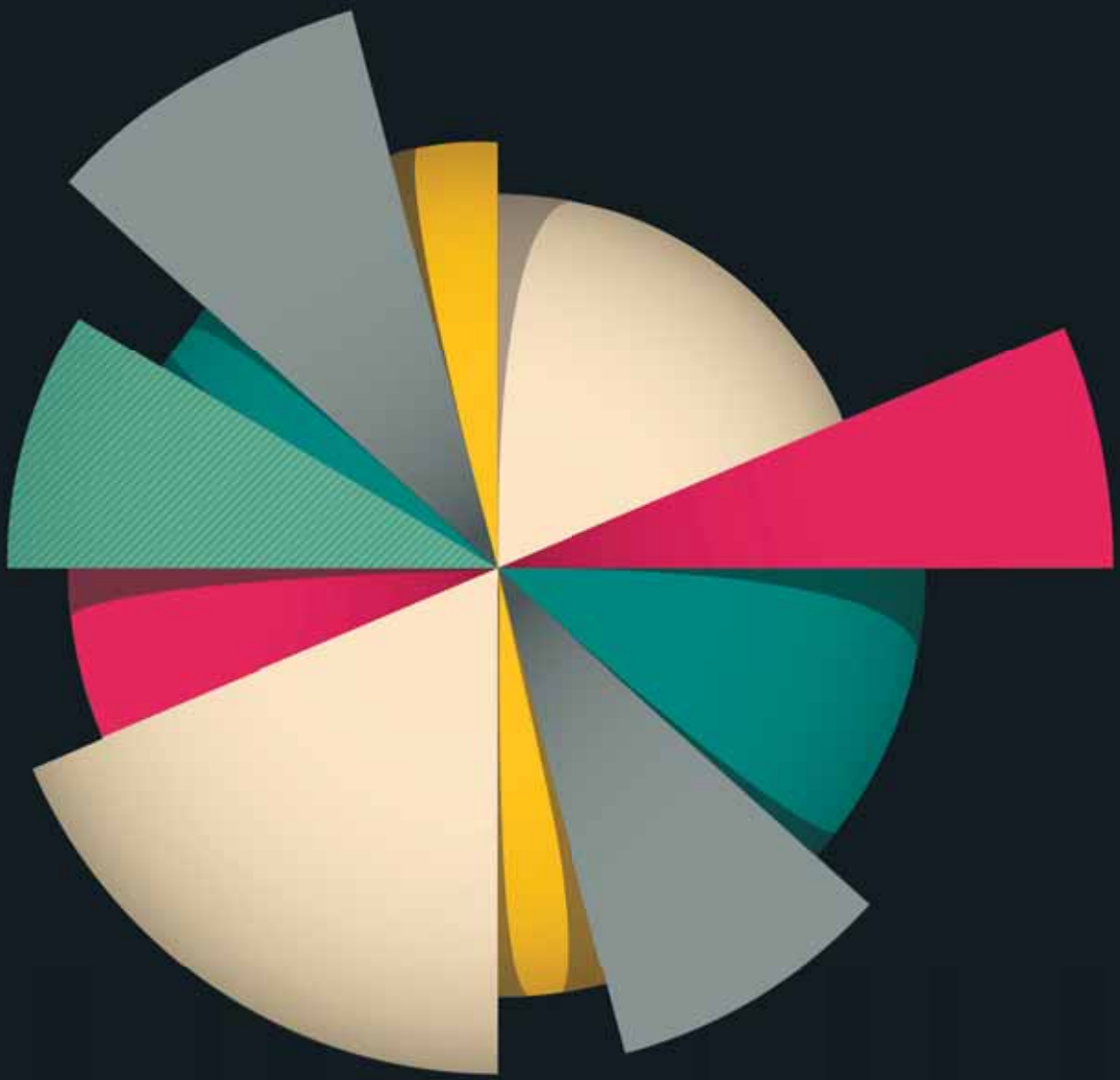
CENGAGE



FUNDAMENTOS DE
PROBABILIDAD Y ESTADÍSTICA

JAY L. DEVORE

FUNDAMENTOS DE PROBABILIDAD Y ESTADÍSTICA

Primera edición

Jay L. Devore
California Polytechnic State University, San Luis Obispo

Joel Ibarra Escutia
Instituto Tecnológico de Toluca

Traducción

Javier León Cárdenas | Jesús Miguel Torres Flores

Revisión técnica

Ana Elizabeth García Hernández
Instituto Politécnico Nacional

Instituto Tecnológico de Celaya
María Josefina Hernández Patiño
María Teresa Villalón Guzmán

Instituto Tecnológico de Durango
María de la Luz Torres Valles

Instituto Tecnológico de Hermosillo
Hilario Mayboca Araujo
Carlos Alberto Pereyda Pierre

Instituto Tecnológico de León
Luz María Trinidad Pérez Alvarado

Instituto Tecnológico de Pachuca
Luis García González

Instituto Tecnológico de Querétaro
Jöns Sánchez Aguilar
Everardo Santiago Rendón

Instituto Tecnológico de San Juan del Río
Saulo Servín Guzmán

Instituto Tecnológico de Toluca
Yolanda Alvarado Pérez
Nelson García García
Mónica Citlali Huerta Zepeda
Carmen Margarita Montiel Leyva

Tecnológico de Estudios Superiores de Jilotepec
Rodolfo Guadalupe Alcántara Rosales

Tecnológico de Estudios Superiores de Jocotitlán
Christopher Gutiérrez Luna
Jorge Ubaldo Jacobo Sánchez
Isaías Vázquez Juárez

Universidad Politécnica de Querétaro
Alejandro Flores Rangel
Alondra Ortiz Verdín

CENGAGE

Australia • Brasil • Corea • España • Estados Unidos • Japón • México • Reino Unido • Singapur

CENGAGE

Fundamentos de probabilidad y estadística.
Primera edición
Jay L. Devore

Director Higher Education Latinoamérica:
Renzo Casapía Valencia

Gerente editorial Latinoamérica:
Jesús Mares Chacón

Editora de desarrollo:
Abril Vega Orozco

Coordinador de manufactura:
Rafael Pérez González

Diseño de portada:
Karla Paola Benítez García

Imagen de portada:
© Vectorwork/Shutterstock

Composición tipográfica:
Heriberto Gachuz Chávez

Esta es una adaptación del libro:
Devore, Jay L. *Probabilidad y estadística para ingeniería y ciencias*. Novena edición. ©2016
ISBN: 978-607-522-828-0
Traducido del libro *Probability and Statistics for Engineering and the Sciences*, Ninth Edition.
Jay L. Devore. Publicado en inglés por Cengage Learning ©2016.
ISBN: 978-1-305-25180-9

Datos para catalogación bibliográfica:
Devore, Jay L. *Fundamentos de probabilidad y estadística*. Primera edición.
ISBN: 978-607-526-663-3

Visite nuestro sitio web en:
http://latinoamerica.cengage.com

Impreso en México
1 2 3 4 5 6 7 20 21 19 18

Para mis queridos nietos
Philip y Elliot quienes son
estadísticamente significativos.

Contenido

Capítulo 8 Análisis de la varianza

Capítulo 9 Regresión lineal simple y correlación

Apéndice de tablas*

* Este material se encuentra disponible en línea. Acceda a www.cengage.com e ingrese con el ISBN de la obra.

Prefacio

Propósito

El uso de modelos de probabilidad y métodos estadísticos para analizar datos se ha convertido en una práctica común en virtualmente todas las disciplinas científicas. Este libro pretende introducir con amplitud aquellos modelos y métodos que con mayor probabilidad encuentran y utilizan los estudiantes en sus carreras de ingeniería y las ciencias naturales. Aun cuando los ejemplos y ejercicios se diseñaron pensando en los científicos y los ingenieros, la mayoría de los métodos tratados son básicos en los análisis estadísticos de muchas otras disciplinas, por lo que los estudiantes de las ciencias administrativas y sociales también se beneficiarán con la lectura de este libro.

Enfoque

Los estudiantes de un curso de estadística diseñado para servir a otras especialidades de estudio al principio es posible que duden del valor y la relevancia del material, pero mi experiencia es que los estudiantes pueden conectarse con la estadística mediante buenos ejemplos y ejercicios que combinen sus experiencias diarias con sus intereses científicos. Así pues, he trabajado duro para encontrar ejemplos reales y no artificiales, que alguien pensó que valía la pena recopilar y analizar. Muchos de los métodos presentados, sobre todo en los últimos capítulos sobre inferencia estadística, se ilustran analizando datos tomados de una fuente publicada y muchos de los ejercicios también implican trabajar con dichos datos. En ocasiones es posible que el lector no esté familiarizado con el contexto de un problema particular (como muchas veces yo lo estuve), pero me di cuenta de que los problemas reales con un contexto un tanto extraño atraen más a los estudiantes que aquellos problemas definitivamente artificiales en un entorno conocido.

Nivel matemático

La exposición es relativamente modesta en función del desarrollo matemático. El uso sustancial del cálculo se hace sólo en el capítulo 4 y en partes de los capítulos 5 y 6. En particular, con excepción de una observación o nota ocasional, el cálculo aparece en la parte de inferencia del libro sólo en la segunda sección del capítulo 6. No se utiliza álgebra matricial en absoluto. Por tanto, casi toda la exposición deberá de ser accesible para aquellos cuyo conocimiento matemático incluye un semestre o dos trimestres de cálculo diferencial e integral.

Ayuda para el aprendizaje de los estudiantes

Aunque el nivel matemático del libro representará poca dificultad para la mayoría de los estudiantes, es posible que el trabajo dirigido hacia la comprensión de los conceptos y la apreciación del desarrollo lógico de la metodología en ocasiones requiera un esfuerzo sustancial. Para ayudar a que los estudiantes ganen en comprensión y apreciación he proporcionado numerosos ejercicios de dificultad variable, desde muchos que implican la aplicación rutinaria del material incluido en el texto hasta algunos que le piden al lector que extienda los conceptos analizados en el texto a situaciones un tanto nuevas. Existen muchos ejercicios más que la mayoría de los profesores desearía asignar durante cualquier curso particular, pero recomiendo que se les pida a los estudiantes que resuelvan un número sustancial de los mismos; en una disciplina de solución de problemas, el compromiso activo de esta clase es la forma más segura de identificar y cerrar las brechas en el entendimiento que inevitablemente surgen.

Para acceder al material adicional del libro, por favor visite www.cengage.com e ingrese con el ISBN de la obra.

Reconocimientos

Mis colegas en Cal Poly me proporcionaron apoyo y retroalimentación invaluables durante el curso de los años. También agradezco a los muchos usuarios de ediciones previas que me sugirieron mejoras (y en ocasiones errores identificados). Una nota especial de agradecimiento va para Matt Carlton por su trabajo en los dos manuales de soluciones, uno para profesores y el otro para estudiantes.

La generosa retroalimentación provista por los siguientes revisores de esta edición y de ediciones previas, ha sido de mucha ayuda para mejorar el libro: Robert L. Armacost, University of Central Florida; Bill Bade, Lincoln Land Community College; Douglas M. Bates, University of Wisconsin–Madison; Michael Berry, West Virginia Wesleyan College; Brian Bowman, Auburn University; Linda Boyle, University of Iowa; Ralph Bravaco, Stonehill College; Linfield C. Brown, Tufts University; Karen M. Bursic, University of Pittsburgh; Lynne Butler, Haverford College; Troy Butler, Colorado State University; Barrett Caldwell, Purdue University; Kyle Caudle, South Dakota School of Mines & Technology; Raj S. Chhikara, University of Houston–Clear Lake; Edwin Chong, Colorado State University; David Clark, California State Polytechnic University at Pomona; Ken Constantine, Taylor University; Bradford Crain, Portland State University; David M. Cresap, University of Portland; Savas Dayanik, Princeton University; Don E. Deal, University of Houston; Annjanette M. Dodd, Humboldt State University; Jimmy Doi, California Polytechnic State University–San Luis Obispo; Charles E. Donaghey, University of Houston; Patrick J. Driscoll, U.S. Military Academy; Mark Duva, University of Virginia; Nassir Eltinay, Lincoln Land Community College; Thomas English, College of the Mainland; Nasser S. Fard, Northeastern University; Ronald Fricker, Naval Postgraduate School; Steven T. Garren, James Madison University; Mark Gebert, University of Kentucky; Harland Glaz, University of Maryland; Ken Grace, Anoka-Ramsey Community College; Celso Grebogi, University of Maryland; Veronica Webster Griffis, Michigan Technological University; José Guardiola, Texas A&M University–Corpus Christi; K. L. D. Gunawardena, University of Wisconsin–Oshkosh; James J. Halavin, Rochester Institute of Technology; James Hartman, Marymount University; Tyler Haynes, Saginaw Valley State University; Jennifer Hoeting, Colorado State University; Wei-Min Huang, Lehigh University; Aridaman Jain, New Jersey Institute of Technology; Roger W. Johnson, South Dakota School of Mines & Technology; Chihwa Kao, Syracuse University; Saleem A. Kassam, University of Pennsylvania; Mohammad T. Khasawneh, State University of New York–Binghamton; Kyungduk Ko, Boise State University; Stephen Kokoska, Colgate University; Hillel J. Kumin, University of Oklahoma; Sarah Lam, Binghamton University; M. Louise Lawson, Kennesaw State University; Jialiang Li, University of Wisconsin–Madison; Wooi K. Lim, William Paterson University; Aquila Lipscomb, The Citadel; Manuel Lladser, University of Colorado at Boulder; Graham Lord, University of California–Los Ángeles; Joseph L. Macaluso, DeSales University; Ranjan Maitra, Iowa State University; David Mathiason, Rochester Institute of Technology; Arnold R. Miller, University of Denver; John J. Millson, University of Maryland; Pamela Kay Miltenberger, West Virginia Wesleyan College; Monica Molsee, Portland State University; Thomas Moore, Naval Postgraduate School; Robert M. Norton, College of Charleston; Steven Pilnick, Naval Postgraduate School; Robi Polikar, Rowan University; Justin Post, North Carolina State University; Ernest Pyle, Houston Baptist University; Xianggui Qu, Oakland University; Kingsley Reeves, University of South Florida; Steve Rein, California Polytechnic State University–San Luis Obispo; Tony Richardson, University of Evansville; Don Ridgeway, North Carolina State University; Larry J. Ringer, Texas A&M University; Nabin Sapkota, University of Central Florida; Robert M. Schumacher, Cedarville University; Ron Schwartz, Florida Atlantic University; Kevan Shafizadeh, California State University– Sacramento; Mohammed Shayib, Prairie View A&M; Alice E. Smith, Auburn University; James MacGregor Smith, University of Massachusetts; Paul J. Smith, University of Maryland; Richard M. Soland, The George Washington University; Clifford Spiegelman, Texas A&M University; Jery Stedinger, Cornell University; David Steinberg, Tel Aviv University; William Thistleton, State University of New York Institute of Technology; J A Stephen Viggiano, Rochester Institute of Technology; G. Geoffrey Vining, University of Florida; Bhutan Wadhwa, Cleveland State University; Gary Wasserman, Wayne State University; Elaine Wenderholm, State University of New York–Oswego; Samuel P. Wilcock, Messiah College; Michael G. Zabetakis, University of Pittsburgh; y Maria Zack, Point Loma Nazarene University.

Preeti Longia Sinha de MPS Limited ha realizado un trabajo excelente al supervisar la producción del libro. Una vez más me veo obligado a expresar mi gratitud a todas aquellas personas en Cengage que han hecho contribuciones importantes a lo largo de mi carrera como escritor de libros de texto. Para esta edición más reciente, un agradecimiento especial a Jay Campbell (por su información oportuna y retroalimentación a través del proyecto), Molly Taylor, Ryan Ahern, Spencer Arritt Cathy Brooks y Andrew Coppola. También apreciamos la labor estelar de todos los representantes de ventas de Cengage Learning que han trabajado para hacer que mis libros sean más visibles para la comunidad estadística. Por último, pero no por ello menor, un sincero agradecimiento a mi esposa Carol por sus décadas de apoyo, y a mis hijas por proporcionar inspiración a través de sus propios logros.

Jay L. Devore

Agradecimientos

Queremos agradecer a todos los profesores que participaron en esta obra, sus aportaciones y sugerencias fueron invaluables para el desarrollo de la misma.

Cengage Latinoamérica

Capítulo

1

Generalidades y estadística descriptiva

INTRODUCCIÓN

Los conceptos y métodos estadísticos no son sólo útiles sino que con frecuencia son indispensables para entender el mundo que nos rodea. Proporcionan formas de obtener ideas nuevas acerca del comportamiento de muchos fenómenos que usted encontrará en el campo de especialización que haya escogido en ingeniería o ciencias.

La estadística como disciplina nos enseña a realizar juicios inteligentes y tomar decisiones informadas en la presencia de incertidumbre y variación. Sin estas habría poca necesidad de métodos estadísticos o de profesionales en estadística. Si los componentes de un tipo particular tuvieran exactamente la misma duración, si todos los resistores producidos por un fabricante tuvieran el mismo valor de resistencia, si las determinaciones del pH en las muestras de suelo de un lugar en particular dieran resultados idénticos, etcétera, entonces una sola observación revelaría toda la información deseada.

La estadística ofrece no sólo métodos para analizar resultados de experimentos una vez que se han realizado sino también sugerencias sobre cómo pueden llevarse a cabo los experimentos de una manera eficiente para mitigar los efectos de la variación y tener una mejor oportunidad de llegar a conclusiones correctas.

1.1 Poblaciones, muestras y procesos

Los ingenieros y científicos constantemente están expuestos a la recolección de hechos o **datos**, tanto en sus actividades profesionales como en sus actividades diarias. La estadística proporciona métodos de organizar y resumir datos, y de obtener conclusiones basadas en la información contenida en los mismos.

Usualmente una investigación se centrará en una colección bien definida de objetos que constituyen una **población** de interés. En un estudio la población podría consistir en todas las cápsulas de gelatina de un tipo particular producidas durante un periodo específico. Otra investigación podría implicar la población compuesta de todos los individuos que obtuvieron una licenciatura de ingeniería durante el último ciclo académico. Cuando la información deseada está disponible para todos los objetos de la población, se tiene lo que se conoce como **censo**. Las restricciones de tiempo, dinero y otros recursos escasos casi siempre hacen que un censo sea impráctico o poco factible. En su lugar, se selecciona un subconjunto de la población —**una muestra**—, de alguna manera recomendada. Así pues, se podría obtener una muestra de cojinetes de una corrida de producción particular como base para investigar si se ajustan a las especificaciones de fabricación; o se podría seleccionar una muestra de los graduados de ingeniería del último año para obtener retroalimentación sobre la calidad de los programas de estudio de ingeniería.

Por lo general existe interés sólo en ciertas características de los objetos de una población: el número de grietas en la superficie de cada recubrimiento, el espesor de cada pared de la cápsula, el género de un graduado de ingeniería, la edad a la cual el individuo se graduó, etcétera. Una característica puede ser categórica, tal como el género o el tipo de funcionamiento defectuoso, o puede ser de naturaleza numérica. En el primer caso el *valor* de la característica es una categoría (p. ej., femenino o soldadura insuficiente), mientras que en el segundo caso, el valor es un número (p. ej., edad = 23 años, o diámetro = 502 cm). Una **variable** es cualquier característica cuyo valor puede cambiar de un objeto a otro en la población. Las últimas letras de nuestro alfabeto, en minúscula, denotarán las variables. Por ejemplo:

x = marca de la calculadora de un estudiante

y = número de visitas a un sitio web particular durante un periodo específico

z = la distancia de frenado de un automóvil en condiciones específicas

Los datos se obtienen al observar una sola variable o dos o más variables simultáneamente. Un conjunto de datos **univariantes** se compone de observaciones realizadas en una sola variable. Por ejemplo, se podría determinar el tipo de transmisión automática (A) o manual (M) en cada uno de diez automóviles recientemente adquiridos con cierto concesionario y el resultado sería el siguiente conjunto de datos categóricos

M A A A M A A M A A

La siguiente muestra del ritmo cardiaco (latidos por minuto) para pacientes de recién ingreso en una unidad de cuidados intensivos para adultos es un conjunto de datos numéricos univariantes:

88 80 71 103 154 132 67 110 60 105

Se tienen datos **bivariantes** cuando se realizan observaciones en cada una de dos variables. El conjunto de datos podría consistir en un par (altura, peso) por cada integrante del equipo de basquetbol, con la primera observación como (72, 168), la segunda como (75, 212), etcétera. Si un ingeniero determina el valor de x = componente de duración y y = razón de la falla del componente, el conjunto de datos resultante es bivariante, con

una variable numérica y otra categórica. Los datos **multivariantes** surgen cuando se realizan observaciones en más de una variable (por tanto, bivariante es un caso especial de multivariante). Por ejemplo, un médico investigador podría determinar la presión sanguínea sistólica, la presión sanguínea diastólica y el nivel de colesterol en suero de cada paciente participante en un estudio. Cada observación sería una terna de números, tal como (120, 80, 146). En muchos conjuntos de datos multivariantes algunas variables son numéricas y otras son categóricas. Por tanto, el número anual dedicado al automóvil de *Consumer Reports* da valores de dichas variables como tipo de vehículo (pequeño, deportivo, compacto, mediano, grande), eficiencia de consumo de combustible en la ciudad y en carretera en millas por galón (mpg), tipo de transmisión (ruedas traseras, ruedas delanteras, cuatro ruedas), etcétera.

Ramas de la estadística

Posiblemente un investigador que ha recopilado datos desee resumir y describir características importantes de los mismos. Esto implica utilizar métodos de **estadística descriptiva**. Algunos de estos son de naturaleza gráfica; la construcción de histogramas, diagramas de caja y gráficas de puntos son ejemplos primordiales. Otros métodos descriptivos implican calcular medidas numéricas, tales como medias, desviaciones estándar y coeficientes de correlación. La amplia disponibilidad de paquetes de software de computación para estadística ha vuelto estas tareas más fáciles de realizar que antes. Las computadoras son mucho más eficientes que los seres humanos para calcular y crear imágenes (una vez que han recibido las instrucciones apropiadas por parte del usuario). Esto significa que el investigador no tendrá que dedicarse al "trabajo tedioso" y tendrá más tiempo para estudiar los datos y extraer información importante. A lo largo de este libro se presentarán los datos de salida de varios paquetes tales como Minitab, SAS, JMP y R. El programa R puede ser descargado sin costo del sitio **http://www.r-project.org**. Este programa ha ganado popularidad entre la comunidad estadística, y existen muchos libros que describen sus diferentes usos (lo cual implica programar, contrario a los menús desplegables de Minitab y JMP).

EJEMPLO 1.1 La caridad es un gran negocio en los Estados Unidos. El sitio web **charitynavigator.com** proporciona información de aproximadamente 6000 organizaciones de caridad y otro gran número de pequeñas organizaciones de beneficiencia. Algunas organizaciones caritativas operan eficientemente, con gastos administrativos y de recaudación de fondos que apenas son un pequeño porcentaje de los gastos totales, mientras que otras gastan un alto porcentaje de lo que obtienen en tal actividad. Enseguida se muestran datos sobre los gastos en la recaudación de fondos como un porcentaje de los gastos totales para una muestra aleatoria de 60 asociaciones de caridad:

6.1	12.6	34.7	1.6	18.8	2.2	3.0	2.2	5.6	3.8
2.2	3.1	1.3	1.1	14.1	4.0	21.0	6.1	1.3	20.4
7.5	3.9	10.1	8.1	19.5	5.2	12.0	15.8	10.4	5.2
6.4	10.8	83.1	3.6	6.2	6.3	16.3	12.7	1.3	0.8
8.8	5.1	3.7	26.3	6.0	48.0	8.2	11.7	7.2	3.9
15.3	16.6	8.8	12.0	4.7	14.7	6.4	17.0	2.5	16.2

Sin organización es difícil tener una idea de las características más importantes de los datos, que podrían significar un valor típico (o representativo): si los valores están muy concentrados en torno a un valor típico o dispersos, si existen brechas en los datos, qué porcentajes de los valores son menores a 20%, etcétera. La figura 1.1 muestra una *gráfica de tallo y hojas* de los datos y un *histograma*. En la sección 1.2 se discutirá la construcción e interpretación de estos resúmenes gráficos; por el momento se espera que se vea cómo los porcentajes están distribuidos sobre el rango de valores de 0 a 100. Es claro que la mayoría

```
Stem-and-leaf of FundRsng N = 60
Leaf Unit = 1.0
     0 | 0111112222333333344
     0 | 55556666666778888
     1 | 0001222244
     1 | 55666789
     2 | 01
     2 | 6
     3 | 4
     3 |
     4 |
     4 | 8
     5 |
     5 |
     6 |
     6 |
     7 |
     7 |
     8 | 3
```

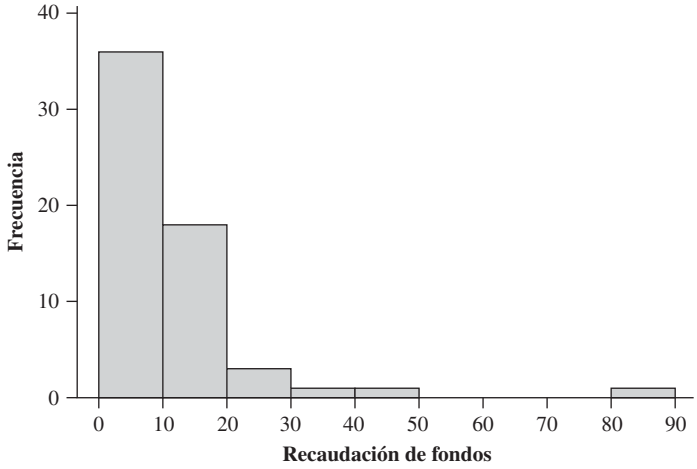


Figura 1.1 Gráfica de tallos y hojas (truncada a diez dígitos) de Minitab e histograma para los datos del porcentaje de recaudación de fondos para caridad

de las organizaciones de caridad en el ejemplo gastan menos de 20% en recaudar fondos y sólo unos pequeños porcentajes podrían ser vistos más allá del límite de una práctica sensible. ■

Después de haber obtenido la muestra de una población, comúnmente un investigador desearía utilizar la información muestral para sacar una conclusión (hacer una inferencia de alguna clase) respecto a la población. Es decir, la muestra es un medio para llegar a un fin en lugar de un fin en sí misma. Las técnicas para generalizar desde una muestra hasta una población se conjuntan en la rama de la **estadística inferencial**.

EJEMPLO 1.2 Las investigaciones sobre la de resistencia de los materiales constituye una rica área de aplicación de métodos estadísticos. El artículo **"Effects of Aggregates and Microfillers on the Flexural Properties of Concrete" (*Magazine of Concrete Research*, 1997: 81-98)** reporta sobre un estudio de propiedades de resistencia de concreto de alto desempeño mediante el uso de superplastificantes y ciertos aglomerantes. La resistencia a la compresión de dicho concreto había sido investigada previamente, pero no se sabía mucho sobre la resistencia a la flexión (una medida de la capacidad de resistir fallas por flexión). Los datos anexos sobre resistencia a la flexión (en megapascales, MPa, donde 1 Pa (Pascal) = 1.45×10^{-4} lb/pulg^2) aparecen en el artículo citado:

5.9 7.2 7.3 6.3 8.1 6.8 7.0 7.6 6.8 6.5 7.0 6.3 7.9 9.0
8.2 8.7 7.8 9.7 7.4 7.7 9.7 7.8 7.7 11.6 11.3 11.8 10.7

Suponga que se desea *estimar* el valor promedio de resistencia a la flexión de todas las vigas que pudieran ser fabricadas de esta manera (si se conceptualiza una población de todas esas vigas, se trata de estimar la media poblacional). Se puede demostrar que con un alto grado de confianza la resistencia media de la población se encuentra entre 7.48 MPa y 8.80 MPa; esto se llama *intervalo de confianza* o *estimación de intervalo*. Alternativamente se podrían utilizar estos datos para predecir la resistencia a la flexión de una *sola* viga de este tipo. Con un alto grado de confianza, la resistencia de una sola viga excederá de 7.35 MPa; el número 7.35 se conoce como *límite de predicción inferior*. ■

El objetivo principal de este libro es presentar e ilustrar métodos de estadística inferencial útiles en el trabajo científico. Los tipos más importantes de procedimientos inferenciales, estimación puntual, comprobación de hipótesis y estimación mediante intervalos de confianza se introducen en los capítulos 5 al 7 y luego se utilizan escenarios más complicados

en los capítulos 8 y 9. El resto de este capítulo presenta métodos de estadística descriptiva que se utilizan mucho en el desarrollo de la inferencia.

Los capítulos 2 al 4 presentan material de la disciplina de probabilidad. Este material finalmente tiende un puente entre las técnicas descriptivas e inferenciales. El dominio de la probabilidad permite entender mejor cómo se desarrollan y utilizan los procedimientos inferenciales, cómo las conclusiones estadísticas pueden ser traducidas e interpretadas en un lenguaje cotidiano, y cuándo y dónde pueden ocurrir errores al aplicar los métodos. La probabilidad y la estadística se ocupan de cuestiones que implican poblaciones y muestras, pero lo hacen de "manera inversa", una respecto a la otra.

En un problema de probabilidad se supone que las propiedades de la población estudiada son conocidas (p. ej., en una población numérica se puede suponer una cierta distribución especificada de valores de la población) y se pueden plantear y responder preguntas respecto a una muestra tomada de una población. En un problema de estadística el experimentador dispone de las características de una muestra y esta información le permite sacar conclusiones respecto a la población. La relación entre las dos disciplinas se resume diciendo que la probabilidad discurre de la población a la muestra (razonamiento deductivo), mientras que la estadística inferencial lo hace de la muestra a la población (razonamiento inductivo). Lo anterior se ilustra en la figura 1.2.

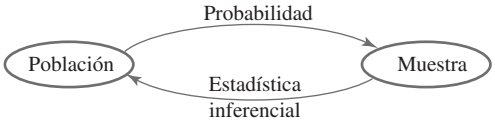


Figura 1.2 Relación entre probabilidad y estadística inferencial

Antes de querer entender lo que una muestra particular dice sobre la población, primero se debe entender la incertidumbre asociada con la toma de una muestra de una población dada. Por esto es que se estudia la probabilidad antes que la estadística.

EJEMPLO 1.3 Para ejemplificar el enfoque contrastante de la probabilidad y la estadística inferencial considere el uso que hacen los automovilistas del cinturón de seguridad manual de regazo en autos que están equipados con sistemas automáticos de cinturones de hombro. (El artículo **"Automobile Seat Belts: Usage Patterns in Automatic Belt Systems", *Human Factors*, 1998: 126-135**, resume datos sobre su uso.) Se podría suponer que probablemente 50% de todos los conductores de automóviles equipados de esta manera, en cierta área metropolitana utilizan regularmente el cinturón de regazo (una suposición sobre la población), por lo que uno puede preguntarse "¿qué tan probable es que una muestra de 100 conductores incluya al menos 70 que utilicen regularmente el cinturón de regazo?", o "¿cuántos de los conductores en una muestra de 100 se puede esperar que utilicen con regularidad el cinturón de regazo?". Por otra parte, en estadística inferencial se dispone de información sobre la muestra; por ejemplo, una muestra de 100 conductores de tales vehículos reveló que 65 utilizan con regularidad su cinturón de regazo. Podemos entonces preguntarnos: "¿Proporciona esto evidencia sustancial para concluir que más de 50% de todos los conductores en dicha área metropolitana utilizan con regularidad el cinturón de regazo?". En el último escenario se intenta utilizar la información sobre la muestra para responder una pregunta respecto a la estructura de toda la población de la cual se seleccionó la muestra. ■

En el ejemplo del cinturón de regazo la población es concreta y está bien definida: todos los conductores de automóviles equipados de una cierta manera en un área metropolitana en particular. En el ejemplo 1.2, sin embargo, las mediciones de resistencia provienen de una muestra de vigas prototipo que no tuvieron que seleccionarse de una población existente. En su lugar conviene pensar en una población compuesta de todas las posibles

mediciones de resistencia que podrían hacerse en condiciones experimentales similares. Tal población se conoce como **población conceptual** o **hipotética.** Existen varias situaciones en las cuales las preguntas encajan en el marco de referencia de la estadística inferencial al conceptualizar una población.

Recopilación de datos

La estadística se ocupa no sólo de la organización y el análisis de datos una vez que han sido recopilados, sino también del desarrollo de las técnicas de recopilación de datos. Si éstos no son apropiadamente reunidos, el investigador será incapaz de responder las preguntas que se tengan consideradas con un razonable grado de confianza. Un problema común es que la población objetivo, aquella sobre la cual se van a sacar conclusiones, puede ser diferente de la población realmente muestreada. Por ejemplo, a los publicistas les gustaría contar con varias clases de información sobre los hábitos de sus clientes potenciales para ver televisión. La información más sistemática de esta clase se obtuvo tras colocar dispositivos de monitoreo en un pequeño número de casas a través de los Estados Unidos. Se ha conjeturado que la colocación de semejantes dispositivos por sí misma modifica el comportamiento del televidente, de modo que las características de la muestra pueden ser diferentes de aquellas de la población objetivo.

Cuando la recopilación de datos implica seleccionar individuos u objetos de un marco, el método más simple para garantizar una selección representativa es tomar una *muestra aleatoria simple*. Esta es una para la cual cualquier subconjunto particular del tamaño especificado (p. ej., una muestra de tamaño 100) tiene la misma oportunidad de ser seleccionada. Por ejemplo, si el marco se compone de 1 000 000 de números en serie, los números 1, 2, …, hasta 1 000 000 podrían ser anotados en hojitas de papel idénticas. Después de reunir los papelitos en una caja y revolverlos perfectamente se sacan uno por uno hasta obtener el tamaño de muestra requerido. De manera alternativa (y preferible), se podría utilizar una tabla de números aleatorios o algún software generador de números aleatorios.

En ocasiones se pueden utilizar otros métodos de muestreo para facilitar el proceso de selección, a fin de obtener información extra o para incrementar el grado de confianza en las conclusiones. Un método como el *muestreo estratificado* implica separar las unidades de la población en grupos que no se traslapen y tomar una muestra de cada uno. Por ejemplo, un fabricante de reproductores de DVD desea información sobre la satisfacción del cliente respecto a las unidades producidas durante el año previo. Si se fabricaran y se vendieran tres modelos diferentes, se seleccionaría una muestra distinta de cada uno de los estratos correspondientes. Esto daría información sobre los tres modelos y garantizaría que ningún modelo estuviera sobrerrepresentado o subrrepresentado en toda la muestra.

Con frecuencia se obtiene una muestra de "conveniencia" seleccionando individuos u objetos sin aleatorización sistemática. Por ejemplo, un conjunto de ladrillos puede ser apilado de tal modo que sea extremadamente difícil seleccionar aquellos que se encuentran en el centro. Si los ladrillos colocados en la parte superior y a los lados de la pila fueran de algún modo diferentes de los demás, los datos muestrales resultantes no representarían la población. A menudo un investigador supondrá que tal muestra de conveniencia representa en forma aproximada una muestra aleatoria, en cuyo caso se utiliza el repertorio de métodos inferenciales de un estadístico; sin embargo, esta es una cuestión de criterio. La mayoría de los métodos aquí analizados se basa en una variación del muestreo aleatorio simple.

Los ingenieros y científicos a menudo reúnen datos realizando alguna clase de experimento. Esto implica decidir cómo asignar varios tratamientos diferentes (tales como fertilizantes o recubrimientos anticorrosivos) a las varias unidades experimentales (parcelas o tramos de tubería). Por otra parte, un investigador puede variar sistemáticamente los niveles o categorías de ciertos factores (p. ej., presión o tipo de material aislante) y observar el efecto en alguna variable de respuesta (tal como rendimiento de un proceso de producción).

EJEMPLO 1.4 Un artículo en el *New York Times* (**27 de enero de 1987**) reporta que el riesgo de sufrir un ataque cardiaco puede disminuirse tomando aspirinas. Esta conclusión se basa en un experimento diseñado que incluía un grupo de control de individuos que consumieron un placebo con apariencia de aspirina pero del que se sabía que era inerte, y un grupo de tratamiento que consumió aspirina de acuerdo con un régimen específico. Los sujetos fueron asignados a cada grupo al azar para protegerlos contra cualquier prejuicio, de modo que se pudieran utilizar métodos basados en la probabilidad para analizar los datos. De los 11 034 individuos del grupo de control, 189 experimentaron subsecuentemente ataques cardiacos, mientras que sólo 104 de los 11 037 en el grupo de aspirina sufrieron un ataque cardiaco. La tasa de incidencia de ataques cardiacos en el grupo de tratamiento fue de casi sólo la mitad de aquella en el grupo de control. Una posible explicación de este resultado es la variación de la probabilidad de que la aspirina en realidad no tiene el efecto deseado y la diferencia observada es sólo una variación típica, del mismo modo que lanzar dos monedas idénticas por lo general producirá diferente número de veces que caiga cara. No obstante, en este caso, los métodos inferenciales sugieren que la variación de la probabilidad por sí misma no puede explicar en forma adecuada la magnitud de la diferencia observada. ■

EJEMPLO 1.5 Un ingeniero desea investigar los efectos tanto del tipo de adhesivo como del material conductor en la fuerza adhesiva cuando se monta un circuito integrado (CI) sobre cierto sustrato. Se consideraron dos tipos de adhesivo y dos materiales conductores. Se realizaron dos observaciones por cada combinación de tipo de adhesivo/material conductor y se obtuvieron los datos siguientes.

Tipo de adhesivo	Material conductor	Fuerza adhesiva observada	Promedio
1	1	82, 77	79.5
1	2	75, 87	81.0
2	1	84, 80	82.0
2	2	78, 90	84.0

En la figura 1.3 se ilustran las fuerzas adhesivas promedio resultantes. El adhesivo tipo 2 mejora la fuerza adhesiva en comparación con el tipo 1 en aproximadamente la misma cantidad siempre que se utiliza uno de los materiales conductores, con la combinación 2, 2 como la mejor. De nuevo se pueden utilizar métodos inferenciales para juzgar si estos efectos son reales o si simplemente se deben a la variación de la probabilidad.

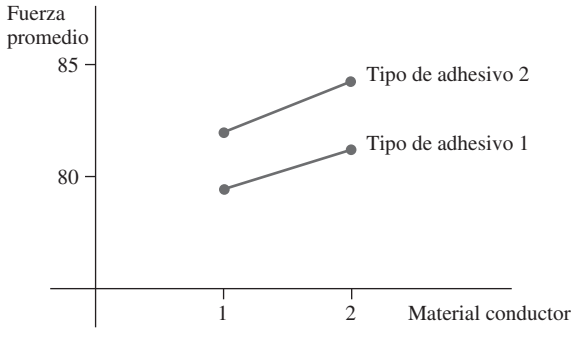


Figura 1.3 Fuerzas adhesivas promedio en el ejemplo 1.5

Suponga además que se consideran dos tiempos de secado y también dos tipos de posrecubrimientos de los circuitos integrados. Existen entonces $2 \cdot 2 \cdot 2 \cdot 2 = 16$ combinaciones

de estos cuatro factores y es posible que el ingeniero no disponga de suficientes recursos para hacer incluso una observación sencilla para cada una de estas combinaciones. ■

EJERCICIOS Sección 1.1 (1–9)

1. Dé una posible muestra de tamaño 4 de cada una de las siguientes poblaciones.
- **a.** Todos los periódicos publicados en los Estados Unidos.
- **b.** Todas las compañías listadas en la Bolsa de Valores de Nueva York.
- **c.** Todos los estudiantes en su colegio o universidad.
- **d.** Todas las calificaciones promedio de los estudiantes en su colegio o universidad.

2. Para cada una de las siguientes poblaciones hipotéticas, dé una muestra posible de tamaño 4:
- **a.** Todas las distancias que podrían resultar cuando usted lanza un balón de futbol americano.
- **b.** Las longitudes de las páginas de los libros publicados de aquí a 5 años.
- **c.** Todas las posibles mediciones de intensidad de los terremotos (escala de Richter) que pudieran registrarse en California durante el siguiente año.
- **d.** Todos los posibles rendimientos (en gramos) de una cierta reacción química realizada en un laboratorio.

3. Considere la población compuesta por todas las computadoras de una cierta marca y modelo y enfóquese en si una de ellas necesita servicio mientras se encuentra dentro del periodo de garantía.
- **a.** Plantee varias preguntas de probabilidad con base en la selección de 100 de estas computadoras.
- **b.** ¿Qué pregunta de estadística inferencial podría ser respondida determinando el número de dichas computadoras en una muestra de tamaño 100 que requieren servicio de garantía?

