores muestrales correspondientes.

20.37. Vuelva a los datos del ejercicio 20.29. Si se toma una muestra total de 160 hogares, averigüe cuántos deben ser del barrio 1 utilizando cada uno de los sistemas siguientes:

a) Afijación proporcional.

b) Afijación óptima, suponiendo que las desviaciones típicas poblacionales de los estratos son iguales que los valores muestrales correspondientes.

20.38. Vuelva al ejercicio 20.30. Si se toma una muestra total de 100 estudiantes, averigüe cuántos son estudiantes de primero y de segundo año utilizando cada uno de los sistemas siguientes:

a) Afijación proporcional.

b) Afijación óptima, suponiendo que las desviaciones típicas poblacionales de los estratos son iguales que los valores muestrales correspondientes.

20.39. Vuelva a los datos del ejercicio 20.32. Si se toma una muestra total de 135 facturas pendientes de cobro, averigüe cuántas deben ser de la división 1 utilizando cada uno de los sistemas siguientes:

a) Afijación proporcional.

b) Afijación óptima, suponiendo que las desviaciones típicas poblacionales de los estratos son iguales que los valores muestrales correspondientes.

20.40. Vuelva a los datos del ejemplo 20.5. Si se toma una muestra total de 100 universidades, averigüe cuántas serán probablemente escuelas universitarias (en vez de facultades) por medio de los siguientes sistemas:

a) Afijación proporcional.

b) Afijación óptima, suponiendo que las desviaciones típicas poblacionales de los estratos son iguales que los valores muestrales correspondientes.

20.5. Elección del tamaño de la muestra

Un importante aspecto de la planificación de cualquier estudio es la elección del número de miembros de la muestra. Hay varios factores que pueden ser relevantes. Si se piensa que con el método utilizado para contactar con los miembros de la muestra probablemente la tasa de falta de respuesta será alta, debe tenerse en cuenta esta posibilidad. En muchos casos, los recursos de los que dispone el investigador, en lo que se refiere a tiempo y dinero,

limitan los resultados. Sin embargo, en este apartado dejamos de lado estas consideraciones y relacionamos el tamaño de la muestra con las varianzas de los estimadores de los parámetros poblacionales y, por consiguiente, con la amplitud de los intervalos de confianza resultantes.

Tamaño de la muestra para el muestreo aleatorio simple: estimación de la media o el total poblacional

Consideremos el problema de estimar la media poblacional a partir de una muestra aleatoria simple de n observaciones. Si la variable aleatoria $\bar{x}$ representa la media muestral, en el Capítulo 7 vimos que la varianza de esta variable aleatoria es

$$\text{Var}(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \times \frac{(N-n)}{(N-1)}$$

Si se conoce la varianza poblacional σ^2, resolviendo la ecuación $\text{Var}(\bar{x})$, podemos hallar el tamaño de la muestra, n, que se necesita para lograr cualquier valor específico de $s_{\bar{x}}^2$ para la varianza de la media muestral. Existen métodos parecidos si la cantidad que nos interesa es el total poblacional.

Tamaño de la muestra: media o total de la población, muestreo aleatorio simple

Consideremos la estimación de la media de una población de N miembros, que tiene la varianza σ^2. Si se especifica la varianza deseada, $s_{\bar{x}}^2$, de la media muestral, **el tamaño de la muestra necesario para estimar la media poblacional por medio de un muestreo aleatorio simple** es

$$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{x}}^2 + \sigma^2} \qquad \textbf{(20.22)}$$

1. A menudo es útil especificar directamente la amplitud de los intervalos de confianza de la media poblacional en lugar de $\sigma_{\bar{x}}^2$. Eso se logra fácilmente, ya que, por ejemplo, el intervalo de confianza al 95 por ciento de la media poblacional tiene una amplitud de aproximadamente $1{,}96\sigma_{\bar{x}}$ a cada lado de la media muestral.
2. Si el objeto de interés es el **total poblacional**, la varianza del estimador muestral de esta cantidad es $N^2\sigma_{\bar{x}}^2$ y el intervalo de confianza al 95 por ciento de ella tiene una amplitud de aproximadamente $1{,}96N\sigma_{\bar{x}}$ a cada lado de la $N\bar{x}$.

Una dificultad obvia que plantea el uso práctico de la ecuación 20.22 es que implica la varianza poblacional, σ^2, que normalmente no se conoce. Sin embargo, un investigador a menudo tiene una idea aproximada de cuál es el valor de esta cantidad. A veces la varianza poblacional puede estimarse a partir de una muestra preliminar de la población.

EJEMPLO 20.7. Créditos hipotecarios (tamaño de la muestra)

Supongamos, como en el ejemplo 20.1, que en una ciudad se solicitaron 1.118 créditos hipotecarios el año pasado y que se toma una muestra aleatoria simple para estimar la cantidad media de créditos hipotecarios. Basándose en estudios anteriores realizados con esas poblaciones, se estima que la desviación típica poblacional es de 20.000 $ aproximadamente. El intervalo de confianza al 95 por ciento de la media poblacional

debe tener una amplitud de 4.000 $ a cada lado de la media muestral. ¿Cuántas observaciones muestrales se necesitan para lograr este objetivo?

Solución

En primer lugar,

$$N = 1.118 \qquad \sigma = 20.000 \qquad 1{,}96\sigma_{\bar{x}} = 4.000$$

El tamaño de la muestra necesario es, pues,

$$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{x}}^2 + \sigma^2} = \frac{(1.118)(20.000)^2}{(1.117)(2.041)^2 + (20.000)^2} = 88{,}5$$

Por lo tanto, debería ser suficiente una muestra aleatoria simple de 89 observaciones para alcanzar nuestro objetivo.

Tamaño de la muestra para el muestreo aleatorio simple: estimación de la proporción poblacional

Consideremos un muestreo aleatorio simple para estimar una proporción poblacional P. Recuérdese que ya hemos visto antes en este libro que

$$\text{Var}(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{P(1-P)}{n} \times \frac{(N-n)}{(N-1)}$$

Despejando n, tenemos el tamaño de la muestra de las ecuaciones 20.23 y 20.24.

Tamaño de la muestra: proporción poblacional, muestreo aleatorio simple

Consideremos la estimación de la proporción P de individuos de una población de tamaño N que poseen un cierto atributo. Si se especifica la varianza deseada, $\sigma_{\hat{p}}^2$, de la proporción muestral, el tamaño de la muestra necesario para estimar la proporción poblacional mediante un muestreo aleatorio simple es

$$n = \frac{NP(1-P)}{(N-1)\sigma_{\hat{p}}^2 + P(1-P)} \qquad \textbf{(20.23)}$$

El mayor valor posible de esta expresión, cualquiera que sea el valor de P, es

$$n_{\max} = \frac{0{,}25N}{(N-1)\sigma_{\hat{p}}^2 + 0{,}25} \qquad \textbf{(20.24)}$$

El intervalo de confianza al 95 por ciento de la proporción poblacional debe tener una amplitud de aproximadamente 1,96 $\sigma_{\hat{p}}$ a cada lado de la proporción muestral.

EJEMPLO 20.8. Estudio sobre la estadística en las universidades (tamaño de la muestra)

Supongamos, al igual que en el ejemplo 20.3, que se toma una muestra aleatoria simple de 1.395 universidades que hay en un país para estimar la proporción en la que la asignatura de estadística para los negocios es anual. Cualquiera que sea la verdadera proporción, el intervalo de confianza al 95 por ciento no debe tener una amplitud de más de 0,04 a cada lado de la proporción muestral. ¿Cuántas observaciones muestrales deben tomarse?

Solución

Sabemos que

$$1{,}96\sigma_{\hat{p}} = 0{,}04$$

o sea

$$\sigma_{\hat{p}} = 0{,}0204$$

El tamaño de la muestra necesario es, pues,

$$n_{\max} = \frac{0{,}25N}{(N-1)\sigma_{\hat{p}}^2 + 0{,}25} = \frac{(0{,}25)(1.395)}{(1.394)(0{,}0204)^2 + 0{,}25} = 420{,}1$$

Por lo tanto, se necesita una muestra de 421 observaciones.

Tamaño de la muestra para un muestreo aleatorio estratificado con un grado de precisión especificado

También es posible obtener fórmulas para hallar el tamaño de la muestra necesario para lograr un grado de precisión especificado cuando se utiliza el muestreo aleatorio estratificado.

Varianza del estimador de la media poblacional, muestreo estratificado

Sea la variable aleatoria $\bar{X}_{st}$ el **estimador de la media poblacional obtenido mediante un muestreo estratificado** y sea $\bar{X}_j$ ($j = 1, 2, ..., K$) las medias muestrales de los estratos individuales. Dado que

$$\bar{X}_{st} = \frac{1}{N}\sum_{j=1}^{K} N_j \bar{X}_j \qquad \textbf{(20.25)}$$

se deduce que la **varianza** de $\bar{X}_{st}$ es

$$\mathrm{Var}(\bar{X}_{st}) = \sigma_{\bar{X}_{st}}^2 = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2 \,\mathrm{Var}(\bar{X}_j) = \frac{1}{N^2}\sum_{j=1}^{K} N_j^2 \frac{\sigma_j^2}{n_j} \times \frac{(N_j - n_j)}{N_j - 1} \qquad \textbf{(20.26)}$$

donde las σ_j^2 son las varianzas poblacionales de los K estratos.

Ahora puede utilizarse la ecuación 20.26, dada cualquier elección de n_1, n_2, ..., n_K, para hallar la varianza correspondiente del estimador de la media poblacional. Sin embargo, el

tamaño total de la muestra, n, necesario para obtener un determinado valor de esta varianza dependerá de la manera en que se repartan las observaciones muestrales entre los estratos. En el apartado 20.4 hemos analizado dos métodos que se emplean frecuentemente, la afijación proporcional y la afijación óptima. En cualquiera de los dos casos, sustituyendo los n_j en la ecuación 20.26, podemos resolver la ecuación resultante y hallar el tamaño de la muestra, n. Los resultados se indican en las ecuaciones 20.27 y 20.28.

Tamaño total de la muestra para estimar la media global (varianzas poblacionales de los estratos especificadas), muestreo aleatorio estratificado

Supongamos que se subdivide una población de N miembros en K estratos que contienen N_1, N_2, ..., N_K miembros. Sea σ_j^2 la varianza poblacional del j-ésimo estrato y supongamos que se desea obtener una **estimación de la media del conjunto de la población**. Si se especifica la varianza deseada, $\sigma^2_{\bar{X}_{st}}$, del estimador muestral, el tamaño total de la muestra necesario, n, se obtiene de la forma siguiente:

1. **Afijación proporcional:**

$$n = \frac{\sum_{j=1}^{K} N_j \sigma_j^2}{N\sigma^2_{\bar{X}_{st}} + \frac{1}{N}\sum_{j=1}^{K} N_j \sigma_j^2} \qquad \textbf{(20.27)}$$

2. **Afijación óptima:**

$$n = \frac{\frac{1}{N}\left(\sum_{j=1}^{K} N_j \sigma_j\right)^2}{N\sigma^2_{\bar{X}_{st}} + \frac{1}{N}\sum_{j=1}^{K} N_j \sigma_j^2} \qquad \textbf{(20.28)}$$

EJEMPLO 20.9. Cadena de restaurantes en tres estados (tamaño de la muestra)

Tomemos, al igual que en el ejemplo 20.4, una muestra aleatoria estratificada para estimar el número medio de pedidos por restaurante de un nuevo plato cuando el número de restaurantes que hay en los tres estados es

$$N_1 = 60 \qquad N_2 = 50 \qquad N_3 = 45$$

Supongamos también que la experiencia de la cadena de restaurantes sugiere que las desviaciones típicas poblacionales de los tres estados es probable que sean aproximadamente

$$\sigma_1 = 13 \qquad \sigma_2 = 11 \qquad \sigma_3 = 9$$

Si se necesita un intervalo de confianza al 95 por ciento de la media poblacional cuya amplitud sea de tres pedidos por restaurante a cada lado de la estimación puntual muestral, ¿cuántas observaciones muestrales se necesitan en total?

Solución

Obsérvese que

$$1{,}96\sigma_{\bar{X}_{st}} = 3, \quad \text{por lo que } \sigma_{\bar{X}_{st}} = 1{,}53$$

$$\sum_{j=1}^{K} N_j\sigma_j^2 = (60(13)^2 + (50)(11)^2 + (45)(9)^2 = 19.835$$

y

$$\frac{1}{N}\left(\sum_{j=1}^{K} N_j\sigma_j\right)^2 = \frac{[(60)(13) + (50)(11) + (45)(9)]^2}{155} = 19.421$$

En el caso de la **afijación proporcional**, el tamaño de la muestra necesario es

$$n = \frac{\sum_{j=1}^{K} N_j\sigma_j^2}{N\sigma_{\bar{X}_{st}}^2 + \frac{1}{N}\sum_{j=1}^{K} N_j\sigma_j^2} = \frac{19.835}{(155)(1{,}53)^2 + 19.835/155} = 40{,}4$$

Por lo tanto, bastará una muestra de 41 observaciones para conseguir el nivel de precisión necesario.

Si se utiliza la **afijación óptima**, el tamaño de la muestra necesario es

$$n = \frac{\frac{1}{N}\left(\sum_{j=1}^{K} N_j\sigma_j\right)^2}{N\sigma_{\bar{X}_{st}}^2 + \frac{1}{N}\sum_{j=1}^{K} N_j\sigma_j^2} = \frac{19.421}{(155)(1{,}53)^2 + 19.835/155} = 39{,}6$$

por lo que puede conseguirse el mismo grado de fiabilidad con 40 observaciones si se utiliza este método de afijación. En este caso concreto, como las desviaciones típicas poblacionales son bastante cercanas, la afijación óptima sólo representa un ahorro muy pequeño en comparación con la afijación proporcional.

EJERCICIOS

Ejercicios aplicados

20.41. Debe estimarse la cantidad media de los 812 créditos hipotecarios solicitados en una ciudad el año pasado. Basándose en la experiencia, una agencia inmobiliaria sabe que es probable que la desviación típica poblacional sea de alrededor de 20.000 $. Si el intervalo de confianza al 95 por ciento de la media poblacional debe tener una amplitud de 2.000 $ a cada lado de la media muestral, ¿cuántas observaciones muestrales se necesitan si se toma una muestra aleatoria simple?

20.42. Un concesionario de automóviles tiene unas existencias de 400 automóviles usados. Para estimar el número medio de kilómetros de estos vehículos, pretende tomar una muestra aleatoria simple de automóviles usados. Los estudios anteriores sugieren que la desviación típica poblacional es de 10.000 kilómetros. El intervalo de confianza al 90 por ciento de la media poblacional debe tener una amplitud de 2.000 kilómetros a cada lado de su estimación muestral. ¿De qué tamaño debe ser la muestra para satisfacer este requisito?

20.43. Un club de campo quiere encuestar a una muestra aleatoria de 320 socios para estimar la proporción que es probable que asista a una función a principio de temporada. El número de observaciones muestrales debe ser lo suficientemente grande para garantizar que el intervalo de confianza al 99 por ciento de la población tiene una amplitud máxima de 0,05 a cada lado de la proporción muestral. ¿De qué tamaño debe ser la muestra?

20.44. Un profesor de una clase de 417 alumnos está considerando la posibilidad de hacer un examen final que los alumnos puedan realizar en casa. Quiere tomar una muestra aleatoria de alumnos para estimar la proporción que prefiere este tipo de examen. Si el intervalo de confianza al 90 por ciento de la proporción poblacional debe tener una amplitud máxima de 0,04 a cada lado de la proporción muestral, ¿de qué tamaño debe ser la muestra?

20.45. Un auditor quiere estimar el valor medio de las facturas pendientes de cobro de una empresa. La población se divide en cuatro estratos, que contienen 500, 400, 300 y 200 facturas, respectivamente. Basándose en la experiencia, se estima que las desviaciones típicas de los valores de estos estratos serán 150 $, 200 $, 300 $ y 400 $, respectivamente. Si el intervalo de confianza al 90 por ciento de la media del conjunto de la población debe tener una amplitud de 25 $ a cada lado de la estimación muestral, halle el tamaño total de la muestra necesario utilizando tanto la afijación proporcional como la óptima.

20.46. Debe estimarse la renta media de los hogares de una ciudad que puede dividirse en tres distritos. La tabla muestra la información relevante.

Distrito	Tamaño de la población	Desviación típica estimada ($)
1	1.150	4.000
2	2.120	6.000
3	930	8.000

Si el intervalo de confianza al 95 por ciento de la media poblacional debe tener una amplitud de 500 $ a cada lado de la estimación muestral, halle el número de observaciones muestrales que se necesitan en total utilizando la afijación proporcional y la óptima.

20.6. Otros métodos de muestreo

Hemos analizado brevemente el muestreo aleatorio simple y el estratificado. Éstos no son los únicos métodos que se utilizan para elegir una muestra. En este apartado se analizan algunos otros.

Muestreo por conglomerados

Supongamos que un investigador quiere estudiar una población que se encuentra repartida en una amplia zona geográfica, como una gran ciudad o una región. Si se utiliza una muestra aleatoria simple o una muestra aleatoria estratificada, se plantean dos problemas inmediatos. En primer lugar, para extraer la muestra, el investigador necesita una lista razonablemente precisa de los miembros de la población. Puede no disponer de esa lista o es posible que pueda conseguirla con un elevado coste. En segundo lugar, aunque el investigador posea una lista de la población, los miembros de la muestra resultante estarán repartidos casi inevitablemente por una gran zona. En ese caso, será bastante caro que los entrevistadores contacten con cada uno de los miembros de la muestra. Naturalmente, este último problema no se plantea si se envía el cuestionario por correo. Sin embargo, con este medio de contacto también puede ocurrir que la tasa de falta de respuesta sea inaceptablemente alta y que el investigador prefiera por ese motivo las entrevistas personales.

Ante el dilema de no tener una lista fiable de la población o querer hacer entrevistas personales con miembros de la muestra cuando los recursos presupuestarios son limitados, el investigador puede recurrir a otro método de muestreo que se conoce con el nombre de *muestreo por conglomerados*. Este método es atractivo cuando una población puede subdividirse en unidades relativamente pequeñas y geográficamente compactas llamadas *conglomerados*. Por ejemplo, una ciudad podría subdividirse en distritos o en barrios, incluso aunque no se disponga de una lista completa de los residentes o de los hogares.

En un muestreo por conglomerados, se selecciona una muestra aleatoria simple de la población y se contacta con cada individuo de cada uno de los conglomerados de la muestra; es decir, se realiza un censo completo en cada uno de los conglomerados elegidos. En las siguientes ecuaciones, mostramos cómo pueden hacerse inferencias válidas sobre la media poblacional y la proporción poblacional a partir de los resultados de una muestra de conglomerados.

Estimadores en el muestreo por conglomerados

Se subdivide una población en M conglomerados, se selecciona una muestra aleatoria simple de m de estos conglomerados y se obtiene información de cada miembro de los conglomerados de la muestra. Sean $n_1, n_2, ..., n_n$ el número de miembros de la población que hay en los m conglomerados de la muestra. Sean las medias de estos conglomerados $\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$ y las proporciones de miembros de los conglomerados que poseen un atributo de interés $P_1, P_2, ..., P_m$. El objetivo es estimar la media μ y la proporción P de la población total.

1. Utilizando métodos de estimación insesgada, tenemos que

$$\bar{x}_c = \frac{\sum_{i=1}^{k} n_j \bar{x}_i}{\sum_{i=1}^{m} n_i} \tag{20.29}$$

y

$$\hat{p}_c = \frac{\sum_{i=1}^{k} n_j P_i}{\sum_{i=1}^{m} n_i} \tag{20.30}$$

2. Las estimaciones de la varianza de estos estimadores, basadas en métodos de estimación insesgada, son

$$\hat{\sigma}^2_{\bar{x}_c} = \frac{M - m}{Mm\bar{n}^2} \left(\frac{\sum_{i=1}^{m} n_j^2 (\bar{x}_i - \bar{x}_c)^2}{m - 1} \right) \tag{20.31}$$

y

$$\hat{\sigma}^2_{\hat{p}_c} = \frac{M - m}{Mm\bar{n}^2} \left(\frac{\sum_{i=1}^{m} n_j^2 (P_i - \hat{p}_c)^2}{m - 1} \right) \tag{20.32}$$

donde $\bar{n} = \sum_{i=1}^{m} n_j / m$ es el número medio de individuos que hay en los conglomerados de la muestra.

Basándose en estos estimadores, se obtienen los intervalos de confianza utilizando el muestreo por conglomerados.

Estimación de la media poblacional, muestreo por conglomerados

Siempre que el tamaño de la muestra es grande, el **intervalo de confianza** al 100(1 − α)% **de la media poblacional utilizando el muestreo por conglomerados** es

$$\bar{x}_c - z_{\alpha/2}\hat{\sigma}_{\bar{x}_c} < \mu < \bar{x}_c + z_{\alpha/2}\hat{\sigma}_{\bar{x}_c} \qquad \textbf{(20.33)}$$

También se hallan intervalos de confianza de la proporción poblacional utilizando el muestreo por conglomerados.

Estimación de la proporción poblacional, muestreo por conglomerados

Siempre que el tamaño de la muestra es grande, el **intervalo de confianza** al 100(1 − α)% **de la proporción poblacional utilizando el muestreo por conglomerados** es

$$\hat{p}_c - z_{\alpha/2}\hat{\sigma}_{\hat{p}_c} < P < \hat{p}_c + z_{\alpha/2}\hat{\sigma}_{\hat{p}_c} \qquad \textbf{(20.34)}$$

Obsérvese que pueden hacerse inferencias con una información previa relativamente pequeña sobre la población. Lo único que se necesita es una división en conglomerados identificables. No es necesario saber cuál es el número total de miembros de la población. Basta con saber cuál es el número que hay en cada uno de los conglomerados *de la muestra* y éste puede averiguarse durante el estudio, ya que se toma un censo completo en cada conglomerado de la muestra. Además, dado que los miembros de la muestra están geográficamente cerca unos de otros dentro de los conglomerados, es relativamente barato para los entrevistadores contactar con ellos.

EJEMPLO 20.10. Muestreo por conglomerados en el caso de las rentas familiares (estimación)

Income Clusters

Se toma una muestra aleatoria simple de 20 manzanas de una zona residencial que contiene un total de 1.100 manzanas. A continuación, se entra en contacto con cada hogar de las manzanas de la muestra y se obtiene información sobre la renta familiar. El fichero de datos **Income Clusters** contiene la renta anual media y la proporción de familias que tienen una renta de menos de 15.000 $ al año y que viven en las manzanas de la muestra. Estime la renta familiar media y la proporción de familias que tienen una renta de menos de 15.000 $ al año en esta zona residencial.

Solución

Se sabe que

$$m = 20 \quad \text{y} \quad M = 1.100$$

El número total de hogares que hay en la muestra es

$$\sum_{i=1}^{m} n_i = (23 + 31 + \cdots + 41) = 607$$

Para obtener estimaciones puntuales,

$$\sum_{i=1}^{m} n_i \bar{x}_i = (23)(26.283) + (31)(19.197) + \cdots + (41)(16.493) = 15.848.158$$

y

$$\sum_{i=1}^{m} n_i \hat{p}_i = (23)(0{,}1304) + (31)(0{,}4516) + \cdots + (41)(0{,}3659) = 153$$

Nuestras estimaciones puntuales son, pues,

$$\bar{x}_c = \frac{\sum n_i \bar{x}_i}{\sum n_i} = \frac{15.848.158}{607} = 26.109$$

$$\hat{p}_c = \frac{\sum n_i \hat{p}_i}{\sum n_i} = \frac{153}{607} = 0{,}2521$$

Por lo tanto, basándose en esta evidencia muestral, se estima que en esta zona residencial la renta anual media de los hogares es de 26.109 $ y el 25,21 por ciento de los hogares tiene una renta de menos de 15.000 $ al año.

Para obtener estimaciones de intervalos de la media poblacional, el tamaño medio de los conglomerados debe ser

$$\bar{n} = \frac{\sum n_i}{m} = \frac{607}{20} = 30{,}35$$

Además,

$$\frac{\sum_{i=1}^{m} n_i^2 (\bar{x}_i - \bar{x}_c)^2}{m-1} = \frac{(23)^2(26.283 - 26.109)^2 + \cdots + (41)^2(16.493 - 26.109)^2}{19} = 69.270.551.000$$

por lo que

$$\sigma^2_{\bar{x}_c} = \frac{M - m}{Mm\bar{n}^2} \times \frac{\sum (n_i^2 (\bar{x}_i - \bar{x}_c)^2}{m - 1} = \frac{(980)(69.270.551.000)}{(1.000)(20)(30{,}35)^2} = 3.684.914$$

y tomando la raíz cuadrada,

$$\hat{\sigma}_{\bar{x}} = 1.920$$

El intervalo de confianza al 95 por ciento de la media poblacional es

$$26.109 - (1{,}96)(1.920) < \mu < 26.109 + (1{,}96)(1.920)$$

o sea

$$22.346 < \mu < 29.872$$

El intervalo de confianza al 95 por ciento de la renta media de todas las familias de esta zona va, pues, de 22.346 $ a 29.872 $.

Para obtener estimaciones de intervalos de la proporción poblacional,

$$\frac{\sum_{i=1}^{m} n_i^2(\hat{p}_i - \hat{p}_c)^2}{m-1} = \frac{(23)^2(0{,}1304 - 0{,}02521)^2 + \cdots + (41)^2(0{,}3659 - 0{,}2521)^2}{19} = 38{,}1547$$

De donde

$$\hat{\sigma}^2_{\hat{p}_c} = \frac{M-m}{Mm\bar{n}^2}\left(\frac{\sum_{i=1}^{m} n_i^2(\hat{p}_i - \hat{p}_c)^2}{m-1}\right)$$

$$= \frac{(980)(38{,}1547)}{(1.000)(20)(30{,}35)^2} = 0{,}0020297$$

y tomando la raíz cuadrada,

$$\hat{\sigma}_{\hat{p}_c} = 0{,}0451$$

El intervalo de confianza al 95 por ciento de la proporción poblacional es

$$0{,}2521 - (1{,}96)(0{,}0451) < P < 0{,}2521 + (1{,}96)(0{,}0451)$$

o sea

$$0{,}164 < P < 0{,}340$$

Nuestro intervalo de confianza al 95 por ciento del porcentaje de hogares cuya renta anual es de menos de 15.000 $ va de 16,4 a 34,0 por ciento.

El muestreo por conglomerados se parece superficialmente al muestreo estratificado. En ambos casos, la población se divide primero en subgrupos. Sin embargo, la similitud es bastante ilusoria. En el muestreo aleatorio estratificado, se toma una muestra de *cada estrato* de la población en un intento de garantizar que se da el debido peso a importantes segmentos de la población. En cambio, en el muestreo por conglomerados se toma una muestra aleatoria de *conglomerados*, por lo que algunos conglomerados no tienen miembros en la muestra. Dado que dentro de los conglomerados los miembros de la población probablemente son bastante homogéneos, se corre el riesgo de que importantes subgrupos de la población no estén representados en absoluto o estén muy subrepresentados en la muestra final. En consecuencia, aunque la gran ventaja del muestreo por conglomerados se encuentra en su comodidad, esta comodidad puede muy bien conseguirse a costa de una imprecisión mayor de las estimaciones muestrales. Otra distinción entre el muestreo por conglomerados y el muestreo estratificado es que en el primero se toma un *censo completo* de miembros del conglomerado, mientras que en el segundo se toma una *muestra aleatoria* de miembros del estrato. Sin embargo, esta diferencia no es esencial. De hecho, a veces un investigador puede extraer una muestra aleatoria de miembros de un conglomerado en lugar de tomar un censo completo.

Muestreo bietápico

En muchas investigaciones, la población no se encuesta en una sola etapa sino que a menudo es cómodo realizar primero un estudio piloto en el que se contacta con una propor-

ción relativamente pequeña de los miembros de la muestra y se analizan los resultados obtenidos antes de realizar la mayor parte del estudio. El principal inconveniente de ese método es que puede llevar mucho tiempo. Sin embargo, tiene varias ventajas que compensan este factor. Una de las ventajas importantes es que el investigador puede probar, con un pequeño coste, el cuestionario propuesto para asegurarse de que las distintas preguntas se entienden perfectamente. El estudio piloto también puede sugerir otras preguntas cuya importancia se había pasado por alto. Además, este estudio también debe dar una estimación de la tasa probable de falta de respuesta. Si ésta fuera inaceptablemente alta, podría ser deseable modificar algo el método para recabar las respuestas.

La realización de un estudio bietápico, comenzando con un estudio piloto, se conoce con el nombre de **muestreo bietápico**. Este enfoque tiene otras dos ventajas. En primer lugar, si se emplea un muestreo aleatorio estratificado, el estudio piloto puede utilizarse para obtener estimaciones de las varianzas de los distintos estratos. Éstas pueden utilizarse, a su vez, para estimar la afijación óptima de la muestra a los distintos estratos. En segundo lugar, los resultados del estudio piloto pueden utilizarse para estimar el número de observaciones necesarias para obtener estimadores de los parámetros poblacionales con un nivel especificado de precisión. Los ejemplos siguientes sirven para ilustrar estas cuestiones. Consideremos una sencilla situación en la que se utiliza una muestra aleatoria simple para estimar una media poblacional. Al principio, la información sobre esta población es relativamente escasa, por lo que se realiza una encuesta piloto para hacerse una idea del tamaño que debe tener la muestra.

EJEMPLO 20.11. Valor medio de las facturas pendientes de cobro (tamaño de la muestra)

Un auditor desea estimar el valor medio de las facturas pendientes de cobro en una población total de 1.120 facturas. Quiere hallar un intervalo de confianza al 95 por ciento de la media poblacional que tenga una amplitud de aproximadamente 4 $ a cada lado de la media muestral. Para empezar, toma una muestra aleatoria simple de 100 facturas y observa una desviación típica muestral de 30,27 $. ¿Cuántas facturas más debe tener la muestra?

Solución

En el apartado 20.5, hemos visto que el tamaño de la muestra necesario es

$$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{x}}^2 + \sigma^2}$$

donde $N = 1.120$ es el número de miembros de la población en este caso. Para que el intervalo de confianza al 95 por ciento tenga la amplitud exigida,

$$1{,}96\sigma_{\bar{x}} = 4$$

por lo que $\sigma_{\bar{x}}$, la desviación típica de la media muestral, debe ser

$$\sigma_{\bar{x}} = \frac{4}{1{,}96} = 2{,}04$$

La desviación típica poblacional, σ, se desconoce. Sin embargo, como consecuencia del estudio inicial de 100 facturas pendientes de cobro, se estima que es 30,27. El número total de observaciones muestrales necesario es, pues,

$$n = \frac{N\sigma^2}{(N-1)\sigma_{\bar{x}}^2 + \sigma^2} = \frac{(1.120)(30{,}27)^2}{(1.119)(2{,}04)^2 + (30{,}27)^2} = 184{,}1$$

Dado que ya se han tomado 100 observaciones, serán suficientes 85 más para satisfacer el objetivo del auditor.

EJEMPLO 20.12. Renta (tamaño de la muestra)

Un investigador quiere tomar una muestra aleatoria estratificada para estimar la renta familiar media de una ciudad en la que el número de familias que hay en cada uno de los tres distritos es

$$N_1 = 1.150 \qquad N_2 = 2.120 \qquad N_3 = 930$$

Para empezar, el investigador hace un estudio piloto, tomando una muestra de 30 hogares de cada distrito y obteniendo desviaciones típicas muestrales de 3.657 $, 6.481 $ y 8.403 $, respectivamente. Supóngase que el objetivo es obtener, con el tamaño más pequeño posible, un intervalo de confianza al 95 por ciento de la media poblacional que tenga una amplitud de 500 $ a cada lado de la estimación muestral. ¿Cuántas observaciones adicionales deben tomarse en cada distrito?

Solución

El requisito de que debe conseguirse un grado especificado de precisión con el menor número de observaciones muestrales posible implica que debe utilizarse la afijación óptima. Recuérdese que en la ecuación 20.20 hemos visto que los números n_1, n_2 y n_3 que deben muestrearse en los tres estratos son los siguientes:

$$n_j = \frac{N_j\sigma_j}{\sum_{i=1}^{K} N_i\sigma_i} \times n \qquad (j = 1, 2, 3)$$

donde las σ_i son las desviaciones típicas poblacionales de los estratos. Utilizando nuestras estimaciones muestrales en lugar de estas cantidades,

$$n_1 = \frac{(1.150)(3.657)}{(1.150)(3.657) + (2.120)(6.481) + (930)(8.403)} \times n = 0{,}163n$$

$$n_2 = \frac{(2.120)(6.481)}{(1.150)(3.657) + (2.120)(6.481) + (930)(8.403)} \times n = 0{,}533n$$

$$n_3 = \frac{(930)(8.403)}{(1.150)(3.657) + (2.120)(6.481) + (930)(8.403)} \times n = 0{,}303n$$

Hemos especificado las propiedades de la muestra total que debe afijarse a cada estrato con el sistema óptimo. Queda por averiguar el número total n de observaciones muestrales.

Métodos de muestreo no probabilísticos

Hemos analizado algunos sistemas de muestreo en los que es posible especificar la probabilidad de que se extraiga una determinada muestra de la población. Esta característica de los métodos de muestreo permite hacer inferencias estadísticas válidas basadas en los resultados muestrales. De lo contrario, no podrían obtenerse estimaciones puntuales insesgadas e intervalos de confianza con un contenido probabilístico especificado que tuvieran una estricta validez estadística.

No obstante, en muchas aplicaciones prácticas se utilizan **métodos no probabilísticos** para seleccionar miembros de la muestra, principalmente por comodidad. Supongamos, por ejemplo, que queremos evaluar las reacciones de los estudiantes de nuestra universidad a algún tema de interés. Una posibilidad sería preguntar a nuestros amigos cuál es su opinión. Este grupo no constituiría una muestra aleatoria de la población de todos los estudiantes. Por lo tanto, si analizamos los datos como si procedieran de una muestra aleatoria, la inferencia resultante carecería de validez estadística.

Las organizaciones que realizan encuestas utilizan a menudo una versión más sofisticada del enfoque que acabamos de describir, llamada **muestreo por cuotas**. Se asignan encuestadores a un lugar y se les dice que contacten con un número especificado de personas de una determinada edad, raza y sexo. Estas cuotas asignadas representan las proporciones del conjunto de la población que se consideran adecuadas. Sin embargo, una vez decididas las cuotas, los entrevistadores tienen flexibilidad para elegir los miembros de la muestra. Su elección normalmente no es aleatoria. El muestreo por cuotas puede producir y a menudo produce estimaciones bastante precisas de los parámetros poblacionales. Su inconveniente es que, como no se elige la muestra utilizando métodos probabilísticos, no existe una forma válida de averiguar la fiabilidad de las estimaciones resultantes.

EJERCICIOS

Ejercicios aplicados

20.47. Una empresa de estudios de mercado quiere estimar la cantidad semanal media de tiempo que están encendidos los televisores en los hogares de una ciudad que contiene 65 barrios. Se selecciona una muestra aleatoria simple de 10 barrios y se pregunta a cada hogar de cada barrio de la muestra. La tabla adjunta muestra los resultados.

Barrio	Número de hogares	Tiempo medio de uso del televisor (horas)
1	28	29,6
2	35	18,4
3	18	32,7
4	52	26,3
5	41	22,4
6	38	31,6
7	36	19,7
8	30	23,8
9	23	25,4
10	42	24,1

a) Halle una estimación puntual de la media poblacional de la cantidad de tiempo que están encendidos los televisores en esta ciudad.

b) Halle el intervalo de confianza al 90 por ciento de la media poblacional.

20.48. Un dirigente sindical quiere estimar el valor medio de las primas pagadas a los administrativos de una empresa en el primer mes de un nuevo plan. Esta empresa tiene 52 subdivisiones y se toma una muestra aleatoria simple de 8. A continuación, se obtiene información de las nóminas de cada administrativo de cada subdivisión de la muestra. La tabla adjunta muestra los resultados.

Subdivisión	Número de administrativos	Prima media (dólares)
1	69	83
2	75	64
3	41	42
4	36	108
5	59	136
6	82	102
7	64	95
8	71	98

a) Halle una estimación puntual de la prima media por administrativo de este mes.
b) Halle el intervalo de confianza al 99 por ciento de la media poblacional.

20.49. En el estudio del ejercicio 20.47, se pregunta a los hogares si tienen televisión por cable. La tabla adjunta muestra el número que tiene televisión por cable.

Barrio	1	2	3	4	5	6	7	8	9	10
Número	12	11	10	29	15	13	20	14	9	26

a) Halle una estimación puntual de la proporción de todos los hogares de la ciudad que tienen televisión por cable.
b) Halle el intervalo de confianza al 90 por ciento de esta proporción poblacional.

20.50. En el estudio del ejercicio 20.48, se preguntó a los administrativos de las ocho subdivisiones de la muestra si estaban satisfechos con el funcionamiento del plan de primas. La tabla adjunta muestra los resultados.

Subdivisión	1	2	3	4	5	6	7	8
Número satisfecho	24	25	11	21	35	44	30	34

a) Halle una estimación puntual de la proporción de todos los administrativos satisfechos con el plan de primas.
b) Halle el intervalo de confianza al 95 por ciento de esta proporción poblacional.

20.51. Una ciudad está dividida en 50 subdivisiones geográficas. Se necesita una estimación de la proporción de los hogares de la ciudad interesados en un nuevo servicio de jardinería. Una muestra aleatoria de tres subdivisiones contiene 611, 521 y 734 hogares, respectivamente. El número que expresa interés por el servicio es 128, 131 y 172, respectivamente. Halle el intervalo de confianza al 90 por ciento de la proporción de todos los hogares de la ciudad interesada en el servicio de jardinería.

20.52. Un banco tiene 720 créditos hipotecarios para la adquisición de viviendas en situación de morosidad. Necesita una estimación del valor catastral medio de estas viviendas. Al principio, se considera una muestra aleatoria de 20 y se halla una desviación típica muestral de 37.600 $. Si el banco requiere un intervalo de confianza al 90 por ciento de la media poblacional que tenga una amplitud de 5.000 $ a cada lado de la media muestral, ¿cuántas viviendas más deben considerarse?

20.53. Una universidad tiene 3.200 estudiantes de grado y 800 estudiantes de postgrado. Los investigadores tienen interés en saber cuánto dinero se gastan estos estudiantes en un año en libros de texto. Al principio se toman muestras aleatorias simples de 30 estudiantes de grado y 30 de postgrado. Las desviaciones típicas muestrales de las cantidades gastadas son 40 $ y 58 $, respectivamente. Se necesita un intervalo de confianza al 90 por ciento de la media del conjunto de la población que tenga una amplitud de 5 $ a cada lado de la estimación puntual muestral. Estime el menor número total de observaciones muestrales adicionales necesario para lograr este objetivo.

20.54. Una empresa tiene una flota de 480 automóviles: 100 pequeños, 180 de tamaño intermedio y 200 grandes. Para estimar los costes totales anuales medios de reparación de estos automóviles, se toma una muestra aleatoria preliminar de 10 automóviles de cada tipo. Las desviaciones típicas muestrales de los costes de reparación son 105 $ en el caso de los automóviles pequeños, 162 $ en el de los automóviles de tamaño intermedio y 183 $ en el de los automóviles grandes. Se necesita un intervalo de confianza al 95 por ciento del coste total anual medio de reparación por automóvil que tenga una amplitud de 20 $ a cada lado de la estimación puntual muestral. Estime el menor número total de observaciones muestrales adicionales que deben tomarse.

RESUMEN

En este capítulo, hemos centrado la atención en el problema de un investigador que quiere descubrir algo de una población que no es necesariamente grande. El investigador pretende recoger información solamente de un subconjunto de miembros de la población y pide asesoramiento para hacerlo. En primer lugar, deben considerarse los pasos necesarios en un plan de muestreo. A continuación, deben distinguirse los errores de muestreo y los errores ajenos al muestreo; deben formularse ecuaciones para estimar una media poblacional, un total poblacional y una proporción poblacional para el muestreo aleatorio simple, así como para el

muestreo estratificado; debe decidirse el tamaño de la muestra para estimar una media poblacional, un total poblacional y una proporción poblacional utilizando el muestreo aleatorio simple o el muestreo estratificado si se especifica la varianza deseada de la media muestral; debe considerarse el muestreo por conglomerados y las ecuaciones establecidas para hallar los intervalos de confianza de la media poblacional y de la proporción poblacional, si el tamaño de la muestra es grande. Hemos mencionado brevemente el método de muestreo bietápico y el método de muestreo no probabilístico.

Dado que la estadística se ocupa en gran parte de los problemas que plantean las afirmaciones sobre una población a partir de la información muestral, nos interesa comprender este capítulo. Para un análisis más detallado de los diseños de muestreo, véanse las notas que se encuentran al final de este capítulo.

TÉRMINOS CLAVE

estimación:
- media poblacional, aleatorio, 820
- media poblacional, conglomerado, 845
- media poblacional, estratificado, 827
- proporción poblacional, aleatorio, 823
- proporción poblacional, conglomerado, 845
- proporción poblacional, estratificado, 831
- total poblacional, aleatorio, 821
- total poblacional, estratificado, 829

error ajeno al muestreo, 817
error de muestreo, 817
factor de corrección en el caso de una población finita, 820
métodos no probabilísticos, 850
muestreo aleatorio simple, 819
muestreo aleatorio estratificado, 826
muestreo por conglomerados, 844
muestreo por cuotas, 850
muestreo bietápico, 848
muestreo sistemático, 819

tamaño de la muestra:
- afijación óptima, 834
- afijación proporcional, 833
- media poblacional, aleatorio, 838
- media poblacional, estratificado, 840
- proporción poblacional, aleatorio, 839

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

20.55. Ha recibido el encargo de diseñar y realizar una encuesta en su ciudad sobre la eficacia de una campaña publicitaria por radio destinada a promocionar una nueva película.

a) Explique qué haría.
b) Analice las posibilidades de que haya errores ajenos al muestreo y los medios para reducir lo más posible su importancia.
c) ¿Hasta qué punto espera que la falta de respuesta sea un problema en esta encuesta?

20.56. Basándose en una muestra aleatoria de 10 miembros de su clase, estime la cantidad media de dinero que gastan los miembros de la clase en libros de texto cada cuatrimestre.

20.57. Explique minuciosamente la distinción entre muestreo aleatorio estratificado y muestreo por conglomerados. Ponga ejemplos de problemas de muestreo en los que podría ser útil cada una de estas técnicas.

20.58. Se hace un examen a 90 estudiantes y se toma una muestra aleatoria de 10 calificaciones:

93 71 62 75 81 63 87 59 84 72

a) Halle el intervalo de confianza al 90 por ciento de la media poblacional de las calificaciones.
b) Sin hacer los cálculos, indique si el intervalo de confianza al 95 por ciento de la media poblacional sería más amplio o más estrecho que el obtenido en el apartado (a).

20.59. Una empresa tiene 272 facturas pendientes de cobro en una determinada categoría. Se toma una muestra aleatoria de 50 facturas. La media muestral es de 492,36 $ y la desviación típica muestral es de 149,92 $.

a) Halle el intervalo de confianza al 99 por ciento de la media poblacional del valor de estas facturas pendientes de cobro.
b) Halle el intervalo de confianza al 95 por ciento del valor total de estas facturas pendientes de cobro.
c) Indique sin hacer los cálculos si el intervalo de confianza al 90 por ciento del total poblacional sería más amplio o más estrecho que el intervalo obtenido en el apartado (b).

20.60. En el Senado de Estados Unidos hay 100 senadores. Se obtuvo información de los individuos

responsables de gestionar la correspondencia de 61 despachos de senadores. De éstos, 38 indicaron que debían recibir un número mínimo de cartas sobre una cuestión antes de escribir una carta en respuesta.

a) Suponga que estas observaciones constituyen una muestra aleatoria de la población y halle el intervalo de confianza al 90 por ciento de la proporción de despachos de senadores que siguen esta política.

b) En realidad, *no* se obtuvo información de una muestra aleatoria de despachos de senadores. Se enviaron cuestionarios a los 100 despachos, pero sólo respondieron 61. ¿Cómo influye esta información en su respuesta al apartado (a)? Véase la referencia bibliográfica 2.

20.61. Una empresa tiene 148 representantes de ventas. Se toma una muestra aleatoria de 60 y se observa que en el caso de 36 de los miembros de la muestra, el volumen de pedidos de este mes es mayor que el del mismo mes del año pasado. Halle el intervalo de confianza al 95 por ciento de la proporción poblacional de representantes de ventas que tienen un volumen de pedidos mayor.

20.62. Una empresa tiene tres subdivisiones, en las que hay un total de 970 directivos. Se toman muestras aleatorias independientes de directivos de cada subdivisión y se halla el número de años que lleva en la empresa cada miembro de las muestras. La tabla adjunta muestra los resultados.

	Subdivisión 1	**Subdivisión 2**	**Subdivisión 3**
N_i	352	287	331
n_i	30	20	30
$\bar{x}_i$	9,2	12,3	13,5
s_i	4,9	6,4	7,6

a) Halle el intervalo de confianza al 99 por ciento del número medio de años que llevan en la empresa los directivos de la subdivisión 1.

b) Halle el intervalo de confianza al 99 por ciento del número medio de años que llevan en la empresa todos los directivos.

20.63. De las 300 páginas de un libro, 180 son principalmente poco técnicas, mientras que el resto es técnico. Se toman muestras aleatorias independientes de páginas técnicas y no técnicas y se anota el número de erratas por página. La tabla resume los resultados.

	Técnicas	**No técnicas**
N_i	120	180
n_i	20	20
$\bar{x}_i$	1,6	0,74
s_i	0,98	0,56

a) Halle el intervalo de confianza al 95 por ciento del número medio de erratas por página de este libro.

b) Halle el intervalo de confianza al 99 por ciento del número total de erratas del libro.

20.64. En el análisis del ejercicio 20.63, se observa que 9 de las páginas técnicas de la muestra y 15 de las páginas no técnicas de la muestra no contienen ninguna errata. Halle el intervalo de confianza al 90 por ciento de la proporción de todas las páginas de este libro que no contiene erratas.

20.65. Vuelva a los datos del ejercicio 20.62. Si se toma una muestra de un total de 80 directivos, averigüe cuántos miembros de la muestra pertenecerían a la subdivisión 1 utilizando cada uno de los siguientes sistemas:

a) La afijación proporcional y

b) La afijación óptima, suponiendo que las desviaciones típicas de los estratos son iguales que las cantidades muestrales correspondientes.

20.66. Vuelva a los datos del ejercicio 20.63. Si se toma una muestra de un total de 40 páginas, averigüe cuántas páginas de la muestra serían técnicas utilizando cada uno de los siguientes sistemas:

a) La afijación proporcional y

b) La afijación óptima, suponiendo que las desviaciones típicas de los estratos son iguales que las cantidades muestrales correspondientes.

20.67. Se pretende tomar una muestra de los estudiantes de su universidad para conocer su opinión sobre la cantidad de espacio que hay en la biblioteca. Se decide utilizar una muestra estratificada por año: estudiantes de primer año, de segundo año, etc. Analice los factores que se tendrían en cuenta para decidir el número de observaciones muestrales que deben tomarse en cada estrato.

20.68. Un concesionario de automóviles tiene unas existencias de 328 automóviles usados. Hay que estimar el número medio de kilómetros de

estos vehículos. La experiencia dice que es probable que la desviación típica poblacional sea de unos 12.000 kilómetros. Si el intervalo de confianza al 90 por ciento de la media poblacional debe tener una amplitud de 2.000 kilómetros a cada lado de la media muestral, ¿de qué tamaño debe ser la muestra si se emplea el muestreo aleatorio simple?

20.69. Debe tomarse una muestra aleatoria simple de 527 estudiantes de administración de empresas de una universidad para estimar la proporción que es partidaria de que se ponga más énfasis en la ética empresarial en el programa de estudios. ¿Cuántas observaciones son necesarias para garantizar que el intervalo de confianza al 95 por ciento de la proporción poblacional tiene una amplitud máxima de 0,06 a cada lado de la proporción muestral?

20.70. Suponga que la junta electoral debe ayudar a resolver un conflicto electoral entre dos candidatos (o quizá una persona debe hacer de experto estadístico en un juicio relacionado con el resultado de unas reñidas elecciones). Son muchas las cuestiones que se plantean. ¿Deben recontarse todos los votos de todas las circunscripciones? Si sólo se recuentan los de algunas, ¿cuáles? Analice las ventajas y los inconvenientes de algunos diseños muestrales que podrían utilizarse para seleccionar los votos que van a recontarse.

Bibliografía

1. Cochran, W. G., *Sampling Techniques*, Nueva York, Wiley, 1977, 3.ª ed.
2. Culnan, M. J., «Processing Unstructured Organizational Transactions: Mail Handling in the U.S. Senate», *Organizational Science*, 3, 1992, págs. 117-137.
3. Deming, W. E., *Sample Design in Business Research*, Nueva York, Wiley, 1960.
4. Hogg, Robert y Allen T. Craig, *Introduction to Mathematical Statistics*, Nueva York, Macmillan, 1977, 4.ª ed.
5. Kish, Leslie, *Survey Sampling*, Nueva York, Wiley, 1965.
6. Levy, Paul S. y Stanley Lemeshow, *Sampling of Populations*: *Methods and Applications*, Nueva York, Wiley, 1991.
7. *Minitab for Windows Version 13*, State College, PA, Minitab, Inc., 2000.
8. Schaeffer, Richard L., William Mendenhall y Lyman Ott, *Elementary Survey Sampling*, Belmont, CA, Duxbury Press, 1996, 5.ª ed.

Capítulo 21

Teoría estadística de la decisión

Esquema del capítulo

Introducción

Podría decirse que el tema de este capítulo recoge la esencia de los problemas de gestión que se plantean en cualquier organización. De hecho, su aplicabilidad va mucho más allá, ya que afecta a muchos aspectos de nuestra vida diaria. Analizaremos situaciones en las que una persona, un grupo o una empresa tienen varios cursos de acción posibles y deben elegir uno de ellos en un mundo en el que hay incertidumbre sobre la futura conducta de los factores que determinan las consecuencias del curso de acción que se elija. En este capítulo analizamos cuatro criterios para tomar decisiones. El criterio maximin y el criterio de la pérdida de oportunidades minimax son criterios no probabilísticos para tomar decisiones. Es decir, estos criterios «no tienen en cuenta la probabilidad de los resultados de cada alternativa; centran meramente la atención en el valor monetario de los resultados» (véase la referencia bibliográfica 4). Dos criterios para tomar decisiones que incluyen información sobre las probabilidades de que se produzca cada resultado son el criterio del valor monetario esperado y el criterio de la utilidad esperada.

21.1. La toma de decisiones en condiciones de incertidumbre

Todos nos vemos obligados a actuar en un entorno cuyo rumbo futuro es incierto. Por ejemplo, podemos estar considerando la posibilidad de ir a un partido de fútbol, pero dudamos porque existe la posibilidad de que llueva. Si *supiéramos* que no va a llover, iríamos al partido; si estuviéramos *seguros* de que va a llover durante varias horas, no iríamos. Pero no podemos predecir con absoluta seguridad el tiempo que va a hacer, por lo que debemos tomar la decisión contemplando un incierto futuro. Por poner otro ejemplo, en algún momento al final de los estudios universitarios, el estudiante tiene que decidir qué va a hacer cuando se gradúe. Es posible que ya tenga varias ofertas de empleo. Hacer el doctorado también es una posibilidad. La decisión es claramente importante. Recabará, desde luego, información sobre las opciones. Sabrá qué sueldos de partida se ofrecen y se habrá enterado de cuáles son las actividades de las empresas entre las que puede elegir y de cómo encaja en esas actividades.

Sin embargo, nadie tiene una idea muy clara de dónde estará dentro de uno o dos años si acepta una determinada oferta. Esta importante decisión se toma, pues, en condiciones de incertidumbre sobre el futuro.

En el mundo empresarial, a menudo existen circunstancias de este tipo, como muestran los siguientes ejemplos:

1. En una recesión, una empresa debe decidir si despide o no a algunos trabajadores. Si la recesión económica va a ser breve, puede ser preferible quedarse con estos trabajadores, que pueden ser difíciles de sustituir cuando mejore la demanda. Sin embargo, si se prolonga la recesión, conservarlos sería caro. Desgraciadamente, el arte de la predicción económica no ha llegado a la fase en la que es posible predecir con un alto grado de certeza la duración o la gravedad de una recesión.
2. Un inversor puede creer que los tipos de interés han alcanzado un máximo. En ese caso, los bonos a largo plazo parecerían muy atractivos. Sin embargo, es imposible estar seguro de cómo evolucionarán en el futuro, y si continuaran subiendo, la decisión de invertir en bonos a largo plazo sería subóptima.
3. Los contratistas a menudo deben hacer ofertas para conseguir la adjudicación de un proyecto. Tienen que decidir la cuantía de la oferta. En este caso, hay dos cuestiones inciertas. En primer lugar, el contratista no sabe de qué cuantía tiene que ser la oferta para conseguir el contrato. En segundo lugar, no puede estar seguro de cuánto le costará cumplir el contrato. De nuevo, a pesar de la incertidumbre, debe tomar alguna decisión.
4. El coste de hacer prospecciones petroleras en alta mar es enorme y, a pesar de contar con excelente asesoramiento geológico, las compañías petroleras no saben, antes de hacer las prospecciones, si se descubrirá una cantidad comercialmente viable. La decisión de hacer o no prospecciones petroleras debe tomarse en un entorno incierto.

Nuestro objetivo es estudiar los métodos para abordar el tipo de problemas de toma de decisiones que acabamos de describir. Una persona que tiene que tomar una decisión se enfrenta a un número finito, K, de *acciones* posibles, que llamaremos a_1, a_2, ..., a_K. En el momento en que tiene que elegir una acción, no sabe cómo evolucionará en el futuro un factor que determinará las consecuencias de la acción elegida. Se supone que un número finito, H, de *estados de la naturaleza* posibles puede caracterizar las posibilidades de este factor. Éstos se representan por medio de s_1, s_2, ..., s_H. Por último, se supone que la persona que tiene que tomar la decisión es capaz de especificar la recompensa monetaria o *ren-*

dimiento de cada combinación acción-estado de la naturaleza. Sea M_{ij} el rendimiento de la acción a_i en el supuesto de que ocurra el estado de la naturaleza s_j. Las acciones, los estados de la naturaleza, los rendimientos monetarios y las tablas de rendimientos forman parte del marco general para analizar cualquier problema de toma de decisiones.

Marco para analizar los problemas de toma de decisiones

1. La persona que tiene que tomar una decisión tiene K cursos de **acción** posibles: a_1, a_2, ..., a_K. Las acciones a veces se llaman alternativas.
2. Hay H **estados de la naturaleza** inciertos posibles: s_1, s_2, ..., s_H. Los estados de la naturaleza son los resultados posibles que el que toma la decisión no controla. A veces se llaman sucesos.
3. Cada combinación posible acción-estado de la naturaleza tiene un resultado que representa un beneficio o una pérdida, llamado **rendimiento** monetario, M_{ij}, que corresponde a la acción a_i y al estado de la naturaleza s_j. La tabla de todos los resultados de un problema de decisión se llama **tabla de rendimientos**.

La Tabla 21.1 muestra la forma general de una tabla de rendimientos.

Tabla 21.1. Tabla de rendimientos de un problema de decisión en el que hay K acciones posibles y H estados de la naturaleza posibles.

Acción	Estado de la naturaleza			
a_i/s_i	s_1	s_2	$\cdots$	s_H
a_1	M_{11}	M_{12}	$\cdots$	M_{1H}
a_2	M_{21}	M_{22}	$\cdots$	M_{2H}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_K	M_{K1}	M_{K2}	$\cdots$	M_{KH}

Cuando una persona que tiene que tomar una decisión se encuentra ante distintos cursos de acción, la elección correcta dependerá en gran medida de los objetivos. Es posible describir varias líneas de ataque que se han empleado en la solución de problemas de toma de decisiones empresariales. Sin embargo, debe tenerse presente que cada problema tiene sus propias características y que los objetivos de los que toman las decisiones pueden variar considerablemente y ser, de hecho, bastante complejos. Se plantea una situación de este tipo cuando se observa la posición de un directivo intermedio de una gran empresa. En la práctica, sus objetivos pueden ser algo distintos de los de la empresa. Al tomar decisiones, es muy probable que sea consciente de su propia posición, así como del bien general de la empresa.

A pesar del carácter individual de los problemas de toma de decisiones, es posible eliminar algunas acciones que no se considerarán en ningún caso.

Acciones admisibles e inadmisibles

Si el rendimiento de una acción a_j es al menos tan alto como el de a_i, cualquiera que sea el estado de la naturaleza, y si el rendimiento de a_j es mayor que el de a_i al menos en un estado de la naturaleza, se dice que la acción a_j *domina* a la acción a_i. Se dice que cualquier acción que es dominada de esta forma es **inadmisible**. Las acciones inadmisibles se eliminan de la lista de posibilidades antes de seguir analizando un problema de toma de decisiones. Se dice que cualquier acción que no es dominada por alguna otra y que, por lo tanto, no es inadmisible es **admisible**.

En este capítulo nos basaremos en el ejemplo siguiente.

EJEMPLO 21.1. Un fabricante de teléfonos móviles (acciones admisibles)

Consideremos un fabricante que planea introducir un nuevo teléfono móvil. Puede elegir entre cuatro procesos de producción, A, B, C y D, que van desde una modificación relativamente pequeña de las instalaciones existentes hasta una gran ampliación de la planta. La decisión sobre el curso de acción debe tomarse en un momento en el que no se conoce la demanda posible del producto. Por comodidad, decimos que esta demanda potencial puede ser «baja», «moderada» o «alta». También se supone que el fabricante puede calcular para cada proceso de producción el beneficio durante la vida de la inversión correspondiente a cada uno de los tres niveles de demanda. La Tabla 21.2 muestra estos niveles de beneficios (en dólares) para cada combinación proceso de producción-nivel de demanda. Averigüe si hay alguna acción inadmisible.

Tabla 21.2. Beneficios estimados de un fabricante de teléfonos móviles correspondientes a diferentes combinaciones de proceso-demanda.

Acción	**Estado de la naturaleza**		
Proceso de producción	**Demanda baja**	**Demanda moderada**	**Demanda alta**
A	70.000	120.000	200.000
B	80.000	120.000	180.000
C	100.000	125.000	160.000
D	100.000	120.000	150.000

Solución

En este ejemplo, hay cuatro acciones posibles que corresponden a los cuatro procesos de producción posibles y tres estados de la naturaleza posibles que corresponden a los tres niveles de demanda del producto posibles.

Consideremos el proceso de producción D de la Tabla 21.2. El rendimiento de este proceso será exactamente igual que el de C si hay un bajo nivel de demanda y más bajo que el del proceso C si el nivel de demanda es moderado o alto. Por lo tanto, no tiene sentido elegir la opción D, ya que hay otra opción con la que los rendimientos no pueden ser menores y podrían ser mayores. Dado que la acción C es necesariamente al menos tan rentable como la D y posiblemente más, se dice que la acción C *domina* a la D. Dado que el proceso de producción D es dominado por otra alternativa, el proceso de producción C, se dice que el D es *inadmisible*. Esta acción no debe seguir considerándose, ya que sería subóptimo adoptarla. Por lo tanto, se eliminará y, en el análisis posterior del problema, sólo se considerará la posibilidad de adoptar el proceso A, el B o el C.

El problema de toma de decisiones esbozado es esencialmente de carácter discreto. Es decir, sólo hay un número finito de alternativas y un número finito de estados de la naturaleza posibles. Sin embargo, muchos problemas prácticos son continuos. Por ejemplo, es posible que sea mejor medir el estado de la naturaleza en un continuo que describirlo por medio de una serie de posibilidades discretas. En el ejemplo del fabricante de teléfonos móviles, es posible prever un intervalo de niveles posibles de demanda en lugar de especificar simplemente tres niveles. En algunos problemas, como mejor se re-

presentan las acciones posibles es en un continuo; por ejemplo, en el caso en el que un contratista debe decidir la cuantía de la oferta para conseguir la adjudicación de un contrato. En el resto de este capítulo centramos la atención en el caso discreto. Los *principios* que implica el análisis del caso continuo no son diferentes. Sin embargo, los detalles de ese análisis se basan en el cálculo y no se examinan más aquí.

EJERCICIOS

Ejercicios básicos

21.1. Un inversor está considerando tres alternativas —un certificado de depósito, un fondo de acciones de bajo riesgo y un fondo de acciones de alto riesgo— para una inversión de 20.000 $. Considera tres estados de la naturaleza posibles:

s_1: mercado de valores fuerte
s_2: mercado de valores moderado
s_3: mercado de valores débil

La tabla de rendimientos (en dólares) es la siguiente:

Acción	Estado de la naturaleza		
Alternativas de inversión posibles	s_1	s_2	s_3
Certificado de depósito	1.200	1.200	1.200
Fondo de acciones de bajo riesgo	4.300	1.200	−600
Fondo de acciones de alto riesgo	6.600	800	−1.500

¿Es inadmisible alguna de estas acciones?

21.2. Un fabricante de desodorantes está a punto de ampliar la capacidad de producción para fabricar un nuevo producto. Tiene cuatro procesos de producción alternativos. La tabla adjunta muestra los beneficios estimados, en dólares, de estos procesos correspondientes a tres niveles de demanda del producto posibles.

Acción	Estado de la naturaleza		
Proceso de producción	**Demanda baja**	**Demanda moderada**	**Demanda alta**
A	100.000	350.000	900.000
B	150.000	400.000	700.000
C	250.000	400.000	600.000
D	250.000	400.000	550.000

¿Es inadmisible alguna de estas acciones?

21.2. Soluciones que no implican la especificación de probabilidaes: criterio maximin, criterio de la pérdida de oportunidades minimax

Antes de elegir el proceso de producción, es probable que nuestro fabricante de teléfonos móviles se pregunte cuáles son las probabilidades de que se materialice realmente cada uno de estos niveles de demanda. Este capítulo se ocupa en su mayor parte de analizar las soluciones a un problema de toma de decisiones que requiere la especificación de las probabilidades de los resultados correspondientes a los diversos estados de la naturaleza. Sin embargo, en este apartado se presentan dos criterios de decisión que no se basan en esas probabilidades y que, en realidad, no tienen ningún contenido probabilístico. Estos enfoques (y otros del mismo tipo) sólo dependen, más bien, de la estructura de la tabla de rendimientos.

Los dos métodos examinados en este apartado se llaman *criterio maximin* y *criterio de la pérdida de oportunidades minimax*. Se examinan en relación con la tabla de rendimientos del fabricante de teléfonos móviles del ejemplo 21.1 dejando de lado la estrategia inad-

misible de elegir el proceso de producción D. El fabricante debe elegir, pues, entre las tres acciones posibles, enfrentándose a tres estados de la naturaleza posibles.

Criterio maximin

Consideremos el peor resultado posible de cada acción, cualquiera que sea el estado de la naturaleza que se materialice. El *peor resultado* es simplemente el menor rendimiento que es razonable pensar que podría obtenerse. El **criterio maximin** selecciona la acción que tiene el rendimiento mínimo, es decir, *maximizamos* el rendimiento *mínimo*.

En el caso del problema del fabricante de teléfonos móviles, el menor rendimiento, cualquiera que sea el proceso de producción que se emplee, se obtiene cuando el nivel de demanda es bajo. Es evidente que, como muestra la Tabla 21.3, el valor máximo de estos rendimientos mínimos es 100.000 $. Se obtiene si se utiliza el proceso de producción C. Por lo tanto, el criterio maximin selecciona el proceso de producción C.

Tabla 21.3. Aplicación del criterio maximin al ejemplo 21.1.

Acción	**Estado de la naturaleza**			**Rendimiento mínimo**
Proceso de producción	**Demanda baja**	**Demanda moderada**	**Demanda alta**	**Rendimiento mínimo de cada proceso**
A	70.000	120.000	200.000	70.000
B	80.000	120.000	180.000	80.000
C	100.000	125.000	160.000	**100.000 (máximo)**

Dado que el valor máximo del rendimiento mínimo de cada proceso de producción es 100.000 $, se deduce que con el criterio maximin se selecciona el proceso de producción C como curso de acción.

EJEMPLO 21.2. Oportunidad de inversión (maximin)

Un inversor quiere elegir entre invertir 10.000 $ durante un año a un tipo de interés garantizado del 12 por ciento e invertir la misma cantidad durante ese periodo en una cartera de acciones ordinarias. Si elige el tipo de interés fijo, tendrá con seguridad un rendimiento de 1.200 $. Si elige la cartera de acciones, el rendimiento dependerá del comportamiento del mercado durante el año. Si el mercado está boyante, se espera un beneficio de 2.500 $; si el mercado se mantiene estable, el beneficio esperado es de 500 $; y si está deprimido, se espera una pérdida de 1.000 $. Elabore la tabla de rendimientos de este inversor y halle la elección de la acción mediante el criterio maximin.

Solución

La Tabla 21.4 muestra los rendimientos (en dólares); un rendimiento negativo indica una pérdida.

El rendimiento mínimo de la inversión a un tipo de interés fijo es de 1.200 $, ya que éste es el rendimiento que se obtendrá independientemente de lo que ocurra en la bolsa de valores. El rendimiento mínimo de la cartera de acciones es una pérdida de 1.000 $, o sea, un rendimiento de -1.000 $, que se produce cuando el mercado está deprimido. Dado que el mayor rendimiento mínimo es el de la inversión a un tipo de interés fijo, se deduce que se selecciona el tipo de interés fijo como curso de acción mediante el criterio maximin.

Tabla 21.4. Aplicación del criterio maximin al ejemplo 21.2.

Acción	Estado de la naturaleza			Rendimiento mínimo
Opción de inversión	**Estado boyante**	**Estado estable**	**Estado deprimido**	**Rendimiento mínimo de cada opción de inversión**
Tipo de interés fijo	1.200	1.200	1.200	**1.200 (máximo)**
Cartera de acciones	2.500	500	− 1.000	− 1.000

En estos ejemplos, se observa claramente la forma general de la regla de decisión basada en el criterio maximin. El objetivo del criterio maximin es *maximizar* el rendimiento *mínimo*.

Regla de decisión basada en el criterio maximin

Supongamos que una persona que tiene que tomar una decisión tiene que elegir entre K acciones admisibles $a_1, a_2, ..., a_K$, dados H estados de la naturaleza posibles $s_1, s_2, ..., s_H$. Sea M_{ij} el rendimiento correspondiente a la i-ésima acción y el j-ésimo estado de la naturaleza. Debe buscarse el menor rendimiento posible de cada acción. Por ejemplo, en el caso de la acción a_1, éste es el menor de $M_{11}, M_{12}, ..., M_{1H}$. Sea este mínimo M_1^*, donde

$$M_1^* = \text{Min}\,(M_{11}, M_{12}, ..., M_{1H})$$

En términos más generales, el menor rendimiento posible de la acción a_i viene dado por

$$M_i^* = (M_{i1}, M_{i2}, ..., M_{iH})$$

El **criterio maximin** selecciona la acción a_i cuyo M_1^* es mayor (es decir, la acción cuyo rendimiento mínimo es mayor).

La característica positiva del criterio maximin para tomar decisiones es que genera el mayor rendimiento posible que puede *garantizarse*. Si se utiliza el proceso de producción C, el fabricante de teléfonos móviles tiene *asegurado* un rendimiento de al menos 100.000 $, cualquiera que sea al final el nivel de demanda. Asimismo, en el caso del inversor del ejemplo 21.2, la elección del tipo de interés fijo genera un beneficio *seguro* de 1.200 $. En ninguno de los dos ejemplos, ninguna acción alternativa puede *garantizar* tanto.

Sin embargo, es precisamente dentro de esta garantía donde surgen las reservas sobre el criterio maximin, ya que a menudo debe pagarse un precio por esa garantía. El precio es aquí la pérdida de la oportunidad de percibir un rendimiento mayor, eligiendo alguna otra acción, *por muy improbable* que parezca que es la peor situación posible. Así, por ejemplo, el fabricante de teléfonos móviles puede estar casi seguro de que la demanda será alta, en cuyo caso el proceso de producción C sería una mala elección, ya que genera el menor rendimiento con este nivel de demanda.

Puede considerarse, pues, que el criterio maximin es una estrategia muy cauta para elegir entre distintas acciones alternativas. Esa estrategia puede ser adecuada en algunas circunstancias, pero sólo un pesimista extremo la utilizaría invariablemente. Por este motivo, a veces se llama *criterio del pesimismo*. «El criterio maximin se utiliza frecuentemente en situaciones en las que el planificador piensa que no puede permitirse equivocarse (la planificación militar podría ser un ejemplo, al igual que la inversión de los ahorros de toda nuestra vida). El planificador elige una decisión que obtenga los mejores resultados posibles en el peor caso posible (más pesimista)» (véase la referencia bibliográfica 1).

Criterio de la pérdida de oportunidades minimax

La persona que tiene que tomar decisiones y quiere utilizar el *criterio de la pérdida de oportunidades minimax* debe imaginar que se encuentra en una situación en la que ha elegido una acción y se ha producido uno de los estados de la naturaleza. Puede mirar la decisión tomada con satisfacción o con decepción porque, tal como se han desarrollado las cosas, habría sido preferible una acción alternativa. La persona que toma decisiones determina entonces el «*pesar*» o *pérdida de oportunidades* de no tomar la mejor decisión, dado el estado de la naturaleza, y elabora una tabla de pérdidas.

Tabla de pérdidas de oportunidades

Supongamos que elaboramos una tabla de rendimientos de forma rectangular, en la que las filas corresponden a las acciones y las columnas a los estados de la naturaleza. Si se resta cada rendimiento de la tabla del rendimiento mayor *de su columna*, la tabla resultante se llama **tabla de pérdidas de oportunidades**.

Considerando la diferencia entre el rendimiento monetario efectivo de una decisión y el rendimiento óptimo correspondiente al mismo estado de la naturaleza, la persona que toma decisiones puede seleccionar la acción que *minimiza* la *máxima* pérdida.

Regla de decisión basada en el criterio criterio de la pérdida de oportunidades minimax

Dada la tabla de pérdidas, las acciones dictadas por el **criterio de la pérdida de oportunidades minimax** se encuentran de la forma siguiente:

1. Se halla en cada fila (acción), la máxima pérdida.
2. Se elige la acción correspondiente al *mínimo* de estas pérdidas *máximas*.

El **criterio de la pérdida de oportunidades minimax** selecciona la acción cuya pérdida máxima es menor; es decir, el criterio de la pérdida de oportunidades minimax produce la menor pérdida de oportunidades posible que puede garantizarse.

Consideremos de nuevo el caso del fabricante de de teléfonos móviles del ejemplo 21.1. Mostraremos que se selecciona el proceso B mediante el criterio de la pérdida de oportunidades minimax. Supongamos que el nivel de demanda del nuevo producto es bajo. En ese caso, la mejor elección de una acción habría sido el proceso de producción C, que generaba un rendimiento de 100.000 $. Si se hubiera elegido esa acción, el fabricante habría tenido una pérdida de 0. Si se hubiera elegido el proceso A, el beneficio resultante habría sido de 70.000 $ solamente. El grado de pérdida del fabricante, en este caso, es la diferencia entre el mejor rendimiento que podría haberse obtenido (100.000 $) y el rendimiento de lo que finalmente fue una peor elección. Por lo tanto, la pérdida sería igual a 100.000 $ − 70.000 $ = 30.000 $. Asimismo, dada una baja demanda, si se hubiera elegido el proceso B, la pérdida sería

$$100.000\ \$ - 80.000\ \$ = 20.000\ \$$$

Continuando de esta forma, se calculan las pérdidas que implican el nivel moderado de demanda y el nivel alto de demanda. En cada caso, la pérdida es igual a 0 en el caso de la mejor elección de la acción (el proceso C en el caso de la demanda moderada y el A en el

de la demanda alta). Estas pérdidas de oportunidades por no tomar la mejor decisión, dado un estado de la naturaleza, se muestran en la Tabla 21.5, cuya última columna indica la máxima pérdida de un proceso dado.

Es evidente que el criterio de la pérdida de oportunidades minimax selecciona el proceso de producción B, ya que la pérdida máxima de este proceso es la menor de los procesos A, B y C.

Ni el criterio maximin ni el criterio de la pérdida de oportunidades minimax permiten a la persona que toma las decisiones introducir en el proceso de toma de decisiones sus opiniones personales como la probabilidad de que se produzcan los estados de la naturaleza. Dado que la mayoría de los problemas empresariales prácticos se producen en un entorno con el que está al menos algo familiarizado el responsable de tomar las decisiones, eso representa un despilfarro de pericia. En el siguiente apartado analizamos las probabilidades de los resultados de cada acción alternativa.

Tabla 21.5. Aplicación del criterio de la pérdida de oportunidades minimax al ejemplo 21.1.

Acción	Estado de la naturaleza			Pérdida
Proceso de producción	Demanda baja	Demanda moderada	Demanda alta	Pérdida máxima de cada proceso
A	30.000	5.000	0	30.000
B	20.000	5.000	20.000	**20.000 (mínimo)**
C	0	0	40.000	40.000

EJERCICIOS

Ejercicios básicos

21.3. Considere el ejercicio 21.1, en el que un inversor está considerando tres alternativas —un certificado de depósito, un fondo de acciones de bajo riesgo y un fondo de acciones de alto riesgo— para hacer una inversión de 20.000 $. Considera tres estados de la naturaleza posibles:

s_1: mercado de valores fuerte
s_2: mercado de valores moderado
s_3: mercado de valores débil

La tabla de rendimientos (en dólares) es la siguiente:

Acción	Estado de la naturaleza		
Alternativas de inversión posibles	s_1	s_2	s_3
Certificado de depósito	1.200	1.200	1.200
Fondo de acciones de bajo riesgo	4.300	1.200	−600
Fondo de acciones de alto riesgo	6.600	800	−1.500

a) ¿Qué acción se selecciona mediante el criterio maximin?

b) ¿Qué acción se selecciona mediante el criterio de la pérdida de oportunidades minimax?

21.4. Considere el fabricante de desodorantes del ejercicio 21.2 que está a punto de ampliar la capacidad de producción para fabricar un nuevo producto. Tiene cuatro procesos de producción alternativos. La tabla adjunta muestra los beneficios estimados, en dólares, de estos procesos correspondientes a tres niveles de demanda del producto posibles.

Acción	Estado de la naturaleza		
Proceso de producción	Demanda baja	Demanda moderada	Demanda alta
A	100.000	350.000	900.000
B	150.000	400.000	700.000
C	250.000	400.000	600.000
D	250.000	400.000	550.000

a) ¿Qué acción se selecciona mediante el criterio maximin?

b) ¿Qué acción se selecciona mediante el criterio de la pérdida de oportunidades minimax?

21.5. Otro criterio para seleccionar una decisión es el **criterio maximax**, llamado a veces **criterio del**

optimismo. Este criterio elige la acción que tiene el mayor rendimiento posible.

a) ¿Qué acción elegiría el fabricante de teléfonos móviles con los rendimientos de la Tabla 21.2 según este criterio?

b) ¿Y el inversor del ejemplo 21.2?

Ejercicios aplicados

21.6. El fabricante de teléfonos móviles del ejemplo 21.1 tiene tres acciones admisibles: los procesos A, B y C. Cuando se consideran conjuntamente, se elige el proceso B según el criterio de la pérdida de oportunidades minimax. Suponga ahora que hay una cuarta alternativa admisible, el proceso de producción E. Los rendimientos estimados de esta acción son 60.000 $ en el caso en el que la demanda es baja, 115.000 $ en el que es moderada y 220.000 $ en el que es alta. Demuestre que cuando se consideran conjuntamente los procesos A, B, C y E, se elige el A según el criterio de la pérdida de oportunidades minimax. Por lo tanto, aunque la introducción del proceso E entre las acciones no lleva a elegir ese proceso, sí lleva a elegir una acción diferente a la que se habría elegido. Comente el atractivo intuitivo del criterio de la pérdida de oportunidades minimax a la luz de este ejemplo.

21.7. Considere un problema de decisión que tiene dos acciones posibles y dos estados de la naturaleza.

a) Ponga un ejemplo de una tabla de rendimientos en la que ambas acciones son admisibles y se elige la misma acción tanto según el criterio maximin como según el criterio de la pérdida de oportunidades minimax.

b) Ponga un ejemplo de una tabla de rendimientos según la cual se eligen diferentes acciones según el criterio maximin y según el criterio de la pérdida de oportunidades minimax.

21.8. Considere un problema de decisión que tiene dos acciones admisibles y dos estados de la naturaleza posibles. Describa la forma que debe tener la tabla de rendimientos para que se elija la misma acción con el criterio maximin que con el criterio de la pérdida de oportunidades minimax.

21.9. Un empresario tiene la posibilidad de abrir una zapatería en centro comercial consolidado y de éxito. Pero también puede abrirla con un coste más bajo en un nuevo centro, que acaba de inaugurarse. Si resulta que el nuevo centro tiene mucho éxito, se espera que los beneficios anuales que obtenga la zapatería por estar en ese centro sean de 130.000 $. Si el centro sólo tiene un éxito moderado, los beneficios anuales serían de 60.000 $. Si no tiene éxito, la pérdida anual sería de 10.000 $. Los beneficios que se espera obtener abriendo la zapatería en el centro comercial consolidado también dependen en alguna medida del grado de éxito del nuevo, ya que los clientes podrían sentirse atraídos por él. Si el nuevo centro no tuviera éxito, los beneficios esperados de la zapatería situada en el centro consolidado serían de 90.000 $. Sin embargo, si el nuevo centro tuviera un éxito moderado, los beneficios esperados serían de 70.000 $, mientras que si tuviera mucho éxito serían de 30.000 $.

a) Elabore la tabla de rendimientos del problema de toma de decisiones del dueño de esta zapatería.

b) ¿Qué acción se elige según el criterio maximin?

c) ¿Qué acción se elige según el criterio de la pérdida de oportunidades minimax?

21.3. Valor monetario esperado; TreePlan

Un importante ingrediente del análisis de muchos problemas de toma de decisiones empresariales probablemente sea la valoración que hace el responsable de tomarlas de la probabilidad de que se produzcan los distintos estados de la naturaleza relevantes en la determinación del rendimiento final. Los criterios analizados en el apartado 21.2 no permiten incorporar este tipo de valoración al proceso de toma de decisiones. Sin embargo, un directivo casi siempre tendrá una buena impresión del entorno en el que se toma la decisión y querrá tenerlo en cuenta antes de decidir un curso de acción. El análisis de este apartado supone que cada estado de la naturaleza tiene una *probabilidad* de ocurrencia y demostrará cómo se emplean estas probabilidades para tomar una decisión.

Generalmente, cuando hay H estados de la naturaleza posibles, debe asignarse una probabilidad a cada uno. Estas probabilidades se representan por medio de P_1, P_2, ..., P_H, por lo que la probabilidad P_j corresponde al estado de la naturaleza s_j. La Tabla 21.6 muestra el planteamiento general de este problema de toma de decisiones.

Tabla 21.6. Rendimientos con probabilidades de los estados de la naturaleza.

Acción	**Estado de la naturaleza**			
a_i/s_i	$s_1(P_1)$	$s_2(P_2)$	$\cdots$	$s_H(P_H)$
a_1	M_{11}	M_{12}	$\cdots$	M_{1H}
a_2	M_{21}	M_{22}	$\cdots$	M_{2H}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_K	M_{K1}	M_{K2}	$\cdots$	M_{KH}

Dado que debe ocurrir uno y sólo uno de los estados de la naturaleza, estas probabilidades suman necesariamente 1, por lo que

$$\sum_{j=1}^{H} P_j = 1$$

Cuando la persona que toma la decisión elige una acción, verá que cada elección tiene una probabilidad específica de recibir el rendimiento correspondiente y, por lo tanto, podrá calcular el *rendimiento esperado* de cada acción. El rendimiento esperado de esta acción es, pues, la suma de los rendimientos individuales, ponderados por sus probabilidades. Estos rendimientos esperados a menudo se llaman **valores monetarios esperados** de las acciones.

Criterio del valor monetario esperado (*VME*)

Supongamos que una persona que tiene que tomar una decisión tiene K acciones posibles, a_1, a_2, ..., a_K y se enfrenta a H estados de la naturaleza. Sea M_{ij} el rendimiento correspondiente a la i-ésima acción y el j-ésimo estado y P_j la probabilidad de que ocurra el j-ésimo estado de la naturaleza, cumpliéndose que $\sum_{j=1}^{H} P_j = 1$. El **valor monetario esperado** de la acción a_i, $VME(a_i)$, es

$$VME(a_i) = P_1M_{i1} + P_2M_{i2} + \cdots + P_HM_{iH} = \sum_{j=1}^{H} P_jM_{ij} \qquad \textbf{(21.1)}$$

El **criterio del valor monetario esperado** adopta la acción que tiene el mayor valor monetario esperado; es decir, dada una elección entre acciones alternativas, el criterio del *VME* dicta la elección de la acción cuyo *VME* es mayor.

Volvamos al fabricante de teléfonos móviles del ejemplo 21.1 y calculemos el *VME* de cada uno de los procesos de producción. El fabricante probablemente tendrá alguna experiencia en el mercado de su producto y, basándose en esa experiencia, podría hacerse una idea de la probabilidad de que la demanda sea baja, moderada o alta. Supongamos que sabe que el 10 por ciento de todas las veces que se ha introducido antes este tipo de producto

tuvo una baja demanda, el 50 por ciento tuvo una demanda moderada y el 40 por ciento tuvo una demanda alta. A falta de más información, es razonable postular, en el caso de la introducción de este nuevo tipo de teléfono móvil, las siguientes probabilidades de los estados de la naturaleza:

$$P_1 = P(s_1) = \text{probabilidad de que la demanda sea baja} = 0{,}1$$
$$P_2 = P(s_2) = \text{probabilidad de que la demanda sea moderada} = 0{,}5$$
$$P_3 = P(s_3) = \text{probabilidad de que la demanda sea alta} = 0{,}4$$

Dado que debe ocurrir uno y sólo uno de los estados de la naturaleza, estas probabilidades suman necesariamente 1; es decir, los estados de la naturaleza son mutuamente excluyentes y colectivamente exhaustivos. Estas probabilidades se añaden a la tabla de rendimientos (Tabla 21.2) y dan la Tabla 21.7.

Tabla 21.7. Rendimientos y probabilidades de los estados de la naturaleza correspondientes al ejemplo 21.1 del fabricante de teléfonos móviles.