4.
- **a.** Dé tres ejemplos diferentes de poblaciones concretas y tres ejemplos distintos de poblaciones hipotéticas.
- **b.** Por cada una de sus poblaciones concretas e hipotéticas dé un ejemplo de una pregunta de probabilidad y un ejemplo de pregunta de estadística inferencial.

5. Muchas universidades y colegios han instituido programas de instrucción suplementaria (IS) en los cuales un facilitador regularmente se reúne con un pequeño grupo de estudiantes inscritos en el curso para promover discusiones sobre el material incluido en el curso y mejorar el dominio de la materia. Suponga que los estudiantes inscritos en un largo curso de estadística (¿de qué más?) se dividen al azar en un grupo de control que no participará en la instrucción suplementaria y en un grupo de tratamiento que sí participará. Al final del curso se determina la calificación total de cada estudiante en el curso.
- **a.** ¿Son las calificaciones del grupo IS muestra de una población existente? De ser así, ¿de cuál se trata? De no ser así, ¿cuál es la población conceptual pertinente?
- **b.** ¿Cuál piensa que es la ventaja de dividir al azar a los estudiantes en los dos grupos en lugar de permitir que cada estudiante elija el grupo al que desea unirse?
- **c.** ¿Por qué los investigadores no pusieron a todos los estudiantes en el grupo de tratamiento? [*Nota*: El artículo **"Supplemental Instruction: An Effective Component of Student Affairs Programming" (*J. of College Student Devel*., 1997: 577-586)** aborda el análisis de datos de varios programas de instrucción suplementaria.]

6. El sistema de la Universidad Estatal de California (CSU, por sus siglas en inglés) consta de 23 campus universitarios, desde la Estatal de San Diego en el sur hasta la Estatal Humboldt cerca de la frontera con Oregon. Un administrador de la CSU desea hacer una inferencia sobre la distancia promedio entre la ciudad natal de los estudiantes y sus campus universitarios. Describa y discuta varios diferentes métodos de muestreo que pudieran ser empleados. ¿Sería éste un estudio enumerativo o un estudio analítico? Explique su razonamiento.

7. Cierta ciudad se divide naturalmente en diez distritos. ¿Cómo podría un valuador de bienes raíces seleccionar una muestra de casas unifamiliares que pudiera ser utilizada como base para desarrollar una ecuación y así predecir el valor estimado a partir de características tales como antigüedad, tamaño, número de baños, distancia a la escuela más cercana, etcétera? ¿El estudio es enumerativo o analítico?

8. La cantidad de flujo a través de una válvula solenoide en el sistema de control de emisiones de un automóvil es una característica importante. Se realizó un experimento para estudiar cómo la velocidad de flujo depende de tres factores: la longitud de la armadura, la fuerza del resorte y la profundidad de la bobina. Se eligieron dos niveles diferentes (alto y bajo) de cada factor y se realizó una sola observación del flujo por cada combinación de niveles.
- **a.** ¿Cuántas observaciones conformaron el conjunto de datos resultante?
- **b.** ¿Este estudio es enumerativo o analítico? Explique su razonamiento.

9. En un famoso experimento, realizado en 1882, Michelson y Newcomb obtuvieron 66 observaciones del tiempo que requería la luz para viajar entre dos lugares en Washington, D.C. Algunas de las mediciones (codificadas en cierta manera) fueron, 31, 23, 32, 36, –2, 26, 27 y 31.
- **a.** ¿Por qué no son idénticas estas mediciones?
- **b.** ¿Es este un estudio enumerativo? ¿Por qué sí o por qué no?

1.2 Métodos pictóricos y tabulares en estadística descriptiva

La estadística descriptiva se divide en dos temas generales. En esta sección se considera la representación de un conjunto de datos mediante técnicas visuales. En las secciones 1.3 y 1.4 se desarrollarán algunas medidas numéricas para conjuntos de datos. Es posible que usted ya conozca muchas técnicas visuales; tablas de frecuencia, hojas de contabilidad, histogramas, gráficas de pastel, gráficas de barras, diagramas de puntos y similares. Aquí se seleccionan algunas de estas técnicas que son más útiles y pertinentes para la probabilidad y la estadística inferencial.

Notación

Alguna notación general facilitará la aplicación de métodos y fórmulas a una amplia variedad de problemas prácticos. El número de observaciones en una muestra única, es decir, el *tamaño de la muestra*, a menudo será denotado por n, de modo que $n = 4$ para la muestra de universidades {Stanford, Iowa State, Wyoming, Rochester} y también para la muestra de lecturas de pH {6.3, 6.2, 5.9, 6.5}. Si se consideran dos muestras al mismo tiempo, m y n o n_1 y n_2 se pueden utilizar para denotar los números de observaciones. En un experimento para comparar la eficiencia térmica de dos tipos diferentes de motores diésel se obtienen las siguientes muestras {29.7, 31.6, 30.9} y {28.7, 29.5, 29.4, 30.3} en este caso $m = 3$ y $n = 4$.

Dado un conjunto de datos compuesto de n observaciones de alguna variable x, las observaciones individuales serán denotadas por x_1, x_2, x_3,…, x_n. El subíndice no guarda ninguna relación con la magnitud de una observación particular. Por tanto x_1 en general no será la observación más pequeña del conjunto, ni x_n será la más grande. En muchas aplicaciones x_1 será la primera observación realizada por el experimentador, x_2 la segunda y así sucesivamente. La observación i-ésima del conjunto de datos será denotada por x_i.

Gráficas de tallos y hojas

Considere un conjunto de datos numéricos x_1, x_2,…, x_n para el cual cada x_i se compone de al menos dos dígitos. Una forma rápida de obtener la representación visual informativa del conjunto de datos es construir una *gráfica de tallos y hojas*.

Pasos para construir una gráfica de tallos y hojas

1. Seleccione uno o más de los primeros dígitos para los valores de tallo. Los segundos dígitos se convierten en hojas.
2. Enumere los posibles valores de tallos en una columna vertical.
3. Anote la hoja para cada observación junto al correspondiente valor de tallo.
4. Indique las unidades para tallos y hojas en algún lugar de la gráfica.

Para un conjunto de datos que se compone de calificaciones de exámenes, cada uno entre 0 y 100, la calificación de 83 tendría un tallo de 8 y una hoja de 3. Si todas las calificaciones del examen están en 90, 80 y 70 (¡el sueño del profesor!), usar los diez dígitos como el tallo daría una gráfica de sólo tres filas. En este caso es deseable estirar la gráfica

repitiendo dos veces el valor de cada tallo, 9H, 9L, 8H, . . . , 7L, una vez para las hojas altas 9, . . . , 5 y otra vez para las hojas bajas 4, . . . , 0. Después de una calificación de 93 tendría un tallo de 9L y una hoja de 3. En general, se recomienda una gráfica basada en tallos entre 5 y 20.

EJEMPLO 1.6 Una queja común entre los estudiantes universitarios es que no duermen lo suficiente. El artículo **"Class Start Times, Sleep, and Academic Performance in College: A Path Analysis" (*Chronobiology Intl.*, 2012: 318-335)** investigó los factores que afectan el tiempo de sueño. La gráfica de tallo y hojas de la figura 1.4 muestra el número promedio de horas de sueño diario durante un periodo de dos semanas para una muestra de 253 estudiantes.

5L	00
5H	6889
6L	000111123444444
6H	55556778899999
7L	000011111112222223333333344444444
7H	55555555666666666667777778888888889999999999999999
8L	000000000000111111222222222222222233333333333444444444444444
8H	555555556666666667777778888888889999999999
9L	00001111111222223334
9H	666678999
10L	00
10H	56

Tallo: dígitos uno
Hoja: dígitos diez

Figura 1.4 Gráfica de tallo y hojas para un tiempo promedio de sueño por día

La primera observación en la fila superior de la gráfica es 5.0, correspondiente a un tallo de 5 y una hoja de 0, y la última observación en la parte inferior de la pantalla es 10.6. Observe que en ausencia de un contexto, sin la identificación del tallo y los dígitos de la hoja en la gráfica, no sabríamos si la observación del tallo 7 y hoja 9 es 0.79, 7.9 o 79. Las hojas en cada fila se ordenan de menor a mayor; esto se realiza comúnmente mediante paquetes de software pero no es necesario si se ha creado a mano una gráfica.

La gráfica sugiere que un valor típico o representativo del tiempo de sueño se encuentra en la fila 8L del tallo, quizás sea 8.1 u 8.2. Las observaciones no aparecen muy centradas en torno a este valor típico, como sería el caso si todos los estudiantes durmieran entre 7.5 y 9.5 horas en promedio. La gráfica parece subir suavemente a una cresta en la fila 8L y luego declinar suavemente (conjeturamos que la cresta menor en la fila 6L desaparecería si se tuvieran más datos disponibles). La forma general de la gráfica es bastante simétrica, teniendo gran parecido a una curva en forma de campana; no se extiende más en una dirección que en otra. Los dos valores más pequeños y los dos más grandes parecen estar un poco separados del resto de los datos, tal vez ligeramente, pero ciertamente no son extremos "atípicos". Una referencia en el artículo citado sugiere que los individuos en este grupo de edad necesitan alrededor de 8.4 horas de sueño por día. Al parecer, un porcentaje sustancial de los estudiantes en la muestra no duerme lo suficiente. ■

Una gráfica de tallos y hojas aporta información sobre los siguientes aspectos de los datos:

- identificación de un valor típico o representativo
- grado de dispersión en torno al valor típico
- presencia de brechas en los datos

- grado de simetría en la distribución de los valores
- número y localización de crestas
- presencia de cualquier *valor atípico* de la gráfica

EJEMPLO 1.7 La figura 1.5 presenta gráficas de tallos y hojas de una muestra aleatoria de longitudes de campos de golf (yardas) designados por ***Golf Magazine*** como los de mayor desafío en los Estados Unidos. Entre la muestra de 40 campos, el más corto es de 6 433 yardas de largo y el más largo es de 7 280. Las longitudes parecen estar distribuidas de una manera más o menos uniforme dentro del rango de valores presentes en la muestra. Obsérvese que la selección de tallo, en este caso de un solo dígito (6 o 7) o de tres (643,…, 728), produciría una gráfica no informativa, primero porque son pocos tallos y segundo porque son demasiados.

64	35 64 33 70	Tallo: Dígitos de millares y centenas
65	26 27 06 83	Hojas: Dígitos de decenas y unidades
66	05 94 14	
67	90 70 00 98 70 45 13	
68	90 70 73 50	
69	00 27 36 04	
70	51 05 11 40 50 22	
71	31 69 68 05 13 65	
72	80 09	

(a)

```
Stem-and-leaf of yardage  N = 40
Leaf Unit = 10
          4          64  3367
          8          65  0228
         11          66  019
         18          67  0147799
        (4)          68  5779
         18          69  0023
         14          70  012455
          8          71  013666
          2          72  08
```

(b)

Figura 1.5 Gráficas de tallos y hojas de la longitud de los campos de golf: (a) hojas de dos dígitos; (b) gráfica Minitab de hojas con truncamiento a un dígito

Los programas computacionales de estadística en general no producen gráficas con tallos de dígitos múltiples. La gráfica Minitab que aparece en la figura 1.5(b) es resultado de *truncar* cada observación al borrar los dígitos uno. ■

Gráficas de puntos

Una gráfica de puntos es un atractivo resumen de datos numéricos cuando el conjunto de datos es razonablemente pequeño o cuando existen pocos valores de distintos datos. Cada observación está representada por un punto sobre la ubicación correspondiente en una escala de medición horizontal. Cuando un valor ocurre más de una vez, existe un punto por cada ocurrencia y estos puntos se apilan verticalmente. Como con la gráfica de tallos y hojas, una gráfica de puntos aporta información sobre localización, dispersión, extremos y brechas.

EJEMPLO 1.8 Existe una creciente preocupación en los Estados Unidos debido a que no se gradúan suficientes estudiantes de la universidad. Los Estados Unidos solían ser el número 1 en el mundo en porcentaje de adultos con títulos universitarios, pero recientemente ha descendido al lugar 16. Aquí se presentan datos acerca del porcentaje de personas de entre 25 y 34 años de edad en cada estado que tenían algún tipo de grado de educación superior, a partir de 2010 (se enumeran en orden alfabético, se incluye el Distrito de Columbia):

31.5	32.9	33.0	28.6	37.9	43.3	45.9	37.2	68.8	36.2	35.5
40.5	37.2	45.3	36.1	45.5	42.3	33.3	30.3	37.2	45.5	54.3
37.2	49.8	32.1	39.3	40.3	44.2	28.4	46.0	47.2	28.7	49.6
37.6	50.8	38.0	30.8	37.6	43.9	42.5	35.2	42.2	32.8	32.2
38.5	44.5	44.6	40.9	29.5	41.3	35.4				

La figura 1.6 muestra una gráfica de puntos para los datos. Los puntos correspondientes a algunos valores muy cercanos (por ejemplo, 28.6 y 28.7) se han apilado verticalmente para evitar la aglomeración. Hay claramente una enorme variabilidad de un estado a otro. El valor más alto, para D.C., es obviamente un extremo atípico, y los otros cuatro valores en el extremo superior de los datos son candidatos a valores atípicos leves (MA, MN, Nueva York y ND). También hay un grupo de estados en el extremo inferior, situado principalmente en el sur y el suroeste. El porcentaje global para todo el país es de 39.3%; este no es un promedio simple de los 51 números, sino un promedio ponderado por tamaño de la población.

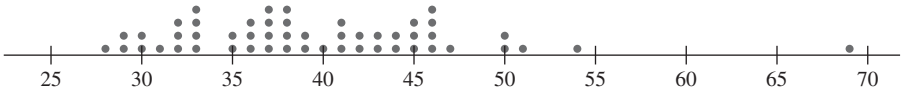


Figura 1.6 Gráfica de puntos para los datos del ejemplo 1.8

Una gráfica de puntos puede ser bastante enfadosa de construir y se ve muy saturada cuando el número de observaciones es grande. La siguiente técnica es muy adecuada en estas situaciones.

Histogramas

Para determinar el valor de una variable algunos datos numéricos se obtienen contando (el número de citatorios de tráfico que una persona recibió durante el año pasado, el número de personas que solicitan empleo durante un periodo específico), mientras que otros datos se obtienen tomando mediciones (el peso de un individuo, el tiempo de reacción a un estímulo particular). La prescripción para trazar un histograma es, en general, diferente en estos dos casos.

DEFINICIÓN

Una variable numérica es **discreta** si su conjunto de valores posibles es finito o si se puede enumerar en una secuencia infinita (una en la cual exista un primer número, un segundo número y así sucesivamente). Una variable numérica es **continua** si sus valores posibles abarcan un intervalo completo sobre la recta numérica.

Una variable discreta x casi siempre resulta de haber contado, en cuyo caso los posibles valores son 0, 1, 2, 3, …, o algún subconjunto de estos enteros. De la toma de mediciones surgen variables continuas. Por ejemplo, si x es el pH de una sustancia química, en teoría x podría ser cualquier número entre 0 y 14: 7.0, 7.03, 7.032, y así sucesivamente. Desde luego, en la práctica existen limitaciones en el grado de precisión de cualquier instrumento de medición, por lo que es posible que no se puedan determinar el pH, el tiempo de reacción, la altura y la concentración con un número arbitrariamente grande de decimales. Sin embargo, con la perspectiva de crear modelos matemáticos de distribuciones de datos, conviene imaginar todo un conjunto continuo de valores posibles.

Considere los datos compuestos de las observaciones de una variable discreta x. La **frecuencia** de cualquier valor particular x es el número de veces que ocurre un valor en el conjunto de datos. La **frecuencia relativa** de un valor es la fracción o proporción de las veces que ocurre el valor:

$$\text{frecuencia relativa de un valor} = \frac{\text{número de veces que ocurre el valor}}{\text{número de observaciones en el conjunto de datos}}$$

Suponga, por ejemplo, que el conjunto de datos se compone de 200 observaciones de $x =$ el número de cursos que un estudiante está tomando en este semestre. Si 70 de estos valores x son 3, entonces

$$\text{frecuencia del valor } x \text{ 3:} \quad 70$$

$$\text{frecuencia relativa del valor } x \text{ 3:} \quad \frac{70}{200} = 0.35$$

Si se multiplica una frecuencia relativa por 100 se obtiene un porcentaje; en el ejemplo de los cursos universitarios, 35% de los estudiantes de la muestra están tomando tres cursos. Las frecuencias relativas, o porcentajes, por lo general interesan más que las frecuencias mismas. En teoría, las frecuencias relativas deberán sumar 1, pero en la práctica la suma puede diferir un poco de 1 debido al redondeo. Una **distribución de frecuencia** es una tabla con las frecuencias o las frecuencias relativas, o ambas.

Construcción de un histograma para datos discretos

En primer lugar, se determinan la frecuencia y la frecuencia relativa de cada valor x. Luego se marcan los valores x posibles en una escala horizontal. Sobre cada valor se traza un rectángulo cuya altura es la frecuencia relativa (o alternativamente, la frecuencia) de dicho valor: Los rectángulos deben medir lo mismo de ancho.

Esta construcción garantiza que el *área* de cada rectángulo sea proporcional a la frecuencia relativa del valor. Por tanto, si las frecuencias relativas de $x = 1$ y $x = 5$ son 0.35 y 0.07, respectivamente, el área del rectángulo por encima de 1 es cinco veces el área del rectángulo por encima de 5.

EJEMPLO 1.9 ¿Qué tan inusual es un juego de béisbol sin *hit* o de un solo *hit* en las ligas mayores y con qué frecuencia un equipo pega más de 10, 15 o incluso 20 *hits*? La tabla 1.1 es una distribución de frecuencia del número de *hits* por equipo y por cada uno de los juegos de nueve episodios que se jugaron entre 1989 y 1993.

Tabla 1.1 Distribución de frecuencia de hits en juegos de nueve entradas

Hits/juego	Número de juegos	Frecuencia relativa	Hits/juego	Número de juegos	Frecuencia relativa
0	20	0.0010	14	569	0.0294
1	72	0.0037	15	393	0.0203
2	209	0.0108	16	253	0.0131
3	527	0.0272	17	171	0.0088
4	1048	0.0541	18	97	0.0050
5	1457	0.0752	19	53	0.0027
6	1988	0.1026	20	31	0.0016
7	2256	0.1164	21	19	0.0010
8	2403	0.1240	22	13	0.0007
9	2256	0.1164	23	5	0.0003
10	1967	0.1015	24	1	0.0001
11	1509	0.0779	25	0	0.0000
12	1230	0.0635	26	1	0.0001
13	834	0.0430	27	1	0.0001
				19 383	1.0005

El histograma correspondiente en la figura 1.7 se eleva suavemente hasta una sola cresta y luego declina. El histograma se extiende un poco más hacia la derecha (hacia valores mayores) que hacia la izquierda, un ligero "asimétrico positivo".

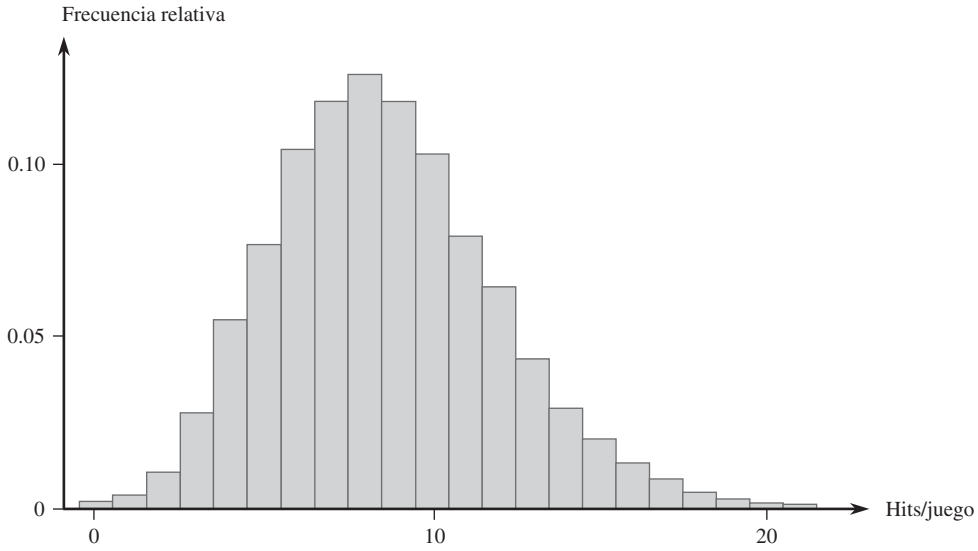


Figura 1.7 Histograma del número de hits por juego de nueve entradas

Con la información tabulada o con el histograma mismo se puede determinar lo siguiente:

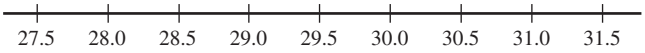
$$\begin{aligned}\text{proporción de juegos de dos hits a lo sumo} &= \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{para } x=0\end{matrix} + \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{para } x=1\end{matrix} + \begin{matrix}\text{frecuencia}\\ \text{relativa}\\ \text{para } x=2\end{matrix}\\ &= 0.0010 + 0.0037 + 0.0108 = 0.0155\end{aligned}$$

De manera similar,

$$\text{proporción de juegos con entre 5 y 10 hits (inclusive)} = 0.0752 + 0.1026 + \cdots + 0.1015 = 0.6361$$

Esto es, aproximadamente 64% de todos los juegos fueron de entre 5 y 10 *hits* (inclusive). ■

La construcción de un histograma para datos continuos (mediciones) implica subdividir el eje de medición entre un número adecuado de **intervalos de clase** o **clases**, de tal suerte que cada observación quede contenida exactamente en una clase. Suponga, por ejemplo, que se hacen 50 observaciones de x = eficiencia de consumo de combustible de un automóvil (mpg), la menor de las cuales es 27.8 y la mayor 31.4. Se podrían utilizar los límites de clase 27.5, 28.0, 28.5, … y 31.55 como se muestra a continuación:



Una dificultad potencial es que de vez en cuando una observación está en un límite de clase, por consiguiente, no cae exactamente en un intervalo, por ejemplo, 29.0. Una forma de tratar este problema es utilizar límites como 27.55, 28.05,…, 31.55. La adición de centésimas a los límites de clase evita que las observaciones queden en los límites resultantes. Otro método es utilizar las clases 27.5 – <28.0, 28.0 – <28.5,…, 31.0 – <31.5. En ese caso 29.0 queda en la clase 29.0 – <29.5 y no en la clase 28.5 – <29.0. En otras palabras, con esta convención una observación que queda en el límite se coloca en el intervalo a la *derecha* del mismo. Así es como Minitab construye un histograma.

Construcción de un histograma para datos continuos: clases con ancho igual

Se determinan la frecuencia y la frecuencia relativa de cada clase. Se marcan los límites de clase sobre un eje de medición horizontal. Sobre cada intervalo de clase se traza un rectángulo cuya altura es la frecuencia relativa correspondiente (o frecuencia).

EJEMPLO 1.10 Las compañías generadoras de electricidad requieren información sobre el consumo de los clientes para obtener pronósticos precisos de la demanda. Investigadores de Wisconsin Power and Light determinaron el consumo de energía (en BTU) durante un periodo particular con una muestra de 90 hogares que utilizan gas. Se calculó un valor de consumo ajustado como sigue:

$$\text{consumo ajustado} = \frac{\text{consumo}}{\text{(clima, en grados-días) (área de la casa)}}$$

Esto dio como resultado los siguientes datos (una parte del conjunto de datos guardados FURNACE.MTW está disponible en Minitab), los cuales se ordenaron desde el valor más pequeño al más grande.

2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

En la figura 1.8 la característica del histograma que más llama la atención es su parecido a una curva en forma de campana, con el punto de simetría aproximadamente en 10.

Clase	1–<3	3–<5	5–<7	7–<9	9–<11	11–<13	13–<15	15–<17	17–<19
Frecuencia	1	1	11	21	25	17	9	4	1
Frecuencia relativa	0.011	0.011	0.122	0.233	0.278	0.189	0.100	0.044	0.011

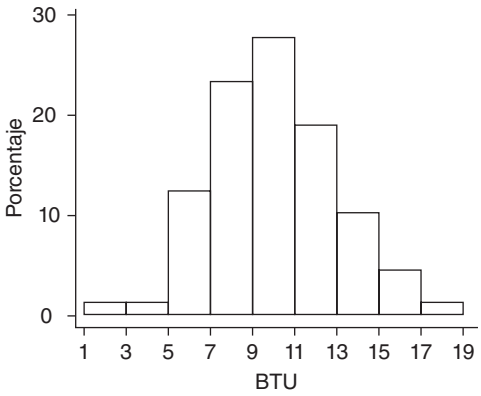


Figura 1.8 Histograma de los datos de consumo de energía del ejemplo 1.10

De acuerdo con el histograma,

$$\begin{pmatrix}\text{proporción de}\\ \text{observaciones}\\ \text{menores que 9}\end{pmatrix} \approx 0.01 + 0.01 + 0.12 + 0.23 = 0.37 \text{ (valor exacto} = \frac{34}{90} = 0.378)$$

La frecuencia relativa para la clase 9–<11 es aproximadamente 0.27, entonces se estima que aproximadamente la mitad de esta, o 0.135, queda entre 9 y 10. Por tanto,

$$\begin{pmatrix}\text{proporción de observaciones}\\ \text{menores que 10}\end{pmatrix} \approx 0.37 + 0.135 = 0.505 \text{ (poco más de 50\%)}$$

El valor exacto de esta proporción es 47/90 = 0.522. ■

No existen reglas inviolables en cuanto al número de clases o a la selección de las mismas. Entre 5 y 20 será satisfactorio para la mayoría de los conjuntos de datos. En general, mientras más grande es el número de observaciones en un conjunto de datos, más clases deberán utilizarse. Una regla empírica razonable es

$$\text{número de clases} \approx \sqrt{\text{número de observaciones}}$$

Es posible que las clases con ancho igual no sean una opción sensible si hay regiones en la escala de medición con una alta concentración de valores y otras donde los datos son muy escasos. La figura 1.9 muestra una gráfica de puntos de dicho conjunto de datos; hay una alta concentración en el medio y relativamente pocas observaciones que se extienden a ambos lados. Con un pequeño número de clases con ancho igual, casi todas las observaciones quedan exactamente en una o dos de las clases. Si se utiliza un número grande de clases con ancho igual, las frecuencias de muchas clases serán cero. Una buena opción es utilizar intervalos más anchos cerca de las observaciones extremas e intervalos más angostos en la región de alta concentración.

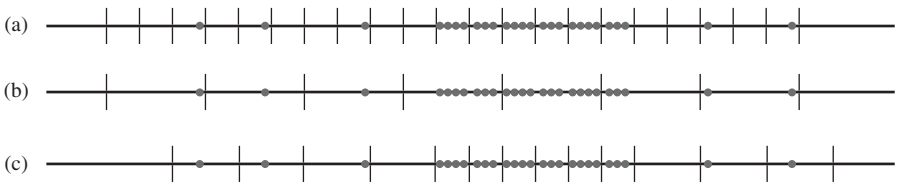


Figura 1.9 Selección de intervalos de clase para datos de "densidad variable": (a) intervalos de ancho igual muy cortos; (b) algunos intervalos de ancho igual; (c) intervalos de ancho desigual

Construcción de un histograma para datos continuos: clases con ancho desigual

Después de determinar las frecuencias y las frecuencias relativas, se calcula la altura de cada rectángulo mediante la fórmula

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa de la clase}}{\text{ancho de clase}}$$

Las alturas del rectángulo resultante se conocen usualmente como *densidades* y la escala vertical es la **escala de densidades**. Esta prescripción también funcionará cuando las clases tengan anchos iguales.

EJEMPLO 1.11 La corrosión del acero de refuerzo es un serio problema en las estructuras de concreto en ambientes afectados por condiciones climáticas severas. Por ello los investigadores han analizado el uso de barras de refuerzo fabricadas de un material compuesto. Se realizó un estudio para desarrollar directrices para adherir barras de refuerzo reforzadas con fibra de vidrio al concreto (**"Design Recommendations for Bond of GFRP Rebars to Concrete",** ***J. of Structural Engr*****., 1996: 247-254**). Considere las siguientes 48 observaciones de mediciones de fuerza adhesiva:

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0	13.4	17.1	9.3	5.6
5.7	5.4	5.2	5.1	4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9	8.2	10.7	14.2	7.6
5.2	5.5	5.1	5.0	5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

Clase	2 –<4	4–<6	6–<8	8–<12	12–<20	20–<30
Frecuencia	9	15	5	9	8	2
Frecuencia relativa	0.1875	0.3125	0.1042	0.1875	0.1667	0.0417
Densidad	0.094	0.156	0.052	0.047	0.021	0.004

El histograma resultante se muestra en la figura 1.10. La cola derecha o superior se alarga mucho más que la izquierda o inferior, un sustancial alejamiento de la simetría.

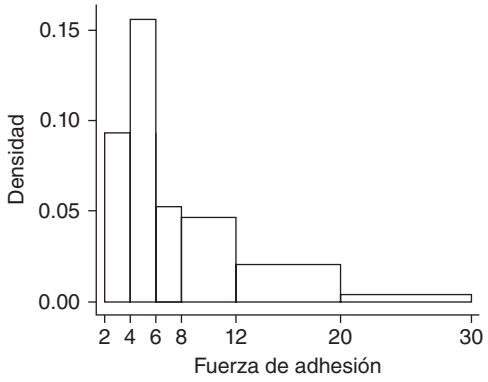


Figura 1.10 Histograma Minitab de densidad para la fuerza de adhesión del ejemplo 1.11 ■

Cuando las clases tienen anchos desiguales, sin utilizar una escala de densidades se obtendrá una gráfica con áreas distorsionadas. Para clases con anchos iguales el divisor es el mismo en cada cálculo de densidad y la aritmética adicional simplemente implica cambiar la escala en el eje vertical (es decir, el histograma que utiliza frecuencia relativa y el que utiliza densidad tendrán exactamente la misma apariencia). Un histograma de densidad tiene una propiedad interesante. Si se multiplican ambos miembros de la fórmula para la densidad por el ancho de clase, se obtiene

$$\begin{matrix}\text{frecuencia}\\ \text{relativa}\end{matrix} = \begin{matrix}\text{(ancho de clase)}\\ \text{(densidad)}\end{matrix} = \begin{matrix}\text{(ancho del rectángulo)}\\ \text{(altura del rectángulo)}\end{matrix} = \begin{matrix}\text{área del}\\ \text{rectángulo}\end{matrix}$$

Es decir, *el área de cada rectángulo es la frecuencia relativa de la clase correspondiente*. Además, puesto que la suma de frecuencias relativas debe ser 1, *el área total de todos los rectángulos en un histograma de densidad es* 1. Siempre es posible trazar un

histograma de modo que el área sea igual a la frecuencia relativa (esto es cierto también para un histograma de datos discretos); simplemente se utiliza la escala de densidad. Esta propiedad desempeñará un papel importante al crear modelos de distribución en el capítulo 4.

Formas de histograma

Los histogramas se presentan en varias formas. Un histograma **unimodal** es el que se eleva a una sola cresta y luego declina. Uno **bimodal** tiene dos crestas diferentes. Puede ocurrir bimodalidad cuando el conjunto de datos se compone de observaciones de dos clases, bastante diferentes, de individuos u objetos. Por ejemplo, considere un gran conjunto de datos compuesto de los tiempos de manejo de automóviles en el trayecto entre San Luis Obispo, California y Monterey, California (sin contar el tiempo que se utilice para visitar lugares de interés, en comer, etc.). Este histograma mostraría dos crestas, una para los autos que toman la ruta interior (aproximadamente 2.5 horas) y otra para los que recorren la costa (3.5-4 horas). La bimodalidad no se presenta automáticamente en dichas situaciones. Sólo si los dos distintos histogramas están "muy alejados" respecto a sus dispersiones, la bimodalidad ocurrirá en el histograma de datos combinados. Por consiguiente, un conjunto de datos grande compuesto de las estaturas de los estudiantes universitarios no producirá un histograma bimodal porque la altura típica de los hombres, que aproximadamente es de 69 pulgadas, no está demasiado por encima de la altura típica de las mujeres, que es aproximadamente de 64-65 pulgadas. Se dice que un histograma con más de dos crestas es **multimodal**. Por supuesto, el número de crestas dependerá de la selección de intervalos de clase, en particular, con un pequeño número de observaciones. Mientras más grande es el número de clases, más probable es que se manifiesten bimodalidad o multimodalidad.

EJEMPLO 1.12 La figura 1.11(a) muestra un histograma Minitab de los pesos (en libras, lb) de los 124 jugadores que figuraban en las listas de los 49's de San Francisco y de los Patriotas de Nueva Inglaterra (equipos que al autor le gustaría ver reunidos en el Súper Tazón) el 20 de noviembre de 2009. La figura 1.11(b) es un histograma suavizado (que en realidad se llama *densidad estimada*) de los datos del paquete de software R. Tanto el histograma como el histograma suavizado muestran tres picos diferentes; el primero a la derecha es para los *linieros*, el del centro corresponde al peso de los *apoyadores* y el pico de la izquierda es para todos los demás jugadores (receptores abiertos, mariscales de campo, etc.).

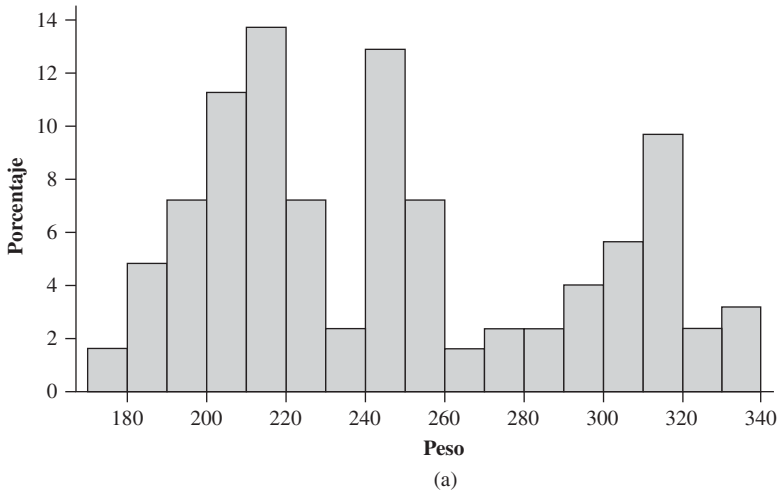


Figura 1.11 Peso de los jugadores de la NFL. (a) histograma y (b) histograma suavizado

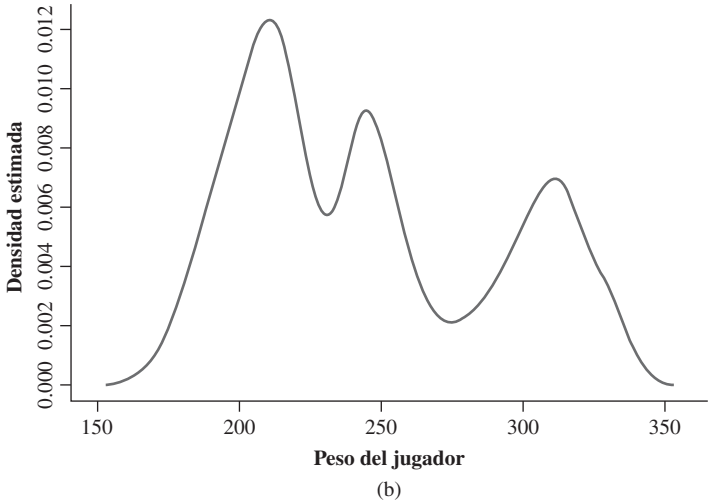


Figura 1.11 (continuación)

Un histograma es **simétrico** si la mitad izquierda es una imagen en espejo de la mitad derecha. Un histograma unimodal es **positivamente asimétrico** si la cola derecha o superior se alarga en comparación con la cola izquierda o inferior, y **negativamente asimétrico** si el alargamiento es hacia la izquierda. La figura 1.12 muestra histogramas "suavizados", que se obtuvieron superponiendo una curva suavizada sobre los rectángulos e ilustran las varias posibilidades.

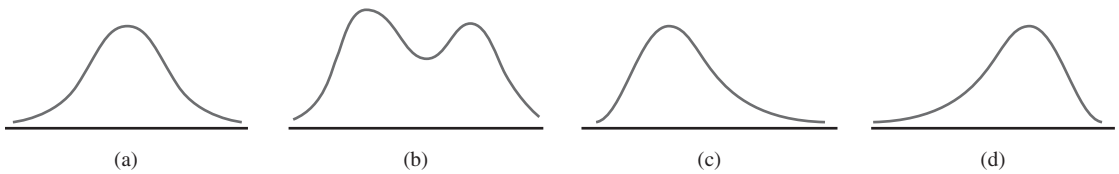


Figura 1.12 Histogramas suavizados: (a) unimodal simétrico; (b) bimodal; (c) positivamente asimétrico y (d) negativamente asimétrico

Datos cualitativos

Tanto una distribución de frecuencia como un histograma pueden ser construidos cuando el conjunto de datos es de naturaleza *cualitativa* (categórico). En algunos casos habrá un ordenamiento natural de las clases, por ejemplo, estudiantes de primer año, de segundo, de tercero, de cuarto y graduados, mientras que en otros casos el orden será arbitrario, por ejemplo, católico, judío, protestante, etcétera. Con estos datos categóricos los intervalos sobre los cuales se construyen los rectángulos deberán ser de ancho igual.

EJEMPLO 1.13 El **Public Policy Institute of California** realizó una encuesta telefónica entre 2501 residentes adultos durante abril de 2006 para indagar lo que pensaban sobre varios aspectos de la educación pública K-12. Una pregunta fue "En general, ¿cómo calificaría la calidad de las escuelas públicas de su vecindario hoy en día?". La tabla 1.2 muestra las frecuencias y las frecuencias relativas, y la figura 1.13 muestra el histograma correspondiente (gráfica de barras).

Tabla 1.2 Distribución de frecuencia para los datos de la calificación de las escuelas

Calificación	Frecuencia	Frecuencia relativa
A	478	0.191
B	893	0.357
C	680	0.272
D	178	0.071
F	100	0.040
No sabe	172	0.069
	2501	1.000

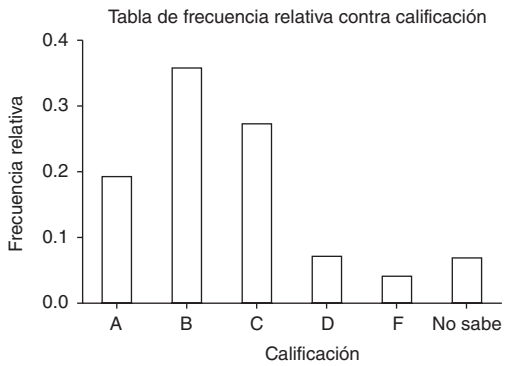


Figura 1.13 Histograma Minitab de los datos de la calificación

Más de la mitad de los encuestados otorgó una calificación A o B y sólo poco más de 10% otorgó una calificación D o F. Los porcentajes de los padres de niños que asisten a escuelas públicas fueron un poco más favorables para las escuelas: 24%, 40%, 24%, 6%, 4% y 2%. ■

Datos multivariantes

En general los datos multivariantes son más difíciles de describir de forma visual. Más adelante se muestran varios métodos para ello, particularmente gráficas de dispersión para datos numéricos bivariantes.

EJERCICIOS Sección 1.2 (10–32)

10. Considere los datos de resistencia de las vigas del ejemplo 1.2.
- **a.** Construya una gráfica de tallos y hojas de los datos. ¿Cuál parece ser el valor de resistencia representativo? ¿Parecen estar las observaciones altamente concentradas en torno al valor representativo, o algo dispersas?
- **b.** ¿Parece la gráfica razonablemente simétrica en torno a un valor representativo, o describiría su forma de otra manera?
- **c.** ¿Habrá algunos valores de resistencia extremos?
- **d.** ¿Qué proporción de las observaciones de resistencia en esta muestra exceden de 10 MPa?