Acción	**Estado de la naturaleza**		
Proceso de producción	**Demanda baja (P = 0,10)**	**Demanda moderada (P = 0,50)**	**Demanda alta (P = 0,40)**
A	70.000	120.000	200.000
B	80.000	120.000	180.000
C	100.000	125.000	160.000

Si el fabricante de teléfonos móviles adopta el proceso de producción A, recibirá un rendimiento de 70.000 $ con una probabilidad de 0,1, 120.000 $ con una probabilidad de 0,5 y 200.000 $ con una probabilidad de 0,4. En el caso del fabricante de teléfonos móviles, los valores monetarios esperados de las tres acciones admisibles son:

$$VME \text{ (Proceso A)} = (0{,}1)(70.000) + (0{,}5)(120.000) + (0{,}4)(200.000) = 147.000\ \$$$
$$VME \text{ (Proceso B)} = (0{,}1)(80.000) + (0{,}5)(120.000) + (0{,}4)(180.000) = 140.000\ \$$$
$$VME \text{ (Proceso C)} = (0{,}1)(100.000) + (0{,}5)(125.000) + (0{,}4)(160.000) = 136.500\ \$$$

El fabricante de teléfonos móviles elegiría el proceso de producción A. Es interesante señalar que ni el criterio maximin ni el criterio de la pérdida de oportunidades minimax llevan a esta elección. Sin embargo, se ha añadido la información de que parece que hay muchas más probabilidades de que el nivel de demanda sea alto que de que sea bajo, por lo que el proceso A es una opción relativamente atractiva.

Árboles de decisión

El análisis de un problema de decisión por medio del criterio del valor monetario esperado puede representarse gráficamente mediante un mecanismo llamado **árbol de decisión**. Cuando se analizan decisiones en condiciones de riesgo, el diagrama del árbol es un instrumento gráfico que obliga a la persona que toma las decisiones a «examinar todos los resultados posibles, incluidos los desfavorables. También la obliga a tomar decisiones de una manera lógica y consecutiva» (véase la referencia bibliográfica 4). Los árboles de decisión son especialmente útiles cuando debe tomarse una sucesión de decisiones. Todos contienen

- □ **Nodos de decisión (o de acción).** Estos cuadrados indican que debe tomarse una decisión y a veces se llaman nodos cuadrados.
- ○ **Nodos de sucesos (estados de la naturaleza).** Estos empalmes circulares, de los que salen *ramas*, representan un estado de la naturaleza posible, al que se asigna la probabilidad correspondiente. Estos nodos a veces se llaman nodos circulares.
- | **Nodos terminales.** Una barra vertical representa el final de la rama decisión-suceso. Originalmente, se utilizaba un triángulo para representar este punto. A veces no se representa de ninguna forma.

Después de definir rigurosamente un problema, la persona que toma la decisión traza el árbol de decisión, asigna probabilidades a los sucesos (estados de la naturaleza) posibles y estima el rendimiento de cada combinación decisión-suceso posible (cada combinación de acción y estado de la naturaleza). Ahora el responsable de tomar la decisión está preparado para encontrar la decisión óptima. Ese proceso se llama «resolver el árbol» (véase la referencia bibliográfica 1). Para resolver un árbol de decisión, hay que trabajar hacia atrás (lo que se llama *plegar* el árbol). Calculemos el valor monetario esperado (*VME*) de cada estado de la naturaleza comenzando por la parte situada más a la derecha del árbol de decisión y retrocediendo hasta los nodos de decisión situados a la izquierda.

La Figura 21.1 muestra un diagrama de árbol del fabricante de teléfonos móviles. Se dan los siguientes pasos para elegir la acción que tiene el mayor *VME*:

1. Comenzando por el lado izquierdo de la figura, vemos que salen ramas del **nodo de decisión** (indicado con un cuadrado) que representan las tres acciones posibles: proceso A, proceso B y proceso C. A continuación, salen los **nodos de sucesos** (representados por un círculo), de los que salen ramas que representan los estados de la naturaleza (los niveles de demanda) posibles.

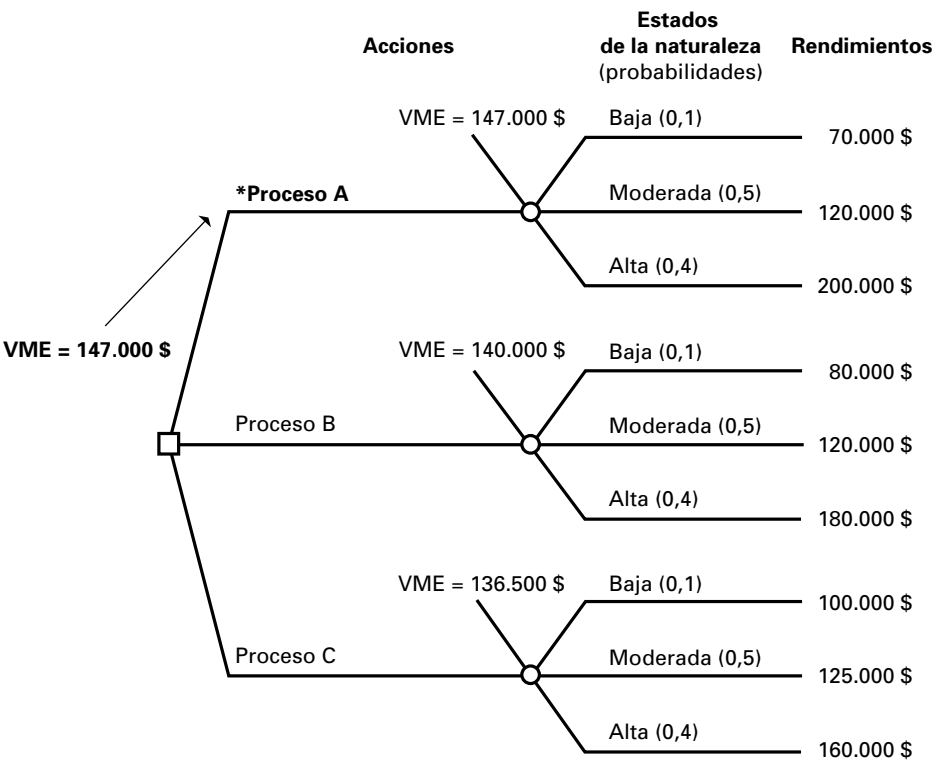


Figura 21.1. Árbol de decisión del fabricante de teléfonos móviles (*la acción que tiene el máximo *VME*).

2. Se asigna la *probabilidad correspondiente* a cada estado de la naturaleza (baja, moderada, alta).
3. En la parte situada más a la derecha se insertan los *rendimientos* correspondientes a las combinaciones acción-estado de la naturaleza.
4. Los cálculos se realizan de *derecha a izquierda*, comenzando por estos rendimientos. Se calcula en cada empalme circular la suma de las probabilidades de las distintas ramas multiplicadas por su rendimiento. De esa manera, se obtiene el *VME de cada acción*.
5. La *decisión óptima* es la que tiene el *VME* más alto y se indica en el punto en el que hay un cuadrado. Por lo tanto, se elige el proceso A mediante el criterio del valor monetario esperado. La elección de esta acción da como resultado un valor monetario esperado o beneficio esperado de 147.000 $ para el fabricante de teléfonos móviles.

La utilización de TreePlan para resolver un árbol de decisión

TreePlan, desarrollado por Michael Middleton (véase la referencia bibliográfica 3) e incluido en este libro, es un complemento de Excel que puede utilizarse para trazar árboles de decisión. Calcula el *VME* e indica la decisión óptima. Entre en la página web www.treeplan.com para la documentación y los detalles que permitirán continuar utilizando este complemento una vez concluido este curso (véase la referencia bibliográfica 5).

EJEMPLO 21.3. Oportunidad de inversión (criterio del *VME*)

El inversor del ejemplo 21.2 tenía que decidir entre una inversión a un tipo de interés fijo y una cartera de acciones. Supongamos que este inversor es, de hecho, muy optimista sobre la futura evolución del mercado de valores y cree que la probabilidad de que el mercado esté boyante es 0,6, mientras que la probabilidad de cada uno de los otros dos estados es 0,2. La tabla adjunta muestra los rendimientos y las probabilidades de los estados de la naturaleza:

Acción	**Estado de la naturaleza**		
Inversión	**Estado boyante ($P = 0{,}60$)**	**Estado estable ($P = 0{,}20$)**	**Estado deprimido ($P = 0{,}20$)**
Tipo de interés fijo	1.200	1.200	1.200
Cartera de acciones	2.500	500	−1.000

¿Qué inversión debe elegir según el criterio del valor monetario esperado?

Solución

Dado que el rendimiento de la inversión a un tipo de interés fijo es de 1.200 $, independientemente de lo que ocurra en la bolsa de valores, el valor monetario esperado de esta inversión es 1.200 $. El *VME* de la cartera de acciones es

$$VME \text{ (Cartera de acciones)} = (0{,}6)(2.500) + (0{,}2)(500) + (0{,}2)(-1.000) = 1.400\ \$$$

Dado que éste es el valor monetario esperado más alto, el inversor elegirá la *cartera de acciones ordinarias*, según el criterio del valor monetario esperado.

Resolvamos ahora este ejemplo con el TreePlan. Una vez instalado el TreePlan, la forma más fácil de acceder a él es abrir una nueva hoja de cálculo Excel y pulsar Ctrl-t (el árbol comenzará donde aparezca el cursor; asegúrese de que tiene suficiente espacio para la tabla de decisión y para el árbol). Pulse en «New Tree» y aparecerá el árbol con dos nodos de decisión (Figura 21.2). El árbol de decisión completo se encuentra en la Figura 21.3.

A continuación, analizamos un problema que requiere una *sucesión* de decisiones.

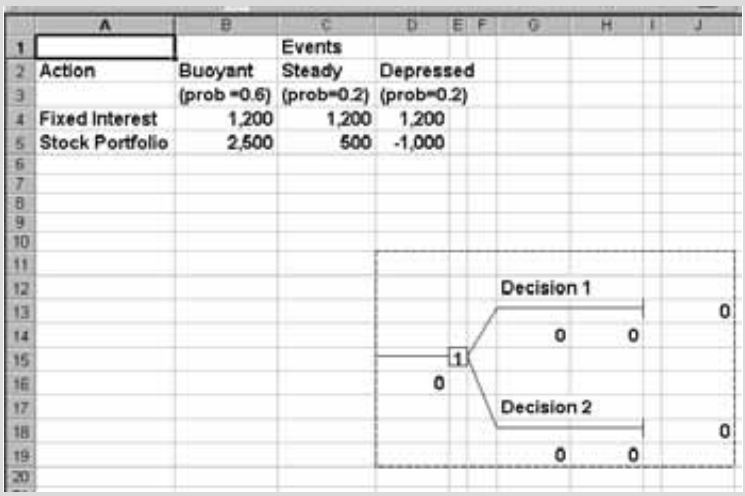


Figura 21.2. Inicio del programa TreePlan.

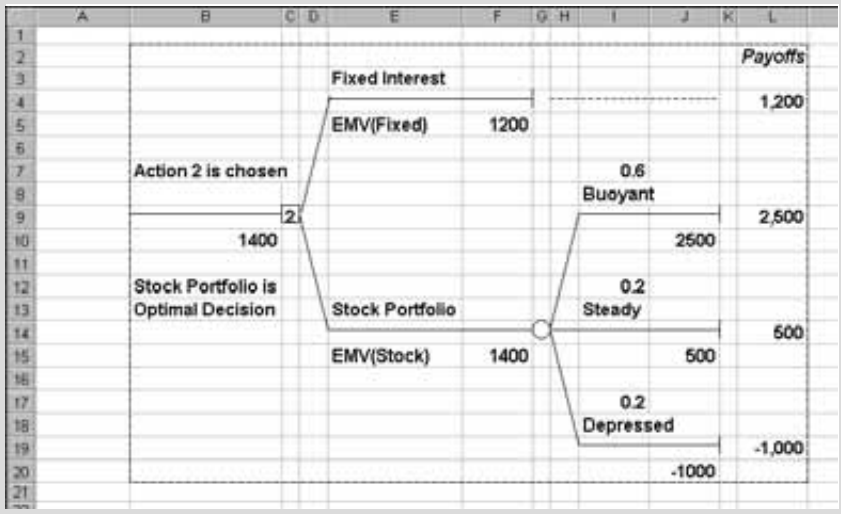


Figura 21.3. Árbol de decisión del ejemplo 21.3 elaborado utilizando TreePlan; decisión óptima: seleccionar la cartera de acciones.

EJEMPLO 21.4. Fabricante de medicamentos (criterio del *VME*)

Un fabricante de medicamentos tiene los derechos de patente de una nueva fórmula que reduce los niveles de colesterol. El fabricante puede vender la patente por 50.000 $ o realizar pruebas intensivas sobre la eficacia del medicamento. El coste de realizar estas pruebas es de 10.000 $. Si se observa que el medicamento es ineficaz, no se comercializará y el coste de las pruebas se considerará una pérdida. Hasta ahora, las pruebas realizadas con medicamentos de este tipo han sido eficaces en un 60 por ciento e ineficaces en un 40 por ciento.

Si las pruebas revelaran ahora que el medicamento es eficaz, el fabricante tiene de nuevo dos opciones. Puede vender los derechos de patente y los resultados de las pruebas por 120.000 $ o puede comercializar él mismo el medicamento. Si lo comercializa, se estima que los beneficios generados por las ventas (excluidos los costes de las pruebas) ascenderán a 180.000 $ si la campaña de ventas tiene mucho éxito, pero sólo a 90.000 $ si tiene un éxito moderado. Se estima que estos dos niveles de penetración en el mercado son igual de probables. Según el criterio del valor monetario esperado, ¿qué debe hacer el fabricante del medicamento?

Solución

Lo mejor es abordar el problema construyendo un árbol de decisión. La Figura 21.4 muestra el árbol completo.

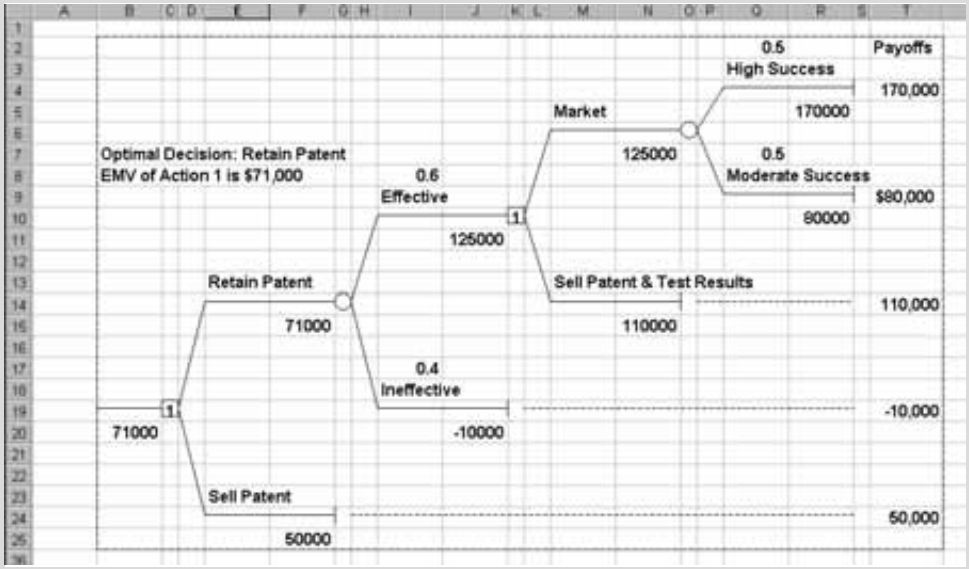


Figura 21.4. Árbol de decisión del ejemplo 21.4; decisión óptima: conservar la patente y, si las pruebas demuestran que el medicamento es eficaz, comercializarlo (*VME* = 71.000 $).

El fabricante puede decidir vender la patente, en cuyo caso no tiene que hacer nada más, o quedársela y realizar pruebas sobre la eficacia del medicamento. Hay dos estados de la naturaleza posibles: el medicamento es eficaz (con una probabilidad de 0,6) o es ineficaz (con una probabilidad de 0,4). En el segundo caso, ahí termina todo. Sin embargo, si el medicamento demuestra ser eficaz, hay que tomar una segunda decisión: comercializarlo o vender los derechos de patente y los resultados de las pruebas. Si se

adopta la primera opción, el nivel de éxito de la comercialización determina el resultado final, que puede ser moderado o alto (cada uno con una probabilidad de 0,5).

A continuación, se examinan los rendimientos de todas las combinaciones acción-estado de la naturaleza. Comencemos por la parte inferior del árbol de decisión. Si la decisión inicial del fabricante es vender la patente, recibe 50.000 $. Si se queda con ella, pero el medicamento resulta ineficaz, el fabricante tiene una pérdida de 10.000 $, que es el coste de las pruebas. Esta pérdida se muestra como un rendimiento negativo de esa cuantía. Si se observa que el medicamento es eficaz y se vende la patente y los resultados de las pruebas, el fabricante recibe 120.000 $, de los que debe restarse el coste de las pruebas, por lo que queda un rendimiento de 110.000 $. Por último, si se comercializa el medicamento, los rendimientos en los casos de éxito moderado y grande son 90.000 $ y 180.000 $, respectivamente, menos el coste de las pruebas, por lo que quedan 80.000 $ y 170.000 $, respectivamente.

Una vez llegados a este punto, el problema de decisión se resuelve yendo hacia atrás de derecha a izquierda. Este paso es necesario, ya que no puede saberse cuál es la acción que debe elegirse en el primer punto de decisión hasta que se conoce el valor monetario esperado de la mejor opción en el segundo punto de decisión.

Comencemos, pues, suponiendo que inicialmente se conserva la patente y que las pruebas demuestran que el medicamento es eficaz. Si se vende la patente y los resultados de las pruebas, se obtiene un beneficio de 110.000 $. El valor monetario esperado de la comercialización del medicamento es

$$VME = (0{,}5)(170.000) + (0{,}5)(80.000) = 125.000\ \$$$

Dado que es de más de 110.000 $, la mejor opción en esta fase, según el criterio del valor monetario esperado, es comercializar el medicamento. Esta cantidad se introduce, pues, en el nodo cuadrado del segundo punto de decisión y se considera que es el rendimiento que obtiene el fabricante si su decisión inicial es conservar la patente y las pruebas indican que el medicamento es eficaz. Aquí mostramos la tabla de rendimientos correspondiente a la decisión inicial con las probabilidades de los estados de la naturaleza. El valor monetario esperado de la venta de la patente son los 50.000 $ seguros, mientras que el valor monetario esperado de conservar la patente es $(0{,}6)(125.000) + + (0{,}4)(-10.000) = 71.000$ $. En ese caso, según el criterio del valor monetario esperado, debe conservarse la patente.

	Estado de la naturaleza	
Acción	**Medicamento eficaz ($P = 0{,}60$)**	**Medicamento ineficaz ($P = 0{,}40$)**
Conservar la patente	125.000	− 10.000
Vender la patente	50.000	50.000

Si el objetivo del fabricante es maximizar el valor monetario esperado (es decir, el beneficio esperado), debe conservar la patente. Si las pruebas demuestran que el medicamento es eficaz, el fabricante debe comercializarlo. Esta estrategia genera un beneficio esperado de 71.000 $.

En la Figura 21.4 se obtiene el mismo resultado utilizando el TreePlan.

Análisis de sensibilidad

En el caso del fabricante de teléfonos móviles, éste ha seleccionado el proceso de producción A utilizando el criterio del valor monetario esperado. Esta decisión se basa en el rendimiento estimado de cada combinación acción-estado de la naturaleza y en la probabilidad estimada de que ocurra cada estado de la naturaleza. Sin embargo, a menudo la persona que tiene que tomar una decisión no está segura de esas estimaciones, por lo que es útil preguntarse en qué intervalo de especificaciones de un problema de decisión es óptima una determinada acción según el criterio del valor monetario esperado. El **análisis de sensibilidad** trata de responder a esas preguntas y el caso más sencillo es aquel en el que se permite que varíe una única especificación del problema.

Para ilustrarlo, supongamos que el fabricante de teléfonos móviles está de acuerdo con que la probabilidad de que la demanda sea alta es de 0,4, pero está menos seguro en el caso de los otros dos estados de la naturaleza. Sea P la probabilidad de que la demanda sea baja, por lo que la probabilidad de que sea moderada debe ser $(0{,}6 - P)$. Según el criterio del valor monetario esperado, ¿en qué intervalo de valores de P sería óptima la adopción del proceso A? Utilizando los rendimientos de la Tabla 21.7, los valores monetarios esperados son

$$VME(\text{A}) = (P)(70.000) + (0{,}6 - P)(120.000) + (0{,}4)(200.000) = 152.000 - 50.000P$$
$$VME(\text{B}) = (P)(80.000) + (0{,}6 - P)(120.000) + (0{,}4)(180.000) = 144.000 - 40.000P$$
$$VME(\text{C}) = (P)(100.000) + (0{,}6 - P)(125.000) + (0{,}4)(160.000) = 139.000 - 25.000P$$

La elección del proceso A seguirá siendo óptima siempre que el *VME* correspondiente sea mayor que el de cada uno de los otros dos procesos. Por lo tanto, para que se prefiera el proceso A al proceso B, debe cumplirse que

$$152.000 - 50.000P \geqslant 144.000 - 40.000P$$

o sea

$$8.000 \geqslant 10.000P$$

por lo que

$$P \leqslant 0{,}8$$

Este resultado debe cumplirse, ya que, según nuestros supuestos, la probabilidad de que la demanda sea baja no puede ser de más de 0,6. Asimismo, para que se prefiera el proceso A al proceso B,

$$152.000 - 50.000P \geqslant 139.000 - 25.000P$$

o sea

$$13.000 \geqslant 25.000P$$

por lo que

$$P \leqslant 0{,}52$$

Si los rendimientos son los que indica la Tabla 21.7 y la probabilidad de que la demanda sea alta es 0,4, entonces la mejor elección según el criterio del valor monetario esperado es el proceso de producción A, siempre que la probabilidad de que la demanda sea baja no sea de más de 0,52.

Supongamos ahora que el fabricante de teléfonos móviles no está seguro del rendimiento estimado de 200.000 $ si elige el proceso A y la demanda es alta. Veamos en qué intervalo de rendimientos el proceso A será la elección óptima, cuando se mantienen todas las demás especificaciones del problema en sus niveles iniciales, mostrados en la Tabla 21.7. Si M es el rendimiento del proceso A cuando la demanda es alta, el valor monetario esperado de este proceso es

$$VME(\text{A}) = (0{,}1)(70.000) + (0{,}5)(120.000) + 0{,}4M = 67.000 + 0{,}4M$$

Los valores monetarios esperados de los procesos B y C son, al igual que antes, de 140.000 $ y 136.500 $. Por lo tanto, el proceso A será la mejor elección según el criterio del valor monetario esperado, siempre que

$$67.000 + 0{,}4M \geqslant 140.000$$

o sea

$$0{,}4M \geqslant 73.000$$

o sea

$$M \geqslant 182.500$$

Si todas las demás especificaciones siguen siendo las que muestra la Tabla 21.7, se seleccionará el proceso de producción A según el criterio del valor monetario esperado, siempre que el rendimiento del proceso A cuando la demanda es alta sea al menos de 182.500 $.

EJERCICIOS

Ejercicios aplicados

21.10. Un estudiante ya tiene ofertas de trabajo. Ahora debe decidir si va a otra entrevista en otra empresa. Considera que el tiempo y el esfuerzo de acudir a otra entrevista tienen un coste de 500 $, en los que incurrirá independientemente de que acepte el trabajo que ofrece esa empresa. Si el empresario ofrece un puesto preferible a sus demás alternativas, se consideraría que es un beneficio que vale 5.000 $ (de los que debe restarse el coste de 500 $). De lo contrario, habría despilfarrado el tiempo y el esfuerzo.

a) Elabore la tabla de rendimientos del problema de decisión del estudiante.

b) Suponga que el estudiante cree que la probabilidad de que este empresario le ofrezca un trabajo preferible a otras alternativas es de 0,05. Según el criterio del valor monetario esperado, ¿debe ir a ver a este empresario?

21.11. Un directivo tiene que elegir entre dos acciones, a_1 y a_2. Hay dos estados de la naturaleza posibles, s_1 y s_2. La tabla adjunta muestra los rendimientos. Si el directivo cree que los dos estados de la naturaleza son igual de probables, ¿qué acción debe elegir, según el criterio del valor monetario esperado?

	Estado de la naturaleza	
Acción	s_1	s_2
a_1	72.000	51.000
a_2	78.000	47.000

21.12. El inversor del ejercicio 21.1 cree que la probabilidad de que la bolsa de valores esté fuerte es de 0,2, la probabilidad de que esté moderada es de 0,5 y la probabilidad de que esté débil es 0,3.

a) ¿Qué acción debe elegir según el criterio del valor monetario esperado?

b) Construya el árbol de decisión del problema del inversor.

21.13. El fabricante de desodorantes del ejercicio 21.2 sabe que históricamente el 30 por ciento de los nuevos productos de este tipo ha tenido una elevada demanda, el 40 por ciento ha tenido una

demanda moderada y el 30 por ciento ha tenido una demanda baja.

a) Según el criterio del valor monetario esperado, ¿qué proceso de producción debe utilizarse?

b) Construya el árbol de decisión del problema de este fabricante.

21.14. Considere un problema de decisión con dos acciones admisibles y dos estados de la naturaleza posibles, que tienen ambos la misma probabilidad de ocurrir.

a) Averigüe si es verdadera o falsa cada una de las siguientes afirmaciones en un problema de ese tipo.

i. La acción elegida según el criterio del valor monetario esperado siempre será igual que la acción elegida según el criterio maximin.

ii. La acción elegida según el criterio del valor monetario esperado siempre será igual que la acción elegida según el criterio de la pérdida de oportunidades minimax.

iii. La acción elegida según el criterio del valor monetario esperado siempre será aquella que tenga el mayor rendimiento medio posible.

b) ¿Sería su respuesta sobre la afirmación (iii) del apartado (a) la misma si los dos estados de la naturaleza no tuvieran la misma probabilidad de ocurrir?

21.15. Un problema de decisión tiene K acciones posibles y H estados de la naturaleza posibles. Si una de estas acciones es inadmisible, demuestre que no puede elegirse según el criterio del valor monetario esperado.

21.16. El empresario del ejercicio 21.9 cree que la probabilidad de que el nuevo centro comercial tenga mucho éxito es de 0,4, que la probabilidad de que tenga un éxito moderado es de 0,4 y que la probabilidad de que no tenga éxito es de 0,2.

a) Según el criterio del valor monetario esperado, ¿dónde debe abrir la zapatería?

b) Construya el árbol de decisión.

21.17. Vuelva al problema de decisión de los ejercicios 21.1, 21.3 y 21.12. Este inversor está de acuerdo con la valoración de que la probabilidad de que el mercado esté fuerte es de 0,2. Sin embargo, está menos seguro de las valoraciones de la probabilidad de los otros dos estados de la naturaleza. ¿En qué intervalo de probabilidades de que el mercado de valores esté débil da el criterio del valor monetario esperado la elección de la acción del ejercicio 21.12?

21.18. Vuelva al problema del fabricante de desodorantes de los ejercicios 21.2, 21.4 y 21.13.

a) El fabricante está de acuerdo con la valoración de que la probabilidad de que la demanda sea baja es de 0,3, pero está menos seguro de las probabilidades de los otros dos niveles de demanda. ¿En qué intervalo de probabilidades de que la demanda sea moderada generará el criterio del valor monetario esperado la elección de la acción del ejercicio 21.13?

b) Considere dado el resto de las especificaciones del problema de los ejercicios 21.2 y 21.13. ¿En qué intervalo de beneficios de una demanda alta cuando se utiliza el proceso A dará el criterio del valor monetario esperado la elección de la acción del ejercicio 21.13?

21.19. Vuelva al problema del empresario de los ejercicios 21.9 y 21.16.

a) El dueño de la zapatería está contento con la valoración de que la probabilidad de que el nuevo centro comercial no tenga éxito es de 0,2, pero está menos seguro de las valoraciones de la probabilidad de los otros dos estados de la naturaleza. ¿En qué intervalo de probabilidades de que el nuevo centro comercial tenga mucho éxito llevará el criterio del valor monetario esperado a la elección de la acción del ejercicio 21.16?

b) Suponiendo que las demás especificaciones del problema son las de los ejercicios 21.9 y 21.16, ¿en qué intervalo de niveles de beneficios correspondientes a la instalación en el nuevo centro si resulta que tiene mucho éxito llevará el criterio del valor monetario esperado a la elección de la acción del ejercicio 21.16?

21.20. Un fabricante recibe habitualmente contratos para entregar grandes pedidos de piezas a la industria automovilística. El proceso de producción del fabricante es tal que cuando funciona correctamente, el 10 por ciento de todas las piezas producidas no satisface las especificaciones de la industria. Sin embargo, es propenso a tener un determinado fallo, cuya presencia puede comprobarse al comienzo de una serie de producción. Cuando el proceso funciona con este fallo, el 30 por ciento de las piezas producidas no satisface las especificaciones de la industria. El fabricante ofrece piezas para un contrato por el que obtendrá un beneficio de 20.000 $ si sólo

es defectuoso el 10 por ciento de las piezas y un beneficio de 12.000 $ si es defectuoso el 30 por ciento de las piezas. El coste de comprobar el fallo es de 1.000 $ y, si se observa que es necesaria una reparación, ésta cuesta otros 2.000 $. Si se incurre en estos costes, deben restarse del beneficio. Históricamente, se ha observado que el proceso de producción funciona correctamente el 80 por ciento del tiempo. El fabricante debe decidir si comprueba el proceso al comienzo de una serie de producción.

a) Según el criterio del valor monetario esperado, ¿cuál es la decisión óptima?
b) Construya el árbol de decisión.
c) Suponga que no se sabe cuál es la proporción de ocasiones en las que el proceso de producción funciona correctamente. ¿En qué intervalo de valores de esta proporción sería óptima la decisión seleccionada en el apartado (a) según el criterio del valor monetario esperado?

21.21. Un contratista tiene que decidir si presenta una oferta para la adjudicación de un proyecto de construcción. El coste de la preparación de la oferta es de 16.000 $. Incurrirá en este coste independientemente de que se le adjudique o no el contrato. El contratista pretende hacer una oferta que generará 110.000 $ de beneficios (menos el coste de la preparación de la oferta). Sabe que el 20 por ciento de las ofertas preparadas de esta forma ha tenido éxito.

a) Elabore la tabla de rendimientos.
b) ¿Debe prepararse y presentarse una oferta según el criterio del valor monetario esperado?
c) ¿En qué intervalo de probabilidades de que la oferta tenga éxito debe prepararse y presentarse una oferta según el criterio del valor monetario esperado?

21.22. El jueves por la tarde, el jefe de una pequeña sucursal de una agencia de alquiler de coches observa que tiene seis coches para alquilar al día siguiente. Sin embargo, puede pedir que le envíen más coches de la central con un coste de 20 $ cada uno. Cada coche que se alquila genera un beneficio esperado de 40 $ (el coste de envío del coche debe restarse de este beneficio). Cada cliente que pide un coche cuando no hay ninguno disponible se cuenta como una pérdida de 10 $ de fondo de comercio. Revisando los datos de los viernes anteriores, el jefe observa que el número de coches solicitados ha ido de 6 a 10; los porcentajes se muestran en la tabla adjunta. El jefe debe decidir si pide coches a la central y, en caso afirmativo, cuántos.

Número de pedidos	6	7	8	9	10
Porcentaje	10	30	30	20	10

a) Elabore la tabla de rendimientos.
b) Si se utiliza el criterio del valor monetario esperado, ¿cuántos coches deben pedirse?

21.23. Un contratista ha decidido presentar una oferta para la adjudicación de un proyecto. Las ofertas deben presentarse en múltiplos de 20.000 $. Se estima que la probabilidad de que se consiga el contrato con una oferta de 240.000 $ es de 0,3, la probabilidad de que se consiga con una oferta de 220.000 $ es de 0,3 y la probabilidad de que se acepte una oferta de 200.000 $ es de 0,5. Se piensa que cualquier oferta de menos de 200.000 $ tendrá éxito con toda seguridad y que cualquier oferta de más de 240.000 $ fracasará con toda seguridad. Si el fabricante consigue el contrato, debe resolver un problema de diseño con dos opciones posibles en esta fase. Puede contratar consultores externos, que le garantizarán una solución satisfactoria, por un precio de 80.000 $. O puede invertir 30.000 $ de sus propios recursos en un intento de resolver el problema internamente; si fracasa este intento, debe contratar a los consultores. Se estima que la probabilidad de resolver con éxito el problema internamente es de 0,6. Una vez que ha resuelto este problema, el coste adicional de cumplir el contrato es de 140.000 $.

a) Este contratista tiene potencialmente dos decisiones que tomar. ¿Cuáles son?
b) Construya el árbol de decisión.
c) ¿Cuál es el curso de acción óptimo según el criterio del valor monetario esperado?

21.24. Considere un problema de decisión con dos acciones, a_1 y a_2, y dos estados de la naturaleza, s_1 y s_2. Sea M_{ij} el rendimiento correspondiente a la acción a_i y el estado de la naturaleza s_j. Suponga que la probabilidad de que ocurra el estado de la naturaleza s_1 es P, por lo que la probabilidad de que ocurra el estado s_2 es $(1 - P)$.

a) Demuestre que se selecciona la acción a_1 según el criterio del *VME* si

$$P(M_{11} - M_{21}) > (1 - P)(M_{22} - M_{12})$$

b) Demuestre, pues, que si a_1 es una acción admisible, existe una probabilidad, P, de que se elija. Sin embargo, si a_1 no es admisible, no puede elegirse, cualquiera que sea el valor de P.

21.4. Información muestral: análisis y valor bayesianos

Las decisiones que se toman en el mundo de la empresa pueden suponer a menudo una cantidad considerable de dinero y el coste de tomar una decisión subóptima puede ser elevado. Ésa es la razón por la que puede muy bien compensarle a la persona que tiene que tomar una decisión hacer un esfuerzo para conseguir la mayor información relevante posible antes de tomar la decisión. En concreto, querrá informarse lo más posible sobre las probabilidades de que ocurran los distintos estados de la naturaleza que determinan el rendimiento final.

Esta característica del examen detenido de un problema de decisión no ha sido evidente hasta ahora en nuestro análisis. El fabricante de teléfonos móviles del apartado 21.3 valoraba las probabilidades de que la demanda del nuevo teléfono móvil fuera baja, moderada y alta en 0,1, 0,5 y 0,4, respectivamente. Sin embargo, esta valoración no reflejaba más que las proporciones históricas logradas por otros productos anteriores. En la práctica, podría muy bien querer realizar algún estudio de mercado sobre las perspectivas del nuevo producto. Con ese estudio, estas *probabilidades a priori* o iniciales de los tres niveles de demanda pueden modificarse y generar nuevas probabilidades, llamadas *probabilidades a posteriori*. La información (en este caso, los resultados del estudio de mercado) que lleva a modificar las probabilidades de los estados de la naturaleza se llama *información muestral*.

Utilización del teorema de Bayes

En el Capítulo 4 explicamos el mecanismo para modificar las probabilidades *a priori* para obtener probabilidades *a posteriori*. Eso se hace por medio del **teorema de Bayes**, que reformulamos por comodidad en el marco de nuestro problema de decisión.

Teorema de Bayes

Sean s_1, s_2, ..., s_H H sucesos mutuamente excluyentes y colectivamente exhaustivos, que corresponden a los H estados de la naturaleza de un problema de decisión. Sea A algún otro suceso. Sea la probabilidad condicionada de que ocurra s_i, dado que ocurre A, $P(s_i|A)$ y la probabilidad de A, dado s_i, $P(A|s_i)$. El **teorema de Bayes** establece que la probabilidad condicionada de s_i, dado A, puede expresarse de la forma siguiente:

$$P(s_i|A) = \frac{P(A|s_i)P(s_i)}{P(A)} = \frac{P(A|s_i)P(s_i)}{P(A|s_1)P(s_1) + P(A|s_2)P(s_2) + \cdots + P(A|s_H)P(s_H)} \quad \textbf{(21.2)}$$

En la terminología de este apartado, $P(s_i)$ es la **probabilidad *a priori*** de s_i y se transforma en la **probabilidad *a posteriori*,** $P(s_i|A)$, dada la **información muestral** de que ha ocurrido el suceso A.

Supongamos ahora que el fabricante de teléfonos móviles contrata a una empresa de estudios de mercado para predecir el nivel de demanda de su nuevo producto. Naturalmente, la empresa le cobrará el estudio. Más adelante en este capítulo, veremos si el rendimiento justifica el coste. La empresa afirma que las perspectivas son «malas», «regulares» o «buenas» en función de su estudio. El análisis del historial de la empresa de estudios de mercado revela la calidad de sus predicciones anteriores en este campo. La Tabla 21.8

Tabla 21.8. Proporción de los distintos tipos de perspectivas según la empresa de estudios de mercado correspondientes a los distintos niveles de la demanda.

Acción	**Estado de la naturaleza**		
Valoración	**Demanda baja (s_1)**	**Demanda moderada (s_2)**	**Demanda alta (s_3)**
Malas	0,6	0,3	0,1
Regulares	0,2	0,4	0,2
Buenas	0,2	0,3	0,7

muestra la proporción de veces que la empresa dijo que las perspectivas eran malas, regulares o buenas correspondiente a cada nivel efectivo de demanda.

Por ejemplo, el 10 por ciento de las veces en que la demanda fue alta, la empresa dijo que las perspectivas eran «malas». Por lo tanto, en la notación de la probabilidad condicionada, representando la demanda baja, moderada y alta por medio de s_1, s_2 y s_3, respectivamente, se deduce que

$$P(\text{malas}|s_1) = 0{,}6 \qquad P(\text{malas}|s_2) = 0{,}3 \qquad P(\text{malas}|s_3) = 0{,}1$$

Es sólo una casualidad que la suma de $P(\text{malas}|s_1) = 0{,}6$, $P(\text{malas}|s_2) = 0{,}3$ y $P(\text{malas}|s_3) = 0{,}1$ sea 1,0. Estas probabilidades condicionadas no tienen que sumar 1. Tomemos, por ejemplo, el caso de «regulares»; obsérvese que la suma de $P(\text{regulares}|s_1) = 0{,}2$, $P(\text{regulares}|s_2) = 0{,}4$ y $P(\text{regulares}|s_3) = 0{,}2$ sólo es 0,8 y no 1,0.

Supongamos ahora que se consulta a la empresa de estudios de mercado y ésta dice que las perspectivas del teléfono móvil son «malas». Dada esta nueva información, las probabilidades *a priori*

$$P(s_1) = 0{,}1 \qquad P(s_2) = 0{,}5 \qquad P(s_3) = 0{,}4$$

de los tres niveles de demanda pueden modificarse utilizando el teorema de Bayes. En el caso de un bajo nivel de demanda, la probabilidad *a posteriori* es

$$P(s_1|\text{malas}) = \frac{P(\text{malas}|s_1)P(s_1)}{P(\text{malas}|s_1)P(s_1) + P(\text{malas}|s_2)P(s_2) + P(\text{malas}|s_3)P(s_3)}$$
$$= \frac{(0{,}6)(0{,}1)}{(0{,}6)(0{,}1) + (0{,}3)(0{,}5) + (0{,}1)(0{,}4)} = \frac{0{,}06}{0{,}25} = 0{,}24$$

Asimismo, en el caso de los otros dos niveles de demanda las probabilidades *a posteriori* son

$$P(s_2|\text{malas}) = \frac{(0{,}3)(0{,}5)}{0{,}25} = 0{,}6$$

$$P(s_3|\text{malas}) = \frac{(0{,}1)(0{,}4)}{0{,}25} = 0{,}16$$

A continuación, pueden utilizarse las probabilidades *a posteriori* para calcular los valores monetarios esperados. La Tabla 21.9 muestra los rendimientos (sin el coste del estudio), junto con las probabilidades *a posteriori* de los tres niveles de demanda. Esta tabla es simplemente una modificación de la 21.7, en la que se han sustituido las probabilidades *a priori* por las probabilidades *a posteriori*.

Tabla 21.9. Rendimientos del fabricante de teléfonos móviles y probabilidades *a posteriori* de los estados de la naturaleza, cuando la empresa de estudios de mercado dice que las perspectivas son «malas».

Acción	**Estado de la naturaleza**		
Proceso de producción	**Demanda baja (P = 0,24)***	**Demanda moderada (P = 0,60)***	**Demanda alta (P = 0,16)***
A	70.000	120.000	200.000
B	80.000	120.000	180.000
C	100.000	1250.000	160.000

* Probabilidades *a posteriori*.

Los valores monetarios esperados de los tres procesos de producción pueden hallarse exactamente de la misma forma que antes. Son los siguientes:

$$VME \text{ (Proceso A)} = (0{,}24)(70.000) + (0{,}60)(120.000) + (0{,}16)(200.000) = 120.800\ \$$$

$$VME \text{ (Proceso B)} = (0{,}24)(80.000) + (0{,}60)(120.000) + (0{,}16)(180.000) = 120.000\ \$$$

$$VME \text{ (Proceso C)} = (0{,}24)(100.000) + (0{,}60)(125.000) + (0{,}16)(160.000) = 124.600\ \$$$

Si la empresa de estudios de mercado considera que las perspectivas son «malas», entonces, según el criterio del valor monetario esperado, debe utilizarse el proceso de producción C. Según la valoración de la empresa de estudios de mercado, la demanda baja es mucho más probable y la demanda alta es considerablemente menos probable que antes. Este cambio de opinión sobre las perspectivas de mercado es suficiente para inducir al fabricante de teléfonos móviles a cambiar su preferencia por el proceso A (basada en las probabilidades *a priori*) por el proceso C.

Siguiendo el mismo razonamiento, es posible saber qué decisiones se tomarían si las perspectivas de éxito del mercado del teléfono móvil se consideraran «regulares» o «buenas». De nuevo, es posible hallar las probabilidades *a posteriori* de los tres niveles de demanda por medio del teorema de Bayes. Si se considera que las perspectivas son «regulares», son

$$P(s_1|\text{regulares}) = \frac{1}{15} \qquad P(s_2|\text{regulares}) = \frac{10}{15} \qquad P(s_3|\text{regulares}) = \frac{4}{15}$$

Si se considera que son «buenas»,

$$P(s_1|\text{buenas}) = \frac{2}{45} \qquad P(s_2|\text{buenas}) = \frac{15}{45} \qquad P(s_3|\text{buenas}) = \frac{28}{45}$$

Utilizando estas probabilidades *a posteriori*, calculamos por medio del programa Excel los valores monetarios esperados de cada uno de los procesos de producción correspondientes a cada valoración. La Tabla 21.10 contiene estos valores monetarios esperados. Podrían variar dependiendo del número de decimales utilizados para expresar las probabilidades *a posteriori*.

Como hemos mostrado antes, si la empresa de estudios de mercado afirma que las perspectivas son «malas», se prefiere el proceso C según el criterio del valor monetario esperado. Si hace otra predicción, se elegirá el proceso A, según este criterio.

Tabla 21.10. Valores monetarios esperados del fabricante de teléfonos móviles correspondientes a tres predicciones posibles realizadas por la empresa de estudios de mercado.

Acción	Estado de la naturaleza (perspectivas)		
Proceso de producción	**Malas**	**Regulares**	**Buenas**
A	120.800	138.000	167.556
B	120.000	133.333	155.556
C	124.600	132.667	145.667

Recuérdese que en el problema del fabricante de teléfonos móviles, cuando se utilizaban las probabilidades *a priori* de los niveles de demanda, la decisión óptima según el criterio del valor monetario esperado era utilizar el proceso A. Puede ocurrir (si la empresa de estudios de mercado dice que las perspectivas son «malas») que se tome una decisión diferente cuando la información muestral lleva a modificar estas probabilidades *a priori*. Por lo tanto, resulta que al fabricante le interesaría consultar a la empresa de estudios de mercado. Naturalmente, si la elección del proceso A hubiera resultado óptima, cualquiera que hubiera sido la predicción, la información muestral posiblemente no tendría ningún valor.

EJEMPLO 21.5. Reconsideración del problema del fabricante de medicamentos (valor monetario esperado)

En el ejemplo, 21.4, un fabricante de medicamentos tenía que decidir si vendía la patente de una fórmula que reducía el colesterol antes de someter el medicamento a una prueba (después, si conservaba la patente y se observaba que el medicamento era eficaz, también tenía que tomar otra decisión, que era comercializar el medicamento o vender la patente y los resultados de la prueba). En el caso de la decisión inicial, los dos estados de la naturaleza eran s_1: el medicamento es eficaz, y s_2: el medicamento es ineficaz. Las probabilidades *a priori* correspondientes, calculadas basándose en la experiencia anterior, son

$$P(s_1) = 0{,}6 \quad \text{y} \quad P(s_2) = 0{,}4$$

El fabricante de medicamentos tiene la opción de realizar con un coste moderado una prueba inicial antes de tomar la primera decisión. La prueba no es infalible. En el caso de medicamentos que después han resultado eficaces, el 60 por ciento de las veces el resultado de la prueba preliminar fue positivo y el resto fue negativo. En el caso de medicamentos ineficaces, el 30 por ciento de las veces el resultado de la prueba preliminar fue positivo y el resto fue negativo. Dados los resultados de la prueba preliminar, ¿qué debe hacer el fabricante? Suponga que sigue siendo posible vender la patente por 50.000 $ si el resultado de la prueba preliminar es negativo.

Solución

Obsérvese, en primer lugar, que si se conserva la patente y las pruebas exhaustivas demuestran que el medicamento es eficaz, entonces en ausencia de información muestral sobre la situación del mercado, la decisión óptima en esta fase es, al igual que en el ejemplo 21.4, comercializar el medicamento. La información suministrada por la prueba preliminar es irrelevante para tomar esa decisión. Sin embargo, podría influir en la decisión inicial de vender o no la patente. Por lo tanto, sólo se considera esta decisión.

Las probabilidades condicionadas de los resultados muestrales, dados los estados de la naturaleza, son

$$P(\text{positivo}|s_1) = 0{,}6 \qquad P(\text{negativo}|s_1) = 0{,}4$$
$$P(\text{positivo}|s_2) = 0{,}3 \qquad P(\text{negativo}|s_2) = 0{,}7$$

Si el resultado de la prueba preliminar es positivo, entonces la probabilidad *a posteriori* del estado s_1 (eficaz), dada esta información, es

$$P(s_1|\text{positivo}) = \frac{P(\text{positivo}|s_1)P(s_1)}{P(\text{positivo}|s_1)P(s_1) + P(\text{positivo}|s_2)P(s_2)} = \frac{(0{,}6)(0{,}6)}{(0{,}6)(0{,}6) + (0{,}3)(0{,}4)} = 0{,}75$$

Además, como las dos probabilidades *a posteriori* deben sumar 1, entonces $P(s_2|\text{positivo}) = 0{,}25$. La tabla de rendimientos adjunta es igual que la del ejemplo 21.4, con la adición de estas probabilidades *a posteriori*.

	Estado de la naturaleza	
Acción	**Medicamento eficaz ($P = 0{,}75$)***	**Medicamento inefizaz ($P = 0{,}25$)***
Conservar la patente	125.000	− 10.000
Vender la patente	50.000	50.000

* Probabilidades *a posteriori*.

El valor monetario esperado, si se vende la patente, es de 50.000 $, mientras que si se conserva, es

$$(0{,}75)(125.000) + (0{,}25)(-10.000) = 91.250\ \$$$

Por lo tanto, si el resultado de la prueba inicial es positivo, la patente debe conservarse, según este criterio.

Consideremos ahora el caso en el que el resultado de la prueba preliminar es negativo. La probabilidad *a posteriori* del estado s_1 es, según el teorema de Bayes,

$$P(s_1|\text{negativo}) = \frac{P(\text{negativo}|s_1)P(s_1)}{P(\text{negativo}|s_1)P(s_1) + P(\text{negativo}|s_2)P(s_2)} = \frac{(0{,}4)(0{,}6)}{(0{,}4)(0{,}6) + (0{,}7)(0{,}4)} = 0{,}4615$$

Por lo tanto, la probabilidad *a posteriori* del estado s_2 es

$$P(s_2|\text{negativo}) = 0{,}5385$$

Una vez más, si se vende la patente, el valor monetario esperado son los 50.000 $ que se recibirán. Si se conserva la patente, el valor monetario esperado de esta decisión es

$$(0{,}4615)(125.000) + (0{,}5385)(-10.000) = 52.302{,}50\ \$$$

Así pues, aunque el resultado de la prueba preliminar sea negativo, la decisión óptima, según el criterio del valor monetario esperado, es conservar la patente.

En este ejemplo, pues, cualquiera que sea la información muestral, la acción elegida es la misma. El fabricante debe conservar la patente cualquiera que sea el resultado de la prueba preliminar. Dado que la información muestral no puede influir en la decisión, no tiene sentido, desde luego, recogerla. De hecho, como la realización de la prueba preliminar tiene costes, sería subóptimo recogerla. Por lo tanto, según el criterio del valor monetario esperado, el fabricante de medicamentos debe conservar la patente y, si las pruebas demuestran que el medicamento es eficaz, debe comercializarlo. La prueba preliminar no debe realizarse.

El valor de la información muestral

Se ha demostrado cómo puede tenerse en cuenta la información muestral en el proceso de toma de decisiones. El valor potencial de esa información se halla, por su puesto, en que permite saber con mayor precisión cuáles son las probabilidades de que ocurra cada uno de los estados de la naturaleza relevantes y eso permite tener una base más sólida para tomar una decisión. En este apartado mostramos cómo puede asignarse un valor *monetario* a la información muestral. Esto es importante, ya que la obtención de información muestral normalmente tiene costes y la persona que debe tomar una decisión quiere saber si los beneficios esperados son mayores que este coste.

El ejemplo 21.5 muestra una situación en la que una misma acción era óptima, cualquiera que fuera el resultado muestral. En ese caso, la información muestral carece claramente de valor, ya que se habría elegido la misma acción sin ella. He aquí la regla general: si la información muestral no puede influir en la elección de la acción, tiene un valor 0.

En el resto de este apartado sólo nos referiremos, pues, a las circunstancias en las que el resultado muestral puede afectar a la elección de la acción. Un caso de ese tipo es nuestro ejemplo del fabricante de teléfonos móviles que está considerando la posibilidad de introducir un nuevo producto. Este fabricante tiene que elegir entre tres procesos de producción y se enfrenta a tres estados de la naturaleza, que representan diferentes niveles de demanda del producto. En el apartado 21.3 hemos mostrado que en ausencia de información muestral y utilizando solamente las probabilidades *a priori*, se selecciona el proceso A que tiene un valor monetario esperado de 147.000 $.

Ahora bien, en la práctica, una vez obtenida la información muestral, la persona que debe tomar una decisión normalmente no sabe qué estado de la naturaleza ocurrirá, pero tiene valoraciones probabilísticas más fundadas de estos estados. Sin embargo, antes de analizar el valor de la información muestral en este modelo general, es útil considerar el caso extremo en el que puede obtenerse **información perfecta**, es decir, el caso en el que la persona que tiene que tomar una decisión puede obtener información que le diga *con seguridad* qué estado ocurrirá. ¿Qué valor tiene esa información perfecta para la persona que debe tomar una decisión?

Valor esperado de la información perfecta, *VEIP*

Supongamos que una persona tiene que elegir entre K acciones posibles y se enfrenta a H estados de la naturaleza, s_1, s_2, ..., s_H. La **información perfecta** corresponde al caso en el que se sabe qué estado de la naturaleza ocurrirá. El valor esperado de la información perfecta se obtiene de la forma siguiente:

1. Se averigua qué acción se elegirá si sólo se utilizan las probabilidades *a priori* $P(s_1)$, $P(s_2)$, ..., $P(s_H)$.

2. Se halla para cada estado de la naturaleza posible, s_i, la diferencia, W_i, entre el rendimiento de la mejor elección de la acción, si se supiera que ocurrirá ese estado, y el rendimiento de la acción que se elegiría sólo si se utilizaran las probabilidades *a priori*. Éste es el **valor de la información perfecta**, cuando se sabe que ocurrirá s_i.
3. El **valor esperado de la información perfecta**, *VEIP*, es, pues,

$$VEIP = P(s_1)W_1 + P(s_2)W_2 + \cdots + P(s_H)W_H \qquad \textbf{(21.3)}$$

Volvamos al caso del fabricante de teléfonos móviles y calculemos el *VEIP*. En este ejemplo, la información perfecta corresponde al caso en el que se sabe cuál será el nivel de demanda de los tres posibles. En ausencia de información muestral y basándose únicamente en las probabilidades *a priori*, se elegirá el proceso A. Sin embargo, volviendo a la Tabla 21.7, si el nivel de demanda es bajo, la mejor elección será el proceso C. Como éste tiene un rendimiento que es 30.000 $ mayor que el del A, el valor de saber que la demanda será baja es de 30.000 $. Asimismo, si se sabe que la demanda será moderada, se elegirá de nuevo el proceso C. En este caso, el rendimiento de la mejor elección es 5.000 $ mayor que el del proceso A, que es, pues, el valor de saber que la demanda será moderada. Si se sabe que la demanda será alta, se elegirá el proceso A. Por lo tanto, esta información carece de valor, ya que se habría tomado la misma decisión sin ella. El valor de la información perfecta depende de la información. El valor esperado de la información perfecta se halla utilizando las probabilidades *a priori* de los distintos estados de la naturaleza.

En el caso del fabricante de teléfonos móviles, las probabilidades *a priori* son 0,1 en el caso en el que la demanda es baja, 0,5 en el caso en el que es moderada y 0,4 en el caso en el que es alta. Se deduce, pues, que para este fabricante el valor de la información perfecta es de 30.000 $ con una probabilidad de 0,1, 5.000 $ con una probabilidad de 0,5 y 0 $ con una probabilidad de 0,4. El valor esperado de la información perfecta es, pues,

$$VEIP = (0{,}1)(30.000) + (0{,}5)(5.000) + (0{,}4)(0) = 5.500\ \$$$

Esta cantidad monetaria representa, pues, el valor esperado para el fabricante de teléfonos móviles de saber cuál será el nivel de demanda.

Cuando los problemas son más complejos, existen programas informáticos para calcular el *VEIP*.

Aunque normalmente no se dispone de información perfecta, puede ser útil calcular su valor esperado. Dado que, naturalmente, ninguna información muestral puede ser mejor que perfecta, su valor esperado no puede ser mayor que el valor esperado de la información perfecta. Por lo tanto, el valor esperado de la información perfecta es un *límite superior* del valor esperado de cualquier información muestral. Por ejemplo, si el fabricante de teléfonos móviles recibe información con un coste de 6.000 $, no es necesario que trate de obtener más información sobre la calidad de esta información. No debería comprarla, por muy fiable que sea, según el criterio del valor monetario esperado, ya que su valor esperado no puede ser de más de 5.500 $.

Consideremos ahora el problema más general de calcular el valor de la información muestral que no es necesariamente perfecta. Consideremos de nuevo el problema de toma de decisiones del fabricante de teléfonos móviles, que tiene la opción de que una empresa de estudios de mercado valore las perspectivas del nuevo teléfono móvil. Estas perspectivas pueden considerarse «malas», «regulares» o «buenas». En el apartado 21.4 hemos mostrado que en los dos últimos casos se elige, aun así, el proceso A. Por lo tanto, si la empresa de estudios de mercado dice que las perspectivas son «regulares» o «buenas», la elección inicial de la acción no varía y no se ganará nada consultando a esta empresa.

Sin embargo, si dice que las perspectivas son «malas», la Tabla 21.10 muestra que la elección óptima es el proceso C. Esta elección óptima generaría un valor monetario esperado de 124.600 $, mientras que el proceso A, que, de no ser así, se habría utilizado, da un valor monetario esperado de 120.800 $. La diferencia entre estas cantidades, 3.800 $, representa la ganancia generada por la información muestral *si la empresa dice que las perspectivas son «malas»*. Las ganancias generadas por la información muestral son 0 $ en el caso en el que las perspectivas son «buenas» o «regulares» y 3.800 $ si son «malas».

Ahora necesitamos saber qué probabilidades hay de que se materialicen estas ganancias, por lo que en nuestro ejemplo debemos hallar la probabilidad de que la empresa de estudios de mercado diga que las perspectivas son «malas». En general, si A representa una parte de la información muestral y s_1, s_2, ..., s_H los H estados de la naturaleza posibles, entonces

$$P(A) = P(A|s_1)P(s_1) + P(A|s_2)P(s_2) + \cdots + P(A|s_H)P(s_H)$$

En el ejemplo del teléfono móvil, si s_1, s_2 y s_3 representan un nivel de demanda bajo, moderado y alto, respectivamente, entonces

$$\begin{array}{lll} P(s_1) = 0{,}1 & P(s_2) = 0{,}5 & P(s_3) = 0{,}4 \\ P(\text{malas}|s_1) = 0{,}6 & P(\text{malas}|s_2) = 0{,}3 & P(\text{malas}|s_3) = 0{,}1 \end{array}$$

Por lo tanto, la probabilidad de que la empresa diga que las perspectivas son «malas» es

$$\begin{aligned} P(\text{malas}) &= P(\text{malas}|s_1)P(s_1) + P(\text{malas}|s_2)P(s_2) + P(\text{malas}|s_3)P(s_3) \\ &= (0{,}6)(0{,}1) + (0{,}3)(0{,}5) + (0{,}1)(0{,}4) = 0{,}25 \end{aligned}$$

De la misma forma, utilizando las probabilidades condicionadas de la Tabla 21.8, las probabilidades de las otras dos valoraciones de la empresa son

$$P(\text{regulares}) = 0{,}30 \qquad P(\text{buenas}) = 0{,}45$$

Por lo tanto, el valor de la información muestral es de 3.800 $ con una probabilidad de 0,25, de 0 $ con una probabilidad de 0,30 y de 0 $ con una probabilidad de 0,45. Se deduce, pues, que el **valor esperado de la información muestral** es

$$VEIM = (0{,}25)(3.800) + (0{,}30)(0) + (0{,}45)(0) = 950\ \$$$

Esta cantidad monetaria representa, pues, el valor esperado de la información muestral para la persona que tiene que tomar una decisión. Según el criterio del valor monetario esperado, esta información muestral merecerá la pena si su coste es menor que su valor esperado. El **valor esperado neto de la información muestral** es la diferencia entre su valor esperado y su coste.

Supongamos que la empresa de estudios de mercado cobra 750 $ por su valoración. El valor esperado neto de esta valoración para el fabricante de teléfonos móviles es, pues, 950 $ − 750 $ = 200 $. Por lo tanto, el rendimiento esperado del fabricante será 200 $ mayor si se compra la información muestral que si no se compra. Esta cantidad representa el valor esperado de tener esa información, teniendo en cuenta su coste. En este caso, la estrategia óptima del fabricante es comprar el informe de la empresa de estudios de mercado y utilizar el proceso de producción A si la empresa dice que las perspectivas son «buenas» o «regulares» y el C si dice que son «malas». El *VME* de esta estrategia es de 147.200 $, es decir, los 147.000 $ que se obtendrían si no se dispusiera de información muestral más el valor esperado neto de la información muestral.

Valor esperado de la información muestral, *VEIM*

Supongamos que una persona tiene que elegir entre K acciones posibles ante H estados de la naturaleza, s_1, s_2, ..., s_H. Puede obtener información muestral. Supongamos que hay M resultados muestrales posibles, A_1, A_2, ..., A_M.

El valor esperado de la información muestral se obtiene de la forma siguiente.

1. Se averigua qué acción se elegiría si sólo se utilizaran las probabilidades *a priori*.
2. Se averiguan las probabilidades de obtener cada resultado muestral:

$$P(A_i) = P(A_i|s_1)P(s_1) + P(A_i|s_2)P(s_2) + \cdots + P(A_i|s_H)P(s_H)$$

3. Se halla para cada resultado muestral posible A_i, la diferencia, V_i, entre el valor monetario esperado de la acción óptima y el de la acción elegida si sólo se utilizan las probabilidades *a priori*. Éste es el **valor de la información muestral**, dado que se observó A_i.
4. El **valor esperado de la información muestral, *VEIM***, es, pues,

$$VEIM = P(A_1)V_1 + P(A_2)V_2 + \cdots + P(A_M)V_M \qquad \textbf{(21.4)}$$

El valor de la información muestral visto por medio de árboles de decisión

El valor esperado de la información muestral puede calcularse de otra forma (equivalente), que es desde el punto de vista aritmético algo más tediosa, pero cómoda para representar el problema por medio de una sucesión de decisiones construyendo un árbol de decisión. La primera decisión que hay que tomar es si conviene obtener la información muestral. A continuación, hay que averiguar cuál de las acciones alternativas debe seguirse.

Para ilustrarlo, consideremos de nuevo el problema del fabricante de teléfonos móviles. La Figura 21.5 muestra los árboles de decisión que se deducen de las tres valoraciones posibles del estudio de mercado. Estos árboles tienen la misma estructura general que la Figura 21.1, con una diferencia esencial: las probabilidades de los tres estados de la naturaleza son las *probabilidades a posteriori*, dada la información muestral específica. Estas probabilidades *a posteriori* se obtuvieron en el apartado 21.4. Ahora se ponderan los rendimientos por las probabilidades *a posteriori* y se obtiene el valor monetario esperado de cada acción, dado cada resultado muestral posible. Éstos son los valores monetarios esperados que muestra la Tabla 21.10. Por último, a la izquierda de cada parte de la Figura 21.5 se encuentra el valor monetario esperado más alto posible de cada resultado muestral.

Esta información se transfiere a la derecha de la Figura 21.6, en la que se analiza la decisión de comprar o no el estudio de mercado. Si no se compra esta información, la parte inferior de la Figura 21.6 muestra un valor monetario esperado de 147.000 $. Esta cantidad se obtiene utilizando las probabilidades *a priori* y procede de la Figura 21.1.

Pasamos ahora a examinar la parte superior de la Figura 21.6; el valor monetario esperado resultante depende del resultado muestral. Las probabilidades son 0,25 en el caso en el que las perspectivas son «malas», 0,30 en el que son «regulares» y 0,45 en el que son «buenas». Por lo tanto, dado que cabe esperar 124.600 $ con una probabilidad de 0,25, 138.000 $ con una probabilidad de 0,30 y 167.000 $ con una probabilidad de 0,45, el rendimiento esperado si se compra la información muestral es

$$(0{,}25)(124.600) + (0{,}30)(138.000) + (0{,}45)(167.556) = 147.950\ \$$$

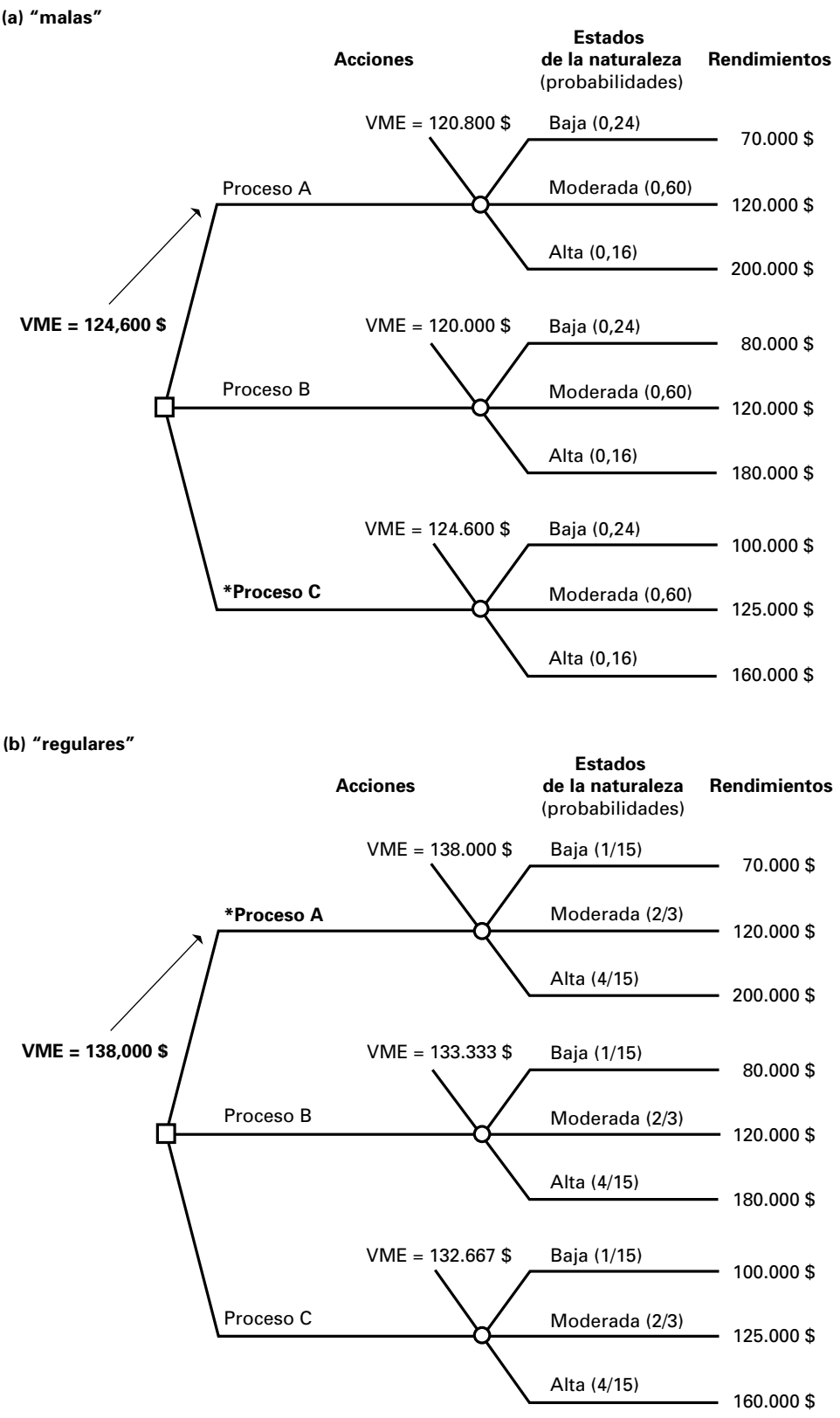


Figura 21.5. Árboles de decisión del fabricante de teléfonos móviles correspondientes a las valoraciones realizadas por la empresa de estudios de mercado de que las perspectivas son (a) «malas», (b) «regulares» y (c) «buenas» (* acción que tiene el máximo *VME*).

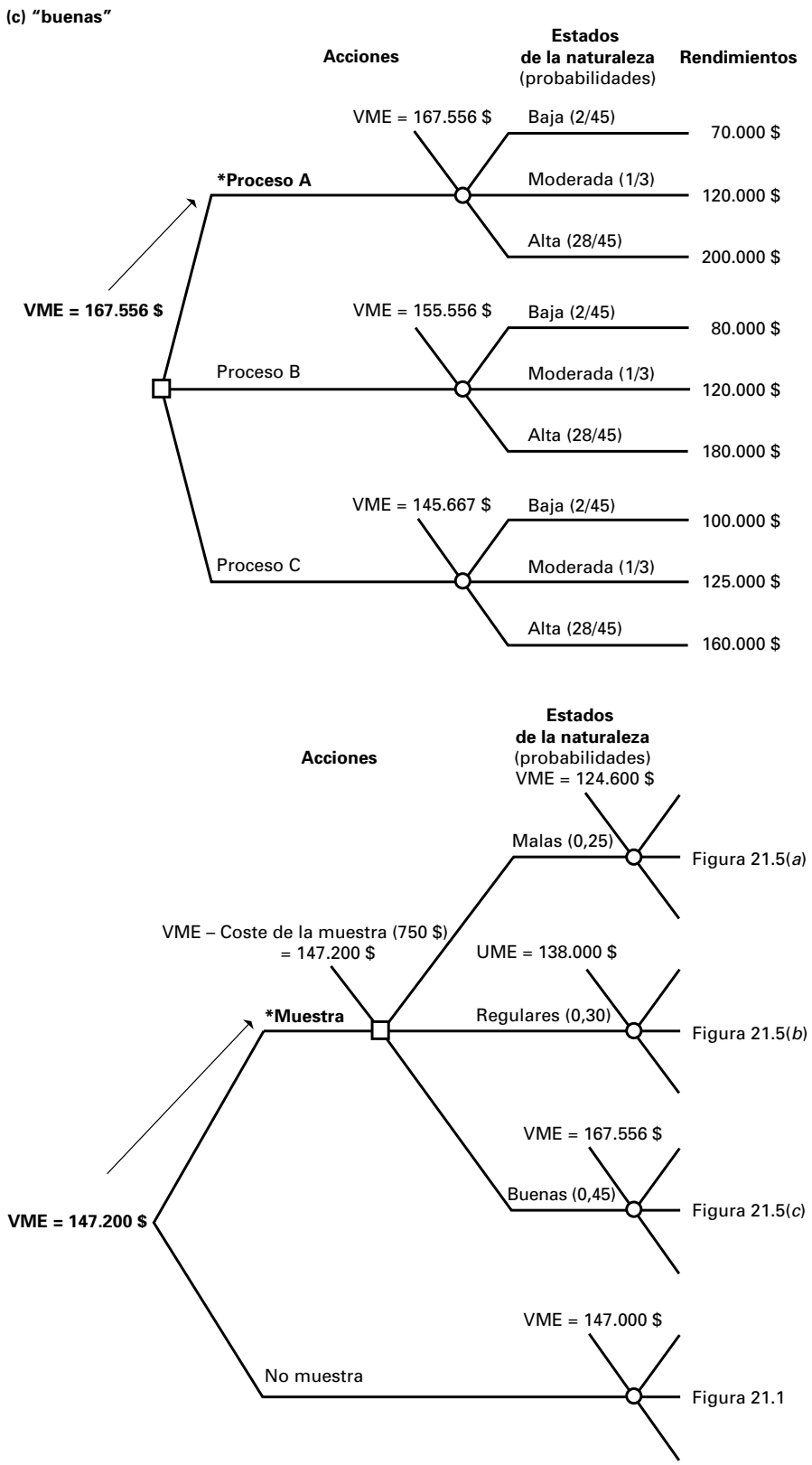


Figura 21.6. Decisión del fabricante de teléfonos móviles de comprar los servicios de la empresa de estudios de mercado (* acción con el máximo *VME*).

Sin embargo, es necesario restar de esta cantidad el coste de 750 $ de la información muestral, por lo que quedan 147.200 $. Dado que esta cantidad es superior al rendimiento esperado cuando no se obtiene información muestral, la mejor estrategia, según el criterio del valor monetario esperado, es comprar los servicios de la empresa de estudios de mercado. La decisión óptima tiene, como se indica a la izquierda de la Figura 21.6, un valor monetario esperado de 147.200 $.

EJERCICIOS

Ejercicios aplicados

21.25. Un fabricante debe decidir si lanza, con un coste de 100.000 $, una campaña publicitaria de un producto cuyas ventas han sido bastante bajas. Se estima que una campaña que tuviera mucho éxito aumentaría los beneficios en 400.000 $ (de los que habría que restar el coste de la campaña) y una campaña que tuviera un éxito moderado los aumentaría en 100.000 $, pero una campaña que no tuviera éxito no los aumentaría nada. Históricamente, el 40 por ciento de todas las campañas parecidas ha tenido mucho éxito, el 30 por ciento ha tenido un éxito moderado y el resto no ha tenido éxito. Este fabricante consulta a un experto en medios de comunicación y le pide que valore la eficacia que puede tener la campaña. El historial de este experto muestra que ha valorado favorablemente el 80 por ciento de las campañas que han tenido mucho éxito, el 40 por ciento de las que han tenido un éxito moderado y el 10 por ciento de las que no han tenido éxito.

a) Halle las probabilidades *a priori* de los tres estados de la naturaleza.
b) En ausencia de un informe del experto en medios de comunicación, ¿debe lanzarse esta campaña publicitaria, según el criterio del *VME*?
c) Halle las probabilidades *a posteriori* de los tres estados de la naturaleza, suponiendo que el experto valora favorablemente la campaña.
d) Dado un informe favorable del experto, ¿debe lanzarse la campaña publicitaria, según el criterio del *VME*?
e) Halle las probabilidades *a posteriori* de los tres estados de la naturaleza, suponiendo que el experto no valora favorablemente la campaña.
f) Si el informe del experto no es favorable, ¿debe lanzarse la campaña publicitaria según el criterio del *VME*?

21.26. Vuelva al ejercicio 21.2. El fabricante de desodorantes tiene cuatro procesos de producción posibles entre los que elegir, dependiendo de la opinión sobre el futuro nivel de demanda. Basándose en la experiencia anterior, las probabilidades *a priori* son de 0,3 en el caso de la demanda alta, de 0,4 en el de la demanda moderada y de 0,3 en el de la demanda baja. La tabla adjunta muestra las proporciones de valoraciones según las cuales las perspectivas son «malas», «regulares» y «buenas»; estas valoraciones han sido realizadas por una empresa de mercado sobre productos similares que han logrado estos niveles de demanda.

Acción	**Estado de la naturaleza**		
Valoración	**Demanda baja**	**Demanda moderada**	**Demanda alta**
Malas	0,5	0,3	0,1
Regulares	0,3	0,4	0,2
Buenas	0,2	0,3	0,7

a) Si no se consulta a la empresa de estudios de mercado, ¿qué acción debe elegirse, según el criterio del *VME*?
b) Halle las probabilidades *a posteriori* de los tres niveles de demanda, suponiendo que la empresa de estudios de mercado dice que las perspectivas son «malas».
c) ¿Qué acción debe elegirse, según el criterio del *VME*, si la empresa de estudios de mercado dice que las perspectivas son «malas»?
d) Halle las probabilidades *a posteriori* de los tres niveles de demanda, suponiendo que la empresa de estudios de mercado dice que las perspectivas son «regulares».
e) ¿Qué acción debe elegirse, según el criterio del *VME*, si la empresa de estudios de mercado dice que las perspectivas son «regulares»?
f) Halle las probabilidades *a posteriori* de los tres niveles de demanda, suponiendo que la empresa de estudios de mercado dice que las perspectivas son «buenas».

g) ¿Qué acción debe elegirse, según el criterio del *VME*, si la empresa de estudios de mercado dice que las perspectivas son «buenas»?

21.27. El empresario del ejercicio 21.9 tiene dos cursos de acción posibles. Su decisión se basa en su opinión sobre el éxito probable del nuevo centro comercial. Históricamente, el 40 por ciento de los centros de este tipo ha tenido mucho éxito, el 40 por ciento ha tenido un éxito moderado y el 20 por ciento no ha tenido éxito. Una empresa de consultoría hace valoraciones de las perspectivas de este tipo de centro comercial. La tabla adjunta muestra la proporción de valoraciones según las cuales las perspectivas son «buenas», «regulares» y «malas», dado el resultado obtenido realmente.

Acción	**Estado de la naturaleza (nivel de éxito)**		
Valoración	**Mucho éxito**	**Éxito moderado**	**Ningún éxito**
Buenas	0,6	0,3	0,2
Razonables	0,3	0,4	0,3
Malas	0,1	0,3	0,5

a) ¿Cuáles son las probabilidades *a priori* de los tres estados de la naturaleza?
b) Si el empresario no busca asesoramiento de la empresa de consultoría, ¿qué acción debe elegir, según el criterio del *VME*?
c) ¿Cuáles son las probabilidades *a posteriori* de los tres estados de la naturaleza, suponiendo que la empresa de consultoría dice que las perspectivas son «buenas»?
d) Según el criterio del *VME*, suponiendo que la empresa de consultoría dice que las perspectivas son «buenas», ¿qué curso de acción debe adoptar?
e) ¿Cuáles son las probabilidades *a posteriori* de los tres estados de la naturaleza, suponiendo que la empresa de consultoría dice que las perspectivas son «regulares»?
f) Según el criterio del *VME*, suponiendo que la empresa de consultoría dice que las perspectivas son «regulares», ¿qué curso de acción debe adoptar?
g) ¿Cuáles son las probabilidades *a posteriori* de los tres estados de la naturaleza, suponiendo que la empresa de consultoría dice que las perspectivas son «malas»?
h) Si se sigue el criterio del *VME*, ¿qué acción debe elegirse, suponiendo que la empresa de consultoría dice que las perspectivas son «malas»?

21.28. Considere el fabricante de medicamentos del ejemplo 21.5, que tiene que decidir si vende la patente de un medicamento que reduce el colesterol antes de probarlo. En el ejemplo hemos visto que, cualquiera que sea el resultado de una prueba preliminar de la eficacia del medicamento, la decisión óptima era conservar la patente. Después, este fabricante desarrollaba una prueba preliminar superior, que podía realizarse de nuevo con un coste moderado. En el caso de los medicamentos que después resultaban eficaces, esta nueva prueba daba un resultado positivo el 80 por ciento de las veces, mientras que obtenía un resultado positivo solamente un 10 por ciento de los medicamentos que resultaban ineficaces.

a) Halle las probabilidades *a posteriori* de los dos estados de la naturaleza, dado un resultado positivo de esta nueva prueba preliminar.
b) Según el criterio del *VME*, ¿debe venderse la patente si el resultado de la nueva prueba es positivo?
c) Halle las probabilidades *a posteriori* de los dos estados de la naturaleza, dado un resultado negativo de esta nueva prueba preliminar.
d) Según el criterio del *VME*, ¿debe venderse la patente si el resultado de la nueva prueba es negativo?

21.29. En el ejercicio 21.20, un proveedor de piezas para la industria automovilística tenía que decidir si comprobaba el proceso de producción en busca de un fallo antes de empezar una serie de producción. Los dos estados de la naturaleza eran

s_1: la reparación no es necesaria (el 10 por ciento de todas las piezas producidas no cumple las especificaciones)
s_2: la reparación es necesaria (el 30 por ciento de todas las piezas producidas no cumple las especificaciones)

Las probabilidades *a priori*, basadas en los datos históricos de este proceso de producción, son

$$P(s_1) = 0{,}8 \quad \text{y} \quad P(s_2) = 0{,}2$$

El fabricante, antes de iniciar una nueva serie de producción, puede producir una pieza y ver si cumple las especificaciones, basando la decisión de comprobar o no el proceso de producción en la información muestral resultante.

a) Si la pieza comprobada cumple las especificaciones, ¿cuáles son las probabilidades *a posteriori* de los estados de la naturaleza?

b) Si la pieza comprobada cumple las especificaciones, ¿debe comprobarse el proceso de producción según el criterio del *VME*?
c) Si la pieza comprobada no cumple las especificaciones, ¿cuáles son las probabilidades *a posteriori* de los estados de la naturaleza?
d) Si la pieza comprobada no cumple las especificaciones, ¿debe comprobarse el proceso de producción según el criterio del *VME*?

21.30. Continuando con el ejercicio 21.29, suponga ahora que antes de tomar la decisión de comprobar o no el proceso de producción, se fabrican *dos* piezas y se examinan.
a) Si no es necesaria realmente una reparación, ¿cuáles son las probabilidades de que ambas piezas, una de ellas o ninguna no cumpla las especificaciones?
b) Calcule las mismas probabilidades que en el apartado (a), suponiendo que es necesario realmente reparar el proceso de producción.
c) Calcule las probabilidades *a posteriori* de los estados de la naturaleza y averigüe la acción óptima según el criterio del valor monetario esperado, dada cada una de las siguientes circunstancias:
i. Ninguna de las dos piezas cumple las especificaciones.
ii. Sólo una incumple las especificaciones.
iii. Ninguna de las piezas incumple las especificaciones.

21.31. Una fábrica de bombillas envía grandes pedidos de bombillas a grandes usuarios industriales. Cuando el proceso de producción funciona correctamente (lo cual ocurre el 90 por ciento del tiempo), el 10 por ciento de todas las bombillas producidas tiene un defecto. Sin embargo, el proceso puede tener de vez en cuando algún fallo y, en ese caso, la tasa de bombillas defectuosas es del 20 por ciento. La fábrica considera que el coste, en fondo de comercio, de un envío con una tasa más alta de bombillas defectuosas a un usuario industrial es de 5.000 $. Si se sospecha que un envío contiene esta proporción más alta de bombillas defectuosas, puede venderlo a una cadena de tiendas de descuento, aunque eso supone una reducción de los beneficios de 600 $, independientemente de que el envío contenga o no una elevada proporción de bombillas defectuosas. Las decisiones de esta empresa se toman siguiendo el criterio del *VME*.
a) Se prepara un envío. En ausencia de más información, ¿debe enviarse a un usuario industrial o a una cadena de descuento?
b) Suponga que se comprueba una bombilla del envío. Averigüe adónde debe enviarse en cada una de las circunstancias siguientes:
i. Esta bombilla tiene defectos.
ii. Esta bombilla no tiene defectos.
c) Suponga que se comprueban dos bombillas del envío. Averigüe adónde debe enviarse en cada una de las circunstancias siguientes:
i. Ambas bombillas tienen defectos.
ii. Sólo una bombilla tiene defectos.
iii. Ninguna de las dos bombillas tiene defectos.
d) Indique sin hacer los cálculos cómo puede abordarse este problema de decisión si se comprueban 100 bombillas antes de enviarlas.

21.32. Vuelva al problema del inversor del ejercicio 21.1.
a) Explique qué se entiende por «información perfecta» en el contexto del problema de este inversor.
b) Las probabilidades *a priori* de que la bolsa de valores esté fuerte son de 0,2, las de que esté moderada son de 0,5 y las de que esté débil son de 0,3. ¿Cuál es el valor esperado de la información perfecta para este inversor?

21.33. En el caso del fabricante de desodorantes del ejercicio 21.2, las probabilidades *a priori* de que la demanda sea alta son de 0,3, las de que sea moderada son de 0,4 y las de que sea baja son de 0,3. Halle el *VEIP* de este fabricante.

21.34. En el caso del empresario del ejercicio 21.9, las probabilidades *a priori* de que el nuevo centro comercial tenga mucho éxito son de 0,4, las de que tenga un éxito moderado son de 0,4 y las de que no tenga éxito son de 0,2. ¿Cuál es el valor esperado de la información perfecta para el empresario?

21.35. El fabricante de piezas de automóvil del ejercicio 21.20 debe decidir si comprueba el proceso de producción antes de comenzar una nueva serie de producción. Dado que el proceso de producción funciona correctamente el 80 por ciento del tiempo, ¿cuál es el valor de la información perfecta para este fabricante?

21.36. Antes de demostrar cómo se halla el valor esperado de la información muestral, hemos analizado por separado la determinación del valor esperado de la información perfecta. En realidad, no era necesario, ya que la información perfecta no es más que un tipo especial de información muestral. Dado el método general para hallar el

valor esperado de la información muestral, muestre cómo especializarlo al caso de la información perfecta.

21.37. Vuelva al ejercicio 21.25. El fabricante está considerando la posibilidad de hacer una campaña publicitaria y busca primero el asesoramiento de un experto en medios de comunicación.

a) ¿Qué valor esperado tiene para el fabricante el asesoramiento del experto en medios de comunicación?

b) El experto cobra 5.000 $. ¿Cuál es el valor esperado neto del asesoramiento del experto?

c) Este fabricante se enfrenta a un problema de decisión en dos etapas. Primero, debe decidir si compra asesoramiento al experto. A continuación, debe decidir si lanza la campaña publicitaria. Construya el árbol de decisión completo e indique qué debe hacer el fabricante.

21.38. Vuelva al ejercicio 21.26. Halle los mayores honorarios que debe pagar el fabricante de desodorantes a la empresa de estudios de mercado, según el criterio del valor monetario esperado.

21.39. Vuelva al ejercicio 21.27. Halle el valor esperado que tiene para el empresario una valoración de las perspectivas del centro comercial realizada por la empresa de consultoría.