11. En el artículo **"Bolted Connection Design Values Based on European Yield Model" (*J. of Structural Engr.*, 1993: 2169-2186)** se publican los valores de gravedad específica de varios tipos de madera que se utilizan en la construcción:

0.31 0.35 0.36 0.36 0.37 0.38 0.40 0.40 0.40
0.41 0.41 0.42 0.42 0.42 0.42 0.42 0.43 0.44
0.45 0.46 0.46 0.47 0.48 0.48 0.48 0.51 0.54
0.54 0.55 0.58 0.62 0.66 0.66 0.67 0.68 0.75

Construya una gráfica de tallos y hojas con tallos repetidos y comente sobre cualquier característica interesante de la gráfica.

12. Los datos adjuntos de granulometrías (nm) de CeO_2 bajo ciertas condiciones experimentales fueron leídos de una gráfica en el artículo **"Nanoceria—Energetics of Surfaces, Interfaces and Water Adsorption" (*J. of the Amer. Ceramic Soc.*, 2011: 3992-3999):**

3.0 – < 3.5	3.5 – < 4.0	4.0 – < 4.5	4.5 – < 5.0	5.0 – < 5.5
5	15	27	34	22
5.5 – < 6.0	6.0 – < 6.5	6.5 – < 7.0	7.0 – < 7.5	7.5 – < 8.0
14	7	2	4	1

a. ¿Qué proporción de las observaciones son menores de 5?

b. ¿Qué proporción de las observaciones son al menos 6?

c. Construya un histograma con frecuencia relativa en el eje vertical y comente las características interesantes. En particular, la distribución de tamaños de partícula ¿parece razonablemente simétrica o algo sesgada? [*Nota*: Los investigadores ajustan los datos a una distribución logarítmica-normal; esto se analiza en el capítulo 4.]

d. Construya un histograma con la densidad en el eje vertical y compárelo con el histograma del inciso c).

13. Las propiedades mecánicas permisibles para el diseño estructural de vehículos aeroespaciales metálicos requieren un método aprobado para analizar estadísticamente los datos de prueba empíricos. El artículo **"Establishing Mechanical Property Allowables for Metals" (*J. of Testing and Evaluation*, 1998: 293-299)** utilizó los datos anexos sobre resistencia a la tensión última ($kg/pulg^2$) como base para abordar las dificultades que se presentan en el desarrollo de dicho método.

122.2	124.2	124.3	125.6	126.3	126.5	126.5	127.2	127.3
127.5	127.9	128.6	128.8	129.0	129.2	129.4	129.6	130.2
130.4	130.8	131.3	131.4	131.4	131.5	131.6	131.6	131.8
131.8	132.3	132.4	132.4	132.5	132.5	132.5	132.5	132.6
132.7	132.9	133.0	133.1	133.1	133.1	133.1	133.2	133.2
133.2	133.3	133.3	133.5	133.5	133.5	133.8	133.9	134.0
134.0	134.0	134.0	134.1	134.2	134.3	134.4	134.4	134.6
134.7	134.7	134.7	134.8	134.8	134.8	134.9	134.9	135.2
135.2	135.2	135.3	135.3	135.4	135.5	135.5	135.6	135.6
135.7	135.8	135.8	135.8	135.8	135.8	135.9	135.9	135.9
135.9	136.0	136.0	136.1	136.2	136.2	136.3	136.4	136.4
136.6	136.8	136.9	136.9	137.0	137.1	137.2	137.6	137.6
137.8	137.8	137.8	137.9	137.9	138.2	138.2	138.3	138.3
138.4	138.4	138.4	138.5	138.5	138.6	138.7	138.7	139.0
139.1	139.5	139.6	139.8	139.8	140.0	140.0	140.7	140.7
140.9	140.9	141.2	141.4	141.5	141.6	142.9	143.4	143.5
143.6	143.8	143.8	143.9	144.1	144.5	144.5	147.7	147.7

a. Construya una gráfica de tallos y hojas de los datos eliminando (truncando) los dígitos de décimos y luego repitiendo cada valor de tallo cinco veces (una vez para las hojas 1 y 2, una segunda vez para las hojas 3 y 4, etc.). ¿Por qué es relativamente fácil identificar un valor de resistencia representativo?

b. Construya un histograma utilizando clases con ancho igual con la primera clase que tiene un límite inferior de 122 y un límite superior de 124. Enseguida comente sobre cualquier característica interesante del histograma.

14. El conjunto de datos adjunto se compone de observaciones del flujo de una regadera (L/min) para una muestra de $n = 129$ casas en Perth, Australia **("An Application of Bayes Methodology to the Analysis of Diary Records in a Water Use Study", *J. Amer. Stat. Assoc.*, 1987: 705-711)**:

4.6	12.3	7.1	7.0	4.0	9.2	6.7	6.9	11.5	5.1
11.2	10.5	14.3	8.0	8.8	6.4	5.1	5.6	9.6	7.5
7.5	6.2	5.8	2.3	3.4	10.4	9.8	6.6	3.7	6.4
8.3	6.5	7.6	9.3	9.2	7.3	5.0	6.3	13.8	6.2
5.4	4.8	7.5	6.0	6.9	10.8	7.5	6.6	5.0	3.3
7.6	3.9	11.9	2.2	15.0	7.2	6.1	15.3	18.9	7.2
5.4	5.5	4.3	9.0	12.7	11.3	7.4	5.0	3.5	8.2
8.4	7.3	10.3	11.9	6.0	5.6	9.5	9.3	10.4	9.7
5.1	6.7	10.2	6.2	8.4	7.0	4.8	5.6	10.5	14.6
10.8	15.5	7.5	6.4	3.4	5.5	6.6	5.9	15.0	9.6
7.8	7.0	6.9	4.1	3.6	11.9	3.7	5.7	6.8	11.3
9.3	9.6	10.4	9.3	6.9	9.8	9.1	10.6	4.5	6.2
8.3	3.2	4.9	5.0	6.0	8.2	6.3	3.8	6.0	

a. Construya una gráfica de tallos y hojas de los datos.

b. ¿Cuál es una velocidad de flujo o gasto típico o representativo?

c. La gráfica ¿parece estar altamente concentrada o dispersa?

d. ¿Es la distribución de valores razonablemente simétrica? Si no, ¿cómo describiría el alejamiento de la simetría?

e. ¿Describiría alguna observación como alejada del resto de los datos (un valor atípico)?

15. Los tiempos de duración de las películas estadounidenses ¿difieren de alguna manera de las del cine francés? El autor investigó esta cuestión seleccionando aleatoriamente 25 películas recientes de cada tipo, lo que resulta en los siguientes tiempos de duración (min):

Am:	94	90	95	93	128	95	125	91	104	116	162	102	90
	110	92	113	116	90	97	103	95	120	109	91	138	
Fr:	123	116	90	158	122	119	125	90	96	94	137	102	
	105	106	95	125	122	103	96	111	81	113	128	93	92

Construya una gráfica de tallos y hojas *comparativa* y haga una lista de tallos a la mitad de la página, y luego ubique las hojas Am a la izquierda y las Fr a la derecha. A continuación comente las características interesantes de la gráfica.

16. El artículo citado en el ejemplo 1.2 también dio las observaciones de resistencia adjuntas para los cilindros:

6.1 5.8 7.8 7.1 7.2 9.2 6.6 8.3 7.0 8.3
7.8 8.1 7.4 8.5 8.9 9.8 9.7 14.1 12.6 11.2

a. Construya una gráfica de tallos y hojas comparativa (véase el ejercicio previo) de los datos de la viga y el cilindro y luego responda las preguntas de los incisos b) al d) del ejercicio 10 para las observaciones de los cilindros.

b. ¿En qué formas son similares los dos lados de la gráfica? ¿Existen diferencias obvias entre las observaciones de la viga y las observaciones del cilindro?

c. Construya una gráfica de puntos de los datos del cilindro.

17. Los datos adjuntos proceden de un estudio de contubernios en las licitaciones dentro de la industria de la construcción (**"Detection of Collusive Behavior",** ***J. of Construction Engr. and Mgmnt*****, 2012: 1251-1258**).

Núm. Concursantes	**Núm. Contratos**
2	7
3	20
4	26
5	16
6	11
7	9
8	6
9	8
10	3
11	2

a. ¿Qué proporción de contratos implica a lo más a cinco concursantes? ¿Y al menos a cinco concursantes?

b. ¿Qué proporción de contratos implica entre cinco y 10 concursantes, inclusive? ¿Y estrictamente entre cinco y 10 concursantes?

c. Construya un histograma y comente las características interesantes.

18. Cada corporación tiene un consejo de directores. El número de personas en un consejo varía de una empresa a otra. Uno de los autores del artículo **"Does Optimal Corporate Board Size Exist? An Empirical Analysis" (*J. of Applied Finance*, 2010: 57-69)** proporciona los datos del número de directores en cada consejo, en una muestra aleatoria de 204 corporaciones.

Núm. de directores:	4	5	6	7	8	9
Frecuencia:	3	12	13	25	24	42
Núm. de directores:	10	11	12	13	14	15
Frecuencia:	23	19	16	11	5	4
Núm. de directores:	16	17	21	24	32	
Frecuencia:	1	3	1	1	1	

a. Construya un histograma de los datos con base en frecuencias relativas y comente cualquier característica interesante.

b. Construya una distribución de frecuencia en la cual se incluyan en la última fila todos los consejos con al menos 18 directores. ¿Si esta distribución se muestra en el citado artículo, podría dibujar un histograma? Explique.

c. ¿Qué proporción de estas corporaciones tienen a lo más 10 directores?

d. ¿Qué proporción de estas empresas tiene más de 15 directores?

19. Se determinó el número de partículas contaminantes en una oblea de silicio antes de cierto proceso de enjuague para cada oblea en una muestra de tamaño 100 y se obtuvieron las siguientes frecuencias:

Número de partículas	0	1	2	3	4	5	6	7
Frecuencia	1	2	3	12	11	15	18	10
Número de partículas	8	9	10	11	12	13	14	
Frecuencia	12	4	5	3	1	2	1	

a. ¿Qué proporción de las obleas de la muestra tuvo al menos una partícula? ¿Y al menos cinco partículas?

b. ¿Qué proporción de las obleas de la muestra tuvo entre cinco y diez partículas, inclusive? ¿Y estrictamente entre cinco y diez partículas?

c. Trace un histograma con la frecuencia relativa en el eje vertical. ¿Cómo describiría la forma del histograma?

20. El artículo **"Determination of Most Representative Subdivision" (*J. of Energy Engr.*, 1993: 43-55)** proporciona datos sobre varias características de subdivisiones que podrían utilizarse para decidir si se suministra energía eléctrica mediante líneas elevadas o por medio de líneas subterráneas. He aquí los valores de la variable x = longitud total de calles dentro de una subdivisión:

1280 5320 4390 2100 1240 3060 4770
1050 360 3330 3380 340 1000 960
1320 530 3350 540 3870 1250 2400
960 1120 2120 450 2250 2320 2400
3150 5700 5220 500 1850 2460 5850
2700 2730 1670 100 5770 3150 1890
510 240 396 1419 2109

a. Construya una gráfica de hojas y tallos con el dígito de los millares como tallo y el dígito de las centenas como las hojas, y comente sobre las diferentes características de la gráfica.

b. Construya un histograma con los límites de clase, 0, 1000, 2000, 3000, 4000, 5000 y 6000. ¿Qué proporción de subdivisiones tiene una longitud total menor que 2000? ¿Entre 2000 y 4000? ¿Cómo describiría la forma del histograma?

21. El artículo citado en el ejercicio 20 también aporta los siguientes valores de las variables y = número de calles cerradas y z = número de intersecciones:

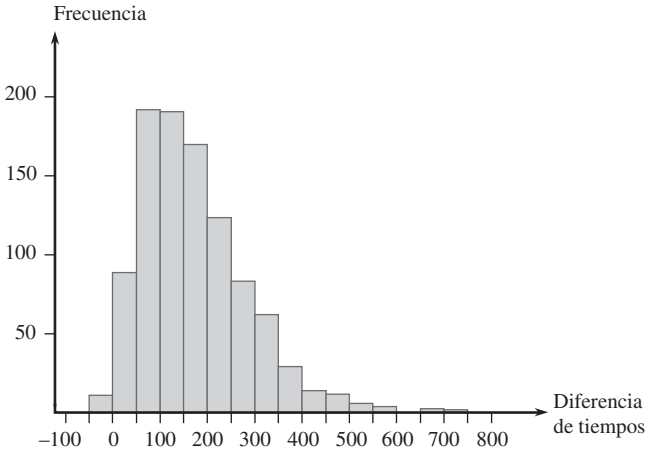
y	1	0	1	0	0	2	0	1	1	1	2	1	0	0	1	1	0	1	1
z	1	8	6	1	1	5	3	0	0	4	4	0	0	1	2	1	4	0	4
y	1	1	0	0	0	1	1	2	0	1	2	2	1	1	0	2	1	1	0
z	0	3	0	1	1	0	1	3	2	4	6	6	0	1	1	8	3	3	5
y	1	5	0	3	0	1	1	0	0										
z	0	5	2	3	1	0	0	0	3										

a. Construya un histograma con los datos y. ¿Qué proporción de estas subdivisiones no tenía calles cerradas? ¿Al menos una calle cerrada?

b. Construya un histograma con los datos z. ¿Qué proporción de estas subdivisiones tenía cuando mucho cinco intersecciones? ¿Y menos de cinco intersecciones?

22. ¿Cómo varía la velocidad de un corredor durante un maratón (una distancia de 42.195 km)? Considere determinar tanto el tiempo de recorrido de los primeros 5 km como el tiempo de recorrido entre los 35 y los 40 km, y luego reste el primer tiempo del segundo. Un valor positivo de esta diferencia corresponde a un corredor que avanza más lento hacia el final de la carrera. El histograma adjunto está basado en los tiempos de corredores que participaron en varios maratones japoneses (**"Factors Affecting Runners' Maraton Performance", *Chance*, otoño de 1993: 24-30**). ¿Cuáles son algunas características interesantes de este histograma? ¿Cuál es un valor de diferencia típico? ¿Aproximadamente qué proporción de los participantes corren la última distancia más rápido que la primera?

Histograma para el ejercicio 22



23. El artículo **"Statistical Modeling of the Time Course of Tantrum Anger" (*Annals of Applied Stats*, 2009: 1013-1034)** analiza cómo la intensidad de la ira en los berrinches de los niños puede estar relacionada con la duración de la rabieta, así como con los indicadores de comportamiento, tales como gritar, arañar y empujar o tirar. Se proporciona siguiente la distribución de frecuencias (y también el histograma correspondiente):

0–<2:	136	2–<4:	92	4–<11:	71
11–<20:	26	20–<30:	7	30–<40:	3

Construya un histograma y comente sobre las características interesantes.

24. El conjunto de datos adjuntos consiste en observaciones de resistencia al esfuerzo cortante (lb) de soldaduras de puntos ultrasónicas aplicadas en un cierto tipo de lámina alclad. Construya un histograma de frecuencia relativa basado en diez clases de ancho igual con límites 4000, 4200, … [El histograma concordará con el que se muestra en **"Comparison of Properties of Joints Prepared by Ultrasonic Welding and Other Means" (*J. of Aircraft*, 1983: 552-556)**. Comente sobre sus características.

5434	4948	4521	4570	4990	5702	5241
5112	5015	4659	4806	4637	5670	4381
4820	5043	4886	4599	5288	5299	4848
5378	5260	5055	5828	5218	4859	4780
5027	5008	4609	4772	5133	5095	4618
4848	5089	5518	5333	5164	5342	5069
4755	4925	5001	4803	4951	5679	5256
5207	5621	4918	5138	4786	4500	5461
5049	4974	4592	4173	5296	4965	5170
4740	5173	4568	5653	5078	4900	4968
5248	5245	4723	5275	5419	5205	4452
5227	5555	5388	5498	4681	5076	4774
4931	4493	5309	5582	4308	4823	4417
5364	5640	5069	5188	5764	5273	5042
5189	4986					

25. Una transformación de valores de datos mediante alguna función matemática, tal como $\sqrt{x}$ o $1/x$ a menudo produce un conjunto de números con "mejores" propiedades estadísticas que los datos originales. En particular, es posible encontrar una función para la cual el histograma de valores transformados es más simétrico (o, incluso, mejor, más como una curva en forma de campana) que los datos originales. Por ejemplo, el artículo **"Time Lapse Cinematographic Analysis of Beryllium-Lung Fibroblast Interactions" (*Environ. Research*, 1983: 34-43)** reportó los resultados de los experimentos diseñados para estudiar el comportamiento de ciertas células individuales que habían estado expuestas a berilio. Una importante característica de dichas células individuales es su tiempo de interdivisión (IDT, por sus siglas en inglés). Se determinaron tiempos de interdivisión de un gran número de células, tanto en condiciones expuestas (tratamiento) como en no expuestas (control).

Los autores del artículo utilizaron una transformación logarítmica, es decir, valor transformado = log(valor original). Considere los siguientes tiempos de interdivisión representativos.

IDT	$\log_{10}$(IDT)	IDT	$\log_{10}$(IDT)	IDT	$\log_{10}$(IDT)
28.1	1.45	60.1	1.78	21.0	1.32
31.2	1.49	23.7	1.37	22.3	1.35
13.7	1.14	18.6	1.27	15.5	1.19
46.0	1.66	21.4	1.33	36.3	1.56
25.8	1.41	26.6	1.42	19.1	1.28
16.8	1.23	26.2	1.42	38.4	1.58
34.8	1.54	32.0	1.51	72.8	1.86
62.3	1.79	43.5	1.64	48.9	1.69
28.0	1.45	17.4	1.24	21.4	1.33
17.9	1.25	38.8	1.59	20.7	1.32
19.5	1.29	30.6	1.49	57.3	1.76
21.1	1.32	55.6	1.75	40.9	1.61
31.9	1.50	25.5	1.41		
28.9	1.46	52.1	1.72		

Use los intervalos de clase 10–<20, 20–<30,… para construir un histograma de los datos originales. Use los intervalos 1.1–<1.2, 1.2–<1.3,… para hacer lo mismo con los datos transformados. ¿Cuál es el efecto de la transformación?

26. En la actualidad se está utilizando la difracción retrodispersada de electrones en el estudio de fenómenos de fractura. La siguiente información sobre ángulo de desorientación (grados) se extrajo del artículo **"Observations on the Faceted Initiation Site in the Dwell-Fatigue Tested Ti-6242 Alloy: Crystallographic Orientation and Size Effects" (*Metallurgical and Materials Trans*., 2006: 1507-1518)**.

Clase:	0–<5	5–<10	10–<15	15–<20
Frecuencia relativa:	0.177	0.166	0.175	0.136
Clase:	20–<30	30–<40	40–<60	60–<90
Frecuencia relativa:	0.194	0.078	0.044	0.030

a. ¿Será verdad que más de 50% de los ángulos muestreados son más pequeños de 15°, como se afirma en el artículo?

b. ¿Qué proporción de los ángulos muestreados son al menos de 30°?

c. ¿Aproximadamente qué proporción de los ángulos está entre 10 y 25°?

d. Construya un histograma y comente sobre cualquier característica interesante.

27. El artículo **"Study on the Life Distribution of Microdrills" (*J. of Engr. Manufacture*, 2002: 301-305)** reporta las siguientes observaciones, listadas en orden ascendente, sobre la duración de las brocas (número de agujeros que fresa una broca antes de romperse) cuando se fresaron agujeros en una cierta aleación de latón.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

a. ¿Por qué una distribución de frecuencia no puede estar basada en los intervalos de clase 0–50, 50–100, 100–150, y así sucesivamente?

b. Construya una distribución de frecuencia e histograma de los datos con los límites de clase 0, 50, 100,…, y luego comente sobre las características interesantes.

c. Construya una distribución de frecuencia y el histograma de los logaritmos naturales de las observaciones de duración y comente sobre las características interesantes.

d. ¿Qué proporción de las observaciones de duración en esta muestra es menor de 100? ¿Qué proporción de las observaciones es al menos de 200?

28. La distribución de frecuencia adjunta en energía depositada (mJ) fue extraída del artículo **"Experimental Analysis of Laser-Induced Spark Ignition of Lean Turbulent Premixed Flames" (*Combustion and Flame*, 2013: 1414-1427)**.

1.0–<2.0	5	2.0–<2.4	11
2.4–<2.6	13	2.6–<2.8	30
2.8–<3.0	46	3.0–<3.2	66
3.2–<3.4	133	3.4–<3.6	141
3.6–<3.8	126	3.8–<4.0	92
4.0–<4.2	73	4.2–<4.4	38
4.4–<4.6	19	4.6–<5.0	11

a. ¿Qué proporción de estos ensayos de ignición da como resultado una energía depositada de menos de 3 mJ?

b. ¿Qué proporción de estos ensayos de ignición resulta en una energía depositada de al menos 4 mJ?

c. Aproximadamente, ¿qué proporción de ensayos resulta en una energía depositada de al menos 3.5 mJ?

d. Construya un histograma y comente acerca de su forma.

29. En el artículo **"Finding Occupational Accident Patterns in the Extractive Industry Using a Systematic Data Mining Approach" (*Reliability Engr. and System Safety*, 2012: 108-122)** se presentaron las siguientes categorías por tipo de actividad física, cuando ocurrió un accidente industrial:

A. Trabajo con herramientas manuales
B. Movimiento
C. Portar a mano
D. Manipulación de objetos
E. Operación de una máquina
F. Otros

Construya una distribución de frecuencia, incluyendo frecuencias relativas y un histograma para los datos adjuntos de 100 accidentes (los porcentajes concuerdan con los del artículo citado):

A B D A A F C A C B E B A C
F D B C D A A C B E B C E A
B A A A B C C D F D B B A F
C B A C B E E D A B C E A A
F C B D D D B D C A F A A B
D E A E D B C A F A C D D A
A B A F D C A C B F D A E A
C D

30. Un **diagrama de Pareto** es una variación de un histograma de datos categóricos producidos por un estudio de control de calidad. Cada categoría representa un tipo diferente de no conformidad del producto o problema de producción. Las categorías se ordenaron de tal modo que en el extremo izquierdo aparezca la categoría con la frecuencia más grande, enseguida la categoría con la segunda frecuencia más grande, y así sucesivamente. Suponga que se obtiene la siguiente información sobre no conformidades en paquetes de circuito: componentes averiados, 126; componentes incorrectos, 210; soldadura insuficiente, 67; soldadura excesiva, 54; componente faltante, 131. Construya un diagrama de Pareto.

31. La **frecuencia acumulada** y la frecuencia relativa acumulada de un intervalo de clase particular son la suma de las frecuencias y las frecuencias relativas, respectivamente, del intervalo y todos los intervalos que quedan debajo de él. Si, por ejemplo, tenemos cuatro intervalos con frecuencias 9, 16, 13 y 12, entonces las frecuencias acumuladas serán 9, 25, 38 y 50; y las frecuencias relativas acumuladas serán 0.18, 0.50, 0.76 y 1.00. Calcule las frecuencias acumuladas y las frecuencias relativas acumuladas de los datos del ejercicio 24.

32. La carga de fuego (MJ/m^2) es la energía calorífica que podría ser liberada por cada metro cuadrado de área de piso debido a la combustión del contenido y la propia estructura. El artículo **"Fire Loads in Office Buildings" (*J. of Structural Engr.*, 1997: 365-368)** dio los siguientes porcentajes acumulados (tomados de una gráfica) de cargas de fuego en una muestra de 388 cuartos:

Valor	0	150	300	450	600
% acumulado	0	19.3	37.6	62.7	77.5
Valor	750	900	1050	1200	1350
% acumulado	87.2	93.8	95.7	98.6	99.1
Valor	1500	1650	1800	1950	
% acumulado	99.5	99.6	99.8	100.0	

a. Construya un histograma de frecuencia relativa y comente sobre las características interesantes.

b. ¿Qué proporción de cargas de fuego es menor de 600? ¿Y al menos menor de 1200?

c. ¿Qué proporción de las cargas está entre 600 y 1200?

1.3 Medidas de ubicación

Los resúmenes visuales de datos son herramientas excelentes para obtener impresiones y percepciones preliminares. Un análisis de datos más formal a menudo requiere el cálculo y la interpretación de medidas resumidas numéricas. Es decir, se trata de extraer varios números resumidos a partir de los datos, números que podrían servir para caracterizar el conjunto de datos y comunicar algunas de sus características prominentes. El interés principal se concentrará en los datos numéricos; al final de la sección aparecen algunos comentarios respecto a los datos categóricos.

Suponga, entonces, que el conjunto de datos es de la forma $x_1, x_2, \ldots, x_n$, donde cada x_i es un número. ¿Qué características del conjunto de números son de mayor interés y merecen énfasis? Una importante característica de un conjunto de números es su ubicación y en particular su centro. Esta sección presenta métodos para describir la ubicación de un conjunto de datos; en la sección 1.4 se regresará a los métodos para medir la variabilidad en un conjunto de números.

La media

Para un conjunto dado de números $x_1, x_2, \ldots, x_n$, la medida más conocida y útil del centro es la *media* o el promedio aritmético del conjunto. Como casi siempre pensaremos que los números x_i constituyen una muestra, a menudo se hará referencia al promedio aritmético como la *media muestral* y se la denotará mediante $\bar{x}$.

DEFINICIÓN

La **media muestral** $\bar{x}$ de las observaciones $x_1, x_2, \ldots, x_n$ está dada por

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

El numerador de $\bar{x}$ se escribe más informalmente como $\sum x_i$, donde la suma incluye todas las observaciones muestrales.

Para reportar $\bar{x}$ se recomienda utilizar una precisión decimal de un dígito más que la precisión de los números x_i. Por consiguiente, si las observaciones son distancias de detención con $x_1 = 125$, $x_2 = 131$, y así sucesivamente, se podría tener $\bar{x} = 127.3$ pies.

EJEMPLO 1.14 En los últimos años ha habido un creciente interés comercial en el uso de lo que se conoce como *concreto internamente curado*. Este concreto comúnmente tiene inclusiones porosas en forma de agregado ligero (LWA). El artículo **"Characterizing Lightweight Aggregate Desorption at High Relative Humidities Using a Pressure Plate Apparatus" (*J. of Materials in Civil Engr*, 2012: 961-969)** reporta sobre un estudio en el cual los investigadores examinaron diversas propiedades físicas de 14 especímenes LWA. Estos son los porcentajes de absorción de agua de los especímenes durante 24 horas:

$x_1 = 16.0$	$x_2 = 30.5$	$x_3 = 17.7$	$x_4 = 17.5$	$x_5 = 14.1$
$x_6 = 10.0$	$x_7 = 15.6$	$x_8 = 15.0$	$x_9 = 19.1$	$x_{10} = 17.9$
$x_{11} = 18.9$	$x_{12} = 18.5$	$x_{13} = 12.2$	$x_{14} = 6.0$	

La figura 1.14 muestra una gráfica de puntos de los datos; un porcentaje de absorción de agua en medio de la decena entre diez y veinte parece ser "típico". Con $\sum x_i = 229.0$, la media muestral es

$$\bar{x} = \frac{229.0}{14} = 16.36$$

Una interpretación física de la media muestral nos indica cómo se evalúa el centro de una muestra. Cada punto en la gráfica de puntos se considera la representación de un peso de 1 lb. Entonces un punto de apoyo colocado con su punta en el eje horizontal estará en equilibrio precisamente cuando se encuentra en $\bar{x}$ (véase la figura 1.14). Por lo que la media muestral puede considerarse el punto de equilibrio de la distribución de las observaciones.

Figura 1.14 Gráfica de puntos de los datos del ejemplo 1.14 ■

Así como $\bar{x}$ representa el valor promedio de las observaciones incluidas en una muestra, es posible calcular el promedio de todos los valores de la población. Este promedio se conoce como la **media de la población** y se denota por la letra griega μ. Cuando existen N valores de la población (una población finita), entonces μ = (suma de los valores de población N)/N. En los capítulos 3 y 4 se dará una definición más general de μ, que se aplica a poblaciones tanto finitas como (conceptualmente) infinitas. Así como $\bar{x}$ es una medida interesante e importante de la ubicación de la muestra, μ es una interesante e importante característica (con frecuencia la más importante) de una

población. Una de nuestras primeras tareas en inferencia estadística será presentar métodos basados en la media muestral para sacar conclusiones respecto una media de población. Por ejemplo, podríamos usar la media muestral $\bar{x} = 16.36$ calculada en el ejemplo 1.14 como una *estimación puntual* (un solo número que es nuestra "mejor" conjetura) de μ = el porcentaje de absorción de agua promedio verdadera para todos los especímenes tratados como se describe.

La media sufre de una deficiencia que, en algunas circunstancias, la convierte en una medida inapropiada del centro: su valor puede ser afectado en gran medida por la presencia incluso de un solo valor extremo (una observación inusualmente grande o pequeña). Por ejemplo, si en una muestra hay nueve empleados que ganan \$50 000 al año y un empleado cuyo salario anual es de \$150 000, el salario promedio de la muestra es \$60 000; en realidad este valor no parece representar los datos. En estas situaciones es conveniente recurrir a una medida menos sensible a los valores de $\bar{x}$ y por el momento propondremos una. Sin embargo, aunque $\bar{x}$ sí tiene este defecto potencial sigue siendo la medida más ampliamente utilizada, básicamente porque existen muchas poblaciones para las cuales un valor atípico extremo en la muestra sería altamente improbable. Cuando se muestrea una población como esa (una población normal o en forma de campana es el ejemplo más importante), la media muestral tenderá a ser estable y bastante representativa de la muestra.

La mediana

La palabra *mediana* es sinónimo de "medio" y la media muestral es en realidad el valor medio una vez que se ordenan las observaciones de la más pequeña a la más grande. Cuando las observaciones están denotadas por $x_1, x_2, \ldots, x_n$, se utilizará el símbolo $\tilde{x}$ para representar la mediana muestral.

DEFINICIÓN

La mediana muestral se obtiene ordenando primero las n observaciones de la más pequeña a la más grande (con cualesquiera valores repetidos incluidos de modo que cada observación muestral aparezca en la lista ordenada). Entonces,

$$\tilde{x} = \begin{cases} \text{El valor medio único si } n \text{ es impar} & = \left(\dfrac{n+1}{2}\right)^{\text{ésimo}} \text{ valor ordenado} \\ \text{El promedio de los dos valores medios si } n \text{ es par} & = \text{promedio de } \left(\dfrac{n}{2}\right)^{\text{ésimo}} \text{ y } \left(\dfrac{n}{2}+1\right)^{\text{ésimo}} \text{ valores ordenados} \end{cases}$$

EJEMPLO 1.15 Quienes no están familiarizados con la música clásica pueden creer que las instrucciones de un compositor para la reproducción de una pieza en particular son tan específicas que la duración no depende en absoluto del los intérpretes. Sin embargo, normalmente hay mucho espacio para la interpretación y para que los directores de orquesta y músicos puedan sacar el máximo provecho de ello. El autor se dirigió al sitio web **ArkivMusic.com** y seleccionó una muestra de 12 grabaciones de la Sinfonía # 9 de Beethoven ("Coral", una obra impresionante y hermosa), y generó las duraciones siguientes (en minutos) clasificadas en orden creciente:

62.3 62.8 63.6 65.2 65.7 66.4 67.4 68.4 68.8 70.8 75.7 79.0

He aquí una gráfica de puntos de los datos:

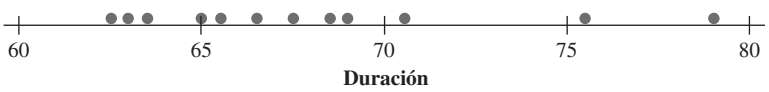


Figura 1.15 Gráfica de puntos de los datos para el ejemplo 1.15

Puesto que $n = 12$ es par, la mediana de la muestra es el promedio de los $n/2 = 6^{o}$ y $(n/2 + 1) = 7^{o}$ valores de la lista ordenada:

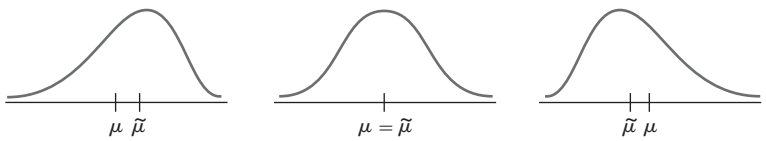
$$\tilde{x} = \frac{66.4 + 67.4}{2} = 66.90$$

Note que si la observación más grande, 79.0, no hubiera aparecido en la muestra, la mediana muestral resultante de las $n = 11$ observaciones restantes habría sido el valor medio 66.4 (el $[n + 1]/2 = 6^{o}$ valor ordenado, es decir, el sexto valor contado desde cualquier extremo de la lista ordenada). La media muestral es $\bar{x} = \sum x_i = 816.1/12 = 68.01$, la cual es poco más de un minuto más grande que la mediana. La media se sale un poco respecto a la mediana, ya que la muestra "se extiende" un poco más en el extremo superior que en el extremo inferior. ■

Los datos del ejemplo 1.15 ilustran una importante propiedad de $\tilde{x}$ en contraste con $\bar{x}$. La mediana muestral es muy insensible a los valores atípicos. Si, por ejemplo, las dos x_i más grandes se incrementan desde 75.7 y 79.0 hasta 85.7 y 89.0, respectivamente, $\tilde{x}$ no se vería afectada. Por tanto, en el tratamiento de valores atípicos, $\bar{x}$ y $\tilde{x}$ no son extremos opuestos de un espectro. Ambas cantidades describen el lugar donde se centran los datos, pero en general no serán iguales porque se enfocan en aspectos diferentes de la muestra.

Análogo a $\tilde{x}$ como valor medio de la muestra existe un valor medio de la población, la **mediana poblacional**, denotada por $\tilde{\mu}$. Tal como con $\bar{x}$ y μ, puede pensarse en utilizar la mediana muestral $\tilde{x}$ para hacer una inferencia sobre $\tilde{\mu}$. En el ejemplo 1.15 se podría utilizar $\tilde{x} = 66.90$ como una estimación de la mediana de tiempo para la población de todas las grabaciones.

La media μ y la mediana $\tilde{\mu}$ poblacionales en general no serán idénticas. Si la distribución de la población es positivamente o negativamente asimétrica, como se ilustra en la figura 1.16, entonces $\mu \neq \tilde{\mu}$. Cuando es este el caso, al hacer inferencias primero se debe decidir cuál de las dos características de la población es de mayor interés y luego proceder como corresponda.



(a) Sesgo negativo (b) Simétrica (c) Sesgo positivo

Figura 1.16 Tres formas diferentes de distribución de la población

Otras medidas de ubicación: cuartiles, percentiles y medias recortadas

La mediana (poblacional o muestral) divide el conjunto de datos en dos partes iguales. Para obtener medidas de ubicación más finas se dividen los datos en más de dos partes. Tentativamente, los cuartiles dividen el conjunto de datos en cuatro partes iguales y las observaciones arriba del tercer cuartil constituyen el cuarto superior del conjunto de

datos, el segundo cuartil es idéntico a la mediana y el primer cuartil separa el cuarto inferior de los tres cuartos superiores. Asimismo, un conjunto de datos (muestra o población) puede ser incluso más finamente dividido mediante percentiles, el 99º percentil separa el más alto 1% del más bajo 99% y así sucesivamente. A menos que el número de observaciones sea un múltiplo de 100, se debe tener cuidado al obtener percentiles. En el capítulo 4 se utilizarán percentiles en conexión con ciertos modelos de poblaciones infinitas.

La media es bastante sensible a un solo valor extremo mientras que la mediana es insensible a muchos valores atípicos. Puesto que el comportamiento extremo de uno u otro tipo podría ser indeseable se consideran brevemente medidas alternativas que no son ni sensibles como $\bar{x}$ ni tan insensibles como $\tilde{x}$. Para motivar estas alternativas observe que $\bar{x}$ y $\tilde{x}$ se encuentran en extremos opuestos de la misma "familia" de medidas. La media es el promedio de todos los datos, mientras que la mediana resulta de eliminar todos excepto uno o dos valores medios y luego promediar. Parafraseando, la media implica recortar 0% de cada extremo de la muestra, mientras que en el caso de la mediana se recorta la cantidad máxima posible de cada extremo. Una **media recortada** es un compromiso entre $\bar{x}$ y $\tilde{x}$. Una media 10% recortada, por ejemplo, se calcularía eliminando el 10% más pequeño y el 10% más grande de la muestra para luego promediar lo que queda.

EJEMPLO 1.16 La producción de Bidri es una artesanía tradicional de India. Las artesanías Bidri (tazones, recipientes, etc.) se funden en una aleación que contiene principalmente zinc y algo de cobre. Considere las siguientes observaciones sobre el contenido de cobre (%) de una muestra de artefactos Bidri tomada del Museo Victoria y Albert de Londres **("Enigmas of Bidri", *Surface Engr*., 2005: 333-339)**, enlistadas en orden creciente:

2.0	2.4	2.5	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	4.8	5.3	10.1

La figura 1.17 es una gráfica de puntos de los datos. Una característica prominente es el valor atípico único en el extremo superior; la distribución está un tanto más dispersa en la región de valores grandes que en el caso de valores pequeños. La media muestral y la mediana son 3.65 y 3.35, respectivamente. Se obtiene una media recortada con un porcentaje de recorte de 100(2/26) = 7.7% al eliminar las dos observaciones más pequeñas y las dos más grandes; esto da $\bar{x}_{tr(7.7)} = 3.42$. El recorte en este caso elimina el valor extremo más grande y, por tanto, acerca la media recortada hacia la mediana.

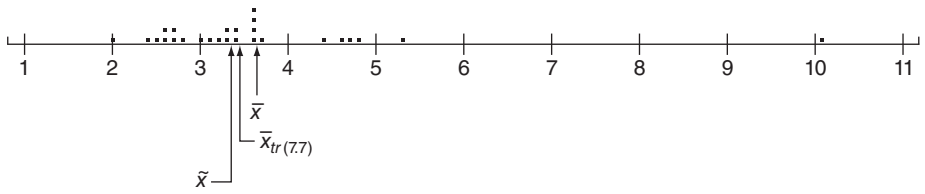


Figura 1.17 Gráfica de puntos del contenido de cobre para el ejemplo 1.16 ■

Una media recortada con un porcentaje de recorte moderado, algo entre 5 y 25%, producirá una medida del centro que no es ni tan sensible a los valores atípicos como la media ni tan insensible como la mediana. Si el porcentaje de recorte deseado es 100a% y na no es un entero, la media recortada debe ser calculada por interpolación. Por ejemplo, considere a = 0.10 para un porcentaje de recorte de 10% y $n = 26$ como en el ejemplo 1.16. Entonces $\bar{x}_{tr(10)}$ sería el promedio ponderado apropiado de la media recortada 7.7% calculada allí y la media recortada 11.5% que resulta de recortar tres observaciones de cada extremo.

Datos categóricos y proporciones muestrales

Cuando los datos son categóricos, una distribución de frecuencia o una distribución de frecuencia relativa proporcionan un resumen tabular efectivo de los datos. Las cantidades resumidas numéricas naturales en esta situación son las frecuencias individuales y las frecuencias relativas. Por ejemplo, si para estudiar la preferencia de marcas se realiza una encuesta de personas que poseen cámara digital y cada persona en la muestra identifica la marca de cámara que usa, entonces se podría contar el número de personas que tienen Canon, Sony, Kodak, etcétera. Considere muestrear una población dividida en dos partes, una consistente en sólo dos categorías (tal como votó o no votó en la última elección, si posee o no una cámara digital, etc.). Si x denota el número en la muestra que cae en la categoría 1, entonces el número en la categoría 2 es $n - x$. La frecuencia relativa o proporción muestral en la categoría 1 es x/n y la *proporción muestral* en la categoría 2 es $1 - x/n$. Designemos con 1 una respuesta que cae en la categoría 1 y con 0 una que cae en la categoría 2. Un tamaño de muestra de $n = 10$ podría dar entonces las respuestas 1, 1, 0, 1, 1, 1, 0, 0, 1, 1. La media muestral de esta muestra numérica es (como la cantidad de números $1s = x = 7$)

$$\frac{x_1 + \cdots + x_n}{n} = \frac{1 + 1 + 0 + \cdots + 1 + 1}{10} = \frac{7}{10} = \frac{x}{n} = \text{proporción muestral}$$

Más generalmente, *centre la atención en una categoría particular y codifique los resultados de modo que se anote un* 1 *para una observación comprendida en la categoría y un* 0 *para una observación no comprendida en la categoría. Entonces la proporción muestral de observaciones comprendidas en la categoría es la media muestral de la secuencia de los* 1 *y los* 0. Por consiguiente se puede utilizar una media muestral para resumir los resultados de una muestra categórica. Estos comentarios también se aplican a situaciones en las cuales las categorías se definen agrupando valores en una muestra o población numérica (p. ej., podría existir interés en saber si las personas han tenido su automóvil actual durante al menos 5 años, en lugar de estudiar el tiempo exacto que se ha poseído algo).