21.40. Vuelva al ejercicio 21.28. Antes de decidir si vende la patente de la nueva fórmula para reducir el colesterol, el fabricante de medicamentos realiza una nueva prueba preliminar. Halle el valor esperado que tiene para el fabricante el resultado de la prueba.

21.41. Vuelva al ejercicio 21.29. El proveedor de piezas de automóvil puede producir y examinar una pieza antes de decidir si comprueba el proceso de producción. ¿Cuál es el *VEIM*?

21.42. Considere la fábrica de bombillas del ejercicio 21.31. La empresa puede comprobar una bombilla o más antes de decidir si envía un pedido a un usuario industrial o a una cadena de descuento.

a) ¿Qué valor esperado tiene para la empresa la comprobación de una bombilla?

b) ¿Qué valor esperado tiene para la empresa la comprobación de dos bombillas?

c) ¿Cuál es la diferencia entre los valores esperados de comprobar dos bombillas y una bombilla?

d) Si la primera bombilla comprobada es defectuosa, ¿cuál es el valor esperado de comprobar la segunda?

e) Si la primera bombilla comprobada no es defectuosa, ¿cuál es el valor esperado de comprobar la segunda?

21.5. Introducción del riesgo: análisis de la utilidad

El criterio del valor monetario esperado para tomar decisiones tiene muchas aplicaciones prácticas. Es decir, en muchos casos, una persona o una empresa creen que la acción que ofrece el mayor valor monetario esperado es el curso de acción preferido. Sin embargo, no siempre es así, como lo demuestran los ejemplos siguientes.

1. Muchas personas compran un seguro de vida a plazo con el que, con un gasto relativamente pequeño, los beneficiarios de la persona asegurada son indemnizados generosamente en caso de muerte durante la vigencia de la póliza. Actualmente, las compañías de seguros pueden calcular la probabilidad que tiene una persona de cualquier edad de morir durante un periodo de tiempo específico. Por lo tanto, fijan sus tarifas de manera que el precio de la póliza sea mayor que la cantidad de dinero que esperan pagar en caso de fallecimiento. La diferencia cubre los costes de la compañía de seguros y genera, en promedio, un margen de beneficio. Se deduce, pues, que para la persona asegurada el rendimiento esperado de la póliza del seguro de vida es menor que su coste. Por lo tanto, si todo el mundo tomara decisiones siguiendo el criterio del valor monetario esperado, el seguro de vida a plazo no se compraría. No obstante, muchas personas lo compran, lo que demuestra que están dispuestas a sacrificar algunos rendimientos esperados a cambio de tener la seguridad de que sus herederos tendrán un colchón financiero en caso de fallecimiento.

2. Supongamos que un inversor está considerando la posibilidad de comprar acciones de un grupo o más de empresas cuyas perspectivas considera brillantes. En principio, es posible postular los distintos estados de la naturaleza que influirán en los rendimientos de la inversión en cada una de estas empresas. De esta forma, podría averiguarse cuál es el valor monetario esperado de una inversión de una cantidad fija en cada empresa. Según el criterio del valor monetario esperado, el inversor debería invertir todo el capital de que dispone en la empresa cuyo valor monetario esperado es mayor. En realidad, muchos inversores en la bolsa de valores no siguen esa estrategia sino que reparten su dinero en efectivo en una cartera de acciones. El abandono de la opción de «poner todos los huevos en la misma cesta», aunque genera un rendimiento esperado menor, protege de la posibilidad de perder mucho dinero si resulta que las acciones de la empresa que tiene el mayor rendimiento esperado marchan mal. Al optar por una cartera de acciones, el inversor muestra su disposición a sacrificar algún valor monetario esperado a cambio de que las probabilidades de experimentar grandes pérdidas financieras sean menores.

En cada uno de estos ejemplos, la persona que toma las decisiones ha mostrado una preferencia por un criterio de elección distinto del valor monetario esperado y en cada circunstancia esta preferencia parece muy razonable. Los dos ejemplos tienen un denominador común, además de los rendimientos esperados. En ambos casos, la persona que toma decisiones quiere tener en cuenta el *riesgo*. El comprador de un seguro de vida a plazo está dispuesto a aceptar un rendimiento esperado negativo a cambio de la posibilidad de tener un gran rendimiento positivo en caso de fallecimiento. De esa forma, expresa una **preferencia por el riesgo** (naturalmente, se protege del riesgo de que su familia salga mal parada económicamente por su fallecimiento). En cambio, el inversor que, al repartir su inversión en una cartera de acciones, acepta un rendimiento esperado menor para reducir las posibilidades de experimentar una gran pérdida muestra **aversión al riesgo**.

El criterio del valor monetario esperado no es adecuado ni para las personas que prefieren el riesgo ni para las que son reacias a él. Afortunadamente, no es demasiado difícil modificarlo para abordar las situaciones en las que el riesgo es un factor relevante. La idea es esencialmente sustituir los rendimientos monetarios por cantidades que reflejen no sólo las cantidades monetarias que van a recibirse sino también la actitud de la persona hacia el riesgo.

El concepto de utilidad

En el ejemplo 21.3 hemos analizado el problema de un inversor que elige entre una inversión a un tipo de interés garantizado y una cartera de acciones. La primera generaría un rendimiento de 1.200 $, mientras que la segunda generaría un rendimiento de 2.500 $ y 500 $ si la bolsa de valores estuviera boyante o se mantuviera estable, pero una pérdida de 1.000 $ si estuviera deprimida. Este inversor creía que las probabilidades respectivas de estos tres estados de la naturaleza eran 0,6, 0,2 y 0,2. En ese caso, el valor monetario esperado de elegir la cartera de acciones era 1.400 $, que era 200 $ mayor que el de la inversión a un tipo de interés fijo. En esta coyuntura, necesitamos averiguar si este rendimiento esperado mayor compensa el riesgo de perder 1.000 $, como ocurriría si el mercado estuviera deprimido. Un inversor muy rico, que pudiera sufrir con comodidad esa pérdida, decidiría casi con toda seguridad que compensa el riesgo. Sin embargo, la postura de una persona relativamente pobre, para la cual una pérdida de 1.000 $ sería desastrosa, puede ser muy distinta. En el caso de ese inversor, los rendimientos deben ser sustituidos por

algunas otras cantidades que reflejen mejor la catástrofe que supondría una pérdida de 1.000 $. Estas cantidades deben medir el valor o *utilidad* que tiene para el inversor una pérdida de 1.000 $ en comparación, por ejemplo, con una ganancia de 500 $ o de 2.500 $.

Los estudios pioneros de investigadores como Von Neumann y Morgenstern (véase la referencia bibliográfica 6) mejoraron el concepto de utilidad, que aún hoy desempeña un papel fundamental en economía. El análisis de la utilidad constituye la base para solucionar problemas de decisión en presencia de preferencia o de aversión al riesgo. Para emplearlo, sólo se necesitan unos supuestos bastante suaves y normalmente bastante razonables. Supongamos que una persona se enfrenta a varios rendimientos posibles, que pueden ser o no monetarios. Se supone que puede ordenar (posiblemente con empates) la utilidad o satisfacción que le reportaría cada uno. Así, si prefiere el rendimiento A al B y el B al C, debe preferir el A al C.

También se supone que si prefiere el rendimiento A al B y el B al C, existe un juego de azar que ofrece A con una probabilidad P y C con una probabilidad $(1 - P)$, tal que al individuo le dará igual aceptar el juego que recibir B con seguridad. Dados estos y otros supuestos generalmente inocuos en cuyos detalles no es necesario que nos detengamos, es posible mostrar que la persona racional elige la acción cuya utilidad esperada es mayor. Por consiguiente, el problema de decisión se analiza exactamente igual que en los apartados anteriores, *pero con utilidades en lugar de rendimientos*. Es decir, se construye una tabla de utilidad en lugar de una tabla de rendimientos y, a continuación, se emplean las probabilidades de los estados de la naturaleza para comparar las utilidades esperadas.

Veamos ahora cómo se averiguan las utilidades correspondientes a los distintos rendimientos. Los rendimientos posibles en orden ascendente en el caso de nuestro inversor son -1.000 $, 500 $, 1.200 $ y 2.500 $. El primer paso es obtener una función de utilidad.

Cómo se obtiene una función de utilidad

Supongamos que una persona puede recibir varios rendimientos alternativos. La transformación de los rendimientos en **utilidades** se realiza de la forma siguiente:

1. Las unidades en las que se mide la utilidad son arbitrarias. Por lo tanto, puede fijarse una escala como convenga. Sea L el rendimiento más bajo de todos y H el más alto. Asignamos la utilidad 0 al rendimiento L y la utilidad 100 al rendimiento H.
2. Sea I cualquier rendimiento comprendido entre L y H. Hallamos la probabilidad P tal que la persona es indiferente entre las siguientes alternativas:
   - **a)** Recibir el rendimiento I con seguridad.
   - **b)** Recibir el rendimiento H con la probabilidad P y el rendimiento L con la probabilidad $(1 - P)$.
3. La utilidad que tiene para el individuo el rendimiento I es, pues, $100P$. La curva que relaciona la utilidad y el rendimiento se llama **función de utilidad**.

El primer paso no tiene ningún misterio y nos permite tener una cómoda medida para medir la utilidad. La elección de los números 0 y 100 para representar la utilidad del menor rendimiento y la del mayor es totalmente arbitraria. Podría muy bien utilizarse cualquier otro par de números, mientras la utilidad del rendimiento mayor sea mayor que la del menor, sin afectar al resto del análisis.

A efectos prácticos, el segundo paso es el más difícil, debido en parte a que presupone que el individuo puede manipular las probabilidades de una manera coherente. En la práctica, la probabilidad debe averiguarse mediante el método de prueba y error, haciendo preguntas como «¿preferiría recibir I con seguridad o participar en un juego de azar en el que

podría recibir *H* con una probabilidad de 0,9 y *L* con una probabilidad de 0,1?». O quizá «¿preferiría recibir *I* con seguridad o participar en un juego de azar en el que podría obtener *H* con una probabilidad de 0,8 y *L* con una probabilidad de 0,2?». Este proceso continúa hasta que se alcanza el punto de indiferencia.

La lógica del último paso es bastante sencilla. Dado que *H* tiene una utilidad de 100 y *L* tiene una utilidad de 0, la *utilidad esperada* si se obtiene *H* con una probabilidad de *P* y *L* con una probabilidad de $(1 - P)$ es

$$100P + 0(1 - P) = 100P$$

Dado que el individuo es indiferente entre este juego y recibir *I* con seguridad, la utilidad del rendimiento *I* es $100P$.

Volvamos ahora a nuestro inversor. En primer lugar, asignamos una utilidad de 0 al menor rendimiento, -1.000 \$, y una utilidad de 100 al mayor, 2.500 \$.

Queda por averiguar las utilidades de los rendimientos intermedios, 500 \$ y 1.200 \$. Se averiguan planteando al individuo una serie de preguntas, como «preferiría recibir 500 \$ con seguridad o participar en un juego en el que podría ganar 2.500 \$ con una probabilidad *P* y perder 1.000 \$ con una probabilidad de $(1 - P)$?». Se prueba con diferentes valores de la probabilidad *P* hasta que se halla el valor con el que el individuo es indiferente entre las dos alternativas. Este proceso se repite en el caso del rendimiento de 1.200 \$.

Supongamos que el inversor es indiferente entre un rendimiento de 500 \$ y el juego de azar que tiene una $P = 0{,}6$ y entre un rendimiento de 1.200 \$ y el juego que tiene una $P = 0{,}8$. Las utilidades de los rendimientos intermedios son, pues,

Rendimiento 500 \$: Utilidad = (100)(0,6) = 60
Rendimiento 1.200 \$: Utilidad = (100)(0,8) = 80

En la Figura 21.7 representamos por medio de puntos las cuatro utilidades de este inversor en relación con los rendimientos correspondientes.

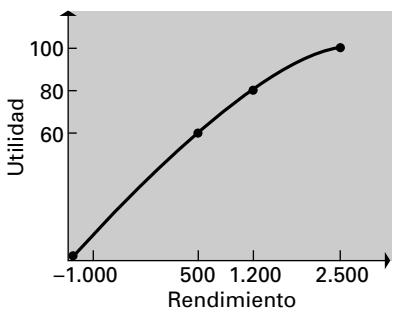


Figura 21.7. Función de utilidad de un inversor.

Trazamos una curva por estos puntos para indicar la forma general de la función de utilidad de este inversor. La forma de esta curva es interesante, ya que caracteriza la actitud del inversor hacia el riesgo. Como no podía ser de otra forma, la utilidad aumenta a medida que aumenta el rendimiento. Obsérvese, sin embargo, que la *tasa de aumento* de la utilidad es mayor en los rendimientos más bajos y disminuye a medida que aumenta el rendimiento. Eso significa un desagrado por los rendimientos más bajos que es más que acorde con su cantidad monetaria, lo que indica una *aversión* al riesgo. Esta aversión puede verse en la actitud del inversor hacia los juegos de azar que le proponen. Por ejemplo, el inversor es indiferente entre un rendimiento seguro de 500 \$ y un juego en el que puede

ganar 2.500 $ con una probabilidad de 0,6 y perder 1.000 $ con una probabilidad de 0,4. El valor monetario esperado de este juego es

$$(0,6)(2.500) + (0,4)(-1.000) = 1.100\ \$$$

que es considerablemente mayor que el rendimiento seguro preferido de 500 $. La cuantía de esta diferencia es una medida del grado de aversión al riesgo.

La forma de la Figura 21.7 es característica de la aversión al riesgo.

Según Friedman y Savage, «una importante clase de reacciones de los individuos al riesgo puede racionalizarse mediante una extensión bastante simple del análisis ortodoxo de la utilidad» (véase la referencia bibliográfica 2). Desarrollaron gráficos de funciones de utilidad similares a los tres tipos de funciones de utilidad que se muestran en la Figura 21.8.

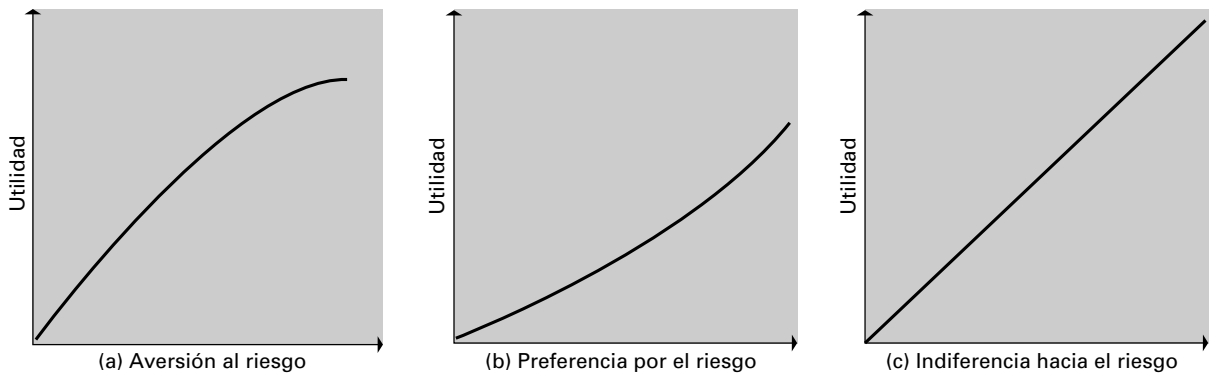


Figura 21.8. Funciones de utilidad: (a) aversión al riesgo; (b) preferencia por el riesgo; (c) indiferencia hacia el riesgo.

La función de la parte (a) de la figura, en la que la utilidad aumenta a una tasa *decreciente* a medida que aumenta el rendimiento, tiene la misma forma que la Figura 21.7, reflejando una vez más una *aversión* al riesgo. En la parte (b) de la figura, la utilidad aumenta a una tasa *creciente* a medida que los rendimientos son mayores. Eso implica un gusto por los rendimientos más altos que es más que acorde con las cantidades monetarias en cuestión, lo que muestra una *preferencia* por el riesgo. Por último, la parte (c) de la Figura 21.8 muestra el caso intermedio en el que la utilidad aumenta a una tasa *constante* en el caso de todos los rendimientos. En este caso, los valores monetarios de los rendimientos constituyen una verdadera medida de su utilidad para el individuo, que demuestra así **indiferencia hacia el riesgo**.

Las tres curvas de la Figura 21.8 caracterizan la aversión al riesgo, la preferencia por el riesgo y la indiferencia hacia el riesgo. Sin embargo, un individuo no tiene por qué mostrar solamente una de estas actitudes ante toda la variedad de rendimientos posibles.

La Figura 21.9 ilustra una situación más compleja. En esta figura, en los rendimientos comprendidos entre M_1 y M_2, la función de utilidad tiene la forma de la Figura 21.8(a), lo que indica una aversión al riesgo entre estos rendimientos. Sin embargo, en el caso de los rendimientos comprendidos entre M_2 y M_3, esta función de utilidad tiene la forma de la Figura 21.8(b). Por lo tanto, entre estos rendimientos el individuo muestra una preferencia por el riesgo. Por último, en el caso de los rendimientos más altos, entre M_3 y M_4, la posición se invierte de nuevo y el individuo es renuente al riesgo. Esa función de utilidad puede surgir en los problemas prácticos. Por ejemplo, un inversor puede muy bien ser reacio a

experimentar grandes pérdidas y estar dispuesto al mismo tiempo a aceptar algún riesgo para obtener un rendimiento positivo bastante alto en lugar de un rendimiento moderado. Sin embargo, si puede lograrse un rendimiento satisfactoriamente alto con un riesgo moderado, puede ser reacio a arriesgarse mucho más ante la posibilidad de obtener un rendimiento aún mayor.

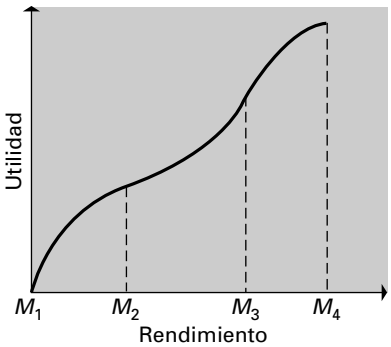


Figura 21.9. Función de utilidad que muestra una aversión al riesgo entre los rendimientos M_1 y M_2, y los rendimientos M_3 y M_4 y una preferencia por el riesgo entre los rendimientos M_2 y M_3.

Criterio de la utilidad esperada para tomar decisiones

Una vez halladas las utilidades, no queda más que resolver el problema de decisión averiguando el curso de acción que tiene la utilidad esperada más alta. Las utilidades esperadas se obtienen como siempre, empleando las probabilidades de los estados de la naturaleza, como se muestra en la ecuación 21.5.

El criterio de la utilidad esperada

Supongamos que una persona tiene K acciones posibles, a_1, a_2, ..., a_K, y se enfrenta a H estados de la naturaleza. Sea U_{ij} la utilidad correspondiente a la i-ésima acción y el j-ésimo estado y P_j la probabilidad de que ocurra el j-ésimo estado de la naturaleza. En ese caso, la **utilidad esperada**, $UE(a_i)$, de la acción a_i es

$$UE(a_i) = P_1U_{i1} + P_2U_{i2} + \cdots + P_HU_{iH} = \sum_{j=1}^{H} P_jU_{ij} \qquad \textbf{(21.5)}$$

Dada una elección entre acciones alternativas, el **criterio de la utilidad esperada** dicta la elección de la acción cuya utilidad esperada es mayor. Partiendo de unos supuestos generalmente razonables, puede demostrarse que una persona racional debe adoptar este criterio.

Si el individuo es indiferente al riesgo, el criterio de la utilidad esperada y el criterio del valor monetario esperado son equivalentes.

La Tabla 21.11 muestra las utilidades y las probabilidades de los estados de la naturaleza de nuestro inversor.

Si se elige la inversión a un tipo de interés fijo, está garantizada una utilidad de 80, cualquiera que sea el estado de la naturaleza. En el caso de la cartera de acciones, la utilidad esperada es

$$(0{,}6)(100) + (0{,}2)(60) + (0{,}2)(0) = 0{,}72$$

Dado que esta cantidad es menor que 80, este inversor debe invertir a un tipo de interés fijo, según el criterio de la utilidad esperada.

Tabla 21.11. Utilidades y probabilidades de los estados de la naturaleza de un inversor.

Acción	Estado del mercado		
Inversión	**Estado boyante ($P = 0{,}60$)**	**Estado estable ($P = 0{,}20$)**	**Estado deprimido ($P = 0{,}20$)**
Tipo de interés fijo	80	80	80
Cartera de acciones	100	60	0

En el ejemplo 21.3 se seleccionó la inversión en la cartera de acciones según el criterio del valor monetario esperado. Sin embargo, la introducción en el análisis de otro factor —el grado de aversión de este inversor al riesgo— lleva a la conclusión de que la opción del tipo de interés fijo es la mejor. Este ejemplo sirve para ilustrar que a veces, cuando el riesgo es un factor importante, el criterio del valor monetario esperado no es adecuado para resolver problemas de decisión.

El criterio de la utilidad esperada es el más aplicable e intelectualmente defendible de todos los introducidos para abordar problemas de decisión.

Su principal inconveniente radica en la dificultad para extraer información sobre qué juegos de azar se consideran igual de atractivos que los diferentes rendimientos asegurados. Este tipo de información es esencial para averiguar las utilidades. En una amplia variedad de problemas en los que puede suponerse con seguridad que el individuo es indiferente al riesgo, el criterio del valor monetario esperado sigue siendo aplicable. Ése sería normalmente el caso, por ejemplo, de una pequeña proporción del ingreso total de la empresa. Sin embargo, si (como puede ocurrir en el desarrollo de una nueva compañía aérea, por ejemplo) las posibles pérdidas de un proyecto pueden poner en peligro una empresa, las utilidades deben reflejar correctamente la aversión al riesgo. Una empresa puede intentar repartir este riesgo creando proyectos de colaboración con otras empresas del sector o con posibles clientes.

EJERCICIOS

Ejercicios aplicados

21.43. Una persona se enfrenta a un problema en el que los rendimientos posibles (en dólares) son

1.000 3.000 6.000 9.000 10.000 12.000

Se asigna la utilidad 0 al rendimiento de 1.000 \$ y la utilidad 100 al rendimiento de 12.000 \$. Esta persona es indiferente al riesgo en el caso de los rendimientos comprendidos en ese intervalo.

a) Halle las utilidades de los cuatro rendimientos intermedios.

b) Halle en el caso del rendimiento intermedio la probabilidad P de que el individuo sea indiferente entre recibir I con seguridad y una apuesta en la que se reciben 12.000 \$ con una probabilidad P y 1.000 \$ con una probabilidad $(1 - P)$.

21.44. El empresario del ejercicio 21.9 tiene seis rendimientos posibles (en dólares):

-10.000 30.000 60.000 70.000 90.000 130.000

Asigne una utilidad de 0 a una pérdida de 10.000 \$ y una utilidad de 100 a un beneficio de 130.000 \$. La tabla adjunta muestra para el caso de cada rendimiento intermedio la probabilidad P de que el empresario sea indiferente entre recibir I con seguridad y un juego de azar en el que recibiría 130.000 \$ con una probabilidad P y perdería 10.000 \$ con una probabilidad $(1 - P)$.

Rendimiento	30.000	60.000	70.000	90.000
P	0,35	0,60	0,70	0,85

a) ¿Cuáles son las utilidades de los rendimientos intermedios?

b) Suponga que las probabilidades de que el nuevo centro comercial tenga mucho éxito, tenga un éxito moderado y no tenga éxito son 0,4, 0,4 y 0,2, respectivamente. ¿Qué acción debería elegirse si se quiere maximizar la utilidad esperada?

21.45. El empresario del ejercicio 21.44 no sabe qué valor P asignar a la indiferencia entre recibir 30.000 $ con seguridad y un juego de azar en el que recibiría 130.000 $ con una probabilidad P y perdería 10.000 $ con una probabilidad $(1 - P)$. Suponiendo que el resto de las especificaciones del problema son las del ejercicio 21.44, ¿en qué intervalo de valores de esta probabilidad generará el criterio de la utilidad esperada la misma elección de la acción?

21.46. Considere el contratista del ejercicio 21.21. En realidad, este contratista es indiferente entre presentar y no presentar una oferta. ¿Qué implica eso sobre la función de utilidad del contratista?

RESUMEN

Este capítulo pretende ser una introducción al análisis de las decisiones. Todos debemos vivir y trabajar en un entorno cuyo futuro es incierto. La toma de decisiones de las empresas no es una excepción. Hemos analizado el marco de un problema de decisión, hemos estudiado varios criterios para seleccionar una acción óptima, hemos analizado el valor de la información muestral y hemos examinado las situaciones en las que la persona que tiene que tomar una decisión puede estar más interesada en tener en cuenta el riesgo que en maximizar los valores monetarios esperados. En la segunda situación, hemos examinado una función de utilidad. En este capítulo, hemos analizado cuatro criterios para tomar decisiones: maximin, pérdida de oportunidades minimax, valor monetario esperado y utilidad esperada. Hemos utilizado el TreePlan para construir árboles de decisión.

TÉRMINOS CLAVE

acción, 856
acción admisible, 857
acción inadmisible, 857
análisis de sensibilidad, 872
árboles de decisión, 866
aversión al riesgo, 891
criterio de la pérdida de oportunidades minimax, 862
criterio de la utilidad esperada, 895
criterio del valor monetario esperado, 865
criterio maximin, 860
estados de la naturaleza, 857
función de utilidad, 892
información perfecta, 881
indiferencia al riesgo, 894
nodos de decisión, 867
nodos de sucesos, 867
nodos terminales, 867
preferencia por el riesgo, 891
probabilidad *a priori*, 876
tabla de pérdida de oportunidades, 862
tabla de pérdidas, 862
tabla de rendimientos, 857
teorema de Bayes, 876
TreePlan, 868
valor de la información muestral, 881
valor de la información perfecta, 881
valor esperado de la información perfecta, 882
valor esperado neto de la información muestral, 883
valor monetario esperado, 865
VEIM, 884
VEIP, 881
VME, 865

EJERCICIOS Y APLICACIONES DEL CAPÍTULO

21.47. Un consultor está considerando la posibilidad de presentar ofertas detalladas para la adjudicación de dos contratos. La preparación de la oferta para el primero cuesta 100 $, mientras que la preparación de la oferta para el segundo cuesta 150 $. Si se acepta la oferta para el primer contrato y se realiza el trabajo, el beneficio es de 800 $. Si se acepta la oferta para el segundo contrato y se realiza el trabajo, el beneficio es de 1.200 $. Los costes de la preparación de la oferta deben restarse de estos beneficios. El consultor puede presentar, si lo desea, ofertas para los dos contratos. Sin embargo, no tiene los recursos necesarios para realizar los dos trabajos simultáneamente. Si presenta una oferta, ésta es aceptada y el consultor no puede rea-

lizar el trabajo, lo contabiliza como un coste de 200 $ de pérdida de fondo de comercio. En el proceso de toma de decisiones, hay cuatro estados de la naturaleza posibles:

s_1: se rechazan ambas ofertas
s_2: se acepta la oferta para el primer contrato y se rechaza la oferta para el segundo
s_3: se acepta la oferta para el segundo contrato y se rechaza la oferta para el primero
s_4: se aceptan ambas ofertas

a) El consultor tiene cuatro cursos de acción posibles. ¿Cuáles son?
b) Elabore la tabla de rendimientos del problema de decisión de este consultor.
c) ¿Qué acción se elige según el criterio maximin?
d) ¿Qué acción se elige según el criterio de la pérdida de oportunidades minimax?

21.48. Vuelva al ejercicio 21.47. El consultor cree que la probabilidad de que se acepte la oferta para el primer contrato es de 0,7 y la probabilidad de que se acepte la oferta para el segundo es de 0,4. También cree que la aceptación de una oferta es independiente de la aceptación de la otra.

a) ¿Cuáles son las probabilidades de los cuatro estados de la naturaleza?
b) Según el criterio del valor monetario esperado, ¿qué acción debe elegir el consultor y cuál es el valor monetario esperado de esta acción?
c) Construya el árbol de decisión del problema del consultor.
d) ¿Cuál es el valor esperado de la información perfecta para este consultor?
e) El consultor tiene la posibilidad de conseguir «información privilegiada» sobre las perspectivas de la oferta para el primer contrato. Esta información es totalmente fiable en el sentido de que le permitiría saber con seguridad qué oferta se aceptaría. Sin embargo, no dispone de más información sobre las perspectivas de la oferta para el segundo contrato. ¿Cuál es el valor esperado de esta «información privilegiada»?

21.49. Vuelva a los ejercicios 21.47 y 21.48. Este consultor se enfrenta a nueve rendimientos posibles (en dólares):

$$-250 \quad -150 \quad 0 \quad 550 \quad 700 \quad 750 \quad 950 \quad 1.950$$

Se asigna una utilidad de 0 a una pérdida de 250 $ y una utilidad de 100 a un beneficio de 1.050 $. La tabla adjunta muestra las probabilidades, P, de cada rendimiento intermedio, I, por las que el consultor es indiferente entre un rendimiento de I con seguridad y un juego de azar en el que ganaría 1.050 $ con la probabilidad P y perdería 250 $ con la probabilidad $(1 - P)$. Según el criterio de la utilidad esperada, ¿qué acción debe elegir el consultor y cuál es la utilidad esperada de esa acción?

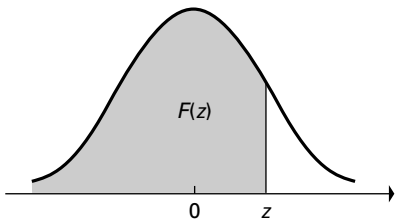
Rendimiento	−150	−100	0	550	700	750	950
P	0,05	0,10	0,20	0,65	0,70	0,75	0,85

Bibliografía

1. Eppen, G. D., F. J. Gould *et al*., *Introductory Management Science*: *Decision Modeling with Spreadsheets*, Upper Saddle River, NJ, Prentice Hall, 1998, 5.ª ed.
2. Friedman, Milton y L. J. Savage, «The Utility Analysis of Choices Involving Risk», *Journal of Political Economy*, 56, 1948, págs. 279-304.
3. Middleton, Michael, profesor, University of San Francisco, www.usaf.edu/∼middleton.
4. Render, Barry y Ralph M. Stair, Jr., *Quantitative Analysis for Management*, Upper Saddle River, NJ, Prentice Hall, 2000, 7.ª ed.
5. TreePlan Documentation, disponible en www.treeplan.com.
6. Von Neumann, John y Oskar Morgenstern, *The Theory of Games and Economic Behavior*, Princeton, NJ, Princeton University Press, 1953, 3.ª ed.

TABLAS DEL APÉNDICE

Tabla 1. Función de distribución acumulada de la distribución normal estándar.



Z	F(z)	Z	F(z)	Z	F(z)	Z	F(z)	Z	F(z)	Z	F(z)
0,00	0,5000										
0,01	0,5040	0,31	0,6217	0,61	0,7291	0,91	0,8186	1,21	0,8869	1,51	0,9345
0,02	0,5080	0,32	0,6255	0,62	0,7324	0,92	0,8212	1,22	0,8888	1,52	0,9357
0,03	0,5120	0,33	0,6293	0,63	0,7357	0,93	0,8238	1,23	0,8907	1,53	0,9370
0,04	0,5160	0,34	0,6331	0,64	0,7389	0,94	0,8264	1,24	0,8925	1,54	0,9382
0,05	0,5199	0,35	0,6368	0,65	0,7422	0,95	0,8289	1,25	0,8944	1,55	0,9394
0,06	0,5239	0,36	0,6406	0,66	0,7454	0,96	0,8315	1,26	0,8962	1,56	0,9406
0,07	0,5279	0,37	0,6443	0,67	0,7486	0,97	0,8340	1,27	0,8980	1,57	0,9418
0,08	0,5319	0,38	0,6480	0,68	0,7517	0,98	0,8365	1,28	0,8997	1,58	0,9429
0,09	0,5359	0,39	0,6517	0,69	0,7549	0,99	0,8389	1,29	0,9015	1,59	0,9441
0,10	0,5398	0,40	0,6554	0,70	0,7580	1,00	0,8413	1,30	0,9032	1,60	0,9452
0,11	0,5438	0,41	0,6591	0,71	0,7611	1,01	0,8438	1,31	0,9049	1,61	0,9463
0,12	0,5478	0,42	0,6628	0,72	0,7642	1,02	0,8461	1,32	0,9066	1,62	0,9474
0,13	0,5517	0,43	0,6664	0,73	0,7673	1,03	0,8485	1,33	0,9082	1,63	0,9484
0,14	0,5557	0,44	0,6700	0,74	0,7704	1,04	0,8508	1,34	0,9099	1,64	0,9495
0,15	0,5596	0,45	0,6736	0,75	0,7734	1,05	0,8531	1,35	0,9115	1,65	0,9505
0,16	0,5636	0,46	0,6772	0,76	0,7764	1,06	0,8554	1,36	0,9131	1,66	0,9515
0,17	0,5675	0,47	0,6803	0,77	0,7794	1,07	0,8577	1,37	0,9147	1,67	0,9525
0,18	0,5714	0,48	0,6844	0,78	0,7823	1,08	0,8599	1,38	0,9162	1,68	0,9535
0,19	0,5753	0,49	0,6879	0,79	0,7852	1,09	0,8621	1,39	0,9177	1,69	0,9545
0,20	0,5793	0,50	0,6915	0,80	0,7881	1,10	0,8643	1,40	0,9192	1,70	0,9554
0,21	0,5832	0,51	0,6950	0,81	0,7910	1,11	0,8665	1,41	0,9207	1,71	0,9564
0,22	0,5871	0,52	0,6985	0,82	0,7939	1,12	0,8686	1,42	0,9222	1,72	0,9573
0,23	0,5910	0,53	0,7019	0,83	0,7967	1,13	0,8708	1,43	0,9236	1,73	0,9582
0,24	0,5948	0,54	0,7054	0,84	0,7995	1,14	0,8729	1,44	0,9251	1,74	0,9591
0,25	0,5987	0,55	0,7088	0,85	0,8023	1,15	0,8749	1,45	0,9265	1,75	0,9599
0,26	0,6026	0,56	0,7123	0,86	0,8051	1,16	0,8770	1,46	0,9279	1,76	0,9608
0,27	0,6064	0,57	0,7157	0,87	0,8078	1,17	0,8790	1,47	0,9292	1,77	0,9616
0,28	0,6103	0,58	0,7190	0,88	0,8106	1,18	0,8810	1,48	0,9306	1,78	0,9625
0,29	0,6141	0,59	0,7224	0,89	0,8133	1,19	0,8830	1,49	0,9319	1,79	0,9633
0,30	0,6179	0,60	0,7257	0,90	0,8159	1,20	0,8849	1,50	0,9332	1,80	0,9641

Tabla 1. Función de distribución acumulada de la distribución normal estándar (*continuación*).

Z	*F(z)*	*Z*	*F(z)*	*Z*	*F(z)*	*Z*	*F(z)*	*Z*	*F(z)*	*Z*	*F(z)*
1,81	0,9649	2,21	0,9864	2,61	0,9955	3,01	0,9987	3,41	0,9997	3,81	0,9999
1,82	0,9656	2,22	0,9868	2,62	0,9956	3,02	0,9987	3,42	0,9997	3,82	0,9999
1,83	0,9664	2,23	0,9871	2,63	0,9957	3,03	0,9988	3,43	0,9997	3,83	0,9999
1,84	0,9671	2,24	0,9875	2,64	0,9959	3,04	0,9988	3,44	0,9997	3,84	0,9999
1,85	0,9678	2,25	0,9878	2,65	0,9960	3,05	0,9989	3,45	0,9997	3,85	0,9999
1,86	0,9686	2,26	0,9881	2,66	0,9961	3,06	0,9989	3,46	0,9997	3,86	0,9999
1,87	0,9693	2,27	0,9884	2,67	0,9962	3,07	0,9989	3,47	0,9997	3,87	0,9999
1,88	0,9699	2,28	0,9887	2,68	0,9963	3,08	0,9990	3,48	0,9997	3,88	0,9999
1,89	0,9706	2,29	0,9890	2,69	0,9964	3,09	0,9990	3,49	0,9998	3,89	1,0000
1,90	0,9713	2,30	0,9893	2,70	0,9965	3,10	0,9990	3,50	0,9998	3,90	1,0000
1,91	0,9719	2,31	0,9896	2,71	0,9966	3,11	0,9991	3,51	0,9998	3,91	1,0000
1,92	0,9726	2,32	0,9898	2,72	0,9967	3,12	0,9991	3,52	0,9998	3,92	1,0000
1,93	0,9732	2,33	0,9901	2,73	0,9968	3,13	0,9991	3,53	0,9998	3,93	1,0000
1,94	0,9738	2,34	0,9904	2,74	0,9969	3,14	0,9992	3,54	0,9998	3,94	1,0000
1,95	0,9744	2,35	0,9906	2,75	0,9970	3,15	0,9992	3,55	0,9998	3,95	1,0000
1,96	0,9750	2,36	0,9909	2,76	0,9971	3,16	0,9992	3,56	0,9998	3,96	1,0000
1,97	0,9756	2,37	0,9911	2,77	0,9972	3,17	0,9992	3,57	0,9998	3,97	1,0000
1,98	0,9761	2,38	0,9913	2,78	0,9973	3,18	0,9993	3,58	0,9998	3,98	1,0000
1,99	0,9767	2,39	0,9916	2,79	0,9974	3,19	0,9993	3,59	0,9998	3,99	1,0000
2,00	0,9772	2,40	0,9918	2,80	0,9974	3,20	0,9993	3,60	0,9998		
2,01	0,9778	2,41	0,9920	2,81	0,9975	3,21	0,9993	3,61	0,9998		
2,02	0,9783	2,42	0,9922	2,82	0,9976	3,22	0,9994	3,62	0,9999		
2,03	0,9788	2,43	0,9925	2,83	0,9977	3,23	0,9994	3,63	0,9999		
2,04	0,9793	2,44	0,9927	2,84	0,9977	3,24	0,9994	3,64	0,9999		
2,05	0,9798	2,45	0,9929	2,85	0,9978	3,25	0,9994	3,65	0,9999		
2,06	0,9803	2,46	0,9931	2,86	0,9979	3,26	0,9994	3,66	0,9999		
2,07	0,9808	2,47	0,9932	2,87	0,9979	3,27	0,9995	3,67	0,9999		
2,08	0,9812	2,48	0,9934	2,88	0,9980	3,28	0,9995	3,68	0,9999		
2,09	0,9817	2,49	0,9936	2,89	0,9981	3,29	0,9995	3,69	0,9999		
2,10	0,9821	2,50	0,9938	2,90	0,9981	3,30	0,9995	3,70	0,9999		
2,11	0,9826	2,51	0,9940	2,91	0,9982	3,31	0,9995	3,71	0,9999		
2,12	0,9830	2,52	0,9941	2,92	0,9982	3,32	0,9996	3,72	0,9999		
2,13	0,9834	2,53	0,9943	2,93	0,9983	3,33	0,9996	3,73	0,9999		
2,14	0,9838	2,54	0,9945	2,94	0,9984	3,34	0,9996	3,74	0,9999		
2,15	0,9842	2,55	0,9946	2,95	0,9984	3,35	0,9996	3,75	0,9999		
2,16	0,9846	2,56	0,9948	2,96	0,9985	3,36	0,9996	3,76	0,9999		
2,17	0,9850	2,57	0,9949	2,97	0,9985	3,37	0,9996	3,77	0,9999		
2,18	0,9854	2,58	0,9951	2,98	0,9986	3,38	0,9996	3,78	0,9999		
2,19	0,9857	2,59	0,9952	2,99	0,9986	3,39	0,9997	3,79	0,9999		
2,20	0,9861	2,60	0,9953	3,00	0,9986	3,40	0,9997	3,80	0,9999		

Tabla 2. Función de probabilidad de la distribución binomial.

La tabla muestra la probabilidad de que se obtengan x éxitos en n pruebas independientes, cada una con una probabilidad de éxito P. Por ejemplo, la probabilidad de que haya cuatro éxitos en ocho pruebas independientes, cada una con una probabilidad de éxito de 0,35, es 0,1875.

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0312
	1	0,2036	0,3280	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1562
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0004	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1562
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0312
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0312
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0312
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703

Tabla 2. Función de probabilidad de la distribución binomial (*continuación*).

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0226	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0015	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
	2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056

Tabla 2. Función de probabilidad de la distribución binomial (*continuación*).

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0348	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	0	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
	1	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
	2	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018
	3	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0888	0,0468	0,0215	0,0085
	4	0,0061	0,0514	0,1311	0,2001	0,2552	0,2040	0,1553	0,1014	0,0572	0,0278
	5	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
	6	0,0001	0,0028	0,0180	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
	7	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
	8	0,0000	0,0001	0,0009	0,0055	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
	9	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
	10	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
	11	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0029	0,0085
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	0	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
	1	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
	2	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
	3	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
	4	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
	5	0,0010	0,0175	0,0068	0,1361	0,1914	0,2081	0,1849	0,1379	0,0875	0,0472
	6	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
	7	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
	8	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1134	0,1606	0,1883	0,1855
	9	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855

Tabla 2. Función de probabilidad de la distribución binomial (*continuación*).

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	10	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1484
	11	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
	12	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	0	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
	1	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
	2	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
	3	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0031
	4	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
	5	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
	6	0,0002	0,0052	0,0301	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708
	7	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
	8	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
	9	0,0000	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
	10	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
	11	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
	12	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0044	0,0134	0,0327
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	0	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
	1	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
	2	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
	3	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
	4	0,0112	0,0798	0,1714	0,2182	0,2023	0,1419	0,0909	0,0467	0,0203	0,0074
	5	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
	6	0,0002	0,0069	0,0374	0,0955	0,1574	0,1916	0,1844	0,1451	0,0949	0,0518
	7	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
	8	0,0000	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
	9	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
	10	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
	11	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
	12	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
	13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	0	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
	1	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
	2	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
	3	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
	4	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
	5	0,0022	0,0319	0,1028	0,1746	0,2023	0,1789	0,1272	0,0746	0,0365	0,0148
	6	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
	7	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
	8	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201

Tabla 2. Función de probabilidad de la distribución binomial (*continuación*).

n	*x*	*P*									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	9	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
	10	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
	11	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
	12	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
	13	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0150	0,0370
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
	16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
	17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
	18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
	19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Tabla 3. Probabilidades binomiales acumuladas.

La tabla muestra la probabilidad de que se obtengan x éxitos o menos en n pruebas independientes cada una con una probabilidad de éxito P. Por ejemplo, la probabilidad de que se obtengan dos éxitos o menos en cuatro pruebas independientes, cada una con una probabilidad de éxito de 0,365, es 0,874.

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,500
	20	0,902	0,81	0,722	0,64	0,562	0,49	0,422	0,36	0,302	0,25
	1	0,998	0,99	0,978	0,96	0,937	0,91	0,877	0,84	0,797	0,75
	2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
3	0	0,857	0,729	0,614	0,512	0,422	0,343	0,275	0,216	0,166	0,125
	1	0,993	0,972	0,939	0,896	0,844	0,784	0,718	0,648	0,575	0,500
	2	1,00	0,999	0,997	0,992	0,984	0,973	0,957	0,936	0,909	0,875
	3	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
4	0	0,815	0,656	0,522	0,41	0,316	0,24	0,179	0,13	0,092	0,062
	1	0,986	0,948	0,89	0,819	0,738	0,652	0,563	0,475	0,391	0,312
	2	1,00	0,996	0,988	0,973	0,949	0,916	0,874	0,821	0,759	0,687
	3	1,00	1,00	0,999	0,998	0,996	0,992	0,985	0,974	0,959	0,937
	4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
5	0	0,774	0,59	0,444	0,328	0,237	0,168	0,116	0,078	0,05	0,031
	1	0,977	0,919	0,835	0,737	0,633	0,528	0,428	0,337	0,256	0,187
	2	0,999	0,991	0,973	0,942	0,896	0,837	0,765	0,683	0,593	0,500
	3	1,00	1,00	0,998	0,993	0,984	0,969	0,946	0,913	0,869	0,812
	4	1,00	1,00	1,00	1,00	0,999	0,998	0,995	0,99	0,982	0,969
	5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
6	0	0,735	0,531	0,377	0,262	0,178	0,118	0,075	0,047	0,028	0,016
	1	0,967	0,886	0,776	0,655	0,534	0,42	0,319	0,233	0,164	0,109
	2	0,998	0,984	0,953	0,901	0,831	0,744	0,647	0,544	0,442	0,344
	3	1,00	0,999	0,994	0,983	0,962	0,93	0,883	0,821	0,745	0,656
	4	1,00	1,00	1,00	0,998	0,995	0,989	0,978	0,959	0,93	1,891
	5	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,996	0,992	0,984
	6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
7	0	0,698	0,478	0,321	0,21	0,133	0,082	0,049	0,028	0,015	0,008
	1	0,956	0,85	0,717	0,577	0,445	0,329	0,234	0,159	0,102	0,062
	2	0,996	0,974	0,926	0,852	0,756	0,647	0,532	0,42	0,316	0,227
	3	1,00	0,997	0,988	0,967	0,929	0,874	0,80	0,71	0,608	0,500
	4	1,00	1,00	0,999	0,995	0,987	0,971	0,944	0,904	0,847	0,773
	5	1,00	1,00	1,00	1,00	0,999	0,996	0,991	0,981	0,964	0,937
	6	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,996	0,992
	7	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
8	0	0,663	0,43	0,272	0,168	0,10	0,058	0,032	0,017	0,008	0,004
	1	0,943	0,813	0,657	0,503	0,367	0,255	0,169	0,106	0,063	0,035
	2	0,994	0,962	0,895	0,797	0,679	0,552	0,428	0,315	0,22	0,145
	3	1,00	0,995	0,979	0,944	0,886	0,806	0,706	0,594	0,477	0,363
	4	1,00	1,00	0,997	0,99	0,973	0,942	0,894	0,826	0,74	0,637
	5	1,00	1,00	1,00	0,999	0,996	0,989	0,975	0,95	0,912	0,855
	6	1,00	1,00	1,00	1,00	1,00	0,999	0,996	0,991	0,982	0,965
	7	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,996
	8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
9	0	0,63	0,387	0,232	0,134	0,075	0,04	0,021	0,01	0,005	0,002
	1	0,929	0,775	0,599	0,436	0,30	0,196	0,121	0,071	0,039	0,020
	2	0,992	0,947	0,859	0,738	0,601	0,463	0,337	0,232	0,15	0,090
	3	0,999	0,992	0,966	0,914	0,834	0,73	0,609	0,483	0,361	0,254
	4	1,00	0,999	0,994	0,98	0,951	0,901	0,828	0,733	0,621	0,500
	5	1,00	1,00	0,999	0,997	0,99	0,975	0,946	0,901	0,834	0,746
	6	1,00	1,00	1,00	1,00	0,999	0,996	0,989	0,975	0,95	0,910
	7	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,996	0,991	0,980
	8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998
	9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000

Tabla 3. Probabilidades binomiales acumuladas (*continuación*).

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,500
10	0	0,599	0,349	0,197	0,107	0,056	0,028	0,013	0,006	0,003	0,001
	1	0,914	0,736	0,544	0,376	0,244	0,149	0,086	0,046	0,023	0,011
	2	0,988	0,93	0,82	0,678	0,526	0,383	0,262	0,167	0,10	0,055
	3	0,999	0,987	0,95	0,879	0,776	0,65	0,514	0,382	0,266	0,172
	4	1,00	0,998	0,99	0,967	0,922	0,85	0,75	1,633	0,504	0,377
	5	1,00	1,00	0,999	0,994	0,98	0,953	0,905	0,834	0,738	0,623
	6	1,00	1,00	1,00	0,999	0,996	0,989	0,974	0,945	0,898	0,828
	7	1,00	1,00	1,00	1,00	1,00	0,998	0,995	0,988	0,973	0,945
	8	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,995	0,989
	9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999
	10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
11	0	0,569	0,314	0,167	0,086	0,042	0,02	0,009	0,004	0,00	1,000
	1	0,898	0,697	0,492	0,322	0,197	0,113	0,06	1,03	0,014	0,006
	2	0,985	0,9	1,779	0,617	0,455	0,313	0,20	0,119	0,065	0,033
	3	0,998	0,98	1,93	1,839	0,713	0,57	0,426	0,296	0,19	1,113
	4	1,00	0,997	0,984	0,95	0,885	0,79	0,668	0,533	0,397	0,274
	5	1,00	1,00	0,997	0,988	0,966	0,922	0,85	1,753	0,633	0,500
	6	1,00	1,00	1,00	0,998	0,992	0,978	0,95	0,90	1,826	0,726
	7	1,00	1,00	1,00	1,00	0,999	0,996	0,988	0,97	1,939	0,887
	8	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,994	0,985	0,967
	9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,994
	10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
	11	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
12	0	0,54	0,282	0,142	0,069	0,032	0,014	0,006	0,002	0,00	1,000
	1	0,882	0,659	0,443	0,275	0,158	0,085	0,042	0,02	0,008	0,003
	2	0,98	0,889	0,736	0,558	0,39	1,253	0,15	1,083	0,042	0,019
	3	0,998	0,974	0,908	0,795	0,649	0,493	0,347	0,225	0,134	0,073
	4	1,00	0,996	0,976	0,927	0,842	0,724	0,583	0,438	0,304	0,194
	5	1,00	0,999	0,995	0,98	1,946	0,882	0,787	0,665	0,527	0,387
	6	1,00	1,00	0,999	0,996	0,986	0,96	1,915	0,842	0,739	0,613
	7	1,00	1,00	1,00	0,999	0,997	0,99	1,974	0,943	0,888	0,806
	8	1,00	1,00	1,00	1,00	1,00	0,998	0,994	0,985	0,964	0,927
	9	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,992	0,981
	10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,997
	11	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
	12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
13	0	0,513	0,254	0,12	1,055	0,024	0,0	1,004	0,00	1,00	0,000
	1	0,865	0,62	1,398	0,234	0,127	0,064	0,03	0,013	0,005	0,002
	2	0,975	0,866	0,692	0,502	0,333	0,202	0,113	0,058	0,027	0,011
	3	0,997	0,966	0,882	0,747	0,584	0,42	1,278	0,169	0,093	0,046
	4	1,00	0,994	0,966	0,90	1,794	0,654	0,50	1,353	0,228	0,133
	5	1,00	0,999	0,992	0,97	0,92	0,835	0,716	0,574	0,427	0,291
	6	1,00	1,00	0,999	0,993	0,976	0,938	0,87	1,77	1,644	0,50
	7	1,00	1,00	1,00	0,999	0,994	0,982	0,954	0,902	0,82	1,709
	8	1,00	1,00	1,00	1,00	0,999	0,996	0,987	0,968	0,93	0,867
	9	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,992	0,98	0,954
	10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,996	0,989
	11	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998
	12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
14	0	0,488	0,229	0,103	0,044	0,018	0,007	0,002	0,00	1,00	0,000
	1	0,847	0,585	0,357	0,198	0,10	1,047	0,02	1,008	0,003	0,001
	2	0,97	0,842	0,648	0,448	0,28	1,16	1,084	0,04	0,017	0,006
	3	0,996	0,956	0,853	0,698	0,52	1,355	0,22	0,124	0,063	0,029
	4	1,00	0,99	1,953	0,87	0,742	0,584	0,423	0,279	0,167	0,090
	5	1,00	0,999	0,988	0,956	0,888	0,78	1,64	1,486	0,337	0,212

Tabla 3. Probabilidades binomiales acumuladas (*continuación*).

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,500
	6	1,00	1,00	0,998	0,988	0,962	0,907	0,816	0,692	0,546	0,395
	7	1,00	1,00	1,00	0,998	0,99	0,969	0,925	0,85	0,74	1,605
	8	1,00	1,00	1,00	1,00	0,998	0,992	0,976	0,942	0,88	1,788
	9	1,00	1,00	1,00	1,00	1,00	0,998	0,994	0,982	0,957	0,910
	10	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,996	0,989	0,971
	11	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998	0,994
	12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999
	13	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
15	0	0,463	0,206	0,087	0,035	0,013	0,005	0,002	0,00	0,00	0,000
	1	0,829	0,549	0,319	0,167	0,08	0,035	0,014	0,005	0,002	0,000
	2	0,964	0,816	0,604	0,398	0,236	0,127	0,062	0,027	0,01	1,004
	3	0,995	0,944	0,823	0,648	0,46	1,297	0,173	0,09	1,042	0,018
	4	0,999	0,987	0,938	0,836	0,686	0,515	0,352	0,217	0,12	0,059
	5	1,00	0,998	0,983	0,939	0,852	0,722	0,564	0,403	0,26	1,151
	6	1,00	1,00	0,996	0,982	0,943	0,869	0,755	0,6	1,452	0,304
	7	1,00	1,00	0,999	0,996	0,983	0,95	0,887	0,787	0,654	0,500
	8	1,00	1,00	1,00	0,999	0,996	0,985	0,958	0,905	0,818	0,696
	9	1,00	1,00	1,00	1,00	0,999	0,996	0,988	0,966	0,923	0,849
	10	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,99	1,975	0,941
	11	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,998	0,994	0,982
	12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,996
	13	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
16	0	0,44	0,185	0,074	0,028	0,0	1,003	0,00	1,00	0,00	0,000
	1	0,811	1,515	0,284	0,14	1,063	0,026	0,0	1,003	0,00	1,000
	2	0,957	0,789	0,56	1,352	0,197	0,099	0,045	0,018	0,007	0,002
	3	0,993	0,932	0,79	0,598	0,405	0,246	0,134	0,065	0,028	0,011
	4	0,999	0,983	0,92	1,798	0,63	0,45	0,289	0,167	0,085	0,038
	5	1,00	0,997	0,976	0,918	0,8	1,66	0,49	0,329	0,198	0,105
	6	1,00	0,999	0,994	0,973	0,92	0,825	0,688	0,527	0,366	0,227
	7	1,00	1,00	0,999	0,993	0,973	0,926	0,84	1,716	0,563	0,402
	8	1,00	1,00	1,00	0,999	0,993	0,974	0,933	0,858	0,744	0,598
	9	1,00	1,00	1,00	1,00	0,998	0,993	0,977	0,942	0,876	0,773
	10	1,00	1,00	1,00	1,00	1,00	0,998	0,994	0,98	1,95	1,895
	11	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,995	0,985	0,962
	12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,989
	13	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998
	14	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
17	0	0,418	0,167	0,063	0,023	0,008	0,002	0,00	1,00	0,00	0,000
	1	0,792	0,482	0,252	0,118	0,05	0,019	0,007	0,002	0,00	1,000
	2	0,95	0,762	0,52	0,3	1,164	0,077	0,033	0,012	0,004	0,001
	3	0,99	1,917	0,756	0,549	0,353	0,202	0,103	0,046	0,018	0,006
	4	0,999	0,978	0,90	1,758	0,574	0,389	0,235	0,126	0,06	0,025
	5	1,00	0,995	0,968	0,894	0,765	0,597	0,42	0,264	0,147	0,072
	6	1,00	0,999	0,992	0,962	0,893	0,775	0,619	0,448	0,29	0,166
	7	1,00	1,00	0,998	0,989	0,96	0,895	0,787	0,64	1,474	0,315
	8	1,00	1,00	1,00	0,997	0,988	0,96	0,90	1,80	1,663	0,500
	9	1,00	1,00	1,00	1,00	0,997	0,987	0,962	0,908	0,817	0,685
	10	1,00	1,00	1,00	1,00	0,999	0,997	0,988	0,965	0,917	0,834
	11	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,989	0,97	0,928
	12	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,99	1,975
	13	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,998	0,994
	14	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999
	15	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
18	0	0,397	0,15	0,054	0,018	0,006	0,002	0,00	0,00	0,00	0,000
	1	0,774	0,45	0,224	0,099	0,039	0,014	0,005	0,00	1,00	0,000

Tabla 3. Probabilidades binomiales acumuladas (*continuación*).

n	x	P									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,500
	2	0,942	0,734	0,48	0,27	1,135	0,06	0,024	0,008	0,003	0,001
	3	0,989	0,902	0,72	0,50	1,306	0,165	0,078	0,033	0,012	0,004
	4	0,998	0,972	0,879	0,716	0,519	0,333	0,189	0,094	0,04	1,015
	5	1,00	0,994	0,958	0,867	0,717	0,534	0,355	0,209	0,108	0,048
	6	1,00	0,999	0,988	0,949	0,86	1,722	0,549	0,374	0,226	0,119
	7	1,00	1,00	0,997	0,984	0,943	0,859	0,728	0,563	0,39	1,240
	8	1,00	1,00	0,999	0,996	0,98	1,94	0,86	1,737	0,578	0,407
	9	1,00	1,00	1,00	0,999	0,995	0,979	0,94	0,865	0,747	0,593
	10	1,00	1,00	1,00	1,00	0,999	0,994	0,979	0,942	0,872	0,760
	11	1,00	1,00	1,00	1,00	1,00	0,999	0,994	0,98	0,946	0,881
	12	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,994	0,982	0,952
	13	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,995	0,985
	14	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,996
	15	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999
	16	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
19	0	0,377	0,135	0,046	0,014	0,004	0,00	1,00	0,00	0,00	0,000
	1	0,755	0,42	0,198	0,083	0,03	1,0	1,003	0,00	1,00	0,000
	2	0,933	0,705	0,44	1,237	0,11	1,046	0,017	0,005	0,002	0,000
	3	0,987	0,885	0,684	0,455	0,263	0,133	0,059	0,023	0,008	0,002
	4	0,998	0,965	0,856	0,673	0,465	0,282	0,15	0,07	0,028	0,010
	5	1,00	0,99	1,946	0,837	0,668	0,474	0,297	0,163	0,078	0,032
	6	1,00	0,998	0,984	0,932	0,825	0,666	0,48	1,308	0,173	0,084
	7	1,00	1,00	0,996	0,977	0,923	0,818	0,666	0,488	0,317	0,180
	8	1,00	1,00	0,999	0,993	0,97	1,916	0,815	0,667	0,494	0,324
	9	1,00	1,00	1,00	0,998	0,99	1,967	0,913	0,814	0,67	1,500
	10	1,00	1,00	1,00	1,00	0,998	0,989	0,965	0,912	0,816	0,676
	11	1,00	1,00	1,00	1,00	1,00	0,997	0,989	0,965	0,913	0,820
	12	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,988	0,966	0,916
	13	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,989	0,968
	14	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,997	0,990
	15	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999	0,998
	16	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000
20	0	0,358	0,122	0,039	0,012	0,003	0,00	1,00	0,00	0,00	0,000
	1	0,736	0,392	0,176	0,069	0,024	0,008	0,002	0,00	1,00	0,000
	2	0,925	0,677	0,405	0,206	0,09	1,035	0,012	0,004	0,00	1,000
	3	0,984	0,867	0,648	0,41	1,225	0,107	0,044	0,016	0,005	0,001
	4	0,997	0,957	0,83	0,63	0,415	0,238	0,118	0,05	1,019	0,006
	5	1,00	0,989	0,933	0,804	0,617	0,416	0,245	0,126	0,055	0,021
	6	1,00	0,998	0,978	0,913	0,786	0,608	0,417	0,25	0,13	0,058
	7	1,00	1,00	0,994	0,968	0,898	0,772	0,60	1,416	0,252	0,132
	8	1,00	1,00	0,999	0,99	0,959	0,887	0,762	0,596	0,414	0,252
	9	1,00	1,00	1,00	0,997	0,986	0,952	0,878	0,755	0,59	1,412
	10	1,00	1,00	1,00	0,999	0,996	0,983	0,947	0,872	0,75	1,588
	11	1,00	1,00	1,00	1,00	0,999	0,995	0,98	0,943	0,869	0,748
	12	1,00	1,00	1,00	1,00	1,00	0,999	0,994	0,979	0,942	0,868
	13	1,00	1,00	1,00	1,00	1,00	1,00	0,998	0,994	0,979	0,942
	14	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,998	0,994	0,979
	15	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,998	0,994
	16	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,999
	17	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,000

Tabla 4. Valores de $e^{-\lambda}$.

λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$	λ	$e^{-\lambda}$
0,00	1,000000	2,60	0,074274	5,10	0,006097	7,60	0,000501
0,10	0,904837	2,70	0,067206	5,20	0,005517	7,70	0,000453
0,20	0,818731	2,80	0,060810	5,30	0,004992	7,80	0,000410
0,30	0,740818	2,90	0,055023	5,40	0,004517	7,90	0,000371
0,40	0,670320	3,00	0,049787	5,50	0,004087	8,00	0,000336
0,50	0,606531	3,10	0,045049	5,60	0,003698	8,10	0,000304
0,60	0,548812	3,20	0,040762	5,70	0,003346	8,20	0,000275
0,70	0,496585	3,30	0,036883	5,80	0,003028	8,30	0,000249
0,80	0,449329	3,40	0,033373	5,90	0,002739	8,40	0,000225
0.90	0,406570	3,50	0,030197	6,00	0,002479	8,50	0,000204
1,00	0,367879	3,60	0,027324	6,10	0,002243	8,60	0,000184
1,10	0,332871	3,70	0,024724	6,20	0,002029	8,70	0,000167
1,20	0,301194	3,80	0,022371	6,30	0,001836	8,80	0,000151
1,30	0,272532	3,90	0,020242	6,40	0,001661	8,90	0,000136
1,40	0,246597	4,00	0,018316	6,50	0,001503	9,00	0,000123
1,50	0,223130	4,10	0,016573	6,60	0,001360	9,10	0,000112
1,60	0,201897	4,20	0,014996	6,70	0,001231	9,20	0,000101
1,70	0,182684	4,30	0,013569	6,80	0,001114	9,30	0,000091
1,80	0,165299	4,40	0,012277	6,90	0,001008	9,40	0,000083
1,90	0,149569	4,50	0,011109	7,00	0,000912	9,50	0,000075
2,00	0,135335	4,60	0,010052	7,10	0,000825	9,60	0,000068
2,10	0,122456	4,70	0,009095	7,20	0,000747	9,70	0,000061
2,20	0,110803	4,80	0,008230	7,30	0,000676	9,80	0,000056
2,30	0,100259	4,90	0,007447	7,40	0,000611	9,90	0,000050
2,40	0,090718	5,00	0,006738	7,50	0,000553	10,00	0,000045
2,50	0,082085						

Tabla 5. Probabilidades de Poisson individuales.

	Tasa Media de Llegada λ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,1
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0	0,0	0,0	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031
6	0,0	0,0	0,0	0,0	0,0	0,0	0,0001	0,0002	0,0003	0,0005
7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0001

	Tasa Media de Llegada λ									
	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353
1	0,3662	0,3614	0,3543	0,3452	0,3347	0,3230	0,3106	0,2975	0,2842	0,2707
2	0,2014	0,2169	0,2303	0,2417	0,2510	0,2584	0,2640	0,2678	0,2700	0,2707
3	0,0738	0,0867	0,0998	0,1128	0,1255	0,1378	0,1496	0,1607	0,1710	0,1804
4	0,0203	0,0260	0,0324	0,0395	0,0471	0,0551	0,0636	0,0723	0,0812	0,0902
5	0,0045	0,0062	0,0084	0,0111	0,0141	0,0176	0,0216	0,0260	0,0309	0,0361
6	0,0008	0,0012	0,0018	0,0026	0,0035	0,0047	0,0061	0,0078	0,0098	0,0120
7	0,0001	0,0002	0,0003	0,0005	0,0008	0,0011	0,0015	0,0020	0,0027	0,0034
8	0,0	0,0	0,0001	0,0001	0,0001	0,0002	0,0003	0,0005	0,0006	0,0009
9	0,0	0,0	0,0	0,0	0,0	0,0	0,0001	0,0001	0,0001	0,0002

	Tasa Media de Llegada λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,2572	0,2438	0,2306	0,2177	0,2052	0,1931	0,1815	0,1703	0,1596	0,1494
2	0,2700	0,2681	0,2652	0,2613	0,2565	0,2510	0,2450	0,2384	0,2314	0,2240
3	0,1890	0,1966	0,2033	0,2090	0,2138	0,2176	0,2205	0,2225	0,2237	0,2240
4	0,0992	0,1082	0,1169	0,1254	0,1336	0,1414	0,1488	0,1557	0,1622	0,1680
5	0,0417	0,0476	0,0538	0,0602	0,0668	0,0735	0,0804	0,0872	0,0940	0,1008
6	0,0146	0,0174	0,0206	0,0241	0,0278	0,0319	0,0362	0,0407	0,0455	0,0504
7	0,0044	0,0055	0,0068	0,0083	0,0099	0,0118	0,0139	0,0163	0,0188	0,0216
8	0,0011	0,0015	0,0019	0,0025	0,0031	0,0038	0,0047	0,0057	0,0068	0,0081
9	0,0003	0,0004	0,0005	0,0007	0,0009	0,0011	0,0014	0,0018	0,0022	0,0027
10	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0004	0,0005	0,0006	0,0008
11	0,0	0,0	0,0	0,0	0,0	0,0001	0,0001	0,0001	0,0002	0,0002
12	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0001

	Tasa Media de Llegada λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0334	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1397	0,1304	0,1217	0,1135	0,1057	0,0984	0,0915	0,0850	0,0789	0,0733
2	0,2165	0,2087	0,2008	0,1929	0,1850	0,1771	0,1692	0,1615	0,1539	0,1465
3	0,2237	0,2226	0,2209	0,2186	0,2158	0,2125	0,2087	0,2046	0,2001	0,1954
4	0,1733	0,1781	0,1823	0,1858	0,1888	0,1912	0,1931	0,1944	0,1951	0,1954
5	0,1075	0,1140	0,1203	0,1264	0,1322	0,1377	0,1429	0,1477	0,1522	0,1563
6	0,0555	0,0608	0,0662	0,0716	0,0771	0,0826	0,0881	0,0936	0,0989	0,1042
7	0,0246	0,0278	0,0312	0,0348	0,0385	0,0425	0,0466	0,0508	0,0551	0,0595
8	0,0095	0,0111	0,0129	0,0148	0,0169	0,0191	0,0215	0,0241	0,0269	0,0298
9	0,0033	0,0040	0,0047	0,0056	0,0066	0,0076	0,0089	0,0102	0,0116	0,0132
10	0,0010	0,0013	0,0016	0,0019	0,0023	0,0028	0,0033	0,0039	0,0045	0,0053
11	0,0003	0,0004	0,0005	0,0006	0,0007	0,0009	0,0011	0,0013	0,0016	0,0019
12	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005	0,0006
13	0,0	0,0	0,0	0,0	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
14	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0001

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	4,1	4,2	4,3	4,4	4,54	0,64	0,74	0,84	0,95	0,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0679	0,0630	0,0583	0,0540	0,0500	0,0462	0,0427	0,0395	0,0365	0,0337
2	0,1393	0,1323	0,1254	0,1188	0,1125	0,1063	0,1005	0,0948	0,0894	0,0842
3	0,1904	0,1852	0,1798	0,1743	0,1687	0,1631	0,1574	0,1517	0,1460	0,1404
4	0,1951	0,1944	0,1933	0,1917	0,1898	0,1875	0,1849	0,1820	0,1789	0,1755
5	0,1600	0,1633	0,1662	0,1687	0,1708	0,1725	0,1738	0,1747	0,1753	0,1755
6	0,1093	0,1143	0,1191	0,1237	0,1281	0,1323	0,1362	0,1398	0,1432	0,1462
7	0,0640	0,0686	0,0732	0,0778	0,0824	0,0869	0,0914	0,0959	0,1002	0,1044
8	0,0328	0,0360	0,0393	0,0428	0,0463	0,0500	0,0537	0,0575	0,0614	0,0653
9	0,0150	0,0168	0,0188	0,0209	0,0232	0,0255	0,0281	0,0307	0,0334	0,0363
10	0,0061	0,0071	0,0081	0,0092	0,0104	0,0118	0,0132	0,0147	0,0164	0,0181
11	0,0023	0,0027	0,0032	0,0037	0,0043	0,0049	0,0056	0,0064	0,0073	0,0082
12	0,0008	0,0009	0,0011	0,0013	0,0016	0,0019	0,0022	0,0026	0,0030	0,0034
13	0,0002	0,0003	0,0004	0,0005	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013
14	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002	0,0003	0,0003	0,0004	0,0005

	Tasa Media de Llegada λ									
	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0311	0,0287	0,0265	0,0244	0,0225	0,0207	0,0191	0,0176	0,0162	0,0149
2	0,0793	0,0746	0,0701	0,0659	0,0618	0,0580	0,0544	0,0509	0,0477	0,0446
3	0,1348	0,1293	0,1239	0,1185	0,1133	0,1082	0,1033	0,0985	0,0938	0,0892
4	0,1719	0,1681	0,1641	0,1600	0,1558	0,1515	0,1472	0,1428	0,1383	0,1339
5	0,1753	0,1748	0,1740	0,1728	0,1714	0,1697	0,1678	0,1656	0,1632	0,1606
6	0,1490	0,1515	0,1537	0,1555	0,1571	0,1584	0,1594	0,1601	0,1605	0,1606
7	0,1086	0,1125	0,1163	0,1200	0,1234	0,1267	0,1298	0,1326	0,1353	0,1377
8	0,0692	0,0731	0,0771	0,0810	0,0849	0,0887	0,0925	0,0962	0,0998	0,1033
9	0,0392	0,0423	0,0454	0,0486	0,0519	0,0552	0,0586	0,0620	0,0654	0,0688
10	0,0200	0,0220	0,0241	0,0262	0,0285	0,0309	0,0334	0,0359	0,0386	0,0413
11	0,0093	0,0104	0,0116	0,0129	0,0143	0,0157	0,0173	0,0190	0,0207	0,0225
12	0,0039	0,0045	0,0051	0,0058	0,0065	0,0073	0,0082	0,0092	0,0102	0,0113
13	0,0015	0,0018	0,0021	0,0024	0,0028	0,0032	0,0036	0,0041	0,0046	0,0052
14	0,0006	0,0007	0,0008	0,0009	0,0011	0,0013	0,0015	0,0017	0,0019	0,0022

	Tasa Media de Llegada λ									
	6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0137	0,0126	0,0116	0,0106	0,0098	0,0090	0,0082	0,0076	0,0070	0,0064
2	0,0417	0,0390	0,0364	0,0340	0,0318	0,0296	0,0276	0,0258	0,0240	0,0223
3	0,0848	0,0806	0,0765	0,0726	0,0688	0,0652	0,0617	0,0584	0,0552	0,0521
4	0,1294	0,1249	0,1205	0,1162	0,1118	0,1076	0,1034	0,0992	0,0952	0,0912
5	0,1579	0,1549	0,1519	0,1487	0,1454	0,1420	0,1385	0,1349	0,1314	0,1277
6	0,1605	0,1601	0,1595	0,1586	0,1575	0,1562	0,1546	0,1529	0,1511	0,1490
7	0,1399	0,1418	0,1435	0,1450	0,1462	0,1472	0,1480	0,1486	0,1489	0,1490
8	0,1066	0,1099	0,1130	0,1160	0,1188	0,1215	0,1240	0,1263	0,1284	0,1304
9	0,0723	0,0757	0,0791	0,0825	0,0858	0,0891	0,0923	0,0954	0,0985	0,1014
10	0,0441	0,0469	0,0498	0,0528	0,0558	0,0588	0,0618	0,0649	0,0679	0,0710
11	0,0244	0,0265	0,0285	0,0307	0,0330	0,0353	0,0377	0,0401	0,0426	0,0452
12	0,0124	0,0137	0,0150	0,0164	0,0179	0,0194	0,0210	0,0227	0,0245	0,0263
13	0,0058	0,0065	0,0073	0,0081	0,0089	0,0099	0,0108	0,0119	0,0130	0,0142
14	0,0025	0,0029	0,0033	0,0037	0,0041	0,0046	0,0052	0,0058	0,0064	0,0071

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003
1	0,0059	0,0054	0,0049	0,0045	0,0041	0,0038	0,0035	0,0032	0,0029	0,0027
2	0,0208	0,0194	0,0180	0,0167	0,0156	0,0145	0,0134	0,0125	0,0116	0,0107
3	0,0492	0,0464	0,0438	0,0413	0,0389	0,0366	0,0345	0,0324	0,0305	0,0286
4	0,0874	0,0836	0,0799	0,0764	0,0729	0,0696	0,0663	0,0632	0,0602	0,0573
5	0,1241	0,1204	0,1167	0,1130	0,1094	0,1057	0,1021	0,0986	0,0951	0,0916
6	0,1468	0,1445	0,1420	0,1394	0,1367	0,1339	0,1311	0,1282	0,1252	0,1221
7	0,1489	0,1486	0,1481	0,1474	0,1465	0,1454	0,1442	0,1428	0,1413	0,1396
8	0,1321	0,1337	0,1351	0,1363	0,1373	0,1381	0,1388	0,1392	0,1395	0,1396
9	0,1042	0,1070	0,1096	0,1121	0,1144	0,1167	0,1187	0,1207	0,1224	0,1241
10	0,0740	0,0770	0,08	0,0829	0,0858	0,0887	0,0914	0,0941	0,0967	0,0993
11	0,0478	0,0504	0,0531	0,0558	0,0585	0,0613	0,0640	0,0667	0,0695	0,0722
12	0,0283	0,0303	0,0323	0,0344	0,0366	0,0388	0,0411	0,0434	0,0457	0,0481
13	0,0154	0,0168	0,0181	0,0196	0,0211	0,0227	0,0243	0,0260	0,0278	0,0296
14	0,0078	0,0086	0,0095	0,0104	0,0113	0,0123	0,0134	0,0145	0,0157	0,0169
15	0,0037	0,0041	0,0046	0,0051	0,0057	0,0062	0,0069	0,0075	0,0083	0,0090
16	0,0016	0,0019	0,0021	0,0024	0,0026	0,0030	0,0033	0,0037	0,0041	0,0045
17	0,0007	0,0008	0,0009	0,0010	0,0012	0,0013	0,0015	0,0017	0,0019	0,0021
18	0,0003	0,0003	0,0004	0,0004	0,0005	0,0006	0,0006	0,0007	0,0008	0,0009
19	0,0001	0,0001	0,0001	0,0002	0,0002	0,0002	0,0003	0,0003	0,0003	0,0004

	Tasa Media de Llegada λ									
	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0025	0,0023	0,0021	0,0019	0,0017	0,0016	0,0014	0,0013	0,0012	0,0011
2	0,01	0,0092	0,0086	0,0079	0,0074	0,0068	0,0063	0,0058	0,0054	0,0050
3	0,0269	0,0252	0,0237	0,0222	0,0208	0,0195	0,0183	0,0171	0,0160	0,0150
4	0,0544	0,0517	0,0491	0,0466	0,0443	0,0420	0,0398	0,0377	0,0357	0,0337
5	0,0882	0,0849	0,0816	0,0784	0,0752	0,0722	0,0692	0,0663	0,0635	0,0607
6	0,1191	0,1160	0,1128	0,1097	0,1066	0,1034	0,1003	0,0972	0,0941	0,0911
7	0,1378	0,1358	0,1338	0,1317	0,1294	0,1271	0,1247	0,1222	0,1197	0,1171
8	0,1395	0,1392	0,1388	0,1382	0,1375	0,1366	0,1356	0,1344	0,1332	0,1318
9	0,1256	0,1269	0,1280	0,1290	0,1299	0,1306	0,1311	0,1315	0,1317	0,1318
10	0,1017	0,1040	0,1063	0,1084	0,1104	0,1123	0,1140	0,1157	0,1172	0,1186
11	0,0749	0,0776	0,0802	0,0828	0,0853	0,0878	0,0902	0,0925	0,0948	0,0970
12	0,0505	0,0530	0,0555	0,0579	0,0604	0,0629	0,0654	0,0679	0,0703	0,0728
13	0,0315	0,0334	0,0354	0,0374	0,0395	0,0416	0,0438	0,0459	0,0481	0,0504
14	0,0182	0,0196	0,0210	0,0225	0,0240	0,0256	0,0272	0,0289	0,0306	0,0324
15	0,0098	0,0107	0,0116	0,0126	0,0136	0,0147	0,0158	0,0169	0,0182	0,0194
16	0,0050	0,0055	0,0060	0,0066	0,0072	0,0079	0,0086	0,0093	0,0101	0,0109
17	0,0024	0,0026	0,0029	0,0033	0,0036	0,0040	0,0044	0,0048	0,0053	0,0058
18	0,0011	0,0012	0,0014	0,0015	0,0017	0,0019	0,0021	0,0024	0,0026	0,0029
19	0,0005	0,0005	0,0006	0,0007	0,0008	0,0009	0,0010	0,0011	0,0012	0,0014

	Tasa Media de Llegada λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005	0,0005
2	0,0046	0,0043	0,0040	0,0037	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
3	0,0140	0,0131	0,0123	0,0115	0,0107	0,01	0,0093	0,0087	0,0081	0,0076
4	0,0319	0,0302	0,0285	0,0269	0,0254	0,0240	0,0226	0,0213	0,0201	0,0189
5	0,0581	0,0555	0,0530	0,0506	0,0483	0,0460	0,0439	0,0418	0,0398	0,0378
6	0,0881	0,0851	0,0822	0,0793	0,0764	0,0736	0,0709	0,0682	0,0656	0,0631
7	0,1145	0,1118	0,1091	0,1064	0,1037	0,1010	0,0982	0,0955	0,0928	0,0901
8	0,1302	0,1286	0,1269	0,1251	0,1232	0,1212	0,1191	0,1170	0,1148	0,1126

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
9	0,1317	0,1315	0,1311	0,1306	0,13	0,1293	0,1284	0,1274	0,1263	0,1251
10	0,1198	0,1210	0,1219	0,1228	0,1235	0,1241	0,1245	0,1249	0,1250	0,1251
11	0,0991	0,1012	0,1031	0,1049	0,1067	0,1083	0,1098	0,1112	0,1125	0,1137
12	0,0752	0,0776	0,0799	0,0822	0,0844	0,0866	0,0888	0,0908	0,0928	0,0948
13	0,0526	0,0549	0,0572	0,0594	0,0617	0,0640	0,0662	0,0685	0,0707	0,0729
14	0,0342	0,0361	0,0380	0,0399	0,0419	0,0439	0,0459	0,0479	0,05	0,0521
15	0,0208	0,0221	0,0235	0,0250	0,0265	0,0281	0,0297	0,0313	0,0330	0,0347
16	0,0118	0,0127	0,0137	0,0147	0,0157	0,0168	0,0180	0,0192	0,0204	0,0217
17	0,0063	0,0069	0,0075	0,0081	0,0088	0,0095	0,0103	0,0111	0,0119	0,0128
18	0,0032	0,0035	0,0039	0,0042	0,0046	0,0051	0,0055	0,0060	0,0065	0,0071
19	0,0015	0,0017	0,0019	0,0021	0,0023	0,0026	0,0028	0,0031	0,0034	0,0037

	Tasa Media de Llegada λ									
	10,1	10,2	10,3	10,4	10,5	10,6	10,7	10,8	10,9	11,0
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,0000
1	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002
2	0,0021	0,0019	0,0018	0,0016	0,0015	0,0014	0,0013	0,0012	0,0011	0,0010
3	0,0071	0,0066	0,0061	0,0057	0,0053	0,0049	0,0046	0,0043	0,0040	0,0037
4	0,0178	0,0168	0,0158	0,0148	0,0139	0,0131	0,0123	0,0116	0,0109	0,0102
5	0,0360	0,0342	0,0325	0,0309	0,0293	0,0278	0,0264	0,0250	0,0237	0,0224
6	0,0606	0,0581	0,0558	0,0535	0,0513	0,0491	0,0470	0,0450	0,0430	0,0411
7	0,0874	0,0847	0,0821	0,0795	0,0769	0,0743	0,0718	0,0694	0,0669	0,0646
8	0,1103	0,1080	0,1057	0,1033	0,1009	0,0985	0,0961	0,0936	0,0912	0,0888
9	0,1238	0,1224	0,1209	0,1194	0,1177	0,1160	0,1142	0,1124	0,1105	0,1085
10	0,1250	0,1249	0,1246	0,1241	0,1236	0,1230	0,1222	0,1214	0,1204	0,1194
11	0,1148	0,1158	0,1166	0,1174	0,1180	0,1185	0,1189	0,1192	0,1193	0,1194
12	0,0966	0,0984	0,1001	0,1017	0,1032	0,1047	0,1060	0,1072	0,1084	0,1094
13	0,0751	0,0772	0,0793	0,0814	0,0834	0,0853	0,0872	0,0891	0,0909	0,0926
14	0,0542	0,0563	0,0584	0,0604	0,0625	0,0646	0,0667	0,0687	0,0708	0,0728
15	0,0365	0,0383	0,0401	0,0419	0,0438	0,0457	0,0476	0,0495	0,0514	0,0534
16	0,0230	0,0244	0,0258	0,0272	0,0287	0,0303	0,0318	0,0334	0,0350	0,0367
17	0,0137	0,0146	0,0156	0,0167	0,0177	0,0189	0,0200	0,0212	0,0225	0,0237
18	0,0077	0,0083	0,0089	0,0096	0,0104	0,0111	0,0119	0,0127	0,0136	0,0145
19	0,0041	0,0045	0,0048	0,0053	0,0057	0,0062	0,0067	0,0072	0,0078	0,0084
20	0,0021	0,0023	0,0025	0,0027	0,0030	0,0033	0,0036	0,0039	0,0043	0,0046

	Tasa Media de Llegada λ									
	11,1	11,2	11,3	11,4	11,5	11,6	11,7	11,8	11,9	12,0
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
2	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004
3	0,0034	0,0032	0,0030	0,0028	0,0026	0,0024	0,0022	0,0021	0,0019	0,0018
4	0,0096	0,0090	0,0084	0,0079	0,0074	0,0069	0,0065	0,0061	0,0057	0,0053
5	0,0212	0,0201	0,0190	0,0180	0,0170	0,0160	0,0152	0,0143	0,0135	0,0127
6	0,0393	0,0375	0,0358	0,0341	0,0325	0,0310	0,0295	0,0281	0,0268	0,0255
7	0,0623	0,0600	0,0578	0,0556	0,0535	0,0514	0,0494	0,0474	0,0455	0,0437
8	0,0864	0,0840	0,0816	0,0792	0,0769	0,0745	0,0722	0,0700	0,0677	0,0655
9	0,1065	0,1045	0,1024	0,1003	0,0982	0,0961	0,0939	0,0917	0,0895	0,0874
10	0,1182	0,1170	0,1157	0,1144	0,1129	0,1114	0,1099	0,1082	0,1066	0,1048
11	0,1193	0,1192	0,1189	0,1185	0,1181	0,1175	0,1169	0,1161	0,1153	0,1144
12	0,1104	0,1112	0,1120	0,1126	0,1131	0,1136	0,1139	0,1142	0,1143	0,1144
13	0,0942	0,0958	0,0973	0,0987	0,1001	0,1014	0,1025	0,1036	0,1046	0,1056
14	0,0747	0,0767	0,0786	0,0804	0,0822	0,0840	0,0857	0,0874	0,0889	0,0905

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	11,1	11,2	11,3	11,4	11,5	11,6	11,7	11,8	11,9	12,0
15	0,0553	0,0572	0,0592	0,0611	0,0630	0,0649	0,0668	0,0687	0,0706	0,0724
16	0,0384	0,0401	0,0418	0,0435	0,0453	0,0471	0,0489	0,0507	0,0525	0,0543
17	0,0250	0,0264	0,0278	0,0292	0,0306	0,0321	0,0336	0,0352	0,0367	0,0383
18	0,0154	0,0164	0,0174	0,0185	0,0196	0,0207	0,0219	0,0231	0,0243	0,0255
19	0,0090	0,0097	0,0104	0,0111	0,0119	0,0126	0,0135	0,0143	0,0152	0,0161
20	0,0050	0,0054	0,0059	0,0063	0,0068	0,0073	0,0079	0,0084	0,0091	0,0097

	Tasa Media de Llegada λ									
	12,1	12,2	12,3	12,4,	12,5	12,6	12,7	12,8	12,9	13,0
4	0,0050	0,0046	0,0043	0,0041	0,0038	0,0035	0,0033	0,0031	0,0029	0,0027
5	0,0120	0,0113	0,0107	0,0101	0,0095	0,0089	0,0084	0,0079	0,0074	0,0070
6	0,0242	0,0230	0,0219	0,0208	0,0197	0,0187	0,0178	0,0169	0,0160	0,0152
7	0,0419	0,0402	0,0385	0,0368	0,0353	0,0337	0,0323	0,0308	0,0295	0,0281
8	0,0634	0,0612	0,0591	0,0571	0,0551	0,0531	0,0512	0,0493	0,0475	0,0457
9	0,0852	0,0830	0,0808	0,0787	0,0765	0,0744	0,0723	0,0702	0,0681	0,0661
10	0,1031	0,1013	0,0994	0,0975	0,0956	0,0937	0,0918	0,0898	0,0878	0,0859
11	0,1134	0,1123	0,1112	0,1100	0,1087	0,1074	0,1060	0,1045	0,1030	0,1015
12	0,1143	0,1142	0,1139	0,1136	0,1132	0,1127	0,1121	0,1115	0,1107	0,1099
13	0,1064	0,1072	0,1078	0,1084	0,1089	0,1093	0,1096	0,1098	0,1099	0,1099
14	0,0920	0,0934	0,0947	0,0960	0,0972	0,0983	0,0994	0,1004	0,1013	0,1021
15	0,0742	0,0759	0,0777	0,0794	0,0810	0,0826	0,0841	0,0856	0,0871	0,0885
16	0,0561	0,0579	0,0597	0,0615	0,0633	0,0650	0,0668	0,0685	0,0702	0,0719
17	0,0399	0,0416	0,0432	0,0449	0,0465	0,0482	0,0499	0,0516	0,0533	0,0550
18	0,0268	0,0282	0,0295	0,0309	0,0323	0,0337	0,0352	0,0367	0,0382	0,0397
19	0,0171	0,0181	0,0191	0,0202	0,0213	0,0224	0,0235	0,0247	0,0259	0,0272
20	0,0103	0,0110	0,0118	0,0125	0,0133	0,0141	0,0149	0,0158	0,0167	0,0177

	Tasa Media de Llegada λ									
	13,1	13,2	13,3	13,4	13,5	13,6	13,7	13,8	13,9	14,0
5	0,0066	0,0062	0,0058	0,0055	0,0051	0,0048	0,0045	0,0042	0,0040	0,0037
6	0,0144	0,0136	0,0129	0,0122	0,0115	0,0109	0,0103	0,0097	0,0092	0,0087
7	0,0269	0,0256	0,0245	0,0233	0,0222	0,0212	0,0202	0,0192	0,0183	0,0174
8	0,0440	0,0423	0,0407	0,0391	0,0375	0,0360	0,0345	0,0331	0,0318	0,0304
9	0,0640	0,0620	0,0601	0,0582	0,0563	0,0544	0,0526	0,0508	0,0491	0,0473
10	0,0839	0,0819	0,0799	0,0779	0,0760	0,0740	0,0720	0,0701	0,0682	0,0663
11	0,0999	0,0983	0,0966	0,0949	0,0932	0,0915	0,0897	0,0880	0,0862	0,0844
12	0,1091	0,1081	0,1071	0,1060	0,1049	0,1037	0,1024	0,1011	0,0998	0,0984
13	0,1099	0,1098	0,1096	0,1093	0,1089	0,1085	0,1080	0,1074	0,1067	0,1060
14	0,1028	0,1035	0,1041	0,1046	0,1050	0,1054	0,1056	0,1058	0,1060	0,1060
15	0,0898	0,0911	0,0923	0,0934	0,0945	0,0955	0,0965	0,0974	0,0982	0,0989
16	0,0735	0,0751	0,0767	0,0783	0,0798	0,0812	0,0826	0,0840	0,0853	0,0866
17	0,0567	0,0583	0,0600	0,0617	0,0633	0,0650	0,0666	0,0682	0,0697	0,0713
18	0,0412	0,0428	0,0443	0,0459	0,0475	0,0491	0,0507	0,0523	0,0539	0,0554
19	0,0284	0,0297	0,0310	0,0324	0,0337	0,0351	0,0365	0,0380	0,0394	0,0409
20	0,0186	0,0196	0,0206	0,0217	0,0228	0,0239	0,0250	0,0262	0,0274	0,0286

	Tasa Media de Llegada λ									
	14,1	14,2	14,3	14,4	14,5	14,6	14,7	14,8	14,9	15,0
6	0,0082	0,0078	0,0073	0,0069	0,0065	0,0061	0,0058	0,0055	0,0051	0,0048
7	0,0165	0,0157	0,0149	0,0142	0,0135	0,0128	0,0122	0,0115	0,0109	0,0104
8	0,0292	0,0279	0,0267	0,0256	0,0244	0,0234	0,0223	0,0213	0,0204	0,0194
9	0,0457	0,0440	0,0424	0,0409	0,0394	0,0379	0,0365	0,0351	0,0337	0,0324
10	0,0644	0,0625	0,0607	0,0589	0,0571	0,0553	0,0536	0,0519	0,0502	0,0486
11	0,0825	0,0807	0,0789	0,0771	0,0753	0,0735	0,0716	0,0698	0,0681	0,0663

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	14,1	14,2	14,3	14,4	14,5	14,6	14,7	14,7	14,7	15,0
12	0,0970	0,0955	0,0940	0,0925	0,0910	0,0894	0,0878	0,0861	0,0845	0,0829
13	0,1052	0,1043	0,1034	0,1025	0,1014	0,1004	0,0992	0,0981	0,0969	0,0956
14	0,1060	0,1058	0,1057	0,1054	0,1051	0,1047	0,1042	0,1037	0,1031	0,1024
15	0,0996	0,1002	0,1007	0,1012	0,1016	0,1019	0,1021	0,1023	0,1024	0,1024
16	0,0878	0,0889	0,0900	0,0911	0,0920	0,0930	0,0938	0,0946	0,0954	0,0960
17	0,0728	0,0743	0,0757	0,0771	0,0785	0,0798	0,0811	0,0824	0,0836	0,0847
18	0,0570	0,0586	0,0602	0,0617	0,0632	0,0648	0,0663	0,0677	0,0692	0,0706
19	0,0423	0,0438	0,0453	0,0468	0,0483	0,0498	0,0513	0,0528	0,0543	0,0557
20	0,0298	0,0311	0,0324	0,0337	0,0350	0,0363	0,0377	0,0390	0,0404	0,0418
21	0,0200	0,0210	0,0220	0,0231	0,0242	0,0253	0,0264	0,0275	0,0287	0,0299
22	0,0128	0,0136	0,0143	0,0151	0,0159	0,0168	0,0176	0,0185	0,0194	0,0204
23	0,0079	0,0084	0,0089	0,0095	0,0100	0,0106	0,0113	0,0119	0,0126	0,0133
24	0,0046	0,0050	0,0053	0,0057	0,0061	0,0065	0,0069	0,0073	0,0078	0,0083

	Tasa Media de Llegada λ									
	15,1	15,2	15,3	15,4	15,5	15,6	15,7	15,8	15,9	16,0
7	0,0098	0,0093	0,0088	0,0084	0,0079	0,0075	0,0071	0,0067	0,0063	0,0060
8	0,0186	0,0177	0,0169	0,0161	0,0153	0,0146	0,0139	0,0132	0,0126	0,0120
9	0,0311	0,0299	0,0287	0,0275	0,0264	0,0253	0,0243	0,0232	0,0223	0,0213
10	0,0470	0,0454	0,0439	0,0424	0,0409	0,0395	0,0381	0,0367	0,0354	0,0341
11	0,0645	0,0628	0,0611	0,0594	0,0577	0,0560	0,0544	0,0527	0,0512	0,0496
12	0,0812	0,0795	0,0778	0,0762	0,0745	0,0728	0,0711	0,0695	0,0678	0,0661
13	0,0943	0,0930	0,0916	0,0902	0,0888	0,0874	0,0859	0,0844	0,0829	0,0814
14	0,1017	0,1010	0,1001	0,0993	0,0983	0,0974	0,0963	0,0953	0,0942	0,0930
15	0,1024	0,1023	0,1021	0,1019	0,1016	0,1012	0,1008	0,1003	0,0998	0,0992
16	0,0966	0,0972	0,0977	0,0981	0,0984	0,0987	0,0989	0,0991	0,0992	0,0992
17	0,0858	0,0869	0,0879	0,0888	0,0897	0,0906	0,0914	0,0921	0,0928	0,0934
18	0,0720	0,0734	0,0747	0,0760	0,0773	0,0785	0,0797	0,0808	0,0819	0,0830
19	0,0572	0,0587	0,0602	0,0616	0,0630	0,0645	0,0659	0,0672	0,0686	0,0699
20	0,0432	0,0446	0,0460	0,0474	0,0489	0,0503	0,0517	0,0531	0,0545	0,0559
21	0,0311	0,0323	0,0335	0,0348	0,0361	0,0373	0,0386	0,0400	0,0413	0,0426
22	0,0213	0,0223	0,0233	0,0244	0,0254	0,0265	0,0276	0,0287	0,0298	0,0310
23	0,0140	0,0147	0,0155	0,0163	0,0171	0,0180	0,0188	0,0197	0,0206	0,0216
24	0,0088	0,0093	0,0099	0,0105	0,0111	0,0117	0,0123	0,0130	0,0137	0,0144
25	0,0053	0,0057	0,0061	0,0064	0,0069	0,0073	0,0077	0,0082	0,0087	0,0092

	Tasa Media de Llegada λ									
	16,1	16,2	16,3	16,4	16,5	16,6	16,7	16,8	16,9	17,0
7	0,0057	0,0054	0,0051	0,0048	0,0045	0,0043	0,0040	0,0038	0,0036	0,0034
8	0,0114	0,0108	0,0103	0,0098	0,0093	0,0088	0,0084	0,0080	0,0076	0,0072
9	0,0204	0,0195	0,0187	0,0178	0,0171	0,0163	0,0156	0,0149	0,0142	0,0135
10	0,0328	0,0316	0,0304	0,0293	0,0281	0,0270	0,0260	0,0250	0,0240	0,0230
11	0,0481	0,0466	0,0451	0,0436	0,0422	0,0408	0,0394	0,0381	0,0368	0,0355
12	0,0645	0,0628	0,0612	0,0596	0,0580	0,0565	0,0549	0,0534	0,0518	0,0504
13	0,0799	0,0783	0,0768	0,0752	0,0736	0,0721	0,0705	0,0690	0,0674	0,0658
14	0,0918	0,0906	0,0894	0,0881	0,0868	0,0855	0,0841	0,0828	0,0814	0,0800
15	0,0986	0,0979	0,0971	0,0963	0,0955	0,0946	0,0937	0,0927	0,0917	0,0906
16	0,0992	0,0991	0,0989	0,0987	0,0985	0,0981	0,0978	0,0973	0,0968	0,0963
17	0,0939	0,0944	0,0949	0,0952	0,0956	0,0958	0,0960	0,0962	0,0963	0,0963
18	0,0840	0,0850	0,0859	0,0868	0,0876	0,0884	0,0891	0,0898	0,0904	0,0909
19	0,0712	0,0725	0,0737	0,0749	0,0761	0,0772	0,0783	0,0794	0,0804	0,0814
20	0,0573	0,0587	0,0601	0,0614	0,0628	0,0641	0,0654	0,0667	0,0679	0,0692
21	0,0439	0,0453	0,0466	0,0480	0,0493	0,0507	0,0520	0,0533	0,0547	0,0560

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	16,1	16,2	16,3	16,4	16,5	16,6	16,7	16,8	16,9	17,0
22	0,0322	0,0333	0,0345	0,0358	0,0370	0,0382	0,0395	0,0407	0,0420	0,0433
23	0,0225	0,0235	0,0245	0,0255	0,0265	0,0276	0,0287	0,0297	0,0309	0,0320
24	0,0151	0,0159	0,0166	0,0174	0,0182	0,0191	0,0199	0,0208	0,0217	0,0226
25	0,0097	0,0103	0,0108	0,0114	0,0120	0,0127	0,0133	0,0140	0,0147	0,0154

	Tasa Media de Llegada λ									
	17,1	17,2	17,3	17,4	17,5	17,6	17,7	17,8	17,9	18,0
8	0,0068	0,0064	0,0061	0,0058	0,0055	0,0052	0,0049	0,0046	0,0044	0,0042
9	0,0129	0,0123	0,0117	0,0112	0,0107	0,0101	0,0097	0,0092	0,0088	0,0083
10	0,0221	0,0212	0,0203	0,0195	0,0186	0,0179	0,0171	0,0164	0,0157	0,0150
11	0,0343	0,0331	0,0319	0,0308	0,0297	0,0286	0,0275	0,0265	0,0255	0,0245
12	0,0489	0,0474	0,0460	0,0446	0,0432	0,0419	0,0406	0,0393	0,0380	0,0368
13	0,0643	0,0628	0,0612	0,0597	0,0582	0,0567	0,0553	0,0538	0,0524	0,0509
14	0,0785	0,0771	0,0757	0,0742	0,0728	0,0713	0,0699	0,0684	0,0669	0,0655
15	0,0895	0,0884	0,0873	0,0861	0,0849	0,0837	0,0824	0,0812	0,0799	0,0786
16	0,0957	0,0951	0,0944	0,0936	0,0929	0,0920	0,0912	0,0903	0,0894	0,0884
17	0,0963	0,0962	0,0960	0,0958	0,0956	0,0953	0,0949	0,0945	0,0941	0,0936
18	0,0914	0,0919	0,0923	0,0926	0,0929	0,0932	0,0934	0,0935	0,0936	0,0936
19	0,0823	0,0832	0,0840	0,0848	0,0856	0,0863	0,0870	0,0876	0,0882	0,0887
20	0,0704	0,0715	0,0727	0,0738	0,0749	0,0760	0,0770	0,0780	0,0789	0,0798
21	0,0573	0,0586	0,0599	0,0612	0,0624	0,0637	0,0649	0,0661	0,0673	0,0684
22	0,0445	0,0458	0,0471	0,0484	0,0496	0,0509	0,0522	0,0535	0,0547	0,0560
23	0,0331	0,0343	0,0354	0,0366	0,0378	0,0390	0,0402	0,0414	0,0426	0,0438
24	0,0236	0,0246	0,0255	0,0265	0,0275	0,0286	0,0296	0,0307	0,0318	0,0328
25	0,0161	0,0169	0,0177	0,0185	0,0193	0,0201	0,0210	0,0218	0,0227	0,0237

	Tasa Media de Llegada λ									
	18,1	18,2	18,3	18,4	18,5	18,6	18,7	18,8	18,9	19,0
9	0,0079	0,0075	0,0072	0,0068	0,0065	0,0061	0,0058	0,0055	0,0053	0,0050
10	0,0143	0,0137	0,0131	0,0125	0,0120	0,0114	0,0109	0,0104	0,0099	0,0095
11	0,0236	0,0227	0,0218	0,0209	0,0201	0,0193	0,0185	0,0178	0,0171	0,0164
12	0,0356	0,0344	0,0332	0,0321	0,0310	0,0299	0,0289	0,0278	0,0269	0,0259
13	0,0495	0,0481	0,0468	0,0454	0,0441	0,0428	0,0415	0,0403	0,0390	0,0378
14	0,0640	0,0626	0,0611	0,0597	0,0583	0,0569	0,0555	0,0541	0,0527	0,0514
15	0,0773	0,0759	0,0746	0,0732	0,0719	0,0705	0,0692	0,0678	0,0664	0,0650
16	0,0874	0,0864	0,0853	0,0842	0,0831	0,0820	0,0808	0,0796	0,0785	0,0772
17	0,0931	0,0925	0,0918	0,0912	0,0904	0,0897	0,0889	0,0881	0,0872	0,0863
18	0,0936	0,0935	0,0934	0,0932	0,0930	0,0927	0,0924	0,0920	0,0916	0,0911
19	0,0891	0,0896	0,0899	0,0902	0,0905	0,0907	0,0909	0,0910	0,0911	0,0911
20	0,0807	0,0815	0,0823	0,0830	0,0837	0,0844	0,0850	0,0856	0,0861	0,0866
21	0,0695	0,0706	0,0717	0,0727	0,0738	0,0747	0,0757	0,0766	0,0775	0,0783
22	0,0572	0,0584	0,0596	0,0608	0,0620	0,0632	0,0643	0,0655	0,0666	0,0676
23	0,0450	0,0462	0,0475	0,0487	0,0499	0,0511	0,0523	0,0535	0,0547	0,0559
24	0,0340	0,0351	0,0362	0,0373	0,0385	0,0396	0,0408	0,0419	0,0431	0,0442
25	0,0246	0,0255	0,0265	0,0275	0,0285	0,0295	0,0305	0,0315	0,0326	0,0336

	Tasa Media de Llegada λ									
	19,1	19,2	19,3	19,4	19,5	19,6	19,7	19,8	19,9	20,0
10	0,0090	0,0086	0,0082	0,0078	0,0074	0,0071	0,0067	0,0064	0,0061	0,0058
11	0,0157	0,0150	0,0144	0,0138	0,0132	0,0126	0,0121	0,0116	0,0111	0,0106
12	0,0249	0,0240	0,0231	0,0223	0,0214	0,0206	0,0198	0,0191	0,0183	0,0176
13	0,0367	0,0355	0,0344	0,0333	0,0322	0,0311	0,0301	0,0291	0,0281	0,0271
14	0,0500	0,0487	0,0474	0,0461	0,0448	0,0436	0,0423	0,0411	0,0399	0,0387
15	0,0637	0,0623	0,0610	0,0596	0,0582	0,0569	0,0556	0,0543	0,0529	0,0516
16	0,0760	0,0748	0,0735	0,0723	0,0710	0,0697	0,0684	0,0671	0,0659	0,0646

Tabla 5. Probabilidades de Poisson individuales (*continuación*).

	Tasa Media de Llegada λ									
	19,1	19,2	19,3	19,4	19,5	19,6	19,7	19,8	19,9	20,0
17	0,0854	0,0844	0,0835	0,0825	0,0814	0,0804	0,0793	0,0782	0,0771	0,0760
18	0,0906	0,0901	0,0895	0,0889	0,0882	0,0875	0,0868	0,0860	0,0852	0,0844
19	0,0911	0,0910	0,0909	0,0907	0,0905	0,0903	0,0900	0,0896	0,0893	0,0888
20	0,0870	0,0874	0,0877	0,0880	0,0883	0,0885	0,0886	0,0887	0,0888	0,0888
21	0,0791	0,0799	0,0806	0,0813	0,0820	0,0826	0,0831	0,0837	0,0842	0,0846
22	0,0687	0,0697	0,0707	0,0717	0,0727	0,0736	0,0745	0,0753	0,0761	0,0769
23	0,0570	0,0582	0,0594	0,0605	0,0616	0,0627	0,0638	0,0648	0,0659	0,0669
24	0,0454	0,0466	0,0477	0,0489	0,0500	0,0512	0,0523	0,0535	0,0546	0,0557
25	0,0347	0,0358	0,0368	0,0379	0,0390	0,0401	0,0412	0,0424	0,0435	0,0446

	Tasa Media de Llegada λ									
	20,1	20,2	20,3	20,4	20,5	20,6	20,7	20,8	20,9	21,0
10	0,0055	0,0053	0,0050	0,0048	0,0045	0,0043	0,0041	0,0039	0,0037	0,0035
11	0,0101	0,0097	0,0092	0,0088	0,0084	0,0080	0,0077	0,0073	0,0070	0,0067
12	0,0169	0,0163	0,0156	0,0150	0,0144	0,0138	0,0132	0,0127	0,0122	0,0116
13	0,0262	0,0253	0,0244	0,0235	0,0227	0,0219	0,0211	0,0203	0,0195	0,0188
14	0,0376	0,0365	0,0353	0,0343	0,0332	0,0322	0,0311	0,0301	0,0292	0,0282
15	0,0504	0,0491	0,0478	0,0466	0,0454	0,0442	0,0430	0,0418	0,0406	0,0395
16	0,0633	0,0620	0,0607	0,0594	0,0581	0,0569	0,0556	0,0543	0,0531	0,0518
17	0,0748	0,0736	0,0725	0,0713	0,0701	0,0689	0,0677	0,0665	0,0653	0,0640
18	0,0835	0,0826	0,0817	0,0808	0,0798	0,0789	0,0778	0,0768	0,0758	0,0747
19	0,0884	0,0879	0,0873	0,0868	0,0861	0,0855	0,0848	0,0841	0,0834	0,0826
20	0,0888	0,0887	0,0886	0,0885	0,0883	0,0881	0,0878	0,0875	0,0871	0,0867
21	0,0850	0,0854	0,0857	0,0860	0,0862	0,0864	0,0865	0,0866	0,0867	0,0867
22	0,0777	0,0784	0,0791	0,0797	0,0803	0,0809	0,0814	0,0819	0,0824	0,0828
23	0,0679	0,0688	0,0698	0,0707	0,0716	0,0724	0,0733	0,0741	0,0748	0,0756
24	0,0568	0,0579	0,0590	0,0601	0,0611	0,0622	0,0632	0,0642	0,0652	0,0661
25	0,0457	0,0468	0,0479	0,0490	0,0501	0,0512	0,0523	0,0534	0,0545	0,0555

Tabla 6. Probabilidades de Poisson acumuladas.

	Tasa Media de Llegada λ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781	0,8442	0,8088	0,7725	0,7358
2	0,9998	0,9989	0,9964	0,9921	0,9856	0,9769	0,9659	0,9526	0,9371	0,9197
3	1,0000	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9865	0,9810
4	1,0000	1,0000	1,0000	0,9999	0,9998	0,9996	0,9992	0,9986	0,9977	0,9963
5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9998	0,9997	0,9994
6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999
7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

	Tasa Media de Llegada λ									
	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,3329	0,3012	0,2725	0,2466	0,2231	0,2019	0,1827	0,1653	0,1496	0,1353
1	0,6990	0,6626	0,6268	0,5918	0,5578	0,5249	0,4932	0,4628	0,4337	0,4060
2	0,9004	0,8795	0,8571	0,8335	0,8088	0,7834	0,7572	0,7306	0,7037	0,6767
3	0,9743	0,9662	0,9569	0,9463	0,9344	0,9212	0,9068	0,8913	0,8747	0,8571
4	0,9946	0,9923	0,9893	0,9857	0,9814	0,9763	0,9704	0,9636	0,9559	0,9473
5	0,9990	0,9985	0,9978	0,9968	0,9955	0,9940	0,9920	0,9896	0,9868	0,9834
6	0,9999	0,9997	0,9996	0,9994	0,9991	0,9987	0,9981	0,9974	0,9966	0,9955
7	1,0000	1,0000	0,9999	0,9999	0,9998	0,9997	0,9996	0,9994	0,9992	0,9989
8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9998	0,9998
9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

	Tasa Media de Llegada λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,1225	0,1108	0,1003	0,0907	0,0821	0,0743	0,0672	0,0608	0,0550	0,0498
1	0,3796	0,3546	0,3309	0,3084	0,2873	0,2674	0,2487	0,2311	0,2146	0,1991
2	0,6496	0,6227	0,5960	0,5697	0,5438	0,5184	0,4936	0,4695	0,4460	0,4232
3	0,8386	0,8194	0,7993	0,7787	0,7576	0,7360	0,7141	0,6919	0,6696	0,6472
4	0,9379	0,9275	0,9162	0,9041	0,8912	0,8774	0,8629	0,8477	0,8318	0,8153
5	0,9796	0,9751	0,9700	0,9643	0,9580	0,9510	0,9433	0,9349	0,9258	0,9161
6	0,9941	0,9925	0,9906	0,9884	0,9858	0,9828	0,9794	0,9756	0,9713	0,9665
7	0,9985	0,9980	0,9974	0,9967	0,9958	0,9947	0,9934	0,9919	0,9901	0,9881
8	0,9997	0,9995	0,9994	0,9991	0,9989	0,9985	0,9981	0,9976	0,9969	0,9962
9	0,9999	0,9999	0,9999	0,9998	0,9997	0,9996	0,9995	0,9993	0,9991	0,9989
10	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9998	0,9998	0,9997
11	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
12	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

	Tasa Media de Llegada λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,0450	0,0408	0,0369	0,0334	0,0302	0,0273	0,0247	0,0224	0,0202	0,0183
1	0,1847	0,1712	0,1586	0,1468	0,1359	0,1257	0,1162	0,1074	0,0992	0,0916
2	0,4012	0,3799	0,3594	0,3397	0,3208	0,3027	0,2854	0,2689	0,2531	0,2381
3	0,6248	0,6025	0,5803	0,5584	0,5366	0,5152	0,4942	0,4735	0,4532	0,4335
4	0,7982	0,7806	0,7626	0,7442	0,7254	0,7064	0,6872	0,6678	0,6484	0,6288
5	0,9057	0,8946	0,8829	0,8705	0,8576	0,8441	0,8301	0,8156	0,8006	0,7851
6	0,9612	0,9554	0,9490	0,9421	0,9347	0,9267	0,9182	0,9091	0,8995	0,8893
7	0,9858	0,9832	0,9802	0,9769	0,9733	0,9692	0,9648	0,9599	0,9546	0,9489
8	0,9953	0,9943	0,9931	0,9917	0,9901	0,9883	0,9863	0,9840	0,9815	0,9786
9	0,9986	0,9982	0,9978	0,9973	0,9967	0,9960	0,9952	0,9942	0,9931	0,9919
10	0,9996	0,9995	0,9994	0,9992	0,9990	0,9987	0,9984	0,9981	0,9977	0,9972
11	0,9999	0,9999	0,9998	0,9998	0,9997	0,9996	0,9995	0,9994	0,9993	0,9991
12	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9997
13	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999
14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,0166	0,0150	0,0136	0,0123	0,0111	0,0101	0,0091	0,0082	0,0074	0,0067
1	0,0845	0,0780	0,0719	0,0663	0,0611	0,0563	0,0518	0,0477	0,0439	0,0404
2	0,2238	0,2102	0,1974	0,1851	0,1736	0,1626	0,1523	0,1425	0,1333	0,1247
3	0,4142	0,3954	0,3772	0,3594	0,3423	0,3257	0,3097	0,2942	0,2793	0,2650
4	0,6093	0,5898	0,5704	0,5512	0,5321	0,5132	0,4946	0,4763	0,4582	0,4405
5	0,7693	0,7531	0,7367	0,7199	0,7029	0,6858	0,6684	0,6510	0,6335	0,6160
6	0,8786	0,8675	0,8558	0,8436	0,8311	0,8180	0,8046	0,7908	0,7767	0,7622
7	0,9427	0,9361	0,9290	0,9214	0,9134	0,9049	0,8960	0,8867	0,8769	0,8666
8	0,9755	0,9721	0,9683	0,9642	0,9597	0,9549	0,9497	0,9442	0,9382	0,9319
9	0,9905	0,9889	0,9871	0,9851	0,9829	0,9805	0,9778	0,9749	0,9717	0,9682
10	0,9966	0,9959	0,9952	0,9943	0,9933	0,9922	0,9910	0,9896	0,9880	0,9863
11	0,9989	0,9986	0,9983	0,9980	0,9976	0,9971	0,9966	0,9960	0,9953	0,9945
12	0,9997	0,9996	0,9995	0,9993	0,9992	0,9990	0,9988	0,9986	0,9983	0,9980
13	0,9999	0,9999	0,9998	0,9998	0,9997	0,9997	0,9996	0,9995	0,9994	0,9993
14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

	Tasa Media de Llegada λ									
	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8	5,9	6,0
0	0,0061	0,0055	0,0050	0,0045	0,0041	0,0037	0,0033	0,0030	0,0027	0,0025
1	0,0372	0,0342	0,0314	0,0289	0,0266	0,0244	0,0224	0,0206	0,0189	0,0174
2	0,1165	0,1088	0,1016	0,0948	0,0884	0,0824	0,0768	0,0715	0,0666	0,0620
3	0,2513	0,2381	0,2254	0,2133	0,2017	0,1906	0,1800	0,1700	0,1604	0,1512
4	0,4231	0,4061	0,3895	0,3733	0,3575	0,3422	0,3272	0,3127	0,2987	0,2851
5	0,5984	0,5809	0,5635	0,5461	0,5289	0,5119	0,4950	0,4783	0,4619	0,4457
6	0,7474	0,7324	0,7171	0,7017	0,6860	0,6703	0,6544	0,6384	0,6224	0,6063
7	0,8560	0,8449	0,8335	0,8217	0,8095	0,7970	0,7841	0,7710	0,7576	0,7440
8	0,9252	0,9181	0,9106	0,9027	0,8944	0,8857	0,8766	0,8672	0,8574	0,8472
9	0,9644	0,9603	0,9559	0,9512	0,9462	0,9409	0,9352	0,9292	0,9228	0,9161
10	0,9844	0,9823	0,9800	0,9775	0,9747	0,9718	0,9686	0,9651	0,9614	0,9574
11	0,9937	0,9927	0,9916	0,9904	0,9890	0,9875	0,9859	0,9841	0,9821	0,9799
12	0,9976	0,9972	0,9967	0,9962	0,9955	0,9949	0,9941	0,9932	0,9922	0,9912
13	0,9992	0,9990	0,9988	0,9986	0,9983	0,9980	0,9977	0,9973	0,9969	0,9964
14	0,9997	0,9997	0,9996	0,9995	0,9994	0,9993	0,9991	0,9990	0,9988	0,9986

	Tasa Media de Llegada λ									
	6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8	6,9	7,0
0	0,0022	0,0020	0,0018	0,0017	0,0015	0,0014	0,0012	0,0011	0,0010	0,0009
1	0,0159	0,0146	0,0134	0,0123	0,0113	0,0103	0,0095	0,0087	0,0080	0,0073
2	0,0577	0,0536	0,0498	0,0463	0,0430	0,0400	0,0371	0,0344	0,0320	0,0296
3	0,1425	0,1342	0,1264	0,1189	0,1118	0,1052	0,0988	0,0928	0,0871	0,0818
4	0,2719	0,2592	0,2469	0,2351	0,2237	0,2127	0,2022	0,1920	0,1823	0,1730
5	0,4298	0,4141	0,3988	0,3837	0,3690	0,3547	0,3406	0,3270	0,3137	0,3007
6	0,5902	0,5742	0,5582	0,5423	0,5265	0,5108	0,4953	0,4799	0,4647	0,4497
7	0,7301	0,7160	0,7017	0,6873	0,6728	0,6581	0,6433	0,6285	0,6136	0,5987
8	0,8367	0,8259	0,8148	0,8033	0,7916	0,7796	0,7673	0,7548	0,7420	0,7291
9	0,9090	0,9016	0,8939	0,8858	0,8774	0,8686	0,8596	0,8502	0,8405	0,8305
10	0,9531	0,9486	0,9437	0,9386	0,9332	0,9274	0,9214	0,9151	0,9084	0,9015
11	0,9776	0,9750	0,9723	0,9693	0,9661	0,9627	0,9591	0,9552	0,9510	0,9467
12	0,9900	0,9887	0,9873	0,9857	0,9840	0,9821	0,9801	0,9779	0,9755	0,9730
13	0,9958	0,9952	0,9945	0,9937	0,9929	0,9920	0,9909	0,9898	0,9885	0,9872
14	0,9984	0,9981	0,9978	0,9974	0,9970	0,9966	0,9961	0,9956	0,9950	0,9943

Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8	7,9	8,0
0	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005	0,0005	0,0004	0,0004	0,0003
1	0,0067	0,0061	0,0056	0,0051	0,0047	0,0043	0,0039	0,0036	0,0033	0,0030
2	0,0275	0,0255	0,0236	0,0219	0,0203	0,0188	0,0174	0,0161	0,0149	0,0138
3	0,0767	0,0719	0,0674	0,0632	0,0591	0,0554	0,0518	0,0485	0,0453	0,0424
4	0,1641	0,1555	0,1473	0,1395	0,1321	0,1249	0,1181	0,1117	0,1055	0,0996
5	0,2881	0,2759	0,2640	0,2526	0,2414	0,2307	0,2203	0,2103	0,2006	0,1912
6	0,4349	0,4204	0,4060	0,3920	0,3782	0,3646	0,3514	0,3384	0,3257	0,3134
7	0,5838	0,5689	0,5541	0,5393	0,5246	0,5100	0,4956	0,4812	0,4670	0,4530
8	0,7160	0,7027	0,6892	0,6757	0,6620	0,6482	0,6343	0,6204	0,6065	0,5925
9	0,8202	0,8096	0,7988	0,7877	0,7764	0,7649	0,7531	0,7411	0,7290	0,7166
10	0,8942	0,8867	0,8788	0,8707	0,8622	0,8535	0,8445	0,8352	0,8257	0,8159
11	0,9420	0,9371	0,9319	0,9265	0,9208	0,9148	0,9085	0,9020	0,8952	0,8881
12	0,9703	0,9673	0,9642	0,9609	0,9573	0,9536	0,9496	0,9454	0,9409	0,9362
13	0,9857	0,9841	0,9824	0,9805	0,9784	0,9762	0,9739	0,9714	0,9687	0,9658
14	0,9935	0,9927	0,9918	0,9908	0,9897	0,9886	0,9873	0,9859	0,9844	0,9827
15	0,9972	0,9969	0,9964	0,9959	0,9954	0,9948	0,9941	0,9934	0,9926	0,9918
16	0,9989	0,9987	0,9985	0,9983	0,9980	0,9978	0,9974	0,9971	0,9967	0,9963
17	0,9996	0,9995	0,9994	0,9993	0,9992	0,9991	0,9989	0,9988	0,9986	0,9984
18	0,9998	0,9998	0,9998	0,9997	0,9997	0,9996	0,9996	0,9995	0,9994	0,9993
19	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9998	0,9998	0,9998	0,9997
20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9999	0,9999	0,9999	0,9999

	Tasa Media de Llegada λ									
	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8	8,9	9,0
0	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001	0,0001
1	0,0028	0,0025	0,0023	0,0021	0,0019	0,0018	0,0016	0,0015	0,0014	0,0012
2	0,0127	0,0118	0,0109	0,0100	0,0093	0,0086	0,0079	0,0073	0,0068	0,0062
3	0,0396	0,0370	0,0346	0,0323	0,0301	0,0281	0,0262	0,0244	0,0228	0,0212
4	0,0940	0,0887	0,0837	0,0789	0,0744	0,0701	0,0660	0,0621	0,0584	0,0550
5	0,1822	0,1736	0,1653	0,1573	0,1496	0,1422	0,1352	0,1284	0,1219	0,1157
6	0,3013	0,2896	0,2781	0,2670	0,2562	0,2457	0,2355	0,2256	0,2160	0,2068
7	0,4391	0,4254	0,4119	0,3987	0,3856	0,3728	0,3602	0,3478	0,3357	0,3239
8	0,5786	0,5647	0,5507	0,5369	0,5231	0,5094	0,4958	0,4823	0,4689	0,4557
9	0,7041	0,6915	0,6788	0,6659	0,6530	0,6400	0,6269	0,6137	0,6006	0,5874
10	0,8058	0,7955	0,7850	0,7743	0,7634	0,7522	0,7409	0,7294	0,7178	0,7060
11	0,8807	0,8731	0,8652	0,8571	0,8487	0,8400	0,8311	0,8220	0,8126	0,8030
12	0,9313	0,9261	0,9207	0,9150	0,9091	0,9029	0,8965	0,8898	0,8829	0,8758
13	0,9628	0,9595	0,9561	0,9524	0,9486	0,9445	0,9403	0,9358	0,9311	0,9261
14	0,9810	0,9791	0,9771	0,9749	0,9726	0,9701	0,9675	0,9647	0,9617	0,9585
15	0,9908	0,9898	0,9887	0,9875	0,9862	0,9848	0,9832	0,9816	0,9798	0,9780
16	0,9958	0,9953	0,9947	0,9941	0,9934	0,9926	0,9918	0,9909	0,9899	0,9889
17	0,9982	0,9979	0,9977	0,9973	0,9970	0,9966	0,9962	0,9957	0,9952	0,9947
18	0,9992	0,9991	0,9990	0,9989	0,9987	0,9985	0,9983	0,9981	0,9978	0,9976
19	0,9997	0,9997	0,9996	0,9995	0,9995	0,9994	0,9993	0,9992	0,9991	0,9989
20	0,9999	0,9999	0,9998	0,9998	0,9998	0,9998	0,9997	0,9997	0,9996	0,9996

	Tasa Media de Llegada λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
0	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0000
1	0,0011	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0005	0,0005
2	0,0058	0,0053	0,0049	0,0045	0,0042	0,0038	0,0035	0,0033	0,0030	0,0028
3	0,0198	0,0184	0,0172	0,0160	0,0149	0,0138	0,0129	0,0120	0,0111	0,0103
4	0,0517	0,0486	0,0456	0,0429	0,0403	0,0378	0,0355	0,0333	0,0312	0,0293
5	0,1098	0,1041	0,0986	0,0935	0,0885	0,0838	0,0793	0,0750	0,0710	0,0671
6	0,1978	0,1892	0,1808	0,1727	0,1649	0,1574	0,1502	0,1433	0,1366	0,1301

Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9	10,0
7	0,3123	0,3010	0,2900	0,2792	0,2687	0,2584	0,2485	0,2388	0,2294	0,2202
8	0,4426	0,4296	0,4168	0,4042	0,3918	0,3796	0,3676	0,3558	0,3442	0,3328
9	0,5742	0,5611	0,5479	0,5349	0,5218	0,5089	0,4960	0,4832	0,4705	0,4579
10	0,6941	0,6820	0,6699	0,6576	0,6453	0,6329	0,6205	0,6080	0,5955	0,5830
11	0,7932	0,7832	0,7730	0,7626	0,7520	0,7412	0,7303	0,7193	0,7081	0,6968
12	0,8684	0,8607	0,8529	0,8448	0,8364	0,8279	0,8191	0,8101	0,8009	0,7916
13	0,9210	0,9156	0,9100	0,9042	0,8981	0,8919	0,8853	0,8786	0,8716	0,8645
14	0,9552	0,9517	0,9480	0,9441	0,9400	0,9357	0,9312	0,9265	0,9216	0,9165
15	0,9760	0,9738	0,9715	0,9691	0,9665	0,9638	0,9609	0,9579	0,9546	0,9513
16	0,9878	0,9865	0,9852	0,9838	0,9823	0,9806	0,9789	0,9770	0,9751	0,9730
17	0,9941	0,9934	0,9927	0,9919	0,9911	0,9902	0,9892	0,9881	0,9870	0,9857
18	0,9973	0,9969	0,9966	0,9962	0,9957	0,9952	0,9947	0,9941	0,9935	0,9928
19	0,9988	0,9986	0,9985	0,9983	0,9980	0,9978	0,9975	0,9972	0,9969	0,9965
20	0,9995	0,9994	0,9993	0,9992	0,9991	0,9990	0,9989	0,9987	0,9986	0,9984

	Tasa Media de Llegada λ									
	10,1	10,2	10,3	10,4	10,5	10,6	10,7	10,8	10,9	11,0
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0005	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002
2	0,0026	0,0023	0,0022	0,0020	0,0018	0,0017	0,0016	0,0014	0,0013	0,0012
3	0,0096	0,0089	0,0083	0,0077	0,0071	0,0066	0,0062	0,0057	0,0053	0,0049
4	0,0274	0,0257	0,0241	0,0225	0,0211	0,0197	0,0185	0,0173	0,0162	0,0151
5	0,0634	0,0599	0,0566	0,0534	0,0504	0,0475	0,0448	0,0423	0,0398	0,0375
6	0,1240	0,1180	0,1123	0,1069	0,1016	0,0966	0,0918	0,0872	0,0828	0,0786
7	0,2113	0,2027	0,1944	0,1863	0,1785	0,1710	0,1636	0,1566	0,1498	0,1432
8	0,3217	0,3108	0,3001	0,2896	0,2794	0,2694	0,2597	0,2502	0,2410	0,2320
9	0,4455	0,4332	0,4210	0,4090	0,3971	0,3854	0,3739	0,3626	0,3515	0,3405
10	0,5705	0,5580	0,5456	0,5331	0,5207	0,5084	0,4961	0,4840	0,4719	0,4599
11	0,6853	0,6738	0,6622	0,6505	0,6387	0,6269	0,6150	0,6031	0,5912	0,5793
12	0,7820	0,7722	0,7623	0,7522	0,7420	0,7316	0,7210	0,7104	0,6996	0,6887
13	0,8571	0,8494	0,8416	0,8336	0,8253	0,8169	0,8083	0,7995	0,7905	0,7813
14	0,9112	0,9057	0,9	0,8940	0,8879	0,8815	0,8750	0,8682	0,8612	0,8540
15	0,9477	0,9440	0,9400	0,9359	0,9317	0,9272	0,9225	0,9177	0,9126	0,9074
16	0,9707	0,9684	0,9658	0,9632	0,9604	0,9574	0,9543	0,9511	0,9477	0,9441
17	0,9844	0,9830	0,9815	0,9799	0,9781	0,9763	0,9744	0,9723	0,9701	0,9678
18	0,9921	0,9913	0,9904	0,9895	0,9885	0,9874	0,9863	0,9850	0,9837	0,9823
19	0,9962	0,9957	0,9953	0,9948	0,9942	0,9936	0,9930	0,9923	0,9915	0,9907
20	0,9982	0,9980	0,9978	0,9975	0,9972	0,9969	0,9966	0,9962	0,9958	0,9953

	Tasa Media de Llegada λ									
	11,1	11,2	11,3	11,4	11,5	11,6	11,7	11,8	11,9	12,0
0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
1	0,0002	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
2	0,0011	0,0010	0,0009	0,0009	0,0008	0,0007	0,0007	0,0006	0,0006	0,0005
3	0,0046	0,0042	0,0039	0,0036	0,0034	0,0031	0,0029	0,0027	0,0025	0,0023
4	0,0141	0,0132	0,0123	0,0115	0,0107	0,0100	0,0094	0,0087	0,0081	0,0076
5	0,0353	0,0333	0,0313	0,0295	0,0277	0,0261	0,0245	0,0230	0,0217	0,0203
6	0,0746	0,0708	0,0671	0,0636	0,0603	0,0571	0,0541	0,0512	0,0484	0,0458
7	0,1369	0,1307	0,1249	0,1192	0,1137	0,1085	0,1035	0,0986	0,0940	0,0895
8	0,2232	0,2147	0,2064	0,1984	0,1906	0,1830	0,1757	0,1686	0,1617	0,1550
9	0,3298	0,3192	0,3089	0,2987	0,2888	0,2791	0,2696	0,2603	0,2512	0,2424
10	0,4480	0,4362	0,4246	0,4131	0,4017	0,3905	0,3794	0,3685	0,3578	0,3472
11	0,5673	0,5554	0,5435	0,5316	0,5198	0,5080	0,4963	0,4847	0,4731	0,4616
12	0,6777	0,6666	0,6555	0,6442	0,6329	0,6216	0,6102	0,5988	0,5874	0,5760
13	0,7719	0,7624	0,7528	0,7430	0,7330	0,7230	0,7128	0,7025	0,6920	0,6815
14	0,8467	0,8391	0,8313	0,8234	0,8153	0,8069	0,7985	0,7898	0,7810	0,7720

Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	11,1	11,2	11,3	11,4	11,5	11,6	11,7	11,8	11,9	12,0
15	0,9020	0,8963	0,8905	0,8845	0,8783	0,8719	0,8653	0,8585	0,8516	0,8444
16	0,9403	0,9364	0,9323	0,9280	0,9236	0,9190	0,9142	0,9092	0,9040	0,8987
17	0,9654	0,9628	0,9601	0,9572	0,9542	0,9511	0,9478	0,9444	0,9408	0,9370
18	0,9808	0,9792	0,9775	0,9757	0,9738	0,9718	0,9697	0,9674	0,9651	0,9626
19	0,9898	0,9889	0,9879	0,9868	0,9857	0,9845	0,9832	0,9818	0,9803	0,9787
20	0,9948	0,9943	0,9938	0,9932	0,9925	0,9918	0,9910	0,9902	0,9893	0,9884
	Tasa Media de Llegada λ									
	12,1	12,2	12,3	12,4	12,5	12,6	12,7	12,8	12,9	13,0
5	0,0191	0,0179	0,0168	0,0158	0,0148	0,0139	0,0130	0,0122	0,0115	0,0107
6	0,0433	0,0410	0,0387	0,0366	0,0346	0,0326	0,0308	0,0291	0,0274	0,0259
7	0,0852	0,0811	0,0772	0,0734	0,0698	0,0664	0,0631	0,0599	0,0569	0,0540
8	0,1486	0,1424	0,1363	0,1305	0,1249	0,1195	0,1143	0,1093	0,1044	0,0998
9	0,2338	0,2254	0,2172	0,2092	0,2014	0,1939	0,1866	0,1794	0,1725	0,1658
10	0,3368	0,3266	0,3166	0,3067	0,2971	0,2876	0,2783	0,2693	0,2604	0,2517
11	0,4502	0,4389	0,4278	0,4167	0,4058	0,3950	0,3843	0,3738	0,3634	0,3532
12	0,5645	0,5531	0,5417	0,5303	0,5190	0,5077	0,4964	0,4853	0,4741	0,4631
13	0,6709	0,6603	0,6495	0,6387	0,6278	0,6169	0,6060	0,5950	0,5840	0,5730
14	0,7629	0,7536	0,7442	0,7347	0,7250	0,7153	0,7054	0,6954	0,6853	0,6751
15	0,8371	0,8296	0,8219	0,8140	0,8060	0,7978	0,7895	0,7810	0,7724	0,7636
16	0,8932	0,8875	0,8816	0,8755	0,8693	0,8629	0,8563	0,8495	0,8426	0,8355
17	0,9331	0,9290	0,9248	0,9204	0,9158	0,9111	0,9062	0,9011	0,8959	0,8905
18	0,9600	0,9572	0,9543	0,9513	0,9481	0,9448	0,9414	0,9378	0,9341	0,9302
19	0,9771	0,9753	0,9734	0,9715	0,9694	0,9672	0,9649	0,9625	0,9600	0,9573
20	0,9874	0,9863	0,9852	0,9840	0,9827	0,9813	0,9799	0,9783	0,9767	0,9750
21	0,9934	0,9927	0,9921	0,9914	0,9906	0,9898	0,9889	0,9880	0,9870	0,9859
22	0,9966	0,9963	0,9959	0,9955	0,9951	0,9946	0,9941	0,9936	0,9930	0,9924
23	0,9984	0,9982	0,9980	0,9978	0,9975	0,9973	0,9970	0,9967	0,9964	0,9960
	Tasa Media de Llegada λ									
	13,1	13,2	13,3	13,4	13,5	13,6	13,7	13,8	13,9	14,0
5	0,0101	0,0094	0,0088	0,0083	0,0077	0,0072	0,0068	0,0063	0,0059	0,0055
6	0,0244	0,0230	0,0217	0,0204	0,0193	0,0181	0,0171	0,0161	0,0151	0,0142
7	0,0513	0,0487	0,0461	0,0438	0,0415	0,0393	0,0372	0,0353	0,0334	0,0316
8	0,0953	0,0910	0,0868	0,0828	0,0790	0,0753	0,0718	0,0684	0,0652	0,0621
9	0,1593	0,1530	0,1469	0,1410	0,1353	0,1297	0,1244	0,1192	0,1142	0,1094
10	0,2432	0,2349	0,2268	0,2189	0,2112	0,2037	0,1964	0,1893	0,1824	0,1757
11	0,3431	0,3332	0,3234	0,3139	0,3045	0,2952	0,2862	0,2773	0,2686	0,2600
12	0,4522	0,4413	0,4305	0,4199	0,4093	0,3989	0,3886	0,3784	0,3684	0,3585
13	0,5621	0,5511	0,5401	0,5292	0,5182	0,5074	0,4966	0,4858	0,4751	0,4644
14	0,6649	0,6546	0,6442	0,6338	0,6233	0,6128	0,6022	0,5916	0,5810	0,5704
15	0,7547	0,7456	0,7365	0,7272	0,7178	0,7083	0,6987	0,6890	0,6792	0,6694
16	0,8282	0,8208	0,8132	0,8054	0,7975	0,7895	0,7813	0,7730	0,7645	0,7559
17	0,8849	0,8791	0,8732	0,8671	0,8609	0,8545	0,8479	0,8411	0,8343	0,8272
18	0,9261	0,9219	0,9176	0,9130	0,9084	0,9035	0,8986	0,8934	0,8881	0,8826
19	0,9546	0,9516	0,9486	0,9454	0,9421	0,9387	0,9351	0,9314	0,9275	0,9235
20	0,9732	0,9713	0,9692	0,9671	0,9649	0,9626	0,9601	0,9576	0,9549	0,9521
21	0,9848	0,9836	0,9823	0,9810	0,9796	0,9780	0,9765	0,9748	0,9730	0,9712
22	0,9917	0,9910	0,9902	0,9894	0,9885	0,9876	0,9866	0,9856	0,9845	0,9833
23	0,9956	0,9952	0,9948	0,9943	0,9938	0,9933	0,9927	0,9921	0,9914	0,9907
	Tasa Media de Llegada λ									
	14,1	14,2	14,3	14,4	14,5	14,6	14,7	14,8	14,9	15,0
6	0,0134	0,0126	0,0118	0,0111	0,0105	0,0098	0,0092	0,0087	0,0081	0,0076
7	0,0299	0,0283	0,0268	0,0253	0,0239	0,0226	0,0214	0,0202	0,0191	0,0180
8	0,0591	0,0562	0,0535	0,0509	0,0484	0,0460	0,0437	0,0415	0,0394	0,0374

Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	14,1	14,2	14,3	14,4	14,5	14,6	14,7	14,8	14,9	15,0
9	0,1047	0,1003	0,0959	0,0918	0,0878	0,0839	0,0802	0,0766	0,0732	0,0699
10	0,1691	0,1628	0,1566	0,1507	0,1449	0,1392	0,1338	0,1285	0,1234	0,1185
11	0,2517	0,2435	0,2355	0,2277	0,2201	0,2127	0,2054	0,1984	0,1915	0,1848
12	0,3487	0,3391	0,3296	0,3203	0,3111	0,3021	0,2932	0,2845	0,2760	0,2676
13	0,4539	0,4434	0,4330	0,4227	0,4125	0,4024	0,3925	0,3826	0,3728	0,3632
14	0,5598	0,5492	0,5387	0,5281	0,5176	0,5071	0,4967	0,4863	0,4759	0,4657
15	0,6594	0,6494	0,6394	0,6293	0,6192	0,6090	0,5988	0,5886	0,5783	0,5681
16	0,7472	0,7384	0,7294	0,7204	0,7112	0,7020	0,6926	0,6832	0,6737	0,6641
17	0,8200	0,8126	0,8051	0,7975	0,7897	0,7818	0,7737	0,7656	0,7573	0,7489
18	0,8770	0,8712	0,8653	0,8592	0,8530	0,8466	0,8400	0,8333	0,8265	0,8195
19	0,9193	0,9150	0,9106	0,9060	0,9012	0,8963	0,8913	0,8861	0,8807	0,8752
20	0,9492	0,9461	0,9430	0,9396	0,9362	0,9326	0,9289	0,9251	0,9211	0,9170
21	0,9692	0,9671	0,9650	0,9627	0,9604	0,9579	0,9553	0,9526	0,9498	0,9469
22	0,9820	0,9807	0,9793	0,9779	0,9763	0,9747	0,9729	0,9711	0,9692	0,9673
23	0,9899	0,9891	0,9882	0,9873	0,9863	0,9853	0,9842	0,9831	0,9818	0,9805
24	0,9945	0,9941	0,9935	0,9930	0,9924	0,9918	0,9911	0,9904	0,9896	0,9888
25	0,9971	0,9969	0,9966	0,9963	0,9959	0,9956	0,9952	0,9947	0,9943	0,9938

	Tasa Media de Llegada λ									
	15,1	15,2	15,3	15,4	15,5	15,6	15,7	15,8	15,9	16,0
7	0,0170	0,0160	0,0151	0,0143	0,0135	0,0127	0,0120	0,0113	0,0106	0,0100
8	0,0355	0,0337	0,0320	0,0304	0,0288	0,0273	0,0259	0,0245	0,0232	0,0220
9	0,0667	0,0636	0,0607	0,0579	0,0552	0,0526	0,0501	0,0478	0,0455	0,0433
10	0,1137	0,1091	0,1046	0,1003	0,0961	0,0921	0,0882	0,0845	0,0809	0,0774
11	0,1782	0,1718	0,1657	0,1596	0,1538	0,1481	0,1426	0,1372	0,1320	0,1270
12	0,2594	0,2514	0,2435	0,2358	0,2283	0,2209	0,2137	0,2067	0,1998	0,1931
13	0,3537	0,3444	0,3351	0,3260	0,3171	0,3083	0,2996	0,2911	0,2827	0,2745
14	0,4554	0,4453	0,4353	0,4253	0,4154	0,4056	0,3959	0,3864	0,3769	0,3675
15	0,5578	0,5476	0,5374	0,5272	0,5170	0,5069	0,4968	0,4867	0,4767	0,4667
16	0,6545	0,6448	0,6351	0,6253	0,6154	0,6056	0,5957	0,5858	0,5759	0,5660
17	0,7403	0,7317	0,7230	0,7141	0,7052	0,6962	0,6871	0,6779	0,6687	0,6593
18	0,8123	0,8051	0,7977	0,7901	0,7825	0,7747	0,7668	0,7587	0,7506	0,7423
19	0,8696	0,8638	0,8578	0,8517	0,8455	0,8391	0,8326	0,8260	0,8192	0,8122
20	0,9128	0,9084	0,9039	0,8992	0,8944	0,8894	0,8843	0,8791	0,8737	0,8682
21	0,9438	0,9407	0,9374	0,9340	0,9304	0,9268	0,9230	0,9190	0,9150	0,9108
22	0,9652	0,9630	0,9607	0,9583	0,9558	0,9532	0,9505	0,9477	0,9448	0,9418
23	0,9792	0,9777	0,9762	0,9746	0,9730	0,9712	0,9694	0,9674	0,9654	0,9633
24	0,9880	0,9871	0,9861	0,9851	0,9840	0,9829	0,9817	0,9804	0,9791	0,9777
25	0,9933	0,9928	0,9922	0,9915	0,9909	0,9902	0,9894	0,9886	0,9878	0,9869

	Tasa Media de Llegada λ									
	16,1	16,2	16,3	16,4	16,5	16,6	16,7	16,8	16,9	17,0
8	0,0208	0,0197	0,0186	0,0176	0,0167	0,0158	0,0149	0,0141	0,0133	0,0126
9	0,0412	0,0392	0,0373	0,0355	0,0337	0,0321	0,0305	0,0290	0,0275	0,0261
10	0,0740	0,0708	0,0677	0,0647	0,0619	0,0591	0,0565	0,0539	0,0515	0,0491
11	0,1221	0,1174	0,1128	0,1084	0,1041	0,0999	0,0959	0,0920	0,0883	0,0847
12	0,1866	0,1802	0,1740	0,1680	0,1621	0,1564	0,1508	0,1454	0,1401	0,1350
13	0,2664	0,2585	0,2508	0,2432	0,2357	0,2285	0,2213	0,2144	0,2075	0,2009
14	0,3583	0,3492	0,3402	0,3313	0,3225	0,3139	0,3054	0,2971	0,2889	0,2808
15	0,4569	0,4470	0,4373	0,4276	0,4180	0,4085	0,3991	0,3898	0,3806	0,3715
16	0,5560	0,5461	0,5362	0,5263	0,5165	0,5067	0,4969	0,4871	0,4774	0,4677
17	0,6500	0,6406	0,6311	0,6216	0,6120	0,6025	0,5929	0,5833	0,5737	0,5640
18	0,7340	0,7255	0,7170	0,7084	0,6996	0,6908	0,6820	0,6730	0,6640	0,6550
19	0,8052	0,7980	0,7907	0,7833	0,7757	0,7681	0,7603	0,7524	0,7444	0,7363
20	0,8625	0,8567	0,8508	0,8447	0,8385	0,8321	0,8257	0,8191	0,8123	0,8055
21	0,9064	0,9020	0,8974	0,8927	0,8878	0,8828	0,8777	0,8724	0,8670	0,8615
22	0,9386	0,9353	0,9319	0,9284	0,9248	0,9210	0,9171	0,9131	0,9090	0,9047

Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	16,1	16,2	16,3	16,4	16,5	16,6	16,7	16,8	16,9	17,0
23	0,9611	0,9588	0,9564	0,9539	0,9513	0,9486	0,9458	0,9429	0,9398	0,9367
24	0,9762	0,9747	0,9730	0,9713	0,9696	0,9677	0,9657	0,9637	0,9616	0,9594
25	0,9859	0,9849	0,9839	0,9828	0,9816	0,9804	0,9791	0,9777	0,9763	0,9748
26	0,9920	0,9913	0,9907	0,9900	0,9892	0,9884	0,9876	0,9867	0,9858	0,9848

	Tasa Media de Llegada λ									
	17,1	17,2	17,3	17,4	17,5	17,6	17,7	17,8	17,9	18,0
8	0,0119	0,0112	0,0106	0,0100	0,0095	0,0089	0,0084	0,0079	0,0075	0,0071
9	0,0248	0,0235	0,0223	0,0212	0,0201	0,0191	0,0181	0,0171	0,0162	0,0154
10	0,0469	0,0447	0,0426	0,0406	0,0387	0,0369	0,0352	0,0335	0,0319	0,0304
11	0,0812	0,0778	0,0746	0,0714	0,0684	0,0655	0,0627	0,0600	0,0574	0,0549
12	0,1301	0,1252	0,1206	0,1160	0,1116	0,1074	0,1033	0,0993	0,0954	0,0917
13	0,1944	0,1880	0,1818	0,1758	0,1699	0,1641	0,1585	0,1531	0,1478	0,1426
14	0,2729	0,2651	0,2575	0,2500	0,2426	0,2354	0,2284	0,2215	0,2147	0,2081
15	0,3624	0,3535	0,3448	0,3361	0,3275	0,3191	0,3108	0,3026	0,2946	0,2867
16	0,4581	0,4486	0,4391	0,4297	0,4204	0,4112	0,4020	0,3929	0,3839	0,3751
17	0,5544	0,5448	0,5352	0,5256	0,5160	0,5065	0,4969	0,4875	0,4780	0,4686
18	0,6458	0,6367	0,6275	0,6182	0,6089	0,5996	0,5903	0,5810	0,5716	0,5622
19	0,7281	0,7199	0,7115	0,7031	0,6945	0,6859	0,6773	0,6685	0,6598	0,6509
20	0,7985	0,7914	0,7842	0,7769	0,7694	0,7619	0,7542	0,7465	0,7387	0,7307
21	0,8558	0,8500	0,8441	0,8380	0,8319	0,8255	0,8191	0,8126	0,8059	0,7991
22	0,9003	0,8958	0,8912	0,8864	0,8815	0,8765	0,8713	0,8660	0,8606	0,8551
23	0,9334	0,9301	0,9266	0,9230	0,9193	0,9154	0,9115	0,9074	0,9032	0,8989
24	0,9570	0,9546	0,9521	0,9495	0,9468	0,9440	0,9411	0,9381	0,9350	0,9317
25	0,9732	0,9715	0,9698	0,9680	0,9661	0,9641	0,9621	0,9599	0,9577	0,9554
26	0,9838	0,9827	0,9816	0,9804	0,9791	0,9778	0,9764	0,9749	0,9734	0,9718
27	0,9905	0,9898	0,9891	0,9883	0,9875	0,9866	0,9857	0,9848	0,9837	0,9827

	Tasa Media de Llegada λ									
	18,1	18,2	18,3	18,4	18,5	18,6	18,7	18,8	18,9	19,0
9	0,0146	0,0138	0,0131	0,0124	0,0117	0,0111	0,0105	0,0099	0,0094	0,0089
10	0,0289	0,0275	0,0262	0,0249	0,0237	0,0225	0,0214	0,0203	0,0193	0,0183
11	0,0525	0,0502	0,0479	0,0458	0,0438	0,0418	0,0399	0,0381	0,0363	0,0347
12	0,0881	0,0846	0,0812	0,0779	0,0748	0,0717	0,0688	0,0659	0,0632	0,0606
13	0,1376	0,1327	0,1279	0,1233	0,1189	0,1145	0,1103	0,1062	0,1022	0,0984
14	0,2016	0,1953	0,1891	0,1830	0,1771	0,1714	0,1658	0,1603	0,1550	0,1497
15	0,2789	0,2712	0,2637	0,2563	0,2490	0,2419	0,2349	0,2281	0,2214	0,2148
16	0,3663	0,3576	0,3490	0,3405	0,3321	0,3239	0,3157	0,3077	0,2998	0,2920
17	0,4593	0,4500	0,4408	0,4317	0,4226	0,4136	0,4047	0,3958	0,3870	0,3784
18	0,5529	0,5435	0,5342	0,5249	0,5156	0,5063	0,4970	0,4878	0,4786	0,4695
19	0,6420	0,6331	0,6241	0,6151	0,6061	0,5970	0,5879	0,5788	0,5697	0,5606
20	0,7227	0,7146	0,7064	0,6981	0,6898	0,6814	0,6729	0,6644	0,6558	0,6472
21	0,7922	0,7852	0,7781	0,7709	0,7636	0,7561	0,7486	0,7410	0,7333	0,7255
22	0,8494	0,8436	0,8377	0,8317	0,8256	0,8193	0,8129	0,8065	0,7998	0,7931
23	0,8944	0,8899	0,8852	0,8804	0,8755	0,8704	0,8652	0,8600	0,8545	0,8490
24	0,9284	0,9249	0,9214	0,9177	0,9139	0,9100	0,9060	0,9019	0,8976	0,8933
25	0,9530	0,9505	0,9479	0,9452	0,9424	0,9395	0,9365	0,9334	0,9302	0,9269
26	0,9701	0,9683	0,9665	0,9646	0,9626	0,9606	0,9584	0,9562	0,9539	0,9514
27	0,9816	0,9804	0,9792	0,9779	0,9765	0,9751	0,9736	0,9720	0,9704	0,9687

	Tasa Media de Llegada λ									
	19,1	19,2	19,3	19,4	19,5	19,6	19,7	19,8	19,9	20,0
10	0,0174	0,0165	0,0157	0,0149	0,0141	0,0134	0,0127	0,0120	0,0114	0,0108
11	0,0331	0,0315	0,0301	0,0287	0,0273	0,0260	0,0248	0,0236	0,0225	0,0214
12	0,0580	0,0556	0,0532	0,0509	0,0488	0,0467	0,0446	0,0427	0,0408	0,0390
13	0,0947	0,0911	0,0876	0,0842	0,0809	0,0778	0,0747	0,0717	0,0689	0,0661

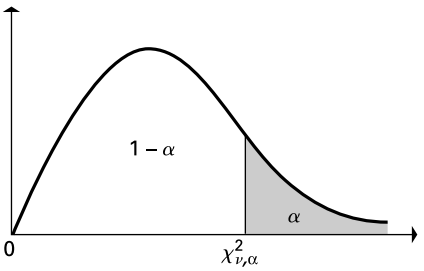
Tabla 6. Probabilidades de Poisson acumuladas (*continuación*).

	Tasa Media de Llegada λ									
	19,1	19,2	19,3	19,4	19,5	19,6	19,7	19,8	19,9	20,0
14	0,1447	0,1397	0,1349	0,1303	0,1257	0,1213	0,1170	0,1128	0,1088	0,1049
15	0,2084	0,2021	0,1959	0,1899	0,1840	0,1782	0,1726	0,1671	0,1617	0,1565
16	0,2844	0,2768	0,2694	0,2621	0,2550	0,2479	0,2410	0,2342	0,2276	0,2211
17	0,3698	0,3613	0,3529	0,3446	0,3364	0,3283	0,3203	0,3124	0,3047	0,2970
18	0,4604	0,4514	0,4424	0,4335	0,4246	0,4158	0,4071	0,3985	0,3899	0,3814
19	0,5515	0,5424	0,5333	0,5242	0,5151	0,5061	0,4971	0,4881	0,4792	0,4703
20	0,6385	0,6298	0,6210	0,6122	0,6034	0,5946	0,5857	0,5769	0,5680	0,5591
21	0,7176	0,7097	0,7016	0,6935	0,6854	0,6772	0,6689	0,6605	0,6521	0,6437
22	0,7863	0,7794	0,7724	0,7653	0,7580	0,7507	0,7433	0,7358	0,7283	0,7206
23	0,8434	0,8376	0,8317	0,8257	0,8196	0,8134	0,8071	0,8007	0,7941	0,7875
24	0,8888	0,8842	0,8795	0,8746	0,8697	0,8646	0,8594	0,8541	0,8487	0,8432
25	0,9235	0,9199	0,9163	0,9126	0,9087	0,9048	0,9007	0,8965	0,8922	0,8878
26	0,9489	0,9463	0,9437	0,9409	0,9380	0,9350	0,9319	0,9288	0,9255	0,9221
27	0,9670	0,9651	0,9632	0,9612	0,9591	0,9570	0,9547	0,9524	0,9500	0,9475

	Tasa Media de Llegada λ									
	20,1	20,2	20,3	20,4	20,5	20,6	20,7	20,8	20,9	21,0
10	0,0102	0,0097	0,0092	0,0087	0,0082	0,0078	0,0074	0,0070	0,0066	0,0063
11	0,0204	0,0194	0,0184	0,0175	0,0167	0,0158	0,0150	0,0143	0,0136	0,0129
12	0,0373	0,0356	0,0340	0,0325	0,0310	0,0296	0,0283	0,0270	0,0257	0,0245
13	0,0635	0,0609	0,0584	0,0560	0,0537	0,0515	0,0493	0,0473	0,0453	0,0434
14	0,1010	0,0973	0,0938	0,0903	0,0869	0,0836	0,0805	0,0774	0,0744	0,0716
15	0,1514	0,1464	0,1416	0,1369	0,1323	0,1278	0,1234	0,1192	0,1151	0,1111
16	0,2147	0,2084	0,2023	0,1963	0,1904	0,1847	0,1790	0,1735	0,1682	0,1629
17	0,2895	0,2821	0,2748	0,2676	0,2605	0,2536	0,2467	0,2400	0,2334	0,2270
18	0,3730	0,3647	0,3565	0,3484	0,3403	0,3324	0,3246	0,3168	0,3092	0,3017
19	0,4614	0,4526	0,4438	0,4351	0,4265	0,4179	0,4094	0,4009	0,3926	0,3843
20	0,5502	0,5413	0,5325	0,5236	0,5148	0,5059	0,4972	0,4884	0,4797	0,4710
21	0,6352	0,6267	0,6181	0,6096	0,6010	0,5923	0,5837	0,5750	0,5664	0,5577
22	0,7129	0,7051	0,6972	0,6893	0,6813	0,6732	0,6651	0,6569	0,6487	0,6405
23	0,7808	0,7739	0,7670	0,7600	0,7528	0,7456	0,7384	0,7310	0,7235	0,7160
24	0,8376	0,8319	0,8260	0,8201	0,8140	0,8078	0,8016	0,7952	0,7887	0,7822
25	0,8833	0,8787	0,8739	0,8691	0,8641	0,8591	0,8539	0,8486	0,8432	0,8377
26	0,9186	0,9150	0,9114	0,9076	0,9037	0,8997	0,8955	0,8913	0,8870	0,8826
27	0,9449	0,9423	0,9395	0,9366	0,9337	0,9306	0,9275	0,9242	0,9209	0,9175

	Tasa Media de Llegada λ									
	21,1	21,2	21,3	21,4	21,5	21,6	21,7	21,8	21,9	22,0
11	0,0123	0,0116	0,0110	0,0105	0,0099	0,0094	0,0090	0,0085	0,0080	0,0076
12	0,0234	0,0223	0,0213	0,0203	0,0193	0,0184	0,0175	0,0167	0,0159	0,0151
13	0,0415	0,0397	0,0380	0,0364	0,0348	0,0333	0,0318	0,0304	0,0291	0,0278
14	0,0688	0,0661	0,0635	0,0610	0,0586	0,0563	0,0540	0,0518	0,0497	0,0477
15	0,1072	0,1034	0,0997	0,0962	0,0927	0,0893	0,0861	0,0829	0,0799	0,0769
16	0,1578	0,1528	0,1479	0,1432	0,1385	0,1340	0,1296	0,1253	0,1211	0,1170
17	0,2206	0,2144	0,2083	0,2023	0,1965	0,1907	0,1851	0,1796	0,1743	0,1690
18	0,2943	0,2870	0,2798	0,2727	0,2657	0,2588	0,2521	0,2454	0,2389	0,2325
19	0,3760	0,3679	0,3599	0,3519	0,3440	0,3362	0,3285	0,3209	0,3134	0,3060
20	0,4623	0,4537	0,4452	0,4367	0,4282	0,4198	0,4115	0,4032	0,3950	0,3869
21	0,5490	0,5403	0,5317	0,5230	0,5144	0,5058	0,4972	0,4887	0,4801	0,4716
22	0,6322	0,6238	0,6155	0,6071	0,5987	0,5902	0,5818	0,5733	0,5648	0,5564
23	0,7084	0,7008	0,6930	0,6853	0,6774	0,6695	0,6616	0,6536	0,6455	0,6374
24	0,7755	0,7687	0,7619	0,7550	0,7480	0,7409	0,7337	0,7264	0,7191	0,7117
25	0,8321	0,8264	0,8206	0,8146	0,8086	0,8025	0,7963	0,7900	0,7836	0,7771
26	0,8780	0,8734	0,8686	0,8638	0,8588	0,8537	0,8486	0,8433	0,8379	0,8324
27	0,9139	0,9103	0,9065	0,9027	0,8988	0,8947	0,8906	0,8863	0,8820	0,8775

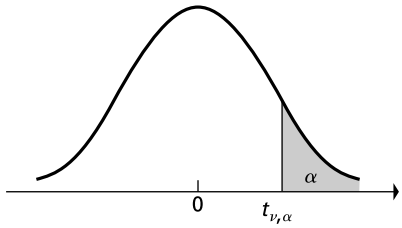
Tabla 7. Puntos de corte de la función de distribución ji-cuadrado.



La tabla muestra, para algunas probabilidades α, los valores de la $\chi^2_{\nu,\alpha}$ tales que $P(\chi^2_\nu > \chi^2_{\nu,\alpha}) = \alpha$, donde χ^2_ν es una variable aleatoria ji-cuadrado con ν grados de libertad. Por ejemplo, la probalidad de que una variable alatoria ji-cuadrado con 10 grados de libertad sea mayor que 15,99 es 0,100.

ν	α									
	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	$0{,}0^4393$	$0{,}0^3157$	$0{,}0^3982$	$0{,}0^2393$	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,60
3	0,072	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,145	1,61	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,541	0,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,4	104,2
80	51,17	53,54	57,15	60,39	64,28	96,58	101,9	106,6	112,3	116,3
90	59,20	61,75	65,65	69,13	73,29	107,6	113,1	118,1	124,1	128,3
100	67,33	70,06	74,22	77,93	82,36	118,5	124,3	129,6	135,8	140,2

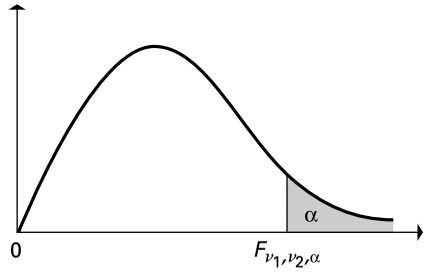
Tabla 8. Puntos de corte de la distribución *t* de Student.



La tabla muestra, para algunas probabilidades α, los valores de la $t^2_{\nu,\alpha}$ tales que $P(t_\nu > t_{\nu,\alpha}) = \alpha$, donde t_ν es una variable aleatoria t de Student con ν grados de libertad. Por ejemplo, la probalidad de que una variable aleatoria t de Student con 10 grados de libertad sea mayor que 1,372 es 0,10.

ν	α				
	0,100	0,050	0,025	0,010	0,005
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750
40	1,303	1,684	2,021	2,423	2,704
60	1,296	1,671	2,000	2,390	2,660
∞	1,282	1,645	1,960	2,326	2,576

Tabla 9. Puntos de corte de la distribución *F*.



Las tablas muestran, para las probabilidades $\alpha = 0{,}5$ y $\alpha = 0{,}01$, los valores de $F_{\nu_1\nu_2,\alpha}$ tales que $P(F_{\nu_1\nu_2} > F_{\nu_1\nu_2,\alpha}) = \alpha$, donde $F_{\nu_1\nu_2}$ es una variable aleatoria F con ν_1 grados de libertad en el numerador y ν_2 grados de libertad en el denominador. Por ejemplo, la probabilidad de que una variable $F_{3,7}$ sea mayor que 4,35 es 0,05.

$\alpha = 0{,}05$

Denominador ν_2	Numerador ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88

Tabla 9. Puntos de corte de la distribución *F* (*continuación*).

	$\alpha = 0{,}05$																		
Denominador ν_2	Numerador ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,67	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,58	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Tabla 9. Puntos de corte de la distribución *F* (*continuación*).

$\alpha = 0{,}01$																			
Denominador ν_2	Numerador ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4.052	4.999,5	5.403	5.625	5.764	5.859	5.928	5.982	6.022	6.056	6.106	6.157	6.209	6.235	6.261	6.287	6.313	6.339	6.366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,48	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,30	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,30	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	3,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	2,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Tabla 10. Puntos de corte de la distribución del estadístico de contraste de Wilcoxon.

En la tabla aparecen los números T_α tales que $P(R \leq T_\alpha) = \alpha$, correspondientes a una muestra de tamaño n y distintos valores de α, siendo la distribución de la variable aleatoria T la del estadístico de contraste de Wilcoxon según la hipótesis nula.

n	α				
	0,005	0,010	0,025	0,050	0,100
4	0	0	0	0	1
5	0	0	0	1	3
6	0	0	1	3	4
7	0	1	3	4	6
8	1	2	4	6	9
9	2	4	6	9	11
10	4	6	9	11	15
11	6	8	11	14	18
12	8	10	14	18	22
13	10	13	18	22	27
14	13	16	22	26	32
15	16	20	26	31	37
16	20	24	30	36	43
17	24	28	35	42	49
18	28	33	41	48	56
19	33	38	47	54	63
20	38	44	53	61	70

Permiso de reproducción de R. L. McCormack, «Extended tables of the Wilcoxon matched pairs signed rank statistics», *Journal of the American statistical Association*, 60, 1965.

Tabla 11. Puntos de corte de la distribución del coeficiente de correlación de orden de Spearman.

En la tabla aparecen los números $r_{s,\alpha}$ tales que $P(r_s > r_{s,\alpha}) = \alpha$, correspondientes a una muestra de tamaño n y algunos valores de α, siendo la distribución de la variable aleatoria r_s la del coeficiente de correlación de orden de Spearman según la hipótesis nula de que no existe ninguna relación n,

n	α			
	0,050	0,025	0,010	0,005
5	0,900	—	—	—
6	0,829	0,886	0,943	—
7	0,714	0,786	0,893	—
8	0,643	0,738	0,833	0,881
9	0,600	0,683	0,783	0,833
10	0,564	0,648	0,745	0,794
11	0,523	0,623	0,736	0,818
12	0,497	0,591	0,703	0,780
13	0,475	0,566	0,673	0,745
14	0,457	0,545	0,646	0,716
15	0,441	0,525	0,623	0,689
16	0,425	0,507	0,601	0,666
17	0,412	0,490	0,582	0,645
18	0,399	0,476	0,564	0,625
19	0,388	0,462	0,549	0,608
20	0,377	0,450	0,534	0,591
21	0,368	0,438	0,521	0,576
22	0,359	0,428	0,508	0,562
23	0,351	0,418	0,496	0,549
24	0,343	0,409	0,485	0,537
25	0,336	0,400	0,475	0,526
26	0,329	0,392	0,465	0,515
27	0,323	0,385	0,456	0,505
28	0,317	0,377	0,448	0,496
29	0,311	0,370	0,440	0,487
30	0,305	0,364	0,432	0,478

Permiso de reproducción de E. G. Olds, «Distribution of sums of squares of rank differences for small samples», *Annals of Mathematical Statistics*, 9, 1938.

Tabla 12. Puntos de corte de la distribución del estadístico de contraste de Durbin-Watson.

Sea d_α el número tal que $P(d < d_\alpha) = \alpha$, donde la variable aleatoria d tiene la distribución del estadístico de Durbin-Watson según la hipótesis nula de que no existe ninguna autocorrelación en los errores de regresión. Las tablas muestran para las probabilidades $\alpha = 0{,}05$ y $\alpha = 0{,}01$, correspondientes a los números de variables independientes, K, los valores de d_L y d_U tales que $d_L \leq d_\alpha \leq d_U$, cuando el número de observaciones es n.

	$\alpha = 0{,}05$									
n	K									
	1		2		3		4		5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	1,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Tabla 12. Puntos de corte de la distribución del estadístico de contraste de Durbin-Watson (*continuación*).

	$\alpha = 0{,}01$									
n	K									
	1		2		3		4		5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
16	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
17	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
18	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
19	0,93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
20	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
21	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
22	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
23	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
24	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
25	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
26	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
27	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
28	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
29	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
30	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
31	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
32	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
33	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
34	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
35	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
36	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
37	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
38	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
39	1,24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
40	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
45	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
50	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
55	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
60	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
65	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
70	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
75	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
80	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
85	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
90	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
95	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
100	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65

Tabla 13. Constantes de los gráficos de control.

	GRÁFICOS $\bar{X}$				GRÁFICOS s				GRÁFICOS R					
n	A	A_2	A_3	c_4	B_3	B_4	B_5	B_6	d_2	d_3	D_1	D_2	D_3	D_4
2	2,121	1,880	2,659	0,7979	0	3,267	0	2,606	1,128	0,853	0	3,686	0	3,267
3	1,732	1,023	1,954	0,8862	0	2,568	0	2,276	1,693	0,888	0	4,358	0	2,574
4	1,500	0,729	1,628	0,9213	0	2,266	0	2,088	2,059	0,880	0	4,698	0	2,282
5	1,342	0,577	1,427	0,9400	0	2,089	0	1,964	2,326	0,864	0	4,918	0	2,114
6	1,225	0,483	1,287	0,9515	0,030	1,970	0,029	1,874	2,534	0,848	0	5,078	0	2,004
7	1,134	0,419	1,182	0,9594	0,118	1,882	0,113	1,806	2,704	0,833	0,204	5,204	0,076	1,924
8	1,061	0,373	1,099	0,9650	0,185	1,815	0,179	1,751	2,847	0,820	0,388	5,306	0,136	1,864
9	1,000	0,337	1,032	0,969	0,239	1,761	0,232	1,707	2,970	0,808	0,547	5,393	0,184	1,816
10	0,949	0,308	0,975	0,9727	0,284	1,716	0,276	1,669	3,078	0,797	0,687	5,469	0,223	1,777
11	0,905	0,285	0,927	0,9754	0,321	1,679	0,313	1,637	3,173	0,787	0,811	5,535	0,256	1,744
12	0,866	0,266	0,886	0,9776	0,354	1,646	0,346	1,610	3,258	0,778	0,922	5,594	0,283	1,717
13	0,832	0,249	0,850	0,9794	0,382	1,618	0,374	1,585	3,336	0,770	1,025	5,647	0,307	1,693
14	0,802	0,235	0,817	0,9810	0,406	1,594	0,399	1,563	3,407	0,763	1,118	5,696	0,328	1,672
15	0,775	0,223	0,789	0,9823	0,428	1,572	0,421	1,544	3,472	0,756	1,203	5,741	0,347	1,653
16	0,750	0,212	0,763	0,9835	0,448	1,552	0,440	1,526	3,532	0,750	1,282	5,782	0,363	1,637
17	0,728	0,203	0,739	0,9845	0,466	1,534	0,458	1,511	3,588	0,744	1,356	5,820	0,378	1,622
18	0,707	0,194	0,718	0,9854	0,482	1,518	0,475	1,496	3,640	0,739	1,424	5,856	0,391	1,608
19	0,688	0,187	0,698	0,9862	0,497	1,503	0,490	1,483	3,689	0,734	1,487	5,891	0,403	1,597
20	0,671	0,180	0,680	0,9869	0,510	1,490	0,504	1,470	3,735	0,729	1,549	5,921	0,415	1,585
21	0,655	0,173	0,663	0,9876	0,523	1,477	0,516	1,459	3,778	0,724	1,605	5,951	0,425	1,575
22	0,640	0,167	0,647	0,9882	0,534	1,466	0,528	1,448	3,819	0,720	1,659	5,979	0,434	1,566
23	0,626	0,162	0,633	0,9887	0,545	1,455	0,539	1,438	3,858	0,716	1,710	6,006	0,443	1,557
24	0,612	0,157	0,619	0,9892	0,555	1,445	0,549	1,429	3,895	0,712	1,759	6,031	0,451	1,548
25	0,600	0,153	0,606	0,9896	0,565	1,435	0,559	1,420	3,931	0,708	1,806	6,056	0,459	1,541

Fuente: Adaptado de la tabla 27 de ASTM STP 15D ASTM *Manual on Presentation of Data and Control Chart Analysis*. © 1976 American Society for Testing and Materials, Filadelfia, PA.

Tabla 14. Función de distribución acumulada del estadístico del contraste de rachas.

La tabla muestra la probabilidad de que en una serie temporal aleatoria el número de rachas no sea mayor que K cuando el número de observaciones es n.

n	K																		
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
6	0,100	0,300	0,700	0,900	1,000														
8	0,029	0,114	0,371	0,629	0,886	0,971	1,000												
10	0,008	0,040	0,167	0,357	0,643	0,833	0,960	0,992	1,000										
12	0,002	0,013	0,067	0,175	0,392	0,608	0,825	0,933	0,987	0,998	1,000								
14	0,001	0,004	0,025	0,078	0,209	0,383	0,617	0,791	0,922	0,975	0,996	0,999	1,000						
16	0,000	0,001	0,009	0,032	0,100	0,214	0,405	0,595	0,786	0,900	0,968	0,991	0,999	1,000	1,000				
18	0,000	0,000	0,003	0,012	0,044	0,109	0,238	0,399	0,601	0,762	0,891	0,956	0,988	0,997	1,000	1,000	1,000		
20	0,000	0,000	0,001	0,004	0,019	0,051	0,128	0,242	0,414	0,586	0,758	0,872	0,949	0,981	0,996	0,999	1,000	1,000	1,000

RESPUESTAS DE ALGUNOS EJERCICIOS PARES

Capítulo 1

1.2. Varias respuestas. Son decisiones de marketing en condiciones de incertidumbre las decisiones relacionadas con los precios, con la promoción, con la publicidad, con el empaquetado, etc.