Análogo a la proporción muestral x/n de personas u objetos que caen en una categoría particular, represente con p la proporción de aquellos que se hallan en la población entera que caen en la categoría. Tal como con x/n, p es una cantidad entre 0 y 1 y mientras que x/n es una característica de la muestra, p es una característica de la población. La relación entre las dos es igual a la relación entre $\tilde{x}$ y $\tilde{m}$ y entre $\bar{x}$ y μ. En particular, subsecuentemente se utilizará x/n para hacer inferencias sobre p. Si una muestra de 100 estudiantes de una gran universidad revela que 38 tienen computadoras Macintosh, entonces se podría usar $38/100 = 0.38$ como estimación puntual de la proporción de todos los estudiantes de la universidad que tienen una Mac. O se puede preguntar si esta muestra proporciona fuerte evidencia para concluir que al menos 1/3 de todos los estudiantes son dueños de una Mac. Con k categorías ($k > 2$) se pueden utilizar las k proporciones muestrales para responder preguntas sobre las proporciones de población $p_1, \ldots, p_k$.

EJERCICIOS **Sección 1.3 (33–43)**

33. El **1° de mayo de 2009** ***The Montclarian*** reportó los siguientes aumentos a los precios de venta de una muestra de casas en Alameda, CA, después de las que se vendieron el mes anterior (miles de dólares):

590 815 575 608 350 1285 408 540 555 679

a. Calcule e interprete la media y la mediana muestrales.

b. Suponga que la 6ª observación hubiera sido 985 en lugar de 1285. ¿Cómo cambiarían la media y la mediana?

c. Calcule una media recortada 20% eliminando primero las dos observaciones muestrales más pequeñas y las dos más grandes.

d. Calcule una media recortada 15%.

34. La exposición a productos microbianos, especialmente endotoxina, puede tener un impacto en la vulnerabilidad respecto a enfermedades alérgicas. El artículo **"Dust Sampling Methods for Endotoxin–An Essential, But Underestimated Issue" (*Indoor Air*, 2006: 20-27)** consideró temas asociados con la determinación de la concentración de endotoxina. Los siguientes datos sobre concentración (EU/mg) en polvo asentada de una muestra de hogares urbanos y otra de casas campestres fueron proporcionados por los autores del artículo citado.

U: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0

F: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0 9.2 3.0 2.0 0.3

a. Determine la media muestral de cada muestra. ¿Cómo se comparan?

b. Determine la mediana muestral de cada muestra. ¿Cómo se comparan? ¿Por qué la mediana de la muestra urbana es tan diferente de la media de dicha muestra?

c. Calcule la media recortada de cada muestra eliminando la observación más pequeña y la más grande. ¿Cuáles son los porcentajes de recorte correspondientes? ¿Cómo se comparan los valores de estas medias recortadas con las medias y las medianas correspondientes?

35. El mercurio es un contaminante del ambiente persistente y dispersivo en muchos ecosistemas alrededor del mundo. Cuando se libera como un subproducto industrial a menudo encuentra su camino en los sistemas acuáticos donde puede tener efectos deletéreos sobre diferentes especies acuáticas y en aves. Los datos adjuntos en la concentración de mercurio de la sangre (μg/g) para las hembras adultas cerca de ríos contaminados en Virginia se obtuvieron de un gráfico en el artículo **"Mercury Exposure Effects the Reproductive Success of a Free-Living Terrestrial Songbird, the Carolina Wren" (*The Auk*, 2011: 759-769**; esta es una publicación de la American Ornithologists' Union).

0.20 0.22 0.25 0.30 0.34 0.41 0.55 0.56
1.42 1.70 1.83 2.20 2.25 3.07 3.25

a. Determine los valores de la media y la mediana muestrales y explique por qué son diferentes. [*Sugerencia:* $\sum x_1 = 18.55$.]

b. Determine el valor de la media recortada 10% y compare con la media y la mediana.

c. ¿Cuánto podría aumentar la observación 0.20 sin afectar el valor de la mediana de la muestra?

36. Una muestra de 26 trabajadores de plataforma petrolera marina tomó parte en un ejercicio de escape y se obtuvieron los datos adjuntos de tiempo (s) para completar el escape (**"Oxygen Consumption and Ventilation During Escape from an Offshore Platform", *Ergonomics*, 1997: 281-292**):

389 356 359 363 375 424 325 394 402
373 373 370 364 366 364 325 339 393
392 369 374 359 356 403 334 397

a. Construya una gráfica de tallo y hojas de los datos. ¿Cómo sugiere la gráfica que se comparen la media y la mediana muestrales?

b. Calcule los valores de la media y la mediana muestrales [*Sugerencia:* $\sum x_i = 9638$.]

c. ¿En cuánto se podría incrementar el tiempo más largo, actualmente de 424, sin afectar el valor de la mediana muestral? ¿En cuánto se podría disminuir este valor sin afectar el valor de la mediana muestral?

d. ¿Cuáles son los valores de $\bar{x}$ y $\tilde{x}$ cuando las observaciones se vuelven a expresan en minutos?

37. El artículo **"Snow Cover and Temperature Relationships in North America and Eurasia" (*J. Climate and Applied Meteorology*, 1983: 460-469)** utilizó técnicas estadísticas para relacionar la cobertura de la capa de nieve en cada continente con la temperatura promedio continental. Los datos allí presentados incluyeron las siguientes diez observaciones de la capa de nieve en octubre en Eurasia durante el periodo 1970-1979 (en millones de km^2):

6.5 12.0 14.9 10.0 10.7 7.9 21.9 12.5 14.5 9.2

¿Qué reportaría como valor representativo, o típico, de la cobertura de la capa de nieve en octubre durante este periodo y qué motivaría su elección?

38. Los valores de presión sanguínea a menudo se reportan en los 5 mmHg más cercanos (100, 105, 110, etc.). Suponga que los valores de presión sanguínea reales de nueve individuos seleccionados al azar son

118.6 127.4 138.4 130.0 113.7 122.0 108.3
131.5 133.2

a. ¿Cuál es la mediana de los valores de presión sanguínea reportados?

b. Suponga que la presión sanguínea del segundo individuo es 127.6 en lugar de 127.4 (un pequeño cambio en un solo valor). ¿Cómo afecta esto a la mediana de los valores reportados? ¿Qué dice esto sobre la sensibilidad de la mediana al redondeo o al agrupamiento de los datos?

39. La propagación de grietas provocadas por fatiga en diferentes partes de un avión ha sido el tema de extensos estudios en años recientes. Los datos adjuntos se componen de vidas de propagación (horas de vuelo/10^4) para alcanzar un cierto tamaño de las grietas en los orificios para los sujetadores en aviones militares (**"Statistical Crack Propagation in Fastener Holes Under Spectrum Loading", *J. Aircraft*, 1983: 1028-1032**):

0.736 0.863 0.865 0.913 0.915 0.937 0.983 1.007
1.011 1.064 1.109 1.132 1.140 1.153 1.253 1.394

a. Calcule y compare los valores de la media y la mediana muestrales.

b. ¿En cuánto se podría disminuir la observación muestral más grande sin afectar el valor de la mediana?

40. Calcule la mediana muestral, la media recortada 25%, la media recortada 10% y la media muestral de los datos de duración dados en el ejercicio 27 y compare las medidas.

41. Se eligió una muestra de $n = 10$ automóviles y cada una se sometió a una prueba de choque a 5 mph. Se denotó un auto sin daños visibles con S y uno con daños con F y los resultados fueron los siguientes:

S S F S S S F F S S

a. ¿Cuál es el valor de la proporción muestral de éxitos x/n?

b. Reemplace cada S con 1 y cada F con 0. Después calcule $\bar{x}$ de esta muestra numéricamente codificada. ¿Cómo se compara $\bar{x}$ con x/n?

c. Suponga que se decide incluir 15 autos más en el experimento. ¿Cuántos de estos tendrían que ser S para obtener $x/n = 0.80$ para toda la muestra de 25 carros?

42. **a.** Si se agrega una constante c a cada x_i en una muestra y se obtiene $y_i = x_i + c$, ¿cómo se relacionan la media y la mediana muestrales de las y_i con la media y la mediana muestrales de las x_i? Verifique sus conjeturas.

b. Si cada x_i se multiplica por una constante c y se obtiene $y_i = cx_i$, responda la pregunta del inciso a). Nuevamente verifique sus conjeturas.

43. Un experimento para estudiar la duración (en horas) de un cierto tipo de componente implicaba poner diez componentes en operación y observarlos durante 100 horas. Ocho de ellos fallaron durante dicho periodo y se registraron las duraciones. Se denotan con 100+ las duraciones de los dos componentes que continuaron funcionando después de 100 horas. Las observaciones muestrales resultantes fueron:

48 79 100+ 35 92 86 57 100+ 17 29

¿Cuáles de las medidas centrales discutidas en esta sección pueden ser calculadas y cuáles son los valores de dichas medidas? [*Nota:* se dice que los datos obtenidos con este experimento están "censados a la derecha".]

1.4 Medidas de variabilidad

El reporte de una medida de centro sólo da información parcial sobre un conjunto o distribución de datos. Diferentes muestras o poblaciones pueden tener medidas idénticas de centro y aun así diferir una de otra en otras importantes maneras. La figura 1.18 muestra gráficas de puntos de tres muestras con las mismas media y mediana, aunque el grado de dispersión en torno al centro es diferente para las tres muestras. La primera tiene la cantidad más grande de variabilidad, la tercera tiene la cantidad más pequeña y la segunda es intermedia respecto a las otras dos en este aspecto.

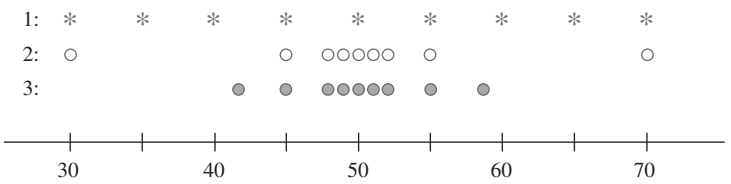


Figura 1.18 Muestras con medidas de centro idénticas pero diferentes cantidades de variabilidad

Medidas de variabilidad de datos muestrales

La medida más simple de variabilidad en una muestra es el **rango**, que es la diferencia entre el valor muestral más grande y el más pequeño. El valor del rango de la muestra 1 en la figura 1.18 es mucho más grande que el de la muestra 3, lo que refleja más variabilidad en la primera muestra que en la tercera. Un defecto del rango, no obstante, es que depende de sólo las dos observaciones más extremas y hace caso omiso de las posiciones de los valores restantes. Las muestras 1 y 2 en la figura 1.18 tienen rangos idénticos aunque, cuando se toman en cuenta las observaciones entre los dos extremos, existe mucho menos variabilidad o dispersión en la segunda muestra que en la primera.

Las medidas principales de variabilidad implican las **desviaciones de la media**, $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$. Es decir, las desviaciones de la media se obtienen restando $\bar{x}$ de cada una de la n observaciones muestrales. Una desviación será positiva si la observación

es más grande que la media (a la derecha de la media sobre el eje de medición) y negativa si la observación es más pequeña que la media. Si todas las desviaciones son pequeñas en magnitud, entonces todas las x_i se aproximan a la media y hay poca variabilidad. Alternativamente, si algunas de las desviaciones son grandes en magnitud, entonces algunas x_i quedan lejos de lo que sugiere una mayor cantidad de variabilidad. Una forma simple de combinar las desviaciones en una sola cantidad es promediarlas. Desafortunadamente, esto es una mala idea:

$$\text{suma de desviaciones} = \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

por lo que la desviación promedio siempre es cero. La verificación utiliza varias reglas estándar de la suma y el hecho de que $\Sigma\bar{x} = \bar{x} + \bar{x} + \cdots + \bar{x} = n\bar{x}$:

$$\sum(x_i - \bar{x}) = \sum x_i - \sum\bar{x} = \sum x_i - n\bar{x} = \sum x_i - n\left(\frac{1}{n}\sum x_i\right) = 0$$

Existen maneras de evitar que las desviaciones negativas y positivas se neutralicen entre sí cuando se combinan. Una posibilidad es trabajar con los valores absolutos de las desviaciones y calcular la desviación absoluta promedio $\Sigma|x_i - \bar{x}|/n$. Debido a que la operación de valor absoluto conduce a un número de dificultades teóricas considere, en cambio, las desviaciones al cuadrado $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$. En vez de utilizar la desviación al cuadrado promedio $\Sigma(x_i - \bar{x})^2/n$, por varias razones se divide la suma de desviaciones al cuadrado entre $n - 1$ en lugar de entre n.

DEFINICIÓN

La **varianza muestral**, denotada por s^2 está dada por

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{S_{xx}}{n - 1}$$

La **desviación estándar muestral**, denotada por s, es la raíz cuadrada (positiva) de la varianza:

$$s = \sqrt{s^2}$$

Observe que s^2 y s son no negativas. La unidad de s es la misma que la de cada una de las x_i. Si, por ejemplo, las observaciones son eficiencias de combustible en millas por galón se podría tener $s = 2.0$ mpg. Una interpretación preliminar de la desviación estándar muestral es que es el tamaño de una desviación típica o representativa de la media muestral dentro de la muestra dada. Por tanto, si $s = 2.0$ mpg algunas x_i en la muestra se aproximan más que 2.0 a $\bar{x}$, en tanto que otras están más alejadas; 2.0 es una desviación representativa (o "estándar") de la eficiencia de combustible media. Si $s = 3.0$ para una segunda muestra de autos de otro tipo, una desviación típica en esta muestra es aproximadamente 1.5 veces la de la primera, una indicación de más variabilidad en la segunda muestra.

EJEMPLO 1.17 El sitio web **www.fueleconomy.gov** contiene gran cantidad de información acerca de las características del combustible de varios vehículos. Además de las calificaciones de millaje de la Environmental Protection Agency (EPA), hay muchos usuarios de vehículos que han informado respecto a sus propios valores de eficiencia de combustible (mpg). Considere la siguiente muestra de $n = 11$ eficiencias para el Ford Focus 2009 equipado con transmisión

automática (para este modelo, la EPA informa de una calificación general de 27 mpg-24 mpg en ciudad y 33 mpg en carretera):

Automóvil	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	
1	27.3	−5.96	35.522	
2	27.9	−5.36	28.730	
3	32.9	−0.36	0.130	
4	35.2	1.94	3.764	
5	44.9	11.64	135.490	
6	39.9	6.64	44.090	
7	30.0	−3.26	10.628	
8	29.7	−3.56	12.674	
9	28.5	−4.76	22.658	
10	32.0	−1.26	1.588	
11	37.6	4.34	18.836	
	$\Sigma x_i = 365.9$	$\Sigma(x_i - \bar{x}) = 0.04$	$\Sigma(x_i - \bar{x})^2 = 314.110$	$\bar{x} = 33.26$

Debido al redondeo la suma de las desviaciones no da exactamente cero. El numerador de s^2 es $S_{xx} = 314.110$, por consiguiente

$$s^2 = \frac{S_{xx}}{n-1} = \frac{314.110}{11-1} = 31.41, \qquad s = 5.60$$

El tamaño de una desviación representativa de la media de la muestra 33.26 es de aproximadamente 5.6 mpg. *Nota:* De las nueve personas que también reportaron hábitos de conducción, sólo tres condujeron más de 80% en la autopista; apostamos a que puede adivinar los automóviles que conducían. Todavía no tenemos idea de por qué los 11 valores registrados exceden la cifra de la EPA, tal vez sólo los conductores con una realmente buena eficiencia de combustible comunican sus resultados. ■

Motivación para s^2

Para explicar el porqué del divisor $n - 1$ en s^2, observe primero que en tanto que s^2 mide la variabilidad muestral, existe una medida de variabilidad en la población llamada **varianza poblacional**. Se utilizará σ^2 (el cuadrado de la letra griega sigma minúscula) para denotar la varianza poblacional y σ para denotar la **desviación estándar poblacional** (la raíz cuadrada de σ^2). El valor de σ se puede interpretar como aproximadamente del tamaño de una desviación típica de μ dentro de toda la población de x valores. Cuando la población es finita y se compone de N valores,

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

la cual es el promedio de todas las desviaciones al cuadrado respecto a la media poblacional (para la población, el divisor es N y no $N - 1$). En los capítulos 3 y 4 se presentan definiciones más generales de σ^2.

Así como se utilizará $\bar{x}$ para hacer inferencias sobre la media poblacional μ, se deberá definir la varianza muestral de modo que pueda ser utilizada para hacer inferencias sobre σ^2. Ahora observe que σ^2 implica desviaciones cuadradas respecto a la media poblacional μ. Si en realidad se conociera el valor de μ, entonces se podría definir la varianza muestral como la desviación al cuadrado promedio de las x_i de la muestra x_i respecto a μ. Sin embargo, el valor de μ casi nunca es conocido, por lo que se debe utilizar el cuadrado de la suma de las

desviaciones respecto a $\bar{x}$. Pero *las x_i tienden a acercarse más a su valor promedio $\bar{x}$ que el promedio poblacional* μ. Para compensar lo anterior se utiliza el divisor $n - 1$ en lugar de n. En otras palabras, si se utiliza un divisor n en la varianza muestral, entonces la cantidad resultante tendería a subestimar σ^2 (en promedio se producen valores demasiado pequeños), mientras que si se divide entre el divisor un poco más pequeño $n - 1$ se corrige esta subestimación.

Es costumbre referirse a s^2 como si estuviera basada en $n - 1$ **grados de libertad** (gl). Esta terminología se deriva del hecho de que aunque s^2 está basada en las n cantidades $x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$, estas suman 0, por lo que al especificar los valores de cualquier $n - 1$ de las cantidades se determina el valor restante. Por ejemplo, si $n = 4$ y $x_1 - \bar{x} = 8$, $x_2 - \bar{x} = 6$ y $x_4 - \bar{x} = -4$, automáticamente $x_3 - \bar{x} = 2$, por lo que sólo tres de los cuatro valores de $x_i - \bar{x}$ son determinados libremente (3 grados de libertad).

Una fórmula para calcular s^2

Es mejor obtener s^2 con software estadístico, o bien utilizar una calculadora que permita ingresar datos en la memoria y luego ver s^2 con un solo golpe de tecla. Si su calculadora no tiene esta capacidad, existe una fórmula alternativa que evita calcular las desviaciones. La fórmula implica a $(\sum x_i)^2$, sumar y luego elevar al cuadrado; y a $\sum x_i^2$, elevar al cuadrado y luego sumar.

Una expresión alternativa para el numerador de s^2 es

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Demostración Debido a que $\bar{x} = \Sigma x_i / n$, $n(\bar{x})^2 = (\Sigma x_i)^2 / n$. Entonces

$$\begin{aligned}\sum (x_i - \bar{x})^2 &= \sum (x_i^2 - 2\bar{x} \cdot x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x} \sum x_i + \sum (\bar{x})^2 \\ &= \sum x_i^2 - 2\bar{x} \cdot n\bar{x} + n(\bar{x})^2 = \sum x_i^2 - n(\bar{x})^2\end{aligned}$$

■

EJEMPLO 1.18 La luxación traumática de rodilla a menudo requiere cirugía para reparar los ligamentos rotos. Una medida de la recuperación es la amplitud de movimiento (medido como el ángulo formado cuando, a partir de la pierna estirada, la rodilla se dobla en la medida de lo posible). Los datos que figuran en el rango de movimiento posquirúrgico aparecen en el artículo **"Reconstruction of the Anterior and Posterior Cruciate Ligaments After Knee Dislocation" (*Amer. J. Sports Med*., 1999: 189-197):**

154 142 137 133 122 126 135 135 108 120 127 134 122

La suma de estas 13 muestras observadas es $\sum x_i = 1695$ y la suma de sus cuadrados es

$$\sum x_i^2 = (154)^2 + (142)^2 + \cdots + (122)^2 = 222.581$$

Por tanto, el numerador de la varianza muestral es

$$S_{xx} = \sum x_i^2 - \left[(\sum x_i)^2\right]/n = 222.581 - (1695)^2/13 = 1579.0769$$

de donde $s^2 = 1579.0769/12 = 131.59$ y $s = 11.47$. ■

Tanto la fórmula de la definición como la fórmula de cálculo para s^2 pueden ser sensibles al redondeo, por lo que en los cálculos intermedios se debe utilizar la mayor precisión decimal que sea posible.

Varias propiedades de s^2 pueden mejorar la comprensión y facilitar el cálculo.

PROPOSICIÓN

Sean $x_1, x_2, \ldots, x_n$ una muestra y c cualquier constante diferente de cero.

1. Si $y_1 = x_1 + c, y_2 = x_2 + c, \ldots, y_n = x_n + c$, entonces $s_y^2 = s_x^2$ y
2. Si $y_1 = cx_1, \ldots, y_n = cx_n$, entonces $s_y^2 = c^2 s_x^2$, $s_y = |c| s_x$

donde s_x^2 es la varianza muestral de las x y s_y^2 es la varianza muestral de las y.

En otras palabras, el resultado 1 dice que si se suma (o se resta) una constante c de cada valor de dato, la varianza no cambia. Esto es intuitivo puesto que la adición o sustracción de c cambia la ubicación del conjunto de datos, pero deja inalteradas las distancias entre los valores de datos. De acuerdo con el resultado 2, la multiplicación de cada x_i por c hace que s^2 sea multiplicada por un factor de c^2. Estas propiedades pueden ser comprobadas al observar en el resultado 1 que $\bar{y} = \bar{x} + c$ y en el resultado 2 que $\bar{y} = c\bar{x}$.

Gráficas de caja

Las gráficas de tallo y hojas y los histogramas transmiten impresiones un tanto generales sobre un conjunto de datos, mientras que un resumen único tal como la media o la desviación estándar se enfoca en sólo un aspecto de los datos. En años recientes se ha utilizado con éxito un resumen gráfico llamado *gráfica de caja* para describir varias de las características más prominentes de un conjunto de datos. Estas características incluyen 1) el centro, 2) la dispersión, 3) el grado y la naturaleza de cualquier alejamiento de la simetría, y 4) la identificación de las observaciones "atípicas" inusualmente alejadas del cuerpo principal de los datos. Puesto que incluso un solo valor extremo puede afectar drásticamente los valores de $\bar{x}$ y s, una gráfica de caja está basada en medidas "resistentes" a la presencia de unos cuantos valores atípicos: la mediana y una medida de variabilidad llamada *distancia entre cuartos*.

DEFINICIÓN

Se ordenan las n observaciones de la más pequeña a la más grande y se separa la mitad más pequeña de la más grande; si n es impar se incluye la mediana en ambas mitades. En tal caso el **cuarto inferior** es la mediana de la mitad más pequeña y **el cuarto superior** es la mediana de la mitad más grande. Una medida de dispersión resistente a los valores atípicos es la distancia entre cuartos f_s, dada por

$$f_s = \text{cuarto superior} - \text{cuarto inferior}$$

En general, la distancia entre cuartos no se ve afectada por las posiciones de las observaciones comprendidas en el 25% más pequeño o el 25% más grande de los datos. Por consiguiente es resistente a valores atípicos. Por tanto, es resistente a valores atípicos. Los cuartos son muy similares a los cuartiles y la cuarta extensión es similar al *rango intercuartil*, la diferencia entre los cuartiles superiores e inferiores. Pero los cuartiles son un poco más enfadosos que los cuartos para calcular a mano, y existen diferentes maneras razonables para calcular los cuartiles (así los valores pueden variar de un programa informático a otro).

La gráfica de caja más simple se basa en el siguiente resumen de cinco números:

x_i más pequeñas cuarto inferior mediana cuarto superior x_i más grandes

Primero, se coloca un rectángulo sobre una escala de medición horizontal; el lado izquierdo del rectángulo está arriba en el cuarto inferior y el derecho en el cuarto superior (por lo que el ancho de la caja $= f_s$). Se coloca un segmento de línea vertical o algún otro símbolo adentro del rectángulo en la ubicación de la mediana; la posición del símbolo de la mediana respecto a los dos lados da información sobre asimetría en el 50% medio de los datos. Por

último, se trazan "bigotes" hacia ambos extremos del rectángulo hacia las observaciones más pequeñas y más grandes. También se puede trazar una gráfica de caja con orientación vertical mediante modificaciones obvias en el proceso de construcción.

EJEMPLO 1.19 Los datos adjuntos se componen de observaciones en el tiempo hasta la falla (miles de horas) para una muestra de turbocompresores de un tipo de motor (de **"The Beta Generalized Weibull Distribution: Properties and Applications",** ***Reliability Engr. and System Safety*****, 2012: 5-15**).

1.6	2.0	2.6	3.0	3.5	3.9	4.5	4.6	4.8	5.0
5.1	5.3	5.4	5.6	5.8	6.0	6.0	6.1	6.3	6.5
6.5	6.7	7.0	7.1	7.3	7.3	7.3	7.7	7.7	7.8
7.9	8.0	8.1	8.3	8.4	8.4	8.5	8.7	8.8	9.0

El resumen de cinco números es el siguiente.

más pequeño:	cuarto inferior:	mediana:	cuarto superior:	más grande:
1.6	5.05	6.5	7.85	9.0

La figura 1.19 muestra la salida de Minitab de una solicitud para describir los datos. Q1 y Q3 son los cuartiles inferiores y superiores, respectivamente, e IQR (rango intercuartil) es la diferencia entre los cuartiles. La media SE es $s/\sqrt{n}$, el "error estándar de la media"; este será importante en nuestro desarrollo subsecuente de varios procedimientos utilizados para hacer inferencias sobre la media de la población m.

```
Variable  Count   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum    IQR
lifetime     40  6.253    0.309  1.956    1.600  5.025   6.500  7.875    9.000  2.850
```

Figura 1.19 Descripción de Minitab de los datos del tiempo de vida del turbocompresor

La figura 1.20 muestra una gráfica de puntos de los datos y una gráfica de caja. Ambas gráficas indican que hay una cantidad razonable de la simetría en el medio 50% de los datos, pero en general los valores estiran más hacia el extremo inferior que hacia el extremo superior, un sesgo negativo. La caja en sí misma es no muy estrecha, lo que indica una buena cantidad de variabilidad en la parte media de los datos, y el bigote inferior es especialmente largo.

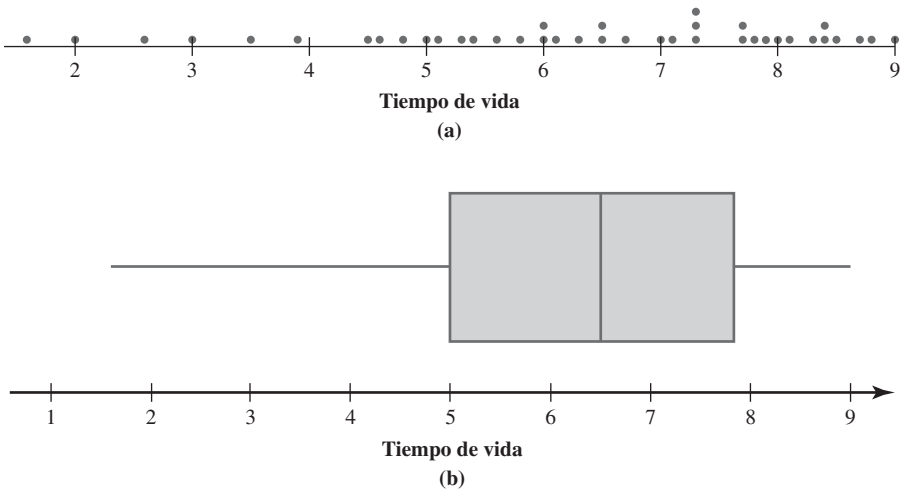


Figura 1.20 (a) Gráfica de puntos y (b) gráfica de caja para los datos del tiempo de vida ■

Gráficas de caja que muestran valores atípicos

Una gráfica de caja puede ser embellecida para indicar explícitamente la presencia de valores atípicos. Muchos procedimientos inferenciales se basan en la suposición de que la distribución de la población es normal (un cierto tipo de curva en forma de campana). Incluso un solo valor apartado extremo que aparezca en la muestra advierte al investigador que tales procedimientos pueden ser no confiables, y la presencia de varios valores atípicos moderados transmite el mismo mensaje.

DEFINICIÓN

Cualquier observación a más de $1.5f_s$ del cuarto más cercano es un valor **atípico**. Un valor atípico es **extremo** si se encuentra a más de $3f_s$ del cuarto más cercano, y **moderado** en caso contrario.

Modifique ahora la construcción previa de una gráfica de caja trazando un bigote que sale de cada extremo de la caja hacia las observaciones más pequeñas y más grandes que *no* son valores atípicos. Cada valor atípico moderado está representado por un círculo cerrado y cada valor atípico extremo, por uno abierto. Algunos programas de computadora estadísticos no distinguen entre valores atípicos moderados y extremos.

EJEMPLO 1.19 La ley Clean Water (agua limpia) y las modificaciones posteriores requieren que todas las aguas en los Estados Unidos alcancen los objetivos de reducción de la contaminación para garantizar que el agua sea "apta para la pesca y para nadar". El artículo **"Spurious Correlation in the USEPA Rating Curve Method for Estimating Pollutant Loads" (*J. of Environ. Engr.*, 2008: 610-618)** ha investigado diferentes técnicas para estimar las cargas contaminantes en las cuencas hidrográficas; los autores "discuten la necesidad imperiosa del uso racional de los métodos estadísticos" para este fin. Entre los datos que se consideran está la siguiente muestra de cargas de NT (nitrógeno total) (kg N/día), a partir de una determinada ubicación en la Bahía de Chesapeake, que aparece aquí en orden creciente.

9.69	13.16	17.09	18.12	23.70	24.07	24.29	26.43
30.75	31.54	35.07	36.99	40.32	42.51	45.64	48.22
49.98	50.06	55.02	57.00	58.41	61.31	64.25	65.24
66.14	67.68	81.40	90.80	92.17	92.42	100.82	101.94
103.61	106.28	106.80	108.69	114.61	120.86	124.54	143.27
143.75	149.64	167.79	182.50	192.55	193.53	271.57	292.61
312.45	352.09	371.47	444.68	460.86	563.92	690.11	826.54
1529.35							

El resumen de las cantidades pertinentes es

$$\tilde{x} = 92.17 \qquad 4°\text{inferior} = 45.64 \qquad 4°\text{superior} = 167.79$$

$$f_s = 122.15 \qquad 1.5f_s = 183.225 \qquad 3f_s = 366.45$$

Restando $1.5f_s$ del 4° inferior da un número negativo y ninguna de las observaciones es negativa, así que no hay valores atípicos en el extremo inferior de los datos. Sin embargo,

$$4° \text{ superior} + 1.5f_s = 351.015 \qquad 4° \text{ superior} + 3f_s = 534.24$$

Por tanto, las cuatro observaciones más grandes: 563.92, 690.11, 826.54 y 1529.35, son valores atípicos extremos; y 352.09, 371.47, 444.68 y 460.86 son valores atípicos moderados.

Los bigotes en la gráfica de caja de la figura 1.22 se extienden hacia afuera de la observación más pequeña, 9.69, en el extremo inferior y 312.45, la observación más grande en el extremo superior que no es un valor apartado. Hay cierta asimetría positiva en la mitad central de los datos (la línea mediana está un poco más cerca del borde izquierdo de la caja que del extremo derecho) y, en general, una gran asimetría positiva.

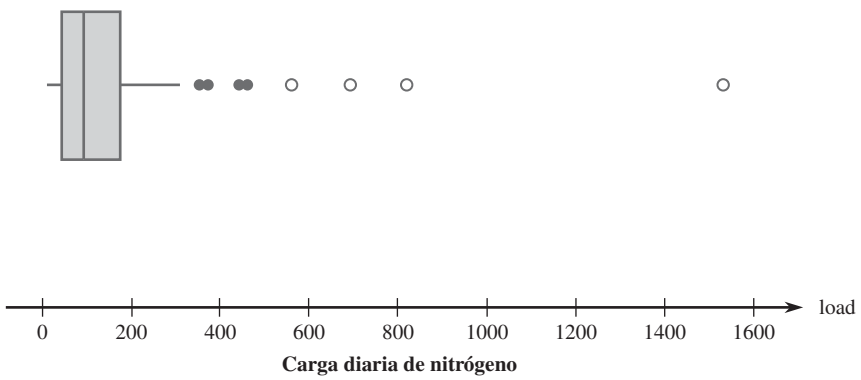


Figura 1.21 Gráfica de caja de los datos de la carga de nitrógeno mostrando los valores atípicos moderados y extremos

Gráficas de caja comparativas

Una gráfica de caja comparativa o lado a lado es una forma muy efectiva de revelar similitudes y diferencias entre dos o más conjuntos de datos compuestos de observaciones de la misma variable, observaciones de eficiencia de consumo de combustible de cuatro tipos distintos de automóviles, rendimientos de cosechas de tres variedades diferentes, etcétera.

EJEMPLO 1.21 Los altos niveles de sodio en los productos alimenticios representan un creciente para la salud. Los datos adjuntos se componen de los valores del contenido de sodio en una porción de cereal para una muestra de cereales fabricado por General Mills, otra muestra fabricada por Kellogg y una tercera muestra producida por Post (consulte el sitio web **http://www.nutritionresource.com/foodcomp2.cfm?id50800** ¡en lugar de visitar el supermercado de su colonia!).

G:	211	408	171	178	359	249	205	203	201	223	234	256	218
K:	143	202	120	229	150	5	207	362	252	275	224		
P:	253	220	212	41	140	215	266	3	214	280			

La figura 1.22 muestra una gráfica de caja comparativa de los datos usando el paquete de software R. El contenido típico de sodio (mediana) es aproximadamente el mismo para las tres empresas. Pero las distribuciones difieren marcadamente en otros aspectos. Los

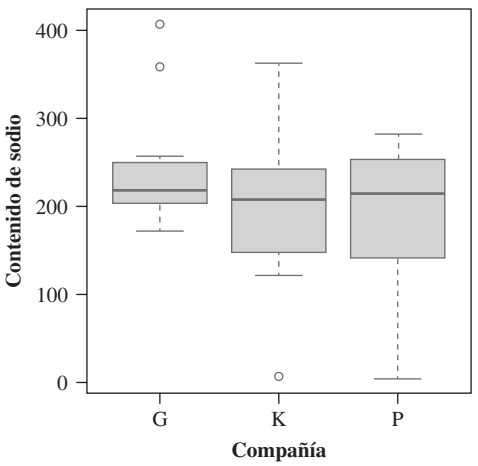


Figura 1.22 Gráfica de caja comparativa de los datos del ejemplo 1.21, obtenida usando R

datos de Kellogg muestran un sesgo positivo considerable tanto en el medio del 50% y, en general, con dos valores atípicos en el extremo superior. Los datos de Kellogg exhiben un sesgo negativo en el medio del 50% y un sesgo positivo en general, con excepción del valor atípico en el extremo inferior (este valor atípico no está identificado por Minitab). Y los datos Post están sesgados negativamente tanto en el medio del 50% y en general con ningún valor atípico. La variabilidad, según lo determinado por la longitud de la caja (aquí el rango intercuartil en lugar de la cuarta extensión) es más pequeña para la marca G y más grande para la marca P, con la marca K intermedia de las otras dos; observando, por otro lado, las desviaciones estándar, s_K y s_P son más o menos lo mismo y ambas mucho más grandes que s_G. ■

EJERCICIOS **Sección 1.4 (44–61)**

44. El polihidroxibutirato (PHB), un polímero semicristalino completamente biodegradable y biocompatible, se obtiene a partir de recursos renovables. Desde una perspectiva de sostenibilidad, el PHB ofrece muchas características atractivas, aunque es más caro de producir que los plásticos estándar. En el artículo **"The Melting Behaviour of Poly(3-Hydroxybutyrate) por DSC. Reproducibility Study" (*Polymer Testing*, 2013: 215–220)** aparecen adjuntos los datos en el punto de fusión (°C) para cada uno de 12 ejemplares del polímero usando un calorímetro de barrido diferencial.

```
180.5  181.7  180.9  181.6  182.6  181.6
181.3  182.1  182.1  180.3  181.7  180.5
```

Calcule lo siguiente:

- **a.** El rango de la muestra.
- **b.** La varianza de la muestra s^2 de la definición. [*Sugerencia*: primero reste 180 de cada observación.]
- **c.** La desviación estándar de la muestra.
- **d.** s^2 utilizando el método directo.

45. Se determinó el valor del módulo de Young (GPa) de placas fundidas compuestas de ciertos sustratos intermetálicos y se obtuvieron las siguientes observaciones muestrales **("Strength and Modulus of a Molybdenum-Coated Ti-25A1-10Nb-3U-1Mo Intermetallic", *J. of Materials Engr. and Performance*, 1997: 46-50):**

116.4 115.9 114.6 115.2 115.8

- **a.** Calcule $\bar{x}$ y las desviaciones de la media.
- **b.** Use las desviaciones calculadas en el inciso a) para obtener la varianza muestral y la desviación estándar muestral.
- **c.** Calcule s^2 utilizando la fórmula computacional para el numerador S_{xx}.
- **d**. Reste 100 de cada observación para obtener una muestra de valores transformados. Ahora calcule la varianza muestral de estos valores transformados y compárela con s^2 de los datos originales.

46. El artículo **"Effects of Short-Term Warming on Low and High Latitude Forest Ant Communities" (*Ecoshpere*, mayo de 2011, artículo 62)** describe un experimento en el que las observaciones de diferentes características fueron hechas usando minicámaras de tres tipos: 1) enfriamiento (marcos de policloruro de vinilo [PVC] cubiertos con tela de cortina), 2) control (sólo marcos de PVC) y 3) calentamiento (marcos de PVC recubiertos de plástico). Uno de los autores del artículo amablemente suministra los datos adjuntos de la diferencia entre las temperaturas del aire y del suelo (°C).

Enfriamiento	Control	Calentamiento
1.59	1.92	2.57
1.43	2.00	2.60
1.88	2.19	1.93
1.26	1.12	1.58
1.91	1.78	2.30
1.86	1.84	0.84
1.90	2.45	2.65
1.57	2.03	0.12
1.79	1.52	2.74
1.72	0.53	2.53
2.41	1.90	2.13
2.34		2.86
0.83		2.31
1.34		1.91
1.76		

- **a.** Compare las medidas del centro para las tres muestras diferentes.
- **b.** Calcule, interprete y compare las desviaciones estándar para las tres muestras diferentes.
- **c.** La cuarta extendida para las tres muestras ¿transmite el mismo mensaje como lo hacen las desviaciones estándar sobre variabilidad relativa?
- **d.** Construya una gráfica de caja comparativa (la cual se incluye en el artículo citado) y comente cualquier característica interesante.

47. Zinfandel es un popular vino tinto varietal producido casi exclusivamente en California. Es bastante controvertido entre los conocedores del vino porque su contenido en alcohol varía bastante de un productor a otro. En mayo de 2013 el autor visitó la página web **klwines.com**, seleccionó al azar

10 zinfandels entre los 325 disponibles y obtuvo los siguientes valores de contenido de alcohol (%):

14.8	14.5	16.1	14.2	15.9
13.7	16.2	14.6	13.8	15.0

a. Calcule e interprete varias medidas del centro.
b. Calcule la varianza muestral utilizando la fórmula definitoria.
c. Calcule la varianza muestral utilizando la fórmula de acceso directo después de restar 13 de cada observación.

48. El ejercicio 34 presentó los siguientes datos sobre concentración de endotoxina en polvo asentada obtenidos con una muestra de casas urbanas y una muestra de casas campestres:

U: 6.0 5.0 11.0 33.0 4.0 5.0 80.0 18.0 35.0 17.0 23.0
C: 4.0 14.0 11.0 9.0 9.0 8.0 4.0 20.0 5.0 8.9 21.0
9.2 3.0 2.0 0.3

a. Determine el valor de la desviación estándar muestral de cada muestra, interprete estos valores y luego contraste la variabilidad en las dos muestras. [*Sugerencia:* $\sum x_i = 237.0$ para la muestra urbana y $= 128.4$ para la muestra campestre; y $\sum x_i^2 = 10\,079$ para la muestra urbana y 1617.94 para la muestra campestre.]
b. Calcule la distancia entre cuartos de cada muestra y compare. ¿La distancia entre cuartos transmite el mismo mensaje sobre la variabilidad que las desviaciones estándar? Explique.
c. Los autores del artículo citado también proporcionan concentraciones de endotoxina en el polvo presente en bolsas captadoras de polvo:

U: 34.0 49.0 13.0 33.0 24.0 24.0 35.0 104.0 34.0 40.0 38.0 1.0
C: 2.0 64.0 6.0 17.0 35.0 11.0 17.0 13.0 5.0 27.0 23.0
28.0 10.0 13.0 0.2

Construya una gráfica de caja comparativa (como se hizo en el artículo citado) y compare y contraste las cuatro muestras.