1.4. **a)** Varias respuestas. Un parámetro poblacional podría ser la verdadera media poblacional de la renta de todas las familias que viven en una ciudad.
b) Varias respuestas. Un parámetro poblacional podría ser la verdadera desviación típica poblacional de todas las acciones que cotizan en una bolsa de valores.
c) Varias respuestas. Un parámetro poblacional podría ser la verdadera media poblacional de los costes de todas las reclamaciones que recibe en un año dado una compañía de seguros médicos.
d) Varias respuestas. Un parámetro poblacional podría ser la verdadera media poblacional de los valores de todas las facturas pendientes de cobro de una empresa.

1.6. **a)** La población son todos los vuelos programados de la compañía en el aeropuerto de Nueva York.
b) La muestra son los 200 vuelos seleccionados aleatoriamente.
c) El estadístico es el 1,5% que se observó que salía tarde en los 200 vuelos seleccionados aleatoriamente.
d) 1,5% es un estadístico muestral.

1.8. **a)** Descriptiva: para describir la información sobre la muestra de una semana.
b) Inferencial: para estimar el verdadero porcentaje de todos los empleados que llegan tarde a trabajar.
c) Inferencial: para predecir las relaciones entre los años de experiencia y la escala salarial.

Capítulo 2

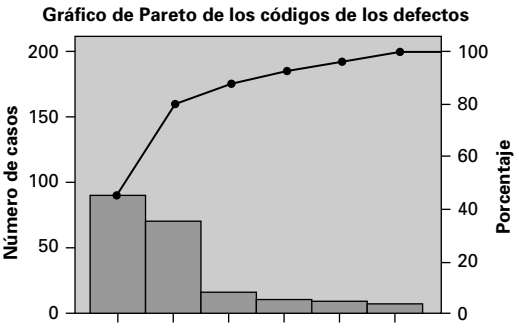
2.2. **a)** Datos categóricos. Los niveles de medición son cualitativos-nominales. Respuesta sí/no.
b) Datos categóricos. Los niveles de medición son cualitativos-nominales.
c) Datos numéricos. Generalmente, se considera que las cantidades monetarias son continuas, aunque podamos agrupar las cantidades monetarias y tratarlas como si fueran discretas.

2.4. **a)** Categóricos-Cualitativos-ordinales
b) Numéricos-Cuantitativos-discretos
c) Categóricos-Cualitativos-nominales
d) Categóricos-Cualitativos-nominales

2.6. **a)** Categóricos-Cualitativos-nominales
b) Numéricos-Cuantitativos-discretos
c) Categóricos-Cualitativos-nominales; respuesta sí/no
d) Categóricos-Cualitativos-ordinales

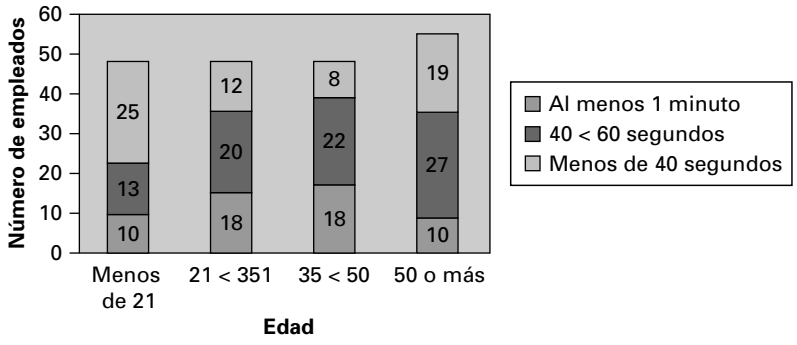
2.8. **a)** Varias respuestas-Variable categórica con respuestas ordinales: preocupación por la salud
b) Varias respuestas-Variable categórica con respuestas nominales: sexo

2.10.

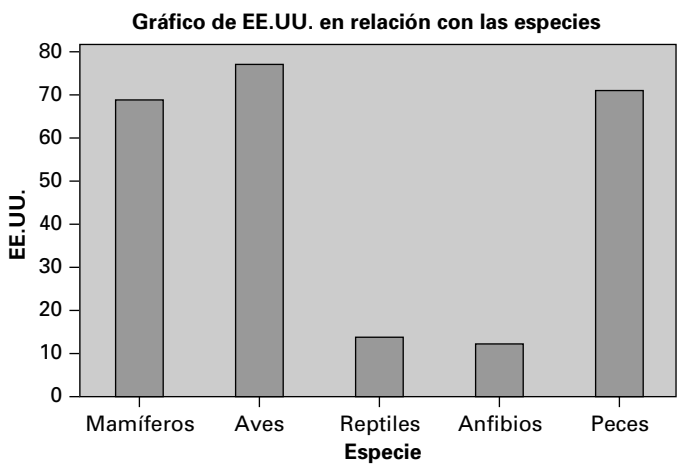


Defect Circle Ex2.10	D	B	C	A	E	Other
Count	90	70	15	10	8	7
Percent	45.0	35.0	7.5	5.0	4.0	3.5
Cum%	45.0	80.0	87.5	92.5	96.5	100.0

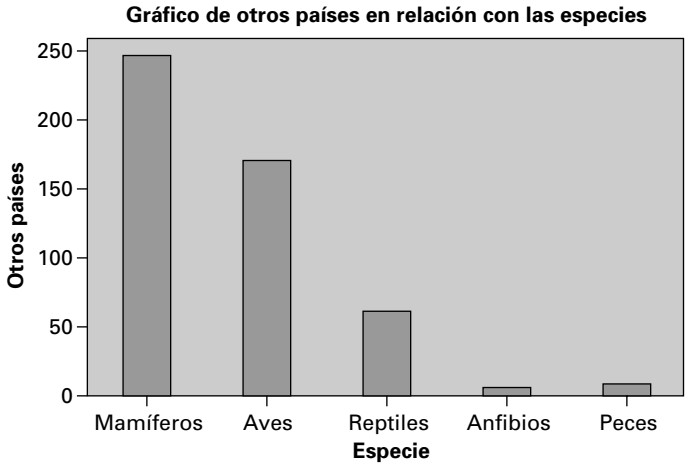
2.12.



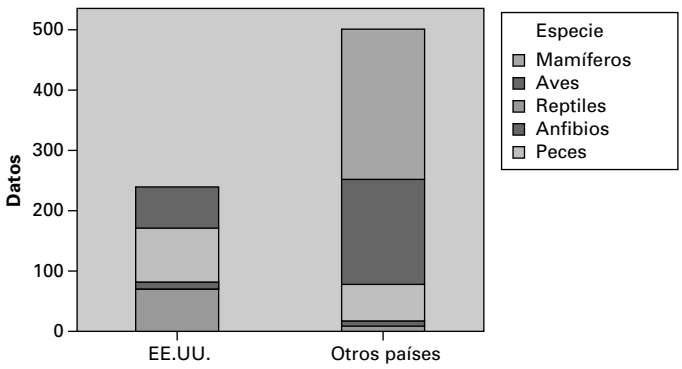
2.14. a)



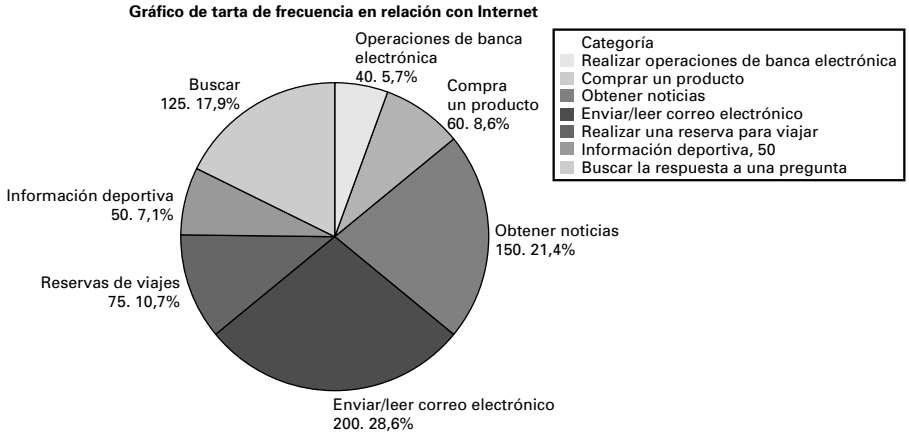
b)



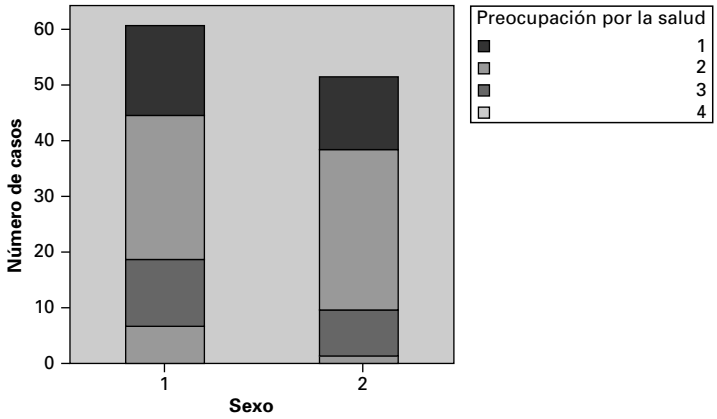
c)



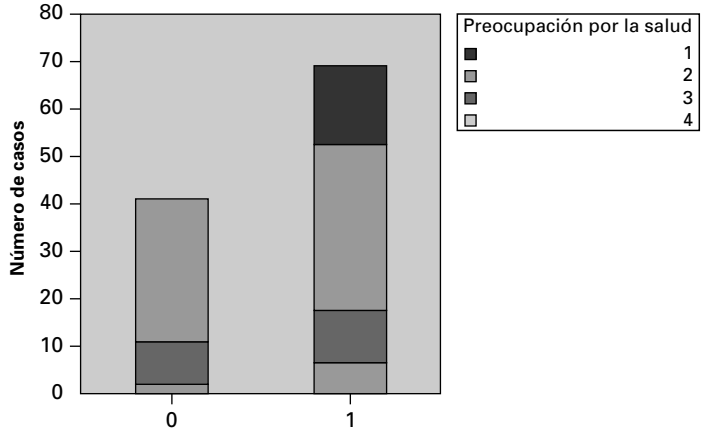
2.16. Describir los datos gráficamente



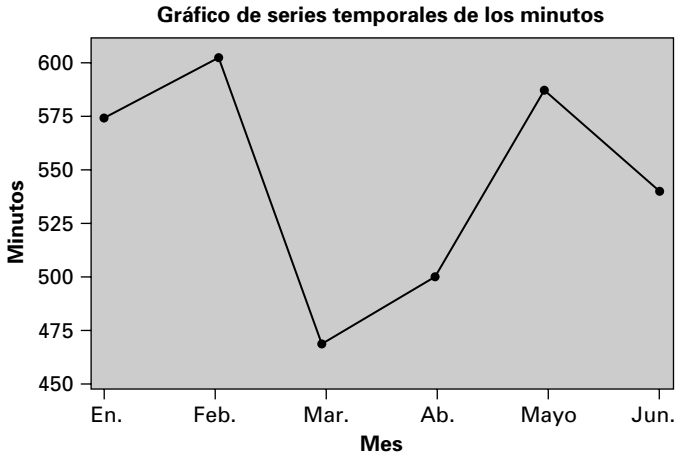
2.18. a) **Gráfico del sexo (1 = hombre, 2 = mujer) y preocupación por la salud**



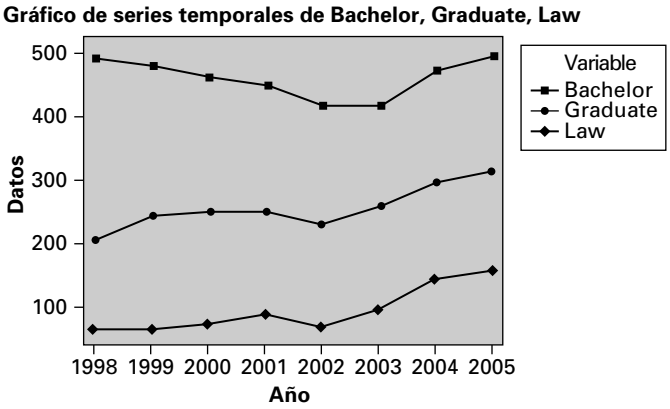
b) **Deseo de suplementos proteínicos (0=No; 1=Sí) y nivel de preocupación por la salud**



2.20.

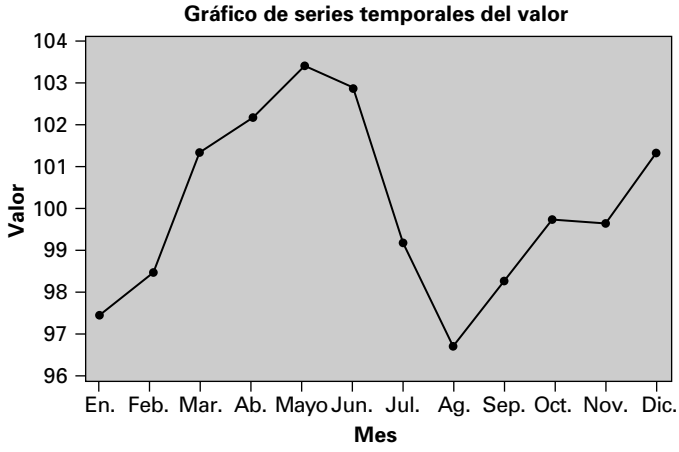


2.22. a)

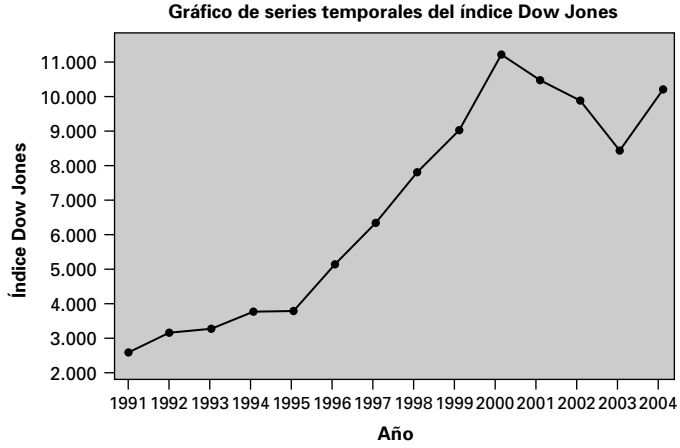


b) El número de títulos de «law» y «graduate» está aumentando. El número de títulos de «bachelor» disminuyó entre 1998 y 2002, se estabilizó en 2003 y comenzó a mostrar una tendencia ascendente en 2004. Es posible que convenga limitar el número de alumnos si las clases son demasiado numerosas o si hay demasiados alumnos.

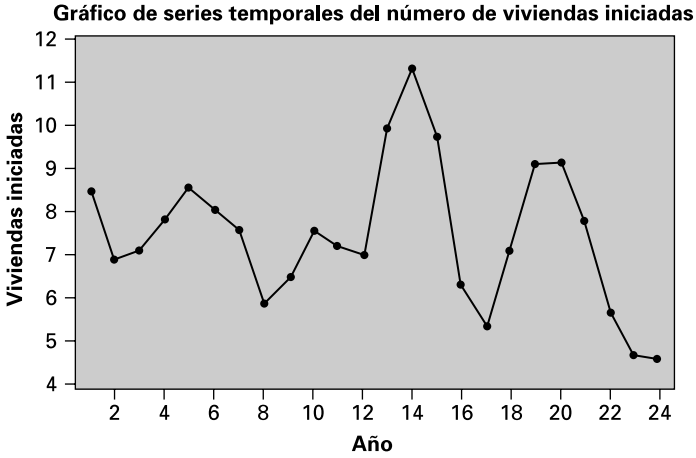
2.24.



2.26.



2.28.

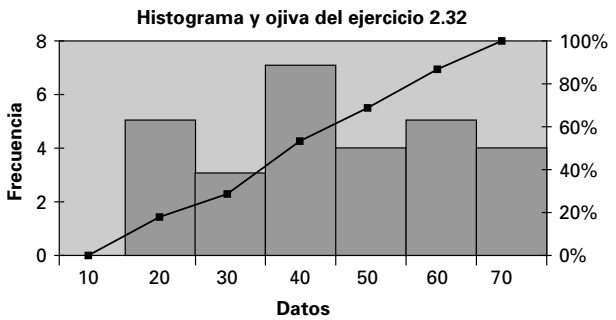


2.30. **a)** 5-7 clases **b)** 7-8 clases **c)** 8-10 clases
d) 8-10 clases **e)** 10-11 clases

2.32. **a)**

Clases	Frecuencia
10 < 20	5
20 < 30	3
30 < 40	7
40 < 50	4
50 < 60	5
60 < 70	4

b) histograma y **c)** ojiva



d) **Diagrama de tallo y hojas: datos**

```
Stem-and-leaf of Data N = 28
Leaf Unit = 1.0

 5    1  23557
 8    2  148
(7)   3  2567799
 13   4  0144
 9    5  14699
 4    6  2455
```

2.34.

Clases	Frecuencia	A) Frecuencia relativa	B) Frecuencia acumulada	C) Frecuencia acumulada relativa
0 < 10	8	16,33%	8	16,33%
10 < 20	10	20,41%	18	36,74%
20 < 30	13	26,53%	31	63,27%
30 < 40	12	24,49%	43	87,76%
40 < 50	6	12,24%	49	100,00%
Total	49	100,00%		

2.36. Varias respuestas - una posibilidad es utilizar 7 clases con una amplitud de 0,1

Clases	Frecuencia	% acumulado
3,5 < 3,6	1	1,33%
3,6 < 3,7	8	12,00%
3,7 < 3,8	29	50,67%
3,8 < 3,9	22	80,00%
3,9 < 4,0	13	97,33%
4,0 < 4,10	1	98,67%
4,10 < 4,20	1	100,00%

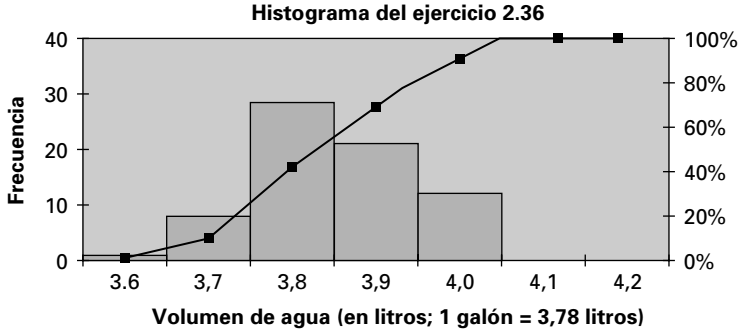
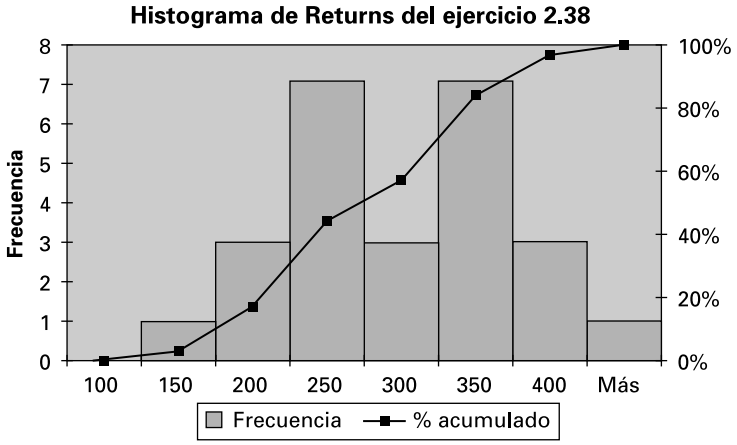


Diagrama de tallo y hojas: peso

```
Stem-and-leaf of Weights N = 28
Leaf Unit = 0.010

 1    35  7
 3    36  34
 9    36  577799
 21   37  111122344444
(17)  37  55566777777889999
 37   38  0111112222244
 24   38  556677899
 15   39  01334444
 7    39  56689
 2    40
 2    40  6
 1    41  1
```

2.38. a) Histograma y **c)** Ojiva de los datos de **Returns**

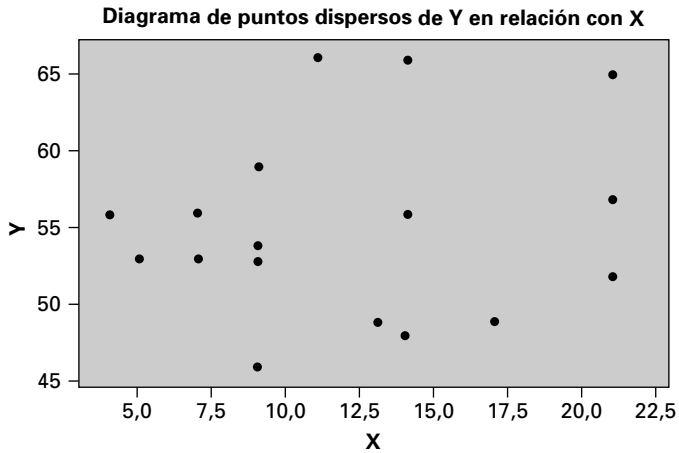


b)

Diagrama de tallo y hojas: Returns

```
Stem-and-leaf of Weights N = 25
Leaf Unit = 10

 1   1  3
 4   1  899
 11  2  0014444
(3)  2  589
 11  3  0000122
 4   3  689
 1   4
 1   4
 1   5  0
```

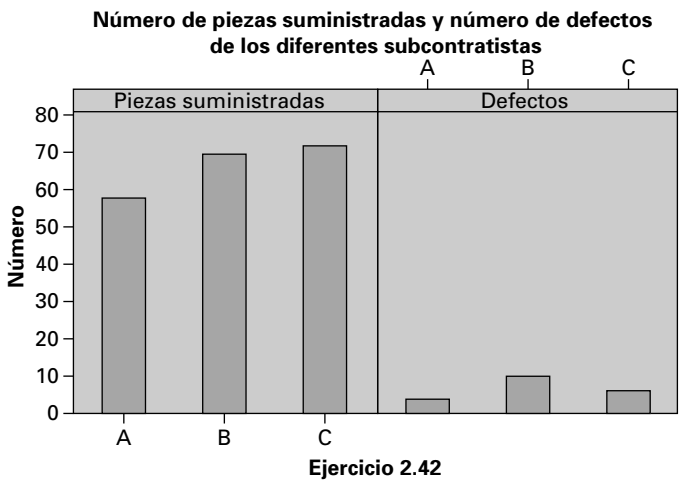
2.40.



2.42. a)

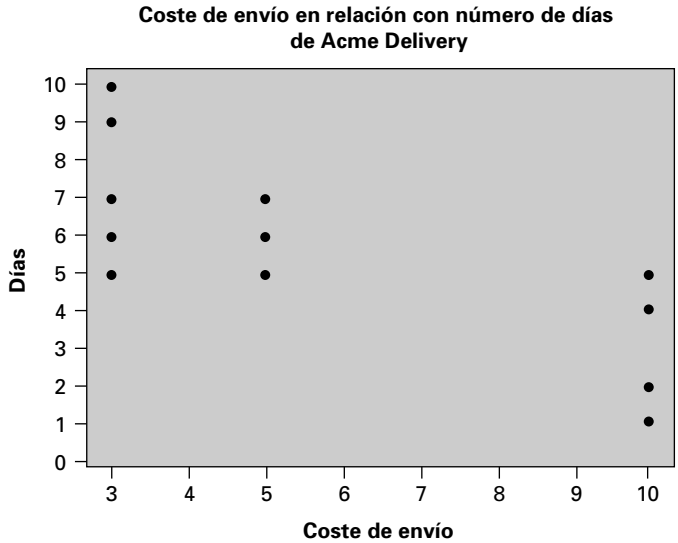
Subcontratista	Piezas defectuosas	Piezas no defectuosas	Piezas suministradas
A	4	54	58
B	10	60	70
C	6	66	72
Total	20	180	200

b)



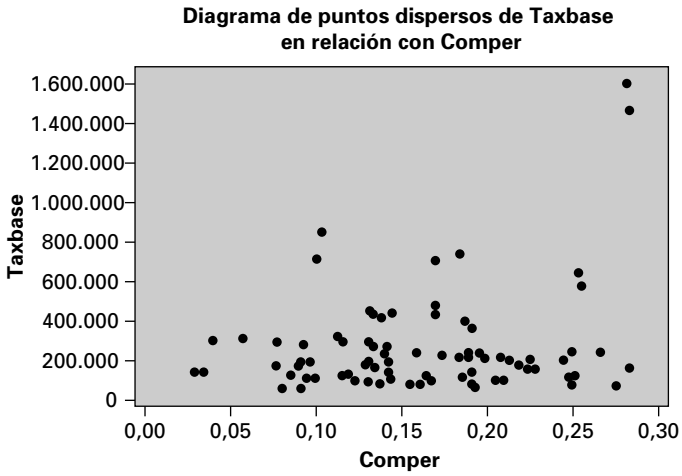
Ejercicio 2.42

2.44.



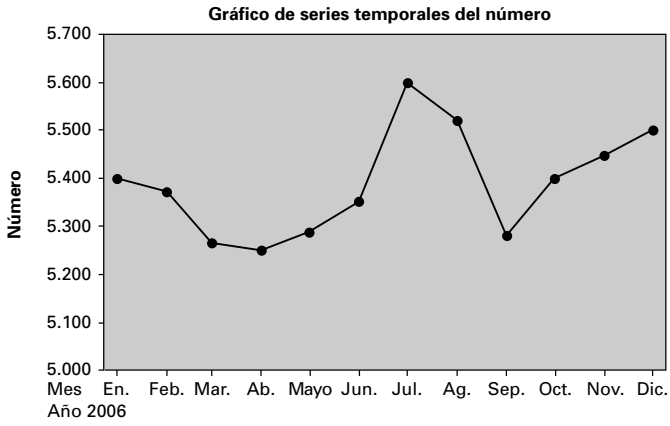
La relación parece que es negativa; sin embargo, el tiempo de envío correspondiente a cada uno de los tres costes de envío —ordinario, 3 $; urgente, 5 $; y superurgente, 10 $— es muy variable.

2.46.

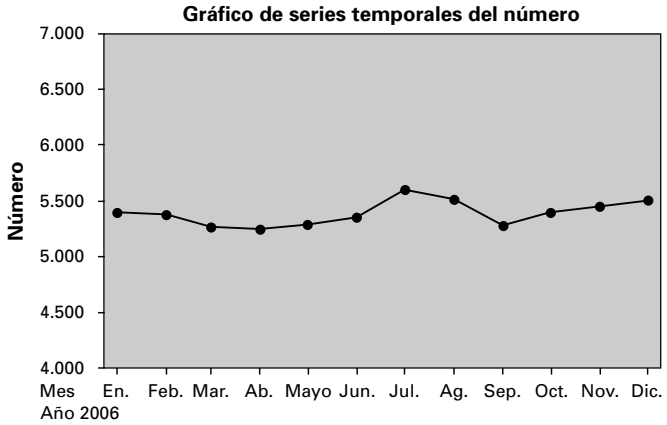


No existe ninguna relación entre las dos variables, por lo que no existen pruebas de que aumente la base imponible poniendo énfasis en atraer a un porcentaje mayor de propiedades comerciales. Los dos puntos extremos del lado de la derecha del gráfico podrían utilizarse para argumentar que la existencia de una gran cantidad de propiedades comerciales aumenta la base imponible. Sin embargo, ese argumento es contrario a la pauta global de los datos.

2.48. a)

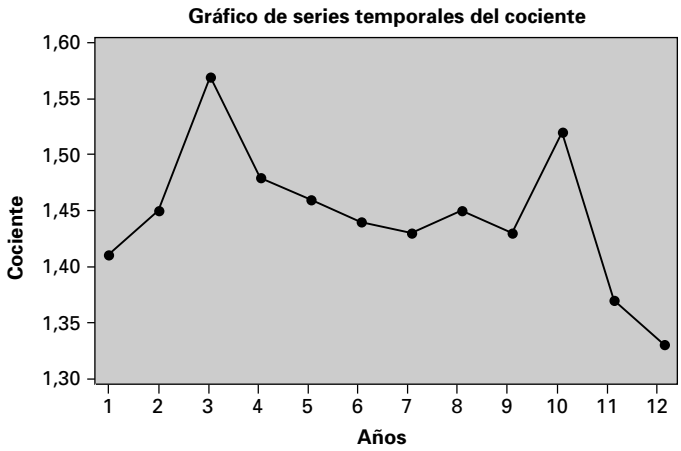


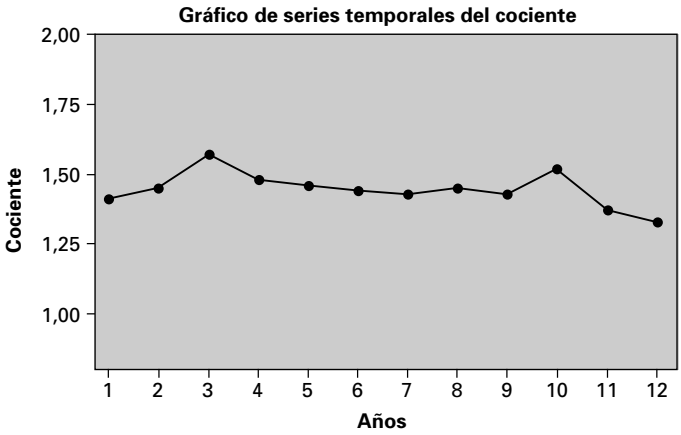
b)



c) Las diferencias entre los dos gráficos se deben, entre otras cosas, a la variabilidad de la serie de datos. Uno de los gráficos sugiere que hay mayor variabilidad en la series de datos, mientras que el otro sugiere que la línea es relativamente plana. Téngase presente la escala en la que se realizan las mediciones.

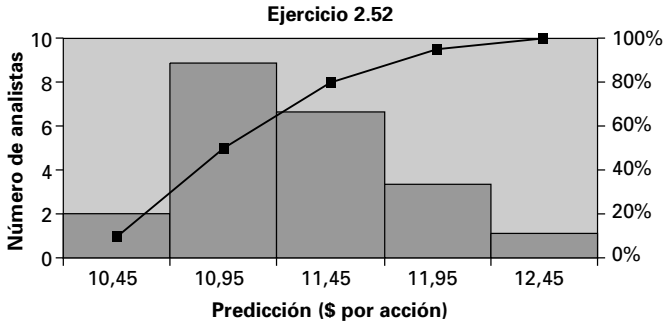
2.50.





Las diferencias entre los dos gráficos se deben, entre otras cosas, a la variabilidad de la serie de datos. Uno de los gráficos sugiere que hay mayor variabilidad en la series de datos, mientras que el otro sugiere que la línea es relativamente plana. Téngase presente la escala en la que se realizan las mediciones.

2.52. a)

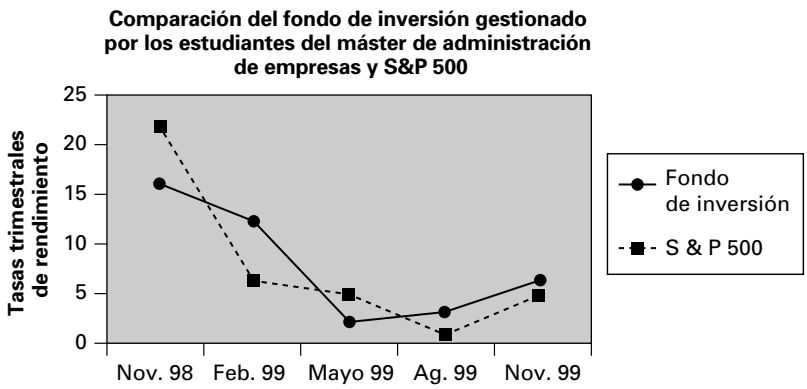


Las respuestas a b), c) y d) son:

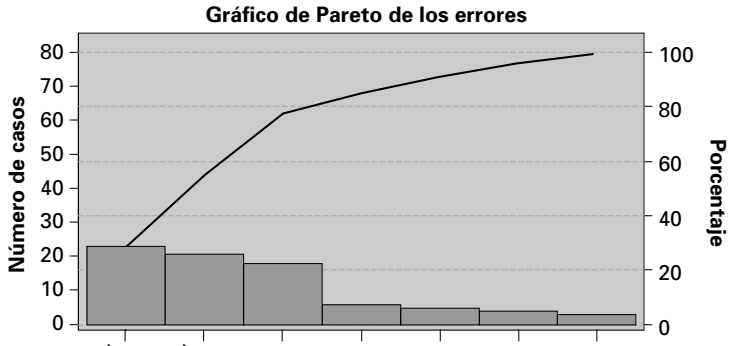
Predicción de los beneficios por acción	Frecuencia	Frec. relativa	Frec. acumulada	% acumulado
9,95	2	0,1	2	10,00%
10,45	8	0,4	10	50,00%
10,95	6	0,3	16	80,00%
11,45	3	0,15	19	95,00%
11,95	1	0,05	20	100,00%

d) Las frecuencias relativas acumuladas se encuentran en la última de la tabla anterior. Estas cifras indican el porcentaje de analistas que predicen ese nivel de beneficios por acción y todas las clases anteriores, incluida la clase considerada.

2.54.



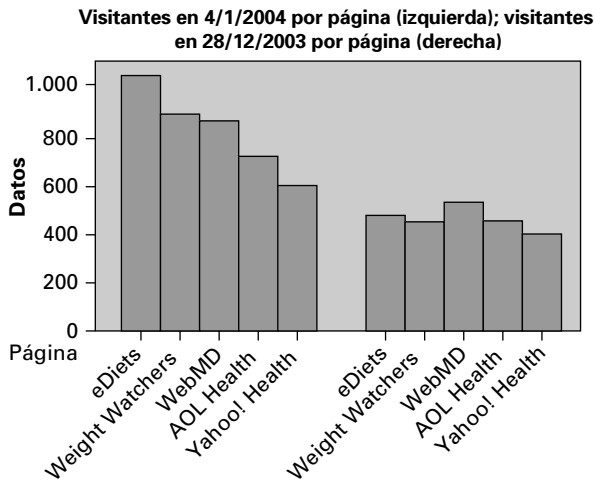
2.56. a)



Error	Escribir mal el nombre del titular	Escribir mal el número de la finca	Finca situada fuera de los límites de la provincia	Escrituras recibidas después de imprimir	Propiedad vendida	Descripción legal errónea o incompleta	Otros
Número de casos	23	21	18	6	5	4	3
Porcentaje	28,7	26,3	22,5	7,5	6,3	5,0	3,8
% acum.	28,7	55,0	77,5	85,0	91,3	96,3	100,0

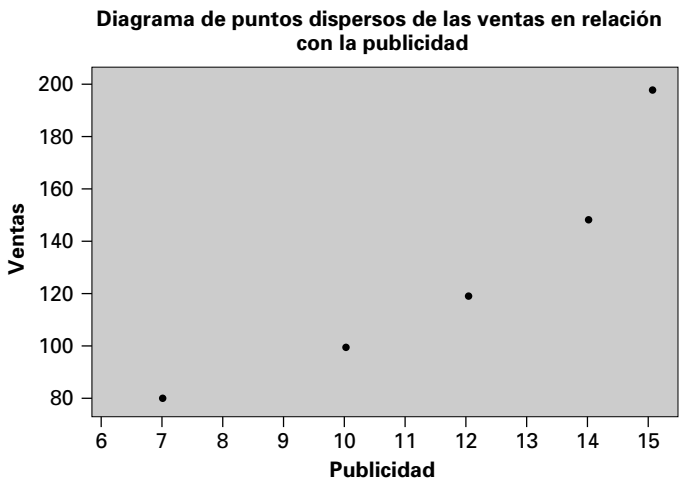
b) Entre las recomendaciones debería encontrarse un análisis del proceso de introducción de los datos. Éstos eran introducidos por personas que no tenían ninguna información sobre ellos. Una importante recomendación es que debe formarse al personal encargado de introducirlos. El aumento del tamaño de los monitores utilizados por el personal que introduce los datos también reduciría el número de errores.

2.58.

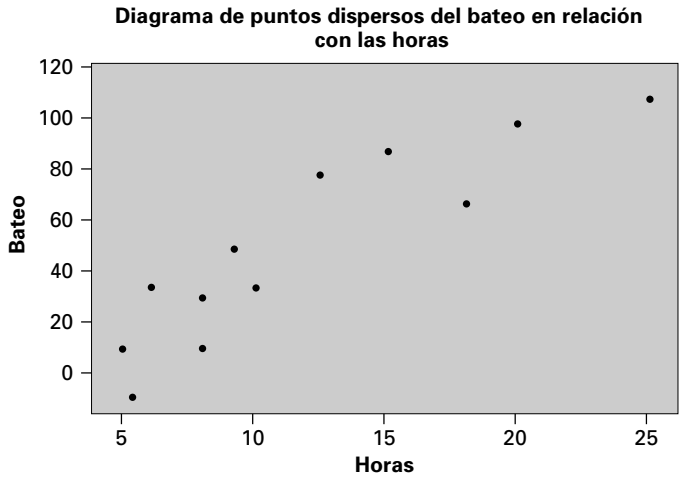


Los aumentos del tráfico semanal registrados entre 2003 y 2004 podrían deberse a que el número total de usuarios de Internet ha aumentado, a que es mayor la información sobre las páginas de Internet dedicadas a la salud o al envejecimiento de la población perteneciente a la explosión de la natalidad que la ha llevado a interesarse más por los temas de salud.

2.60.



2.62.

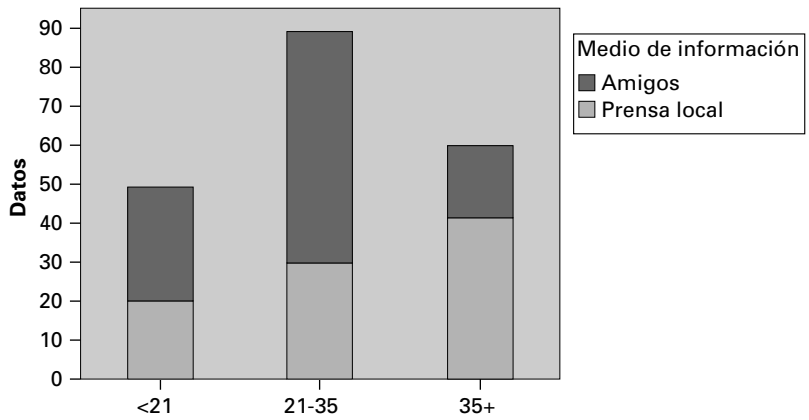


Parece que el número de horas semanales del programa especial de entrenamiento está relacionado positivamente con el cambio de las medias de bateo con respecto a la temporada anterior.

2.64. a)

Edad	Amigos	Prensa local	Subtotal
< 21 años	30	20	50
21–35	60	30	90
> 35	18	42	60
Subtotal	108	92	200

b) **Gráfico de <21, 21-35, 35+ en relación con medio de información**



2.66. a)

Preocupación por la salud	Hombres	Mujeres	Subtotal
Mucha	16	13	55
Moderada	26	29	55
Poca	12	8	20
No mucha	7	2	9
Subtotal	61	52	113

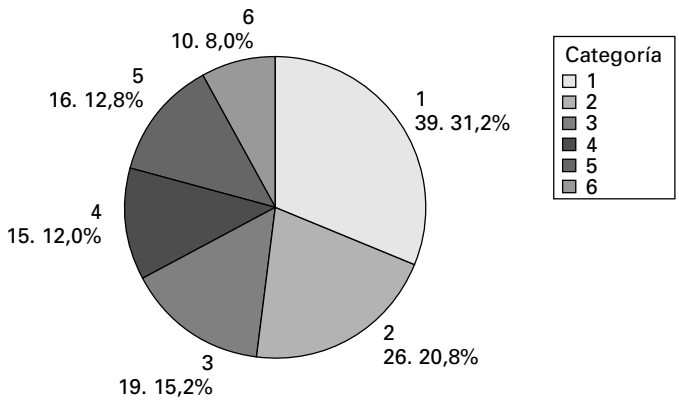
b) ¿Le gusta tomar un suplemento proteínico con su batido?

Preocupación por la salud	No	Sí	Subtotal
Mucha	12	17	29
Moderada	19	36	55
Poca	9	11	20
No mucha	2	7	9
Subtotal	42	71	113

2.68. a)

Método de pago	L	M	Mi	J	V	S	Total
Am Ex	7	0	3	4	3	6	23
MC	1	4	4	2	4	9	24
Visa	6	6	4	5	8	10	39
Efectivo	3	1	0	0	3	9	16
Otros	2	0	4	4	7	6	23
Subtotal	19	11	15	15	25	40	125

b) **Gráfico de tarta de las preferencias por el color de las rosas**



Capítulo 3

3.2. a) 12 **b)** 13 **c)** 8

3.4. a) 5,94 **b)** 6,35

c) La distribución es relativamente simétrica, ya que la media de 5,94 está relativamente cerca de la mediana de 6,35. Dado que la media es algo menor que la mediana, la distribución está algo sesgada hacia la izquierda.

3.6. a) 53,57. La demanda media de botellas de un galón es de 53,57, que es el punto que equilibra la distribución. La mediana de 55 indica que la mitad de la distribución tenía un volumen de ventas de más de 55 botellas y la mitad tenía un volumen de ventas inferior a esa cantidad. No existe una única moda en la distribución.

b) Comente la simetría o el sesgo. Dado que la media es algo menor que la mediana, la distribución está algo sesgada hacia la izquierda.

3.8. a) 25,58 **b)** 22,50 **c)** 22

3.10. **a)** 8,545 **b)** 9,0
c) La distribución está algo sesgada hacia la izquierda, ya que la media es algo menor que la mediana.

3.12. $s^2 = 5{,}143$ y $s = 2{,}268$

3.14. $\bar{x} = 9$; $s^2 = 2{,}5$; $s = 1{,}581$; $CV = 17{,}57$

3.16. **a)** RIC = 24,25; $Q_1 = 49{,}5$; $Q_3 = 73{,}75$
b) 77,2
c) 83,64

3.18. **a)** 190 y 310. Al menos el 8,9% de las observaciones se encuentra dentro de 3 desviaciones típicas de la media.
b) 210 y 290. Al menos el 75% de las observaciones se encuentra dentro de 2 desviaciones típicas de la media.
c) 230 y 270. Al menos el 0% de las observaciones se encuentra dentro de 1 desviación típica de la media.

3.20. **a)** $\mu_{\text{accciones}} = 8{,}16$, $\mu_{\text{letrasT}} = 5{,}786$
El rendimiento porcentual anual medio de las acciones es mayor que el de las letras del Tesoro de Estados Unidos.
b) $\sigma_{\text{accciones}} = 20{,}648$, $\sigma_{\text{letrasT}} = 1{,}362$
La variabilidad de las letras del Tesoro de Estados Unidos es mucho mayor que el rendimiento de las acciones.

3.22. **a)** rango = 0,54, desviación típica = 0,1024, varianza = 0,010486
b) Resumen de cinco números:

Min	Q1	Mediana	Q3	Max
3,57	3,74	3,79	3,87	4,11

c) RIC = 0,13. Nos dice que el rango del 50% central de la distribución es 0,13.
d) 0,02689, o sea, 2,689%.

3.24. **a)** $s = 3{,}8696$
b) La distribución tiene forma de campana. Por lo tanto, se aplica la regla empírica. Es de esperar que alrededor del 95% de la distribución esté dentro de $+/-2$ desviaciones típicas de la media.

3.26. **a)** 4,2 **b)** 4,583

3.28. 32.299,519

3.30. **a)** 1,40 **b)** $s^2 = 3{,}0612$, $s = 1{,}7496$

3.32. **a)** 11,025 **b)** 0,520

3.34. **a)** 261,54545 **b)** 17,370

3.36. **a)** 1.392,5 **b)** 0,9930

3.38. **a)** 4,268 **b)** 0,128
c) Débil relación positiva entre el número de dosis del medicamento y el número de días necesarios para la recuperación total. Se recomienda una dosis baja o nula.

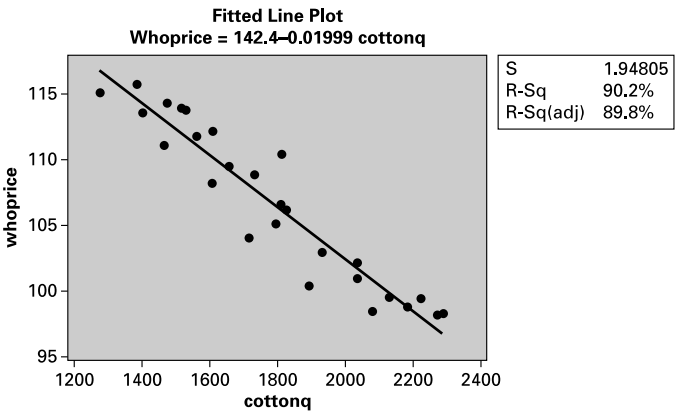
3.40. **a)** 0,65
b) 4,40
c) $\hat{y} = b_0 + b_1 x = 4{,}40 + 0{,}65x$

3.42. **a)** Covarianza = −99,762, Correlación = −0,927136
b) $b_1 = -18{,}217$. Estimamos que si el precio de un litro de pintura aumenta un dólar, la cantidad vendida en siete días de operaciones disminuiría en 18,217 litros de pintura.
c) $b_0 = 268{,}70$. Si el precio de la pintura fuera de 0 $ por litro, esperaríamos que se vendieran 268,7 litros en siete días de operaciones. Interprete con cautela los datos: observe que estamos extrapolando los resultados a puntos situados fuera del rango de datos observados.
d) 141,181

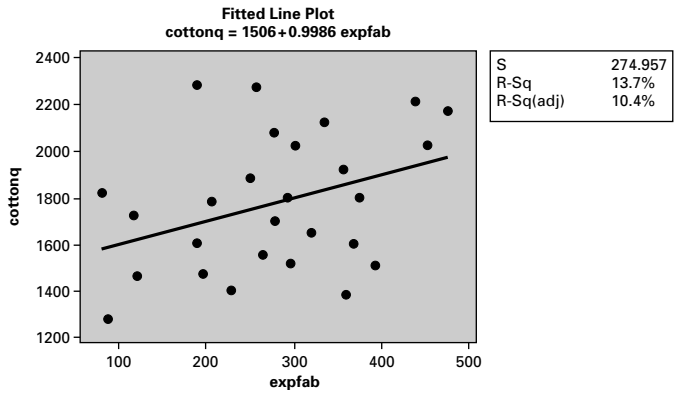
3.44. **a)** Cov $(x, y) = 9{,}96429 \quad r = 0{,}985$
b) $b_1 = 3{,}695$ y $b_0 = -3{,}69536$
c) La ecuación de regresión da una estimación de la influencia de la experiencia adicional en las ventas semanales (en cientos de dólares). Parece que a medida que aumenta la experiencia, las ventas semanales también aumentan. Esta estimación se basa en una experiencia de 2 a 6 años y unas ventas semanales de 400 $ a 2.000 $.

3.46. **a)** 18,1325
b) Varianza muestral = 204,7017, $s = 14{,}307$

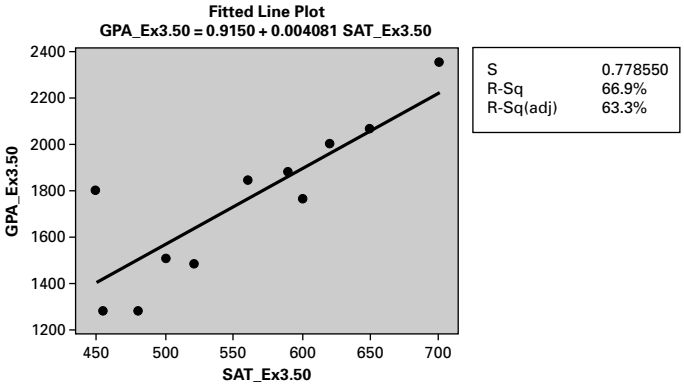
3.48. **a)**



b) $\hat{y} = 142{,}398 - 0{,}0199937x$; el efecto marginal es −0,0199937
c)



3.50. a)



El sentido es positivo y existe una correlación relativamente estrecha ($r = 0{,}818$) entre las dos variables. Existe una relación positiva entre las notas obtenidas en la prueba de matemáticas y la calificación media obtenida al terminar los estudios.

b) $b_1 = 0{,}004081$. Estimamos que por cada aumento de la nota de la prueba de matemáticas de un punto, la calificación media obtenida al terminar los estudios aumenta en 0,004081.

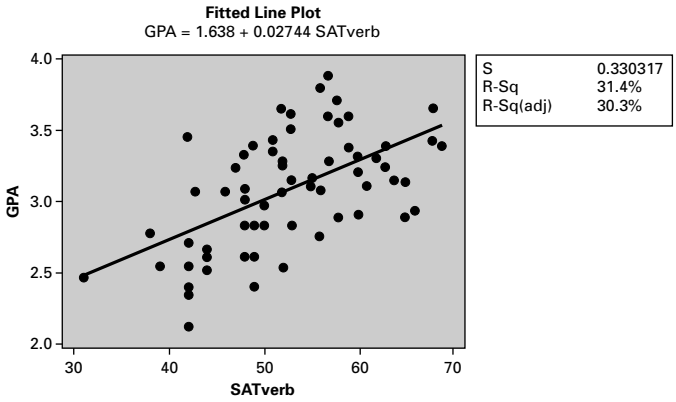
c) $b_0 = 0{,}9150$

d) 3,078

e) Basándose en los datos, ¿es posible predecir cuál será la calificación media obtenida al terminar los estudios si la nota de la prueba de matemáticas es 375?

El valor de 375 de la nota de la prueba de matemáticas está fuera del rango de datos observados. Tendríamos que extrapolar a puntos situados fuera del rango de datos observados para hacer una afirmación sobre la calificación media obtenida al terminar los estudios. Los resultados situados fuera del rango de datos observados son mucho menos fiables.

3.52. a)



b) Describir los datos numéricamente

Covariances: GPA, SATverb

```
                GPA SATverb
GPA        0.169284
SATverb    1.791637  65.293985
```

Correlations: GPA, SATverb

```
Pearson correlation of GPA and SATverb=0.560
P—Value=0.000
```

Regression Analysis: GPA versus SATverb

```
The regression equation is
GPA=1.64 + 0.0274 SATverb
```

c) 3,96

3.54. **a)** De 195,46 a 394,54
b) De 137,50 a 452,50

3.56. **a)** De 23.000 a 35.000
b) De 15.583,59 a 42.416,41

Capítulo 4

4.2. **a)** (E_3, E_9)
b) $(E_1, E_2, E_3, E_7, E_8, E_9)$
c) La unión de A y B no es colectivamente exhaustiva: no contiene todos los puntos muestrales posibles.

4.4. **a)** (E_3, E_6)
b) $(E_3, E_4, E_5, E_6, E_9, E_{10})$
c) La unión de A y B no es colectivamente exhaustiva: no contiene todos los puntos muestrales posibles.

4.6. **a)** $(A \cap B)$ es el suceso de que el índice sube los dos días, que es O_1. $(\bar{A} \cap B)$ es el suceso de que el índice no sube el primer día, pero sube el segundo, que es O_3. La unión de estos dos, o sea, que ocurra el suceso O_1 o O_3, es por definición el suceso B: el índice sube el segundo día.
b) Dado que $(\bar{A} \cap B)$ es el suceso de que el índice no sube el primer día pero sube el segundo, o sea, el suceso es O_3, y como A es el suceso de que el índice sube el primer día, la unión de ambos es O_2, es decir, o bien el índice no sube el primer día pero sube el segundo, o bien el índice sube el primer día o ambos. Ésta es la definición de $A \cup B$.

4.8. 0,53

4.10. 0,3709

4.12. 0,0123

4.14. **a)** 0,54 **b)** 0,18
c) Un complementario es el suceso de que la tasa de rendimiento no es de más del 10%.
d) 0,46
e) La intersección de más de un 10% y el rendimiento será negativo es el conjunto nulo o vacío.
f) 0
g) La unión de A y B es el suceso de que las tasas de rendimiento son de menos de -10%, entre -10% y 0%, entre 10% y 20% y más de 20%.
h) 0,72
i) A y B son mutuamente excluyentes porque su intersección es el conjunto vacío.
j) A y B no son colectivamente exhaustivos porque su unión no es igual a 1

4.16. A y $\bar{A}$ del ejercicio 4.1 no son mutuamente excluyentes. Dado que $P(A) = 0{,}68$ y $P(\bar{A}) = 0{,}75$, compruebe si $P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 0{,}68 + 0{,}75 = 1{,}43 > 1$. Por lo tanto, si dos sucesos no son mutuamente excluyentes, la probabilidad de su unión no puede ser igual a la suma de sus probabilidades.

4.18. **a)** 0,87 **b)** 0,35
c) Por el tercer postulado de la probabilidad, la suma de las probabilidades de todos los resultados del espacio muestral debe ser 1.

4.20. 0

4.22. 0,75

4.24. 0,80, A y B son independientes dado que la $P(A|B)$ de 0,80 es igual a la $P(A)$ de 0,80.

4.26. 0,625, A y B no son independientes, ya que la $P(A|B)$ de 0,625 no es igual a la $P(A)$ de 0,70.

4.28. **a)** 5.040 **b)** 0,0001984

4.30. 0,00833

4.32. 0,0167

4.34. 28

4.36. **a)** 150 **b)** 0,2667 **c)** 0,20

4.38. 0,35

4.40. **a)** No, los dos sucesos no son mutuamente excluyentes porque $P(A \cap B) \neq 0$
b) No, los dos sucesos no son colectivamente exhaustivos porque $P(A \cup B) \neq 1$
c) No, los dos sucesos no son estadísticamente independientes porque $P(A \cap B) = 0{,}15 \neq \neq 0{,}06 = P(A)P(B)$.

4.42. 0,069

4.44. **a)** 0,556 **b)** 0,8333

4.46. 0,1292

4.48. **a)** 0,9 **b)** 0,88 **c)** 0,925

4.50. **a)** 0,867
b) Compruebe si $P(A \cap B) = P(A)P(B)$. Dado que $0{,}04 \neq 0{,}06$, los dos sucesos no son sucesos independientes.

4.52. 0,2

4.54. 0,05

4.56. 0,05

4.58. 0,1667

4.60. 0,40

4.62. Ventaja $= \dfrac{0{,}5}{1 - 0{,}5}$ = ventaja de 1 a 1

4.64. 2,00

4.66. **a)** 0,12
b) 0,7037
c) Compruebe si $P(F \cap N) = P(F)P(N)$. Dado que $0{,}19 \neq 0{,}2133$, los dos sucesos no son independientes.
d) 0,3333
e) Compruebe si $P(I \cap 0) = P(I)P(0)$. Dado que $0{,}07 \neq 0{,}0399$, los dos sucesos no son independientes.
f) 0,79
g) 0,27
h) 0,87

4.68. **a)** 0,25 **b)** 0,32 **c)** 0,16 **d)** 0,125 **e)** 0,2121

4.70. **a)** 0,32 **b)** 0,25 **c)** 0,375 **d)** 0,48 **e)** 0,4375
f) No, ya que $P(A \cap \text{sí}) = 0{,}12 \neq P(A)P(\text{sí}) = 0{,}08$.

4.72. **a)** 0,76 **b)** 0,77 **c)** 0,4348

4.74. **a)** 0,025 **b)** 0,445 **c)** 0,2697

4.76. **a)** 0,475 **b)** 0,3684 **c)** 0,8571

4.78. 0,375

4.80. 0,3636

4.82. 0,6667

4.84. 0,6923

4.86. 0,444

4.88. **a)** Verdadera **b)** Falsa **c)** Verdadera **d)** Verdadera **e)** Verdadera **f)** Verdadera **g)** Falsa

4.90. El teorema de Bayes es un resumen de la relación entre un suceso específico que ha ocurrido y el efecto que produce en un suceso posterior. La ocurrencia del suceso específico es la información *a priori* o «probabilidad *a priori*» que se conoce. Esta información *a priori* puede analizarse para comprender el efecto en la probabilidad de un suceso posterior. El suceso posterior es la «probabilidad *a posteriori*».

4.92. Varias respuestas. Por definición, la *probabilidad conjunta* es la probabilidad de que dos sucesos ocurran conjuntamente, por ejemplo, P(mujer y licenciado en Filología Hispánica).
La *probabilidad marginal* es la probabilidad de un suceso individual, por ejemplo, P(mujer).
La *probabilidad condicionada* es la probabilidad de ocurrencia de un suceso, dado que ha ocurrido otro, por ejemplo, P(mujer dado licenciado en Filología Hispánica).

4.94. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - [P(A|B)P(B)] =$
$= P(A) + P(B)[1 - P(A|B)]$

4.96. **a)** 0,4211 **b)** 0,6316 **c)** 0,2526

4.98. **a)** 0,125 **b)** 0,3571 **c)** 0,875
d) No, ya que $P(PC \cap E)$, que es $0{,}125 \neq 0{,}175$, que es $P(PC)P(E)$.
e) No, su intersección no es cero, por lo que los dos sucesos no pueden ser mutuamente excluyentes. $P(PC \cap E) = 0{,}125 \neq 0$.
f) No, la probabilidad de su unión no es igual a 1. $P(PC \cup E) = P(PC) + P(E) - P(PC \cap E) =$
$= 0{,}35 + 0{,}5 - 0{,}125 = 0{,}725$, que es menor que 1.

4.100. **a)** 0,48 **b)** 0,11 **c)** 0,7273
d) No, compruebe que $P(H \cap PG) = 0{,}8 \neq 0{,}088 = P(H)P(PG)$.
e) 0,191

4.102. **a)** 1.820 **b)** 0,089

4.104. **a)** 0,075 **b)** 0,1429 **c)** 10!90!/100!

4.106. 0,2581

4.108. 0,6364

4.110. **a)** 0,4526 **b)** 0,6632 **c)** 0,9406
d) No, ya que $P(L \cap FP) = (0{,}78)(0{,}27) = 0{,}2106 \neq 0{,}1791 = P(L)P(FP) = (0{,}6632)(0{,}27)$

4.112. **a)** 0,58 **b)** 0,6034 **c)** 0,3966

4.114. 0,0128

4.116. 0,5085

Capítulo 5

5.2. Variable aleatoria discreta

5.4. Variable aleatoria discreta

5.6. Ventas totales, gastos publicitarios, ventas del competidor

5.8. Discreta

5.10. Distribución de probabilidad del número de caras en un lanzamiento

X–número de caras	$P(x)$
0	0,5
1	0,5

5.12. Varias respuestas

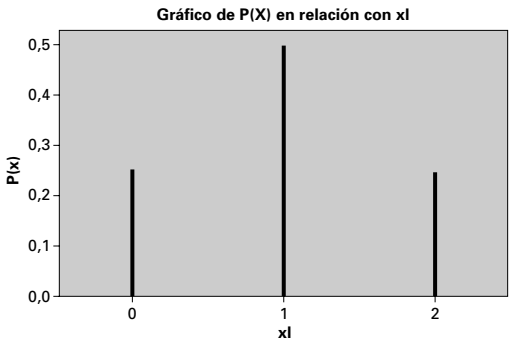
X – n.º de veces que faltará a clase	$P(x)$	$F(x)$
0	0,65	0,65
1	0,15	0,80
2	0,10	0,90
3	0,09	0,99
4	0,01	1,00

5.14. a) Función de probabilidad acumulada

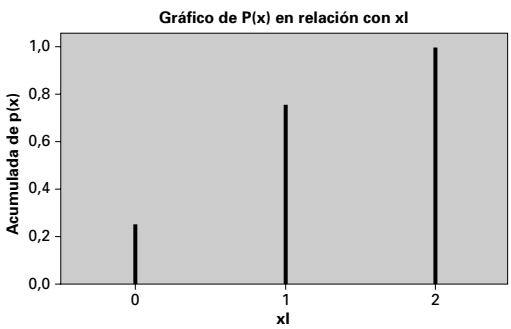
X	0	1	2	3	4	5	6	7	8	9
$P(x)$	0,10	0,08	0,07	0,15	0,12	0,08	0,10	0,12	0,08	0,10
$F(x)$	0,10	0,18	0,25	0,40	0,52	0,60	0,70	0,82	0,90	1,00

b) 0,48 **c)** 0,57

5.16. a) Función de distribución de probabilidad

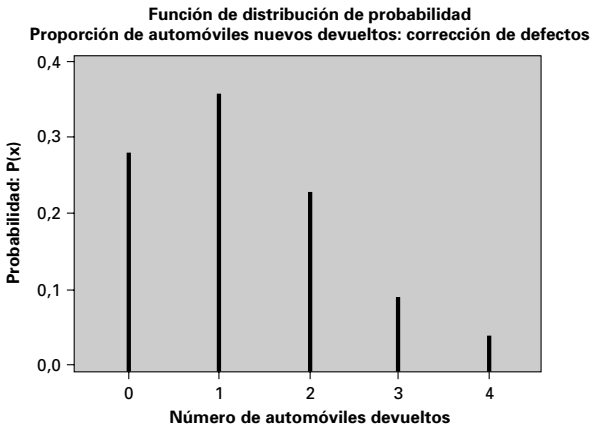


b) Función de probabilidad acumulada

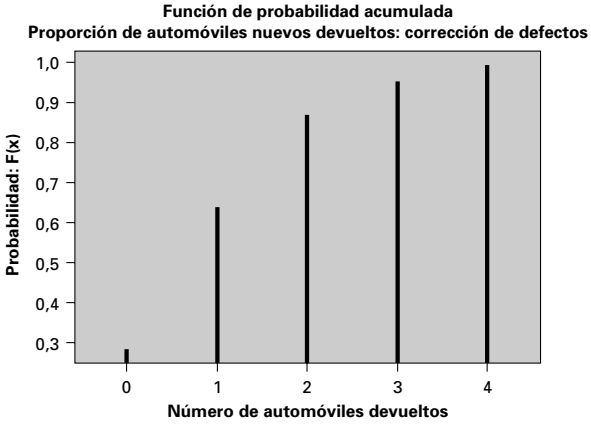


c) $\mu_x = 1{,}00$ **d)** $\sigma_x^2 = 0{,}50$

5.18. a) Función de probabilidad

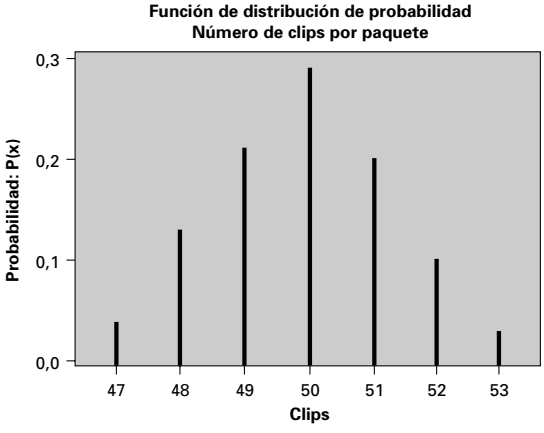


b) Función de probabilidad acumulada

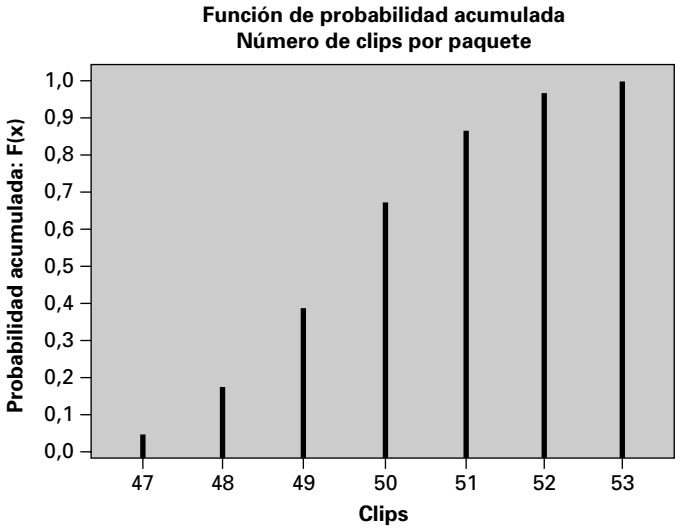


c) 1,25 defectos
d) 1,1675 defectos

5.20. a) Función de probabilidad



b) Función de probabilidad acumulada



c) 0,70
d) 0,8556
e) $\mu = 49{,}9 \quad \sigma_x = 1{,}3964$

Microsoft Excel - Book1

File Edit View Insert Format Tools OIC Data Window Help

W12 =

	M	N	O	P	Q
1	Clips	P(x)	F(x)	Mean	Variance
2	47	0.04	0.04	1.88	0.3364
3	48	0.13	0.17	6.24	0.4693
4	49	0.21	0.38	10.29	0.1701
5	50	0.29	0.67	14.5	0.0029
6	51	0.20	0.87	10.2	0.242
7	52	0.10	0.97	5.2	0.441
8	53	0.03	1.00	1.59	0.2883
9		1.00		49.9	1.95
10					

f) $\mu = 0{,}342$ \$, $\sigma_\pi = 0{,}0279$ \$

5.22. a) Función de probabilidad

X	0	1	2
P(x)	0,81	0,18	0,01

b) $P(Y = 0) = 153/190$
$P(Y = 1) = 36/190$
$P(Y = 2) = 1/190$
La respuesta del apartado b) es diferente de la respuesta del apartado a) porque en el b) la probabilidad de seleccionar una pieza defectuosa la segunda vez depende del resultado obtenido la primera vez.

c) $\mu = 0{,}2$ defectos, $\sigma_x^2 = 0{,}18$
d) $\mu = 0{,}2$ defectos, $\sigma_y^2 = 0{,}1705$

5.24. «Uno más uno» $E(X) = 1{,}3125$
«Falta de dos tiros libres» $E(X) = 1{,}50$
La «falta de dos tiros libres» tiene un valor esperado más alto.

5.26. $\mu = 3{,}29 \quad \sigma = 1{,}1515$

5.28. **a)** $\mu = 1{,}82$, $\sigma = 1{,}0137$
b) Coste: $\mu = 2.730$ \$, $\sigma = 1.520{,}559$ \$

5.30. $\mu_x = 0{,}5$
$\sigma_x^2 = 0{,}25$

5.32. $P(x = 7) = 0{,}06181$, $P(x < 6) = 0{,}7805$

5.34. $P(x = 12) = 0{,}1873$, $P(x < 6) = 0{,}000269$

5.36. **a)** $P(x \geqslant 1) = 0{,}7627$
b) $P(x \geqslant 3) = 0{,}1035$

5.38. $P(x \geqslant 4) = 0{,}5$

5.40. **a)** $P(x = 5) = 0{,}0102$
b) $P(x \geqslant 3) = 0{,}3174$
c) $P(x \geqslant 2) = 0{,}5248$
d) $E(X) = 2$ partidos. A menos, por supuesto, que usted sea un hincha de los Verdes y, en ese caso, *desearía* que los Verdes ganaran todos los partidos, pero *esperaría* que ganaran dos.
e) $E(X) = 2{,}6$ partidos.

5.42. **a)** 0,5248 **b)** $E(X) = 1{,}6$, $\sigma_x = 0{,}9798$

5.44. **a)** $E(X) = 64$, $\sigma_x = 7{,}871$
b) $E(X) = 640$ \$, $\sigma_z = 78{,}71$ \$

5.46. **a)** $E(X) = 483{,}6$, $\sigma_x = 10{,}3146$
b) $E(Z) = 967{,}20$ \$, $\sigma_z = 20{,}6292$ \$

5.48. Las reglas de aceptación tienen las siguientes probabilidades:

(i) Regla 1: $P(X = 0) = (0{,}8)^{10} = 0{,}1074$
(ii) Regla 2: $P(X \leqslant 1) = (0{,}8)^{20} + 20(0{,}2)(0{,}8)^{19} = 0{,}0692$

La segunda regla de aceptación tiene la menor probabilidad de aceptar un envío que contenga un 20% de componentes defectuosos.

5.50. 0,210376

5.52. 0,151769

5.54. 0,1999

5.56. 0,3808

5.58. 0,2619

5.60. 0,1336

5.62. 0,857614

5.64. 0,4232

5.66. 0,7898

5.68. 0,0884

5.70. 0,9380

5.72. Hay dos modelos posibles: la distribución de Poisson es adecuada cuando el almacén recurre a uno de los muchos miles de camioneros independientes, donde el número medio de «éxitos» es relativamente pequeño. Sin embargo, en el caso del supuesto de una pequeña flota de 10 camiones con una probabilidad de 0,1 de que llegue cualquiera durante una hora dada, la distribución binomial es el modelo más adecuado. Ambos modelos dan probabilidades similares, aunque no idénticas.

Función de distribución acumulada

```
Poisson with mean = 1

 x  P( X <= x )
 0      0.36788
 1      0.73576
 2      0.91970
 3      0.98101
 4      0.99634
 5      0.99941
 6      0.99992
 7      0.99999
 8      1.00000
 9      1.00000
10      1.00000
```

Función de distribución acumulada

```
Binomial with n = 10 and p = 0.1

 x  P( X <= x )
 0      0.34868
 1      0.73610
 2      0.92981
 3      0.98720
 4      0.99837
 5      0.99985
 6      0.99999
 7      1.00000
 8      1.00000
 9      1.00000
10      1.00000
```

5.74. a)

		X		
Y		1	2	$P(y)$
	0	0,2	0,25	0,45
	1	0,3	0,25	0,55
$P(x)$		0,5	0,5	1

b) $\text{Cov}(X, Y) = -0{,}025$, $\rho = -0{,}1005$

5.76. a) Calcule las distribuciones de probabilidad marginal de X e Y.

		X		
Y		1	2	$P(y)$
	0	0,3	0,2	0,5
	1	0,25	0,25	0,5
$P(x)$		0,55	0,45	1

b) $\text{Cov}(X, Y) = -0{,}025$, $\rho = -0{,}0326$

c) $\mu_W = 3{,}4$, $\sigma_W^2 = 9{,}75$

5.78. a)

		X		
Y		1	2	$P(y)$
	0	0,25	0,25	0,5
	1	0,25	0,25	0,5
$P(x)$		0,5	0,5	1

b) $\text{Cov}(X, Y) = 0{,}0$, $\rho = 0{,}0$

c) $\rho_W = 2{,}0$, $\sigma_W^2 = 0{,}50$

5.80. a)

		X		
Y		1	2	$P(y)$
	0	0	0,6	0,6
	1	0,4	0	0,4
$P(x)$		0,4	0,6	1

b) $\text{Cov}(X, Y) = -0{,}24$, $\rho = -1{,}00$

c) $\mu_W = 1{,}6$, $\sigma_W^2 = 0{,}48$

5.82. a) $Px(0) = 0{,}22$ $Px(1) = 0{,}26$ $Px(2) = 0{,}43$ $Px(3) = 0{,}09$ $\mu_x = 1{,}39$

b) $Py(0) = 0{,}23$ $Py(1) = 0{,}21$ $Py(2) = 0{,}30$ $Py(3) = 0{,}26$ $\mu_y = 1{,}59$

c) $P_{Y|X}(0|3) = 0{,}1111$ $P_{Y|X}(1|3) = 0{,}1111$
$P_{Y|X}(2|3) = 0{,}3333$ $P_{Y|X}(3|3) = 0{,}4444$

d) $\text{Cov}(X, Y) = 0{,}3399$

e) No, porque $\text{Cov}(X, Y) \neq 0$

5.84. a) $Py(0) = 0{,}12$ $Py(1) = 0{,}24$ $Py(2) = 0{,}23$ $Py(3) = 0{,}23$ $Py(4) = 0{,}18$

b) $P_{Y|X}(y|3) = 1/26;\ 3/26;\ 6/26;\ 8/26;\ 8/26$

c) No, porque $Px, y(3, 4) = 0{,}08 \neq 0{,}0468 = Px(3)Py(4)$

5.86. a)

Y/X	0	1	Total
0	0,704	0,168	0,872
1	0,096	0,032	0,128
Total	0,80	0,20	1,00

b) $P_{Y|X}(y|0) = 0{,}88;\ 0{,}12$

c) $Px(0) = 0{,}80$ $Px(1) = 0{,}20$ $Py(0) = 0{,}872$ $Py(1) = 0{,}128$

d) $\text{Cov}(X, Y) = 0{,}0064$
La covarianza indica que hay una relación positiva entre X e Y; es más probable que los profesores no estén en su despacho el viernes que los demás días.

5.88. El número total de quejas (quejas por la comida + quejas por el servicio) tiene una media de (1,36 + 1,64) = 3,00. Si los dos tipos de quejas son independientes, la varianza del total de quejas es igual a la suma de la varianza de los dos tipos de quejas, ya que la covarianza sería cero. (0,8104 + 0,7904) = 1,6008. La desviación típica es la raíz cuadrada de la varianza = 1,26523.

Si el número de quejas por la comida y por el servicio no son independientes entre sí, la covarianza ya no sería cero. La media seguiría siendo igual; sin embargo, la desviación típica variaría. La varianza de la suma de los dos tipos de quejas se convierte en la varianza de uno más la varianza del otro más el doble de la covarianza.

5.90. **a)** No, no necesariamente. Las tasas de rendimiento del fondo de inversión siguen una determinada distribución de probabilidad y no todas las tasas de rendimiento serán iguales al valor esperado.
b) Depende no sólo del valor esperado del rendimiento sino también del riesgo de cada fondo y de lo renuente al riesgo que sea el cliente.

5.92. **a)** 2,21
b) 1,3513
c) Sueldo medio = 913 \$. Desviación típica del sueldo = 405,39 \$
d) Para ganar 1.000 \$ o más, el vendedor debe vender al menos 3 automóviles.
$P(X \geqslant 3) = 0{,}16 + 0{,}12 + 0{,}07 = 0{,}35$

5.94. **a)** Covarianza positiva: gastos de consumo y renta disponible
b) Covarianza negativa: precio de los automóviles y número de automóviles vendidos
c) Covarianza cero: índice bursátil y precipitaciones en Brasil

5.96. **a)** 0,17
b) $\mu_x = 2{,}59$, $\mu_y = 1{,}1$
c) $\text{Cov}(X, Y) = 0{,}191$. Eso implica que existe una relación positiva entre el número de años en la universidad y el número de visitas a un museo el año anterior.

5.98. **a)** 0,3369 **b)** 0,5931
c) $\mu = 44$. La proporción es 0,55. $\sigma = 4{,}4497$. La proporción es 0,05562

5.100. Para evaluar la eficacia de la capacidad del analista, halle aleatoriamente la probabilidad de que x sea mayor o igual que 3. $P(x \geqslant 3) = 0{,}16683$

5.102. **a)** $P(0) = 0{,}09072$ **b)** $P(x \geqslant 3) = 0{,}2213$

5.104. $P(x = 0) = 0{,}0907$. Sea Y el número de paradas en ambas cadenas. Halle la $P(Y \geqslant 1) = = 0{,}99177$

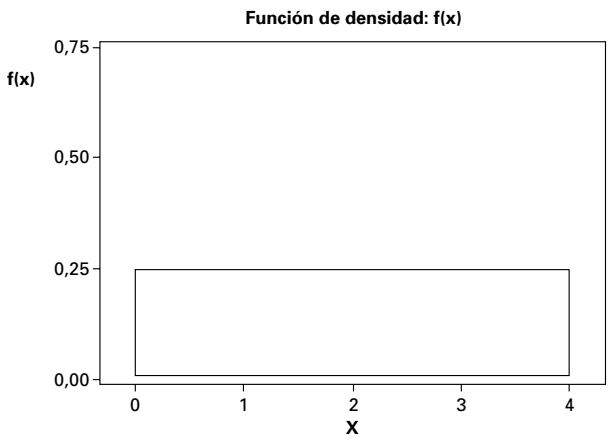
5.106. $\mu_W = 10$, $\sigma_W^2 = 22{,}5$

Capítulo 6

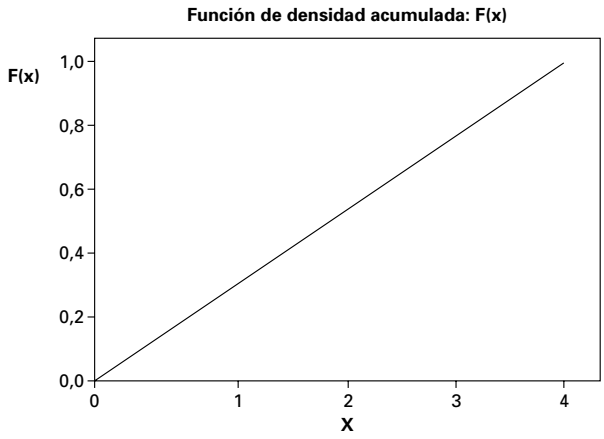
6.2. 0,45

6.4. 0,35

6.6. **a)**



b)



c) 0,25
d) 0,25

6.8. **a)** 0,2 **b)** $0,4 < P(X < 400) < 0,6$

6.10. $\mu_W = 900$, $\sigma_W^2 = 360$

6.12. $\mu_W = 4.000$, $\sigma_W^2 = 8.100$

6.14. $\mu_Y = 26,4$ millones de dólares, $\sigma_Y = 1$ millón de dólares

6.16. $\mu_Y = 54.000$ \$, $\sigma_Y = 14.400$ \$

6.18. **a)** 0,52 **b)** $-0,67$ **c)** 0,84 **d)** $-0,25$

6.20. **a)** 0,9772 **b)** 0,3674 **c)** = 0,0062 **d)** 92,8 **e)** X = 70 y 90

6.22. **a)** 0,6554 **b)** 0,6554
c) El gráfico debe mostrar la propiedad de la simetría: las áreas de las colas equidistantes de la media deben ser iguales.
d) 0,6006
e) El área situada debajo de la curva normal es igual a 0,8 en el caso de un número infinito de intervalos: basta comenzar en un punto que sea marginalmente más alto. El intervalo más corto será el que esté centrado en el valor cero. La z que corresponde a un área de 0,8 centrada en la media es $Z = \pm 1,28$. De este resultado se deduce un intervalo igual a la media más/menos 64 \$, o sea, [36 \$, 444 \$]

6.24. **a)** 0,2266 **b)** 0,2266 **c)** 0,5468
d) (i) El gráfico debe mostrar la propiedad de la simetría: las áreas de las colas equidistantes de la media deben ser iguales.
e) (ii) Las respuestas a los apartados a), b), c) suman uno debido a que los sucesos abarcan toda el área situada debajo de la curva normal que, por definición, debe ser igual a 1.

6.26. **a)** 0,2148 **b)** 0,1587 **c)** 0,3692
d) La respuesta al apartado a) será mayor debido a que 10 gramos está más cerca de la media que 15. Por lo tanto, el área que quedaría por debajo de 10 gramos sería mayor que la que quedaría por encima de 15 gramos.

6.28. 0,668

6.30. $\mu = 15,265$, $\sigma^2 = 14,317$

6.32. En el caso de la inversión A, la probabilidad de que el rendimiento sea de más del 10%

$$P\left(Z > \frac{10 - 10{,}4}{1{,}2}\right) = P(Z > -0{,}33) = F_Z\,(0{,}33) = 0{,}6293$$

En el caso de la inversión B, la probabilidad de que el rendimiento sea de más del 10%

$$P\left(Z > \frac{10 - 11{,}0}{4}\right) = P(Z > -0{,}25) = F_Z\,(0{,}28) = 0{,}5987$$

Por lo tanto, la inversión A es mejor.

6.34. **a)** 98,8 **b)** 183,6 **c)** 0,9487

6.36. **a)** 0,3721 **b)** 522,4 **c)** 400 − 439 **d)** 520 − 559 **e)** 0,2922

6.38. 0,4990

6.40. **a)** 0,0054 **b)** 0,0002 **c)** 0,9892
d) $X = 1.573{,}741 \approx 1.574$ éxitos
e) $X = 1.616{,}46 \approx 1.616$ éxitos

6.42. **a)** 0,000 **b)** 0,0005 **c)** 0,9990
d) $P = 38{,}971\%$ **e)** P = 41,642%

6.44. **a)** 0,0475 **b)** 0,3372

6.46. 0,0207

6.48. 0,2877

6.50. 0,864665

6.52. 0,2019

6.54. 0,4866

6.56. 0,3012

6.58. **a)** $P(X > 3) = 1 - [1 - e^{-(3/\mu)}] = e^{-3\lambda}$, ya que $\lambda = 1/\mu$
b) $P(X > 6) = 1 - [1 - e^{-(6/\mu)}] = e^{-(6/\mu)} = e^{-6\lambda}$
c) $P(X > 6|X > 3) = P(X > 6/P(X > 3) = e^{-6\lambda}/e^{-3\lambda}] = e^{-3\lambda}$
La probabilidad de una ocurrencia en un periodo de tiempo en el futuro no está relacionada con la cantidad de tiempo que ha pasado desde la ocurrencia más reciente.

6.60. $\mu_W = 1.300$, $\sigma_W^2 = 4.900$

6.62. $\mu_W = 1.700$, $\sigma_W^2 = 4.900$

6.64. $\mu_x = 28.000$, $\sigma_x = 12.000$

6.66. $\mu_Y = 162.000$
$\sigma_Y = 18.027{,}76$

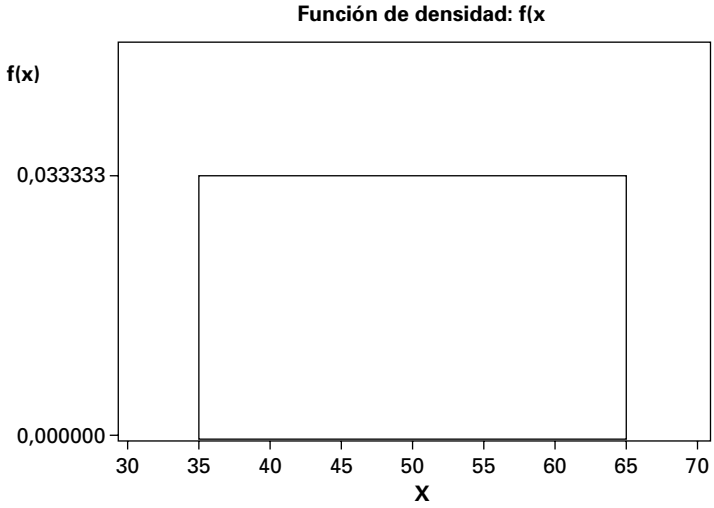
6.68. El cálculo de la media es correcto, pero las desviaciones típicas de dos variables aleatorias no pueden sumarse. Para hallar la desviación típica correcta, se suman las varianzas y se toma la raíz cuadrada. La desviación típica: $\sigma = \sqrt{5(16)^2} = 35{,}7771$

6.70. **a)** $\mu_W = 2.850$, $\sigma_W^2 = 992.500$
b) $\mu_W = 2.850$, $\sigma_W^2 = 332.500$

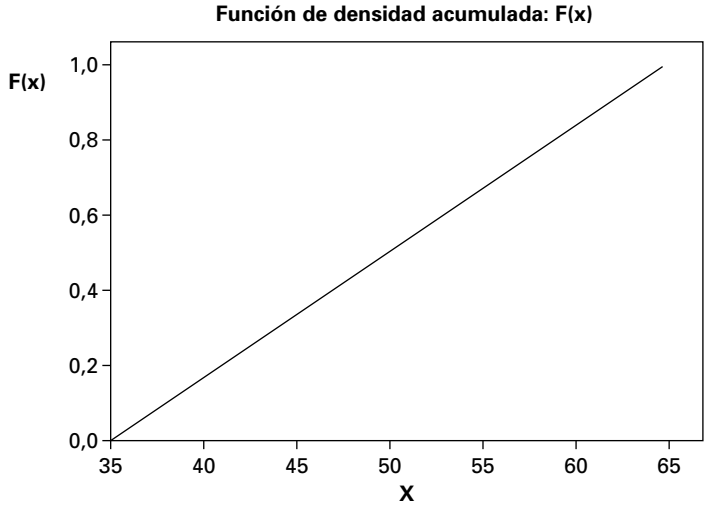
6.72. **a)** $\mu_W = 100$, $\sigma_W^2 = 256{,}90465$
b) 0,3483

6.74. **a)** $\mu_W = -5$, $\sigma_W^2 = 21{,}79449$
b) 0,4090

6.76. **a)**



b) Función de densidad acumulada



c) 10/30
d) 50

6.78. **a)** $\mu_Y = 3.360$ **b)** $\sigma_Y = 80$

6.80. Dado que las varianzas de los beneficios predichos y del error de predicción son ambas positivas y dado que la varianza de los beneficios efectivos es igual a la suma de las varianzas de los beneficios predichos y el error de predicción, la varianza de los beneficios predichos debe ser menor que la varianza de los beneficios efectivos.

6.82. **a)** 0,2119 **b)** 0,3759 **c)** 3,24 **d)** 0,7190 **e)** 0,3789

6.84. **a)** 0,3085
b) 0,6826
c) $X_i = 149{,}35$

d) 0,9916
e) 0,0417
f) 90 − 109
g) 130 − 149

6.86. 0,0436

6.88. $P(Z < 6{,}45) \approx 1{,}0000$

6.90. 0,0084

6.92. **a)** 0,0475

b) 236,95 (237 oyentes)

6.94. 0,975

6.96. **a)** $\mu_W = 200$, $\sigma_W^2 = 3.204{,}919$
b) Opción 1: $\sigma_1^2 = 3.813{,}744$, Opción 2: $\sigma_2^2 = 2.665{,}66$
Para reducir la varianza de la cartera, seleccionar la Opción 2.

6.98. **a)** 0,1020 **b)** 0,2764

Capítulo 7

7.2. **a)**

Función de densidad

```
Binomial with n = 2 and p = 0.5
 x  p( X = x )
 0       0.25
 1       0.50
 2       0.25
```

b)

Función de densidad

```
Binomial with n = 4 and p = 0.5
 x  p( X = x )
 0     0.0625
 1     0.2500
 2     0.3750
 3     0.2500
 4     0.0625
```

c)

Función de densidad

```
Binomial with n = 10 and p = 0.5
 x  P( X = x )
 0    0.000977
 1    0.009766
 2    0.043945
 3    0.117188
 4    0.205078
 5    0.246094
 6    0.205078
 7    0.117188
 8    0.043945
 9    0.009766
10    0.000977
```

7.4. La respuesta debería señalar que se cometerán errores en la realización de un censo de toda la población y en la selección de una muestra. Es posible mejorar la precisión por medio de métodos de muestreo en lugar de realizar un censo completo (véase la referencia bibliográfica Hogan, 90). Utilizando la información muestral, podemos hacer inferencias válidas sobre toda la población sin incurrir en el tiempo y el gasto necesarios para realizar un censo.

7.6. **a)** $\mu_{\bar{x}} = \mu = 100$, $\sigma_{\bar{x}}^2 = 30$
b) 0,0505
c) 0,7337
d) 0,8997

7.8. **a)** $\mu_{\bar{x}} = \mu = 400$, $\sigma_{\bar{x}}^2 = 45{,}7143$
b) 0,0384
c) 0,7016
d) 0,0516

7.10. **a)** $E(\bar{X}) = \mu_{\bar{x}} = 1.200$
b) $\sigma_{\bar{x}}^2 = 17.778$
c) $\sigma_{\bar{x}} = 33{,}33$
d) 0,1292

7.12. **a)** 0,9772 **b)** 0,5762 **c)** 0,3108 **d)** 114.000 \$ − 116.000 \$
e) Incluso cuando las poblaciones no siguen una distribución normal, la distribución de las medias muestrales en el muestreo es normal si la muestra tiene un tamaño n suficiente. Dado que n es $\geqslant 30$, puede suponerse que la distribución de las medias muestrales en el muestreo sigue una distribución normal.

7.14. **a)** $\sigma_{\bar{x}}^2 = 5{,}5$ **b)** 0,9909 **c)** 0,8980 **d)** 0,4329
e) Mayores, mayores, menores. El gráfico muestra que el error típico de las medias muestrales disminuye conforme aumenta el tamaño de la muestra.

7.16. **a)** $\sigma_{\bar{x}} = 4$ **b)** 0,1056 **c)** 0,1587 **d)** 0,4532

7.18. **a)** Diferencia = 0,2632
b) Diferencia = −0,2048
c) Diferencia = ±0,2304

7.20. **a)** $n = 68$ **b)** menor **c)** mayor

7.22. **a)** $N = 20$, factor de corrección $= \dfrac{0}{19}$

$N = 40$, factor de corrección $= \dfrac{20}{39}$

$N = 100$, factor de corrección $= \dfrac{80}{90}$

$N = 1.000$, factor de corrección $= \dfrac{980}{999}$

$N = 10.000$, factor de corrección $= \dfrac{9.980}{9.999}$

b) Cuando el tamaño de la población (N) es igual al tamaño de la muestra (n), entonces la media estimada es igual a la media poblacional y el error típico es cero. Cuando el tamaño de la muestra es relativamente pequeño en comparación con el de la población, el factor de corrección tiende a 1 y es menos importante en el cálculo del error típico.
c) El factor de corrección tiende a un valor de 1 y es cada vez menos importante como factor de modificación cuando el tamaño de la muestra disminuye en relación con el tamaño de la población.

7.24. **a)** 0,2546 **b)** 0,0951 **c)** 0,0086

7.26. **a)** 0,1539 **b)** 0,0122 **c)** 0,8339

7.28. **a)** 0,1112 **b)** 0,0071 **c)** 0,8372

7.30. **a)** 0,424 **b)** 0,00244 **c)** 0,0494 **d)** 0,0618

7.32. **a)** 0,20 **b)** 0,000889 **c)** 0,0298 **d)** 0,0465

7.34. 0,7372

7.36. **a)** 0,0351 **b)** 0,9222 **c)** 0,4314 **d)** Mayores, mayores

7.38. σ_p alcanza su valor más alto cuando $p = 0{,}5$. En este caso, $\sigma_p = \sqrt{\frac{(0{,}5)(0{,}5)}{100}} = 0{,}05$

7.40. **a)** 0,0395
b) Diferencia = 0,0506
c) Diferencia = 0,065
d) Diferencia = 0,0409

7.42. 0,0057

7.44. **a)** 0,03934 **b)** 0,0384 **c)** 0,0054

7.46. $P(Z > 4{,}61) \approx 0{,}0000$

7.48. **a)** 0,1587
b) $s^2 < 57{,}702$
c) $s^2 > 151{,}879$

7.50. Entre 0,01 y 0,025 (0,0201 exactamente)

7.52. **a)** Algo superior a 0,1 (0,1187 exactamente)
b) Entre 0,01 y 0,025 (0,0118 exactamente)

7.54. **a)** Entre 0,025 y 0,05 (0,0428 exactamente)
b) Inferior a 0,005 (0,0004 exactamente)

7.56.

Estadísticos descriptivos: C1, C2, C3, C4, C5, C6, C7, C8, …

Variable	Media	Varianza
C1	4.500	3.667
C2	4.75	4.92
C3	5.00	6.67
C4	4.75	4.92
C5	5.00	6.67
C6	5.25	7.58
C7	5.25	4.92
C8	5.50	6.33
C9	5.75	6.92
C10	5.75	6.92
C11	5.750	1.583
C12	6.000	2.667
C13	6.250	2.917
C14	6.250	2.917
C15	6.750	0.917

$$\bar{x} = \frac{70{,}518}{15} = 4{,}7012 \quad E(s^2) = {}^{15(3{,}91667)}\!/_{(14)} = 4{,}1964$$

que no es igual a $\sigma^2 = \frac{47}{12} = 3{,}91667$

7.58. **a)** 163,11%
b) La probabilidad de que la varianza muestral esté comprendida entre 30 y 211,33% de la varianza poblacional es 0,95.
c) El intervalo del apartado b) es menor.

7.60. **a)** 41,55%
b) 50,73%
c) La probabilidad de que la varianza muestral esté comprendida entre 34,727 y 199,27% de la varianza poblacional es 0,95.

7.62. Inferior a 0,90 (0,5438 exactamente)

7.64. **a)** 15 muestras posibles
b) (41, 39), (41, 35), (41, 35), (41, 33), (41, 38), (39, 35), (39, 35), (39,33), (39, 38), (35, 35), (35, 33), (35, 38), (35, 33), (35, 38), (33, 38)
c) $\frac{2}{15}$ para 34 y 36,5

$\frac{1}{15}$ para el resto

d) $34P_{\bar{X}}(34) = 34\,\frac{2}{15} = 4{,}5333$ $\quad$ $37P_{\bar{X}}(37) = 37\,\frac{3}{15} = 7{,}4$

$35P_{\bar{X}}(35) = \frac{35}{15} = 2{,}3333$ $\quad$ $38P_{\bar{X}}(38) = 38\,\frac{2}{15} = 5{,}0667$

$35{,}5P_{\bar{X}}(35{,}5) = \frac{35{,}5}{15} = 2{,}3667$ $\quad$ $38{,}5P_{\bar{X}}(38{,}5) = \frac{38{,}5}{15} = 2{,}5667$

$36P_{\bar{X}}(36) = \frac{36}{15} = 2{,}4$ $\quad$ $39{,}5P_{\bar{X}}(39{,}5) = \frac{39{,}5}{15} = 2{,}6333$

$36{,}5P_{\bar{X}}(36{,}5) = 36{,}5\,\frac{2}{15} = 4{,}8667$ $\quad$ $40P_{\bar{X}}(40) = \frac{40}{15} = 2{,}6667$

La media de la distribución de la media muestral en el muestreo es $\Sigma\,\bar{x}P_{\bar{X}}(\bar{x}) = 36{,}8333$, que es exactamente igual que la media poblacional: $\frac{1}{N}\,\Sigma x_i = 36{,}8333$. Éste es el resultado que cabría esperar teniendo en cuenta el teorema del límite central.

7.66. **a)** 0,0668 **b)** 0,7745 **c)** 445,6 **d)** 394,4
e) $s_x = 123{,}1868$ **f)** $s_x = 75{,}966$ **g)** Menor

7.68. **a)** 0,0228 **b)** 0,9544 **c)** $X_{\text{i}} = 13{,}3825$
d) $s_x = 8{,}1414$ **e)** Menor

7.70. Sea $n = N$; entonces $\bar{X} = \mu_x$:

$$E\left[\sum_{i-1}^{N}(X_i - \bar{X})^2\right] = n\sigma_x^2 - n\,\frac{\sigma_x^2}{n}\,\frac{N-n}{N-1} = n\sigma_x^2 - \frac{N-n}{N-1}\sigma_x^2 =$$

$$= \frac{\sigma_x^2}{N-1}\,(nN - n - N + n) = \frac{N\sigma_x^2}{N-1}\,(n-1)$$

Por lo tanto, $E\left[\frac{1}{n-1}\sum(X_i - \bar{X})^2\right] = \frac{1}{n-1}\,E\left[\sum(X_i - \bar{X})^2\right] = \frac{N\sigma_x^2}{N-1}$

7.72. **a)** 0,0262 **b)** 0,3446 **c)** 0,2709 **d)** 0,3210

7.74. 0,005

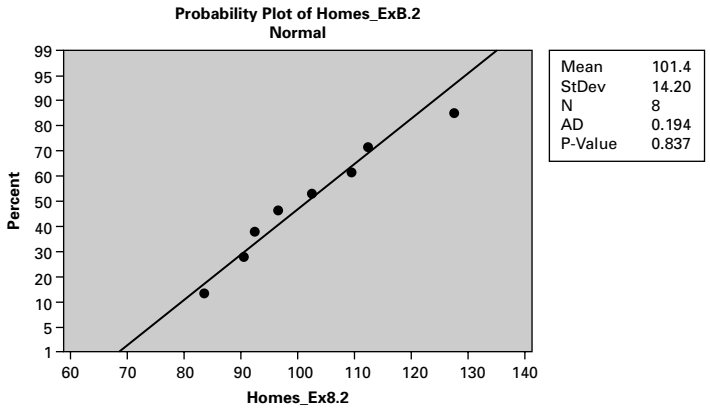
7.76. 0,6826

7.78. **a)** 0,3739 **b)** 0,4397 **c)** Diferencia $= \pm 0{,}0322$

7.80. **a)** Superior a 0,99 (0,9979 exactamente)
b) Entre 0,9 y 0,95 (0,9354 exactamente)

Capítulo 8

8.2. a)



No hay nada que indique la ausencia de normalidad.

b) El estimador puntual insesgado de varianza mínima de la media poblacional es la media muestral: $\bar{x} = 101{,}375$

c) El estimador puntual insesgado de la varianza de la media muestral: $s^2 = 201{,}6964$
$\text{Var}(\bar{X}) = 25{,}2121$

d) $\hat{p} = 0{,}375$

8.4. a) Estimador puntual insesgado de la media poblacional es la media muestral: $\bar{x} = 24{,}42$

b) Estimador puntual insesgado de la varianza poblacional: $s^2 = 85{,}72$

c) Estimador puntual insesgado de la varianza de la media muestral: $\text{Var}(\bar{X}) = 7{,}1433$

d) Estimador insesgado de la proporción poblacional: $\hat{p} = 0{,}25$

e) Estimador insesgado de la varianza de la proporción poblacional: $\text{Var}(\hat{p}) = 0{,}015625$

8.6. a) $E(\bar{X}) = \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) = \frac{\mu}{2} + \frac{\mu}{2} = \mu$

$E(Y) = \frac{1}{4}E(X_1) + \frac{3}{4}E(X_2) = \frac{\mu}{4} + \frac{3\mu}{4} = \mu$

$E(Z) = \frac{1}{3}E(X_1) + \frac{2}{3}E(X_2) = \frac{\mu}{3} + \frac{2\mu}{3} = \mu$

b) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{4}\text{Var}(X_1) + \frac{1}{4}\text{Var}(X_2) = \frac{\sigma^2}{4}$

$\text{Var}(Y) = \frac{1}{16}\text{Var}(X_1) + \frac{9}{16}\text{Var}(X_2) = \frac{5\sigma^2}{8}$

$\text{Var}(Z) = \frac{1}{9}\text{Var}(X_1) + \frac{4}{9}\text{Var}(X_2) = \frac{5\sigma^2}{9}$

$\bar{X}$ es el estimador de máxima eficiencia, ya que $\text{Var}(\bar{X}) < \text{Var}(Y)$ y $\text{Var}(\bar{X}) < \text{Var}(Z)$

c) Eficiencia relativa entre Y y $\bar{X}$: $\frac{\text{Var}(Y)}{\text{Var}(\bar{X})} = 2{,}5$

Eficiencia relativa entre Z y $\bar{X}$: $\frac{\text{Var}(Z)}{\text{Var}(\bar{X})} = 2{,}222$

8.8. **a)** No hay ninguna prueba de que la distribución de los datos proceda de una población que no sigue una distribución normal.
b) El estimador puntual insesgado de varianza mínima de la media poblacional es la media muestral: $\bar{x} = 3{,}8079$
c) El estimador puntual insesgado de varianza mínima de la varianza poblacional es la varianza muestral $s^2 = 0{,}0105$

8.10. **a)** 3,495 **b)** 23,552

8.12. **a)** De 40,2 a 59,8 **b)** De 81,56 a 88,44 **c)** De 506,2652 a 513,73478

8.14. **a)** 1,75 **b)** 0,63246 **c)** 2,2136

8.16. **a)** De 3,9926 a 4,1474 **b)** Menos amplitud
c) Menos amplitud **d)** Más amplitud

8.18. **a)** 13,9182 **b)** 19,007 **c)** 3,493 **d)** 7,5407

8.20. **a)** De 541,424 a 578,576 **b)** De 156,28 a 163,72 **c)** De 49,9474 a 66,9526

8.22. **a)** 83,9685 **b)** 24,1428 **c)** 34,22

8.24. **a)** De 519,379 a 522,517 **b)** Menor

8.26. 5,9152

8.28. De 41.104,28 $ a 44.375,72 $

8.30. **a)** 0,02898 **b)** 0,03761 **c)** 0,010897

8.32. **a)** De 0,079055 a 0,120945 **b)** De 0,0 a 0,031696 **c)** De 0,4555 a 0,5445

8.34. De 0,5846 a 0,8260

8.36. Intervalo de confianza al 95%: de 0,2026 a 0,2974

8.38. 84,14%

8.40. De 0,1079 a 0,2173

8.42. **a)** De 235,4318 a 278,5628
Se supone que la población sigue una distribución normal
b) [95%]: de 230,39 a 283,61
[98%]: de 223,815 a 290,185

8.44. De 29,0229 a 30,9771

8.46. **a)** De 0,0613 a 0,2721 **b)** De 0,0782 a 0,2552

8.48. 95,96%

8.50. **a)** La media muestral $\bar{x} = 8{,}545$ **b)** 1,3568

8.52. **a)** 0,03098
b) Margen de error de un intervalo de confianza al 95% = 0,0607
c) De 0,3457 a 0,4542

8.54. **a)** De 0,5392 a 0,71742 **b)** De 0,1609 a 0,3523
c) De 0,6093 a 0,7535

8.56. **a)** $\bar{x} = 50{,}48$
Se supone que el nivel de confianza es del 95%: de 48,19 a 52,77 años
b) De 0,0267 a 0,1173
c) Se estima la media poblacional por medio de la media muestral = 52,65 $.

8.58. 99,38%

8.60. **a)** 0,9623 **b)** Menor

Capítulo 9

9.2. **a)** 0,79772 **b)** De −2,59766 a −1,00234 **c)** 2,07737

9.4. $27{,}0649 < \mu_x - \mu_y < 47{,}5351$

9.6. De 5,4831 a 14,5169

9.8. **a)** 33,25 **b)** 33,2727 **c)** 21,2105

9.10. **a)** 2,3204 **b)** 3,5026
c) Duplicando el tamaño de las dos muestras, se reduce el margen de error; sin embargo, no se reduce a la mitad.

9.12. De 5,0579 a 9,3421

9.14. De −0,00591 a 0,061907

9.16. **a)** 0,083367 **b)** 0,063062 **c)** 0,056126

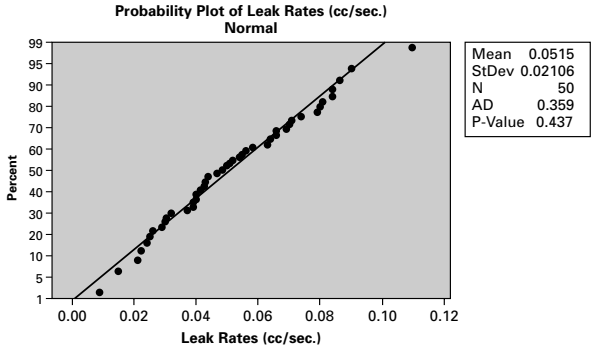
9.18. De 0,0971 a 0,3625

9.20. De −0,3001 a −0,0627

9.22. De −0,314 a 0,0816

9.24. **a)** $9{,}8332 < \sigma^2 < 35{,}036$ **b)** $34{,}9218 < \sigma^2 < 153{,}3546$
c) $126{,}9138 < \sigma^2 < 533{,}446$

9.26. No hay ninguna prueba de la ausencia de normalidad



$3{,}279\text{E-}4 < \sigma^2 < 7{,}238\text{E} - 4$

9.28. De 3,8289 a 14,1167. Se supone que la población sigue una distribución normal.

9.30. **a)** De 2,9852 a 13,8498 **b)** Mayor

9.32. **a)** 427 **b)** 107
c) Para reducir el ME a la mitad, debe cuadruplicarse el tamaño de la muestra.

9.34. **a)** 666 **b)** 271
c) Para aumentar el intervalo de confianza para un margen de error dado, debe aumentarse el tamaño de la muestra.

9.36. 666

9.38. De 25,4893 a 54,5107

9.40. **a)** De −60,21056 a −19,7894 **b)** De −60,669 a −19,331

9.42. De −6,2971 a 2,8971

9.44. De −0,04136 a 0,14295

9.46. De 6,055 a 13,945. Se supone que las dos poblaciones siguen una distribución normal con varianzas iguales y un nivel de confianza del 90%. Dado que los dos límites del intervalo de confianza son positivos, eso es una prueba de que el peso medio de las botellas llenadas con la nueva máquina es mayor que el de las botellas llenadas con la antigua.

9.48. De $-1{,}18066$ a 10,18066

9.50. De 0,23915 a 0,36085

Capítulo 10

10.2. H_0: no está justificada la modificación de los tipos de interés.
H_1: bajar los tipos de interés para estimular la economía.

10.4. **a)** Punto de vista de los europeos:
H_0: los alimentos modificados genéticamente no son seguros.
H_1: son seguros.

b) Punto de vista de los estadounidenses:
H_0: los alimentos modificados genéticamente son seguros.
H_1: no son seguros.

10.6. **a)** Rechazar H_0 si $\bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n} = 108{,}225$
b) Rechazar H_0 si $\bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n} = 110{,}28125$
c) Rechazar H_0 si $\bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n} = 106{,}1998$
d) Rechazar H_0 si $\bar{x} > \bar{x}_c = \mu_0 + z_\alpha \sigma/\sqrt{n} = 107{,}26994$

10.8. El valor crítico $\bar{x}_c$ está más lejos del valor planteado como hipótesis cuanto menor es el tamaño de la muestra n, debido a que el error típico aumenta a medida que el tamaño de la muestra es menor. El valor crítico $\bar{x}_c$ está más lejos del valor planteado como hipótesis cuanto mayor es la varianza poblacional, debido a que el error típico aumenta a medida que es mayor la varianza poblacional.

10.10. **a)** 0,0004 **b)** 0,0475 **c)** 0,0062 **d)** 0,020

10.12. $H_0 : \mu \geqslant 50$; $H_1 : \mu < 50$; rechazar H_0 si $Z_{0,10} < -1{,}28$

$$Z = \frac{48{,}2 - 50}{3/\sqrt{9}} = -1{,}8,$$ por lo tanto, rechazar H_0 al nivel del 10%.

10.14. **a)** Rechazar si $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\,\alpha/2}$, $t = 2{,}00$. Dado que 2,00 es mayor que el valor crítico de 1,711, hay suficientes pruebas para rechazar la hipótesis nula.

b) Rechazar si $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1,\,\alpha/2}$, $t = 2{,}00$. Dado que 2,00 es mayor que el valor crítico de 1,711, hay suficientes pruebas para rechazar la hipótesis nula.

c) Rechazar si $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\,\alpha/2}$, $t = -2{,}50$. Dado que $-2{,}50$ es menor que el valor crítico de $-1{,}711$, hay suficientes pruebas para rechazar la hipótesis nula.

d) Rechazar si $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\,\alpha/2}$, $t = -2{,}22$. Dado que $-2{,}22$ es menor que el valor crítico de $-1{,}711$, hay suficientes pruebas para rechazar la hipótesis nula.

10.16. $H_0: \mu \geqslant 3$; $H_1: \mu < 3$;
$Z = \dfrac{2{,}4 - 3}{1{,}8/\sqrt{100}} = -3{,}33$, p-valor = 0,004; por lo tanto, rechazar H_0 a niveles de significación inferiores a 0,04%; $\alpha = 0{,}04$.

10.18. $H_0: \mu = 0$; $H_1: \mu \neq 0$;
$Z = \dfrac{0{,}078 - 0}{0{,}201/\sqrt{76}} = 3{,}38$, p-valor = 0,0008; por lo tanto, rechazar H_0 a niveles de significación inferiores a 0,08%; $\alpha = 0{,}08$.

10.20. $H_0: \mu = 0$; $H_1: \mu < 0$;
$Z = \dfrac{-2{,}91 - 0}{11{,}33/\sqrt{170}} = -3{,}35$, p-valor = 0,0004; por lo tanto, rechazar H_0 a cualquier nivel habitual de alfa.

10.22. **a)** No, el nivel de confianza del 95% implica que el 2,5% del área se encuentra en cualquiera de las dos colas. Esta situación no corresponde a un contraste de hipótesis de una cola con una alfa del 5% en el que hay un 5% del área en una de las colas.
b) Sí.

10.24. $H_0: \mu = 20$; $H_1: \mu \neq 20$; rechazar H_0 si $|t_{8,\ 0{,}05/2}| > 2{,}306$
$t = \dfrac{20{,}3556 - 20}{0{,}6126/\sqrt{9}} = 1{,}741$; por lo tanto, no rechazar H_0 al nivel del 5%.

10.26. Debe suponerse que los valores poblacionales siguen una distribución normal.
$H_0: \mu \geqslant 50$; $H_1: \mu < 50$; rechazar H_0 si $t_{19,\ 0{,}05} < -1{,}729$
$t = \dfrac{41{,}3 - 50}{12{,}2/\sqrt{20}} = -3{,}189$; por lo tanto, no rechazar H_0 al nivel del 5%.

10.28. **a)** 0,2907 **b)** 0,30427 **c)** 0,28256 **d)** 0,2771

10.30. $H_0: p \leqslant 0{,}25$; $H_1: p > 0{,}25$;
$z = 1{,}79$, p-valor = 0,0367; por lo tanto, rechazar H_0 a una alfa mayor que 3,67%.

10.32. $H_0: p = 0{,}5$; $H_1: p \neq 5$;
$z = -1{,}26$, p-valor = 0,2076. La probabilidad de encontrar una muestra aleatoria cuya proporción muestral esté tan lejos de 0,5 o más si la hipótesis nula es realmente verdadera es 0,2076.

10.34. $H_0: p = 0{,}5$; $H_1: p > 0{,}5$;
$z = 0{,}85$, p-valor, 0,1977; por lo tanto, rechazar H_0 a niveles de alfa superiores a 19,77%.

10.36. $H_0: p \geqslant 0{,}75$; $H_1: p < 75$;
$z = -1{,}87$, p-valor = 0,0307; por lo tanto, rechazar H_0 a los niveles de alfa superiores a 3,07%.

10.38. **a)** 0,8349 **b)** 0,0233 **c)** 0,6876 **d)** 0,8349 **e)** 0,0694

10.40. **a)** Se rechaza H_0 cuando $\dfrac{\bar{X} - 3}{4/\sqrt{64}} > 1{,}645$ o cuando $\bar{X} > 3{,}082$. Dado que la media muestral es 3,07%, que es menor que el valor crítico, la decisión es no rechazar la hipótesis nula.
b) $\beta = 0{,}3594$. La potencia del contraste $= 1 - \beta = 0{,}6406$

10.42. Se rechaza H_0 cuando $\dfrac{p - 0{,}5}{\sqrt{0{,}25/802}} < -1{,}28$ o cuando $p < 0{,}477$
La potencia del contraste $= 1 - \beta = 0{,}9382$

10.44. **a)** Se rechaza H_0 cuando $-1{,}645 > \dfrac{p - 0{,}5}{\sqrt{0{,}25/199}} > 1{,}645$ o cuando $0{,}442 > p > 0{,}558$. Dado que la proporción muestral es 0,5226, que está dentro de los valores críticos, la decisión es no rechazar la hipótesis nula.
b) $\beta = 0{,}1131$.

10.46. **a)** $\alpha = P(Z > 1{,}33) = 0{,}0918$
b) $\alpha = P(Z > 2{,}67) = 0{,}0038$. Obsérvese que el mayor tamaño de la muestra hace que el error típico de la media sea menor.
c) $\beta = 0{,}0668$
d) i) menor, ii) mayor

10.48. El p-valor indica la probabilidad de que el resultado muestral esté tan lejos del valor postulado como hipótesis como el obtenido, suponiendo que la distribución está centrada realmente en la hipótesis nula. Cuanto menor es el p-valor, más contundentes son las pruebas en contra de la hipótesis nula.

10.50 **a)** Falsa **b)** Verdadera **c)** Verdadera **d)** Falsa
e) Falsa **f)** Verdadera **g)** Falsa

10.52. **a)** $\alpha = P(Z < -2) = 0{,}0228$ **b)** $\beta = P(Z > 3) = 0{,}0014$
c) i) menor, ii) menor **d)** i) menor, ii) mayor

10.54. $H_0 : p = 0{,}5$; $H_1 : p \neq 0{,}5$;
$z = -0{,}39$, p-valor = 0,6966; por lo tanto, rechazar H_0 a los niveles superiores a 69,66%.

10.56. $H_0 : p \leqslant 0{,}25$; $H_1 : p > 0{,}25$; rechazar H_0 si $z_{0{,}05} > 1{,}645$
$z = 2{,}356$; por lo tanto, rechazar H_0 al nivel del 5%.

10.58. Modelo de costes, donde W = coste total; $W = 1.000 + 5X$
$\mu_W = 1.000 + 5(400) = 3.000$

$\sigma_W^2 = (5)^2(625) = 15.625$, $\sigma_W = 125$, $\sigma_{\bar{W}} = \dfrac{125}{\sqrt{25}} = 25$

$H_0 : W \leqslant 3.000$; $H_1 : W > 3.000$;

Utilizando los criterios de los estadísticos de contraste: $(3.050 - 3.000)/25 = 2{,}00$, lo que da un p-valor de 0,0228; por lo tanto, rechazar H_0 al nivel del 0,05.

Utilizando los criterios de los estadísticos muestrales: $\bar{X}_{\text{crit}} = 3.000 + (25)(1{,}645) = 3.041{,}1$, $\bar{X}_{\text{calc}} = 3.050$, dado que $\bar{X}_{\text{crit}} = 3.050 > \bar{X}_{\text{crit}} = 3.041{,}1$; por lo tanto, rechazar H_0 al nivel de 0,05.

10.60. Suponer que la población de diferencias pareadas sigue una distribución normal
$H_0 : \mu_x - \mu_y = 0$; $H_1 : \mu_x - \mu_y \neq 0$;
$t = 1{,}961$; por lo tanto, rechazar H_0 al nivel del 10% ya que $1{,}96 > 1{,}796 = t_{(11,\ 0{,}05)}$

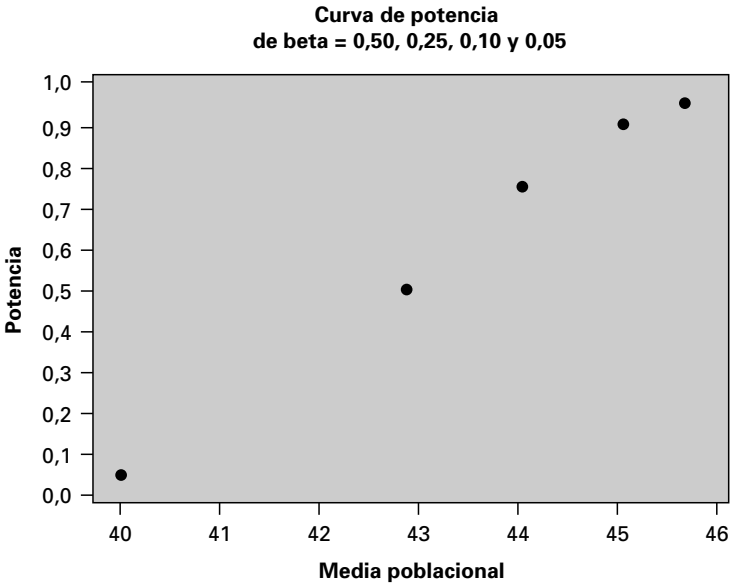
10.62. $H_0 : \mu \leqslant 40$, $H_1 : \mu > 40$; $\bar{X} = 49{,}73 > 42{,}86$ rechazar H_0

One-Sample T: Salmon Weight

```
Test of mu = 40 vs mu > 40

Variable          N      Mean     StDev   SE Mean
Salmon Weigh     39     49.73     10.60      1.70

Variable      95.0% Lower Bound         T       P
Salmon Weigh              46.86      5.73   0.000
```

Rechazar la hipótesis nula y aceptar la alternativa de que el peso medio es significativamente superior a 40
$\bar{X}_{\text{crit}} = H_0 + t_{\text{crit}}(S_{\bar{x}}) = 42{,}8662$

Media poblacional de $\beta = 0{,}50$ (potencia = 0,50): $t_{crit} = 0{,}0$: 42,8662
Media poblacional de $\beta = 0{,}25$ (potencia = 0,75): $t_{crit} = 0{,}681$: 44,0239
Media poblacional de $\beta = 0{,}10$ (potencia = 0,90): $t_{crit} = 1{,}28$: 45,0422
Media poblacional de $\beta = 0{,}05$ (potencia = 0,95): $t_{crit} = 1{,}645$: 45,6627



Capítulo 11

11.2. **a)** $t = -1{,}50$, p-valor = 0,073; no rechazar H_0 a una alfa de 0,05
b) $t = -1{,}00$, p-valor = 0,164; no rechazar H_0 a una alfa de 0,05
c) $t = -2{,}00$, p-valor = 0,028
d) $t = -0{,}75$, p-valor = 0,230

11.4. $z = 7{,}334$; rechazar H_0 a todos los niveles habituales de alfa.

11.6. $z = -1{,}0207$, p-valor = 0,3078; rechazar H_0 a los niveles de alfa superiores a 30,78%.

11.8. $t = 1{,}108$; no rechazar H_0 al nivel de alfa del 10%, ya que $1{,}108 < 1{,}645 = t_{(119,\ 0{,}05)}$

11.10. $t = 2{,}239$, p-valor = 0,0301; rechazar H_0 a los niveles superiores a 3%.

11.12. **a)** $z = -2{,}65$, p-valor = 0,004; rechazar H_0 a todos los niveles habituales de alfa.
b) $z = -1{,}36$, p-valor = 0,0869; rechazar H_0 al nivel de 0,10, pero no al nivel de alfa de 0,05.
c) $z = -2{,}32$, p-valor = 0,0102; rechazar H_0 al nivel de 0,05, pero no al nivel de alfa de 0,01.
d) $z = -3{,}25$, p-valor = 0,0006; rechazar H_0 a todos los niveles habituales de alfa.
e) $z = -1{,}01$, p-valor = 0,1562; no rechazar H_0 a ningún nivel habitual de alfa.

11.14. $z = -6{,}97$; rechazar H_0 a todos los niveles habituales de alfa.

11.16. $z = 2{,}465$; rechazar H_0 al nivel del 5%.

11.18. $z = 0{,}926$; no rechazar H_0 al nivel del 5%.

11.20. **a)** $\chi^2 = 39{,}6$, $\chi^2_{(24,\ 0{,}025)} = 39{,}36$, $\chi^2_{(24,\ 0{,}010)} = 42{,}98$; rechazar H_0 al nivel del 2,5%, pero no al nivel de significación del 1%.
b) $\chi^2 = 46{,}2 = 46{,}2$, $\chi^2_{(28,\ 0{,}025)} = 44{,}46$, $\chi^2_{(28,\ 0{,}010)} = 48{,}28$; rechazar H_0 al nivel del 2,5%, pero no al nivel de significación del 1%.
c) $\chi^2 = 38{,}16$, $\chi^2_{(24,\ 0{,}050)} = 36{,}42$, $\chi^2_{(24,\ 0{,}025)} = 39{,}36$; rechazar H_0 al nivel del 5%, pero no al nivel de significación del 2,5%.

d) $\chi^2 = 24{,}79 = 24{,}79$, $\chi^2_{(30,\ 0{,}100)} = 40{,}26$, $\chi^2_{(40,\ 0{,}100)} = 51{,}81$; no rechazar H_0 a ningún nivel habitual de significación.

11.22. **a)** $s^2 = 5{,}1556$
b) $\chi^2 = 20{,}6224$; rechazar H_0 si $\chi^2_{(9,\ 0{,}05)} > 16{,}92$; rechazar H_0 al nivel del 5%.

11.24. El contraste de hipótesis supone que los valores poblacionales siguen una distribución normal Rechazar H_0 si $\chi^2_{(19,\ 0{,}05)} > 30{,}14$. $\chi^2 = 26{,}4556$; no rechazar H_0 al nivel del 5%.

11.26 **a)** $F = 2{,}451$; rechazar H_0 al nivel del 1%, ya que $2{,}451 > 2{,}11 \approx F_{(44,\ 40,\ 0{,}01)}$
b) $F = 1{,}88$; rechazar H_0 al nivel del 5%, ya que $1{,}88 > 1{,}69 \approx F_{(43,\ 44,\ 0{,}05)}$
c) $F = 2{,}627$; rechazar H_0 al nivel del 1%, ya que $2{,}627 > 2{,}11 \approx F_{(47,\ 40,\ 0{,}01)}$
d) $F = 1{,}90$; rechazar H_0 al nivel del 5%, ya que $1{,}90 > 1{,}79 \approx F_{(24,\ 38,\ 0{,}05)}$

11.28. Rechazar H_0 si $F_{(3{,}6,\ 0{,}05)} > 4{,}76$, $F = 7{,}095$; rechazar H_0 al nivel del 5%.

11.30. $F = 1{,}57$; no rechazar H_0 al nivel del 10%, ya que $1{,}57 < 3{,}18 \approx F_{(9,\ 9,\ 0{,}05)}$

11.32. No. La probabilidad de rechazar la hipótesis nula, dado que es verdadera, es del 5%.

11.34. **a)** Suponer que la población sigue una distribución normal. Rechazar H_0 si $|t_{11,\ 0{,}025)}| > 2{,}201$ $t = -1{,}018$; no rechazar H_0 al nivel del 5%.
b) Suponer que la población sigue una distribución normal. Rechazar H_0 si $\chi^2_{(11,\ 0{,}05)} > 19{,}68$, $\chi^2 = 154{,}19$; rechazar H_0 al nivel del 5%.

11.36. Suponiendo que las varianzas poblacionales son iguales, $t = 1{,}974$; rechazar al nivel del 5%, pero no al nivel del 1%.

11.38. Suponiendo que las varianzas poblacionales son iguales, $t = -0{,}2099$; no rechazar al nivel del 10% o a cualquier nivel habitual de alfa.

11.40. Suponer que la población sigue una distribución normal con varianzas iguales y muestras aleatorias independientes. Rechazar H_0 si $t_{(10,\ 0{,}05)} > 1{,}812$. $t = 3{,}33$; rechazar H_0 al nivel del 5%.

11.42. $z = -2{,}30$, p-valor = 0,0107; rechazar H_0 a los niveles de alfa superiores a 1,07%.

11.44. **a)** Rechazar H_0 si $z_{0{,}05} < -1{,}645$, $z = -1{,}2$; no rechazar H_0 al nivel del 5%.
b) Rechazar H_0 si $|z_{0{,}025}| > 1{,}96$, $z = 0{,}932$; no rechazar H_0 al nivel del 5%.

11.46. Rechazar H_0 si $|z_{0{,}01}| < -2{,}33$, $z = -1{,}19$; no rechazar H_0 al nivel del 1%.

11.48. $z = -1{,}653$, p-valor = 0,0495; rechazar H_0 a niveles de alfa del 4,95%.

11.50. Rechazar H_0 si $F_{11,\ 11,\ 0{,}05} > 2{,}85$. $F = 1{,}05$; por lo tanto, no rechazar H_0

11.52. **a)** Rechazar H_0 si $|z_{0{,}015}| > 2{,}17$, $z = 1{,}987$; no rechazar al nivel del 3%.
b) Rechazar H_0 si $|z_{0{,}03}| > 1{,}88$, $z = 1{,}987$; rechazar al nivel del 3%.

11.54. **a)** $t = 1{,}74$, p-valor = 0,044. El contraste t correspondiente a pares enlazados realizado con los datos originales muestra una diferencia significativa entre las ventas semanales; se observa que la marca 2 es significativamente mayor que la marca 4 al nivel de 0,05.
b) $t = 1{,}42$, p-valor = 0,081. Eliminando el caso atípico más extremo de los datos de la marca 2, la diferencia entre las dos marcas deja de ser significativa al nivel de 0,05.

11.56. Los límites de control son 16,48 y 15,52.

Capítulo 12

12.2. **a)** $t = 2{,}303$, $t_{38,\ 0{,}05} \approx 2{,}021$, $t_{38,\ 0{,}01} \approx 2{,}704$; por lo tanto, rechazar H_0 al nivel del 5%. Pruebas insuficientes para rechazar H_0 al nivel del 1%.
b) $t = 4{,}397$, $t_{58,\ 0{,}05} \approx 2{,}000$, $t_{58,\ 0{,}01} \approx 2{,}660$; por lo tanto, rechazar H_0 al nivel del 1%.
c) $t = 5{,}18$, $t_{43,\ 0{,}05} \approx 2{,}021$, $t_{43,\ 0{,}01} \approx 2{,}704$; por lo tanto, rechazar H_0 al nivel del 1%.

d) $t = 3{,}597$, $t_{23,\ 0{,}05} \approx 2{,}069$, $t_{23,\ 0{,}01} \approx 2{,}807$; por lo tanto, rechazar H_0 al nivel del 1%.

12.4. $t = 7{,}7736$, $t_{47,\ 0{,}05} \approx 1{,}684$, $t_{47,\ 0{,}01} \approx 2{,}423$; por lo tanto, rechazar H_0 a todos los niveles habituales de alfa.

12.6. $t = 4{,}8168$; por lo tanto, rechazar H_0 al nivel del 5%, ya que $4{,}8168 > 1{,}671 \approx t_{66,\ 0{,}05}$

12.8. a)

```
Pearson correlation of Instructor Rating and Expected Grade=0.722, p-value=0,008
```

b) $t = 3{,}2971$; por lo tanto, rechazar H_0 al nivel del 10%, ya que $3{,}2971 > 1{,}372 = t_{10,\ 0{,}10}$

12.10. a) Y varía en $+30$
b) Y varía en -40
c) $\hat{y} = 220$
d) $\hat{y} = 330$
e) Los resultados de la regresión no «demuestran» que los aumentos de los valores de X «causen» un aumento de los valores de Y. La teoría ayuda a establecer las conclusiones sobre la causalidad.

12.12. a) Y varía en $+80$
b) Y varía en -60
c) $\hat{y} = 153$
d) $\hat{y} = 333$
e) Los resultados de la regresión no «demuestran» que los aumentos de los valores de X «causen» un aumento de los valores de Y. La teoría ayuda a establecer las conclusiones sobre la causalidad.

12.14. Una ecuación de regresión poblacional contiene los verdaderos coeficientes de regresión β_i y el verdadero error del modelo ε_i. En cambio, el modelo de regresión estimado consiste en los coeficientes de regresión estimados b_i y el residuo ε_i. La ecuación de regresión poblacional es un modelo que pretende medir el valor efectivo de Y en función de X, mientras que la ecuación de regresión muestral es una estimación del valor predicho de la variable dependiente Y en función de X.

12.16. La constante representa un ajuste del modelo estimado y no del número vendido cuando el precio es cero.

12.18. a) $b_1 = 1{,}80$ $b_0 = 10$ $\hat{y}_i = 10 + 1{,}80x_i$
b) $b_1 = 1{,}30$ $b_0 = 132$ $\hat{y}_i = 132 + 1{,}30x_i$
c) $b_1 = 0{,}975$ $b_0 = 80{,}5$ $\hat{y}_i = 80{,}5 + 0{,}975x_i$
d) $b_1 = 0{,}30$ $b_0 = 47$ $\hat{y}_i = 47 + 0{,}30x_i$
e) $b_1 = 0{,}525$ $b_0 = 152{,}75$ $\hat{y}_i = 152{,}75 + 0{,}525x_i$

12.20. a) $b_1 = 1{,}0737$, $b_0 = -0{,}2336$, $\hat{y}_i = -0{,}2336 + 1{,}0737x_i$
b) Estimamos que por cada aumento de la tasa de rendimiento del índice S&P 500 en una unidad, la tasa de rendimiento de las acciones de la empresa aumentan un 1,07%.
c) Cuando la tasa porcentual de rendimiento del índice S&P 500 es cero, estimamos que la tasa de rendimiento de la empresa es de $-0{,}2336\%$.

12.22. a) $b_1 = 0{,}5143$, $b_0 = 2{,}8854$, $\hat{y}_i = 2{,}8854 + 0{,}5143x$
b) Estimamos que por cada aumento del coste medio de una cena en una unidad, el número de botellas vendidas aumenta en 0,5148%.
c) Sí. Se estima que se venden 2,8854 botellas, independientemente del precio pagado por una cena.

12.24. a) $\hat{y} = 1{,}89 + 0{,}0896x$
b) 0,0896%. Estimamos que por cada ganancia de un 1% obtenida antes del 13 de noviembre, hay una pérdida de 0,0896% el 13 de noviembre.

12.26. **a)** $SCR = 50.000$. $SCE = 50.000$. $s_e^2 = 1.000$, $R^2 = 0{,}50$
b) $SCR = 63.000$. $SCE = 27.000$. $s_e^2 = 540$, $R^2 = 0{,}70$
c) $SCR = 192$. $SCE = 48$. $s_e^2 = 0{,}96$, $R^2 = 0{,}80$
d) $SCR = 60.000$. $SCE = 140.000$. $s_e^2 = 1.944{,}444$, $R^2 = 0{,}30$
e) $SCR = 54.000$. $SCE = 6.000$. $s_e^2 = 157{,}8947$, $R^2 = 0{,}90$

12.28. **a)** $$R^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{\Sigma[b_1(x_i - \bar{x})]^2}{\Sigma(y_i - \bar{y})^2} = b_1^2 \frac{\Sigma(x_i - x)^2}{\Sigma(y_i - \bar{y})^2}$$

b) $$R^2 = b_1^2 \frac{\Sigma(x_i - \bar{x})^2}{\Sigma(y_i - \bar{y})^2} = b_1 \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(y_i - \bar{y})^2} = \frac{[\Sigma(x_i - \bar{x})(y_i - \bar{y})]^2}{\Sigma(x_i - x)^2 \sum(y_i - y)^2} = r^2$$

c) $$b_1 b_1^* = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(y_i - \bar{y})^2} = r^2$$

12.30. **a)** $R^2 = 0{,}1653$, como se vio en el ejercicio 12.7: $r = -0{,}4066$, $r^2 = 0{,}1653 = R^2$

12.32 **a)**

Regression Analysis: Change in Mean a Versus Change in Absent

```
The regression equation is
Change in Mean absence illness = 0.0449 - 0.224 Change in Absentee Rate
Predictor                  Coef    SE Coef     T      P
Constant                0.04485  0.06347  0.71  0.498
Change in Absentee Rate -0.22426  0.05506 -4.07  0.003
S = 0.207325   R-Sq = 64.8%   R-Sq(adj) = 60.9%
```

b) $STC = \Sigma y^2 - n\bar{y}^2 = 1{,}1 - 25(0{,}0)^2 = 1{,}1$
$SCR = \Sigma(\hat{y}_i - \bar{y})^2 = 0{,}713$
$SCE = \Sigma e_i^2 = 0{,}387$
$STC = 1{,}1 = 0{,}713 + 0{,}387 = SCR + SCE$

c) $R^2 = SCR/STC = 0{,}713/1{,}1 = 0{,}648$; el 64,8% de la variación de la variable dependiente, la tasa media de absentismo laboral por enfermedad, puede atribuirse a la variación del cambio de la tasa de absentismo.

12.34. $R^2 = r^2 = 0{,}0121$. El 1,21% de la variación de la variable dependiente de las subidas anuales puede atribuirse a la variación de las evaluaciones de la docencia.

12.36. **a)** $F = 8{,}857$. $F_{\alpha,\,1,\,n-2} = 4{,}170$; por lo tanto, rechazar H_0 al nivel de 0,05
b) $F = 43{,}165$. $F_{\alpha,\,1,\,n-2} = 4{,}00$; por lo tanto, rechazar H_0 al nivel de 0,05
c) $F = 20{,}902$. $F_{\alpha,\,1,\,n-2} = 4{,}281$; por lo tanto, rechazar H_0 al nivel de 0,05

12.38. **a)** $b_1 = 0{,}5391$, $b_0 = 3{,}2958$, $\hat{y}_i = 3{,}2958 + 0{,}5391x_i$
b) De 2.406 a 0,8376

12.40. **a)** $s_e^2 = 144{,}4686$
b) $s_b^2 = 1{,}8991$
c) De $-5{,}0673$ a 0,9991
d) $t = -1{,}476$; por lo tanto, no rechazar H_0 al nivel del 10%, ya que $t = -1{,}476 > -1{,}796 = -t_{11,\,0{,}05}$

12.42. Intervalo de predicción al 95%: (56,467, 97,533)
Intervalo de confianza al 95%: (71,371, 82,629)

12.44. Intervalo de predicción al 95%: (150,331, 165,669)
Intervalo de confianza al 95%: (155,351, 160,649)

12.46 **a)** $t = -7{,}303$; por lo tanto, rechazar H_0 al nivel del 1%, ya que $t = -7{,}303 > -2{,}807 = t_{23,\,0{,}005}$
b) $y_{n+1} = 12{,}6 - 1{,}2(4) = 7{,}8$; intervalo al 90%: (4,4798, 11,1203)

12.48 **a)** $t = 7{,}689$; por lo tanto, rechazar H_0 al nivel del 1%, ya que $t = 7{,}689 > 2{,}878 = t_{18,\ 0{,}005}$
b) $t = 0{,}5278$; por lo tanto, no rechazar H_0 al nivel del 20%, ya que $t = 0{,}5278 < 1{,}33 = t_{18,\ 0{,}10}$

12.50. $t = 5{,}1817$; por lo tanto, rechazar H_0 al nivel del 1%, ya que $t = 5{,}1817 > 2{,}807 = t_{23,\ 0{,}005}$

12.52. $t = 2{,}969$; por lo tanto, rechazar H_0 al nivel del 1%, ya que $t = 2{,}969 > 2{,}947 = t_{15,\ 0{,}005}$

12.54. $t = -10{,}251$; por lo tanto, rechazar H_0 al nivel del 0,5%, ya que $t = -10{,}251 < -3{,}707 = t_{6,\ 0{,}005}$

12.56. El intervalo de confianza al 90% de la predicción del valor efectivo: (de $-11{,}5212$ a 33,337)
El intervalo de confianza al 90% de la predicción del valor esperado: (de 4,8197 a 16,9961)
La distinción entre los dos se encuentra en la incertidumbre sobre el valor esperado o medio en comparación con la incertidumbre sobre un valor específico. Ambos están centrados en el mismo valor; sin embargo, la incertidumbre sobre un valor específico es mayor que en el caso del valor esperado o medio, ya que están incluidas las diferencias tanto del valor esperado como del valor individual con respecto al valor esperado.

12.58. El intervalo de confianza al 90% de la predicción del valor esperado: (de 0,2591 a 0,14211)
El intervalo de confianza al 95% de la predicción del valor esperado: (de 0,1362 a 1,544)

12.60. Obsérvese que los valores calculados de la salida Minitab del análisis de regresión son exactamente los mismos para los cuatro conjuntos de datos, pero en los diagramas de puntos dispersos las pautas de los datos son muy diferentes y, por lo tanto, los modelos también:
El modelo de $Y_1 = f(X_1)$ es un buen ajuste de un modelo lineal.
El modelo de $Y_2 = f(X_2)$ es un modelo no lineal.
El modelo de $Y_3 = f(X_3)$ tiene un importante caso extremo en el valor más alto de $X1$.
El modelo de $Y_4 = f(X_4)$ sólo tiene dos valores de la variable independiente.

12.62. Dos variables aleatorias están correlacionadas positivamente si los valores bajos de una van acompañados de valores bajos de la otra y los valores altos de una van acompañados de valores altos de la otra:
a) Los gastos totales de consumo están correlacionados positivamente con la renta disponible.
b) El precio de un bien o de un servicio está relacionado negativamente con la cantidad vendida.
c) El precio de la mantequilla y las ventas de relojes de pulsera no están correlacionados.

12.64. $t = 2{,}844$; por lo tanto, rechazar H_0 al nivel del 0,5%, ya que $t = 2{,}844 > 2{,}660 \approx t_{51,\ 0{,}005}$

12.66. $t = 2{,}452$; por lo tanto, rechazar H_0 al nivel del 1%, ya que $t = 2{,}452 > 2{,}39 \approx t_{60,\ 0{,}01}$

12.68. Para demostrarlo, sea $x = \bar{x}$ en el caso de la regresión de y con respecto a x, $y = b_0 + b_1 x$
$\hat{y} = b_0 + b_1\bar{x} = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y}$

12.70. **a)** Estimamos que por cada variación de la tasa de inflación en una unidad, el tipo efectivo al contado varía en 0,7916 unidades.
b) $R^2 = 9{,}7\%$. El 9,7% de la variación del tipo efectivo al contado puede atribuirse a las variaciones del tipo al contado predicho por la tasa de inflación.
c) $t = 2{,}8692$; por lo tanto, rechazar H_0 al nivel del 0,5%, ya que $t = 2{,}8692 > 2{,}66 = t_{77,\ 0{,}005}$
d) $t = -0{,}7553$; por lo tanto, no rechazar H_0 a cualquier nivel habitual de alfa.

12.72. **a)** Estimamos que por cada aumento del examen de posición en una unidad, la calificación final de los estudiantes al final del curso aumenta en 0,2875 puntos.
b) El 11,58% de la variación de la calificación final de los estudiantes puede atribuirse a la variación del examen de posición.

c) Los dos métodos son (1) contrastar la significación del coeficiente de la pendiente de la regresión poblacional (β) y (2) contrastar la significación del coeficiente de correlación poblacional (ρ).
(1) $H_0: \beta = 0$, $H_1: \beta > 0$
$t = 6{,}2965$. Por lo tanto, rechazar H_0 a cualquier nivel habitual de alfa.
(2) $H_0: \rho = 0$, $H_1: \rho > 0$, $r = 0{,}3403$, $t = 6{,}3098$. Por lo tanto, rechazar H_0 a cualquier nivel habitual de alfa.

12.74. a) $R^2 = 23{,}88\%$ de la variación de la variable dependiente puede atribuirse a la variabilidad de la variable independiente x.
b) $t = 2{,}6863$; por lo tanto, rechazar H_0 al nivel del 5%, ya que
$t = 2{,}6863 > 2{,}069 = t_{23,0,025}$
c) (De 0,2987 a 2,3013)

12.76. El modelo de regresión lineal todavía podría ser adecuado si la cantidad utilizada de fertilizante estuviera dentro del rango de valores utilizados para estimar la ecuación de regresión. Las ecuaciones de regresión pueden no ser tan útiles para realizar predicciones partiendo de datos situados fuera del rango de los valores muestrales.

12.78. a) El gráfico en el que la variable independiente es el peso del vehículo muestra una leve relación positiva con las muertes en accidente. El R^2 de la regresión simple es de 5,9%. El porcentaje de automóviles importados tiene una leve relación negativa con el R^2 de la regresión simple: 8,1%. La relación entre las muertes y las camionetas es una relación positiva mucho mayor con el R^2 de la regresión simple del 52,7%. Y la antigüedad del automóvil tiene una débil relación negativa con el R^2 de la regresión simple del 17,8%. Todos los gráficos muestran un dato atípico de 0,55 muertes en accidente en la 49.ª observación del conjunto de datos. Este punto es mucho más alto de lo esperado, dados los niveles de las variables independientes.

b)

Regression Analysis: Deaths Versus vehwt

```
The regression equation is
deaths = -0.346+0.000147 vehwt
Predictor       Coef    SE Coef     T      P
Constant     -0.3458     0.3022  -1.14  0.258
vehwt     0.00014697 0.00008528   1.72  0.091

S = 0.0786123 R-Sq = 5.9% R-Sq(adj) = 3.9%
```

Regression Analysis: Deaths Versus impcars

```
The regression equation is
deaths = 0.224-0.00478 impcars
Predictor      Coef   SE Coef     T      P
Constant    0.22371    0.02662  8.40  0.000
impcars   -0.004776   0.002351 -2.03  0.048

S = 0.0777183 R-Sq = 8.1% R-Sq(adj) = 6.1%
```

Regression Analysis: Deaths Versus lghttrks

```
The regression equation is
deaths = 0.0137 + 0.00974 lghttrks
Predictor     Coef   SE Coef    T      P
Constant   0.01375   0.02359 0.58  0.563
lghttrks  0.009742 0.001346 7.24  0.000

S = 0.0557321 R-Sq = 52.7% R-Sq(adj) = 51.7%
```

Regression Analysis: Deaths Versus carage

```
The regression equation is
deaths = 5.26 - 0.0723 carage
Predictor     Coef  SE Coef     T      P
Constant     5.263    1.594  3.30  0.002
carage    -0.07234 0.02266 -3.19  0.003

S = 0.0734818 R-Sq = 17.8% R-Sq(adj) = 16.1%
```

c) Las variables independientes están ordenadas basándose en el R^2 de la regresión simple.

Variable	R^2	Rank
Camionetas	52,7%	1
Antigüedad de los automóviles	17,8%	2
Automóviles importados	8,1%	3
Peso de los vehículos	5,9%	4

Las muertes en accidente están relacionadas positivamente tanto con el peso como con el porcentaje de camionetas. Están relacionadas negativamente con el porcentaje de automóviles importados y la antigüedad del vehículo. Las camionetas tienen la relación lineal más estrecha seguidas de la antigüedad y del peso de los vehículos.

12.80. a) Los diagramas de puntos dispersos muestran que el valor de mercado está relacionado positivamente con el tamaño de la vivienda. Los casos atípicos incluyen varias viviendas cuya valoración basada en su tamaño es mucho mayor de lo esperado. El valor de mercado está relacionado negativamente con el tipo impositivo, aunque se observan algunos casos excepcionales en los que las viviendas tienen la máxima valoración, pero sus tipos impositivos están entre los más bajos.

b)

Regression Analysis: hseval Versus sizehse

```
The regression equation is
hseval = -40.1 + 11.2 sizehse

Predictor          Coef    SE Coef        T        P
Constant         -40.15      10.11    -3.97    0.000
sizehse          11.169      1.844     6.06    0.000

S = 4.188     R-Sq = 29.4%     R-Sq(adj)  =  28.6%
```

Regression Analysis: hseval Versus taxrate

```
The regression equation is
hseval = 26.6 -208 taxrate

Predictor          Coef    SE Coef        T        P
Constant         26.650      1.521    17.52    0.000
taxrate         -207.60      53.27    -3.90    0.000

S = 4.603    R-Sq = 14.7%   R-Sq(adj)  =  13.7%
```

El tamaño de la vivienda es un predictor más fuerte que el tipo impositivo.

c) El hecho de que se bajen o no los tipos impositivos no influye en la valoración tanto como el tamaño de la vivienda.

12.82. a) Inversión fija en vivienda en relación con el tipo de interés preferencial:

Regression Analysis: FRH Versus FBPR

```
The regression equation is
FRH = 132 + 10.5 FBPR
210 cases used 8 cases contain missing values

Predictor      Coef    SE Coef          T       P
Constant    132.004      9.828      13.43   0.000
FBPR         10.529      1.233       8.54   0.000

S = 64.19      R-Sq = 26.0%   R-Sq(adj) = 25.6%
```

Inversión privada fija en vivienda en relación con el tipo de los fondos federales:

Regression Analysis: FRH Versus FFED

```
The regression equation is
FRH = 191 + 5.01 FFED
184 cases used 34 cases contain missing values

Predictor       Coef     SE Coef         T        P
Constant      190.67       10.25     18.60    0.000
FFED           5.013       1.480      3.39    0.001

S = 66.41       R-Sq = 5.9%     R-Sq(adj) = 5.4%
```

El primer modelo de regresión con un coeficiente de determinación del 26,0% hace mejores predicciones que el segundo con un coeficiente de determinación de sólo 5,9%.

b) Intervalos de confianza al 95% del coeficiente de la pendiente:
Tipo de interés preferencial: de 8,112 a 12,946
Tipo de los fondos federales: de 2,112 a 7,914

c) Subida de los tipos de interés de dos puntos:
Tipo de interés preferencial: $\hat{Y} = 153{,}062$
Tipo de los fondos federales: $\hat{Y} = 200{,}696$

d) Intervalos de confianza al 95% dada una subida de los tipos de interés de 2 puntos porcentuales:
Dada una subida de cada tipo de interés de 2%:
Tipo de interés preferencial: de 152,8559 a 153,2681
Tipo de los fondos federales: de 200,5225 a 200,8695

Capítulo 13

13.2. **a)** $\hat{y} = 174$ **b)** $\hat{y} = 181$ **c)** $\hat{y} = 311$ **d)** $\hat{y} = 188$

13.4. **a)** $\hat{y}$ aumenta en 8 **b)** $\hat{y}$ aumenta en 8 **c)** $\hat{y}$ aumenta en 24

13.6. **a)** $b_1 = 0{,}661$: manteniéndose todo lo demás constante, un aumento de la velocidad máxima del avión en un kilómetro por hora aumenta el número esperado de horas del esfuerzo de diseño, según las estimaciones, en 0,661 millones, o sea, en 661.000 horas de trabajo.

b) $b_2 = 0{,}065$: manteniéndose todo lo demás constante, un aumento del peso del avión en una tonelada aumenta el número esperado de horas del esfuerzo de diseño, según las estimaciones, en 0,065 millones, o sea, 65.000 horas de trabajo.

c) $b_3 = -0{,}018$: manteniéndose todo lo demás constante, un aumento del porcentaje de piezas en común con otros modelos provoca una disminución del número esperado de horas del esfuerzo de diseño, según las estimaciones, en 0,018 millones, o sea, 18.000 horas de trabajo.

13.8. **a)** $b_1 = 0{,}052$: manteniéndose todo lo demás constante, un aumento de la renta semanal de cien dólares provoca, según las estimaciones, un aumento del consumo de leche de 0,052 litros semanales. $b_2 = 1{,}14$: manteniéndose todo lo demás constante, un aumento del tamaño de la familia en una persona provoca un aumento estimado del consumo de leche en 1,14 litros a la semana.

b) El término constante b_0 de $-0{,}025$ es el consumo estimado de litros de leche a la semana, dado que la renta semanal de la familia es de 0 dólares y hay 0 miembros en la familia. Esta interpretación probablemente supone extrapolar más allá de los datos muestrales, por lo que no es una interpretación útil.

13.10. **a)** $b_1 = 2{,}000$, $b_2 = 3{,}200$ **b)** $b_1 = -0{,}667$, $b_2 = 1{,}067$
c) $b_1 = 0{,}083$, $b_2 = 0{,}271$ **d)** $b_1 = 0{,}9375$, $b_2 = -0{,}4375$

13.12. **a)**

```
The regression equation is
salesmw2 = − 647363 + 19895 priclec2 + 2.35 numcust2
```

Manteniéndose todo lo demás constante, estimamos que por cada aumento del precio de la electricidad en una unidad, las ventas aumentan en 19.895 megavatios.

Manteniéndose todo lo demás constante, estimamos que por cada cliente residencial adicional que utiliza la electricidad para calentar su casa, las ventas aumentan en 2,353 megavatios.

b)

```
The regression equation is
salesmw2 = − 410202 + 2.20 numcust2
```

Se estima que un cliente residencial más aumenta las ventas de electricidad en 2,2027 megavatios.

Los dos modelos tienen más o menos el mismo poder explicativo; por lo tanto, la introducción del precio como variable no aumenta significativamente el poder explicativo del modelo. Parece que existe una estrecha correlación entre las variables independientes que puede influir en los coeficientes estimados. Las estimaciones de los coeficientes de regresión dependen de las demás variables de predicción del modelo.

c)

```
The regression equation is
salesmw2 = 2312260 - 165275 priclec2 + 56.1 degrday2
```

Manteniéndose todo lo demás constante, un aumento del precio de la electricidad reduce las ventas de electricidad en 165.275 megavatios.

Manteniéndose todo lo demás constante, un aumento de los grados-días (desviación con respecto al tiempo normal) en una unidad aumenta las ventas de electricidad en 56,06 megavatios.

Obsérvese que el coeficiente de la variable del precio ahora es negativo, como cabría esperar, y es significativamente diferente de cero (p-valor = 0,000).

d)

```
The regression equation is
salesmw2 = 293949 + 326 Yd872 + 58.4 degrday2
```

Manteniéndose todo lo demás constante, un aumento de la renta personal disponible en una unidad aumenta las ventas de electricidad en 325,85 megavatios.

Manteniéndose todo lo demás constante, un aumento de los grados-días en una unidad aumenta las ventas de electricidad en 58,36 megavatios.

13.14. a)

```
The regression equation is
horspwr = 23.5 + 0.0154 weight + 0.157 displace
```

Manteniéndose todo lo demás constante, un aumento del peso del automóvil de 100 libras va acompañado de un aumento de la potencia del automóvil de 1,54.

Manteniéndose todo lo demás constante, un aumento del desplazamiento del motor de 10 pulgadas cúbicas va acompañado de un aumento de la potencia del automóvil de 1,57.

b)

```
The regression equation is
horspwr = 16.7 + 0.0163 weight + 0.105 displace + 2.57 cylinder
```

Manteniéndose todo lo demás constante, un aumento del peso del automóvil de 100 libras va acompañado de un aumento de la potencia del automóvil de 1,63.

Manteniéndose todo lo demás constante, un aumento del desplazamiento del motor de 10 pulgadas cúbicas va acompañado de un aumento de la potencia del automóvil de 1,05.

Manteniéndose todo lo demás constante, un cilindro más en el motor va acompañado de un aumento de la potencia del automóvil de 2,57.

Obsérvese que la introducción de la variable independiente del número de cilindros no ha aumentado el poder explicativo del modelo. R^2 ha aumentado marginalmente. El desplazamiento del motor ya no es significativo al nivel de 0,05 (p-valor de 0,074) y el coeficiente estimado del número de cilindros no es significativamente diferente de cero, debido a la estrecha correlación que existe entre las pulgadas cúbicas de desplazamiento del motor y el número de cilindros.

c)

```
The regression equation is
horspwr = 93.6 + 0.00203 weight + 0.165 displace − 1.24 milpgal
```

Manteniéndose todo lo demás constante, un aumento del peso del automóvil de 100 libras va acompañado de un aumento de la potencia del automóvil de 0,203.

Manteniéndose todo lo demás constante, un aumento del desplazamiento del motor de 10 pulgadas cúbicas va acompañado de un aumento de la potencia del automóvil de 1,6475.

Manteniéndose todo lo demás constante, un aumento del ahorro de combustible del vehículo de 1 milla por galón va acompañado de una reducción de la potencia de 1,2392.

Obsérvese que el coeficiente negativo del ahorro de combustible indica la disyuntiva que se espera entre la potencia y el ahorro de combustible. La variable del desplazamiento es significativamente positiva, como cabía esperar; sin embargo, la variable del peso ya no es significativa. Una vez más, sería de esperar que hubiera una estrecha correlación entre las variables independientes.

d)

```
The regression equation is
horspwr = 98.1 - 0.00032 weight + 0.175 displace - 1.32 milpgal + 0.000138 price
```

Manteniéndose todo lo demás constante, un aumento del peso del automóvil de 100 libras va acompañado de un aumento de la potencia del automóvil de 0,032.

Manteniéndose todo lo demás constante, un aumento del desplazamiento del motor de 10 pulgadas cúbicas va acompañado de un aumento de la potencia del automóvil de 1,75.

Manteniéndose todo lo demás constante, un aumento del ahorro de combustible del vehículo de 1 milla por galón va acompañado de una reducción de la potencia del automóvil de 1,32.

Manteniéndose todo lo demás constante, un aumento del precio de 100 $ va acompañado de un aumento de la potencia del automóvil de 0,0138.

e) El poder explicativo ha aumentado marginalmente entre el primer modelo y el último. El coeficiente estimado del precio no es significativamente diferente de cero. El desplazamiento y el ahorro de combustible tienen los signos esperados. El coeficiente del peso tiene el signo incorrecto; sin embargo, no es significativamente diferente de cero (p-valor de 0,953).

13.16. **a)** $s_e^2 = 86{,}207$, $s_e = 9{,}2848$
b) $STC = 9.500$
c) $R^2 = 0{,}7368$, $\bar{R}^2 = 0{,}7187$

13.18. **a)** $s_e^2 = 75{,}0$, $s_e = 8{,}660$
b) $STC = 95.000$
c) $R^2 = 0{,}8421$, $\bar{R}^2 = 0{,}7822$

13.20. **a)** $R^2 = 0{,}5441$; por lo tanto, el 54,41% de la variabilidad del consumo de leche puede atribuirse a las variaciones de la renta semanal y del tamaño de la familia.
b) $\bar{R}^2 = 0{,}5103$
c) $R = 0{,}7376$. Ésta es la correlación muestral entre el valor observado del consumo de leche y el predicho.

13.22. a)
```
The regression equation is
Y profit = 1.55 - 0.000120 X2  offices
```

b)
```
The regression equation is
X1 revenue = - 0.078 + 0.000543 X2 offices
```

c)
```
The regression equation is
Y profit = 1.33 - 0.169 X1 revenue
```

d)
```
The regression equation is
X2  offices = 957 + 163 1X1 revenue
```

13.24. a) IC al 95% de x_1: de 0,4698 a 13,1302
IC al 95% de x_2: de $-0{,}6554$ a 14,554
IC al 95% de x_3: de $-7{,}2 + 2{,}042$ (3,2); de $-13{,}7344$ a $-0{,}6656$

b) Para x_1: $t = 2{,}194$ $t_{30,\ 0{,}05/0{,}01} = 1{,}697,\ 2{,}457$. Rechazar H_0 al nivel del 5%, pero no al nivel del 1%.
Para x_2: $t = -1{,}865$ $t_{30,\ 0{,}05/0{,}01} = 1{,}697,\ 2{,}457$. Rechazar H_0 al nivel del 5%, pero no al nivel del 1%.
Para x_3: $t = -2{,}25$ $t_{30,\ 0{,}05/0{,}01} = 1{,}697,\ 2{,}457$. No rechazar H_0 al nivel del 5% ni al nivel del 1%.

13.26. a) IC al 95% de x_1: de 3,4510 a 32,149
IC al 95% de x_2: de $-0{,}788$ a 54,588
IC al 95% de x_3: de $-16{,}88$ a $-1{,}52$

b) Para x_1: $t = 2{,}507$ $t_{35,\ 0{,}05/0{,}01} \approx 1{,}697,\ 2{,}457$. Rechazar H_0 al nivel del 5% y al nivel del 1%.
Para x_2: $t = 1{,}964$ $t_{35,\ 0{,}05/0{,}01} \approx 1{,}697,\ 2{,}457$. Rechazar H_0 al nivel del 5%, pero no al nivel del 1%.
Para x_3: $t = -2{,}421$ $t_{35,\ 0{,}05/0{,}01} \approx 1{,}697,\ 2{,}457$. No rechazar H_0 al nivel del 5%, pero no al nivel del 1%.

13.28. a) $t = 2{,}26$. $t_{27,\ 0{,}025/0{,}01} = 2{,}052,\ 2{,}473$; por lo tanto, rechazar H_0 al nivel del 2,5%, pero no al nivel del 1%.

b) IC al 90%: de 0,5439 a 1,7361
IC al 95%: de 0,4218 a 1,8582
IC al 99%: de 0,1701 a 2,1099

13.30. a) $t = -0{,}428$, $t_{16,\ 0{,}10} = -1{,}337$; por lo tanto, no rechazar H_0 al nivel del 20%.

b) $F = 13{,}057$, $F_{3,\ 16,\ 0{,}01} = 5{,}29$; por lo tanto, rechazar H_0 al nivel del 1%.

13.32. a) Manteniéndose todo lo demás constante, un aumento de la renta personal per cápita media de 1 $ más provoca un aumento de los ingresos netos per cápita esperados generados por la lotería de 0,04 $.

b) IC al 95%: de 0,2359 a 1,5185

c) $t = -1{,}383$, $t_{24,\ 0{,}10/0{,}05} = -1{,}318,\ -1{,}711$; por lo tanto, rechazar H_0 al nivel del 10%, pero no al nivel del 5%.

13.34. a) IC al 95%: de 0,1805 a 0,2195

b) $t = -1{,}19$, $t_{16,\ 0{,}10} = -1{,}337$; por lo tanto, no rechazar H_0 al nivel del 10%.

13.36. a) IC al 99%: de 0,0173 a 0,0817

b) $t = 0{,}617$, $t_{30,\ 0{,}10} = 1{,}31$; por lo tanto, no rechazar H_0 al nivel del 20%.

c) $t = 2{,}108$, $t_{30,\ 0{,}025/0{,}01} = 2{,}042,\ 2{,}457$; por lo tanto, rechazar H_0 al nivel del 5%, pero no al nivel del 2%.

13.38. **a)** $F = 81{,}955$, $F_{3,\ 23,\ 0{,}01} = 4{,}76$; por lo tanto, rechazar H_0 al nivel del 1%.
b) Tabla del análisis de la varianza:

Fuentes de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente F
Regresor	3,549	3	1,183	81,955
Error	0,332	23	0,014435	
Total	3,881	26		

13.40. **a)** $F = 16{,}113$, $F_{2,\ 27,\ 0{,}01} = 5{,}49$; por lo tanto, rechazar H_0 al nivel del 1%.
b) Tabla del análisis de varianza:

Fuentes de variación	Suma de los cuadrados	Grados de libertad	Media de los cuadrados	Cociente F
Regresor	88,2	2	44,10	16,113
Error	73,9	27	2,737	
Total	162,1	29		

13.42. **a)** $F = 6{,}2449$, $F_{4,\ 24,\ 0{,}01} = 4{,}22$; por lo tanto, rechazar H_0 al nivel del 1%.

13.44. **a)** $F = 217$, $F_{2,\ 16,\ 0{,}01} = 6{,}23$; por lo tanto, rechazar H_0 al nivel del 1%.

13.46.
$$\frac{(SCE^* - SCE)k_1}{SCE/(n-k-1)} = \frac{n-k-1}{k_1}\,\frac{(SCE^* - SCE)/STC}{SCE/STC}$$
$$= \frac{n-k-1}{k_1}\,\frac{1 - R^{2*} - (1-R^2)}{1-R^2}$$
$$= \frac{n-k-1}{k_1}\,\frac{R^2 - R^{2*}}{1-R^2}$$

13.47. **a)**
$$\bar{R}^2 = 1 - \frac{SCE/(n-k-1)}{SCE/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$
$$= \frac{n-1}{n-k-1}R^2 - \frac{k}{n-k-1} = \frac{(n-1)R^2 - k}{n-k-1}$$

b) Dado que $\bar{R}^2 = \dfrac{(n-1)R^2 - k}{n-k-1}$, entonces $R^2 = \dfrac{(n-k-1)\bar{R}^2 + k}{n-1}$

c)
$$\frac{SCR/k}{SCE/(n-k-1)} = \frac{n-k-1}{k}\,\frac{SCR/STC}{SCE/STC}$$
$$= \frac{n-k-1}{k}\,\frac{R^2}{1-R^2} = \frac{n-k-1}{k}\,\frac{[(n-k-1)R^2 + k]/(n-1)}{[n-1-(n-k-1)\bar{R}^2 - k]/(n-1)}$$
$$= \frac{n-k-1}{k}\,\frac{(n-k-1)\bar{R}^2 + k}{(n-k-1)(1-\bar{R}^2)} = \frac{n-k-1}{k}\,\frac{\bar{R}^2 + k}{(1-\bar{R}^2)}$$

13.50. $\hat{Y} = 10{,}638$ kilos

13.52. $\hat{Y} = 2{,}216$ millones de horas de trabajo

13.54. Calcule los valores de y_i cuando $x_i = 1, 2, 4, 6, 8, 10$

x_i	1	2	4	6	8	10
$y_i = 4x^{1,5}$	4	11,3137	32	58,7878	90,5097	126,4611
$y_i = 1 + 2x_i + 2x_i^2$	5	13	41	85	145	221

13.56. Calcule los valores de y_i cuando $x_i = 1, 2, 4, 6, 8, 10$

x_i	1	2	4	6	8	10
$y_i = 4x^{1,5}$	4	11,3137	32	58,7878	90,5097	126,4611
$y_i = 1 + 2x_i + 1{,}7x_i^2$	4,7	11,8	36,2	74,2	125,8	191

13.58. Hay muchas respuestas posibles. Entre las relaciones que pueden aproximarse mediante un modelo cuadrático no lineal hay muchas funciones de oferta, funciones de producción y funciones de coste, incluido el coste medio en relación con el número de unidades producidas.

13.60. **a)** Manteniéndose todo lo demás constante, un aumento del gasto anual de consumo de un 1% va acompañado de un aumento del gasto en viajes de vacaciones del 1,1556%.

Manteniéndose todo lo demás constante, un aumento del tamaño del hogar va acompañado de una disminución de los gastos en viajes de vacaciones del 0,4408%.

b) Una variación de los gastos en viajes de vacaciones del 16,8% puede atribuirse a las variaciones del logaritmo de los gastos totales de consumo y el logaritmo del número de miembros del hogar.

c) De 1,049 a 1,2626

d) $t = -8{,}996$; por lo tanto, rechazar H_0 al nivel del 1%.

13.62. **a)** Manteniéndose todo lo demás constante, un aumento del precio del vacuno del 1% va acompañado de una disminución de las toneladas de vacuno consumidas anualmente en Estados Unidos del 0,529%.

b) Manteniéndose todo lo demás constante, un aumento del precio del porcino del 1% va acompañado de una disminución de las toneladas de vacuno consumidas anualmente en Estados Unidos del 0,217%.

c) $t = 2{,}552$; $t_{25,\,0{,}01} = 2{,}485$; por lo tanto, rechazar H_0 al nivel del 1%.

d) $F = 13{,}466$, $F_{4,\,25,\,0{,}01} = 4{,}18$; por lo tanto, rechazar H_0 al nivel del 1%.

e) Si se ha omitido una variable independiente importante, puede haber un sesgo de especificación. Los coeficientes de regresión obtenidos en el modelo mal especificado serían engañosos.

13.64. **a)** Los coeficientes de los modelos exponenciales pueden estimarse tomando el logaritmo del modelo de regresión múltiple para obtener una ecuación lineal en los logaritmos de las variables.

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \beta_3 \log(X_3) + \beta_4 \log(X)_4 + \log(\varepsilon)$$

Introduciendo las restricciones sobre los coeficientes: $\beta_1 + \beta_2 = 1$, $\beta_2 = 1 - \beta_1$, $\beta_3 + \beta_4 = 1$, $\beta_4 = 1 - \beta_3$

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + [1 - \beta_1] \log(X_2) + \beta_3 \log(X_3) + [1 - \beta_3] \log(X_4) + \log(\varepsilon)$$

Simplificar algebraicamente y estimar los coeficientes. El coeficiente β_2 puede hallarse restando β_1 de 1,0. Asimismo, el coeficiente β_4 puede hallarse restando β_3 de 1,0.

b) La elasticidad constante de Y con respecto a X_4 es el coeficiente del término X_4 de la regresión logarítmica.

13.66.

Results for: GermanImports.xls
Regression Analysis: LogYt Versus LogX1t, LogX2t

```
The regression equation is
LogYt = -4.07 + 1.36 LogX1t + 0.101 LogX2t
Predictor          Coef      SE Coef         T         P    VIF
Constant        -4.0709       0.3100    -13.13    0.000
LogX1t          1.35935      0.03005     45.23    0.000    4.9
LogX2t          0.10094      0.05715      1.77    0.088    4.9

S = 0.04758       R-Sq = 99.7%     R-Sq(adj) = 99.7%
```

13.68. **a)** $\hat{y} = 5{,}78 + 4{,}87x_1$ **b)** $\hat{y} = 1{,}15 + 9{,}51x_1$ **c)** $\hat{y} = 13{,}67 + 8{,}98x_1$

13.70. **a)** Manteniéndose todo lo demás constante, el precio esperado de venta es más alto en 3.219 $ si el piso tiene chimenea.
b) Manteniéndose todo lo demás constante, el precio esperado de venta es más alto en 2.005 $ si el piso tiene suelos de madera.
c) IC al 95%: de 1.362,88 $ a 5.075,12 $.
d) $t = 2{,}611$; $t_{809,\ 0{,}005} = 2{,}576$; por lo tanto, rechazar H_0 al nivel del 0,5%.

13.72. El 35,6% de la variación del rendimiento global en los estudios de postgrado puede atribuirse a la variación de la calificación media de los estudios de grado, a la calificación obtenida en el examen de acceso a la universidad y a si la carta de recomendación es especialmente buena. El modelo global es significativo, ya que podemos rechazar la hipótesis nula de que el modelo no tiene poder explicativo en favor de la hipótesis alternativa de que tiene mucho poder explicativo. Los coeficientes de regresión individuales que son significativamente diferentes de cero son las calificación media en el examen de acceso a la universidad y si la carta de recomendación del estudiante es especialmente buena. El coeficiente de la calificación media en los estudios de grado no es significativo al nivel del 5%.

13.74. **a)** Manteniéndose todo lo demás constante, la valoración media de un curso es 6,21 unidades mayor si interviene un profesor visitante que en caso contrario.
b) $t = 1{,}73$; $t_{20,\ 0{,}05} = 1{,}725$; por lo tanto, rechazar H_0 al nivel del 5%.
c) El 56,9% de la variación de la valoración media del curso puede atribuirse a la variación del porcentaje de tiempo dedicado a sesiones de discusión en grupo, los dólares gastados en la preparación de los materiales del curso, los dólares gastados en comida y bebidas y si interviene un profesor visitante. $F = 6{,}6$, $F_{4,\ 20,\ 0{,}01} = 4{,}43$; por lo tanto, rechazar H_0 al nivel del 1%.
d) De 0,0819 a 0,9581

13.76.

Results for: Student Performance.xls
Regression Analysis: Y Versus X1, X2, X3, X4, X5

```
The regression equation is
Y = 2.00 + 0.0099 X1 + 0.0763 X2 - 0.137 X3 + 0.064 X4 + 0.138 X5
Predictor          Coef      SE Coef         T        P     VIF
Constant          1.997        1.273      1.57    0.132
X1              0.00990      0.01654      0.60    0.556     1.3
X2              0.07629      0.05654      1.35    0.192     1.2
X3             -0.13652      0.06922     -1.97    0.062     1.1
X4               0.0636       0.2606      0.24    0.810     1.4
X5              0.13794      0.07521      1.83    0.081     1.1

S = 0.5416       R-Sq = 26.5%          R-Sq(adj)  =  9.0%
```

El modelo no es significativo (p-valor del contraste $F = 0{,}229$). El modelo sólo explica el 26,5% de la variación de la calificación media utilizando como variables independientes las horas dedicadas a estudiar, las horas dedicadas a preparar los exámenes, las horas pasadas en los bares, el hecho de que los estudiantes tomen o no notas o subrayen cuando leen los libros de texto y el número medio de créditos realizados por semestre. Las únicas variables independientes que son marginalmente significativas (al nivel del 10%, pero no al nivel del 5%) incluyen el número de horas pasadas en los bares y el número medio de créditos. Las otras variables independientes no son significativas a los niveles habituales de alfa.