49. Un estudio de la relación entre edad y varias funciones visuales (tales como agudeza y percepción de profundidad) reportó las siguientes observaciones en el área de la lámina esclerótica (mm^2) de las cabezas del nervio óptico humano (**"Morphometry of Nerve Fiber Bundle Pores in the Optic Nerve Head of the Human", *Experimental Eye Research*, 1988: 559-568):**

2.75 2.62 2.74 3.85 2.34 2.74 3.93 4.21 3.88
4.33 3.46 4.52 2.43 3.65 2.78 3.56 3.01

a. Calcule $\sum x_i$ y $\sum x_i^2$.
b. Use los valores calculados en el inciso a) para calcular la varianza muestral s^2 y luego la desviación estándar muestral s.

50. En 1997 una mujer demandó a un fabricante de teclados de computadora y lo acusó de que sus repetidas lesiones por el esfuerzo eran debido al teclado (*Genessy v. Digital Equipment Corp.*). El jurado le adjudicó $3.5 millones por el dolor y sufrimiento, pero la corte anuló dicha adjudicación por considerarla una compensación irrazonable. Al hacer esta determinación, la corte identificó un grupo "normativo" de 27 casos similares y especificó que una adjudicación razonable estaría dentro de dos desviaciones estándar de la media de las adjudicaciones en los 27 casos. Las 27 adjudicaciones fueron (en el rango de los $1000 dólares) 37, 60, 75, 115, 135, 140, 149, 150, 238, 290, 340, 410, 600, 750, 750, 750, 1050, 1100, 1139, 1150, 1200, 1200, 1250, 1576, 1700, 1825 y 2000, con las cuales $\sum x_i = 20\,179$, $\sum x_i^2 = 24\,657\,511$. ¿Cuál es la cantidad máxima posible que podría ser adjudicada conforme a la regla de dos desviaciones estándar?

51. El artículo **"A Thin-Film Oxygen Uptake Test for the Evaluation of Automotive Crankcase Lubricants" (*Lubric. Engr.*, 1984: 75-83)** reportó los siguientes datos sobre tiempo de inducción de oxidación (min) de varios aceites comerciales:

87 103 130 160 180 195 132 145 211 105 145
153 152 138 87 99 93 119 129

a. Calcule la varianza y la desviación estándar muestrales.
b. Si las observaciones se volvieran a expresar en horas, ¿cuáles serían los valores resultantes de la varianza de la muestra y la desviación estándar muestral? Responda sin llegar a expresarlas nuevamente.

52. Las primeras cuatro desviaciones de la media en una muestra de $n = 5$ tiempos de reacción fueron 0.3, 0.9, 1.0 y 1.3. ¿Cuál es la quinta desviación de la media? Proporcione una muestra para la cual estas sean las cinco desviaciones de la media.

53. Un fondo mutuo es un esquema de inversiones administrado por profesionales que invierten el dinero de muchas personas en una variedad de valores. Los fondos de crecimiento se centran principalmente en el aumento del valor de las inversiones, mientras que los fondos mezclados buscan un equilibrio entre ingresos corrientes y crecimiento. Aquí hay datos sobre la proporción de gastos (gastos en % de los activos, de **www.morningstar.com**) para las muestras de los 20 fondos de gran capitalización equilibrada y 20 fondos de crecimiento de gran capitalización ("gran capitalización" se refiere al tamaño de las empresas en las cuales se invierten los fondos; los tamaños de la población son 825 y 762, respectivamente):

Bal	1.03	1.23	1.10	1.64	1.30
	1.27	1.25	0.78	1.05	0.64
	0.94	2.86	1.05	0.75	0.09
	0.79	1.61	1.26	0.93	0.84
Gr	0.52	1.06	1.26	2.17	1.55
	0.99	1.10	1.07	1.81	2.05
	0.91	0.79	1.39	0.62	1.52
	1.02	1.10	1.78	1.01	1.15

a. Calcule y compare los valores de $\bar{x}$, $\tilde{x}$ y s para los dos tipos de fondos.
b. Construya una gráfica de caja comparativa para los dos tipos de fondos y comente acerca de las características interesantes.

54. El agarre se aplica para producir fuerzas superficiales normales que comprimen el objeto que se quiere aferrar. Los ejemplos incluyen a dos personas dándose la mano, o una enfermera apretando el antebrazo del paciente para detener el sangrado. El artículo **"Investigation of Grip Force, Normal Force, Contact Area, Hand Size, and Handle Size for Cylindrical Handles" (*Human Factors*, 2008: 734-744)** incluye los siguientes datos sobre la fuerza de prensión (N) para una muestra de 42 individuos:

16 18 18 26 33 41 54 56 66 68 87 91 95
98 106 109 111 118 127 127 135 145 147 149 151 168
172 183 189 190 200 210 220 229 230 233 238 244 259
294 329 403

a. Construya un diagrama de tallo y hojas sobre la base de repetir cada valor de tallo dos veces y comente sobre las características interesantes.

b. Determine los valores de los cuartos y distancia entre cuartos.

c. Construya una gráfica de caja basada en el resumen de cinco números y comente sobre sus características.

d. ¿Qué tan grande o pequeña tiene que ser una observación para calificarla como valor atípico? ¿Y cómo valor atípico extremo? ¿Hay valores atípicos?

e. ¿Por cuánto podría disminuir la observación 403, actualmente la más grande, sin afectar f_s?

55. He aquí una gráfica de tallo y hojas de los datos de tiempo de escape introducidos en el ejercicio 36 de este capítulo.

32	55
33	49
34	
35	6699
36	34469
37	03345
38	9
39	2347
40	23
41	
42	4

a. Determine el valor de la distancia entre cuartos.

b. ¿Hay algunos valores atípicos en la muestra? ¿Hay algún valor atípico extremo?

c. Construya una gráfica de caja y comente sobre sus características.

d. ¿En cuánto se podría disminuir la observación más grande, actualmente de 424, sin afectar el valor de la distancia entre cuartos?

56. Los siguientes datos sobre el contenido de alcohol destilado (%) para una muestra de 35 vinos de Oporto fueron extraídos del artículo **"A Method for the Estimation of Alcohol in Fortified Wines Using Hydrometer Baumé and Refractometer Brix" (*Amer. J. Enol. Vitic.*, 2006: 486–490)**. Cada valor es un promedio de dos medidas por duplicado.

16.35 18.85 16.20 17.75 19.58 17.73 22.75 23.78 23.25
19.08 19.62 19.20 20.05 17.85 19.17 19.48 20.00 19.97
17.48 17.15 19.07 19.90 18.68 18.82 19.03 19.45 19.37
19.20 18.00 19.60 19.33 21.22 19.50 15.30 22.25

Utilice los métodos de este capítulo, incluyendo un diagrama de caja que muestre los valores atípicos, para describir y resumir los datos.

57. Se seleccionó una muestra de 20 botellas de vidrio de un tipo particular y se determinó la resistencia de cada botella a la presión interna. Considere la siguiente información parcial sobre la muestra:

mediana = 202.2 cuarto inferior = 196.0
cuarto superior = 216.8

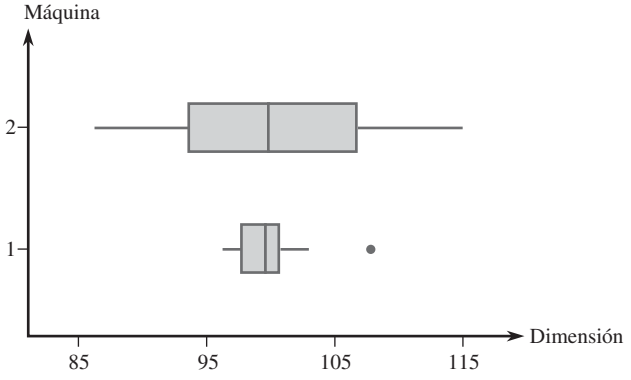
Las tres observaciones más pequeñas	125.8	188.1	193.7
Las tres observaciones más grandes	221.3	230.5	250.2

a. ¿Hay valores atípicos en la muestra? ¿Hay algún valor atípico extremo?

b. Construya una gráfica de caja que muestre los valores atípicos y comente sobre cualquier característica interesante.

58. Una compañía utiliza dos máquinas diferentes para fabricar piezas de cierto tipo. Durante un solo turno se obtuvo una muestra de $n = 20$ piezas producidas por cada máquina y se determinó el valor de una dimensión crítica particular de cada pieza. La gráfica de caja comparativa que aparece en la parte inferior de esta página se construyó con los datos resultantes. Compare y contraste las dos muestras.

Gráfica de caja comparativa para el ejercicio 58



59. Se determinó la concentración de cocaína (mg/L) mediante una muestra de individuos que murieron de delirio excitado (DE) inducido por el consumo de cocaína y mediante una muestra de aquellos que murieron de una sobredosis de cocaína sin delirio excitado; el tiempo de sobrevivencia de las personas en ambos grupos fue a lo sumo de 6 horas. Los datos adjuntos se tomaron de una gráfica de caja comparativa incluida en el artículo **"Fatal Excited Delirium Following Cocaine Use" (*J. of Forensic Sciences*, 1997: 25-31).**

DE	0	0	0	0	0.1	0.1	0.1	0.1	0.2	0.2	0.3	0.3
		0.3	0.4	0.5	0.7	0.8	1.0	1.5	2.7	2.8		
		3.5	4.0	8.9	9.2	11.7	21.0					
DE	0	0	0	0	0	0.1	0.1	0.1	0.1	0.2	0.2	0.2
	0.3	0.3	0.3	0.4	0.5	0.5	0.6	0.8	0.9	1.0		
	1.2	1.4	1.5	1.7	2.0	3.2	3.5	4.1				
	4.3	4.8	5.0	5.6	5.9	6.0	6.4	7.9				
	8.3	8.7	9.1	9.6	9.9	11.0	11.5					
	12.2	12.7	14.0	16.6	17.8							

a. Determine las medianas, los cuartos y las distancias entre cuartos de las dos muestras.

b. ¿Existen algunos valores atípicos en una u otra muestra? ¿Algún valor atípico extremo?

c. Construya una gráfica de caja comparativa y utilícela como base para comparar y contrastar las muestras con DE y sin DE.

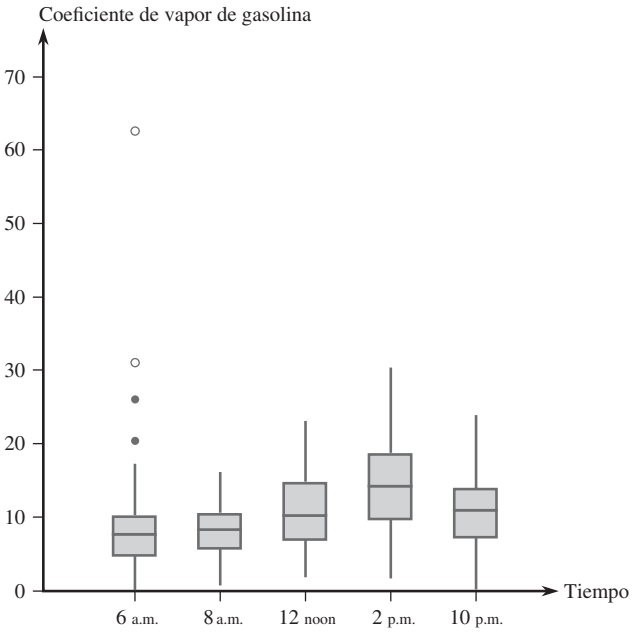
60. Se obtuvieron observaciones de resistencia al estallamiento (lb/pulg^2) mediante pruebas con soldaduras de cierre de tobera, así como con soldaduras para tobera de envases de producción (**"Proper Procedures Are the Key to Welding Radioactive Waste Cannisters",** ***Welding J.*****, agosto de 1997: 61-67).**

Prueba	7200	6100	7300	7300	8000	7400
	7300	7300	8000	6700	8300	
Envase	5250	5625	5900	5900	5700	6050
	5800	6000	5875	6100	5850	6600

Construya una gráfica de caja comparativa y comente sobre las características interesantes (el artículo citado no incluía esta gráfica, pero los autores comentaron haber visto una).

61. La gráfica de caja comparativa adjunta de coeficientes de vapor de gasolina para vehículos en Detroit apareció en el artículo **"Receptor Modeling Approach to VOC Emission Inventory Validation" (*J. of Envir. Engr.*, 1995: 483-490).** Discuta las características interesantes.

Gráfica de caja comparativa para el ejercicio 61



EJERCICIOS SUPLEMENTARIOS (62–79)

62. Considere la siguiente información sobre resistencia a la tensión final (lb/pulg) de una muestra de $n = 4$ probetas de alambre de cobre de zirconio duro (de **"Characterization Methods for Fine Copper Wire",** ***Wire J. Intl.*****, agosto de 1997: 74–80):**

$\bar{x} = 76\,831$ $s = 180$ x_i más pequeña $= 76\,683$

x_i más grande $= 77\,048$

Determine los valores de las dos observaciones muestrales intermedias (¡pero no lo haga mediante conjeturas sucesivas!).

63. Se tomó una muestra de 77 personas que trabajan en una oficina particular y se determinó el nivel de ruido (dBA) experimentado por cada individuo, mediante los siguientes datos (**"Acceptable Noise Levels for Construction Site Offices",** ***Building Serv. Engr. Research and Technology*****, 2009: 87–94).**

55.3	55.3	55.3	55.9	55.9	55.9	55.9	56.1	56.1	56.1	56.1
56.1	56.1	56.8	56.8	57.0	57.0	57.0	57.8	57.8	57.8	57.9
57.9	57.9	58.8	58.8	58.8	59.8	59.8	59.8	62.2	62.2	63.8
63.8	63.8	63.9	63.9	63.9	64.7	64.7	64.7	65.1	65.1	65.1
65.3	65.3	65.3	65.3	67.4	67.4	67.4	67.4	68.7	68.7	68.7
68.7	69.0	70.4	70.4	71.2	71.2	71.2	73.0	73.0	73.1	73.1
74.6	74.6	74.6	74.6	79.3	79.3	79.3	79.3	83.0	83.0	83.0

Use algunos de los métodos estudiados en este capítulo para organizar, describir y resumir estos datos.

64. La corrosión por fricción es un proceso de desgaste que resulta de los movimientos oscilatorios tangenciales de pequeña amplitud en las piezas de una máquina. El artículo **"Grease Effect on Fretting Wear of Mild Steel" (*Industrial Lubrication and***

Tribology, 2008: 67-78) incluye los siguientes datos sobre el desgaste de volumen (10^{-4} mm^3) para los aceites base que tienen cuatro diferentes viscosidades.

Viscocidad			Desgaste			
20.4	58.8	30.8	27.3	29.9	17.7	76.5
30.2	44.5	47.1	48.7	41.6	32.8	18.3
89.4	73.3	57.1	66.0	93.8	133.2	81.1
252.6	30.6	24.2	16.6	38.9	28.7	23.6

a. El *coeficiente de variación muestral* $100s/\bar{x}$ evalúa el grado de variabilidad respecto a la media (específicamente, la desviación estándar como porcentaje de la media). Calcule el coeficiente de variación para la muestra en cada viscosidad. Después, compare los resultados y coméntelos.

b. Construya una gráfica de caja comparativa de los datos y comente las características interesantes.

65. La distribución de frecuencia adjunta de observaciones de resistencia a la fractura (MPa) de barras de cerámica cocidas en un horno particular aparece en el artículo **"Evaluating Tunnel Kiln Performance"** (*Amer. Ceramic Soc. Bull.*, agosto de 1997: 59-63).

Clase	81−<83	83−<85	85−<87	87−<89	89−<91
Frecuencia	6	7	17	30	43
Clase	91−<93	93−<95	95−<97	97−<99	
Frecuencia	28	22	13	3	

a. Construya un histograma basado en frecuencias relativas y comente sobre cualquier característica interesante.

b. ¿Qué proporción de las observaciones de resistencia son al menos de 85? ¿Y menores de 95?

c. Aproximadamente, ¿qué proporción de las observaciones son menores de 90?

66. Una deficiencia del microelemento selenio en la dieta puede impactar negativamente en el crecimiento, la inmunidad, la función muscular y neuromuscular y en la fertilidad. La introducción de suplementos de selenio en vacas lecheras se justifica cuando las pasturas contienen niveles bajos de dicho elemento. Los autores del artículo **"Effects of Short-Term Supplementation with Selenised Yeast on Milk Production and Composition of Lactating Cows"** (*Australian J. of Dairy Tech.*, 2004: 199–203) aportan los siguientes datos sobre la concentración de selenio en la leche (mg/L) obtenidos mediante una muestra de vacas a las que se les administró un suplemento de selenio, y una muestra de control de vacas a las que no se les administró ningún suplemento, tanto al inicio como después de un periodo de 9 días.

Obs	Se inicial	Contenido inicial	Se final	Contenido Final
1	11.4	9.1	138.3	9.3
2	9.6	8.7	104.0	8.8
3	10.1	9.7	96.4	8.8
4	8.5	10.8	89.0	10.1
5	10.3	10.9	88.0	9.6
6	10.6	10.6	103.8	8.6
7	11.8	10.1	147.3	10.4
8	9.8	12.3	97.1	12.4
9	10.9	8.8	172.6	9.3
10	10.3	10.4	146.3	9.5
11	10.2	10.9	99.0	8.4
12	11.4	10.4	122.3	8.7
13	9.2	11.6	103.0	12.5
14	10.6	10.9	117.8	9.1
15	10.8		121.5	
16	8.2		93.0	

a. ¿Parecen similares las concentraciones iniciales de Se en las muestras de suplemento y en las muestras de control? Use varias técnicas de este capítulo para resumir los datos y responder la pregunta.

b. De nuevo use métodos de este capítulo para resumir los datos y luego describa cómo los valores finales de concentración de Se en el grupo de tratamiento difieren de aquellos en el grupo de control.

67. La *estenosis aórtica* se refiere al estrechamiento de la válvula aórtica en el corazón. El artículo **"Correlation Analysis of Stenotic Aortic Valve Flow Patterns Using Phase Contrast MRI"** (*Annals of Biomed. Engr.*, 2005: 878-887) aporta los siguientes datos sobre el diámetro de la raíz aórtica (cm) y el género de una muestra de pacientes con varios grados de estenosis aórtica:

H: 3.7 3.4 3.7 4.0 3.9 3.8 3.4 3.6 3.1 4.0 3.4 3.8 3.5
M: 3.8 2.6 3.2 3.0 4.3 3.5 3.1 3.1 3.2 3.0

a. Compare y contraste los diámetros observados en los dos géneros.

b. Calcule una media recortada 10% de cada una de las dos muestras y compare las demás medidas del centro (para la muestra de hombres se debe utilizar el método de interpolación que se menciona en la sección 1.3).

68. **a.** ¿Con qué valor de c es mínima la cantidad $\sum(x_i - c)^2$? [*Sugerencia:* Saque la derivada respecto a c, iguale a 0 y resuelva.]

b. Utilizando el resultado del inciso a), ¿cuál de las dos cantidades $\sum(x_i - \bar{x})^2$ y $\sum(x_i - \mu)^2$ es más pequeña que la otra (suponiendo que $\bar{x} \neq \mu$)?

69. **a.** Sean a y b constantes y sea $y_i = ax_i + b$ con $i = 1, 2, \ldots, n$. ¿Cuáles son las relaciones entre $\bar{x}$ y $\bar{y}$ y entre s_x^2 y s_y^2?

b. Una muestra de temperaturas para iniciar una cierta reacción química dio un promedio muestral (°C) de 87.3 y una desviación estándar muestral de 1.04. ¿Cuáles son el promedio muestral y la desviación estándar medidos en °F? [*Sugerencia:* $F + \frac{9}{5}C + 32$.]

70. El elevado consumo de energía que ocurre durante el ejercicio continúa incluso después de que termina la actividad. Puesto que las calorías quemadas debido al ejercicio contribuyen a la pérdida de peso además de tener otras consecuencias, es importante entender el proceso. El artículo **"Effect of Weight Training Exercise and Treadmill Exercise on Post-Exercise**

Oxygen Consumption" (*Medicine and Science in Sports and Exercise*, 1998: 518-522) reporta los datos adjuntos tomados de un estudio en el cual se midió el consumo de oxígeno (litros) de forma continua durante 30 minutos de cada uno de 15 sujetos después de un entrenamiento con pesas, así como después de una sesión de ejercicio en una caminadora.

Sujeto	1	2	3	4	5	6	7
Pesas (x)	14.6	14.4	19.5	24.3	16.3	22.1	23.0
Caminadora (y)	11.3	5.3	9.1	15.2	10.1	19.6	20.8
Sujeto	8	9	10	11	12	13	14	15
Pesas (x)	18.7	19.0	17.0	19.1	19.6	23.2	18.5	15.9
Caminadora (y)	10.3	10.3	2.6	16.6	22.4	23.6	12.6	4.4

a. Construya una gráfica de caja comparativa de las observaciones tanto del ejercicio con pesas como en la caminadora y comente sobre lo que ve.

b. Debido a que estos datos se muestran en pares (x, y), con mediciones de x y y de la misma variable en dos condiciones distintas, es natural centrarse en las diferencias que existen en ellos: $d_1 = x_1 - y_1, \ldots, d_n = x_n - y_n$. Construya una gráfica de caja de las diferencias muestrales. ¿Qué sugiere la gráfica?

71. La siguiente es una descripción dada por Minitab de los datos de resistencia del ejercicio 13.

```
Variable   N   Mean Median TrMean  StDev SE Mean
resistencia 153135.39135.40 135.41   4.59    0.37

Variable    Minimum    Maximum        Q1        Q3
resistencia   122.20     147.70    132.95    138.25
```

a. Comente sobre cualquier característica interesante (los cuartiles y los cuartos son virtualmente idénticos en este caso).

b. Construya una gráfica de caja de los datos basada en los cuartiles y comente sobre lo que ve.

72. Los desórdenes y síntomas de ansiedad con frecuencia pueden ser tratados exitosamente con benzodiazepina. Se sabe que los animales expuestos a estrés exhiben una disminución de la ligadura del receptor de benzodiazepina en la corteza frontal. El artículo **"Decreased Benzodiazepine Receptor Binding in Prefrontal Cortex in Combat-Related Posttraumatic Stress Disorder" (*Amer. J. of Psychiatry*, 2000: 1120–1126)** describe el primer estudio de ligadura del receptor de benzodiazepina en individuos que sufren trastorno por estrés postraumático (PTSD, por sus siglas en inglés). Los datos anexos sobre una medición de ligadura al receptor (volumen de distribución ajustado) se tomaron de una gráfica que aparece en el artículo.

PTSD: 10, 20, 25, 28, 31, 35, 37, 38, 38, 39, 39, 42, 46

Saludables: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72

Use varios métodos de este capítulo para describir y resumir los datos.

73. El artículo **"Can We Really Walk Straight?" (*Amer. J. of Physical Anthropology*, 1992: 19-27)** reporta sobre un experimento en el cual a cada uno de 20 hombres saludables se le pidió que caminara en línea recta hacia un punto a 60 m de distancia a velocidad normal. Considere las siguientes observaciones de cadencia (número de pasos por segundo):

0.95 0.85 0.92 0.95 0.93 0.86 1.00 0.92 0.85 0.81
0.78 0.93 0.93 1.05 0.93 1.06 1.06 0.96 0.81 0.96

Use los métodos desarrollados en este capítulo para resumir los datos; incluya una interpretación o discusión en los casos en que sea apropiado. [*Nota*: El autor del artículo utilizó un análisis estadístico un tanto complejo para concluir que las personas no pueden caminar en línea recta para lo cual sugirió varias explicaciones.]

74. La **moda** de un conjunto de datos numéricos es el valor que ocurre con más frecuencia en el conjunto.

a. Determine la moda de los datos de cadencia dados en el ejercicio 73.

b. Para una muestra categórica, ¿cómo definiría la categoría modal?

75. Se seleccionaron especímenes de tres tipos diferentes de cable, se determinó el límite de fatiga (MPa) de cada espécimen y se obtuvieron los datos adjuntos.

Tipo 1	350	350	350	358	370	370	370	371
	371	372	372	384	391	391	392	
Tipo 2	350	354	359	363	365	368	369	371
	373	374	376	380	383	388	392	
Tipo 3	350	361	362	364	364	365	366	371
	377	377	377	379	380	380	392	

a. Construya una gráfica de caja comparativa y comente sobre las similitudes y diferencias.

b. Construya una gráfica de puntos comparativa (una gráfica de puntos de cada muestra con una escala común). Comente sobre las similitudes y diferencias.

c. La gráfica de caja comparativa del inciso a) ¿aporta alguna evaluación informativa de similitudes y diferencias? Explique su razonamiento.

76. Las tres medidas centrales introducidas en este capítulo son la media, la mediana y la media recortada. Dos medidas centrales adicionales que de vez en cuando se utilizan son el *rango medio* el cual es el promedio de las observaciones más pequeñas y más grandes, y el cuarto medio el cual es el promedio de los dos cuartos. ¿Cuáles de estas cinco medidas centrales son resistentes a los efectos de los valores atípicos y cuáles no? Explique su razonamiento.

77. Los autores del artículo **"Predictive Model for Pitting Corrosion in Buried Oil and Gas Pipelines" (*Corrosion* 2009: 332–342)** proporcionan los datos en los cuales basaron sus investigaciones.

a. Considere la muestra siguiente de 61 mediciones de la profundidad máxima, a la cual se corroen los pozos (mm) del tipo de tubería enterrada en suelo de arcilla limo.

0.41	0.41	0.41	0.41	0.43	0.43	0.43	0.48	0.48
0.58	0.79	0.79	0.81	0.81	0.81	0.91	0.94	0.94
1.02	1.04	1.04	1.17	1.17	1.17	1.17	1.17	1.17
1.17	1.19	1.19	1.27	1.40	1.40	1.59	1.59	1.60
1.68	1.91	1.96	1.96	1.96	2.10	2.21	2.31	2.46
2.49	2.57	2.74	3.10	3.18	3.30	3.58	3.58	4.15
4.75	5.33	7.65	7.70	8.13	10.41	13.44		

Construya una gráfica de tallos y hojas en la que los dos valores más grandes se muestran en la última fila HI.

b. Remítase de nuevo el inciso a) y construya un histograma basado en las ocho clases con 0 como el límite inferior de la primera clase y con anchos de clase de 0.5, 0.5, 0.5, 1, 2 y 5, respectivamente.

c. La gráfica de caja comparativa de Minitab que se observa al final de esta columna muestra lotes de la profundidad de los pozos para cuatro tipos diferentes de suelos. Describa sus características importantes.

78. Considere una muestra $x_1, x_2, \ldots, x_n$ y suponga que los valores de $\bar{x}$, s^2 y s han sido calculados.

a. Sea $y_i = x_i - \bar{x}$, con $i = 1, \ldots, n$. ¿Cómo se comparan los valores de s^2 y s de las y_i con los valores correspondientes de las x_i? Explique.

b. Sea $z_i = (x_i - \bar{x})/s$ con $i = 1, \ldots, n$. ¿Cuáles son los valores de la varianza muestral y la desviación estándar muestral de las z_i?

79. Sea que $\bar{x}_n$ y s_n^2 denotan la media y la varianza de la muestra $x_1, x_2, \ldots, x_n$ y $\bar{x}_{n+1}$ y s_{n+1}^2 denotan estas cantidades cuando se agrega una observación adicional a la muestra.

a. Demuestre cómo se puede calcular $\bar{x}_{n+1}$ con x_n y x_{n+1}.

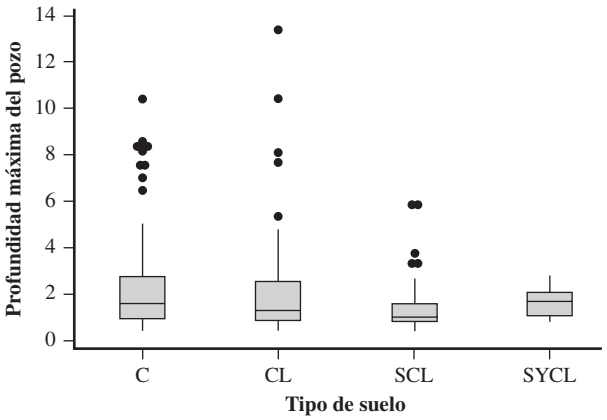
b. Demuestre que

$$ns_{n+1}^2 = (n-1)s_n^2 + \frac{n}{n+1}(x_{n+1} - \bar{x}_n)^2$$

de modo que s_{n+1}^2 pueda ser calculada con x_{n+1}, $\bar{x}_n$ y s_n^2.

c. Suponga que una muestra de 15 torzales de hilo dio por resultado una media muestral del alargamiento del hilo de 12.58 mm y una desviación estándar muestral de 0.512 mm. Un 16[av] de torzal resulta en un valor de alargamiento de 11.8. ¿Cuáles son los valores de la media muestral y la desviación estándar muestral de las 16 observaciones de alargamiento?

Gráfica de caja comparativa para el ejercicio 77



BIBLIOGRAFÍA

Albert, Jim y Maria Rizzo, *R by Example, Springer*, Nueva York, 2012. Una introducción actualizada que trata acerca de la aplicación de técnicas estadísticas más que de los detalles del lenguaje de programación R.

Chambers, John *et al., Graphical Methods for Data Analysis*, Brooks/Cole, Pacific Grove, CA, 1983. Una presentación muy recomendable de diversas metodologías gráficas y pictóricas de estadística.

Cleveland, William, *Visualizing Data*, Hobart Press, Summit, NJ, 1993. Un entretenido recorrido de las técnicas pictóricas.

Freedman, David, Robert Pisani y Roger Purves, *Statistics* (4ª ed.), Norton, Nueva York, 2007. Un estudio excelente y muy "no matématico" razonamiento de la estadística básica y la metodología.

Hoaglin, David, Frederick Mosteller y John Tukey, *Understanding Robust and Exploratory Data Analysis*, Wiley, Nueva York, 1983. Explica por qué y también cómo deben emplearse los métodos exploratorios; es bueno para los detalles de las gráficas de tallo y hojas y para las gráficas de caja.

Moore, David y William Notz, *Statistics: Concepts and Controversies* (7ª ed.), Freeman, San Francisco, 2009. Un libro en edición rústica muy legible y entretenido que contiene un análisis intuitivo de los problemas relacionados con el muestreo y el diseño de experimentos.

Peck, Roxy y Jay Devore, *Statistics: The Exploration and Analysis of Data* (7ª ed.), Cengage Brooks/Cole, Belmont, CA, 2012. Los primeros capítulos brindan un estudio no muy matemático de métodos para describir y resumir datos.

Peck, Roxy *et al.* (eds.), *Statistics: A Guide to the Unknown* (4ª ed.), Cengage Learning, Belmont, CA, 2006. Contiene muchos artículos cortos no matemáticos que describen diversas aplicaciones de la estadística.

Verzani, John, *Using R for Introductory Statistics*, Chapman and Hall/CRC, Boca Ratón, FL, 2005. Una muy buena introducción para el paquete de software R.

Capítulo

2

Probabilidad

INTRODUCCIÓN

El término **probabilidad** se refiere al estudio del azar y la incertidumbre en cualquier situación en la que varios posibles sucesos pueden ocurrir; la disciplina de la probabilidad proporciona métodos para cuantificar las oportunidades y probabilidades asociadas con los varios sucesos. El lenguaje de probabilidad se utiliza constantemente de manera informal tanto en el contexto escrito como en el hablado. Algunos ejemplos incluyen enunciados tales como: "Es probable que el índice Dow-Jones se incremente al final del año", "Existen 50–50 probabilidades de que la persona en posesión de su cargo busque la reelección", "Probablemente se ofrezca al menos una sección del curso el próximo año", "Las probabilidades favorecen la rápida solución de la huelga", "Se espera que se vendan al menos 20 000 boletos para el concierto". En este capítulo se introducen algunos conceptos de probabilidad, se indica cómo pueden ser interpretadas las probabilidades y se demuestra cómo pueden ser aplicadas las reglas de probabilidad para calcular las probabilidades de muchos eventos interesantes. La metodología de la probabilidad permite expresar en lenguaje preciso enunciados informales como los expresados.

El estudio de la probabilidad como una rama de las matemáticas se remonta más de 300 años y su origen se relaciona con cuestiones que implican juegos de azar. Muchos libros se han ocupado exclusivamente de la probabilidad, pero el objetivo en este caso es abarcar sólo la parte de la materia que tiene más aplicación directa en problemas de inferencia estadística.

2.1 Espacios muestrales y eventos

Un **experimento** es cualquier acción o proceso cuyo resultado está sujeto a la incertidumbre. Aunque la palabra *experimento* en general sugiere una situación de prueba cuidadosamente controlada en un laboratorio, aquí se usa en un sentido mucho más amplio. Por tanto, experimentos que pueden ser interesantes incluyen lanzar al aire una moneda una o varias veces, seleccionar una carta o más de un mazo, pesar una hogaza de pan, medir el tiempo del recorrido entre la casa y el trabajo en una mañana particular, obtener tipos de sangre de un grupo de individuos, o medir las resistencias a la compresión de diferentes vigas de acero.

El espacio muestral de un experimento

DEFINICIÓN

El **espacio muestral** de un experimento, denotado por $\mathcal{S}$, es el conjunto de todos los posibles resultados de dicho experimento.

EJEMPLO 2.1 El experimento más simple en el que aplica la probabilidad es aquel con dos posibles resultados. Tal experimento consiste en examinar una soldadura para saber si está defectuosa. El espacio muestral de este experimento se abrevia como $\mathcal{S} = \{N, D\}$, donde N representa no defectuosa, D representa defectuosa y los paréntesis se utilizan para encerrar los elementos de un conjunto. Otro experimento como este implicaría lanzar al aire una tachuela y observar si cae punta arriba (U) o punta abajo (D), con espacio muestral $\mathcal{S} = \{U, D\}$; otro más consistiría en observar el sexo del siguiente niño nacido en el hospital, con $\mathcal{S} = \{M, F\}$. ■

EJEMPLO 2.2 Si se examinan tres soldaduras en secuencia y se anota el resultado de cada examen, entonces un resultado del experimento es cualquier secuencia de letras N y D de longitud 3, por tanto

$$\mathcal{S} = \{NNN, NND, NDN, NDD, DNN, DND, DDN, DDD\}$$

Si se hubiera lanzado una tachuela tres veces, el espacio muestral se obtendría reemplazando N por U en la expresión anterior para $\mathcal{S}$, y con un cambio de notación similar se obtendría el espacio muestral para el experimento en el cual se observan los sexos de tres niños recién nacidos. ■

EJEMPLO 2.3 Dos gasolineras están localizadas en cierta intersección. Cada una dispone de seis bombas de gasolina. Considere el experimento en el cual se determina el número de bombas en uso a una hora particular del día en cada una de las gasolineras. Un resultado experimental especifica cuántas bombas están en uso en la primera gasolinera y cuántas están en uso en la segunda. Un posible resultado es (2, 2), otro es (4, 1) y otro más es (1, 4). Los 49 resultados en $\mathcal{S}$ se muestran en la tabla adjunta. El espacio muestral del experimento en el cual un dado de seis lados es lanzado dos veces se obtiene eliminando la fila 0 y la columna 0 de la tabla, lo cual da 36 resultados.

		Segunda estación						
		0	**1**	**2**	**3**	**4**	**5**	**6**
	0	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
	1	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 0)	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
Primera estación	**3**	(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 0)	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 0)	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

■

EJEMPLO 2.4 Un porcentaje bastante grande de programas C++ escritos en una empresa particular, se compilan en la primera ejecución, pero algunos otros no (un *compilador* es un programa que traduce el código fuente, en este caso programas C++, en lenguaje de máquina para que los programas puedan ser ejecutados). Supongamos que un experimento consiste en seleccionar y compilar programas C++ en este lugar uno por uno hasta encontrar uno que se compile en la primera ejecución. Se denota mediante S (éxito) un programa que se compila en la primera ejecución y uno que no lo hace se denota mediante F (falla). Aunque tal vez no sea muy probable, un posible resultado de este experimento es que los primeros 5 (o 10 o 20 o …) sean F y el siguiente sea S. Es decir, para cualquier entero positivo n, es posible que se tenga que examinar n programas antes de ver la primera S. El espacio muestral es $\mathcal{S} = \{S, FS, FFS, FFFS, \ldots\}$, el cual contiene un número infinito de posibles resultados. La misma forma abreviada del espacio muestral es apropiada para un experimento en el cual, a partir de una hora específica, se anota el sexo de cada recién nacido hasta que nazca un varón. ■

Eventos

En el estudio de la probabilidad, interesan no sólo los resultados individuales de $\mathcal{S}$ sino también varias recopilaciones de los resultados de $\mathcal{S}$.

DEFINICIÓN

> Un **evento** es cualquier recopilación (subconjunto) de resultados contenidos en el espacio muestral $\mathcal{S}$. Un evento es **simple** si consiste en exactamente un resultado y **compuesto** si consiste en más de un resultado.

Cuando se realiza un experimento se dice que ocurre un evento particular A si el resultado experimental resultante está contenido en A. En general, ocurrirá exactamente un evento simple, pero muchos eventos compuestos ocurrirán al mismo tiempo.

EJEMPLO 2.5 Considere un experimento en el cual cada tres vehículos que toman la salida de una autopista particular viran ya sea a la izquierda (L) o a la derecha (R) al final de la rampa de salida. Los ocho posibles resultados que constituyen el espacio muestral son LLL, RLL, LRL, LLR, LRR, RLR, RRL y RRR. Así pues existen ocho eventos simples, entre los cuales están $E_1 = \{LLL\}$, $E_5 = \{LRR\}$. Algunos eventos compuestos incluyen

$$A = \{RLL, LRL, LLR\} = \text{el evento en que sólo uno de los tres vehículos vira a la derecha}$$

$$B = \{LLL, RLL, LRL, LLR\} = \text{el evento en que a lo mas uno de los vehículos vira a la derecha}$$

$$C = \{LLL, RRR\} = \text{el evento en que los tres vehículos viran en la misma dirección}$$

Suponga que cuando se realiza el experimento, el resultado es LLL. Entonces ha ocurrido el evento simple E_1 y, por tanto, también comprende los eventos B y C (pero no A). ■

EJEMPLO 2.6 (Continuación del ejemplo 2.3) Cuando se observa el número de bombas en uso en cada una de dos gasolineras de seis bombas, existen 49 posibles resultados, por lo que existen 49 eventos simples: $E_1 = \{(0, 0)\}$, $E_2 = \{(0, 1)\}$, …, $E_{49} = \{(6, 6)\}$. Ejemplos de eventos compuestos son

$A = \{(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$ = el evento en que el número de bombas en uso es el mismo en ambas gasolineras

$B = \{(0, 4), (1, 3), (2, 2), (3, 1), (4, 0)\}$ = el evento en que el número total de bombas en uso es cuatro

$C = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ = el evento en que a lo sumo una bomba está en uso en cada gasolinera. ■

EJEMPLO 2.7 (Continuación del ejemplo 2.4)

El espacio muestral del experimento del programa de compilación contiene un número infinito de resultados, por lo que existe un número infinito de eventos simples. Los eventos compuestos incluyen

$A = \{S, FS, FFS\}$ = el evento en que cuando mucho se examinan tres programas

$E = \{FS, FFFS, FFFFFS, \ldots\}$ = el evento en que se examina un número par de programas ■

Algunas relaciones de la teoría de conjuntos

Un evento es simplemente un conjunto, así que las relaciones y los resultados de la teoría elemental de conjuntos pueden ser utilizados para estudiar eventos. Se utilizarán las siguientes operaciones para crear eventos nuevos a partir de los ya dados.