13.78. **a)** Una gran correlación entre las variables independientes hace que la varianza de los coeficientes estimados sea alta y tiende a tener un pequeño estadístico t de Student.

Utilizar la regla práctica $|r| > \dfrac{2}{\sqrt{n}}$ para averiguar si la correlación es «grande».

b) No existe ninguna correlación entre las variables independientes. Ningún efecto en los coeficientes estimados.

c) Una gran correlación entre las variables independientes hace que la varianza de los coeficientes estimados sea alta y tiende a tener un pequeño estadístico *t* de Student.

d) Utilizar la regla práctica $|r| > \frac{2}{\sqrt{n}}$ para averiguar si la correlación es «grande».

13.80. La correlación entre la variable independiente y la variable dependiente no es necesariamente una prueba de que el estadístico *t* de Student sea pequeño. Una elevada correlación entre las variables *independientes* podría hacer que el estadístico *t* de Student fuera muy pequeño, debido a que la correlación crea una elevada varianza.

El informe de los ejercicios 13.82 a 13.84 puede escribirse siguiendo el extenso caso práctico basado en los datos del fichero **Cotton** (véase el apartado 13.9).

13.86.

Regression Analysis: y_FemaleLFPR versus x1_income, x2_yrsedu, ...

```
The regression equation is
y_FemaleLFPR = 0.2 + 0.000406 x1_income + 4.84 x2_yrsedu - 1.55 x3_femaleun
Predictor        Coef    SE Coef       T          P       VIF
Constant         0.16      34.91    0.00      0.996
x1_incom   0.0004060  0.0001736    2.34      0.024       1.2
x2_yrsed       4.842      2.813    1.72      0.092       1.5
x3_femal     -1.5543     0.3399   -4.57      0.000       1.3

S = 3.048       R-Sq = 54.3%  R-Sq(adj)= 51.4%
```

13.88.

Regression Analysis: y_manufgrowt versus x1_aggrowth, x2_exportgro, ...

```
The regression equation is
y_manufgrowth = 2.15 + 0.493 x1_aggrowth + 0.270 x2_exportgrowth
            - 0.117 x3_inflation
Predictor          Coef      SE Coef          T         P    VIF
Constant         2.1505       0.9695      2.22     0.032
x1_aggro         0.4934       0.2020      2.44     0.019    1.0
x2_expor        0.26991      0.06494      4.16     0.000    1.0
x3_infla       -0.11709      0.05204     -2.25     0.030    1.0

S = 3.624         R-Sq = 39.3%  R-Sq(adj) = 35.1%
```

13.90. La tabla del análisis de la varianza identifica cómo se descompone la variabilidad total de la variable dependiente (*STC*) entre la parte de la variabilidad que es explicada por el modelo de regresión (*SCR*) y la parte que no es explicada (*SCE*). El coeficiente de determinación (R^2) es el cociente ente *SCR* y *STC*. La tabla del análisis de la varianza también calcula el estadístico *F* del contraste de la significación de la regresión global, es decir, si todos los coeficientes estimados son todos ellos iguales a cero. El *p*-valor correspondiente también se indica generalmente en esta tabla.

13.92. Si uno de los modelos contiene más variables explicativas, *STC* sigue siendo igual en los dos modelos, pero *SCR* es más alto en el modelo que tiene más variables explicativas. Dado que $STC = SCR_1 + SCE_1$, que es equivalente a $SCR_2 + SCE_2$, y dado que $SCR_2 > SCR_1$, entonces $SCE_1 > SCE_2$. Por lo tanto, el coeficiente de determinación es más alto cuando es mayor el número de variables explicativas y el coeficiente de determinación debe interpretarse conjuntamente con el hecho de que los coeficientes de la pendiente de regresión de las variables explicativas sean o no significativamente diferentes de cero.

13.94. $\Sigma e_i = \Sigma (y_i - a - b_1 x_{1i} - b_2 x_{2i})$

$\Sigma e_i = \Sigma (y_i - \bar{y} + b_1 \bar{x}_{1i} + b_2 \bar{x}_{2i} - b_1 x_{1i} - b_2 x_{2i})$

$\Sigma e_i = n\bar{y} - n\bar{y} + nb_1 \bar{x}_1 + nb_2 \bar{x}_2 - nb_1 \bar{x}_1 - nb_2 \bar{x}_2$

$\Sigma e_i = 0$

13.96. **a)** Manteniéndose todo lo demás constante, un aumento de una pregunta provoca una disminución del porcentaje esperado de respuestas recibidas de 1,834. Manteniéndose todo lo demás constante, un aumento de la longitud del cuestionario en una palabra provoca una disminución del porcentaje esperado de respuestas recibidas de 0,016.
b) El 63,7% de la variabilidad del porcentaje de respuestas recibidas puede atribuirse a la variabilidad del número de preguntas realizadas y del número de palabras.
c) $F = 23{,}69$, $F_{2,\ 27,\ 0{,}01} = 5{,}49$; por lo tanto, rechazar H_0 al nivel del 1%.
d) De $-3{,}5938$ a $-0{,}752$
e) $t = -1{,}78$; $t_{27,\ 0{,}05/0{,}025} = -1{,}703,\ -2{,}052$. Por lo tanto, rechazar H_0 al nivel del 5%, pero no al nivel del 2,5%.

13.98.

Regression Analysis: y_rating versus x1_expgrade, x2_Numstudents

```
The regression equation is
y_rating = -0.200 + 1.41 x1_expgrade - 0.0158 x2_Numstudents
Predictor          Coef      SE Coef        T        P    VIF
Constant        -0.2001       0.6968    -0.29    0.777
x1_expgr         1.4117       0.1780     7.93    0.000    1.5
x2_Numst      -0.015791     0.003783    -4.17    0.001    1.5

S = 0.1866        R-Sq = 91.5%  R-Sq(adj) = 90.5%
```

13.100. **a)** Manteniéndose todo lo demás constante, cada punto adicional de la calificación esperada del estudiante provoca un aumento esperado de la calificación efectiva de 0,469.
b) De 2,4752 a 4,26276
c) $t = 2{,}096$, $t_{103,\ 0{,}025} = 1{,}96$; por lo tanto, rechazar H_0 al nivel del 5%.
d) El 68,6% de la variación de las calificaciones del examen se debe a su dependencia lineal de la calificación esperada por el estudiante, las horas semanales dedicadas a estudiar para el curso y la calificación media del estudiante.
e) $F = 75{,}008$, $F_{3,\ 103,\ 0{,}01} = 3{,}95$. Rechazar H_0 a cualquier nivel habitual de alfa.
f) $R = 0{,}82825$
g) $\hat{Y} = 75{,}812$

13.102. **a)** De 110,0795 a 850,0005
b) De 803,4152 a 1.897,1848
c) $t = -4{,}9299$; $t_{2.669,\ 0{,}005} = 2{,}576$; por lo tanto, rechazar H_0 al nivel del 0,5%.
d) $t = 6{,}5142$; $t_{2.669,\ 0{,}005} = 2{,}576$; por lo tanto, rechazar H_0 al nivel del 0,5%.
e) El 52,39% de la variabilidad de los minutos jugados puede atribuirse a la variabilidad de las 9 variables.
f) $R = 0{,}7238$

13.104. Puede escribirse un informe basándose en el caso práctico y contrastando la significación del modelo. Véase el apartado 13.9.

13.106. La matriz de correlaciones indica que algunas de las variables independientes es probable que sean significativas; sin embargo, también es un resultado probable una elevada correlación entre las variables independientes. El modelo de regresión con todas las variables independientes es:

Regression Analysis: Salary Versus age, Experien, ...

```
The regression equation is
Salary = 23725 - 40.3 age + 357 Experien + 263 yrs_asoc + 493 yrs_full
            - 954 Sex_1Fem + 3427 Market + 1188 C8
Predictor        Coef    SE Coef          T        P    VIF
Constant        23725       1524      15.57    0.000
age            -40.29      44.98      -0.90    0.372    4.7
Experien       356.83      63.48       5.62    0.000   10.0
yrs_asoc       262.50      75.11       3.49    0.001    4.0
yrs_full       492.91      59.27       8.32    0.000    2.6
Sex_1Fem       -954.1      487.3      -1.96    0.052    1.3
Market         3427.2      754.1       4.54    0.000    1.1
C8             1188.4      597.5       1.99    0.049    1.1

S = 2332 R-Sq = 88.2%  R-Sq(adj) = 87.6%
```

Dado que la edad no es significativa y tiene los estadísticos t más pequeños, se elimina del modelo: eliminando la edad como variable independiente, el contraste F condicionado de la edad es: $F_{x_2} = 0{,}80$, que es muy inferior a cualquier valor crítico habitual de F. Por lo tanto, se elimina la edad del modelo. El resto de las variables independientes son todas significativas al nivel de significación de 0,05 y, por lo tanto, la ecuación con la edad eliminada es el modelo de regresión final. Análisis de los residuos para averiguar si el supuesto de la linealidad se cumple: el gráfico de los residuos de la experiencia muestra una relación cuadrática relativamente estrecha entre la experiencia y los salarios. Por lo tanto, se genera y se añade al modelo una nueva variable, que tiene en cuenta la relación cuadrática. Ninguno de los demás gráficos de los residuos demuestra claramente la ausencia de linealidad. Se ha añadido a las variables independientes el modelo con el término del cuadrado de la experiencia. El término del cuadrado de la experiencia es estadísticamente significativo; sin embargo, Sex_1Fem ya no es significativa al nivel de 0,05, por lo que se elimina del modelo:

Regression Analysis: Salary Versus Experien, ExperSquared, ...

```
The regression equation is
Salary = 18538 + 888 Experien - 16.3 ExperSquared + 237 yrs_asoc
           + 624 yrs_full + 3982 Market + 1145 C8
Predictor        Coef     SE Coef          T        P       VIF
Constant      18537.8       543.6      34.10    0.000
Experien       887.85       72.32      12.28    0.000      20.4
ExperSqu      -16.275       1.718      -9.48    0.000      16.0
yrs_asoc       236.89       59.11       4.01    0.000       3.9
yrs_full       624.49       48.41      12.90    0.000       2.8
Market         3981.8       602.9       6.60    0.000       1.1
C8             1145.4       466.3       2.46    0.015       1.0

S = 1857        R-Sq = 92.5%     R-Sq(adj) = 92.2%
```

Éste es el modelo final con todas las variables independientes significativas, incluida la transformación cuadrática de la experiencia. Eso indicaría que existe una relación no lineal entre la experiencia y el salario.

13.108. **a)** La matriz de correlaciones indica que las muertes en accidente están relacionadas positivamente con el peso del vehículo y el porcentaje de camionetas y negativamente con el porcentaje de automóviles importados y con la antigüedad de los automóviles. Las camionetas son los que tienen la mayor relación lineal de cualquier variable independiente seguidos de la antigüedad de los automóviles. Es probable que exista una estrecha correlación entre las variables independientes debido a la estrecha correlación entre impcars y el peso de los vehículos.

b)

Regression Analysis: deaths Versus vehwt, impcars, lghttrks, carage

```
The regression equation is
deaths = 2.60 + 0.000064 vehwt - 0.00121 impcars + 0.00833 lghttrks
         - 0.0395 carage
Predictor        Coef     SE Coef          T        P       VIF
Constant        2.597       1.247       2.08    0.043
vehwt       0.0000643   0.0001908       0.34    0.738      10.9
impcars     -0.001213    0.005249      -0.23    0.818      10.6
lghttrks     0.008332    0.001397       5.96    0.000       1.2
carage       -0.03946     0.01916      -2.06    0.045       1.4

S = 0.05334     R-Sq = 59.5%     R-Sq(adj) = 55.8%
```

Las camionetas son una variable positiva significativa. Dado que impcars tiene el menor estadístico *t*, se elimina del modelo. También se elimina por la misma razón el peso de los vehículos y se obtiene el siguiente modelo final:

Regression Analysis: deaths Versus lghttrks, carage

```
The regression equation is
deaths = 2.51 + 0.00883 lghttrks - 0.0352 carage
Predictor        Coef     SE Coef          T        P       VIF
Constant        2.506       1.249       2.01    0.051
lghttrks     0.008835    0.001382       6.39    0.000       1.1
carage       -0.03522     0.01765      -2.00    0.052       1.1

S = 0.05404     R-Sq = 56.5%     R-Sq(adj) = 54.6%
```

En el modelo, las camionetas y la antigüedad de los automóviles son las variables significativas. Obsérvese que la antigüedad de los automóviles es marginalmente significativa (*p*-valor de 0,052) y, por lo tanto, también podría eliminarse del modelo.

c) El modelo de regresión indica que el porcentaje de camionetas es significativo en todos los modelos y, por lo tanto, es un importante predictor en el modelo. La antigüedad de los automóviles y los automóviles importados son predictores marginalmente significativos cuando sólo se incluyen en el modelo las camionetas.

13.110. a) La matriz de correlaciones muestra que no es probable que la elevada correlación entre las variables independientes sea un problema en este modelo, ya que ninguna de las correlaciones entre las variables independientes es relativamente alta.

El intervalo para aplicar el modelo de regresión (medias de las variables $+/-$ dos errores típicos) es:

Hseval	11,11 a 30,94
Sizehse	5,0 a 5,96
Taxhse	32,35 a 227,91
Comper	0,034 a 0,286
Incom72	2.727 a 3.995
Totexp	$1.488.848 +/- 2(1.265.564)$ = no es una buena aproximación

b) Modelos de regresión:

Regression Analysis: hseval Versus sizehse, Taxhse, ...

```
The regression equation is
hseval = - 31.1 + 9.10 sizehse - 0.00058 Taxhse - 22.2 Comper
         + 0.00120 incom72 + 0.000001 totexp
Predictor        Coef     SE Coef          T        P       VIF
Constant       -31.07       10.09      -3.08    0.003
sizehse          9.105       1.927       4.72    0.000       1.3
Taxhse      -0.000584    0.008910      -0.07    0.948       1.2
Comper         -22.197       7.108      -3.12    0.002       1.3
incom72       0.001200    0.001566       0.77    0.445       1.5
totexp     0.00000125  0.00000038       3.28    0.002       1.5

S = 3.785       R-Sq = 45.0%     R-Sq(adj) = 41.7%
```

Taxhse no es significativo; tampoco lo es la renta; sin embargo, eliminando una variable cada vez, eliminar primero Taxhse y después la renta:

Regression Analysis: hseval Versus sizehse, Comper, totexp

```
The regression equation is
hseval = - 29.9 + 9.61 sizehse - 23.5 Comper + 0.000001 totexp
Predictor        Coef     SE Coef          T        P       VIF
Constant      -29.875       9.791      -3.05    0.003
sizehse         9.613       1.724       5.58    0.000       1.1
Comper        -23.482       6.801      -3.45    0.001       1.2
totexp     0.00000138  0.00000033       4.22    0.000       1.1

S = 3.754       R-Sq = 44.6%     R-Sq(adj) = 42.6%
```

Éste es el modelo de regresión final. Todas las variables independientes son significativas. Tanto el tamaño de la vivienda como el gasto público total aumentan el valor de mercado de las viviendas, mientras que el porcentaje de propiedades comerciales tiende a reducir el valor de mercado de las viviendas.

c) En el modelo de regresión final, no se observa que la variable de los impuestos sea significativas y, por lo tanto, es difícil apoyar la afirmación del promotor.

13.112. a) La matriz de correlaciones muestra que ambos tipos de interés producen un efecto positivo significativo en la inversión en viviendas. La oferta monetaria, el PIB y el gasto público también tienen una relación lineal significativa con la inversión en viviendas. Obsérvese la estrecha correlación entre los dos tipos de interés, que, como cabría esperar, plantearía importantes problemas si se incluyen en el modelo de regresión. Por lo tanto, los tipos de interés se desarrollarán en dos modelos independientes.

Regression Analysis: FRH Versus FBPR, FM2, GDPH, GH

```
The regression equation is
FRH = 70.0 - 3.79 FBPR - 0.0542 FM2 + 0.0932 GDPH - 0.165 GH
166 cases used 52 cases contain missing values
Predictor        Coef     SE Coef          T        P       VIF
Constant        70.00       24.87       2.82    0.005
FBPR          -3.7871      0.6276      -6.03    0.000       1.2
FM2         -0.054210    0.009210      -5.89    0.000      46.8
GDPH         0.093223    0.007389      12.62    0.000      58.1
GH           -0.16514     0.03747      -4.41    0.000      28.7

S = 23.42       R-Sq = 86.7%     R-Sq(adj) = 86.3%
```

Éste es el modelo final en el que el tipo preferencial es la variable del tipo de interés, ya que todas las variables independientes son significativas. Obsérvese la significativa multicolinealidad que existe entre las variables independientes.

Regression Analysis: FRH Versus FFED, FM2, GDPH, GH

```
The regression equation is
FRH = 55.0 - 2.76 FFED - 0.0558 FM2 + 0.0904 GDPH - 0.148 GH
166 cases used 52 cases contain missing values
Predictor        Coef     SE Coef          T        P       VIF
Constant        55.00       26.26       2.09    0.038
FFED          -2.7640      0.6548      -4.22    0.000       1.2
FM2         -0.05578     0.01007      -5.54    0.000      50.7
GDPH         0.090402    0.007862      11.50    0.000      59.6
GH           -0.14752     0.03922      -3.76    0.000      28.5

S = 24.61       R-Sq = 85.3%     R-Sq(adj) = 84.9%
```

El modelo en el que el tipo de los fondos federales es la variable del tipo de interés también es el modelo final en el que todas las variables independientes son significativas. De nuevo, la multicolinealidad será un problema en este modelo de regresión.

b) Intervalos de confianza al 95% de los coeficientes de las pendientes del término de los tipos de interés:

Tipo preferencial como variable del tipo de interés: $-3{,}7871 +/- 1{,}23$

Tipo de los fondos federales como variable del tipo de interés: $-2{,}764 +/- 1{,}2834$

13.114. a) Se calcula, en primer lugar, la matriz de correlaciones, que indica que hay varias variables independientes que seguramente tienen mucho poder explicativo en el modelo de regresión. Sería de esperar que la edad, los años de profesor asociado y los años de catedrático («full profesor») fueran significativos:

Regression Analysis: Salary Versus age, yrs_asoc, ...

```
The regression equation is
Salary = 21107 + 105 age + 532 yrs_asoc + 690 yrs_full
           - 1312 Sex_1Fem + 2854 Market + 1101 C8
Predictor        Coef     SE Coef          T        P       VIF
Constant        21107        1599      13.20    0.000
age            104.59       40.62       2.58    0.011       3.1
yrs_asoc       532.27       63.66       8.36    0.000       2.4
yrs_full       689.93       52.66      13.10    0.000       1.7
Sex_1Fem      -1311.8       532.3      -2.46    0.015       1.3
Market         2853.9       823.3       3.47    0.001       1.0
C8             1101.0       658.1       1.67    0.097       1.1

S = 2569        R-Sq = 85.6%     R-Sq(adj) = 85.0%
```

A continuación, eliminando la variable C8, se obtiene el modelo final:

Regression Analysis: Salary Versus age, yrs_asoc, ...

```
The regression equation is
Salary = 21887 + 90.0 age + 539 yrs_asoc + 697 yrs_full
           - 1397 Sex_1Fem + 2662 Market
Predictor        Coef     SE Coef          T        P       VIF
Constant        21887        1539      14.22    0.000
age             90.02       39.92       2.26    0.026       3.0
yrs_asoc       539.48       63.91       8.44    0.000       2.4
yrs_full       697.35       52.80      13.21    0.000       1.7
Sex_1Fem      -1397.2       533.2      -2.62    0.010       1.2
Market         2662.3       820.3       3.25    0.001       1.0

S = 2585        R-Sq = 85.3%     R-Sq(adj) = 84.8%

Analysis of Variance
Source            DF          SS          MS         F        P
Regression         5  5585766862  1117153372    167.14    0.000
Residual Error   144   962459821     6683749
Total            149  6548226683
```

Éste es el modelo final. Todas las variables independientes son significativas y el modelo explica una parte significativa de la variabilidad del salario.

b) Para contrastar la hipótesis de que la tasa de variación de los salarios femeninos en función de la edad es menor que la tasa de variación de los salarios masculinos en función de la edad, se utiliza la variable ficticia Sex_1Fem para ver si el coeficiente de la pendiente de la edad (X_1) es diferente en el caso de los hombres y de las mujeres. Se utiliza el siguiente modelo:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_6 X_4)X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \\ &= \beta_0 + \beta_1 X_1 + \beta_6 X_4 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \end{aligned}$$

Crear la variable X_4X_1 y contrastar la significación condicionada en el modelo de regresión. Si se demuestra que es un importante predictor de los salarios, existen pruebas contundentes para concluir que la tasa de variación de los salarios femeninos en función de la edad es diferente de la de los salarios masculinos:

Regression Analysis: Salary Versus age, femage, ...

```
The regression equation is
Salary = 22082 + 85.1 age + 11.7 femage + 543 yrs_asoc
          + 701 yrs_full - 1878 Sex_1Fem + 2673 Market
Predictor        Coef     SE Coef          T        P       VIF
Constant        22082        1877      11.77    0.000
age              85.07       48.36       1.76    0.081       4.4
femage           11.66       63.89       0.18    0.855      32.2
yrs_asoc        542.85       66.73       8.13    0.000       2.6
yrs_full        701.35       57.35      12.23    0.000       2.0
Sex_1Fem         -1878        2687      -0.70    0.486      31.5
Market          2672.8       825.1       3.24    0.001       1.0

S = 2594        R-Sq = 85.3%     R-Sq(adj) = 84.7%
```

La regresión muestra que la variable recién creada de las mujeres no es significativa. Por lo tanto, no podemos concluir que la tasa de variación de los salarios femeninos en función de la edad es diferente de la tasa de variación de los salarios masculinos.

13.116. a) Existe una relación positiva entre EconGPA y todas las variables independientes, como es de esperar. Obsérvese que existe una estrecha correlación entre la puntuación global de la ACT (ACTcomp) y los componentes individuales, de nuevo, como es de esperar. Por lo tanto, es probable que la estrecha correlación entre las variables independientes sea un serio motivo de preocupación en este modelo de regresión.

Regression Analysis: EconGPA Versus sex, Acteng, ...

```
The regression equation is
EconGPA = - 0.050 + 0.261 sex + 0.0099 Acteng + 0.0064 ACTmath
           + 0.0270 ACTss + 0.0419 ACTcomp + 0.00898 HSPct
71 cases used 41 cases contain missing values
Predictor        Coef     SE Coef          T        P       VIF
Constant      -0.0504      0.6554      -0.08    0.939
sex            0.2611      0.1607       1.62    0.109       1.5
Acteng        0.00991     0.02986       0.33    0.741       2.5
ACTmath       0.00643     0.03041       0.21    0.833       4.3
ACTss         0.02696     0.02794       0.96    0.338       4.7
ACTcomp       0.04188     0.07200       0.58    0.563      12.8
HSPct        0.008978    0.005716       1.57    0.121       1.4

S = 0.4971      R-Sq = 34.1%     R-Sq(adj) = 27.9%
```

Como era de esperar, la multicolinealidad afecta a los resultados. La estrategia de eliminar la variable que tiene el menor estadístico *t* con cada modelo sucesivo provoca la eliminación de las siguientes variables (por orden): (1) ACTmath, (2) ACTeng, (3) ACTss, (4) HSPct. Las dos variables que quedan son el modelo final del sexo y ACTcomp:

Regression Analysis: EconGPA Versus sex, ACTcomp

```
The regression equation is
EconGPA = 0.322 + 0.335 sex + 0.0978 ACTcomp
73 cases used 39 cases contain missing values
Predictor        Coef     SE Coef          T        P       VIF
Constant       0.3216      0.5201       0.62    0.538
sex            0.3350      0.1279       2.62    0.011       1.0
ACTcomp       0.09782     0.01989       4.92    0.000       1.0

S = 0.4931      R-Sq = 29.4%     R-Sq(adj) = 27.3%
```

Las dos variables independientes son significativas.

b) El modelo podría utilizarse para las decisiones de admisión creando un GPA predicho en economía basado en el sexo y en las calificaciones globales en ACT (ACTcomp). Este GPA predicho podría utilizarse junto con otros factores para decidir las admisiones. Obsérvese que este modelo predice que las mujeres obtendrán mejores resultados que los hombres que tienen las mismas calificaciones. La utilización de este modelo como única fuente de información puede llevar a acusar de discriminación.

Capítulo 14

14.2. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + e_i$, donde Y_i = salarios, X_1 = años de experiencia, $X_2 = 1$ para Alemania, 0 en caso contrario, $X_3 = 1$ para Gran Bretaña, 0 en caso contrario, $X_4 = 1$ para Japón, 0 en caso contrario, $X_5 = 1$ para Turquía, 0 en caso contrario. La categoría excluida son los salarios de Estados Unidos.

14.4. **a)** Para cualquier observación, los valores de las variables ficticias suman uno. Dado que la ecuación tiene una constante, hay perfecta multicolinealidad y se puede caer en la «trampa de las variables ficticias».

b) β_3 mide la diferencia esperada entre la demanda del primer trimestre y del cuarto, manteniéndose todo lo demás constante. β_4 mide la diferencia esperada entre la demanda del segundo trimestre y del cuarto, manteniéndose todo lo demás constante. β_5 mide la diferencia esperada entre la demanda del tercer trimestre y la del cuarto, manteniéndose todo lo demás constante.

14.6. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$

donde Y_i = ventas de cereales per cápita
X_1 = precio de los cereales
X_2 = precio de los cereales rivales
X_3 = renta media per cápita
X_4 = % de titulados universitarios
X_5 = temperatura anual media
X_6 = precipitaciones anuales medias
X_7 = 1 para las ciudades situadas al este del Misisipi, 0 en caso contrario
X_8 = 1 para la renta per cápita alta, 0 en caso contrario
X_9 = 1 para la renta per cápita intermedia, 0 en caso contrario
X_{10} = 1 para el noroeste, 0 en caso contrario
X_{11} = 1 para el sudoeste, 0 en caso contrario
X_{12} = 1 para el noreste, 0 en caso contrario
$X_{13} = X_1 X_7$: término de interacción entre el precio y las ciudades del este del Misisipi

La especificación del modelo incluye variables independientes continuas, variables individuales dicotómicas y variables ficticias. Aunque la forma funcional puede ser lineal, podría introducirse la no linealidad basándose en un análisis inicial de los diagramas de puntos dispersos de las relaciones. También podría detectarse una estrecha correlación entre las variables independientes, por ejemplo, la renta per cápita y el porcentaje de titulados universitarios podrían muy bien ser colineales. Podrían realizarse varias iteraciones del modelo para hallar las combinaciones óptimas de variables.

14.8. Definir las siguientes variables del experimento:

Y = remuneración de los trabajadores
X_1 = años de experiencia
X_2 = nivel de clasificación del puesto: 1. aprendiz, 2. profesional, 3. maestro
X_3 = capacidad personal
X_4 = sexo: 1. hombre, 2. mujer
X_5 = raza: 1. blanco, 2. negro, 3. latino

Pueden desarrollarse dos variables dependientes a partir de los datos sobre los salarios. El salario base es uno de los análisis que pueden realizarse. También pueden analizarse los complementos salariales. Se necesitan variables ficticias para analizar la influencia de las clasificaciones de los puestos en el salario. La discriminación puede medirse por medio de la magnitud de la variable ficticia del sexo y la raza. Para cada variable ficticia, se necesitan $(k - 1)$ categorías para evitar la «trampa de las variables ficticias».

14.10. **a)** $\dfrac{\beta_j}{(1-\gamma)} = 3{,}03$

b) $\dfrac{\beta_j}{(1-\gamma)} = 3{,}289$

c) $\dfrac{\beta_j}{(1-\gamma)} = 5{,}556$

d) $\dfrac{\beta_j}{(1-\gamma)} = 6{,}515$

14.12.

Regression Analysis: Y Retail Sales Versus X Income, Ylag1

```
The regression equation is
Y  Retail Sales = 1752 + 0.367 X  Income + 0.053 Ylag1
21 cases used 1 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant       1751.6       500.0       3.50    0.003
X  Incom      0.36734     0.08054       4.56    0.000
Ylag1          0.0533      0.2035       0.26    0.796

S = 153.4       R-Sq = 91.7%     R-Sq(adj) = 90.7%
```

$t = 0{,}2619$; $t_{18,\,0{,}10} = 1{,}33$; por lo tanto, no rechazar H_0 al nivel del 20%.

14.14.

Regression Analysis: Y_%stocks Versus X_Return, Y_lag%stocks

```
The regression equation is
Y_%stocks = 1.65 + 0.228 X_Return + 0.950 Y_lag%stocks
24 cases used 1 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant        1.646       2.414       0.68    0.503
X_Return      0.22776     0.03015       7.55    0.000
Y_lag%st      0.94999     0.04306      22.06    0.000

S = 2.351       R-Sq = 95.9%     R-Sq(adj) = 95.5%
```

14.16.

Regression Analysis: Y_Birth Versus X_1stmarriage, Y_lagBirth

```
The regression equation is
Y_Birth = 21262 + 0.485 X_1stmarriage + 0.192 Y_lagBirth

19 cases used 1 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant        21262        5720       3.72    0.002
X_1stmar       0.4854      0.1230       3.94    0.001
Y_lagBir       0.1923      0.1898       1.01    0.326

S = 2513        R-Sq = 93.7%     R-Sq(adj) = 93.0%
```

14.18.

Regression Analysis: Y_logCons Versus X_LogDI, Y_laglogCons

```
The regression equation is
Y_logCons = 0.405 + 0.373 X_LogDI + 0.558 Y_laglogCons

28 cases used 1 cases contain missing values
Predictor        Coef     SE Coef          T        P
Constant       0.4049      0.1051       3.85    0.001
X_LogDI        0.3734      0.1075       3.47    0.002
Y_laglog       0.5577      0.1243       4.49    0.000

S = 0.03023     R-Sq = 99.6%     R-Sq(adj) = 99.6%
Durbin–Watson statistic = 1.63
```

14.20. **a)** En el caso especial en el que la correlación muestral entre x_1 y x_2 es cero, la estimación de β_1 es la misma independientemente de que se incluya o no x_2 en la ecuación de regresión. En la regresión lineal simple de y con respecto a x_1, el término constante recoge la influencia de x_2 en y, en estas circunstancias especiales.

b) $$b_1 = \frac{\Sigma (x_{2i} - \bar{x}_2)^2 \Sigma (x_{1i} - \bar{x})(y_{1i} - \bar{y}) - \Sigma (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \Sigma (x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\Sigma (x_{1i} - \bar{x}_1)^2 \Sigma (x_{2i} - \bar{x}_2)^2 - [\Sigma (x_{1i} - \bar{x}_1) \Sigma (x_{2i} - \bar{x}_2)]^2}$$

Si la correlación muestral entre x_1 y x_0 es cero, entonces $\Sigma (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = 0$ y la ecuación del coeficiente de la pendiente puede simplificarse. El resultado es

$$b_1 = \frac{\Sigma (x_{1i} - \bar{x}_1)(y_{1i} - \bar{y})}{\Sigma (x_{1i} - \bar{x}_1)^2}$$

que es el coeficiente de la pendiente estimado de la regresión lineal bivariante de y con respecto a x_1.

14.22. El modelo inicial de regresión incluye todas las variables independientes indicadas:

Results for: CITYDAT.XLS
Regression Analysis: hseval Versus Comper, Homper, ...

```
The regression equation is
hseval = - 19.0 - 26.4 Comper - 12.1 Homper - 15.5 Indper
           + 7.22 sizehse + 0.00408 incom72
Predictor        Coef     SE Coef          T        P
Constant       -19.02       13.20      -1.44    0.153
Comper        -26.393       9.890      -2.67    0.009
Homper        -12.123       7.508      -1.61    0.110
Indper        -15.531       8.630      -1.80    0.075
sizehse         7.219       2.138       3.38    0.001
incom72      0.004081    0.001555       2.62    0.010

S = 3.949       R-Sq = 40.1%     R-Sq(adj) = 36.5%
Durbin–Watson statistic = 1.03
```

Excluyendo inicialmente las variables poco significativas Homper e Indper y excluyendo el número mediano de habitaciones por vivienda (Sizehse) se obtiene el modelo final:

Regression Analysis: hseval Versus Comper, incom72

```
The regression equation is
hseval = 4.69 - 20.4 Comper + 0.00585 incom72
Predictor        Coef     SE Coef          T        P
Constant        4.693       5.379       0.87    0.385
Comper        -20.432       7.430      -2.75    0.007
incom72      0.005847    0.001484       3.94    0.000

S = 4.352       R-Sq = 24.7%     R-Sq(adj) = 22.9%
Durbin–Watson statistic = 0.98
```

Obsérvese que el coeficiente del porcentaje de locales comerciales de ambos modelos es negativo; sin embargo, es mayor en el segundo en el que se excluye la variable del número mediano de habitaciones.

14.24. Si x_2 influye mucho en y, la eliminación de y de la ecuación de regresión podría introducir un grave sesgo de especificación. En lugar de eliminar la variable, es preferible reconocer que, aunque el grupo en conjunto es claramente influyente, los datos no contienen información que permita distinguir con un cierto grado de precisión los efectos independientes de cada una de las variables explicativas.

14.26. **a)** La comprobación gráfica de la existencia de heterocedasticidad no indica que exista una fuerte heterocedasticidad.

b) La regresión auxiliar es $e^2 = -63310{,}41 + 13{,}75\hat{y}$
$n = 22$, $R^2 = 0{,}06954$, $nR^2 = 1{,}5299 < 2{,}71 = \chi^2_{1,1}$; por lo tanto, no rechazar H_0 de que los términos de error tienen una varianza constante al nivel del 10%.

14.28. **a)**

Regression Analysis: y Versus X1, X2, X3

```
The regression equation is
y = 0.2 + 0.000406 X1 + 4.84 X2 - 1.55 X3

Predictor       Coef    SE Coef      T      P
Constant        0.16      34.91   0.00  0.996
X1         0.0004060  0.0001736   2.34  0.024
X2             4.842      2.813   1.72  0.092
X3           -1.5543     0.3399  -4.57  0.000

S = 3.04752   R-Sq = 54.3%   R-Sq(adj) = 51.4%
```

b) La comprobación gráfica de la existencia de heterocedasticidad no indica que exista una fuerte heterocedasticidad.

c) La regresión auxiliar es $e^2 = 20{,}34 - 0{,}201\hat{y}$
$n = 50$, $R^2 = 0{,}00322$, $nR^2 = 0{,}161 < 2{,}71 = \chi^2_{1,1}$; por lo tanto, no rechazar H_0 de que los términos de error tienen una varianza constante al nivel del 10%.

14.30. $d = 0{,}50$, $\alpha = 0{,}05$: $d_L = 1{,}26$ y $d_U = 1{,}56$, $\alpha = 0{,}01$: $d_L = 1{,}04$ y $d_U = 1{,}32$
Rechazar la hipótesis nula basada en el contraste de Durbin-Watson a los niveles del 5 y el 1%. Estimación del coeficiente de autocorrelación: 0,75

a) $d = 0{,}80$, $\alpha = 0{,}05$: $d_L = 1{,}26$ y $d_U = 1{,}56$, $\alpha = 0{,}01$: $d_L = 1{,}04$ y $d_U = 1{,}32$
Rechazar la hipótesis nula basada en el contraste de Durbin-Watson a los niveles del 5 y el 1%. Estimación del coeficiente de autocorrelación: 0,60

b) $d = 1{,}10$, $\alpha = 0{,}05$: $d_L = 1{,}26$ y $d_U = 1{,}56$, $\alpha = 0{,}01$: $d_L = 1{,}04$ y $d_U = 1{,}32$
Rechazar la hipótesis nula basada en el contraste de Durbin-Watson al nivel del 5%. El contraste no es concluyente al nivel del 1%.
Estimación del coeficiente de autocorrelación: 0,45

c) $d = 1{,}25$, $\alpha = 0{,}05$: $d_L = 1{,}26$ y $d_U = 1{,}56$, $\alpha = 0{,}01$: $d_L = 1{,}04$ y $d_U = 1{,}32$
Rechazar la hipótesis nula basada en el contraste de Durbin-Watson al nivel del 5%. El contraste no es concluyente al nivel del 1%.

d) $d = 1{,}70$, $\alpha = 0{,}05$: $d_L = 1{,}26$ y $d_U = 1{,}56$, $\alpha = 0{,}01$: $d_L = 1{,}04$ y $d_U = 1{,}32$
No rechazar la hipótesis nula ni al nivel del 5% ni al nivel del 1%. No existen pruebas suficientes de que haya una autocorrelación de los residuos.

14.32. Dado que $\text{Var}(\varepsilon_i) = Kx_i^2$ $(K > 0)$

$$\text{Var}(\varepsilon_i/x_i) = \frac{1}{x_i^2}\text{Var}(\varepsilon_i) = \frac{1}{x_i^2}Kx_i^2 = K$$

Si puede hallarse la relación cuadrática entre la varianza de los términos de error y x_i tal que $\text{Var}(\varepsilon_i) = Kx_i^2$, puede eliminarse el problema de heterocedasticidad dividiendo los dos miembros de la ecuación de regresión por x_i.

14.34. El modelo de regresión del ejercicio 14.13 incluye el valor retardado de la variable dependiente como una variable independiente. En presencia de una variable dependiente retardada utilizada como variable independiente, el estadístico de Durbin-Watson ya no es válido. Hay que utilizar el estadístico h de Durbin:

$$H_0: \rho = 0,\ H_1: \rho > 0,\ r = 1 - \frac{d}{2} = 1 - \frac{1{,}65}{2} = 0{,}175,\ s_c^2 = (0{,}1266)^2 = 0{,}0160$$

$$h = r\sqrt{\frac{n}{1 - n(s_c^2)}} = 0{,}175\sqrt{\frac{27}{1 - 27(0{,}0160)}} = 1{,}21,\ z_1 = 1{,}28.$$ No rechazar H_0 al nivel del 10%.

14.36. $d = 0{,}85$, $\alpha = 0{,}05$: $d_L = 1{,}01$ y $d_U = 1{,}78$, $\alpha = 0{,}01$: $d_L = 0{,}80$ y $d_U = 1{,}53$
Rechazar H_0 al nivel del 5%; el contraste no es concluyente al nivel del 1%.

14.38. $d = 0{,}88$, $\alpha = 0{,}01$: $d_L = 1{,}05$ y $d_U = 1{,}21$. Rechazar H_0 al nivel del 1%; por lo tanto, un modelo de regresión mal especificado con una variable omitida puede dar como resultado la presencia de autocorrelación de los residuos.

14.40. **a)** Variables ficticias: las variables ficticias se utilizan siempre que un factor no es fácil de cuantificar. Por ejemplo, si quisiéramos averiguar cómo afectan las barreras comerciales a las tasas de crecimiento de la producción, podríamos incluir una variable ficticia que tomara el valor de uno cuando se imponen barreras comerciales y cero en caso contrario. De esa manera se podría distinguir entre los diferentes niveles de barreras comerciales.
b) Variables dependientes retardadas: las variables dependientes retardadas son útiles cuando se analizan datos de series temporales. Por ejemplo, podríamos incluir las tasas de crecimiento retardadas en un modelo utilizado para explicar las fluctuaciones de producción.
c) Transformación logarítmica: las transformaciones logarítmicas permiten utilizar técnicas estadísticas inherentemente lineales como la regresión lineal por mínimos cuadrados para estimar funciones no lineales. Por ejemplo, las funciones de costes en las que el coste es una función de la producción normalmente son funciones no lineales. La transformación logarítmica nos permite expresar las relaciones no lineales en forma lineal y, por lo tanto, utilizar técnicas de estimación lineal para el modelo.

14.42. La afirmación no es válida. El sumatorio de varias regresiones lineales (simples) bivariantes no es igual a los resultados obtenidos en una regresión múltiple. Por lo tanto, aunque considerar por separado las variables independientes pueda dar alguna idea de la significación estadística de los efectos individuales, no suministra ninguna información sobre la influencia en la variable dependiente cuando las variables independientes se consideran conjuntamente. Es preferible reconocer que el grupo en su conjunto es claramente influyente, pero los datos no son lo suficientemente informativos para poder distinguir con precisión los efectos de cada variable independiente.

14.44. **a)** $t = 1{,}179$. No rechazar H_0 al nivel del 10%, ya que $t < 1{,}282 \approx t_{84,\ 0{,}1}$
b) $t = 0{,}495$. No rechazar H_0 al nivel del 10%, ya que $t < 1{,}282 \approx t_{84,\ 0{,}1}$
c) La diferencia entre los resultados probablemente se deba a la existencia de multicolinealidad entre los beneficios por acción (x_1) y el flujo de fondos por acción (x_2).

14.46. No ha sido posible obtener ninguna información, ya que ninguno de los coeficientes estimados de χ_5 y χ_6 son significativos; por lo tanto, las estimaciones del modelo no son válidas.

14.48. **a)** Manteniéndose todo lo demás constante, un aumento del valor de los nuevos pedidos de un 1% provoca una disminución esperada del número de quiebras del 0,82%.
b) $d = 0{,}49$, $\alpha = 0{,}01$: $d_L = 1{,}01$ y $d_U = 1{,}42$. Rechazar H_0 al nivel del 1%.

c) Dado que los residuos no están autocorrelacionados, los resultados del contraste de hipótesis del apartado b) no son válidos. El modelo debe volver a estimarse teniendo en cuenta los errores autocorrelacionados.

d) $r = 0{,}755$

14.50. **a)** IC al 95%: $0{,}035 < \beta < 0{,}471$

b) Un aumento de 0,253 $ en el periodo actual, otro aumento de 0,138 $ en el siguiente periodo, un aumento de 0,075 $ dentro de dos periodos, etc. Aumento total esperado de 0,557 $.

c) Obsérvese que, debido a la presencia de una variable dependiente retardada utilizada como variable independiente, el estadístico h de Durbin es relevante.
$h = 0{,}56449$, $z_1 = 1{,}28$; por lo tanto, no rechazar H_0 al nivel del 10%.

14.52.

Regression Analysis: y_log Versus x1_log, x2_log

```
The regression equation is
y_log = - 2.14 + 0.909 x1_log + 0.195 x2_log

Predictor        Coef     SE Coef          T        P
Constant      -2.1415      0.2000     -10.71    0.000
x1_log        0.90947     0.03518      25.85    0.000
x2_log        0.19451     0.07126       2.73    0.018

S = 0.07721     R-Sq = 99.6%     R-Sq(adj) = 99.5%
Durbin–Watson statistic = 1.67
```

$d = 1{,}67$, $\alpha = 0{,}05$: $d_L = 0{,}95$ y $d_U = 1{,}54$. No rechazar H_0 al nivel del 5%.

14.54.

Regression Analysis: y_log Versus x1_log, x2_log, y_laglog_1

```
The regression equation is
y_log = 0.435 - 0.101 x1_log + 0.237 x2_log + 0.666 y_laglog_1
34 cases used 1 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant       0.4352      0.4360       1.00    0.326
x1_log       -0.10116     0.03822      -2.65    0.013
x2_log         0.2365      0.1017       2.32    0.027
y_laglog       0.6658      0.1174       5.67    0.000

S = 0.04039     R-Sq = 75.1%     R-Sq(adj) = 72.6%
Durbin–Watson statistic = 2.22
```

$h = -0{,}8798$, p-valor = 0,3788; no rechazar H_0 a los niveles habituales de alfa.

14.56.

Regression Analysis: y_log Versus x1_log, x2_log, x3_log

```
The regression equation is
y_log = 2.72 - 0.0252 x1_log + 0.315 x2_log + 0.379 x3_log

Predictor        Coef     SE Coef          T        P
Constant      2.71584     0.08821      30.79    0.000
x1_log       -0.02519     0.04049      -0.62    0.543
x2_log        0.31472     0.05689       5.53    0.000
x3_log         0.3788      0.2009       1.89    0.078

S = 0.03611     R-Sq = 91.7%     R-Sq(adj) = 90.2%
Durbin–Watson statistic = 1.75
```

$d = 1{,}75$, $\alpha = 0{,}05$: $d_L = 1{,}00$ y $d_U = 1{,}68$, $\alpha = 0{,}01$: $d_L = 0{,}77$ y $d_U = 1{,}41$. No rechazar H_0 al nivel del 1% o al nivel del 5%.

14.58. a)

Regression Analysis: CSH Versus GDPH

```
The regression equation is
CSH = - 207 + 0.417 GDPH
214 cases used 4 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant     -207.440       6.920     -29.98    0.000
GDPH         0.416931    0.001430     291.66    0.000

S = 44.42       R-Sq = 99.8%     R-Sq(adj) = 99.8%
Durbin-Watson statistic = 0.11
```

$d = 0{,}11$, $\alpha = 0{,}01$: $d_L = 1{,}52$ y $d_U = 1{,}56$

Rechazar H_0 al nivel del 1% y aceptar la alternativa de que existe una autocorrelación positiva de primer orden significativa en los residuos.

El modelo muestra un poder explicativo extraordinariamente grande ($R^2 = 99{,}8\%$); sin embargo, hay una autocorrelación significativa en los residuos ($d = 0{,}11$).

b)

Regression Analysis: CSH Versus GDPH, FBPR, CSH_lag

```
The regression equation is
CSH = - 4.30 + 0.0178 GDPH - 0.504 FBPR + 0.965 CSH_lag
210 cases used 8 cases contain missing values

Predictor        Coef     SE Coef          T        P
Constant       -4.302       2.661      -1.62    0.108
GDPH         0.017760    0.004441       4.00    0.000
FBPR          -0.5040      0.1676      -3.01    0.003
CSH_lag       0.96547     0.01077      89.64    0.000

S = 6.976       R-Sq = 100.0%    R-Sq(adj) = 100.0%

Durbin-Watson statistic = 1.66
```

$r = 0{,}17$, $s_c^{22} = 0{,}000116$, $h = 2{,}494$, p-valor $= 0{,}0128$; por lo tanto, no rechazar H_0 al nivel del 1%; rechazar al nivel del 5%.

La inclusión del valor retardado de la variable dependiente como variable independiente ha reducido el problema de la autocorrelación de los residuos; sin embargo, es probable que como consecuencia haya multicolinealidad entre las variables independientes.

14.60. a)

Regression Analysis: hseval Versus sizehse, taxrate, totexp, Comper

```
The regression equation is
hseval = - 23.4 + 9.21 sizehse - 178 taxrate +0.000001 totexp
         - 20.4 Comper
Predictor        Coef     SE Coef          T        P
Constant      -23.433       8.986      -2.61    0.011
sizehse         9.210       1.564       5.89    0.000
taxrate       -177.53       39.87      -4.45    0.000
totexp     0.00000142  0.00000030       4.80    0.000
Comper        -20.370       6.199      -3.29    0.001

S = 3.400       R-Sq = 55.1%     R-Sq(adj) = 52.9%
Durbin-Watson statistic = 1.20
```

Dado que todas las variables independientes son estadísticamente significativas, dejar todas las variables independientes en el modelo de regresión.

b) La regresión auxiliar es:

Regression Analysis: ResiSq Versus FITS1

```
The regression equation is
ResiSq = - 15.1 + 1.24 FITS1

Predictor        Coef     SE Coef          T        P
Constant       -15.09       11.96      -1.26    0.210
FITS1          1.2370      0.5604       2.21    0.030

S = 19.44       R-Sq = 5.2%      R-Sq(adj) = 4.2%
```

$e^2 = -15{,}1 + 1{,}24\hat{y}$, $n = 90$, $R^2 = 0{,}052$, $nR^2 = 4{,}68 > 3{,}84 = \chi^2_{1,\,0{,}05}$; por lo tanto, rechazar la hipótesis nula de que los términos de error tienen una varianza constante al nivel del 5% y el economista tiene razón en que es probable que la heterocedasticidad sea un problema.

c) Transformar las variables utilizando la población como valor ponderado y volver a hacer el modelo de regresión múltiple.

Regression Analysis: hseval_pop Versus sizehse_pop, taxrate_pop, ...

```
The regression equation is
hseval_pop = 0.000570 + 4.58 sizehse_pop - 158 taxrate_pop
             - 0.000002 totexp_pop - 24.5 comper_pop
Predictor           Coef     SE Coef      T      P
Constant       0.0005700   0.0001863   3.06  0.003
sizehse_pop       4.5845      0.2726  16.82  0.000
taxrate_pop       -157.52       33.47  -4.71  0.000
totexp_pop   -0.00000212  0.00000139  -1.52  0.133
comper_pop        -24.503       4.900  -5.00  0.000

S = 0.000338490   R-Sq = 86.3%   R-Sq(adj) = 85.7%
```

R^2 aumenta de 52,9% a 86,3% y todas las variables independientes siguen siendo significativas con la excepción de la variable totexp_pop.

Capítulo 15

15.2. H_0: $P = 0{,}50$ (no mejoran en general los niveles de comprensión tras la participación en el programa)
H_1: $P > 0{,}50$ (el nivel de comprensión aumenta gracias al programa)
$n = 9$. En el caso de 8 aumentos «Después» del programa y un contraste unilateral, $P(X \geqslant 8) = = 0{,}0176 + 0{,}002 = 0{,}0196$
Por lo tanto, rechazar H_0 a los niveles de alfa superiores a 1,96%.

15.4. H_0: $P = 0{,}50$ (los rendimientos positivos y negativos son igual de probables)
H_1: $P > 0{,}50$ (los rendimientos positivos son más probables)
$\hat{p} = 0{,}6842$, $\mu = 28{,}5$, $\sigma = 3{,}7749$, $z = 2{,}65$
p-valor = 0,0040. Por lo tanto, rechazar H_0 a los niveles de alfa superiores a 0,40%.

15.6. H_0: $P = 0{,}50$ (los economistas están divididos por igual en esta cuestión)
H_1: $P \neq 0{,}50$ (en caso contrario)
$\hat{p} = 0{,}5918$
$\mu = 24{,}5$, $\sigma = 3{,}50$, $z = 1{,}14$
p-valor = 0,2542. Por lo tanto, rechazar H_0 a los niveles de alfa superiores a 25,42%.

15.8. H_0: ninguna preferencia por la cerveza nacional frente a la importada
H_1: se prefiere la cerveza importada.
Contraste de Wilcoxon basado en la ordenación de las diferencias: Diff_15.8

Contraste de la mediana = 0,000000 frente a mediana < 0,000000

	N	N del contraste	Estadístico de Wilcoxon	P	Mediana estimada
Diff_15.8	10	9	7,0	0,038	–1,500

$n = 9$, $T = 7{,}0$, $T_{0,05} = 9$. Por lo tanto, rechazar H_0 a los niveles de alfa superiores a 3,8%.

15.10. H_0: los dos cursos se consideran igual de interesantes.
H_1: el curso de estadística se considera más interesante.
$z = -1{,}73$, p-valor = 0,0418. Por lo tanto, rechazar H_0 a los niveles superiores a 4,18%.

15.12. H_0: tiempo dedicado por igual.
H_1: tiempo no dedicado por igual.
$z = -0{,}57$, p-valor = 0,5686. Por lo tanto, no rechazar H_0 a ningún nivel habitual.

15.14. H_0: no hay diferencia entre los rendimientos.
H_1: la «lista de compra» tiene un rendimiento porcentual mayor (un contraste de una cola)
Suma de los puestos de la «lista de compra» = 137
$z = -2{,}42$, p-valor = 0,0078. Por lo tanto, rechazar H_0 a los niveles superiores a 0,78%.

15.16. H_0: ninguna preferencia entre los expertos en marketing y los expertos en economía financiera
H_1: se prefieren los expertos en economía financiera (contraste de una cola).
$z = -0{,}234$, p-valor = 0,4090. Por lo tanto, rechazar H_0 a los niveles superiores a 40,9%.

15.18. H_0: las tasas porcentuales de rendimiento son iguales.
H_1: los fondos mejor calificados logran mayores tasas de rendimiento
$z = 0{,}64$, p-valor = 0,2611. Por lo tanto, rechazar H_0 a los niveles superiores a 26,11%.

15.20. H_0: el tiempo en días que se tarda en publicar desde finales de año un informe preliminar sobre los beneficios es el mismo en las empresas en las que los informes de auditoría son buenos y en las que no son buenos.
H_1: las empresas cuyos informes de auditoría no son buenos tardan más.
$z = 1{,}86$, p-valor = 0,0314. Por lo tanto, rechazar H_0 a los niveles superiores a 3,14%.

15.22. **a)** Obtener las ordenaciones de las dos variables

Ordenación de los exámenes	Ordenación de los proyectos
6	5,5
1	3,0
4	2,0
5	5,5
10	9,0
2	1,0
3	8,0
7	4,0
9	10,0
8	7,0

Por lo tanto, la correlación de Pearson entre las ordenaciones de las variables es el coeficiente de correlación de orden de Spearman:

Correlations: RankExam, RankProject
Correlación de Pearson entre RankExam y RankProject = 0,717

b) H_0: ninguna relación entre la calificación en el examen y la calificación en el proyecto
H_1: existe una relación (contraste de dos colas)
$n = 10$, $r_{s,\,0{,}025} = 0{,}648$, $r_{s,\,0{,}010} = 0{,}745$
Por lo tanto, rechazar H_0 de que no existe ninguna relación entre las dos variables al nivel de 0,05, pero no al nivel de 0,02 (contraste de dos colas).

15.24. Los contrastes no paramétricos no postulan ningún supuesto sobre la conducta de la distribución de la población. Las ventajas de los contrastes son que los supuestos son menos restrictivos y que pueden utilizarse contrastes más fáciles de calcular utilizando datos nominales u ordinales. Se da menos peso a los casos atípicos.

15.26. H_0: $P = 0{,}50$ (las ventas del año que viene serán iguales que las de este año)
H_1: $P \neq 0{,}50$ (en caso contrario)
$n = 9$. En el caso de 2 «a favor» y un contraste de dos colas, $P(2 \geqslant X \geqslant 7) = 0{,}1798$. Por lo tanto, rechazar H_0 a los niveles de alfa superiores a 17,98%.

15.28. H_0: $P = 0{,}50$ (más estudiantes esperan tener un nivel de vida más alto)
H_1: $P < 0{,}50$ (más estudiantes esperan tener un nivel de vida más bajo que el de sus padres)
$z = -0{,}79$, p-valor = 0,2148. Por lo tanto, rechazar H_0 a los niveles de alfa superiores a 21,48%.

15.30. H_0: los analistas de empresas son más optimistas sobre las perspectivas de sus propias empresas que sobre la economía en general.
H_1: en caso contrario (contraste de una cola)
$T = 11$. Vemos en la Tabla 10 del apéndice, $T_{0,10} = 9$. Por lo tanto, no rechazar H_0 al nivel del 10%.

Capítulo 16

16.2. H_0: el rendimiento de los fondos de inversión tiene la misma probabilidad de estar en los 5 quintiles de rendimiento.
H_1: en caso contrario

Fondos de inversión	20% superior	2.º 20%	3.er 20%	4.º 20%	5.º 20%	Total
Número observado	13	20	18	11	13	75
Probabilidad (H_0)	0,2	0,2	0,2	0,2	0,2	1
Número esperado	15	15	15	15	15	75
Cálculo de la ji-cuadrado	0,266667	1,666667	0,6	1,066667	0,266667	3,8667

$\chi^2 = 3{,}8667$, $\chi^2_{(4,\,0{,}1)} = 7{,}78$. Por lo tanto, no rechazar H_0 al nivel del 10%.

16.4. H_0: la calidad de la producción se ajusta a la pauta habitual.
H_1: en caso contrario

Componente electrónico	Ningún defecto	1 defecto	> 1 defecto	Total
Número observado	458	30	12	500
Probabilidad (H_0)	0,93	0,05	0,02	1
Número esperado	465	25	10	500
Cálculo de la ji-cuadrado	0,105376344	1	0,4	1,505376

$\chi^2 = 1{,}505$, $\chi^2_{(2,\,0{,}05)} = 5{,}99$ $\chi^2_{(2,\,0{,}10)} = 4{,}61$. Por lo tanto, no rechazar H_0 al nivel del 5% o del 10%.

16.6. H_0: la opinión de los estudiantes sobre los cursos de administración de empresas es igual que su opinión sobre todos los cursos.
H_1: en caso contrario

Opinión	Muy útiles	Algo	Inútiles	Total
Número observado	68	18	14	100
Probabilidad (H_0)	0,6	0,2	0,2	1
Número esperado	60	20	20	100
Cálculo de la ji-cuadrado	1,066666667	0,2	1,8	3,066667

$\chi^2 = 3{,}067$, $\chi^2_{(2,\ 0{,}10)} = 4{,}61$. Por lo tanto, no rechazar H_0 al nivel del 10%.

16.8. H_0: los consumidores tienen las mismas preferencias por las 5 bebidas refrescantes
H_1: en caso contrario

Drink16-8	A	B	C	D	E	Total
Número observado	20	25	28	15	27	115
Probabilidad (H_0)	0,2	0,2	0,2	0,2	0,2	1
Número esperado	23	23	23	23	23	115
Cálculo de la ji-cuadrado	0,391304	0,173913	1,086957	2,782609	0,695652	5,130435

$\chi^2 = 5{,}130$, $\chi^2_{(4,\ 0{,}10)} = 7{,}78$. Por lo tanto, no rechazar H_0 al nivel del 10%.

16.10. H_0: las preferencias de los profesores de estadística están repartidas por igual entre los 4 paquetes.
H_1: en caso contrario

Programas informáticos	M	E	S	P	Total
Número observado	100	80	35	35	250
Probabilidad (H_0)	0,25	0,25	0,25	0,25	1
Número esperado	62,5	62,5	62,5	62,5	250
Cálculo de la ji-cuadrado	22,5	4,9	12,1	12,1	51,6

$\chi^2 = 51{,}6$, $\chi^2_{(3,\ 0{,}005)} = 12{,}84$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

16.12. H_0: la distribución poblacional de las llegadas por minuto es una distribución de Poisson.
H_1: en caso contrario

Llegadas	0	1	2	3	4+	Total
Número observado	10	26	35	24	5	100
Probabilidad (H_0)	0,1496	0,2842	0,27	0,171	0,1252	1
Número esperado	14,96	28,42	27	17,1	12,52	100
Cálculo de la ji-cuadrado	1,6445	0,2061	2,3704	2,7842	4,5168	11,522

$\chi^2 = 11{,}52$, $\chi^2_{(3,\ 0{,}01)} = 11{,}34$ $\chi^2_{(3,\ 0{,}005)} = 12{,}84$. Por lo tanto, rechazar H_0 al nivel del 1%, pero no al nivel del 0,5%.

16.14. H_0: la resistencia de los componentes electrónicos sigue una distribución normal.
H_1: en caso contrario
$B = 9{,}625$. Tabla 16.7: puntos críticos del estadístico de Bowman-Shelton; el punto al 5% ($n = 100$) es 4,29. Por lo tanto, rechazar H_0 al nivel del 5%.

16.16. H_0: los saldos mensuales de titulares de una determinada tarjeta de crédito siguen una distribución normal.
H_1: en caso contrario
$B = 6{,}578$. Tabla 16.7: puntos críticos del estadístico de Bowman-Shelton; el punto al 5% ($n = 125$) es 4,34. Por lo tanto, rechazar H_0 al nivel del 5%.

16.18. a) H_0: no existe ninguna relación entre la calificación media y la facultad.
H_1: en caso contrario
Ji-cuadrado = 0,226 + 0,276 + 0,341 + 0,417 + 2,227 + 2,722 = 6,209

GL = 2, p-valor = 0,045, $\chi^2_{(2,\ 0,05)} = 5,99$. Por lo tanto, rechazar H_0 de que no existe ninguna relación al nivel del 5%.

16.20. a)

Método por el que se enteraron de la existencia del nuevo producto			
Edad	Amigos	Anuncios	Total columnas
<21	30	20	50
21–35	60	30	90
35+	18	42	60
Total filas	108	92	200

b) H_0: no existe ninguna relación entre el método por el que se enteraron de la existencia del nuevo producto y su edad.
H_1: en caso contrario
Ji-cuadrado = 0,333 + 0,391 + 2,674 + 3,139 + 6,400 + 7,513 = 20,451
GL = 2, p-valor = 0,000, $\chi^2_{(2,\ 0,005)} = 10,6$. Por lo tanto, rechazar H_0 de que no existe ninguna relación al nivel del 5%.

16.22. H_0: no existe ninguna relación entre las amortizaciones de activos y las fusiones.
H_1: en caso contrario
Ji-cuadrado = 0,527 + 0,286 + 0,514 + 0,279 = 1,607
GL = 1, p-valor = 0,205, $\chi^2_{(1,\ 0,10)} = 2,71$. Por lo tanto, no rechazar H_0 al nivel del 10%.

16.24. H_0: no existe ninguna relación entre la valoración del personal y la carrera estudiada.
H_1: en caso contrario
Ji-cuadrado = 0,186 + 0,010 + 0,188 + 0,943 + 0,008 + 1,867 + 0,620 + 0,843 + + 0,022 + 4,814 + 0,543 + 3,605 = 13,648
GL = 6, p-valor = 0,034, 1 casilla en la que se espera que el número de casos sea inferior a 5,0 $\chi^2_{(6,\ 0,05)} = 12,59$. Por lo tanto, rechazar H_0 al nivel del 5%.

16.26. H_0: no existe ninguna relación entre los programas de doctorado y la carrera estudiada
H_1: en caso contrario
Ji-cuadrado = 8.000 + 1,815 + 1,667 + 8,000 + 1,815 + 1,667 = 22,963
GL = 2, p-valor = 0,000, $\chi^2_{(2,\ 0,005)} = 10,60$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

16.28. H_0: no existe ninguna relación entre las preferencias por los candidatos en las elecciones primarias y el distrito de votación.
H_1: en caso contrario
Ji-cuadrado = 0,660 + 0,168 + 1,565 + 0,578 + 0,098 + 3,235 + 1,174 + 0,117 + 0,196 + + 0,878 + 0,065 + 0,743 = 9,478
GL = 6, p-valor = 0,148, $\chi^2_{(6,\ 0,10)} = 10,64$. Por lo tanto, no rechazar H_0 al nivel del 10%.

16.30. H_0: no existe ninguna relación entre los años de experiencia y las piezas producidas por hora.
H_1: en caso contrario
Ji-cuadrado = 0,000 + 5,000 + 5,000 + 0,000 + 0,000 + 0,000 + 0,000 + 5,000 + + 5,000 = 20,000
GL = 4, p-valor = 0,000, $\chi^2_{(4,0,005)} = 14,86$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

16.32. a) H_0: no existe ninguna relación entre el peso del paquete y la procedencia.
H_1: en caso contrario
Ji-cuadrado = 0,123 + 0,313 + 1,201 + 24,779 + 1,429 + 21,420 + 43,973 + 0,068 + + 70,301 + 2,635 + 1,500 + 11,852 = 179,594
GL = 6, p-valor = 0,000, $\chi^2_{(6,\ 0,005)} = 18,55$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

b) Las combinaciones que tienen la mayor diferencia porcentual entre la frecuencia observada y la esperada son 1) entre las fábricas y los paquetes de 11-75 kilos y 2) entre las fábricas y los paquetes de menos de 3 kilos.

16.34. H_0: no existe ninguna relación entre la antigüedad de la empresa y la opinión del propietario sobre la eficacia de las firmas digitales.
H_1: en caso contrario
Ji-cuadrado = 1,070 + 0,533 + 0,478 + 2,796 + 1,987 + 0,311 + 0,489 + 0,542 + 0,016 = = 8,222
GL = 4, p-valor = 0,084, $\chi^2_{(4,\ 0,05)} = 9,49$ $\chi^2_{(4,\ 0,10)} = 7,78$. Por lo tanto, no rechazar H_0 al nivel del 5%, pero sí al nivel del 10%.

16.36. H_0: no existe ninguna relación entre la razón para trasladarse a Florida y el tipo de sector.
H_1: en caso contrario
Ji-cuadrado = 2,858 + 0,386 + 3,320 + 1,156 + 0,321 + 0,887 + 7,495 + 1,424 + 7,362 = = 25,210
GL = 4 p-valor = 0,000, $\chi^2_{(4,\ 0,005)} = 14,86$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

16.38. H_0: no existe ninguna relación entre las opiniones de que deben controlarse más rigurosamente los anuncios de los productos de adelgazamiento y el consumo de un producto de adelgazamiento rápido.
H_1: en caso contrario
Ji-cuadrado = 6,700 + 7,086 + 9,410 + 9,952 = 33,148
GL = 1, p-valor = 0,000, $\chi^2_{(1,\ 0,005)} = 7,88$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

16.40. H_0: no existe ninguna diferencia entre las preferencias actuales de los clientes y las preferencias anteriores.
H_1: en caso contrario

	A	B	C	D
Frecuencia observada	56	70	28	126
Frecuencia esperada	56	92,4	56	75,6
$(O_i - E_i)^2/E_i$	0	5,43	14	33,6

Estadístico del contraste de la ji-cuadrado = 53,03, $\chi^2_{(3,\ 0,005)} = 12,84$. Por lo tanto, rechazar H_0 al nivel del 0,5%.

16.42. **a)** H_0: no existe ninguna relación entre el curso en el que se encuentran los estudiantes y sus opiniones sobre la ampliación del horario de apertura de la biblioteca.
H_1: en caso contrario
Aquí hemos utilizado el programa Minitab y hemos incluido solamente las respuestas de 340 estudiante que tenían opinión sobre la ampliación del horario de apertura de la biblioteca.

Tabulated statistics: Class, Hours **Extension**

```
Rows: Class   Columns: Hours Extension
                  Yes      No     All
1                  86      53     139
                98.12   40.88  139.00

2                  79      21     100
                70.59   29.41  100.00

3                  46      15      61
                43.06   17.94   61.00

4                  29      11      40
                28.24   11.76   40.00
Missing             0       1       *
                    *       *       *
All               240     100     340
               240.00  100.00  340.00
Cell Contents:      Count
                    Expected count
Pearson Chi-Square = 9.250, DF = 3, p-value = 0.026
Likelihood Ratio Chi-Square = 9.262, DF = 3, P-Value = 0.026
```

$\chi^2_{(3,\ 0{,}025)} = 9{,}35$

No rechazar H_0 al nivel del 2,5%.

b) Entre las recomendaciones se encuentran orientar mejor a los estudiantes de primer año con el fin de que conozcan mejor la biblioteca y las horas en que está abierta. También estaría bien ampliar el horario de apertura, especialmente durante los momentos en que más se utiliza.

16.44. Las respuestas varían.

16.46. H_0: no existe ninguna relación entre el método para hacer la declaración de la renta y la edad de la persona.

H_1: en caso contrario

```
Chi-Sq = 7.143, DF = 4, p-value = 0.129
```

$\chi^2_{(4,\ 0{,}10)} = 7{,}78$. Por lo tanto, no rechazar H_0 al nivel del 10%.

No existe ninguna relación estadísticamente significativa entre el método para hacer la declaración de la renta y la edad de la persona.

Capítulo 17

17.2. $MCG = \dfrac{879}{3}$, $MCD = \dfrac{798}{16}$, $F = \dfrac{293}{49{,}875} = 5{,}87$

$F_{3,\ 16,\ 0{,}05} = 3{,}24$, $F_{3,\ 16,\ 0{,}01} = 5{,}29$

Por lo tanto, rechazar H_0 al nivel del 1%.

17.4. **a)** $\bar{x}_1 = 62$, $\bar{x}_2 = 53$, $\bar{x}_3 = 52$, $SCD = 3.608$, $SCG = 340{,}9375$, $STC = 3.948{,}9375$

b)

One-Way ANOVA: SodaSales Versus CanColor

```
Analysis of Variance for SodaSale
Source     DF        SS        MS        F        P
CanColor    2       341       170     0.61    0.556
Error      13      3608       278
Total      15      3949
```

$F_{2,\ 13,\ 0{,}05} = 3{,}81$; no rechazar H_0 al nivel del 5%.

17.6. **a)** $\bar{x}_1 = 32$, $\bar{x}_2 = 24{,}3333$, $\bar{x}_3 = 34{,}8333$

One-Way ANOVA: Nonconforming Versus Supplier

```
Analysis of Variance for Nonconfo
Source     DF        SS        MS        F        P
Supplier    2     354.1     177.1    10.45    0.001
Error      15     254.2      16.9
Total      17     608.3
```

b) $F_{2,\ 15,\ 0{,}01} = 6{,}36$; rechazar H_0 al nivel del 1%.

17.8. **a)** $\bar{x}_1 = 71{,}7143$, $\bar{x}_2 = 75{,}2857$, $\bar{x}_3 = 76{,}5714$

One-Way ANOVA: Scores Versus Class

```
Analysis of Variance for Scores
Source     DF        SS        MS        F        P
Class       2        89        44     0.28    0.756
Error      18      2813       156
Total      20      2901
```

b) $F_{2,\ 18,\ 0{,}05} = 3{,}55$; no rechazar H_0 al nivel del 5%.

17.10. **a)** $\bar{x}_1 = 11{,}3333$, $\bar{x}_2 = 12{,}5$, $\bar{x}_3 = 8$

One-Way ANOVA: Time Versus Rank

```
Analysis of Variance for Time
Source     DF        SS        MS        F        P
Rank        2     51.40     25.70     3.27    0.074
Error      12     94.33      7.86
Total      14    145.73
```

b) $F_{2,\,12,\,0{,}05} = 3{,}89$; no rechazar H_0 al nivel del 5%.

17.12. **a)** $\bar{x}_1 = 10{,}4017$, $\bar{x}_2 = 7{,}0450$, $\bar{x}_3 = 6{,}7767$

One-Way ANOVA: Fog Versus Mag

```
Analysis of Variance for Fog
Source     DF        SS        MS        F        P
Mag         2     48.96     24.48     4.07    0.039
Error      15     90.13      6.01
Total      17    139.09
```

b) $F_{2,\,15,\,0{,}05} = 3{,}68$; rechazar H_0 al nivel del 5%.

17.14. **a)** $\hat{\mu} = 8{,}0744$
b) $\hat{G}_1 = 2{,}3273$, $\hat{G}_2 = -1{,}0294$, $\hat{G}_3 = -1{,}2977$
c) $\hat{\varepsilon}_{32} = 0{,}7483$

17.16. $W = 8{,}32$, $\chi^2_{(2,\,0{,}05)} = 5{,}99$; por lo tanto, rechazar H_0 al nivel del 5%.

17.18. $W = 1{,}18$, $\chi^2_{(2,\,0{,}10)} = 4{,}61$; por lo tanto, no rechazar H_0 al nivel del 5%.

Kruskal–Wallis Test: SodaSales Versus CanColor

```
Kruskal–Wallis Test on SodaSale
CanColor    N    Median    Ave Rank         Z
1           6     60.00        10.2      1.08
2           5     52.00         7.4     -0.62
3           5     53.00         7.6     -0.51
Overall    16                   8.5
H = 1.18  DF = 2  P = 0.554
H = 1.19  DF = 2  P = 0.553 (adjusted for ties)
```

17.20. $W = 9{,}3772$, $\chi^2_{(2,\,0{,}01)} = 9{,}21$; por lo tanto, rechazar H_0 al nivel del 1%.

Kruskal–Wallis Test: Nonconforming Versus Supplier

```
Kruskal–Wallis Test on Nonconfo
Supplier    N    Median    Ave Rank         Z
1           6     32.00        10.6      0.61
2           6     24.50         4.3     -2.90
3           6     35.00        13.6      2.29
Overall    18                   9.5
H = 9.38  DF = 2  P = 0.009
H = 9.47  DF = 2  P = 0.009 (adjusted for ties)
```

17.22. $W = 0{,}7403$, $\chi^2_{(2,\,0{,}10)} = 4{,}61$; por lo tanto, no rechazar H_0 al nivel del 10%.

17.24. $W = 5{,}2452$, $\chi^2_{(2,\,0{,}10)} = 4{,}61$; por lo tanto, rechazar H_0 al nivel del 10%.

17.26. **a)** La hipótesis nula contrasta la igualdad de las medias poblacionales de las valoraciones realizadas por los distintos grupos.
b) $W = 0{,}17$, $\chi^2_{(2,\,0{,}10)} = 4{,}61$; por lo tanto, no rechazar H_0 al nivel del 10%.

17.28. Contraste de la igualdad de las medias de los H bloques en los que se divide la población: $\frac{MCB}{MCE} = 3{,}597$, $F_{5,\,30,\,0{,}05} = 2{,}53$, $F_{5,\,30,\,0{,}01} = 3{,}70$. Rechazar al nivel del 5%, no rechazar H_0 al nivel del 1% que las medias de los bloques son diferentes.

Contraste de la igualdad de las medias de los K grupos en los que se divide la población: $\frac{MCG}{MCE} = 4{,}91$, $F_{6,\,30,\,0{,}05} = 2{,}42$, $F_{6,\,30,\,0{,}01} = 3{,}47$. Rechazar H_0 al nivel del 1%. Los datos sugieren que las medias de los grupos difieren.

17.30. a)

Two-Way ANOVA: earngrowth Versus OilCo, Analyst

```
Analysis of Variance for earngrow
Source        DF        SS        MS        F        P
OilCo          4      3.30      0.83     0.31    0.866
Analyst        3     31.35     10.45     3.93    0.036
Error         12     31.90      2.66
Total         19     66.55
```

b) $F_{4,\,12,\,0{,}05} = 3{,}26 > 0{,}31$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.32. a)

Two-Way ANOVA: sales Versus Quarter, soup

```
Analysis of Variance for sales
Source        DF        SS        MS        F        P
Quarter        3     615.0     205.0     2.10    0.202
Soup           2       6.2       3.1     0.03    0.969
Error          6     586.5      97.7
Total         11    1207.7
```

b) $F_{2,\,6,\,0{,}05} = 5{,}14 > 0{,}03$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.34. a)

Two-Way ANOVA: Ratings Versus Exam, Text

```
Analysis of Variance for Ratings
Source        DF        SS        MS        F        P
Exam           2    0.2022    0.1011     5.20    0.077
Text           2    0.4356    0.2178    11.20    0.023
Error          4    0.0778    0.0194
Total          8    0.7156
```

b) [libros de texto]: $F_{2,\,4,\,0{,}05} = 6{,}94 < 11{,}20$; por lo tanto, rechazar H_0 al nivel del 5%.
c) [tipo de examen]: $F_{2,\,4,\,0{,}05} = 6{,}94 > 5{,}20$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.36. $\hat{G}_3 = -0{,}1556$
$\hat{B}_1 = 0{,}1778$
$\hat{\varepsilon}_{31} = 0{,}0556$

17.38. a) Completar la tabla ANOVA:

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Fertilizantes	135,6	3	45,20	6,0916
Tipos de suelo	81,7	5	16,34	2,2022
Error	111,3	15	7,42	
Total	328,6	23		

b) [fertilizantes]: $F_{3,\ 15,\ 0,01} = 5,42 < 6,0916$; por lo tanto, rechazar H_0 al nivel del 1%.
c) [tipos de suelo]: $F_{5,\ 15,\ 0,05} = 2,90 > 2,2021$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.40. Dados, por ejemplo, diez pares de observaciones, el estadístico F tendría 1, 9 grados de libertad. El contraste es un contraste de dos colas. Suponiendo que alfa = 0,05, el valor crítico de F sería 5,12. En el caso de un contraste de pares enlazados, los grados de libertad serían 9 y el área de cada cola sería 0,025. El valor crítico de t sería 2,262 (que es la raíz cuadrada del estadístico F de 5,12). Por lo tanto, los dos contrastes son equivalentes.

17.42. *Cociente F: interacción* $= \dfrac{MCI}{MCE} = 0,67$, $F_{20,\ 90,\ 0,05} \approx 1,75$, $F_{20,\ 90,\ 0,01} \approx 2,20$. No rechazar H_0 al nivel del 5%. No existe ninguna interacción significativa entre los grupos de tratamiento A y B. Por lo tanto, continuar contrastando los principales efectos de cada grupo de tratamiento.

Cociente F: tratamiento A $= \dfrac{MSG_A}{MCE} = 5,73$, $F_{4,\ 80,\ 0,05} \approx 2,53$, $F_{4,\ 80,\ 0,01} \approx 3,65$. Rechazar H_0 al nivel del 1%; hay un efecto significativo en el caso del grupo A.

Cociente F: tratamiento B $= \dfrac{MCG_B}{MCE} = 4,00$, $F_{5,\ 80,\ 0,05} \approx 2,37$, $F_{5,80,\ 0,01} \approx 3,34$. Rechazar H_0 al nivel del 1%; hay un efecto significativo en el caso del grupo B.

17.44. **a)** Tabla ANOVA:

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Participantes	364,50	21	17,3571	19,2724
Jueces	0,81	8	0,1013	0,1124
Interacción	4,94	168	0,0294	0,0326
Error	1.069,94	1.188	0,9006	
Total	1.440,19	1.385		

H_0: las medias poblacionales del valor de los 22 participantes son iguales.
H_1: en caso contrario
$F_{21,\ 1.188,\ 0,01} \approx 1,88 < 19,2724$; por lo tanto rechazar H_0 al nivel del 1%.
H_0: las medias poblacionales del valor de los 9 jueces son iguales.
H_1: en caso contrario
$F_{8,\ 1.188,\ 0,05} \approx 1,94 > 0,1124$; por lo tanto, no rechazar H_0 al nivel del 5%.
H_0: no hay ninguna interacción entre los participantes y los jueces.
H_1: en caso contrario
$F_{168,\ 1.188,\ 0,05} \approx 1,22 > 0,0326$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.46. **a)** Tabla ANOVA:

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Tipo de test	57,5556	2	28,7778	4,7091
Sujeto	389,0000	3	129,6667	21,2182
Interacción	586,0000	6	97,66667	15,9818
Error	146,6667	24	6,1111	
Total	1.179,2223	35		

b) H_0: no existe ninguna interacción entre el tipo de sujeto y el tipo de test.
H_1: en caso contrario
$F_{6,\ 24,\ 0{,}01} = 3{,}67 < 15{,}9818$; por lo tanto, rechazar H_0 al nivel del 1%.

17.48. a) El supuesto implícito es que no hay ningún efecto de interacción entre el año de estudios y la valoración de las residencias.

b)

General Linear Model: Ratings Versus Dorm, Year

```
Factor     Type Levels Values
Dorm      fixed      4 A B C D
Year      fixed      4 1 2 3 4

Analysis of Variance for Ratings_, using Adjusted SS for Tests
Source     DF     Seq SS     Adj SS     Adj MS       F      P
Dorm        3     20.344     20.344      6.781    4.91  0.008
Year        3     10.594     10.594      3.531    2.56  0.078
Error      25     34.531     34.531      1.381
Total      31     65.469
```

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente *F*
Residencia	20,344	3	6,781	4,91
Año	10,594	5	3,531	2,56
Error	34,531	25	1,381	
Total	65,469	31		

c) H_0: las medias poblacionales de las valoraciones de las 4 residencias son iguales.
H_1: en caso contrario
$F_{3,\ 25,\ 0{,}01} = 4{,}68 < 4{,}91$; por lo tanto, rechazar H_0 al nivel del 1%.

d) H_0: las medias de las valoraciones de los 4 estudiantes son iguales.
H_1: en caso contrario
$F_{3,\ 25,\ 0{,}05} = 2{,}99 > 2{,}56$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.50.

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente *F*
Color	243,250	2	121,625	11,3140
Región	354,000	3	118,000	10,9767
Interacción	189,750	6	31,625	2,9419
Error	129,000	12	10,750	
Total	916,000	23		

H_0: no existe ninguna interacción entre la región y el color de las latas.
H_1: en caso contrario. $F_{6,\ 12,\ 0{,}01} = 4{,}82 > 2{,}9419$; por lo tanto, no rechazar H_0 al nivel del 1%.

17.52. Un ANOVA de un factor examina el efecto de un único factor (que tiene tres o más condiciones). Un ANOVA bifactorial reconoce situaciones en las que puede ser significativo más de un factor. Ejemplos de ANOVA de un factor son la duración de las baterías de cuatro tipos de teléfonos móviles, el tiempo que tardan en recibir los platos que se piden en cinco restaurantes de comida rápida y el salario de partida de los estudiantes de cuatro especialidades distintas. Ejemplos de ANOVA bifactorial son las diferencias entre las calificaciones medias de cuatro especialidades por sexo, las ventas semanales de un artículo en una tienda de alimenta-

ción según el lugar en el que se coloca en los expositores (alto, medio, bajo) y el tamaño de los anuncios que promueven un producto (grande, medio, pequeño), la fuerza del hormigón según cuatro tipos de cemento y 2 métodos para mezclarlo.

17.54.

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Entre	5.156	2	2.578,000	21,4458
Dentro	120.802	1.005	120,201	
Total	125.967	1.007		

$F_{2,\ 1.005,\ 0,01} = 4{,}61 < 21{,}4458$; por lo tanto, rechazar H_0 al nivel del 1%.

17.56. a)

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Entre	221,3400	3	73,7800	25,6
Dentro	374,6640	130	2,8820	
Total	596,0040	133		

b) H_0: las medias poblacionales de los sueldos de los abogados de los 4 grupos son iguales.
H_1: en caso contrario
$F_{3,\ 130,\ 0,01} \approx 3{,}95 < 25{,}6$; por lo tanto, rechazar H_0 al nivel del 1%.

17.58.

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Entre	11.438,3028	2	5.719,1514	0,7856
Dentro	109.200,000	15	7.280,000	
Total	120.638,3028	17		

H_0: las medias poblacionales de los niveles de ventas de los tres periodos son iguales.
H_1: en caso contrario
$F_{2,\ 15,\ 0,05} = 3{,}68 > 0{,}7856$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.60. $W = 5{,}0543$, $\chi^2_{3,\ 0,10} = 6{,}25$; por lo tanto, no rechazar H_0 al nivel del 10%.

17.62. a) $SCD = \sum_{j=1}^{K} \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

$$= \sum_{j=1}^{K} \left[\sum_{i=1}^{n_i} x_{ij}^2 - 2n_i\bar{x}_i^2 + n_i\bar{x}_i^2 \right]$$

$$= \sum_{j=1}^{K} \sum_{i=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{K} n_i\bar{x}_i^2$$

b) $SCG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x})^2$

$$= \sum_{i=1}^{K} n_i\bar{x}_i^2 - 2\bar{x}\sum_{i=1}^{k} n_i\bar{x}_i + n\bar{x}^2$$

$$= \sum_{i=1}^{K} n_i\bar{x}_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$= \sum_{i=1}^{K} n_i\bar{x}_i^2 - n\bar{x}^2$$

c) $STC = \sum_{i=1}^{K}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$

$$= \sum_{i=1}^{K}\left[\sum_{i=1}^{n_i} x_{ij}^2 - 2\bar{x}\sum_{i=1}^{n_i} x_j + n_i\bar{x}_i^2\right]$$

$$= \sum_{i=1}^{K}\sum_{i=1}^{n_i} (x_{ij}^2 - 2n_i\bar{x}\bar{x}_i + n_ix^2)$$

$$= \sum_{i=1}^{K}\sum_{j=1}^{n_j} x_{ij}^2 - n\bar{x}^2$$

17.64.