DEFINICIÓN

1. El **complemento** de un evento A, denotado por A', es el conjunto de todos los resultados en $\mathcal{S}$ que no están contenidos en A.
2. La **unión** de dos eventos A y B, denotados por $A \cup B$ y leídos "A *o* B", es el evento que consiste en todos los resultados que están *en A o en B o en ambos eventos* (de tal suerte que la unión incluya resultados en los que ocurren tanto A como B, así también resultados donde ocurre exactamente uno), es decir, todos los resultados en al menos uno de los eventos.
3. La **intersección** de dos eventos A y B, denotada por $A \cap B$ y leída "A *y* B", es el evento que consiste en todos los resultados que están *tanto en A como en B*.

EJEMPLO 2.8 (Continuación del ejemplo 2.3)

Para el experimento en el cual se observa el número de bombas en uso en una sola gasolinera de seis bombas, sea $A = \{0, 1, 2, 3, 4\}$, $B = \{3, 4, 5, 6\}$ y $C = \{1, 3, 5\}$. Por tanto

$$A' = \{5, 6\}, \quad A \cup B = \{0, 1, 2, 3, 4, 5, 6\} = \mathcal{S}, \quad A \cup C = \{0, 1, 2, 3, 4, 5\},$$

$$A \cap B = \{3, 4\}, \quad A \cap C = \{1, 3\}, (A \cap C)' = \{0, 2, 4, 5, 6\}$$ ■

EJEMPLO 2.9 (Continuación del ejemplo 2.4)

En el experimento de compilación de programas defina A, B y C como

$$A = \{S, FS, FFS\}, \quad B = \{S, FFS, FFFFS\}, \quad C = \{FS, FFFS, FFFFFS, \ldots\}$$

Por tanto

$$A' = \{FFFS, FFFFS, FFFFFS, \ldots\}, \quad C' = \{S, FFS, FFFFS, \ldots\}$$

$$A \cup B = \{S, FS, FFS, FFFFS\}, \quad A \cap B = \{S, FFS\}$$ ■

En ocasiones A y B no tienen resultados en común, por lo que la intersección de A y B no contiene resultados.

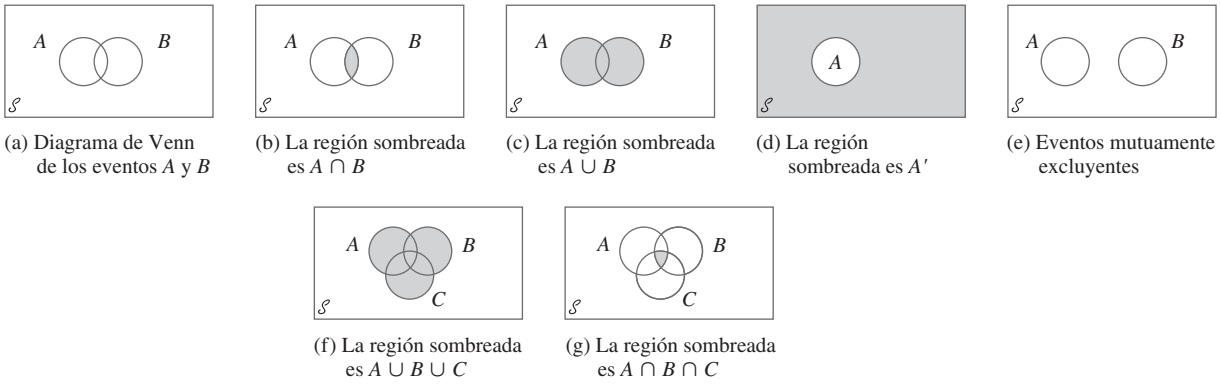
DEFINICIÓN

Sea que $\emptyset$ denote el *evento nulo* (el evento sin resultados). Cuando $A \cap B = \emptyset$, se dice que A y B son eventos **mutuamente excluyentes** o **disjuntos.**

EJEMPLO 2.10 En una pequeña ciudad hay tres distribuidores de automóviles: un distribuidor GM que vende Chevrolet y Buick, un distribuidor Ford que vende Ford y Lincoln, y un distribuidor Toyota. Si un experimento consiste en observar la marca del siguiente automóvil vendido, entonces los eventos {Chevrolet, Buick} y {Ford, Lincoln} son mutuamente excluyentes porque el siguiente automóvil vendido no puede ser a la vez un producto GM y un producto Ford (¡a menos que las empresas se fusionen!). ■

Las operaciones de unión e intersección pueden ser ampliadas a más de dos eventos. Para tres eventos cualesquiera A, B y C, el evento $A \cup B \cup C$ es el conjunto de resultados contenidos en al menos uno de los tres eventos, mientras que $A \cap B \cap C$ es el conjunto de resultados contenidos en los tres eventos. Se dice que los eventos dados A_1, A_2, A_3, son mutuamente excluyentes (disjuntos por pares) si ninguno de dos eventos tiene resultados en común.

Con diagramas de Venn se obtiene una representación pictórica de eventos y manipulaciones con eventos. Para construir un diagrama de Venn se traza un rectángulo cuyo interior representará el espacio muestral $\mathcal{S}$. En tal caso cualquier evento A se representa como el interior de una curva cerrada (a menudo un círculo) contenido en $\mathcal{S}$. La figura 2.1 muestra ejemplos de diagramas de Venn.



(a) Diagrama de Venn de los eventos A y B

(b) La región sombreada es $A \cap B$

(c) La región sombreada es $A \cup B$

(d) La región sombreada es A'

(e) Eventos mutuamente excluyentes

(f) La región sombreada es $A \cup B \cup C$

(g) La región sombreada es $A \cap B \cap C$

Figura 2.1 Diagramas de Venn

EJERCICIOS **Sección 2.1 (1–10)**

1. Cuatro universidades, 1, 2, 3 y 4, están participando en un torneo de basquetbol. En la primera ronda, 1 jugará con 2 y 3 jugará con 4. Acto seguido los dos ganadores jugarán por el campeonato y los dos perdedores también jugarán. Un posible resultado puede ser denotado por 1324 (1 derrota a 2 y 3 derrota a 4 en los juegos de la primera ronda, y luego 1 derrota a 3 y 2 derrota a 4).
   a. Enumere todos los resultados en $\mathcal{S}$.
   b. Sea que A denote el evento en que 1 gana el torneo. Enumere los resultados en A.
   c. Sea que B denote el evento en que 2 gana el juego de campeonato. Enumere los resultados en B.
   d. ¿Cuáles son los resultados en $A \cup B$ y en $A \cap B$? ¿Cuáles son los resultados en A'?

2. Suponga que un vehículo que toma una salida particular de una autopista puede virar a la derecha (R), virar a la izquierda (L) o continuar de frente (S). Observe la dirección de cada uno de tres vehículos sucesivamente.

a. Elabore una lista con todos los resultados en el evento A en que los tres vehículos circulan en la misma dirección.

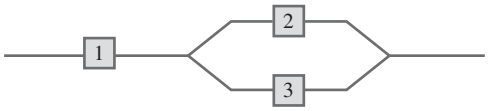
b. Elabore una lista de todos los resultados en el evento B en que los tres vehículos toman direcciones diferentes.

c. Elabore una lista de todos los resultados en el evento C en que sólo dos de los tres vehículos viran a la derecha.

d. Elabore una lista de todos los resultados en el evento D en que dos vehículos circulan en la misma dirección.

e. Enumere los resultados en D', $C \cup D$ y $C \cap D$.

3. Tres componentes están conectados para formar un sistema como se muestra en el diagrama adjunto. Como los componentes del subsistema 2–3 están conectados en paralelo, dicho subsistema funcionará si al menos uno de los dos componentes individuales funciona. Para que todo el sistema funcione, el componente 1 debe hacerlo y, por tanto, el subsistema 2–3 también.



El experimento consiste en determinar la condición de cada componente [S (éxito) para un componente que funciona y F (falla) para un componente que no funciona].

a. ¿Qué resultados están contenidos en el evento A en el que exactamente dos de los tres componentes funcionan?

b. ¿Qué resultados están contenidos en el evento B en el cual al menos dos de los componentes funcionan?

c. ¿Qué resultados están contenidos en el evento C en el que el sistema funciona?

d. Ponga en lista los resultados en C', $A \cup C$, $A \cap C$, $B \cup C$ y $B \cap C$.

4. Cada hipoteca de una muestra de cuatro hipotecas residenciales está clasificada como tasa fija (F) o tasa variable (V).

a. ¿Cuáles son los 16 resultados en $\mathcal{S}$?

b. ¿Qué resultados están en el evento en el que exactamente tres de las hipotecas seleccionadas son de tasa fija?

c. ¿Qué resultados están en el evento en el que las cuatro hipotecas son del mismo tipo?

d. ¿Qué resultados están en el evento en el que a lo más una de las cuatro hipotecas es de tasa variable?

e. ¿Cuál es la unión de eventos en los incisos c) y d), y cuál es la intersección de estos dos eventos?

f. ¿Cuáles son la unión y la intersección de los dos eventos en los incisos b) y c)?

5. Una familia compuesta de tres personas, A, B y C, acude a una clínica médica que siempre tiene disponible un doctor para cada una de las estaciones 1, 2 y 3. Durante cierta semana, cada miembro de la familia visita la clínica una vez y es asignado al azar a una estación. El experimento consiste en registrar la estación para cada miembro. Un resultado es (1, 2, 1) para A a la estación 1, B a la estación 2 y C a la estación 1.

a. Elabore una lista de los 27 resultados en el espacio muestral.

b. Elabore una lista de todos los resultados en el evento en que los tres miembros van a la misma estación.

c. Haga una lista de los resultados en el evento en el que todos los miembros van a diferentes estaciones.

d. Elabore una lista de los resultados en el evento en el que ninguno va a la estación 2.

6. La biblioteca de una universidad dispone de cinco ejemplares de un cierto texto en reserva. Dos ejemplares (1 y 2) son primeras impresiones y los otros tres (3, 4 y 5) son segundas impresiones. Un estudiante examina estos libros en orden aleatorio y se detiene sólo cuando una segunda impresión ha sido seleccionada. Un posible resultado es 5 y otro 213.

a. Ponga en lista los resultados en $\mathcal{S}$.

b. Sea A el evento en el que exactamente un libro debe ser examinado. ¿Qué resultados están en A?

c. Sea B el evento en el que el libro 5 es seleccionado. ¿Qué resultados están en B?

d. Sea C el evento en el que el libro 1 no es examinado. ¿Qué resultados están en C?

7. Un departamento académico acaba de votar en secreto para elegir al jefe del mismo. La urna contiene cuatro boletas con votos para el candidato A y tres con votos para el candidato B. Suponga que estas boletas se sacan de la urna una por una.

a. Haga una lista con todos los posibles resultados.

b. Suponga que mantiene un conteo continuo de las boletas que se retiran de la urna. ¿Para cuáles resultados A se mantiene adelante de B durante todo el conteo?

8. Actualmente una firma constructora de ingeniería trabaja en plantas eléctricas en tres sitios diferentes. Sea que A_i denote el evento en que la planta localizada en el sitio i se completa alrededor de la fecha contratada. Use las operaciones de unión, intersección y complementación para describir cada uno de los siguientes eventos en función de A_1, A_2 y A_3, trace un diagrama de Venn y sombree la región que corresponde a cada uno.

a. Al menos una planta se completa alrededor de la fecha contratada.

b. Todas las plantas se completan alrededor de la fecha contratada.

c. Sólo la planta localizada en el sitio 1 se completa alrededor de la fecha contratada.

d. Exactamente una planta se completa alrededor de la fecha contratada.

e. La planta localizada en el sitio 1 o las otras dos plantas se completan alrededor de la fecha contratada.

9. Use diagramas de Venn para verificar las dos siguientes relaciones para los eventos A y B (estas se conocen como leyes de De Morgan):

a. $(A \cup B)' = A' \cap B'$

b. $(A \cap B)' = A' \cup B'$

[*Sugerencia:* Para cada inciso dibuje un diagrama que corresponda al lado derecho y otro al izquierdo.]

10. **a.** En el ejemplo 2.10 identifique tres eventos que sean mutuamente excluyentes.

b. Suponga que no hay resultado común a los tres eventos A, B y C. ¿Son estos tres eventos mutuamente excluyentes necesariamente? Si su respuesta es afirmativa, explique por qué; si su respuesta es no, dé un contraejemplo valiéndose del experimento del ejemplo 2.10.

2.2 Axiomas, interpretaciones y propiedades de la probabilidad

Dados un experimento y un espacio muestral $\mathcal{S}$, *el objetivo de la probabilidad es asignar a cada evento A un número P(A), llamado la probabilidad del evento A, que dará una medida precisa de la oportunidad de que A ocurra*. Para garantizar que las asignaciones serán consistentes con las nociones intuitivas de la probabilidad, todas las asignaciones deberán satisfacer los siguientes axiomas (propiedades básicas) de probabilidad.

AXIOMA 1 Para cualquier evento A, $P(A) \geq 0$.

AXIOMA 2 $P(\mathcal{S}) = 1$

AXIOMA 3 Si $A_1, A_2, A_3, \ldots$ es un conjunto de eventos disjuntos, entonces

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

Se podría preguntar por qué el tercer axioma no contiene ninguna referencia a un conjunto *finito* de eventos disjuntos. Es porque la propiedad correspondiente para un conjunto finito puede ser derivada de los tres axiomas. Se pretende que la lista de axiomas sea tan corta como sea posible y que no contenga ninguna propiedad que pueda deducirse a partir de las demás que aparecen en la lista. El axioma 1 refleja la noción intuitiva de que la probabilidad de que A ocurra sea no negativa. El espacio muestral es por definición el evento que debe ocurrir cuando se realiza el experimento ($\mathcal{S}$ contiene todos los posibles resultados), así que el axioma 2 dice que la máxima probabilidad posible de 1 está asignada a $\mathcal{S}$. El tercer axioma formaliza la idea que si se desea la probabilidad de que al menos uno de varios eventos ocurra y dado que dos eventos no pueden ocurrir al mismo tiempo, la probabilidad de que al menos ocurra uno es la suma de las probabilidades de los eventos individuales.

PROPOSICIÓN $P(\emptyset) = 0$ donde $\emptyset$ es el evento nulo (el evento que no contiene resultados en absoluto). Esto a su vez implica que la propiedad contenida en el axioma 3 es válida para un conjunto *finito* de eventos disjuntos.

Demostración Primero considere el conjunto infinito $A_1 = \emptyset, A_2 = \emptyset, A_3 = \emptyset, \ldots$. Puesto que $\emptyset \cap \emptyset = \emptyset$, los eventos en este conjunto están disjuntos y $\cup A_i = \emptyset$. El tercer axioma da entonces

$$P(\emptyset) = \Sigma P(\emptyset)$$

Esto puede suceder sólo si $P(\emptyset) = 0$.

Ahora suponga que $A_1, A_2, \ldots, A_k$ son eventos disjuntos y anexe a estos el conjunto infinito $A_{k+1} = \emptyset, A_{k+2} = \emptyset, A_{k+3} = \emptyset, \ldots$. Si de nuevo se invoca el tercer axioma,

$$P\left(\bigcup_{i=1}^{k} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{k} P(A_i)$$

como se deseaba. ■

EJEMPLO 2.11 Considere lanzar una tachuela al aire. Cuando se detenga en el suelo, la punta estará hacia arriba (resultado U) o hacia abajo (resultado D). El espacio muestral de este evento es, por consiguiente, $\mathcal{S} = \{U, D\}$. Los axiomas especifican $P(\mathcal{S}) = 1$, por lo que la asignación de probabilidad se completará determinando $P(U)$ y $P(D)$. Debido a que U y D están disjuntos y su unión es $\mathcal{S}$, la siguiente proposición implica que

$$1 = P(\mathcal{S}) = P(U) + P(D)$$

Se deduce que $P(D) = 1 - P(U)$. Una posible asignación de probabilidades es $P(U) = 0.5$, $P(D) = 0.5$, mientras que otra posible asignación es $P(U) = 0.75$, $P(D) = 0.25$. De hecho, si p representa cualquier número fijo entre 0 y 1, $P(U) = p$, $P(D) = 1 - p$ es una asignación compatible con los axiomas. ■

EJEMPLO 2.12 Considere probar las baterías que salen de la línea de ensamble una por una hasta que se encuentre una con el voltaje dentro de los límites prescritos. Los eventos simples son $E_1 = \{S\}$, $E_2 = \{FS\}$, $E_3 = \{FFS\}$, $E_4 = \{FFFS\}$,…. Suponga que la probabilidad de que cualquier batería resulte satisfactoria es de 0.99. Entonces se puede demostrar que se trata de una asignación de probabilidades a los eventos simples que satisface los axiomas. En particular, puesto que los E_i son disjuntos y $\mathcal{S} = E_1 \cup E_2 \cup E_3 \cup \dots$, debe ser el caso de que

$$\begin{aligned} 1 = P(\mathcal{S}) &= P(E_1) + P(E_2) + P(E_3) + \cdots \\ &= 0.99[1 + 0.01 + (0.01)^2 + (0.01)^3 + \cdots] \end{aligned}$$

Aquí se utilizó la fórmula para la suma de una serie geométrica:

$$a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1 - \mathrm{r}}$$

Sin embargo, otra asignación de probabilidad legítima (de acuerdo con los axiomas) del mismo tipo "geométrico" se obtiene reemplazando 0.99 por cualquier otro número p entre 0 y 1 (y 0.01 por $1 - p$). ■

Interpretación de probabilidad

Los ejemplos 2.11 y 2.12 muestran que los axiomas no determinan por completo una asignación de probabilidades de eventos. Los axiomas sirven sólo para excluir las asignaciones incompatibles con las nociones intuitivas de probabilidad. En el experimento del ejemplo 2.11 de lanzar al aire tachuelas se sugirieron dos asignaciones particulares. La asignación apropiada o correcta depende de la naturaleza de la tachuela y también de la interpretación de probabilidad. La interpretación que más frecuentemente se utiliza y que es más fácil de entender se basa en el concepto de frecuencias relativas.

Considere un experimento que pueda ser realizado repetidamente de una manera idéntica e independiente, y sea A un evento que consiste en un conjunto fijo de resultados del experimento. Ejemplos simples de experimentos repetibles incluyen el lanzamiento al aire de tachuelas y dados que ya se mencionó. Si el experimento se realiza n veces, en algunas de las réplicas ocurrirá el evento A (el resultado estará en el conjunto A) y en otras, no ocurrirá A. Denote con $n(A)$ el número de réplicas en las cuales A sí ocurre. El cociente $n(A)/n$ se conoce como la *frecuencia relativa* de ocurrencia del evento A en la secuencia de n réplicas.

Por ejemplo, sea A el evento de que un paquete enviado al interior del estado de California para ser entregado en dos días en realidad llega en un día. Los resultados de enviar 10 paquetes de este tipo (las primeras 10 repeticiones) son los siguientes:

Paquete #	1	2	3	4	5	6	7	8	9	10
¿Ocurre A?	N	Y	Y	Y	N	N	Y	Y	N	N
Frecuencia relativa de A	0	0.5	0.667	0.75	0.6	0.5	0.571	0.625	0.556	0.5

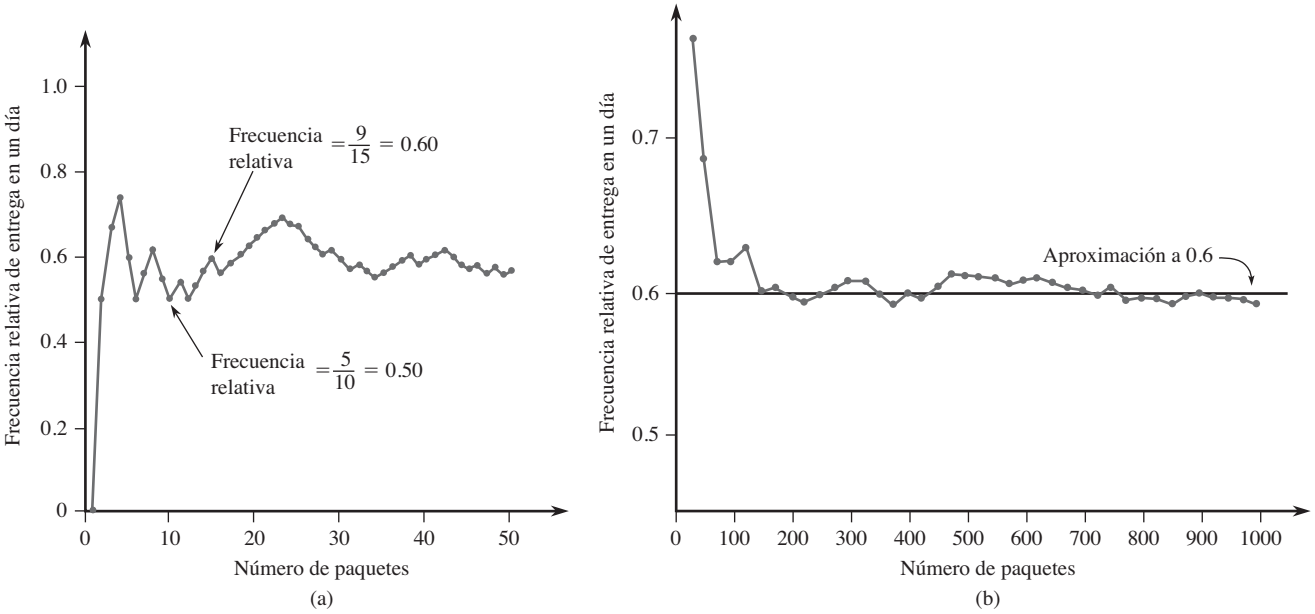


Figura 2.2 Comportamiento de la frecuencia relativa (a) fluctuación inicial (b) estabilización a largo plazo.

La figura 2.2(a) muestra cómo la frecuencia relativa fluctúa sustancialmente en el curso de las primeras 50 repeticiones. Pero puesto que el número de repeticiones sigue aumentando, la figura 2.2(b) ilustra cómo se estabiliza la frecuencia relativa.

En términos más generales, la evidencia empírica con base en los resultados de muchos experimentos repetibles, indica que cualquier frecuencia relativa de este tipo se estabilizará conforme el número de repeticiones n aumente. Es decir, puesto que n se hace arbitrariamente grande, $n(A)/n$ se aproxima a un valor límite, que se denomina *límite* (o *largo plazo*) *de la frecuencia relativa* del evento A. La *interpretación objetiva de probabilidad* identifica esta frecuencia límite en relación con $P(A)$. Suponga que las probabilidades se asignan a los acontecimientos de acuerdo con su límite de frecuencias relativas. Una afirmación como "la probabilidad de que un paquete se entregue en un día de envío es 0.6" significa que de un gran número de paquetes enviados por correo, aproximadamente 60% llegará en un día. Del mismo modo, si B es el evento de que un tipo particular de aparato necesitará servicio mientras la garantía es válida, entonces $P(B) = 0.1$ se interpreta en el sentido de que en el largo plazo, 10% de estos aparatos necesitará un servicio de garantía. Esto no quiere decir exactamente que uno de cada 10 necesitará servicio o que exactamente 10 de cada 100 necesitarán servicio, ya que 10 y 100 no son a largo plazo.

Se dice que esta interpretación de frecuencia relativa de probabilidad es objetiva porque se apoya en una propiedad del experimento y no en algún individuo particular interesado en el experimento. Por ejemplo, dos observadores diferentes de una secuencia de lanzamiento de una moneda deberán utilizar la misma asignación de probabilidad puesto que los observadores no tienen nada que ver con la frecuencia relativa límite. En la práctica, la interpretación no es tan objetiva como podría parecer, puesto que la frecuencia relativa límite de un evento no será conocida. Por tanto, se tendrán que asignar probabilidades con base en creencias sobre la frecuencia relativa límite de los eventos en estudio. Afortunadamente, existen muchos experimentos para los cuales habrá consenso respecto a las asignaciones de probabilidad. Cuando se habla de una moneda imparcial, significa que $P(H) = P(T) = 0.5$ y un dado imparcial es aquel para el cual las frecuencias relativas limitativas de los seis resultados son 1/6, lo cual sugiere las asignaciones de probabilidad $P(\{1\}) = \cdots = P(\{6\}) = 1/6$.

Debido a que la interpretación objetiva de probabilidad se basa en el concepto de frecuencia limitativa, su aplicabilidad está limitada a situaciones experimentales repetibles.

No obstante, el lenguaje de probabilidad a menudo se utiliza en referencia a situaciones que son inherentemente irrepetibles. Incluimos algunos ejemplos: "Las probabilidades de un tratado de paz son buenas"; "Es probable que el contrato le sea otorgado a nuestra compañía"; y "Debido a que su mejor mariscal de campo está lesionado, espero que no anoten más de 10 puntos contra nosotros". En tales situaciones se desearía, como antes, asignar probabilidades numéricas a varios resultados y eventos (p. ej., la probabilidad de que obtengamos el contrato es 0.9). Por consiguiente se debe adoptar una interpretación alternativa de estas probabilidades. Puesto que diferentes observadores pueden tener información y opiniones previas respecto a tales situaciones experimentales, las asignaciones de probabilidad ahora pueden diferir de un individuo a otro. Las interpretaciones en tales situaciones se conocen, por tanto, como *subjetivas*. El libro de Robert Winkler que se cita en las referencias del capítulo ofrece un recuento sencillo de varias interpretaciones subjetivas.

Más propiedades de probabilidad

PROPOSICIÓN

Para cualquier evento A, $P(A) + P(A') = 1$, a partir de lo cual $P(A) = 1 - P(A')$.

Demostración En el axioma 3, sea $k = 2$, $A_1 = A$, y $A_2 = A'$. A partir de la definición de A', $A \cup A' = \mathcal{S}$ mientras A y A' sean eventos disjuntos, $1 = P(\mathcal{S}) = P(A \cup A') = P(A) + P(A')$. ■

Esta proposición es sorprendentemente útil porque se presentan muchas situaciones en las cuales $P(A')$ es más fácil de obtener mediante métodos directos que $P(A)$.

EJEMPLO 2.13 Considere un sistema de cinco componentes idénticos conectados en serie, como se ilustra en la figura 2.3.

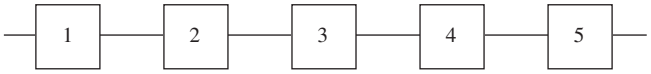


Figura 2.3 Sistema de cinco componentes conectados en serie

Denote con F un componente que falla y con S uno que no falla (por *éxito)*. Sea A el evento en el que el *sistema* falla. Para que ocurra A, al menos uno de los componentes individuales debe fallar. Los resultados en A incluyen *SSFSS* (1, 2, 4 y 5 funcionarán, pero 3 no), *FFSSS*, etcétera. Existen, de hecho, 31 resultados diferentes en A. Sin embargo, A', el evento en que el sistema funciona, consiste en el resultado único *SSSSS*. En la sección 2.5 se verá que si 90% de todos estos componentes no falla y diferentes componentes fallan independientemente uno de otro $P(A') = P(SSSSS) = 0.9^5 = 0.59$. Así, $P(A) = 1 - 0.59 = 0.41$; por tanto, entre un gran número de sistemas como ese, aproximadamente 41% fallará. ■

En general, la proposición anterior es útil cuando el evento de interés puede ser expresado como "al menos …", puesto que en ese caso puede ser más fácil trabajar con el complemento "menos que …" (en algunos problemas es más fácil trabajar con "más que …" que con "cuando mucho …"). Cuando se tenga dificultad al calcular $P(A)$ directamente habrá que pensar en determinar $P(A')$.

PROPOSICIÓN

Para cualquier evento A, $P(A) \leq 1$.

Esto se debe a que $1 = P(A) + P(A') \geq P(A)$ puesto que $P(A') \geq 0$.

Cuando los eventos A y B son mutuamente excluyentes, $P(A \cup B) = P(A) + P(B)$. Para eventos que no son mutuamente excluyentes la adición de $P(A)$ y $P(B)$ da como resultado

un "doble conteo" de los resultados en la intersección. El siguiente resultado, la *regla de la adición* para la probabilidad de una unión doble, muestra cómo corregir esto.

PROPOSICIÓN

Para dos eventos cualesquiera A y B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Demostración Note primero que $A \cup B$ se puede descomponer en dos eventos *excluyentes*, A y $B \cap A'$; la última es la parte de B que queda afuera de A (véase la figura 2.4). Además, por sí mismo B es la unión de los dos eventos excluyentes $A \cap B$ y $A' \cap B$, por tanto, $P(B) = P(A \cap B) + P(A' \cap B)$. Así

$$\begin{aligned} P(A \cup B) &= P(A) + P(B \cap A') = P(A) + [P(B) - P(A \cap B)] \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

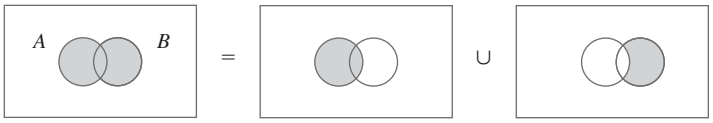


Figura 2.4 Representación de A ∪ B como la unión de dos eventos excluyentes ■

EJEMPLO 2.14 En cierto suburbio residencial 60% de las familias se suscribe al servicio de internet de la compañía de televisión por cable, 80% contrata el servicio de televisión de esa misma compañía y 50% de todas las familias contrata ambos servicios. Si se elige una familia al azar, ¿cuál es la probabilidad de que contrate al menos uno de los dos servicios, y cuál es la probabilidad de que contrate exactamente uno de ambos servicios?

Con $A = \{$se suscribe al servicio de internet$\}$ y $B = \{$se suscribe al servicio de televisión por cable$\}$, la información dada implica que $P(A) = 0.6$, $P(B) = 0.8$ y $P(A \cap B) = 0.5$. La proposición precedente ahora lleva a

$P($se suscribe al menos a uno de los dos servicios$)$

$$= P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.8 - 0.5 = 0.9$$

El evento de que una familia se suscribe sólo al servicio de televisión por cable se escribe como $A' \cap B$ [(no internet) y televisión]. Ahora la figura 2.4 implica que

$$0.9 = P(A \cup B) = P(A) + P(A' \cap B) = 0.6 + P(A' \cap B)$$

a partir de la cual $P(A' \cap B) = 0.3$. Asimismo, $P(A \cap B') = P(A \cup B) - P(B) = 0.1$. Todo esto se ilustra en la figura 2.5, donde se ve que

$$P(\text{exactamente uno}) = P(A \cap B') + P(A' \cap B) = 0.1 + 0.3 = 0.4$$

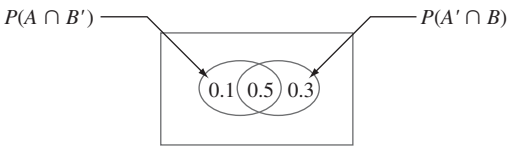


Figura 2.5 Probabilidades para el ejemplo 2.14 ■

La regla de adición para la probabilidad de una unión triple es similar a la regla anterior.

PROPOSICIÓN

Para tres eventos cualesquiera A, B y C,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Esto se puede verificar al examinar el diagrama de Venn de $A \cup B \cup C$, el cual se muestra en la figura 2.6. Cuando $P(A)$, $P(B)$ y $P(C)$ se agregan, las probabilidades de intersección $P(A \cap B)$, $P(A \cap C)$ y $P(B \cap C)$ se cuentan dos veces. Por tanto cada una se debe restar. Pero entonces $P(A \cap B \cap C)$ se ha sumado tres veces y se ha restado también tres veces, por lo que hay que agregarla de nuevo. En general, la probabilidad de unión de k eventos se obtiene sumando las probabilidades de cada uno, restando las probabilidades de doble intersección, sumando las probabilidades de triple intersección, restando las probabilidades de cuádruple intersección y así sucesivamente.

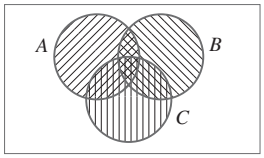


Figura 2.6 $A \cup B \cup C$

Determinación sistemática de probabilidades

Considere un espacio muestral finito o "contablemente infinito" (esto último significa que los resultados pueden ser puestos en lista en una secuencia infinita, por lo que existen un primer resultado, un segundo resultado, un tercer resultado y así sucesivamente; por ejemplo, el escenario de prueba de baterías del ejemplo 2.12). Denote con E_1, E_2, E_3,… los eventos simples correspondientes, cada uno de ellos compuesto de un solo resultado. Una estrategia sensata para el cálculo de probabilidad es determinar primero cada probabilidad del evento simple, con la condición de que $\sum P(E_i) = 1$. Entonces la probabilidad de cualquier evento compuesto A se calcula agregando los $P(E_i)$ para todos los E_i que existen en A:

$$P(A) = \sum_{\text{todas las } E_i \text{ en } A} P(E_i)$$

EJEMPLO 2.15

Durante las horas no pico el tren que viaja entre los suburbios y la ciudad utiliza cinco carros. Suponga que existe el doble de probabilidades de que un usuario seleccione el carro intermedio (#3) a cualquier carro adyacente (#2 o #4) y hay también el doble de probabilidades de que seleccione cualquier carro adyacente a cualquier carro de los extremos (#1 o #5). Sea $p_i = P(\text{carro } i \text{ es seleccionado}) = P(E_i)$. Entonces se tiene $p_3 = 2p_2 = 2p_4$ y $p_2 = 2p_1 = 2p_5 = p_4$. Esto da

$$1 = \sum P(E_i) = p_1 + 2p_1 + 4p_1 + 2p_1 + p_1 = 10p_1$$

lo cual implica $p_1 = p_5 = 0.1$, $p_2 = p_4 = 0.2$, $p_3 = .4$. La probabilidad de que uno de los tres carros intermedios se seleccione (un evento compuesto) es entonces $p_2 + p_3 + p_4 = 0.8$. ■

Resultados igualmente probables

En muchos experimentos compuestos de N resultados es razonable asignar probabilidades iguales a todos los N eventos simples. Esto incluye ejemplos tan obvios como lanzar al aire una moneda o un dado una o dos veces (o cualquier número fijo de veces) o seleccionar una o varias cartas de un mazo de 52 cartas bien barajado. Con $p = P(E_i)$ por cada i,

$$1 = \sum_{i=1}^{N} P(E_i) = \sum_{i=1}^{N} p = p \cdot N \quad \text{por lo que } p = \frac{1}{N}$$

Es decir, si existen N resultados igualmente probables, la probabilidad de cada uno es $1/N$.

Ahora considere un evento A, con $N(A)$ como el número de resultados contenidos en A. En seguida

$$P(A) = \sum_{E_i \text{ en } A} P(E_i) = \sum_{E_i \text{ en } A} \frac{1}{N} = \frac{N(A)}{N}$$

Así, cuando los resultados son igualmente probables, el cálculo de probabilidades se reduce a contar: determine tanto el número de resultados $N(A)$ en A como el número de resultados N en $\mathcal{S}$ y forme su cociente.

EJEMPLO 2.16 Usted tiene en su biblioteca seis libros de misterios y seis de ciencia ficción, todos sin leer. Los tres primeros de cada tipo son de tapa dura y los tres últimos son de bolsillo. Considere la posibilidad de seleccionar al azar uno de los seis libros de misterios y luego seleccionar al azar uno de los seis libros de ciencia ficción para sus vacaciones en Acapulco (después de todo, necesita algo para leer en la playa). Numere los libros de misterio 1, 2,…, 6, y haga lo mismo con los libros de ciencia ficción. A continuación, cada resultado es un par de números, tales como (4, 1) y hay $N = 36$ resultados posibles (para una representación visual de esta situación consulte la tabla en el ejemplo 2.3 y elimine la primera fila y la primera columna). Con la selección al azar, como se ha descrito, los 36 resultados son igualmente probables. Nueve de estos resultados son tales que los libros seleccionados son libros de bolsillo (aquellos en la esquina inferior derecha de la tabla de referencia): (4, 4), (4, 5),…, (6, 6). Así que la probabilidad del evento A de que ambos libros seleccionados sean libros de bolsillo es

$$P(A) = \frac{N(A)}{N} = \frac{9}{36} = 0.25$$ ■

EJERCICIOS Sección 2.2 (11–28)

11. Una compañía de fondos mutuos de inversión ofrece a sus clientes varios fondos: un fondo de mercado de dinero, tres fondos de bonos (a corto, intermedio y largo plazos), dos fondos de acciones (de riesgo moderado y de alto riesgo) y un fondo balanceado. Entre los clientes que poseen acciones en un solo fondo los porcentajes de clientes en los diferentes fondos son como sigue:

Mercado de dinero	20%	Acciones de alto riesgo	18%
Bonos a corto plazo	15%	Acciones de riesgo moderado	25%
Bonos a plazo intermedio	10%	Balanceados	7%
Bonos a largo plazo	5%		

Se selecciona al azar un cliente que posea acciones en solo un fondo.

- **a.** ¿Cuál es la probabilidad de que el cliente seleccionado posea acciones en el fondo balanceado?
- **b.** ¿Cuál es la probabilidad de que el mismo cliente posea acciones en un fondo de bonos?
- **c.** ¿Cuál es la probabilidad de que ese cliente no posea acciones en un fondo de acciones?

12. Considere seleccionar al azar a un estudiante en cierta universidad y que A denote el evento en que el individuo seleccionado tenga una tarjeta de crédito Visa y que B denote el evento análogo para la tarjeta MasterCard. Suponga que $P(A) = 0.6$, y $P(B) = 0.4$.

- **a.** ¿Podría darse el caso de que $P(A \cap B) = 0.5$? ¿Por qué sí o por qué no? [*Sugerencia*: Véase el ejercicio 24.]
- **b.** De aquí en adelante suponga que $P(A \cap B) = 0.3$. ¿Cuál es la probabilidad de que el alumno seleccionado tenga al menos uno de ambos tipos de tarjeta?
- **c.** ¿Cuál es la probabilidad de que el alumno seleccionado no tenga ningún tipo de tarjeta?
- **d.** Describa, en términos de A y B, el evento en que el alumno seleccionado tenga una tarjeta Visa pero no una tarjeta MasterCard y después calcule la probabilidad de este evento.
- **e.** Calcule la probabilidad de que el alumno seleccionado tenga exactamente uno de los dos tipos de tarjeta.

13. Una firma consultora de computación presentó propuestas en tres proyectos. Sea $A_i = \{\text{proyecto otorgado } i\}$, con $i = 1, 2, 3$ y suponga que $P(A_1) = 0.22$, $P(A_2) = 0.25$, $P(A_3) = 0.28$, $P(A_1 \cap A_2) = 0.11$, $P(A_1 \cap A_3) = 0.05$, $P(A_2 \cap A_3) = 0.07$, $P(A_1 \cap A_2 \cap A_3) = 0.01$. Exprese en palabras siguientes eventos y calcule la probabilidad de cada uno:

- **a.** $A_1 \cup A_2$
- **b.** $A'_1 \cap A'_2$ [*Sugerencia:* $(A_1 \cup A_2)' = A'_1 \cap A'_2$]
- **c.** $A_1 \cup A_2 \cup A_3$
- **d.** $A'_1 \cap A'_2 \cap A'_3$
- **e.** $A'_1 \cap A'_2 \cap A_3$
- **f.** $(A'_1 \cap A'_2) \cup A_3$

14. Suponga que 55% de todos los adultos consume regularmente café, 45% consume regularmente refrescos con gas y 70% consume con frecuencia al menos uno de estos dos productos.

a. ¿Cuál es la probabilidad de que un adulto al azar regularmente consuma café y soda?

b. ¿Cuál es la probabilidad de que un adulto al azar no consuma regularmente al menos uno de estos dos productos?

15. Considere el tipo de secadora de ropa (de gas o eléctrica) adquirida por cada uno de cinco clientes diferentes en cierta tienda.

a. Si la probabilidad de que al menos uno de ellos adquiera una secadora eléctrica es 0.428, ¿cuál es la probabilidad de que al menos dos adquieran una secadora eléctrica?

b. Si P(los cinco compran una secadora de gas) = 0.116 y P(los cinco compran una secadora eléctrica) = 0.005, ¿cuál es la probabilidad de que al menos se adquiera una secadora de cada tipo?

16. A un individuo se le presentan tres vasos diferentes de refresco de cola, designados C, D y P. Se le pide que pruebe los tres y que los anote en orden de preferencia. Suponga que se sirvió el mismo refresco de cola en los tres vasos.

a. ¿Cuáles son los eventos simples en este evento de clasificación y qué probabilidad le asignaría a cada uno?

b. ¿Cuál es la probabilidad de que C obtenga el primer lugar?

c. ¿Cuál es la probabilidad de que C obtenga el primer lugar y D el último?

17. Denote con A el evento en que la siguiente solicitud de asesoría de un consultor de software estadístico tiene que ver con el paquete SPSS y que B denote el evento en que la siguiente solicitud de ayuda tiene que ver con SAS. Suponga que $P(A) = 0.30$ y y $P(B) = 0.50$.

a. ¿Por qué no es el caso de que $P(A) + P(B) = 1$?

b. Calcule $P(A')$.

c. Calcule $P(A \cup B)$.

d. Calcule $P(A' \cap B)$.

18. Una cartera contiene cinco billetes de \$10, cuatro de \$5 y seis de \$1 (nada más). Si se seleccionan los billetes uno por uno en orden aleatorio, ¿cuál es la probabilidad de que se deban seleccionar al menos dos billetes para obtener un primer billete de \$10?

19. La inspección visual humana de uniones soldadas en un circuito impreso puede ser muy subjetiva. Una parte del problema se deriva de los numerosos tipos de defectos de soldadura (p. ej., almohadilla seca, visibilidad en escuadra, picaduras) incluso del grado al cual una unión posee uno o más de estos defectos. Por consiguiente, incluso inspectores altamente entrenados pueden discrepar en cuanto a la disposición particular de una unión particular. En un lote de 10 000 uniones el inspector A encontró 724 defectuosas y el inspector B, 751 mientras que 1159 fueron consideradas defectuosas por al menos uno de los inspectores. Suponga que se selecciona una de las 10 000 uniones al azar.

a. ¿Cuál es la probabilidad de que la unión seleccionada no sea juzgada defectuosa por ninguno de los dos inspectores?

b. ¿Cuál es la probabilidad de que la unión seleccionada sea juzgada defectuosa por el inspector B, pero no por el inspector A?

20. En cierta fábrica se trabajan tres turnos diferentes. Durante el año pasado ocurrieron en la fábrica 200 accidentes. Algunos de ellos pueden ser atribuidos, al menos en parte, a condiciones de trabajo inseguras mientras que las otras no se relacionan con las condiciones de trabajo. La tabla adjunta da el porcentaje de accidentes que ocurren en cada tipo de categoría accidente-turno.

		Condiciones inseguras	No vinculados a las condiciones
Turno	*Diurno*	10%	35%
	Mixto	8%	20%
	Nocturno	5%	22%

Suponga que uno de los 200 reportes de accidente se selecciona al azar de un archivo de reportes y que el turno, y el tipo de accidente se han determinado.

a. ¿Cuáles son los eventos simples?

b. ¿Cuál es la probabilidad de que el accidente seleccionado se atribuya a condiciones inseguras?

c. ¿Cuál es la probabilidad de que el accidente seleccionado no haya ocurrido en el turno diurno?

21. Una compañía de seguros ofrece cuatro diferentes niveles de deducible: ninguno, bajo, medio y alto para sus tenedores de pólizas de propietario de casa; y tres diferentes niveles: bajo, medio y alto para sus tenedores de pólizas de automóviles. La tabla adjunta muestra las proporciones de las varias categorías de tenedores de pólizas que tienen ambos tipos de seguro. Por ejemplo, la proporción de individuos con deducible bajo de casa y deducible bajo de automóvil es 0.06 (6% de todos los individuos).

	Propietarios de viviendas			
Auto	**N**	**L**	**M**	**H**
L	0.04	0.06	0.05	0.03
M	0.07	0.10	0.20	0.10
H	0.02	0.03	0.15	0.15

Suponga que se elige al azar un individuo que posee ambos tipos de pólizas.

a. ¿Cuál es la probabilidad de que esta persona tenga un deducible de auto medio y un deducible de casa alto?

b. ¿Cuál es la probabilidad de que tenga un deducible de casa bajo y un deducible de auto bajo?

c. ¿Cuál es la probabilidad de que este individuo se encuentre en la misma categoría de deducibles de casa y auto?

d. Basado en su respuesta para el inciso c), ¿cuál es la probabilidad de que las dos categorías sean diferentes?

e. ¿Cuál es la probabilidad de que el individuo tenga al menos un nivel deducible bajo?

f. Utilizando la respuesta para el inciso e), ¿cuál es la probabilidad de que ningún nivel deducible sea bajo?

22. En la ruta que un automovilista recorre para trasladarse a su trabajo existen dos intersecciones con señalamientos de tránsito.

La probabilidad de que el automovilista tenga que detenerse en la primera señal es .4, la probabilidad análoga para la segunda señal es 0.5 y la probabilidad de que tenga que detenerse en al menos una de las dos señales es 0.7. Cuál es la probabilidad de que tenga que detenerse:

a. ¿En ambas señales?

b. ¿En la primera señal, pero no en la segunda?

c. ¿En exactamente una señal?

23. Las computadoras de seis profesores de la facultad deberán ser reemplazadas. Dos de los profesores seleccionaron computadoras portátiles y los otros cuatro escogieron computadoras de escritorio. Suponga que sólo se pueden realizar dos configuraciones en un día particular y que las dos computadoras que van a ser configuradas se seleccionan al azar de entre las seis (lo cual implica 15 resultados igualmente probables; si las computadoras se numeran 1, 2,…, 6, entonces un resultado se compone de las computadoras 1 y 2; otro resultado, de las computadoras 1 y 3, y así sucesivamente).

a. ¿Cuál es la probabilidad de que las dos configuraciones seleccionadas sean para las computadoras portátiles?

b. ¿Cuál es la probabilidad de que ambas configuraciones seleccionadas sean para las computadoras de escritorio?

c. ¿Cuál es la probabilidad de que al menos una configuración seleccionada sea para una computadora de escritorio?

d. ¿Cuál es la probabilidad de que al menos una computadora de cada tipo sea elegida para ser configurada?

24. Demuestre que si un evento A está contenido en otro evento B (es decir, A es un subconjunto de B), entonces $P(A) \leq P(B)$. [*Sugerencia*: Para tales A y B, A y $B \cap A'$ son disjuntos y $B = A \cup (B \cap A')$, como se puede ver en el diagrama de Veen.] Para A y B en general, ¿qué implica esto respecto a la relación entre $P(A \cap B)$, $P(A)$ y $P(A \cup B)$?

25. Las tres opciones más populares en un cierto tipo de automóvil nuevo son GPS (sistema de posicionamiento global) (A), quemacocos (B) y transmisión automática (C). Si 40% de todos los compradores solicita A, 55% solicita B, 70% pide C, 63% A o B, 77% A o C, 80% B o C y 85% solicita A o B o C, calcule las probabilidades de los siguientes eventos. [*Sugerencia*: *"A* o *B"* es el evento en el que se solicita al menos una de las dos opciones; intente trazar un diagrama de Venn y rotule todas las regiones.]

a. El siguiente comprador solicitará al menos una de las tres opciones.

b. El siguiente comprador no seleccionará ninguna de las tres opciones.

c. El siguiente comprador solicitará sólo transmisión automática y ninguna de las otras dos opciones.

d. El siguiente comprador seleccionará exactamente una de estas tres opciones.

26. Un sistema puede experimentar tres tipos diferentes de defectos. Sea A_i ($i = 1, 2, 3$) el evento en que el sistema tiene un defecto de tipo i. Suponga que

$P(A_1) = 0.12 \quad P(A_2) = 0.07 \quad P(A_3) = 0.05$

$P(A_1 \cup A_2) = 0.13 \quad P(A_1 \cup A_3) = 0.14$

$P(A_2 \cup A_3) = 0.10 \quad P(A_1 \cap A_2 \cap A_3) = 0.01$

a. ¿Cuál es la probabilidad de que el sistema no tenga un defecto de tipo 1?

b. ¿Cuál es la probabilidad de que el sistema tenga defectos de tipo 1 así como de tipo 2?

c. ¿Cuál es la probabilidad de que el sistema tenga defectos de tipo 1 y de tipo 2 pero no de tipo 3?

d. ¿Cuál es la probabilidad de que el sistema tenga a lo más dos de estos defectos?

27. El departamento académico formado por cinco profesores de la facultad: Anderson, Box, Cox, Cramer y Fisher, debe seleccionar a dos de ellos para que participen en un comité de revisión de personal. Puesto que el trabajo requerirá mucho tiempo, ninguno parece ansioso de participar, por lo que se decidió anotar en trozos de papel idénticos el nombre de cada uno y seleccionarlos al azar.

a. ¿Cuál es la probabilidad de que tanto Anderson como Box sean seleccionados? [*Sugerencia*: Mencione los resultados igualmente probables.]

b. ¿Cuál es la probabilidad de que al menos uno de los dos miembros cuyo nombre comienza con *C* sea seleccionado?

c. Si los cinco miembros del cuerpo de profesores han impartido clases en la universidad durante 3, 6, 7, 10 y 14 años, respectivamente, ¿cuál es la probabilidad de que los dos representantes seleccionados tengan al menos 15 años de experiencia académica?

28. En el ejercicio 5 suponga que cualquier individuo que entre a la clínica tiene las mismas probabilidades de ser asignado a cualquiera de las tres estaciones independientemente de a dónde hayan sido asignados otros individuos. Cuál es la probabilidad de que:

a. ¿Los tres miembros de una familia sean asignados a la misma estación?

b. ¿Al menos dos miembros de la familia sean asignados a la misma estación?

c. ¿Cada miembro de la familia sea asignado a una estación diferente?

2.3 Técnicas de conteo

Cuando los diversos resultados de un experimento son igualmente probables (la misma probabilidad es asignada a cada evento simple), la tarea de calcular probabilidades se reduce a contar. Sea N el número de resultados en un espacio muestral y $N(A)$ el número de resultados contenidos en un evento A,

$$P(A) = \frac{N(A)}{N} \tag{2.1}$$

Si una lista de resultados es fácil de obtener y N es pequeña, entonces N y $N(A)$ pueden ser determinadas sin utilizar ningún principio de conteo.

Existen, sin embargo, muchos experimentos en los cuales el esfuerzo implicado al elaborar la lista es prohibitivo porque N es bastante grande. Explotando algunas reglas de conteo generales, es posible calcular probabilidades de la forma (2.1) sin una lista de resultados. Estas reglas también son útiles en muchos problemas que implican resultados que no son igualmente probables. Se utilizarán varias de las reglas desarrolladas aquí al estudiar distribuciones de probabilidad en el siguiente capítulo.

Regla de producto para pares ordenados

La primera regla de conteo se aplica a cualquier situación en la cual un conjunto (evento) se compone de pares ordenados de objetos y se desea contar el número de pares. Por par ordenado se quiere decir que si O_1 y O_2 son objetos, entonces el par (O_1, O_2) es diferente del par (O_2, O_1). Por ejemplo, si un individuo selecciona una línea aérea para viajar de Los Ángeles a Chicago y (luego de realizar algunas transacciones de negocios en Chicago) una segunda para continuar a Nueva York, una posibilidad es (American, United), otra es (United, American) y una más es (United, United).

PROPOSICIÓN

Si el primer elemento u objeto de un par ordenado puede ser seleccionado de n_1 maneras, y si por cada una de estas n_1 maneras el segundo elemento del par puede ser seleccionado de n_2 maneras, entonces el número de pares es n_1n_2.

Una interpretación alternativa consiste en llevar a cabo una operación que consta de dos etapas. Si la primera etapa se puede realizar en cualquiera n_1 maneras y para cada una hay n_2 formas de realizar la segunda etapa, entonces, n_1n_2 es el número de maneras de llevar a cabo las dos etapas en la secuencia.

EJEMPLO 2.17 El propietario de una casa va a llevar a cabo una remodelación y requiere los servicios de un contratista de fontanería y un contratista de electricidad. Si existen 12 contratistas de fontanería y nueve de electricidad disponibles en el área, ¿de cuántas maneras pueden ser elegidos? Si $P_1,\ldots, P_{12}$ son los fontaneros y $Q_1,\ldots, Q_9$ son los electricistas, entonces se desea el número de pares de la forma (P_i, Q_j). Con $n_1 = 12$ y $n_2 = 9$, la regla de producto da $N = (12)(9) = 108$ formas posibles de seleccionar los dos tipos de contratistas. ■

En el ejemplo 2.17 la selección del segundo elemento del par no dependió de cuál primer elemento ocurrió o fue elegido. En tanto exista el mismo número de opciones del segundo elemento por cada primer elemento, la regla de producto es válida incluso cuando el conjunto de posibles segundos elementos depende del primero.

EJEMPLO 2.18 Una familia recién se cambió a una nueva ciudad y requiere los servicios de un obstetra y de un pediatra. Existen dos clínicas médicas de fácil acceso y cada una tiene dos obstetras y tres pediatras. La familia obtendrá los máximos beneficios del seguro de salud si se afilia a una clínica y selecciona a ambos doctores de dicha clínica. ¿De cuántas maneras se puede hacer esto? Denote a los obstetras con O_1, O_2, O_3 y O_4 y a los pediatras con $P_1,\ldots, P_6$. Se desea el número de pares (O_i, P_j) para los cuales O_i y P_j están asociados a la misma clínica. Puesto que existen cuatro obstetras, $n_1 = 4$; y por cada uno existen tres opciones de pediatras, por tanto, $n_2 = 3$. Aplicando la regla de producto se obtiene $N = n_1n_2 = 12$ posibles opciones. ■

En muchos problemas de conteo y probabilidad se puede utilizar una configuración conocida como **diagrama de árbol** para representar pictóricamente todas las posibilidades. En la figura 2.7 aparece el diagrama de árbol asociado al ejemplo 2.18. Partiendo de un punto

localizado en el lado izquierdo del diagrama, por cada posible primer elemento de un par emana un segmento de línea recta hacia la derecha. Cada una de estas rectas se conoce como rama de primera generación. Ahora, para cualquier rama de primera generación se construye otro segmento de recta que emana de la punta de la rama por cada posible opción de un segundo elemento del par. Cada segmento de recta es una rama de segunda generación. Puesto que hay cuatro obstetras, hay cuatro ramas de primera generación, y tres pediatras por cada obstetra resultan en tres ramas de segunda generación que emanan de cada rama de primera generación.

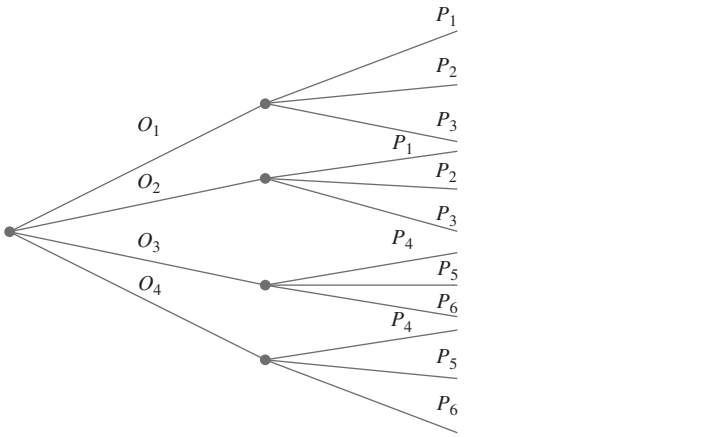


Figura 2.7 Diagrama de árbol para el ejemplo 2.18

Generalizando, suponga que existen n_1 ramas de primera generación y que por cada rama de primera generación existen n_2 ramas de segunda generación. El número total de ramas de segunda generación es entonces n_1n_2. Debido a que el extremo de cada rama de segunda generación corresponde a exactamente un posible par (la selección de un primer elemento y luego de un segundo elemento nos sitúa en el extremo de exactamente una rama de segunda generación) existen n_1n_2 pares, con lo cual se verifica la regla de producto.

La construcción de un diagrama de árbol no depende de que se tenga el mismo número de ramas de segunda generación que emanen de cada rama de primera generación. Si la segunda clínica tuviera cuatro pediatras, entonces habría sólo tres ramas emanando de dos de las ramas de primera generación y cuatro emanando de cada una de las otras dos ramas de primera generación. Un diagrama de árbol puede ser utilizado, por tanto, para representar pictóricamente experimentos diferentes de aquellos en los que se aplica la regla de producto.

Una regla de producto más general

Si se lanza al aire un dado de seis lados cinco veces en sucesión en lugar de sólo dos veces, entonces cada posible resultado es un conjunto ordenado de cinco números como (1, 3, 1, 2, 4) o (6, 5, 2, 2, 2). Un conjunto ordenado de k objetos recibirá el nombre de ***k-tupla*** (por tanto, un par es una 2-tupla y una terna es una 3-tupla). Cada resultado del experimento de lanzar al aire un dado es entonces una 5-tupla.

Regla de producto para *k*-tuplas

Suponga que un conjunto se compone de conjuntos ordenados de k elementos (k-tuplas) y que existen n_1 posibles opciones para el primer elemento; por cada opción del primer elemento, existen n_2 posibles opciones del segundo elemento; …; por cada posible opción de los primeros $k - 1$ elementos, existen n_k opciones del elemento k-ésimo. Existen entonces $n_1n_2 \cdot \cdots \cdot n_k$ posibles k-tuplas.

Una interpretación alternativa consiste en llevar a cabo una operación en k etapas. Si la primera etapa se puede realizar en cualesquiera n_1 maneras, si para cada una de tales maneras hay n_2 formas de realizar la segunda etapa, y si para cada forma de llevar a cabo las dos primeras etapas hay n_3 formas de realizar la tercera fase, y así sucesivamente, entonces $n_1n_2 \cdot \cdots \cdot n_k$ es el número de formas para llevar a cabo toda la k-etapa de operación en secuencia. Esta regla más general también se puede visualizar con un diagrama de árbol. Para el caso $k = 3$ sólo se debe añadir un número adecuado de 3ª generación en las ramas de la punta de cada rama de 2ª generación. Si, por ejemplo, una ciudad universitaria tiene cuatro pizzerías, un complejo de cine con seis pantallas y tres lugares para ir a bailar, entonces habrá cuatro ramas de 1ª generación, seis ramas de 2ª generación que emanan de la punta de cada rama de 1ª generación, y tres ramas de 3ª generación que abren cada rama de 2ª generación. Cada posible 3-tupla corresponde a la punta de una rama de 3ª generación.

EJEMPLO 2.19 (Continuación del ejemplo 2.17)

Suponga que el trabajo de remodelación de la casa implica adquirir primero varios utensilios de cocina. Se adquirirán en la misma tienda y hay cinco tiendas en el área. Mediante las tiendas denotadas con $D_1,\ldots,D_5$, existen $N = n_1n_2n_3 = (5)(12)(9) = 540$ 3-tuplas de la forma (D_i, P_j, Q_k), entonces existen 540 formas de elegir primero una tienda, luego un contratista de fontanería y finalmente un contratista electricista. ■

EJEMPLO 2.20 (Continuación del ejemplo 2.18)

Si cada clínica tiene dos especialistas en medicina interna y dos médicos generales, existen $n_1n_2n_3n_4 = (4)(3)(3)(2) = 72$ formas de seleccionar un doctor de cada tipo, de tal suerte que todos los doctores practiquen en la misma clínica. ■

Permutaciones y combinaciones

Considere un grupo de n individuos u objetos distintos ("distintos" significa que existe alguna característica que diferencia a cualquier individuo u objeto de cualquier otro). ¿Cuántas maneras existen de seleccionar un subconjunto de tamaño k del grupo? Por ejemplo, si un equipo de ligas pequeñas tiene 15 jugadores registrados, ¿cuántas maneras existen de seleccionar nueve jugadores para una alineación inicial? O, si una librería universitaria vende diez computadoras portátiles diferentes, pero tiene espacio para mostrar sólo tres de ellas, de cuántas maneras pueden ser elegidas las tres?

Una respuesta a la pregunta general que se acaba de plantear requiere distinguir entre dos casos. En algunas situaciones, tal como en el escenario del béisbol, el orden de la selección es importante. Por ejemplo, con Ángela como lanzador y Beto como receptor se obtiene una alineación diferente de aquella con Ángela como receptor y Beto como lanzador. No obstante, con frecuencia el orden no es importante y a nadie le interesa qué individuos u objetos son seleccionados, como sería el caso en el escenario para seleccionar las computadoras portátiles.

DEFINICIÓN

> Un subconjunto ordenado se llama **permutación**. El número de permutaciones de tamaño k que se puede formar con los n individuos u objetos en un grupo será denotado con $P_{k,n}$. Un subconjunto no ordenado se llama **combinación**. Una forma de denotar el número de combinaciones es $C_{k,n}$, pero en su lugar se utilizará una notación bastante común en los libros sobre probabilidad: $\binom{n}{k}$, que se lee así: "de n se elige k".

El número de permutaciones se determina utilizando la primera regla de conteo para k-tuplas. Suponga, por ejemplo, que un colegio de ingeniería tiene siete departamentos, denotados con a, b, c, d, e, f y g. Cada departamento tiene un representante frente al consejo de estudiantes del colegio. De estos siete representantes, uno tiene que ser elegido como presidente, otro como vicepresidente y un tercero como secretario. ¿Cuántas maneras existen

de seleccionar a los tres funcionarios? Es decir, ¿cuántas permutaciones de tamaño 3 pueden ser formadas con los 7 representantes? Para responder esta pregunta habrá que pensar en formar una terna (3-tupla) en la que el primer elemento es el presidente, el segundo es el vicepresidente y el tercero es el secretario. Una terna es (a, g, b), otra es (b, g, a) y otra más es (d, f, b). Ahora bien, el presidente puede ser seleccionado en cualesquiera $n_1 = 7$ formas. Por cada forma de seleccionar al presidente, existen $n_2 = 6$ formas de seleccionar al vicepresidente y, por consiguiente, $7 \times 6 = 42$ (pares de presidente, vicepresidente). Por último, por cada forma de seleccionar un presidente y un vicepresidente, existen $n_3 = 5$ formas de seleccionar al secretario. Esto da

$$P_{3,7} = (7)(6)(5) = 210$$

como el número de permutaciones de tamaño 3 que se pueden formar con 7 individuos distintos. Una representación de diagrama de árbol mostraría tres generaciones de ramas.

La expresión para $P_{3,7}$ puede volver a escribirse con ayuda de la *notación factorial*. Recuerde que 7! (que se lee así: "factorial de 7") es una notación compacta para el producto descendente de enteros (7)(6)(5)(4)(3)(2)(1). Más generalmente, para cualquier entero positivo m, $m! = m(m-1)(m-2)\cdot\cdots\cdot(2)(1)$. Esto da $1! = 1$, y también se define $0! = 1$. Por tanto,

$$P_{3,7} = (7)(6)(5) = \frac{(7)(6)(5)(4!)}{(4!)} = \frac{7!}{4!}$$

Al generalizar para un grupo arbitrario de tamaño n y un subconjunto de tamaño k se obtiene

$$P_{k,n} = n(n-1)(n-2)\cdot\cdots\cdot(n-(k-2))(n-(k-1))$$

Al multiplicar y dividir esta ecuación por $(n-k)!$ se obtiene una expresión compacta para el número de permutaciones.

PROPOSICIÓN

$$P_{k,n} = \frac{n!}{(n-k)!}$$

EJEMPLO 2.21 Hay diez asistentes de profesor disponibles para calificar exámenes en un curso de cálculo en una gran universidad. El primer examen se compone de cuatro preguntas y el profesor desea seleccionar un asistente diferente para calificar cada pregunta (sólo un asistente por pregunta). ¿De cuántas maneras se pueden elegir los asistentes? En este caso n = tamaño del grupo = 10 y k = tamaño del subconjunto = 4. El número de permutaciones es

$$P_{4,10} = \frac{10!}{(10-4)!} = \frac{10!}{6!} = 10(9)(8)(7) = 5040$$

Es decir, el profesor podría aplicar 5040 exámenes diferentes de cuatro preguntas utilizando calificadores diferentes para cada una de las preguntas, ¡tiempo en el cual todos los asistentes seguramente habrán terminado sus programas de licenciatura! ■

Considere ahora las combinaciones (es decir, los subconjuntos no ordenados). De nuevo tendrá que remitirse al escenario de consejo estudiantil, y suponga que tres de los siete representantes tienen que ser seleccionados para asistir a una convención estatal. El orden de selección no es importante; lo que importa es cuáles tres son seleccionados. Así que se busca $\binom{7}{3}$, el número de combinaciones de 3 que se pueden formar con los

7 individuos. Considere por un momento las combinaciones a, c, g. Estos tres individuos pueden ser ordenados en $3! = 6$ formas para producir el número de permutaciones:

$$a, c, g \quad a, g, c \quad c, a, g \quad c, g, a \quad g, a, c \quad g, c, a$$

De manera similar, hay $3! = 6$ maneras de ordenar la combinación b, c, e para producir permutaciones y, de hecho, hay 3! modos de ordenar cualquier combinación particular de tamaño 3 para producir permutaciones. Esto implica la siguiente relación entre el número de combinaciones y el número de permutaciones:

$$P_{3,7} = (3!) \cdot \binom{7}{3} \Rightarrow \binom{7}{3} = \frac{P_{3,7}}{3!} = \frac{7!}{(3!)(4!)} = \frac{(7)(6)(5)}{(3)(2)(1)} = 35$$

No sería difícil poner en lista las 35 combinaciones, pero no hay necesidad de hacerlo si sólo interesa cuántas son. Observe que el número de 210 permutaciones excede por mucho el número de combinaciones; ¡el primero es más grande que el segundo por un factor de 3! puesto que así es como cada combinación puede ser ordenada.

Generalizando la línea de razonamiento anterior se obtiene una relación simple entre el número de permutaciones y el número de combinaciones que produce una expresión concisa para la última cantidad.

PROPOSICIÓN

$$\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Note que $\binom{n}{n} = 1$ y $\binom{n}{0} = 1$ puesto que sólo hay una forma de seleccionar un conjunto de n (todos) elementos, o de ningún elemento, y $\binom{n}{1} = n$ puesto que existen n subconjuntos de tamaño 1.

EJEMPLO 2.22 Una lista de reproducción de iPod contiene 100 canciones de las cuales 10 son de los Beatles. Supongamos que la función de reproducción aleatoria se utiliza para reproducir las canciones en orden aleatorio (la aleatoriedad del proceso de barajar es investigada en **"Does Your iPod *Really* Play Favorites?" (*The Amer. Statistician*, 2009: 263–268)**. ¿Cuál es la probabilidad de que la primera canción escuchada de los Beatles sea la quinta canción en reproducirse?

Para que este evento ocurra, debe ser el caso de que las primeras cuatro canciones que se reproducen no sean canciones de los Beatles (NB) y que la quinta canción sí sea de los Beatles (B). El número de maneras de seleccionar las primeras cinco canciones es de 100(99)(98)(97)(96). El número de maneras de seleccionar estas cinco canciones para que las cuatro primeras sean NB y la siguiente sea B es de 90(89)(88)(87)(10). La suposición aleatoria implica que cualquier conjunto particular de 5 canciones de entre las 100 tiene la misma probabilidad de ser seleccionado como los primeros cinco reproducidos al igual que cualquier otro conjunto de cinco canciones; cada resultado es igualmente probable. Por tanto, la probabilidad deseada es el cociente entre el número de resultados para que el evento de interés ocurra con el número de resultados posibles:

$$P(1^{a}\text{ B es la }5^{a}\text{ canción reproducida}) = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 10}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} = \frac{P_{4,90} \cdot (10)}{P_{5,100}} = 0.0679$$

Esta es una línea alternativa de razonamiento que implica combinaciones. En lugar de centrarse en la selección de sólo las primeras cinco canciones, piense en reproducir las 100 canciones en orden aleatorio. El número de formas de elegir 10 de estas canciones que sean (sin tener en cuenta el orden en que se reprodujeron) es $\binom{100}{10}$. Ahora bien, si elegimos 9 de las últimas 95 canciones que sean B, lo cual se puede hacer de $\binom{95}{9}$ maneras, quedarán

cuatro NB y una B para las primeras cinco canciones. Sólo hay una forma más de estas cinco de empezar con cuatro NB y luego seguir con una B (recordemos que estamos considerando subconjuntos *desordenados*). Por tanto,

$$P(\text{1ª B es la 5ª canción reproducida}) = \frac{\binom{95}{9}}{\binom{100}{10}}$$

Es fácil verificar que esta última expresión es, de hecho, idéntica a la primera expresión para la probabilidad deseada, por lo que el resultado numérico es de nuevo 0.0679.

La probabilidad de que una de las primeras cinco canciones que se reproducen sea una canción de los Beatles es

P(la 1ª B es la 1ª o 2ª o 3ª o 4ª o 5ª canción reproducida)

$$= \frac{\binom{99}{9}}{\binom{100}{10}} + \frac{\binom{98}{9}}{\binom{100}{10}} + \frac{\binom{97}{9}}{\binom{100}{10}} + \frac{\binom{96}{9}}{\binom{100}{10}} + \frac{\binom{95}{9}}{\binom{100}{10}} = 0.4162$$

Por tanto, es bastante probable que la canción de los Beatles sea una de las primeras cinco canciones reproducidas. Esta "coincidencia" no es tan sorprendente como podría parecer. ■

EJEMPLO 2.23 El almacén de una universidad recibió 25 impresoras, de las cuales 10 son láser y 15 son modelos de inyección de tinta. Si 6 de estas 25 se seleccionan al azar para que las revise un técnico particular, ¿cuál es la probabilidad de que exactamente 3 de las seleccionadas sean impresoras láser (de modo que las otras 3 sean de inyección de tinta)?

Sea D_3 = {exactamente 3 de las 6 seleccionadas son impresoras de inyección de tinta}. Suponiendo que cualquier conjunto particular de 6 impresoras es tan probable de ser elegido como cualquier otro conjunto de 6, se tienen resultados igualmente probables, por tanto $P(D_3) = N(D_3)/N$, donde N es el número de formas de elegir 6 impresoras de entre las 25 y $N(D_3)$ es el número de formas de elegir 3 impresoras láser y 3 de inyección de tinta. Por tanto $N = \binom{25}{6}$. Para obtener $N(D_3)$ primero se piensa en elegir 3 de las 15 impresoras de inyección de tinta y luego 3 de las impresoras láser. Existen $\binom{15}{3}$ formas de elegir las 3 impresoras de inyección de tinta y $\binom{10}{3}$ formas de elegir las 3 impresoras láser; $N(D_3)$ es ahora el producto de estos dos números (visualice un diagrama de árbol; en realidad aquí se está utilizando el argumento de la regla de producto), por tanto

$$P(D_3) = \frac{N(D_3)}{N} = \frac{\binom{15}{3}\binom{10}{3}}{\binom{25}{6}} = \frac{\frac{15!}{3!12!} \cdot \frac{10!}{3!7!}}{\frac{25!}{6!19!}} = 0.3083$$

Sea D_4 = {exactamente 4 de las 6 impresoras seleccionadas son impresoras de inyección de tinta} y defínanse D_5 y D_6 del mismo modo. Entonces la probabilidad de seleccionar al menos 3 impresoras de inyección de tinta es

$$P(D_3 \cup D_4 \cup D_5 \cup D_6) = P(D_3) + P(D_4) + P(D_5) + P(D_6)$$

$$= \frac{\binom{15}{3}\binom{10}{3}}{\binom{25}{6}} + \frac{\binom{15}{4}\binom{10}{2}}{\binom{25}{6}} + \frac{\binom{15}{5}\binom{10}{1}}{\binom{25}{6}} + \frac{\binom{15}{6}\binom{10}{0}}{\binom{25}{6}} = 0.8530$$

■

EJERCICIOS Sección 2.3 (29–44)

29. A partir de abril de 2006 aproximadamente 50 millones de nombres de dominio web.com fueron registrados (p. ej., yahoo.com).

- **a.** ¿Cuántos nombres de dominio compuestos de exactamente dos letras en secuencia pueden ser formados? ¿Cuántos nombres de dominio de dos letras existen si se permiten dígitos y letras como caracteres? [*Nota*: Ahora es obligatoria una longitud de tres o más de caracteres.]
- **b.** ¿Cuántos nombres de dominio compuestos de tres letras en secuencia existen? ¿Cuántos de esta longitud existen si se permiten igual letras y dígitos? [*Nota*: En la actualidad todos se usan.]
- **c.** Responda las preguntas formuladas en b) para secuencias de cuatro caracteres.
- **d.** A partir de abril de 2006 aún no habían sido reclamadas 97,786 de las secuencias de cuatro caracteres con letras o números. Si se elige al azar un nombre de cuatro caracteres, ¿cuál es la probabilidad de que ya tenga propietario?

30. Un amigo va a ofrecer una fiesta. Sus existencias actuales de vino incluyen 8 botellas de zinfandel, 10 de merlot y 12 de cabernet (él sólo bebe vino tinto), todos de diferentes fábricas vinícolas.

- **a.** Si desea servir 3 botellas de zinfandel y el orden de servicio es importante, ¿cuántas formas existen de hacerlo?
- **b.** Si 6 botellas de vino tienen que ser seleccionadas al azar de entre las 30 para servir, ¿cuántas formas existen de hacerlo?
- **c.** Si se seleccionan al azar 6 botellas, ¿cuántas formas existen de obtener dos botellas de cada variedad?
- **d.** Si se seleccionan 6 botellas al azar, ¿cuál es la probabilidad de que el resultado sea dos botellas de cada variedad?
- **e.** Si se eligen 6 botellas al azar, ¿cuál es la probabilidad de que todas sean de la misma variedad?

31. Beethoven, el compositor, escribió 9 sinfonías, 5 conciertos para piano (música para piano y orquesta) y 32 sonatas para piano (música para piano solo).

- **a.** ¿Cuántas maneras existen de tocar primero una sinfonía de Beethoven y después un concierto para piano de Beethoven?
- **b.** El gerente de una estación de radio decide que en noches sucesivas (7 días a la semana), se tocará una sinfonía de Beethoven seguida por un concierto para piano de Beethoven seguido por una sonata para piano de Beethoven. ¿Durante cuántos años se podría continuar con esta política antes de que exactamente el mismo programa se repitiera?

32. Una tienda de equipos de sonido está ofreciendo un precio especial en un juego completo de componentes (receptor, reproductor de discos compactos, altavoces, tornamesa). Al comprador se le ofrece una selección de fabricante por cada componente.

Receptor: Kenwood, Onkyo, Pioneer, Sony, Sherwood
Reproductor de discos compactos: Onkyo, Pioneer, Sony, Technics
Altavoces: Boston, Infinity, Polk
Tornamesa: Onkyo, Sony, Teac, Technics

Un tablero de distribución en la tienda le permite al cliente conectar cualquier selección de componentes (que consiste en uno de cada tipo). Use las reglas de producto para responder las siguientes preguntas.

- **a.** ¿De cuántas maneras puede ser seleccionado un componente de cada tipo?
- **b.** ¿De cuántas maneras pueden ser seleccionados los componentes, si tanto el receptor como el reproductor de discos compactos tienen que ser Sony?
- **c.** ¿De cuántas maneras pueden ser seleccionados los componentes, si ninguno tiene que ser Sony?
- **d.** ¿De cuántas maneras se puede hacer una selección, si se tiene que incluir al menos un componente Sony?
- **e.** Si alguien mueve los interruptores en el tablero de distribución completamente al azar, ¿cuál es la probabilidad de que el sistema seleccionado contenga al menos un componente Sony? ¿Y exactamente un componente Sony?

33. De nuevo considere el equipo de ligas pequeñas que tiene 15 jugadores en su plantel.

- **a.** ¿Cuántas formas existen de seleccionar 9 jugadores para la alineación inicial?
- **b.** ¿Cuántas formas existen de seleccionar 9 jugadores para la alineación inicial y un orden al bat de los nueve inicialistas?
- **c.** Suponga que 5 de los 15 jugadores son zurdos. ¿Cuántas formas existen de seleccionar 3 jardineros zurdos y tener las otras 6 posiciones ocupadas por jugadores derechos?

34. Las fallas en los teclados de computadora pueden ser atribuidas a defectos eléctricos o mecánicos. Un taller de reparación actualmente cuenta con 25 teclados averiados, de los cuales 6 tienen defectos eléctricos y 19 tienen defectos mecánicos.

- **a.** ¿Cuántas maneras hay de seleccionar al azar cinco de estos teclados para una inspección completa (sin tener en cuenta el orden)?
- **b.** ¿De cuántas maneras puede seleccionarse una muestra de 5 teclados, de manera que sólo dos tengan un defecto eléctrico?
- **c.** Si se selecciona al azar una muestra de 5 teclados, ¿cuál es la probabilidad de que al menos 4 de éstos tengan un defecto mecánico?

35. Una empresa productora emplea 10 trabajadores en el turno de día, 8 en el turno de tarde y 6 en el turno de medianoche. Un consultor de control de calidad seleccionará 5 de estos trabajadores para entrevistarlos. Suponga que la selección se hace de tal modo que cualquier grupo particular de 5 trabajadores tiene

la misma oportunidad de ser seleccionado, al igual que cualquier otro grupo (escogiendo 5 de entre 24 sin reemplazarlos).

a. ¿Cuántas selecciones resultarán en las que los 5 trabajadores seleccionados provengan del turno de día? ¿Cuál es la probabilidad de que los 5 trabajadores seleccionados sean del turno de día?

b. ¿Cuál es la probabilidad de que los 5 trabajadores seleccionados sean del mismo turno?

c. ¿Cuál es la probabilidad de que al menos dos turnos diferentes estén representados entre los trabajadores seleccionados?

d. ¿Cuál es la probabilidad de que al menos uno de los turnos no esté representado en la muestra de trabajadores?

36. Un departamento académico compuesto de cinco profesores limitó su selección para jefe de departamento al candidato *A* o el candidato *B*. Cada miembro votó por alguno de los candidatos. Suponga que en realidad existen tres votos para *A* y dos para *B*. Si los votos se cuentan al azar, ¿cuál es la probabilidad de que *A* se mantenga adelante de *B* durante todo el conteo de votos (p. ej. este evento ocurre si el orden seleccionado es *AABAB* pero no si es *ABBAA*)?

37. Un experimentador está estudiando los efectos de la temperatura, la presión y el tipo de catalizador en la producción de cierta reacción química. Se están considerando tres diferentes temperaturas, cuatro presiones distintas y cinco catalizadores diferentes.

a. Si cualquier experimento particular implica utilizar una temperatura, una presión y un catalizador, ¿cuántos experimentos son posibles?

b. ¿Cuántos experimentos hay que impliquen el uso de la temperatura más baja y dos presiones bajas?

c. Suponga que se tienen que realizar cinco experimentos diferentes el primer día de experimentación. Si los cinco se eligen al azar de entre todas las posibilidades, de modo que cualquier grupo de cinco tenga la misma probabilidad de selección, ¿cuál es la probabilidad de que se utilice un catalizador diferente en cada experimento?

38. Un soneto es un poema de 14 renglones en el que se siguen ciertos patrones de rimas. El escritor Raymond Queneau publicó un libro que sólo contiene 10 sonetos, cada uno en una página diferente. Sin embargo, estos están estructurados de modo que podrían crearse otros sonetos de la siguiente manera: el primer renglón de un soneto podría provenir del primer reglón de cualquiera de las 10 páginas, el segundo renglón podría provenir del segundo renglón de cualquiera de las 10 páginas y así sucesivamente (para este propósito se perforaban los renglones sucesivos).

a. ¿Cuántos sonetos pueden crearse a partir de los 10 del libro?

b. Si se selecciona al azar uno de los sonetos formados a partir del inciso a), ¿cuál es la probabilidad de que ninguno de sus renglones provenga del primer soneto o del último soneto del libro?

39. Una caja en un almacén contiene 15 focos compactos fluorescentes, de los cuales cinco son de 13 *watts*, seis de 18 y cuatro de 23. Suponga que se eligen al azar tres focos.

a. ¿Cuál es la probabilidad de que exactamente dos de los focos seleccionados sean de 23 *watts*?

b. ¿Cuál es la probabilidad de que los tres focos seleccionados sean de los mismos *watts*?

c. ¿Cuál es la probabilidad de que se seleccione un foco de cada tipo?

d. Suponga ahora que los focos tienen que ser seleccionados uno por uno hasta encontrar uno de 23 *watts*. ¿Cuál es la probabilidad de que sea necesario examinar al menos seis focos?

40. Tres moléculas de tipo *A*, tres de tipo *B*, tres de tipo *C* y tres de tipo *D* tienen que ser unidas para formar una cadena molecular. Una cadena molecular como esa es *ABCDABCDABCD* y otra es *BCDDAAABDBCC*.

a. ¿Cuántas moléculas de cadena hay? [*Sugerencia*: Si las tres *A* se distinguieran una de otra (A_1, A_2, A_3) y también las *B*, las *C* y las *D* ¿cuántas moléculas habría? ¿Cómo se reduce este número si se le quitan los subíndices a las *A*?]

b. Supongamos que se selecciona al azar una molécula de cadena del tipo descrito. ¿Cuál es la probabilidad de que las tres moléculas de cada tipo terminen una al lado de la otra (como en *BBBAAADDDCCC*)?