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Consumidores	37.571,5	124	302,996	1,3488
Marcas	32.987,3	2	16.493,65	73,4226
Error	55.710,7	248	224,6399	
Total	126.269,5	374		

H_0: las medias poblacionales de los niveles de percepciones de las tres marcas son iguales.
H_1: en caso contrario
$F_{2,\ 248,\ 0,01} \approx 4{,}79 < 73{,}4226$; por lo tanto, rechazar H_0 al nivel del 1%.

17.66.

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Renta	0,0067	2	0,0033	0,2000
Examen de acceso a la universidad	0,8267	2	0,4133	24,8000
Error	0,0667	4	0,0167	
Total	0,9000	8		

H_0: las medias poblacionales de las calificaciones medias de los tres grupos de renta son iguales.
H_1: en caso contrario
$F_{2,\ 4,\ 0,05} = 6{,}94 > 0{,}2000$; por lo tanto, no rechazar H_0 al nivel del 5%.
H_0: las medias poblacionales de las calificaciones medias de los tres grupos de notas de acceso a la universidad son iguales.
H_1: en caso contrario
$F_{2,\ 4,\ 0,01} = 18{,}0 < 24{,}8$; por lo tanto, rechazar H_0 al nivel del 1%.

Two-way ANOVA: GPA versus SAT, Income

```
Source     DF        SS         MS      F      P
SAT         2  0.826667  0.413333  24.80  0.006
Income      2  0.006667  0.003333   0.20  0.826
Error       4  0.066667  0.016667
Total       8  0.900000

S = 0.1291   R-Sq = 92.59%   R-Sq(adj) = 85.19%

                 Individual 95% CIs For Mean Based on
                 Pooled StDev
SAT        Mean    +---------+---------+---------+---------
High    3.36667                    (------*------)
Mod     2.90000    (------*------)
VeryH   3.63333                              (------*------)
                   +---------+---------+---------+---------
                 2.70      3.00      3.30      3.60

                    Individual 95% CIs For Mean Based on
                    Pooled StDev
Income        Mean  ------+---------+---------+---------+---
High       3.33333      (-------------*-------------)
Low        3.26667  (-------------*-------------)
Mod        3.30000    (-------------*-------------)
                    ------+---------+---------+---------+---
                        3.15      3.30      3.45      3.60
```

17.68. **a)** $\mu = 3{,}333$ **b)** $\hat{G}_2 = 0{,}0$ **c)** $\hat{\beta}_1 = 0{,}0667$ **d)** $\hat{\varepsilon}_{21} = 0{,}1333$

17.70. **a)**

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Precios	0,178	2	0,0890	0,0944
Países	4,365	2	2,1825	2,3151
Interacción	1,262	4	0,3155	0,3347
Error	93,330	99	0,9427	
Total	99,135	107		

H_0: las medias poblacionales de las valoraciones de la calidad correspondientes a los tres niveles de precios son iguales.
H_1: en caso contrario
$F_{2,\ 99,\ 0{,}05} \approx 3{,}07 > 0{,}0944$; por lo tanto, no rechazar H_0 al nivel del 5%.
H_0: las medias poblacionales de las valoraciones de la calidad correspondientes a los tres países son iguales.
H_1: en caso contrario
$F_{2,\ 99,\ 0{,}05} \approx 3{,}07 > 2{,}3151$; por lo tanto, no rechazar H_0 al nivel del 5%.
H_0: No existe ninguna interacción entre el precio y el país.
H_1: en caso contrario
$F_{4,\ 99,\ 0{,}05} \approx 2{,}45 > 0{,}3347$; por lo tanto, no rechazar H_0 al nivel del 5%.

17.72. **a)**

Fuente de variación	Suma de los cuadrados	gl	Media de los cuadrados	Cociente F
Renta	0.0178	2	0,0089	0,5333
Nota de acceso a la universidad	2,2011	2	1,1006	66,0333
Interacción	0,1022	4	0,0256	1,5333
Error	0,1500	9	0,0167	
Total	2,4711	17		

H_0: las medias poblacionales de las calificaciones medias de los tres grupos son iguales.
H_1: en caso contrario
$F_{2,\,9,\,0,05} = 4,26 > 0,5333$; por lo tanto, no rechazar H_0 al nivel del 5%.
H_0: las medias poblacionales de las calificaciones medias de los tres grupos de notas del examen de acceso a la universidad son iguales.
H_1: en caso contrario
$F_{2,\,9,\,0,01} = 8,02 < 66,0333$; por lo tanto, rechazar H_0 al nivel del 1%.
H_0: no existe ninguna interacción entre la renta y el grupo de notas del examen de acceso a la universidad.
H_1: en caso contrario
$F_{4,\,9,\,0,05} = 3,63 > 1,5333$; por lo tanto, rechazar H_0 al nivel del 5%.

Capítulo 18

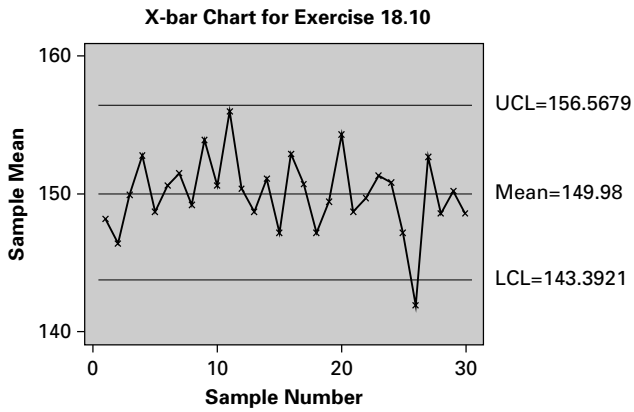
18.2. Varias respuestas

18.4. Varias respuestas

18.6. **a)** $\hat{\sigma} = 5,6517$
b) $LC = \bar{\bar{x}} = 192,6$, $LCI = 186,2044$, $LCS = 198,9956$
c) $LC = 5,42$, $LCI = 0,6504$, $LCS = 10,1896$

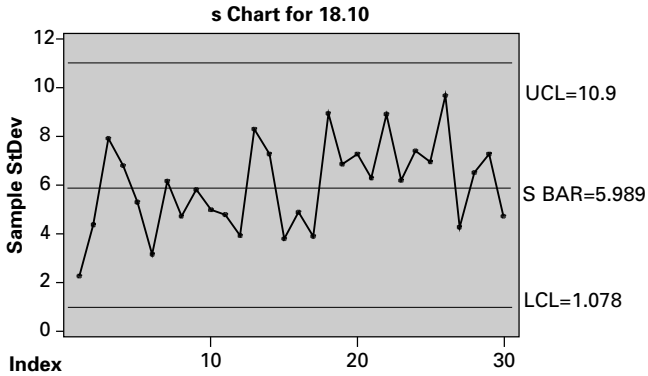
18.8. **a)** $\hat{\sigma} = 1,2746$
b) $LC = 19,86$, $LCI = 18,507$, $LCS = 21,213$
c) $LC = 1,23$, $LCI = 0,2214$, $LCS = 2,2386$

18.10. **a)** $\bar{\bar{x}} = 149,98$
b) $\bar{s} = 5,989$
c) $\hat{\sigma} = 6,2062$
d) $LC = 149,98$, $LCI = 143,3921$, $LCS = 156,5679$
e) El gráfico $\bar{X}$ muestra que la 26.ª observación se encuentra por debajo del *LCI*. Sería una ocurrencia excepcional en un proceso que está bajo control.



f) $LC = 5,989$, $LCI = 1,078$, $LCS = 10,9$

g) Gráfico *s*



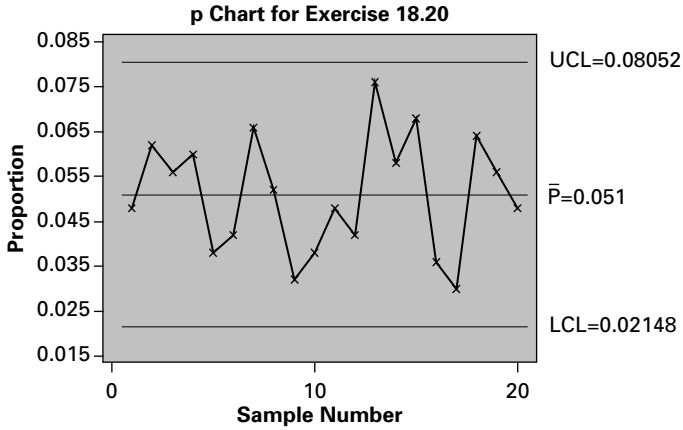
18.12. **a)** (175,6449, 209,5551). Estos valores se encuentran dentro de los límites de tolerancia.
b) $C_p = 1{,}327$. El proceso no es capaz, ya que $C_p < 1{,}33$
c) $C_{pk} = 1{,}321$. El proceso no es capaz, ya que $C_{pk} < 1{,}33$

18.14. **a)** (16,0362, 23,6838). Estos límites están más allá de las tolerancias fijadas por la dirección.
b) $C_p = 0{,}523$. El proceso no es capaz, ya que $C_p < 1{,}33$
c) $C_{pk} = 0{,}486$. El proceso no es capaz, ya que $C_{pk} < 1{,}33$

18.16. **a)** (13,5691, 26,1109). Estos límites están más allá de las tolerancias fijadas por la dirección.
b) $C_p = 0{,}638$. El proceso no es capaz, ya que $C_p < 1{,}33$
c) $C_{pk} = 0{,}612$. El proceso no es capaz, ya que $C_{pk} < 1{,}33$

18.18. $LC = 0{,}018$, $LCI = 0$, $LCS = 0{,}0358$

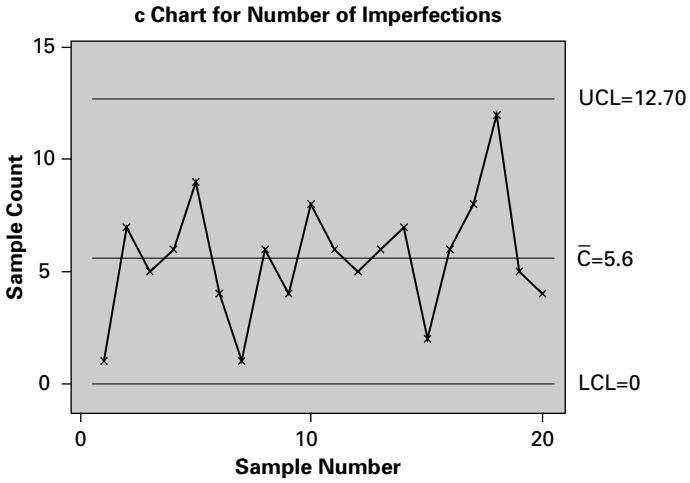
18.20. **a)** $\bar{p} = 0{,}051$, $LC = 0{,}051$, $LCI = 0{,}0215$, $LCS = 0{,}0805$
b) Gráfico *p*



No hay nada que indique que el proceso está fuera de control estadístico.

18.22. **a)** $\bar{c} = 5.6$
b) $LC = 5{,}6$, $LCI = 0$, $LCS = 12{,}70$

c) Gráfico c

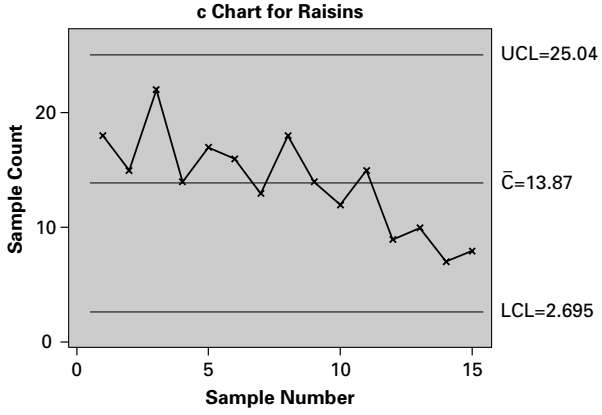


No hay nada que indique que el proceso está fuera de control estadístico.

18.24. **a)** $\bar{c} = 13{,}8667$

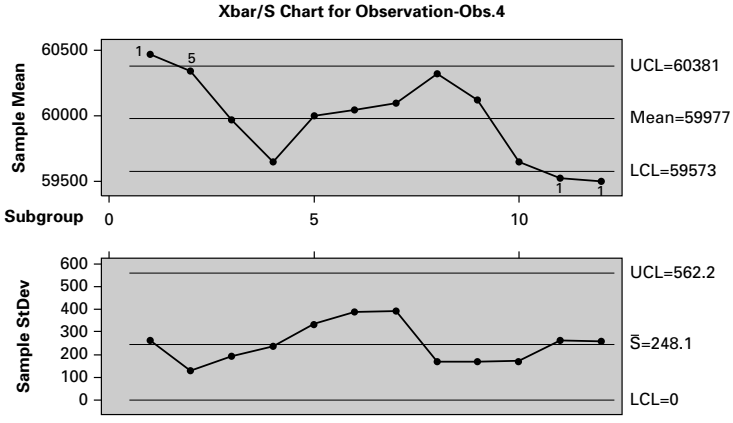
b) $LC = 13{,}8667$, $LCI = 2{,}6953$, $LCS = 25{,}0381$

c) Gráfico c



No hay nada que indique que el proceso está fuera de control; sin embargo, debe continuar vigilándose debido a la reciente tendencia descendente.

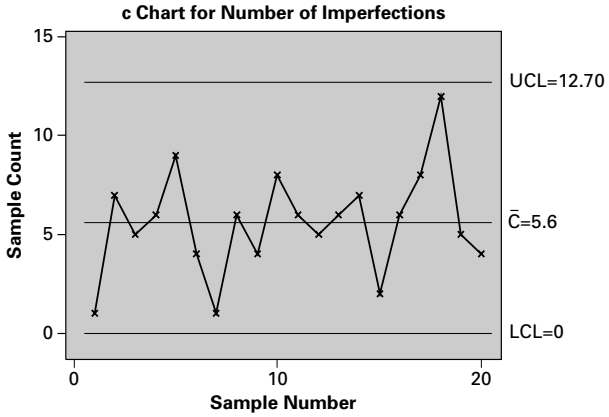
18.26. En los gráficos de control del ejercicio 18.25 y en los siguientes, vemos que el proceso no es estable.



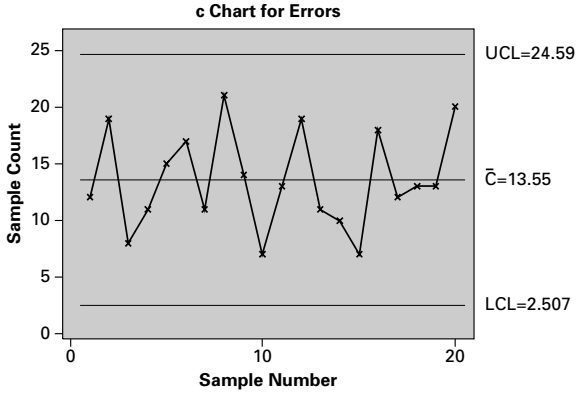
a) El «capability analysis» (normal) no tiene sentido, ya que el proceso de producción de tornillos de precisión no es estable (véase el ejercicio 18.25).

b) El «capability sixpack» (normal) no tiene sentido, ya que el proceso de producción de tornillos de precisión no es estable (véase el ejercicio 18.25).

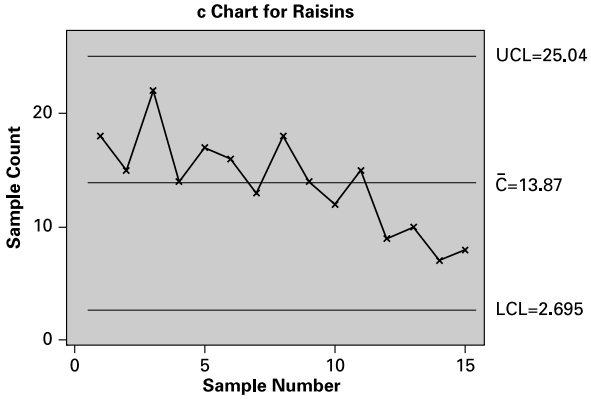
18.28. a)



b)



c)

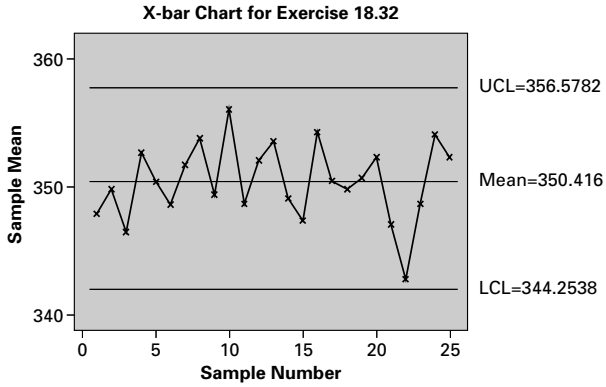


18.30. Pueden cometerse dos tipos de errores: (1) identificar una causa especial o asignable de variación cuando no hay ninguna y (2) no tener en cuenta una causa especial suponiendo que se debe a la variabilidad natural. Si se imponen unos límites de control «demasiado estrechos», se marcan incorrectamente las causas naturales de la variabilidad como una causa especial y el investigador busca una causa especial de la variabilidad que no existe. Los límites de control que son «demasiado amplios» implican que el investigador no va a corregir los procesos que están fuera de control. La utilización de límites de tres sigmas como límites de control fue establecida por Shewhart como un punto intermedio adecuado entre los dos tipos de errores.

18.32. **a)** $\bar{\bar{x}} = 350{,}416$ **b)** $\bar{s} = 5{,}602$ **c)** $\sigma = 5{,}8052$

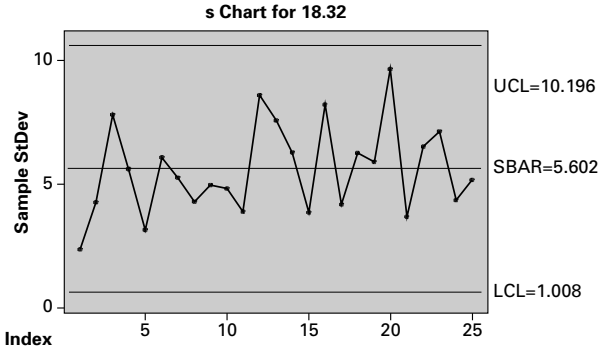
d) $LC = 350{,}416$, $LCI = 344{,}2538$, $LCS = 356{,}5782$

e) El gráfico X-barra no contiene ninguna prueba de que el proceso esté fuera de control.



f) $LC = 5{,}602$, $LCI = 1{,}0084$, $LCS = 10{,}1956$

g) Gráfico s



No se viola ninguna regla del análisis de las pautas. El proceso está bajo control.

h) i) (333,0004, 367,8316) dentro de las tolerancias fijadas por la dirección.

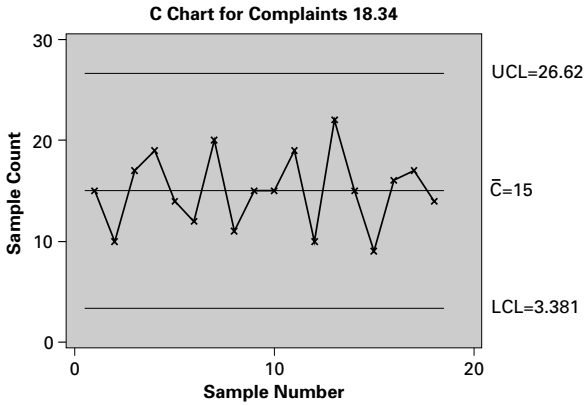
ii) $C_p = 1{,}435 > 1{,}33$. Por lo tanto, el proceso es capaz.

iii) $C_{pk} = 1{,}412 > 1{,}33$. Por lo tanto, el proceso es capaz.

18.34. **a)** $\bar{c} = 15$

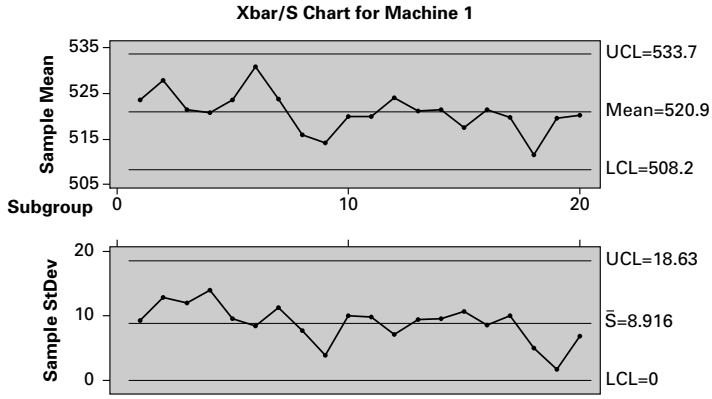
b) $LC = 15$, $LCI = 3{,}381$, $LCS = 26{,}619$

c), d) El gráfico c no contiene ninguna prueba de que un proceso esté fuera de control estadístico.



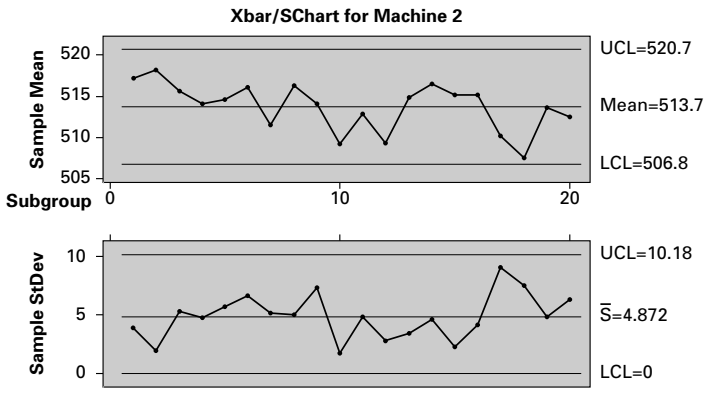
18.36. **a)** Causa común: afecta a todos los trabajadores del proceso
b) Causa común: afecta a todos los trabajadores del proceso
c) Causa asignable
d) Causa asignable
e) Causa asignable

18.38. **a)** Máquina 1: gráfico *X*-barra—*c*:



La máquina 1 no muestra ninguna prueba de estar «fuera de control estadístico».

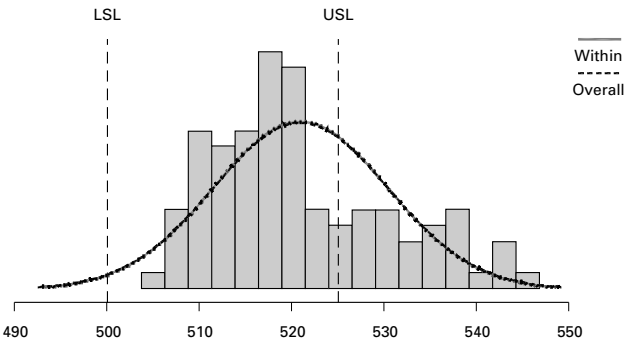
b) Máquina 2: gráfico *X*-barra—*c*:



La máquina 2 tampoco muestra ninguna prueba de estar «fuera de control estadístico».

c) Máquina 1: el «capability analysis» muestra que $C_p = 0{,}44$ y $C_{pk} = 0{,}14$. La máquina 1 no es capaz de cumplir las especificaciones.

Process Capability Analysis for Machine 1



Process Data	
USL	525.000
Target	*
LSL	500.000
Mean	520.948
Sample N	100
StDev (Within)	9.40444
StDev (Overall)	9.47539

Potential (Within) Capability	
Cp	0.44
CPU	0.14
CPL	0.74
Cpk	0.14
Cpm	*

Overall Capability	
Pp	0.44
PPU	0.14
PPL	0.74
Ppk	0.14

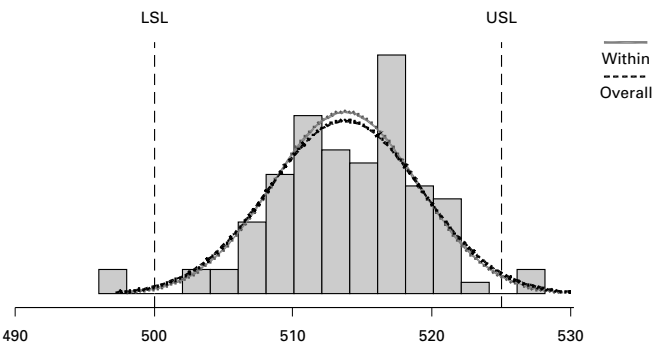
Observed Performance	
PPM < LSL	0.00
PPM > USL	310000.00
PPM Total	310000.00

Exp. "Within" Performance	
PPM < LSL	12958.32
PM > USL	333284.96
PPM Total	346243.28

Exp. "Overall" Performance	
PPM < LSL	13525.58
PPM > USL	334458.84
PPM Total	347984.42

d) Máquina 2: el «capability analysis» muestra que $C_p = 0{,}80$ y $C_{pk} = 0{,}72$. La máquina 2 no es capaz de cumplir las especificaciones.

Process Capability Analysis for Machine 2



Process Data	
USL	525.000
Target	*
LSL	500.000
Mean	513.747
Sample N	100
StDev (Within)	5.23400
StDev (Overall)	5.49945

Potential (Within) Capability	
Cp	0.80
CPU	0.72
CPL	0.88
Cpk	0.72
Cpm	*

Overall Capability	
Pp	0.76
PPU	0.68
PPL	0.83
Ppk	0.68

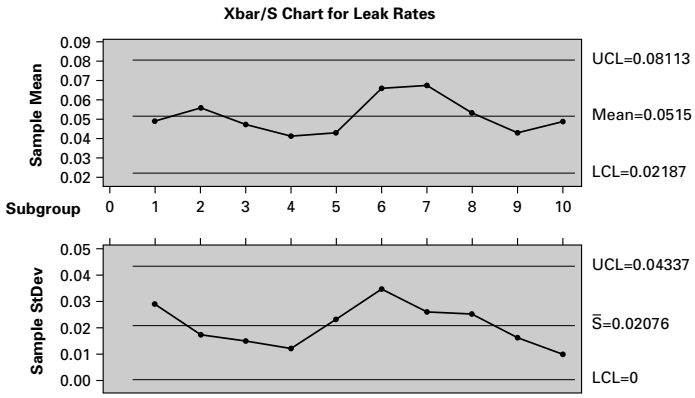
Observed Performance	
PPM < LSL	20000.00
PPM > USL	20000.00
PPM Total	40000.00

Exp. "Within" Performance	
PPM < LSL	4313.63
PPM > USL	15778.32
PPM Total	20091.95

Exp. "Overall" Performance	
PPM < LSL	6214.86
PPM > USL	20368.15
PPM Total	26583.01

e) Ninguna de las dos máquinas es capaz de cumplir las especificaciones. Las dos producen un producto con una variabilidad mayor de lo que exigen los límites de las especificaciones. Obsérvese que la máquina 1 tiene más variabilidad que la 2.

18.40. Gráfico X-barra de los datos **TOC**:



Todos los puntos de datos se encuentran dentro de los límites de control. No se ha violado ninguna regla de análisis de las pautas.

Capítulo 19

19.2. 100,0, 122,6, 123,5, 134,5, 142,5, 140,4, 152,2, 161,2, 188,1, 163,4

19.4. **a)** 100, 102,5, 99,29, 98,21, 100, 99,64, 100, 99,29, 99,29, 100,71, 110,71, 106,07
b) 101,82, 104,36, 101,09, 100, 101,82, 101,45, 101,82, 101,09, 101,09, 102,55 112,73, 108

19.6. **a)** 100, 105,37, 109,6, 112,71, 115,54, 117,23
b) 100, 104,81, 110,49, 112,14, 115,71, 117,47

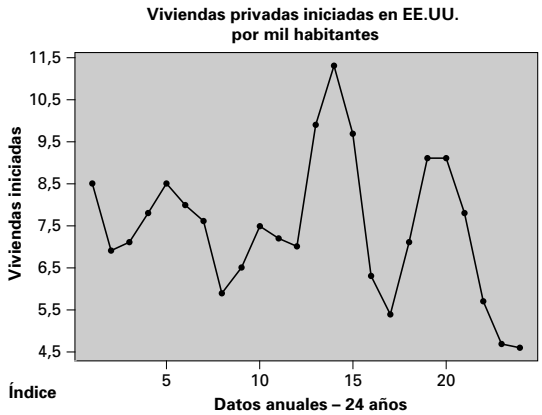
19.8. Un índice de precios de la energía es útil en el sentido de que nos permite decir algo sobre la evolución de los precios de un grupo de mercancías, a saber, los precios de la energía. Un índice ponderado de precios nos permite comparar el coste de un grupo de productos en un periodo con su coste en otros.

19.10. **a)** $Z = -3{,}43\ P(Z < -3{,}43) = 0{,}0003$ **b)** $Z = -2{,}57\ P(Z < -2{,}57) = 0{,}0051$
c) $Z = 3{,}43\ P(Z > 3{,}43) = 0{,}0003$

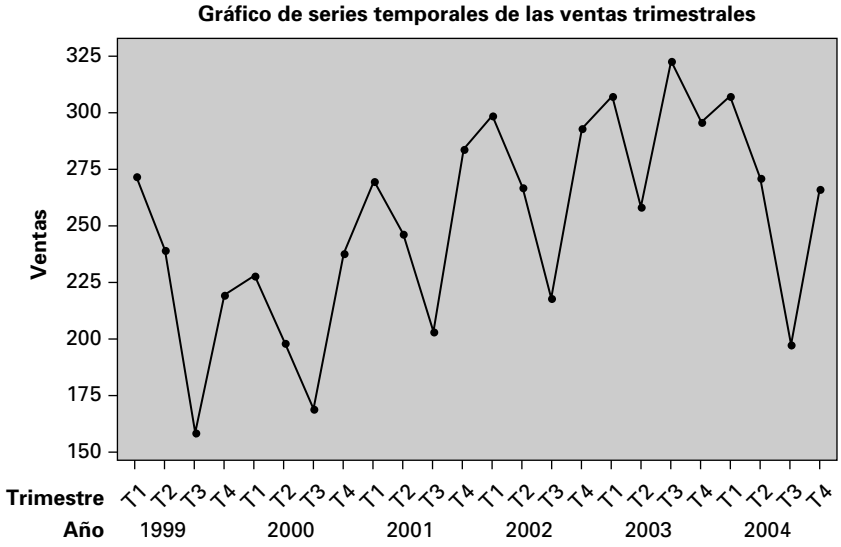
19.12. $R = 7$; no se puede rechazar H_0 a cualquier nivel habitual de significación.

19.14. $R = 9$; no se puede rechazar H_0 a cualquier nivel habitual de significación.

19.16. **a)** $R = 10$; no se puede rechazar H_0 a cualquier nivel habitual de significación.
b) En el gráfico de series temporales adjunto no se observa ninguna conducta cíclica significativa.

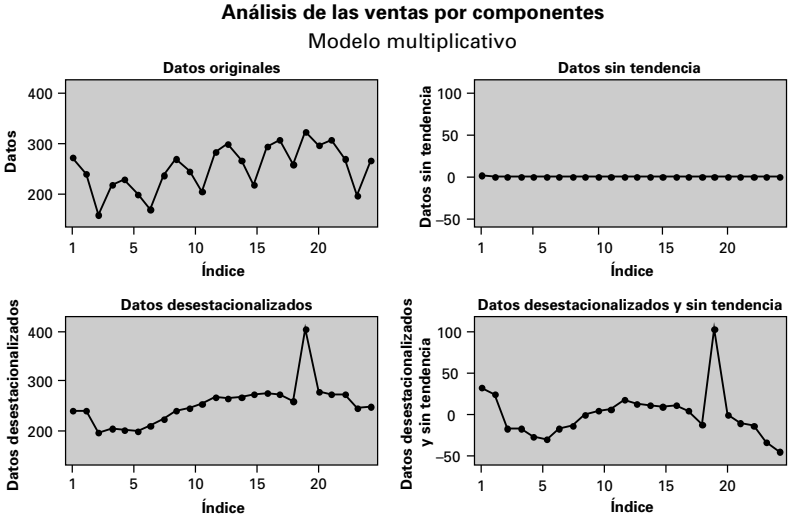


19.18. a) Gráfico de series temporales de las ventas trimestrales
En el gráfico de series temporales se observan pautas evidentes en los datos; fuerte estacionalidad y fuerte tendencia ascendente



b)

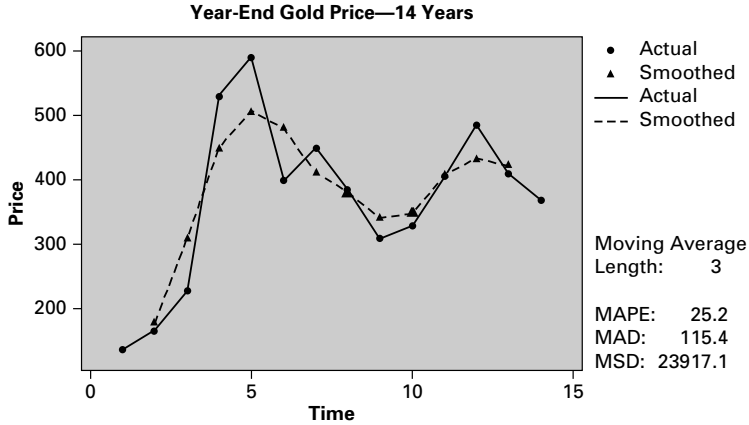
Periodo	MM de 4 periodos	$100\,\frac{X_t}{X_t^*}$	Factor estacional	Serie desestacionalizada
1-1			112,848	241,032
2			99,609	239,938
3	216,500	72,979	79,826	197,930
4	205,875	106,375	107,716	203,312
2-1	202,125	112,802		202,041
2	205,875	96,175		198,777
3	213,500	79,157		211,709
4	224,750	105,895		220,951
3-1	235,000	114,894		239,259
2	245,000	100,408		246,966
3	254,375	79,803		254,302
4	260,625	108,969		263,656
4-1	265,125	112,777		264,958
2	268,125	99,580		268,048
3	270,250	80,666		273,093
4	270,125	108,468		272,011
5-1	270,750	113,389		272,047
2	272,875	94,549		259,013
3	273,250	84,904		290,631
4	274,875	107,685		274,796
6-1	272,125	112,816		272,047
2	264,000	102,652		272,064
3				246,786
4				246,945



Los datos desestacionalizados ya no muestran el ciclo trimestral regular. Hay un punto atípico en el tercer trimestre de 2003. El valor es mucho más alto de lo esperado.

19.20. Media móvil centrada de 3 periodos - precio del oro a finales de año

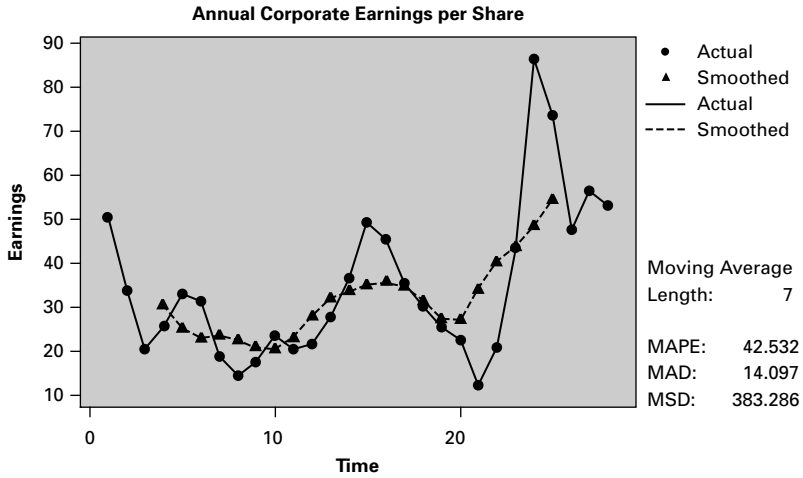
Año	Media móvil de 3 puntos
1	*
2	176,000
3	308,667
4	450,333
5	507,667
6	480,000
7	411,333
8	381,000
9	340,667
10	347,333
11	406,667
12	433,667
13	421,667
14	*



Los datos resultantes muestran una fuerte conducta cíclica.

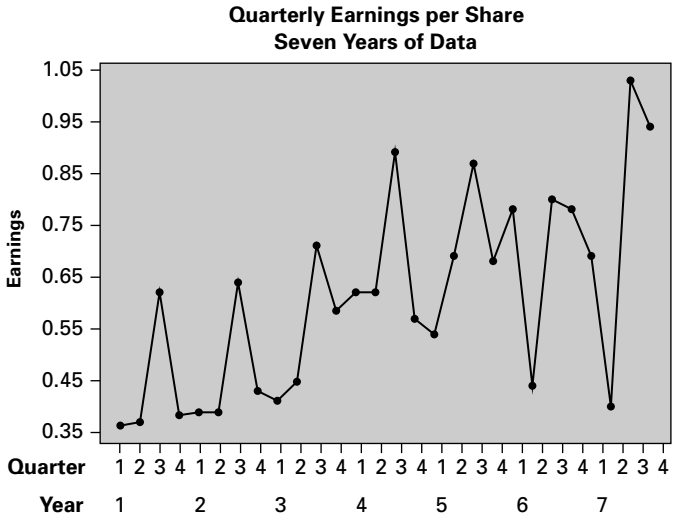
19.22.

Año	MV de 7 puntos
1	*
2	*
3	*
4	30,4429
5	25,3429
6	23,0000
7	23,4286
8	22,7429
9	21,1286
10	20,6000
11	23,1286
12	28,1000
13	32,0857
14	33,8000
15	35,1571
16	35,7143
17	34,9714
18	31,5143
19	27,4571
20	27,0857
21	34,3286
22	40,5429
23	43,7143
24	48,6000
25	54,4286
26	*
27	*
28	*



Los datos suavizados muestran una pauta cíclica.

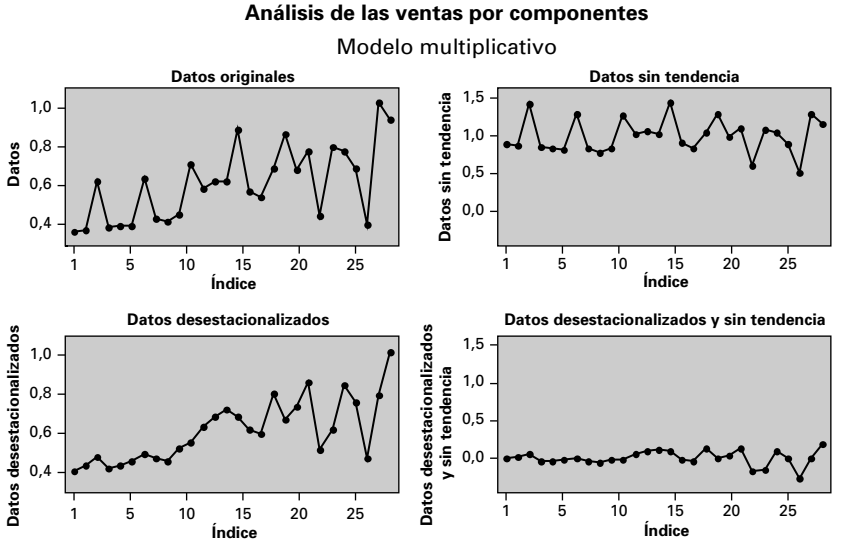
19.24. a)



El gráfico muestra un fuerte componente estacional con máximos en el tercer trimestre.

b)

Periodo	MM de 4 periodos	$100\,\frac{X_t}{X_t^*}$	Factor estacional	Serie desestacionalizada
1-1			90,930	0,3981
2			86,020	0,4301
3	0,438	141,902	130,400	0,4278
4	0,443	86,608	92,649	0,4522
2-1	0,448	86,830		0,4278
2	0,456	85,284		0,4522
3	0,465	137,493		0,4900
4	0,475	90,761		0,4652
3-1	0,491	83,643		0,4520
2	0,520	86,216		0,5208
3	0,565	126,046		0,5460
4	0,613	95,347		0,6303
4-1	0,656	94,458		0,6818
2	0,677	91,581		0,7208
3	0,665	133,935		0,6833
4	0,664	85,843		0,6152
5-1	0,670	80,582		0,5939
2	0,681	101,284		0,8021
3	0,725	120,000		0,6672
4	0,724	93,955		0,7340
6-1	0,684	114,077		0,8578
2	0,688	64,000		0,5115
3	0,689	116,153		0,6135
4	0,673	115,985		0,8419
7-1	0,696	99,102		0,7588
2	0,745	53,691		0,4650
3				0,7899
4				1,0146



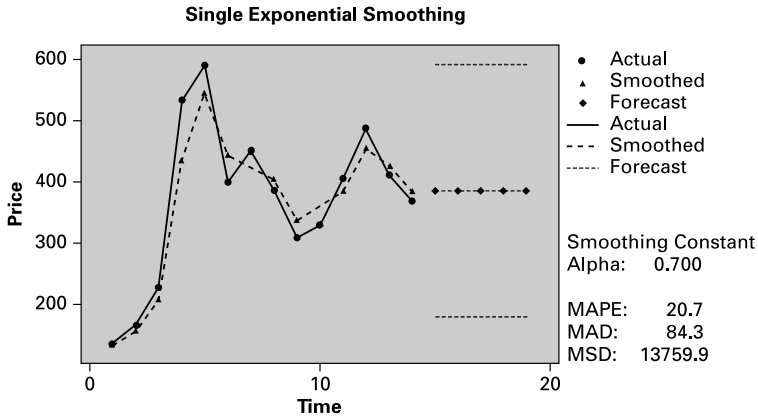
La serie desestacionalizada muestra una tendencia ascendente en los datos con una creciente variabilidad.

19.26.

Periodo	MM de 4 periodos	$100\,\frac{X_t}{X_t^*}$	Factor estacional	Serie desestacionalizada
1-1			93,701	574,165
2			97,648	634,935
3			119,127	729,476
4			101,932	832,911
5			102,038	928,084
6			97,450	955,365
7	767,750	97,167	99,356	750,834
8	773,750	95,638	100,300	737,789
9	775,000	84,387	87,504	747,395
10	770,333	113,457	113,255	771,707
11	759,667	99,912	98,563	770,069
12	743,000	85,734	89,127	714,721
2-1	730,250	87,094		678,752
2	725,750	91,767		682,043
3	721,458	118,233		716,045
4	712,542	105,678		738,730
5	699,542	112,502		771,280
6	689,458	100,224		709,084
7	684,000	99,415		684,406
8	678,917	102,811		695,915
9	668,208	88,745		677,684
10	651,750	110,625		636,614
11	630,917	95,100		608,751
12	611,417	90,609		621,586
3-1	598,167	98,300		627,526
2	583,625	101,435		606,260
3	570,375	117,467		562,426
4	563,542	96,000		530,748

Periodo	MM de 4 periodos	$100\,\frac{X_t}{X_t^*}$	Factor estacional	Serie desestacionalizada
5	558,250	89,387		489,033
6	551,917	92,586		524,373
7				545,512
8				485,545
9				555,403
10				586,286
11				537,730
12				529,583

19.28. Utilice una constante de suavización de 0,7 (alfa de 0,3) en Minitab. Fije el valor inicial de suavización en la media de las primeras observaciones.



19.30. a) Predicciones correspondientes a constantes de suavización de 0,2, 0,4, 0,6, 0,8:

Periodo	Xt	Alfa = 0,2	Alfa = 0,4	Alfa = 0,6	Alfa = 0,8
1	3,63	3,6300	3,6300	3,6300	3,6300
2	3,62	3,6220	3,6240	6,6260	3,6280
3	3,66	3,6524	3,6456	3,6396	3,6344
4	5,31	4,9785	4,6442	4,3078	3,9695
5	6,14	5,9077	5,5417	5,0407	4,4036
6	6,42	6,3175	6,0687	5,5924	4,8069
7	7,01	6,8715	6,6335	6,1594	5,2475
8	6,37	6,4703	6,4754	6,2437	5,4720
9	5,82	5,9501	6,0822	6,0742	5,5416
10	4,98	5,1740	5,4209	5,6365	5,4293
11	3,43	3,7788	4,2263	4,7539	5,0294
12	3,40	3,4758	3,7305	4,2123	4,7035
13	3,54	3,5272	3,6162	3,9434	4,4708
14	1,65	2,0254	2,4365	3,0260	3,9067
15	2,15	2,1251	2,2646	2,6756	3,5553
16	6,09	5,2970	4,5598	4,0414	4,0623
17	5,95	5,8194	5,3939	4,8048	4,4398
18	6,26	6,1719	5,9136	5,3869	4,8039
MAPE		20,795	24,546	30,613	36,6633
MAD		0,8578	1,0216	1,2362	1,4307
MSD		1,6403	1,9071	2,3378	2,8048

b) Dadas las medidas de precisión, elegir un alfa de 0,2 para la «mejor» predicción.

19.32. Si alfa es 1,0, entonces la predicción siempre será igual a la primera observación. $\hat{X}_{t+h} = X_1$

19.34. Utilice 0,7 para el nivel (alfa de 0,3) y 0,5 para la tendencia (beta de 0,5).

19.36.

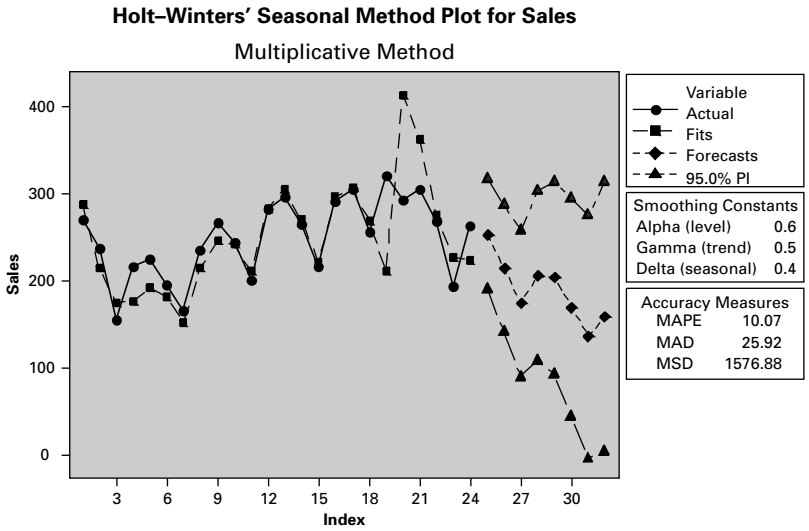
Winters' Method for FoodPrice

Forecasts

Period	Forecast	Lower	Upper
15	125.448	124.599	126.297
16	126.466	125.531	127.402
17	126.967	125.927	128.007

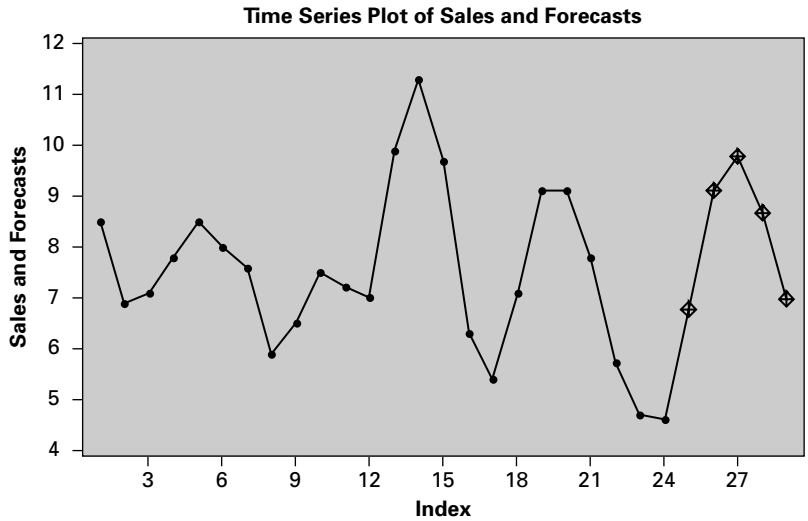
19.38. $\hat{X}_n = 260{,}6644$, $T_n = -8{,}6609$
Predicción para ocho trimestres:

Año	1	2	3	4
7	273,1269	230,6040	177,3303	232,1205
8	2535,5794	197,7740	151,1529	196,5420



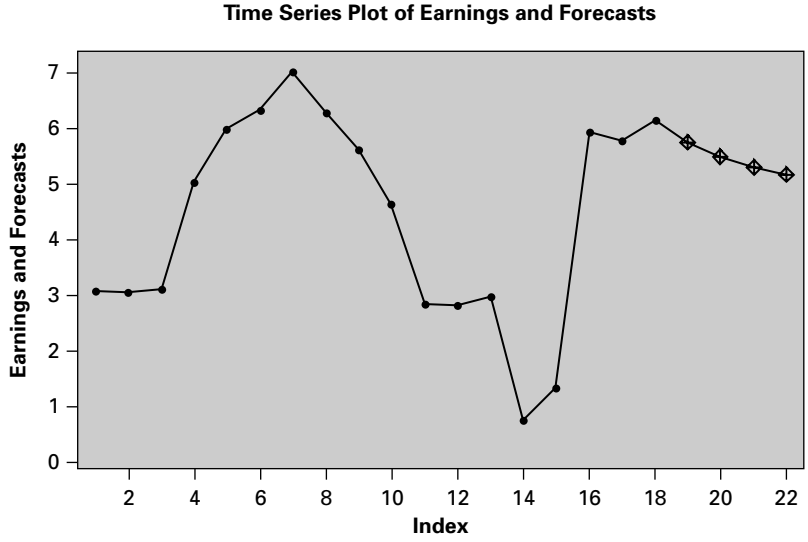
19.40. El modelo autorregresivo de primer orden es:
$\hat{y}_t = 87{,}85 + 0{,}169y_{t-1} + a_t$
$y_{17} = 87{,}85 + 0{,}169(92) = 103{,}398$
$y_{18} = 87{,}85 + 0{,}169(103{,}398) = 105{,}324$
$y_{19} = 87{,}85 + 0{,}169(105{,}324) = 105{,}650$
$y_{20} = 87{,}85 + 0{,}169(105{,}650) = 105{,}705$

19.42. Modelo de 4.º orden:
Estadístico z de $\phi_4 = -0{,}218$. No rechazar H_0 al nivel del 10%.
Modelo de 3.er orden:
Estadístico z de $\phi_3 = -0{,}909$. No rechazar H_0 al nivel del 10%.
Modelo de 2.º orden:
Estadístico z de $\phi_2 = -4{,}621$. Rechazar H_0 al nivel del 10%.
Modelo de 1.er orden:
Predicciones del modelo de segundo orden:
$\hat{y}_{25} = 6{,}776$, $\hat{y}_{26} = 9{,}103$, $\hat{y}_{27} = 9{,}792$, $\hat{y}_{28} = 8{,}670$, $\hat{y}_{29} = 6{,}968$



No habría ningún cambio si se utilizara un nivel de significación del 5% en lugar del 10%; el estadístico z del modelo de segundo orden de $-4{,}621$ es significativo a los niveles del 10 y el 5%.

19.44. *Estadístico z de* $\phi_3 = -0{,}303$. No rechazar H_0 al nivel del 10%.
Modelo de 2.º orden:
Estadístico z de $\phi_2 = -1{,}327$. No rechazar H_0 al nivel del 10%.
Modelo de 1.er orden:
Estadístico z de $\phi_1 = 3{,}664$. Rechazar H_0 al nivel del 10%.
Utilizar el modelo de 1.er orden para hacer predicciones.
$\hat{y}_{19} = 5{,}927$, $\hat{y}_{20} = 5{,}695$, $\hat{y}_{21} = 5{,}534$, $\hat{y}_{22} = 5{,}422$



No habría ningún cambio si se utilizara un nivel de significación del 5% en lugar del 10%; el estadístico z del modelo de primer orden de 3,664 es significativo a los niveles del 10 y el 5%.

19.46. $\hat{X}_{1996} = 202 + 1{,}1(951) - 0{,}48(923) + 0{,}17(867) = 952{,}45$
$\hat{X}_{1997} = 202 + 1{,}1(952{,}45) - 0{,}48(951) = 0{,}17(923) = 950{,}13$
$\hat{X}_{1998} = 202 + 1{,}1(950{,}13) - 0{,}48(952{,}45) + 0{,}17(951) = 951{,}64$

19.48. Modelo de 4.º orden:
Estadístico T de $\phi_4 = -1{,}185$. No rechazar H_0 al nivel del 10%.
Modelo de 3.er orden:
Estadístico T de $\phi_3 = -0{,}846$. No rechazar H_0 al nivel del 10%.
Modelo de 2.º orden:
Estadístico T de $\phi_2 = -1{,}490$. No rechazar H_0 al nivel del 10%.
Modelo de 1.er orden:
Estadístico T de $\phi_1 = -3{,}263$. Rechazar H_0 al nivel del 10%.
Utilizar el modelo de 1.er orden para hacer predicciones.
$\hat{y}_{25} = 0{,}070$, $\hat{y}_{26} = -0{,}001$, $\hat{y}_{27} = 0{,}041$

19.50. Apartados **a)**, **b)**, **c)**:

Periodo	**a)** No ponderado	**b)** Precios Laspeyres	**c)** Cantidades Laspeyres
1	100,00	100,00	100,00
2	110,30	109,72	103,78
3	117,43	115,55	100,04
4	127,52	123,47	105,29
5	143,96	137,18	106,00
6	158,22	149,41	105,85

19.52. Las predicciones se realizan analizando cada componente: tendencial, estacional y cíclico. Una vez analizado y medido cada componente, se incorpora la información al modelo de predicción.

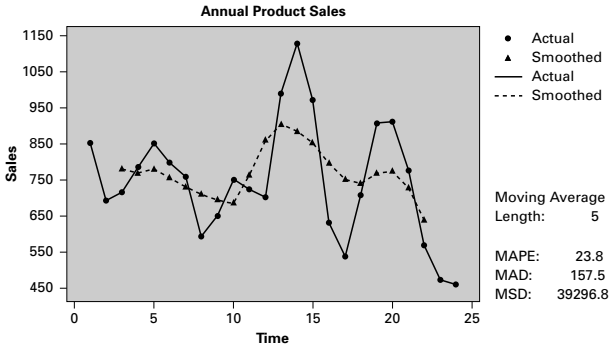
19.54. Una serie temporal desestacionalizada es una serie libre de los efectos de la influencia estacional. Los organismos oficiales realizan grandes esfuerzos para desestacionalizar los datos con el fin de tener una idea más clara de la pauta subyacente.

19.56. **a)** $R = 10$; no se puede rechazar H_0 a cualquier nivel habitual de alfa.
b)

Fuerte conducta cíclica, así como leve tendencia descendente

Años	Ventas	MV de 5 puntos
1	853	*
2	693	*
3	715	779,4
4	785	768,2
5	851	781,2
6	797	756,8
7	758	729,8

Años	Ventas	MV de 5 puntos
8	593	709,8
9	650	695,0
10	751	683,8
11	723	763,4
12	702	859,2
13	991	903,4
14	1129	885,0
15	972	852,2
16	631	795,6
17	538	751,2
18	708	739,2
19	907	768,4
20	912	774,6
21	777	727,6
22	569	638,0
23	473	*
24	459	*



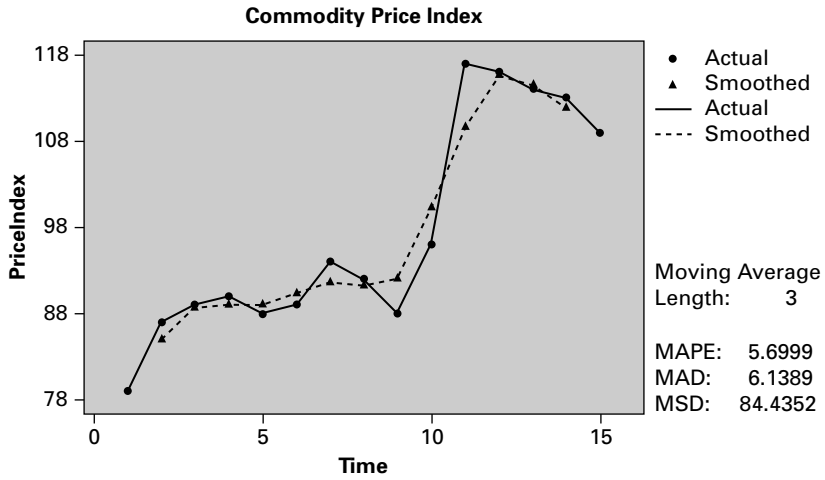
Fuerte tendencia ascendente y conducta cíclica

19.58. a)

Moving Average

```
Data          PriceIndex
Length        15.0000
NMissing      0

Moving Average
Length: 3

Accuracy Measures
MAPE:  5.6999
MAD:   6.1389
MSD:  84.4352
```

Row	Period	PriceIndex	AVER3	Predict	Error
1	1	79	*	*	*
2	2	87	85.000	*	*
3	3	89	88.667	*	*
4	4	90	89.000	85.000	5.0000
5	5	88	89.000	88.667	-0.6667
6	6	89	90.333	89.000	0.0000
7	7	94	91.667	89.000	5.0000
8	8	92	91.333	90.333	1.6667
9	9	88	92.000	91.667	-3.6667
10	10	96	100.333	91.333	4.6667
11	11	117	109.667	92.000	25.0000
12	12	116	115.667	100.333	15.6667
13	13	114	114.333	109.667	4.3333
14	14	113	112.000	115.667	-2.6667
15	15	109	*	114.333	-5.3333

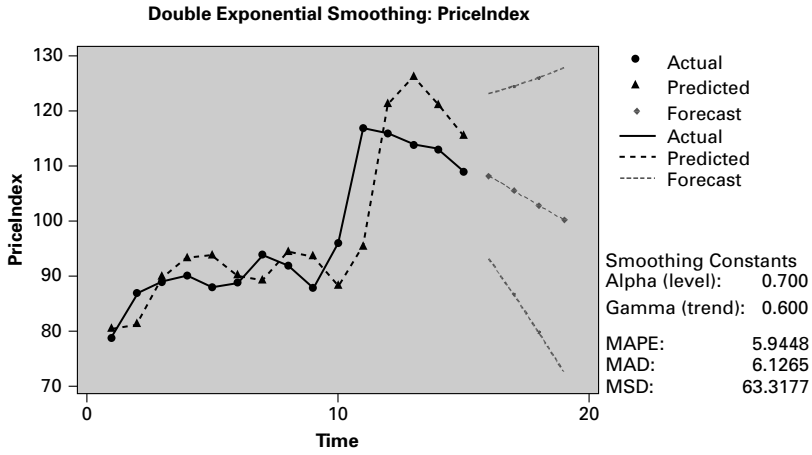


Fuerte tendencia ascendente y conducta cíclica.

19.60.

Double Exponential Smoothing

```
Data         PriceIndex
Length       15.0000
NMissing     0
Smoothing Constants
Alpha (level): 0.7
Gamma (trend): 0.6

Accuracy Measures
MAPE:  5.9448
MAD:   6.1265
MSD:  63.3177
 Row   Time     PriceIndex    Smooth   Predict     Error
   1      1             79    79.345    80.150   -1.1500
   2      2             87    85.298    81.326    5.6737
   3      3             89    89.199    89.662   -0.6621
   4      4             90    90.985    93.285   -3.2848
   5      5             88    89.708    93.692   -5.6920
   6      6             89    89.307    90.024   -1.0235
   7      7             94    92.558    89.193    4.8069
   8      8             92    92.739    94.463   -2.4629
   9      9             88    89.683    93.609   -5.6094
  10     10             96    93.659    88.197    7.8026
  11     11            117   110.535    95.451   21.5491
  12     12            116   117.613   121.378   -5.3776
  13     13            114   117.659   126.197  -12.1970
  14     14            113   115.436   121.120   -8.1201
  15     15            109   110.946   115.487   -6.4866
 Row  Period  Forecast     Lower     Upper
   1      16   108.272   93.2621   123.282
   2      17   105.598   86.6480   124.549
   3      18   102.925   79.6916   126.157
   4      19   100.251   72.5514   127.950
```



19.62. Modelo de 4.º orden:
Estadístico T de $\phi_4 = -0{,}216$. No rechazar H_0 al nivel del 10%.
Modelo de 3.er orden:
Estadístico T de $\phi_3 = 0{,}940$. No rechazar H_0 al nivel del 10%.
Modelo de 2.º orden:
Estadístico T de $\phi_2 = -4{,}590$. Rechazar H_0 al nivel del 10%.
Modelo de 1.er orden:
Estadístico T de $\phi_1 = 3{,}40$. Rechazar H_0 al nivel del 10%.
Utilizar el modelo de 2.º orden para hacer predicciones.
$\hat{y}_{25} = 672{,}829$, $\hat{y}_{26} = 905{,}554$, $\hat{y}_{27} = 979{,}039$

Capítulo 20

20.2. Las respuestas deben referirse a cada uno de los pasos esbozados en la Figura 20.1.

20.4. Las respuestas deben referirse a cada uno de los pasos esbozados en la Figura 20.1.

20.6. Las respuestas deben referirse a cuestiones como (a) la identificación de la población correcta, (b) el sesgo de selección (falta de respuesta), (c) el sesgo de respuesta

20.8. Las respuestas deben referirse a cuestiones como (a) la identificación de la población correcta, (b) el sesgo de selección (falta de respuesta), (c) el sesgo de respuesta

20.10. Dentro de Minitab, vaya a Calc → Make Patterned Data... para generar un conjunto simple de números de tamaño 'n' o 'N'. Introduzca como primer valor 1, como último valor n o N según proceda. Para el ejercicio 20.20, introduzca el último valor $n = 20$. A continuación, utilice Calc → Random Data... Sample from Columns... para generar una muestra aleatoria simple de tamaño 'n'.

20.12. Las mismas instrucciones que en el ejercicio 20.10, con la salvedad de que el último valor $n = 12.723$

20.14. (8,2262, 11,1738)

20.16. (5,4904, 9,0696)

20.18. $\hat{\sigma}_{\bar{x}}^2 = \dfrac{(s)^2}{n}\dfrac{N-n}{N} = \dfrac{s^2}{n}\left[1 - \dfrac{n}{N}\right] = s^2\left[\dfrac{1}{n} - \dfrac{1}{N}\right]$

20.20. $95.849{,}2706 < N\mu < 113.135{,}9294$

20.22. $403{,}2307 < N\mu < 577{,}3407$

20.24. De 0,4884 a 0,6316

20.26. $128{,}688 < Np < 196{,}812$, o sea, entre 129 y 197 tienen intención de hacer el examen final.

20.28. **a)** De 40,806 a 45,794
b) $\bar{x}_{st} = 37{,}3306$
c) Intervalo de confianza al 90%: de 36,0313 a 38,6299
Intervalo de confianza al 95%: de 35,7825 a 38,8787

20.30. **a)** De 2,8435 a 3,3965 **b)** De 3,1431 a 3,5969
c) De 3,0513 a 3,4166

20.32. **a)** $N\bar{x}_{xt} = 81.720$
b) Intervalo de confianza al 95%: $77.542{,}3153 < N\mu < 85.897{,}6847$

20.34. **a)** $\hat{p}_{st} = 0{,}3467$
b) Intervalo de confianza al 90%: de 0,2550 a 0,4383
Intervalo de confianza al 95%: de 0,2375 a 0,4559

20.36. **a)** 56 observaciones **b)** 68 observaciones

20.38. **a)** 55 observaciones **b)** 60 observaciones

20.40. **a)** 74 observaciones **b)** 88 observaciones

20.42. 58 observaciones

20.44. 211 observaciones

20.46. Afijación proporcional: tomar 498 observaciones.
Afijación óptima: tomar 471 observaciones.

20.48. **a)** $\bar{x}_c = 91{,}6761$ **b)** De 70,6920 a 112,6602

20.50. **a)** $\hat{p}_c = 0{,}4507$ **b)** De 0,38 a 0,5214

20.52. Las observaciones muestrales adicionales necesarias son $127 - 20 = 107$

20.54. Las observaciones muestrales adicionales necesarias son $160 - 30 = 130$

20.56. Tema de discusión: varias respuestas.

20.58. **a)** $\bar{x} = 74{,}7$, $s = 11{,}44$, $\hat{\sigma}_{\bar{x}}^2 = 11{,}633$
Intervalo de confianza al 90%: de 69,089 a 80,311
b) El intervalo sería más amplio; el valor de z aumentaría a 1,96.

20.60. **a)** De 0,559 a 0,687
b) Si la información muestral no se selecciona aleatoriamente, las conclusiones resultantes pueden estar sesgadas.

20.62. **a)** De 6,997 a 11,403 **b)** De 0,8247 hasta 13,3444

20.64. De 0,5147 a 0,7453

20.66. **a)** 16 observaciones **b)** 22 observaciones

20.68. 76 observaciones

20.70. Varias respuestas. Las respuestas deben incluir un análisis de las posibilidades de estratificar la población. Como los diferentes países utilizan diferentes sistemas y técnicas de votación, puede ser razonable una estratificación por circunscripciones. El método empleado por la circunscripción también podría utilizarse en la estratificación, por ejemplo, las papeletas en forma de mariposa o el sistema electrónico de votación.

Capítulo 21

21.2. D es dominada por C. Por lo tanto, D es inadmisible.

21.4. **a)** D es dominada por C. Por lo tanto, D es inadmisible y no vuelve a considerarse. El criterio maximin seleccionaría el proceso de producción C:

Acciones	Estados de la naturaleza			
Proceso de prod.	Demanda baja	Demanda moderada	Demanda alta	Rendimiento mín.
A	100.000	350.000	900.000	100.000
B	150.000	400.000	700.000	150.000
C	250.000	400.000	600.000	250.000

b) El criterio de la pérdida de oportunidades minimax seleccionaría el proceso de producción A:

Acciones	Tabla de pérdida de oportunidades			
Proceso de prod.	Demanda baja	Demanda moderada	Demanda alta	Rendimiento mín.
A	150.000	50.000	0	150.000
B	100.000	0	200.000	200.000
C	0	0	300.000	300.000

21.6.

Acciones	Estados de la naturaleza			
Proceso de prod.	Demanda baja	Demanda moderada	Demanda alta	Rendimiento mín.
A	70.000	120.000	200.000	70.000
B	80.000	120.000	180.000	80.000
C	100.000	125.000	160.000	100.000
D*	100.000	120.000	150.000	Inadmisible
E	60.000	115.000	220.000	60.000

*inadmisible

Por lo tanto, utilizando el criterio maximin se elegiría el proceso de producción C.

Acciones	Tabla de pérdida de oportunidades			
Proceso de prod.	Demanda baja	Demanda moderada	Demanda alta	Rendimiento mín.
A	30.000	5.000	20.000	30.000
B	20.000	5.000	40.000	40.000
C	0	0	60.000	60.000
D*				Inadmisible
E	40.000	10.000	0	40.000

*inadmisible

Por lo tanto, utilizando el criterio de la pérdida de oportunidades minimax se elegiría el proceso de producción A.

21.8.

Acción	S1	S2
A1	M_{11}	M_{12}
A2	M_{21}	M_{22}

En ese caso, se elegirá A1 tanto según el criterio maximin como según el criterio de la pérdida de oportunidades minimax si para $M_{11} > M_{21}$ y $M_{12} < M_{22}$ y $(M_{11} - M_{21}) > (M_{22} - M_{12})$

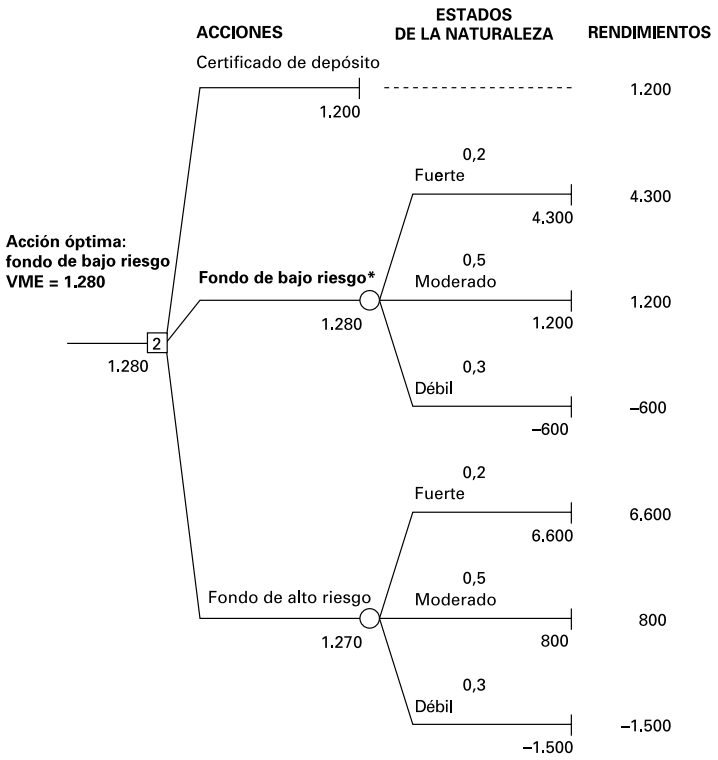
21.10. a)

Acciones	Se ofrece un empleo mejor	No se ofrece un empleo mejor
Acudir a una entrevista	4.500	–500
No acudir a una entrevista	0	0

b) VME (acudir a una entrevista) = -250
VME (no acudir a una entrevista) = 0
Por lo tanto, la acción óptima es no acudir a una entrevista.

21.12. a) VME (certificado de depósito) = 1.200
VME (fondo de acciones de bajo riesgo) = 1.280
VME (fondo de acciones de alto riesgo) = 1.270
Por lo tanto, la acción óptima es fondo de acciones de bajo riesgo.

b) Árbol de decisión:

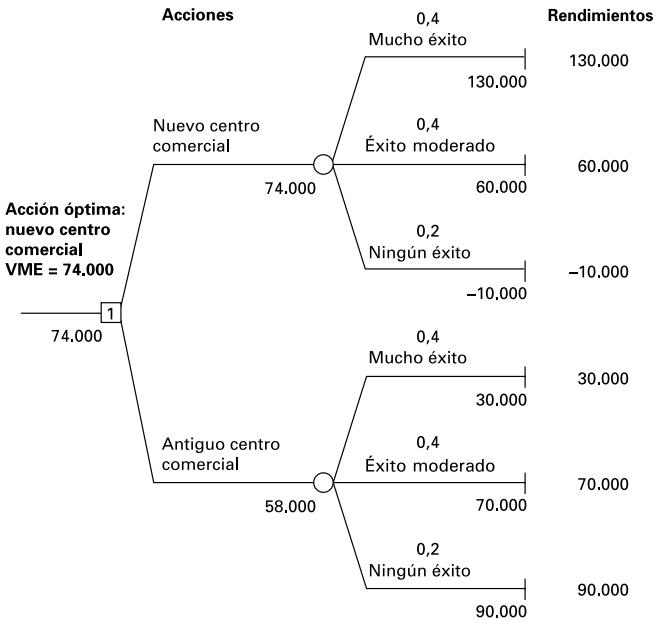


21.14. a) i) Falsa
ii) Verdadera
iii) Verdadera

b) No.

21.16. a) VME (nuevo) = 74.000
VME (antiguo) = 58.000
Por lo tanto, la acción óptima es el nuevo centro.

b) Árbol de decisión:

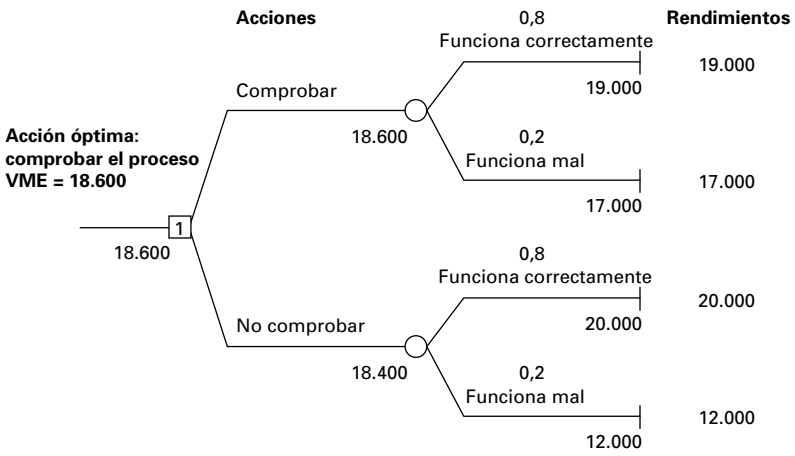


21.18. a) VME(A) = 660.000 − 550.000p
VME(B) = 535.000 − 300.000p
VME(C) = 495.000 − 200.000p
VME(D) = 460.000 − 150.000p
VME(A) > VME(B) cuando $p < 0{,}5$
VME(A) > VME(C) cuando $p < 0{,}471$
VME(A) > VME(D) cuando $p < 0{,}5$
Para $p < 0{,}471$, el criterio del VME elige la acción A, la misma que en el ejercicio 21.13. Obsérvese que D era «inadmisible».

b) VME(A) > VME(B) > VME(C) > VME(D) cuando a > 816.667

21.20. a) VME (comprobar) = 18.600
VME (no comprobar) = 18.400
Por lo tanto, la acción óptima es comprobar el proceso.

b) Árbol de decisión:



c) VME (comprobar) $= 19.000p + 17.000(1 - p) > 20.000p + 12.000(1 - p)$ cuando $p < 5/6$

21.22. a)

Pedidos extra	6	7	8	9	10
0	0	−10	−20	−30	−40
1	−20	20	10	0	−10
2	−40	0	40	30	20
3	−60	−20	20	60	50
4	−80	−40	0	40	80

b) Según el criterio del VME, la acción óptima es pedir 2 automóviles más:

Pedidos extra						
	6	7	8	9	10	VME
0	0(0,1)	−10(0,3)	−20(0,3)	−30(0,2)	−40(0,1)	−19
1	−20(0,1)	20(0,3)	10(0,3)	0(0,2)	−10(0,1)	6
2	−40(0,1)	0(0,3)	40(0,3)	30(0,2)	20(0,1)	16
3	−60(0,1)	−20(0,3)	20(0,3)	60(0,2)	50(0,1)	11
4	−80(0,1)	−40(0,3)	0(0,3)	40(0,2)	80(0,1)	−4

21.24. a) Se elige la acción A1 si $M_{11}p + M_{12}(1 - p) > M_{21}p + (1 - p)M_{22}$, o sea, $p(M_{11} - M_{21}) > > (1 - p)(M_{22} - M_{12})$

b) La acción A1 inadmisible implica que se elegirá A1 sólo si $p > 1$. En suma, para que el apartado a sea verdadero, ambos rendimientos de A1 no pueden ser menores que los correspondientes rendimientos de A2.

21.26. a) La acción óptima según el criterio del *VME* es la acción A.

b) $P(L \mid P) = 0{,}5$
$P(M \mid P) = 0{,}4$
$P(H \mid P) = 0{,}1$

c) $VME(A) = 280.000$
$VME(B) = 305.000$
$VME(C) = 345.000$
Por lo tanto, la acción óptima es la C.

d) $P(L \mid F) = 0{,}2903$
$P(M|F) = 0{,}5161$
$P(H|F) = 0{,}1935$

e) $VME(A) = 383.815$, $VME(B) = 385.435$, $VME(C) = 395.115$
Por lo tanto, la acción óptima es C.

f) $P(L \mid G) = 0{,}1538$
$P(M \mid G) = 0{,}3077$
$P(H \mid G) = 0{,}5385$

g) $VME(A) = 607.992$, $VME(B) = 523.077$, $VME(C) = 484.615$
Por lo tanto, la acción óptima es A.

21.28. a) $P(E \mid P) = 0{,}9231$, $P(\text{no } E \mid P) = 0{,}0769$

b) $VME(S) = 50.000$, $VME(R) = 114.615$. Por lo tanto, la acción óptima es conservar.

c) $P/E \mid N) = 0{,}25$, $P(\text{no E} \mid \text{N}) = 0{,}75$

d) $VME(S) = 50.000$
$VME(R) = 23.750$
Por lo tanto, la acción óptima es vender.

21.30. **a)** $P(2\,|\,10\%) = 0{,}01$, $P(1\,|\,10\%) = 0{,}18$, $P(0\,|\,10\%) = 0{,}81$
b) $P(2\,|\,30\%) = 0{,}09$, $P(1\,|\,30\%) = 0{,}42$, $P(0\,|\,30\%) = 0{,}49$
c) Probabilidad de los estados de un 10% de piezas defectuosas y un 30% de piezas defectuosas:

	N.º de piezas de fectuosas	10% de defectos	30% de defectos
i	2 piezas defectuosas	0,308	0,692
ii	1 piezas defectuosas	0,632	0,368
iii	0 piezas defectuosas	0,869	0,131

VME de las acciones	Comprobar	No comprobar
2 piezas defectuosas	17,616*	14,464
1 piezas defectuosas	18,264*	17,056
0 piezas defectuosas	18,737	18,952*

*acción óptima dada la circunstancia

21.32. **a)** Información perfecta es el caso en el que la persona que debe tomar una decisión es capaz de obtener información para saber con seguridad qué estado ocurrirá.
b) La acción óptima: fondo de acciones de bajo riesgo (véase el problema 21.12)
$VEIP = 0{,}2(6.600 - 4.300) + 0{,}5(0) + 0{,}3(1.200 - (-600)) = 1.000$

21.34. Dado que la acción óptima es nuevo centro
$VEIP = 24.000$

21.36. El valor esperado de la información muestral es

$$\sum_{i=1}^{M} P(A_i)V_i, \quad \text{donde} \quad P(A_i) = \sum_{j=1}^{H} P(A_i\,/s_j)$$

Para información perfecta, $P(A_i\,|\,s_j) = 0$ para $i \neq j$ y $P(A_i\,/s_j) = 1$ para $i = j$; por lo tanto, $P(A_i) = P(s_i)$

21.38. $VEIP = 23.003$

21.40. Dado que la acción óptima es conservar la patente (véase el problema 21.28).
$VEIP = 13.650$

21.42. **a)** $VEIP = 34{,}1$ **b)** $VEIP = 55{,}87$ **c)** La diferencia = 21,77
d) Ninguno **e)** 24,75

21.44. **a)**

Rendimiento	−10.000	30.000	60.000	70.000	90.000	13.000
Utilidad	0	35	60	70	85	100

b) UE(Nuevo) = 64
UE(Antiguo) = 59
Por lo tanto, la acción esperada es Nuevo centro

21.46. $94.000p - 16.000(1 - p) = 0 \rightarrow p = 16/110$

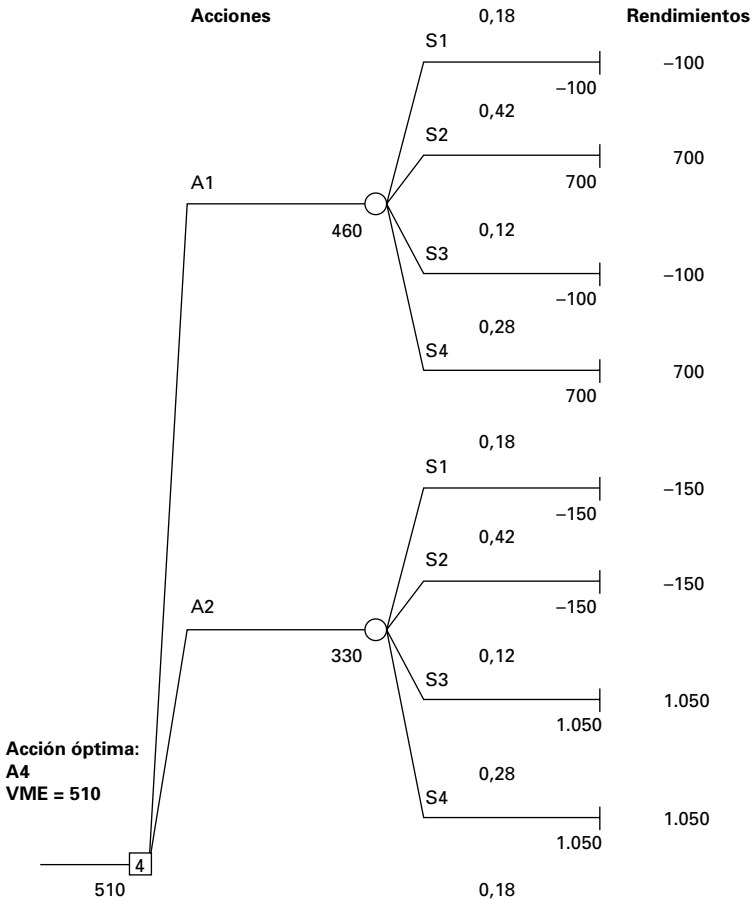
Rendimiento	−160.000	0	94.000
Utilidad	0	160/10	100

Pendiente $(-16.000{,}0) = 0{,}00009$
Pendiente $(0{,}94.000) = 0{,}00105$
Por lo tanto, el contratista tiene una preferencia por el riesgo.

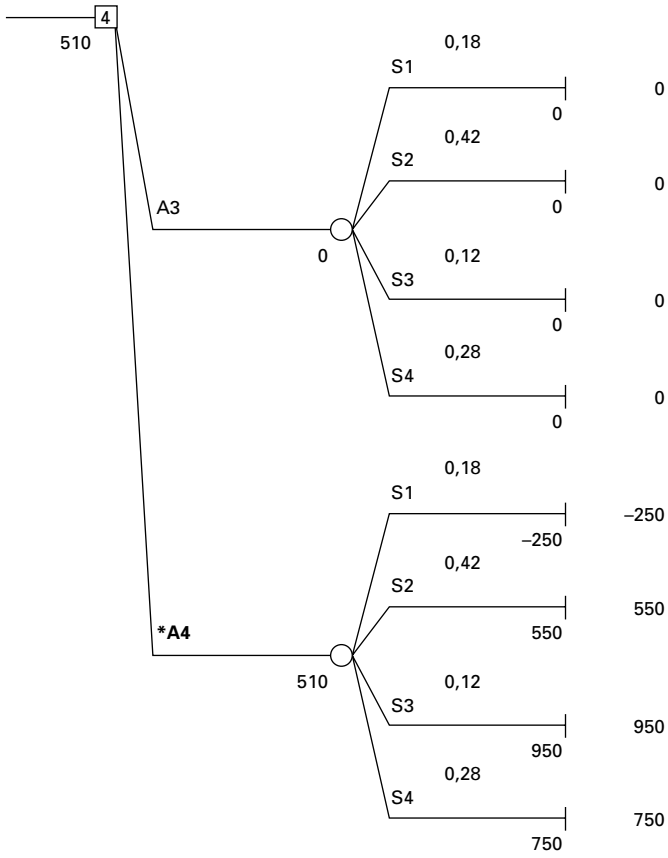
21.48. **a)** $P(S1) = 0{,}3(0{,}6) = 0{,}18$, $P(S2) = 0{,}42$, $P(S3) = 0{,}12$, $P(S4) = 0{,}28$

b) $VME(A1) = 460$, $VME(A2) = 330$, $VME(A3) = 0$, $VME(A4) = 510$
Por lo tanto, la acción óptima es A4.

c) Trace el árbol de decisión:



TreePlan (continuación del problema 21.48):



d) $VEIP = 204$
e) 79

ÍNDICE ANALÍTICO

A

B

C

D

E

F

G

H

I

J

L

M

N

O

P

R

S

T

U

V

W