41. Un número de identificación personal para cajero automático (NIP) consta de cuatro cifras, cada una de 0, 1, 2,…, 8, o 9, en secuencia.

a. ¿Cuántos posibles NIP diferentes hay si no existen restricciones en la elección de dígitos?

b. De acuerdo con un representante en la sucursal local del autor del Chase Bank, hay restricciones en la elección de dígitos. La opción es que se prohíba lo siguiente: *i*) los cuatro dígitos idénticos; *ii*) las secuencias consecutivas de dígitos sean ascendentes o descendentes, como 6543; *iii*) cualquier secuencia de arranque con 19 (años de nacimiento son demasiado fáciles de adivinar). Así que si uno de los NIP en a) es seleccionado al azar, ¿cuál es la probabilidad de que sea uno legítimo (es decir, que no sea una de las secuencias prohibidas)?

c. Alguien ha robado una tarjeta de cajero automático y sabe que los dígitos primero y último del NIP son 8 y 1, respectivamente. Tiene tres intentos antes de que la tarjeta sea retenida por el cajero automático (pero no se da cuenta de eso). Así que selecciona al azar los dígitos 2º y 3º para el primer intento, a continuación selecciona al azar un par de dígitos diferentes para el segundo intento, y otro par de dígitos seleccionados al azar para el tercer intento (el individuo sabe acerca de las restricciones descritas en b) para seleccionar sólo de las posibilidades legítimas). ¿Cuál es la probabilidad de que el individuo tenga acceso a la cuenta?

d. Vuelva a calcular la probabilidad de c) si los dígitos primero y último son 1 y 1, respectivamente.

42. Una alineación titular en el baloncesto se compone de dos defensas, dos delanteros y un centro.

a. Un equipo de la universidad tiene en su lista tres centros, cuatro defensas, cuatro delanteros y un individuo (X) que

puede jugar de defensa o como delantero. ¿Cuántas alineaciones diferentes de inicio se pueden crear? [*Sugerencia*: Considere la posibilidad de alineaciones sin X, luego alineaciones con X como defensa, y a continuación alineaciones con X como delantero.]

b. Ahora supongamos que la lista tiene 5 defensas, 5 delanteros, 3 centros y 2 "jugadores comodín" (X y Y), que pueden jugar de defensas y de delanteros. Si 5 de los 15 jugadores son seleccionados al azar, ¿cuál es la probabilidad de que constituyan una alineación de inicio legítima?

43. En un juego de póker de cinco cartas, una corrida se compone de cinco cartas con denominaciones adyacentes (p. ej. 9 de tréboles, 10 de corazones, comodín de corazones, reina de espadas y rey de tréboles). Suponiendo que los ases pueden estar arriba o abajo, si le reparten una mano de cinco cartas ¿cuál es la probabilidad de que sea una corrida con un 10 como carta alta? ¿Cuál es la probabilidad de que sea una corrida del mismo palo?

44. Demuestre que $\binom{n}{k} = \binom{n}{n-k}$. Dé una interpretación que implique subconjuntos.

2.4 Probabilidad condicional

Las probabilidades asignadas a varios eventos dependen de lo que se sabe sobre la situación experimental cuando se hace la asignación. Enseguida a la asignación inicial puede llegar a estar disponible información parcial pertinente al resultado del experimento. Tal información puede hacer que se revisen algunas de las asignaciones de probabilidad. Para un evento particular A se ha utilizado $P(A)$ para representar la probabilidad asignada a A; ahora se considera $P(A)$ como la probabilidad original o no condicional del evento A.

En esta sección se examina cómo la información de que "un evento B ha ocurrido" afecta la probabilidad asignada a A. Por ejemplo, A podría referirse a un individuo que sufre una enfermedad particular en presencia de ciertos síntomas. Si se realiza un examen de sangre en el individuo y el resultado es negativo (B = examen de sangre negativo), entonces la probabilidad de que tenga la enfermedad cambiará (deberá reducirse, pero no a cero, puesto que los exámenes de sangre no son infalibles). Se utilizará la notación para representar la **probabilidad condicional de A puesto que el evento B ha ocurrido**. B es el "evento condicionante".

Por ejemplo, considere el evento A en el que un estudiante seleccionado al azar en su universidad obtuvo todas las clases deseadas durante el ciclo de inscripciones del semestre anterior. Presumiblemente $P(A)$ no es muy grande. Sin embargo, suponga que el estudiante seleccionado es un atleta con prioridad de inscripción especial (el evento B). Entonces $P(A|B)$ deberá ser sustancialmente más grande que $P(A)$, aunque quizás aún no cerca de 1.

EJEMPLO 2.24 En una planta se ensamblan componentes complejos en dos líneas de ensamble diferentes, A y A'. La línea A utiliza equipo más viejo que A', por lo que es un poco más lenta y menos confiable. Suponga que en un día dado la línea A ensambla 8 componentes, de los cuales 2 han sido identificados como defectuosos (B) y 6 como no defectuosos (B'), mientras que A' ha producido 1 componente defectuoso y 9 no defectuosos. Esta información se resume en la tabla siguiente.

		Condición	
		B	B'
Línea	A	2	6
	A'	1	9

Ajeno a esta información, el gerente de ventas selecciona al azar 1 de estos 18 componentes para una demostración. Antes de la demostración

$$P(\text{componente de la línea } A \text{ seleccionado}) = P(A) = \frac{N(A)}{N} = \frac{8}{18} = 0.44$$

No obstante, si el componente seleccionado resulta defectuoso, entonces el evento B ha ocurrido, por lo que el componente debe haber sido 1 de los 3 de la columna B de la tabla. Debido a que estos 3 componentes son igualmente probables entre sí una vez que B ha ocurrido,

$$P(A|B) = \frac{2}{3} = \frac{2/18}{3/18} = \frac{P(A \cap B)}{P(B)} \tag{2.2}$$

■

En la ecuación (2.2) la probabilidad condicional está expresada como una razón de probabilidades incondicionales. El numerador es la probabilidad de la intersección de los dos eventos, en tanto que el denominador es la probabilidad del evento condicionante B. Un diagrama de Venn ilustra esta relación (figura 2.8).

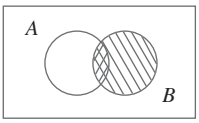


Figura 2.8 Motivación de la definición de probabilidad condicional

Dado que B ha ocurrido, el espacio muestral pertinente ya no es $\mathcal{S}$ pero consta de resultados en B; A ha ocurrido si y sólo si en la intersección ocurrió uno de los resultados, así que la probabilidad condicional de A dado B es proporcional a $P(A \cap B)$. Se utiliza la constante de proporcionalidad $1/P(B)$ para garantizar que la probabilidad $P(B/B)$ del nuevo espacio muestral B sea igual a 1.

Definición de probabilidad condicional

El ejemplo 2.24 demuestra que cuando los resultados son igualmente probables, el cálculo de probabilidades condicionales puede basarse en la intuición. Cuando los experimentos son más complicados, la intuición puede fallar, así que se requiere una definición general de probabilidad condicional que dé respuestas intuitivas en problemas simples. El diagrama de Venn y la ecuación (2.2) sugieren cómo proceder.

DEFINICIÓN

Para dos eventos cualesquiera A y B con $P(B) > 0$, la **probabilidad condicional de A dado que B** ha ocurrido está definida por

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.3}$$

EJEMPLO 2.25 Suponga que de todos los individuos que compran cierta cámara digital, 60% incluye en su compra una tarjeta de memoria opcional, 40% incluye una batería extra y 30%, tanto una tarjeta como una batería. Considere seleccionar al azar un comprador y sean $A = \{\text{tarjeta de memoria adquirida}\}$ y $B = \{\text{batería adquirida}\}$. Entonces $P(A) = 0.60$, $P(B) = 0.40$ y $P(\text{ambas adquiridas}) = P(A \cap B) = 0.30$. Dado que el individuo seleccionado adquirió una batería extra, la probabilidad de que una tarjeta opcional también sea adquirida es

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.30}{0.40} = 0.75$$

Es decir, de todos quienes adquieren una batería extra, 75% adquirió una tarjeta de memoria opcional. Asimismo,

$$P(\text{batería} \mid \text{tarjeta de memoria}) = P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.30}{0.60} = 0.50$$

Observe que $P(A|B) \neq P(A)$ y $P(B|A) \neq P(B)$. ■

El evento cuya probabilidad se desea podría ser una unión o intersección de otros eventos y lo mismo podría decirse del evento condicionante.

EJEMPLO 2.26 Una revista de noticias publica tres columnas tituladas "Arte" (A), "Libros" (B) y "Cine" (C). Los hábitos de lectura de un lector, seleccionado al azar, respecto a estas columnas son

Lee regularmente	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
Probabilidad	0.14	0.23	0.37	0.08	0.09	0.13	0.05

La figura 2.9 ilustra las probabilidades pertinentes.

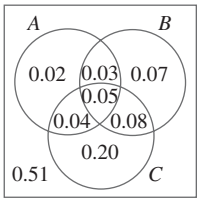


Figura 2.9 Diagrama de Venn para el ejemplo 2.26

Considere las siguientes cuatro probabilidades condicionales

(i) $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.08}{0.23} = 0.348$

(ii) La probabilidad de que el lector seleccionado lea regularmente la columna "Arte", dado que regularmente lee al menos una de las otras dos columnas, es

$$P(A|B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{0.04 + 0.05 + 0.03}{0.47} = \frac{0.12}{0.47} = 0.255$$

(iii) $P(A \mid \text{lee al menos una}) = P(A|A \cup B \cup C) = \dfrac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)}$

$$= \frac{P(A)}{P(A \cup B \cup C)} = \frac{0.14}{0.49} = 0.286$$

(iv) La probabilidad de que el lector seleccionado lea al menos una de las dos primeras columnas, dado que lee la columna de cine es

$$P(A \cup B|C) = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{0.04 + 0.05 + 0.08}{0.37} = 0.459$$ ■

Regla de multiplicación para $P(A \cap B)$

La definición de probabilidad condicional da el siguiente resultado, obtenido multiplicando ambos miembros de la ecuación (2.3) por $P(B)$.

La regla de multiplicación

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Esta regla es importante porque a menudo sucede que se desea $P(A \cap B)$, en tanto que $P(B)$ y $P(A|B)$ pueden ser especificadas a partir de la descripción del problema. La consideración de $P(B|A)$ da $P(A \cap B) = P(B|\text{A}) \cdot P(A)$.

EJEMPLO 2.27 Cuatro individuos han respondido a la solicitud de un banco de sangre para efectuar donaciones. Ninguno de ellos ha donado antes, por lo que sus tipos de sangre son desconocidos. Suponga que únicamente se desea el tipo O+ y sólo uno de los cuatro lo tiene. Si los donadores potenciales se seleccionan en orden aleatorio para determinar su tipo de sangre, ¿cuál es la probabilidad de que al menos tres individuos deban ser examinados para determinar su tipo de sangre y obtener el tipo deseado?

Al identificar $B = \{\text{primer tipo no O+}\}$ y $A = \{\text{segundo tipo no O+}\}$, $P(B) = 3/4$. Dado que el primer tipo no es O+, dos de los tres individuos que quedan tampoco son O+, por tanto $P(A|B) = 2/3$. La regla de multiplicación ahora da

$$\begin{aligned} P(\text{al menos tres individuos fueron examinados para determinar su tipo de sangre}) &= P(A \cap B) \\ &= P(A|B) \cdot P(B) \\ &= \frac{2}{3} \cdot \frac{3}{4} = \frac{6}{12} \\ &= 0.5 \end{aligned}$$

■

La regla de multiplicación es más útil cuando los experimentos se componen de varias etapas en secuencia. El evento condicionante B describe entonces el resultado de la primera etapa y A el resultado de la segunda, de modo que $P(A|B)$, condicionada en lo que ocurra primero, a menudo será conocida. La regla es fácil de ser ampliada a experimentos que implican más de dos etapas. Por ejemplo, considere los tres eventos A_1, A_2 y A_3. La triple intersección de estos eventos se puede representar como la doble intersección $(A_1 \cap A_2) \cap A_3$. Al aplicar la regla de la multiplicación a esta intersección y después a $A_1 \cap A_2$ se obtiene.

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_3|A_1 \cap A_2) \cdot P(A_1 \cap A_2) \\ &= P(A_3|A_1 \cap A_2) \cdot P(A_2|A_1) \cdot P(A_1) \end{aligned} \quad \textbf{(2.4)}$$

Así la probabilidad de la triple intersección es un producto de tres probabilidades, dos de las cuales son condicionales.

EJEMPLO 2.28 Para el experimento de determinación del tipo de sangre del ejemplo 2.27,

$$\begin{aligned} P(\text{el tercer tipo es O+}) &= P(\text{el tercero es}|\text{el primero no es} \cap \text{el segundo no es}) \\ &\quad \cdot P(\text{el segundo no es}|\text{el primero no es}) \cdot P(\text{el primero no es}) \\ &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{4} = 0.25 \end{aligned}$$

■

Cuando el experimento de interés se compone de una secuencia de varias etapas es conveniente representarlas con un diagrama de árbol. Una vez que se tiene un diagrama de árbol apropiado las probabilidades y las probabilidades condicionales pueden ingresarse en las diversas ramas; esto implica el uso repetido de la regla de multiplicación.

EJEMPLO 2.29 Una cadena de tiendas de video vende tres marcas diferentes de reproductores de DVD. De sus ventas de reproductores de DVD, 50% es de la marca 1 (la menos costosa), 30% de la marca 2 y 20% de la marca 3. Cada fabricante ofrece 1 año de garantía en las partes y en mano de obra. Se sabe que 25% de los reproductores de DVD de la marca 1 requiere trabajo de reparación dentro del periodo de garantía, mientras que los porcentajes correspondientes a las marcas 2 y 3 son 20 y 10%, respectivamente.

1. ¿Cuál es la probabilidad de que un comprador seleccionado al azar haya adquirido un reproductor de DVD marca 1 que necesitará reparación mientras se encuentra dentro del plazo de garantía?

2. ¿Cuál es la probabilidad de que un comprador seleccionado al azar haya comprado un reproductor de DVD que necesitará reparación mientras se encuentra dentro del plazo de garantía?

3. Si un cliente regresa a la tienda con un reproductor de DVD que necesita reparación dentro del plazo de garantía, ¿cuál es la probabilidad de que sea un reproductor de DVD marca 1? ¿Y uno marca 2? ¿Y uno marca 3?

La primera etapa del problema implica a un cliente que selecciona una de las tres marcas de reproductor de DVD. Sea A_i = {marca i adquirida}, con i = 1, 2 y 3. Entonces $P(A_1) = 0.50$, $P(A_2) = 0.30$ y $P(A_3) = 0.20$. Una vez que se selecciona una marca de reproductor de DVD, la segunda etapa implica observar si el reproductor seleccionado necesita reparación dentro del plazo de garantía. Con B = {necesita reparación} y B' = {no necesita reparación}, la información dada implica que $P(B|A_1) = 0.25$, $P(B|A_2) = 0.20$ y $P(B|A_3) = 0.10$.

El diagrama de árbol que representa esta situación experimental se muestra en la figura 2.10. Las ramas iniciales corresponden a marcas diferentes de reproductores de DVD; hay dos ramas de segunda generación que emanan de la punta de cada rama inicial, una para "necesita reparación" y la otra para "no necesita reparación". La probabilidad $P(A_i)$ aparece en la rama i-ésima inicial, en tanto que las probabilidades condicionales $P(B|A_i)$ y $P(B'|A_i)$ aparecen en las ramas de la segunda generación. A la derecha de cada rama de segunda generación, correspondiente a la ocurrencia de B, se muestra el producto de probabilidades en las ramas que conducen hacia fuera de dicho punto. Esta es simplemente la regla de multiplicación en acción. La respuesta a la pregunta planteada en 1 es, por tanto, $P(A_1 \cap B) = P(B|A_1) \cdot P(A_1) = 0.125$. La respuesta a la pregunta 2 es

$$\begin{aligned} P(B) &= P[\text{(marca 1 y reparación) o (marca 2 y reparación) o (marca 3 y reparación)}] \\ &= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) \\ &= 0.125 + 0.060 + 0.020 = 0.205 \end{aligned}$$

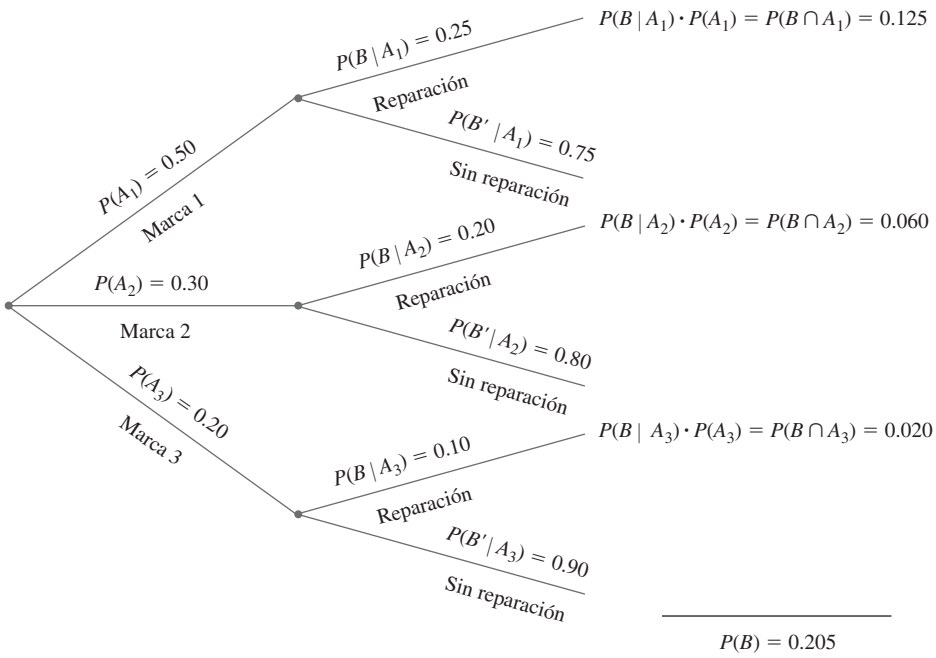


Figura 2.10 Diagrama de árbol para el ejemplo 2.29

Finalmente,

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{0.125}{0.205} = 0.61$$
$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{0.060}{0.205} = 0.29$$

y

$$P(A_3|B) = 1 - P(A_1|B) - P(A_2|B) = 0.10$$

La *probabilidad previa* o inicial de la marca 1 es 0.50. Una vez que se sabe que el reproductor de DVD seleccionado necesitó reparación, la *probabilidad posterior* de la marca 1 se incrementa a 0.61. Esto se debe a que es más probable que los reproductores marca 1 necesiten reparación de garantía que las demás marcas. La probabilidad posterior de la marca 3 es $P(A_3|B) = -0.10$, la cual es mucho menor que la probabilidad previa $P(A_3) = 0.20$. ■

Teorema de Bayes

El cálculo de una probabilidad posterior $P(A_j|B)$ a partir de probabilidades previas dadas $P(A_i)$ y probabilidades condicionales $P(B|A_i)$ ocupa una posición central en la probabilidad elemental. La regla general de dichos cálculos, los que en realidad son una aplicación simple de la regla de multiplicación, se remonta al reverendo Thomas Bayes quien vivió en el siglo XVIII. Para formularla primero se requiere otro resultado. Recuerde que los eventos $A_1,\ldots, A_k$ son mutuamente excluyentes si ninguno de los dos tiene resultados comunes. Los eventos son *exhaustivos* si un A_i debe ocurrir, de modo que $A_1 \cup \ldots \cup A_k = \mathcal{S}$.

Ley de probabilidad total

Sean $A_1,\ldots, A_k$ eventos mutuamente excluyentes y exhaustivos. Así, para cualquier otro evento B,

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k) \\ &= \sum_{i=1}^{k} P(B|A_i)P(A_i) \end{aligned} \quad \textbf{(2.5)}$$

Demostración Puesto que los eventos A_i son mutuamente excluyentes y exhaustivos, si ocurre B debe ser en forma conjunta con uno de los eventos A_i exactamente. Es decir, $B = (A_1 \cap B) \cup \ldots \cup (A_k \cap B)$, donde los eventos $(A_i \cap B)$ son mutuamente excluyentes. Esta "partición de B" se ilustra en la figura 2.11. Por tanto,

$$P(B) = \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

como se deseaba.

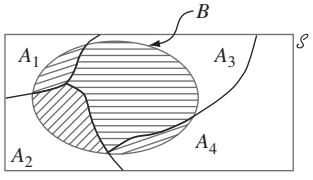


Figura 2.11 Partición de B entre las A_i' mutuamente excluyentes y exhaustivas ■

EJEMPLO 2.30 Una persona tiene 3 cuentas de correo electrónico diferentes. La mayoría de sus mensajes, de hecho 70%, entra en la cuenta #1, mientras que 20% entra en la cuenta #2 y el restante 10%, en la cuenta #3. De los mensajes en la cuenta #1 sólo 1% es *spam* (correo no deseado), mientras que los porcentajes correspondientes a las cuentas #2 y #3 son 2 y 5%, respectivamente. ¿Cuál es la probabilidad de que un mensaje *spam* sea seleccionado al azar?

Para responder esta pregunta, primero establecemos una notación:

$$A_i = \{\text{el mensaje es de la cuenta \#}i\} \text{ para } i = 1, 2, 3, \; B = \{\text{el mensaje es } spam\}$$

Por tanto, los porcentajes dados implican que

$$P(A_1) = 0.70,\ P(A_2) = 0.20,\ P(A_3) = 0.10$$
$$P(B \mid A_1) = 0.01,\ P(B \mid A_2) = 0.02,\ P(B \mid A_3) = 0.05$$

Ahora simplemente es cuestión de sustituir en la ecuación de la ley de probabilidad total:

$$P(B) = (0.01)(0.70) + (0.02)(0.20) + (0.05)(0.10) = 0.016$$

A largo plazo, 1.6% de los mensajes de esta persona serán *spam*. ■

Teorema de Bayes

Sean $A_1, A_2, \ldots, A_k$ un conjunto de eventos mutuamente excluyentes y exhaustivos con probabilidades *previas* $P(A_i)(i = 1, \ldots, k)$. Entonces para cualquier otro evento B para el cual $P(B) > 0$, la probabilidad *posterior* de A_j dado que ha ocurrido B es

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} \quad j = 1, \ldots, k \qquad \textbf{(2.6)}$$

La transición de la segunda a la tercera expresión en (2.6) se apoya en el uso de la regla de multiplicación en el numerador, y la ley de probabilidad total en el denominador. La proliferación de eventos y subíndices en (2.6) puede ser un poco intimidante para los novatos en el tema de la probabilidad. Mientras existan relativamente pocos eventos en la repartición se puede utilizar un diagrama de árbol (como en el ejemplo 2.29) como base para calcular probabilidades posteriores sin jamás referirse de manera explícita al teorema de Bayes.

EJEMPLO 2.31 *Incidencia de una enfermedad rara*. Sólo 1 de 1000 adultos padece una enfermedad rara para la cual se ha creado una prueba de diagnóstico. La prueba es tal que cuando un individuo padece realmente la enfermedad, un resultado positivo se presentará en 99% de las veces mientras que en individuos sin la enfermedad el examen será positivo sólo 2% de las ocasiones (la *sensibilidad* de esta prueba es de 99% y la *especificidad* es de 98%; en contraste, el **tema en el reporte de septiembre 22 de 2012 de *The Lancet*** informa que la primera prueba del VIH para hacer en casa tiene una sensibilidad de sólo 92% y una especificidad de 99.98%). Si un individuo seleccionado al azar se somete a la prueba y el resultado es positivo, ¿cuál es la probabilidad de que el padezca la enfermedad?

Para utilizar el teorema de Bayes, sea A_1 = el individuo padece la enfermedad, A_2 = el individuo no tiene la enfermedad y B = resultado de prueba positivo. Entonces $P(A_1) = 0.001$, $P(A_2) = 0.999$, $P(B \mid A_1) = 0.99$ y $P(B \mid A_2) = 0.02$. El diagrama de árbol para este problema se muestra en la figura 2.12.

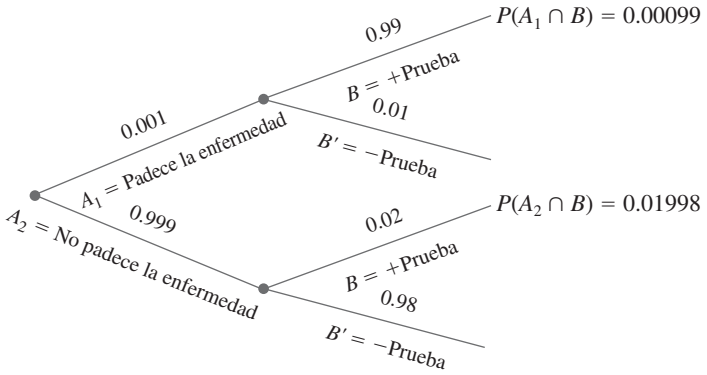


Figura 2.12 Diagrama de árbol para el problema de una enfermedad rara

Junto a cada rama correspondiente a un resultado positivo de prueba, la regla de multiplicación da las probabilidades anotadas. Por consiguiente, $P(B) = 0.00099 + 0.01998 = 0.02097$ a partir de la cual se tiene

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{0.00099}{0.02097} = 0.47$$

Este resultado parece contraintuitivo; la prueba de diagnóstico parece tan precisa que es altamente probable que alguien con un resultado positivo en la prueba padezca la enfermedad, mientras que la probabilidad condicional calculada es de sólo 0.047. Sin embargo, puesto que la enfermedad es rara y la prueba es sólo moderadamente confiable, surgen más resultados de prueba positivos debido a errores y no de individuos enfermos. La probabilidad de tener la enfermedad se ha incrementado por un factor de multiplicación de 47 (desde la probabilidad previa de 0.001 hasta la probabilidad posterior de 0.047), pero para incrementar aún más la probabilidad posterior se requiere una prueba de diagnóstico con tasas de error mucho más pequeñas. ■

EJERCICIOS **Sección 2.4 (45–69)**

45. La población de un país particular se compone de tres grupos étnicos. Cada individuo pertenece a uno de los cuatro grupos sanguíneos principales. La siguiente tabla de *probabilidad conjunta* da la proporción de individuos en las diversas combinaciones de grupo étnico–grupo sanguíneo.

		Grupo sanguíneo			
		O	**A**	**B**	**AB**
Grupo étnico	**1**	0.082	0.106	0.008	0.004
	2	0.135	0.141	0.018	0.006
	3	0.215	0.200	0.065	0.020

Suponga que se selecciona al azar a un individuo de la población y que los eventos se definen como A = {tipo A seleccionado}, B = {tipo B seleccionado} y C = {grupo étnico 3 seleccionado}.

a. Calcule $P(A)$, $P(C)$ y $P(A \cap C)$.

b. Calcule $P(A \mid C)$ y $P(C \mid A)$ y explique en contexto lo que cada una de estas probabilidades representa.

c. Si el individuo seleccionado no tiene sangre de tipo B, ¿cuál es la probabilidad de que pertenezca al grupo étnico 1?

46. Suponga que un individuo es seleccionado al azar de la población de todos los adultos varones que viven en los Estados Unidos. Sea A el evento en que el individuo seleccionado tiene una estatura de más de 6 pies, y sea B el evento en que el individuo seleccionado es un jugador profesional de basquetbol. ¿Cuál piensa que es más grande, $P(A \mid B)$ o $P(B \mid A)$? ¿Por qué?

47. Regrese al escenario de la tarjeta de crédito del ejercicio 12 (sección 2.2), y sea C el evento de que el estudiante seleccionado tiene una tarjeta American Express. Además de $P(A) = 0.6$, $P(B) = 0.4$ y $P(A \cap B) = 0.3$, suponga que $P(C) = 0.2$, $P(A \cap C) = 0.15$, $P(B \cap C) = 0.1$ y $P(A \cap B \cap C) = 0.08$.

a. ¿Cuál es la probabilidad de que el alumno seleccionado tenga al menos uno de los tres tipos de tarjetas?

b. ¿Cuál es la probabilidad de que el alumno seleccionado tenga tanto una tarjeta Visa como una MasterCard, pero no una tarjeta American Express?

c. Calcule e interprete $P(B \mid A)$ y también $P(A \mid B)$.

d. Si sabemos que el alumno seleccionado tiene una tarjeta American Express, ¿cuál es la probabilidad de que también tenga una tarjeta Visa y una MasterCard?

e. Dado que el alumno seleccionado tiene una tarjeta American Express, ¿cuál es la probabilidad que tenga al menos una de los otros dos tipos de tarjetas?

48. Reconsidere la situación del sistema defectuoso descrito en el ejercicio 26 (sección 2.2).

a. Dado que el sistema tiene un defecto de tipo 1, ¿cuál es la probabilidad de que tenga un defecto de tipo 2?

b. Dado que el sistema tiene un defecto de tipo 1, ¿cuál es la probabilidad de que tenga los tres tipos de defecto?

c. Dado que el sistema tiene al menos un tipo de defecto, ¿cuál es la probabilidad de que tenga exactamente un particular tipo de defecto?

d. Dado que el sistema tiene los primeros dos tipos de defecto, ¿cuál es la probabilidad de que no tenga el tercer tipo de defecto?

49. La siguiente tabla proporciona información sobre el tipo de café seleccionado por alguien que compra una taza de café en un kiosco del aeropuerto en particular

	Pequeño	Mediano	Grande
Regular	14%	20%	26%
Descafeinado	20%	10%	10%

Considere la posibilidad de seleccionar al azar a un comprador de café.

a. ¿Cuál es la probabilidad de que la persona adquiera una taza pequeña? ¿Y una taza de café descafeinado?

b. Si nos enteramos de que la persona seleccionada compra una taza de café pequeña, ¿cuál es ahora la probabilidad de que escoja el café descafeinado y cómo se interpreta esta probabilidad?

c. Si nos enteramos de que el individuo seleccionado compró un café descafeinado, ¿cuál es ahora la probabilidad de que haya escogido un tamaño pequeño, y cómo se compara esto con la probabilidad incondicional correspondiente de a)?

50. Una tienda de departamentos vende camisetas deportivas en tres tallas (pequeña, mediana y grande), tres diseños (a cuadros, estampadas y a rayas) y dos largos de manga (larga y corta). Las siguientes tablas dan las proporciones de camisetas vendidas en diferentes combinaciones de cada categoría.

Manga corta

	Diseño		
Talla	A cuadros	Estampada	A rayas
Ch	0.04	0.02	0.05
M	0.08	0.07	0.12
G	0.03	0.07	0.08

Manga larga

	Diseño		
Talla	A cuadros	Estampada	A rayas
Ch	0.03	0.02	0.03
M	0.10	0.05	0.07
G	0.04	0.02	0.08

a. ¿Cuál es la probabilidad de que la siguiente camiseta vendida sea mediana, estampada y de manga larga?

b. ¿Cuál es la probabilidad de que la siguiente camiseta vendida sea estampada y mediana?

c. ¿Cuál es la probabilidad de que la siguiente camiseta vendida sea de manga corta? ¿Y de manga larga?

d. ¿Cuál es la probabilidad de que la talla de la siguiente camiseta vendida sea mediana? ¿Y de que la siguiente camiseta vendida sea estampada?

e. Dado que la camiseta que se acaba de vender es de manga corta a cuadros, ¿cuál era la probabilidad de que fuera mediana?

f. Dado que la camiseta que se acaba de vender era mediana a cuadros, ¿cuál era la probabilidad de que fuera de manga corta? ¿Y de manga larga?

51. De acuerdo con lo publicado el 31 de julio de 2013 en **cnn.com,** después de la muerte de un niño que probó un cacahuate, un estudio de 2010 en la revista ***Pediatrics*** encontró que 8% de los menores de 18 años en los Estados Unidos tienen al menos una alergia alimentaria. Entre las personas con alergia a algún alimento, aproximadamente 39% tenía antecedentes de reacciones graves.

a. Si un menor de 18 años es seleccionado al azar, ¿cuál es la probabilidad de sea alérgico al menos a un alimento y con antecedentes de reacción grave?

b. También se informa que 30% de las personas con alergia en realidad es alérgico a múltiples alimentos. Si un menor de 18 años es seleccionado al azar, ¿cuál es la probabilidad que sea alérgico a múltiples alimentos?

52. Un sistema se compone de bombas idénticas, #1 y #2. Si una falla, el sistema seguirá operando. Sin embargo, debido al esfuerzo adicional, ahora es más probable que antes que la bomba restante falle. Es decir, $r = P(\#2 \text{ falla} \mid \#1 \text{ falla}) > P(\#2 \text{ falla}) = q$. Si al menos una bomba falla alrededor del final de su vida útil en 7% de todos los sistemas, y ambas bombas fallan durante dicho periodo en sólo 1%, ¿cuál es la probabilidad de que la bomba #1 falle durante su vida útil?

53. Un taller repara componentes de audio y de video. Sea A el evento en que el siguiente componente que se recibe para su reparación es un componente de audio, y sea B el evento en que el siguiente componente que se recibe sea un reproductor de discos compactos (así que el evento B está contenido en A). Suponga que $P(A) = 0.6$ y $P(B) = 0.05$. ¿Cuál es $P(B|A)$?

54. En el ejercicio 13, $A_i = \{\text{proyecto otorgado } i\}$, con $i = 1, 2, 3$. Use las probabilidades dadas en dicho ejercicio para calcular

las siguientes probabilidades y explique el significado de cada una.

a. $P(A_2 \mid A_1)$ **b.** $P(A_2 \cap A_3 \mid A_1)$

c. $P(A_2 \cup A_3 \mid A_1)$ **d.** $P(A_1 \cap A_2 \cap A_3 \mid A_1 \cup A_2 \cup A_3)$

55. Las garrapatas del venado pueden ser portadoras de la enfermedad de Lyme o de la erliquiosis granulocítica humana (HGE, por sus siglas en inglés). Con base en un estudio reciente, suponga que 16% de todas las garrapatas en cierto lugar portan la enfermedad de Lyme, 10% portan HGE y 10% de las garrapatas que portan al menos una de estas enfermedades en realidad porta las dos. Si se determina que una garrapata seleccionada al azar ha sido portadora de HGE, ¿cuál es la probabilidad de que la garrapata seleccionada también porte la enfermedad de Lyme?

56. Para los eventos A y B con $P(B) > 0$, demuestre que $P(A|B) + P(A'|B) = 1$.

57. Si $P(B|A) > P(B)$, demuestre que $P(B'|A) < P(B')$. [*Sugerencia*: Sume $P(B'|A)$ a ambos lados de la desigualdad dada y luego utilice el resultado del ejercicio 56.]

58. Demuestre que para tres eventos cualesquiera A, B y C con $P(C) > 0$, $P(A \cup B \mid C) = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$.

59. En una gasolinera, 40% de los clientes utiliza gasolina regular (A_1), 35% usa gasolina plus (A_2) y 25% utiliza premium (A_3). De los clientes que utilizan gasolina regular, sólo 30% llena su tanque (evento B). De los clientes que utilizan plus, 60% llena su tanque, mientras que los que utilizan premium, 50% llena su tanque.

a. ¿Cuál es la probabilidad de que el siguiente cliente pida gasolina plus y llene el tanque ($A_2 \cap B$)?

b. ¿Cuál es la probabilidad de que el siguiente cliente llene el tanque?

c. Si el siguiente cliente llena el tanque, ¿cuál es la probabilidad de que pida gasolina regular? ¿Y de que pida plus? ¿Y de que pida premium?

60. De las aeronaves ligeras que desaparecen en vuelo en cierto país 70% es localizada posteriormente. De las aeronaves que son localizadas, 60% cuenta con un localizador de emergencia, mientras que 90% de las aeronaves no localizadas no cuenta con dicho localizador. Suponga que una aeronave ligera ha desaparecido.

a. Si tiene un localizador de emergencia, ¿cuál es la probabilidad de que no sea localizada?

b. Si no tiene un localizador de emergencia, ¿cuál es la probabilidad de que sí sea localizada?

61. Componentes de cierto tipo son enviados a un distribuidor en lotes de diez. Suponga que 50% de dichos lotes no contiene componentes defectuosos, 30% contiene un componente defectuoso y 20% contiene dos componentes defectuosos. Se seleccionan al azar dos componentes de un lote y se prueban. ¿Cuáles son las probabilidades asociadas con 0, 1 y 2 componentes defectuosos, que están en el lote, en cada una de las siguientes condiciones?

a. Ningún componente probado está defectuoso.

b. Uno de los dos componentes probados está defectuoso. [*Sugerencia*: Trace un diagrama de árbol con tres ramas de primera generación correspondientes a los tres tipos diferentes de lotes.]

62. La compañía Taxi Azul opera 15% de los taxis en cierta ciudad, y Taxi Verde opera el otro 85%. Luego de un accidente en la noche con un taxi, un testigo dijo que el vehículo era azul. Supongamos, sin embargo, que bajo condiciones de visión nocturna, sólo 80% de los individuos puede distinguir correctamente entre un vehículo de color azul y uno verde. ¿Cuál es la probabilidad (posterior) de que la culpa haya sido de un taxi azul? En su respuesta, asegúrese de indicar qué reglas de probabilidad está utilizando. [*Sugerencia*: Un diagrama de árbol podría ayudar. *Nota*: Este ejercicio se basa en un incidente real.]

63. Para los clientes que compran un refrigerador en una tienda de aparatos domésticos, sea A el evento en que el refrigerador fue fabricado en los Estados Unidos, B el evento en que el refrigerador contaba con una máquina de hacer hielos y C el evento en que el cliente adquirió una garantía ampliada. Las probabilidades pertinentes son

$P(A) = 0.75$ $P(B \mid A) = 0.9$ $P(B \mid A') = 0.8$

$P(C \mid A \cap B) = 0.8$ $P(C \mid A \cap B') = 0.6$

$P(C \mid A' \cap B) = 0.7$ $P(C \mid A' \cap B') = 0.3$

a. Construya un diagrama de árbol compuesto de ramas de primera, segunda y tercera generaciones y anote el evento y la probabilidad apropiados junto a cada rama.

b. Calcule $P(A \cap B \cap C)$.

c. Calcule $P(B \cap C)$.

d. Calcule $P(C)$.

e. Calcule $P(A \mid B \cap C)$, la probabilidad de la compra de un refrigerador fabricado en los Estados Unidos, dado que también se adquirieron una máquina de hacer hielos y una garantía ampliada.

64. El editor de comentarios de una cierta revista científica decide si la revisión de cualquier libro en particular debe ser corta (1–2 páginas), mediana (3–4 páginas) o larga (5–6 páginas). Los datos sobre estudios recientes indican que 60% de ellas son cortas, 30% son medianas y el restante 10% son largas. Los comentarios están presentados en Word o LaTeX. Para las revisiones cortas, 80% está en Word, mientras que 50% de las revisiones medianas y 30% de las revisiones largas están también en Word. Supongamos que se selecciona aleatoriamente una revisión reciente.

a. ¿Cuál es la probabilidad de que la revisión seleccionada se presente en formato Word?

b. Si la revisión seleccionada se presenta en formato Word, ¿cuáles son las probabilidades posteriores de que sea corta, mediana o larga?

65. Un gran operador de complejos de tiempo compartido requiere que cualquier persona interesada en hacer una compra primero visite el sitio de interés. Los datos históricos indican que 20% de todos los compradores potenciales seleccionaron un día de visita, 50% eligió una visita de una noche y 30% optó por una visita de dos noches. Además, 10% de los visitantes de un día en última instancia, hizo una compra, 30% de los visitantes de una noche compró una unidad y 20% de los visitantes de dos noches decidió comprar. Supongamos que un visitante es seleccionado al azar y se demuestra que ha realizado una compra. ¿Qué tan probable es que esta persona haya realizado una visita de un día? ¿Y una visita de una noche? ¿Y una visita de dos noches?

66. Considere la siguiente información sobre vacacionistas (basada en parte en una encuesta reciente de Travelocity): 40% revisa su correo electrónico de trabajo, 30% utiliza un teléfono celular para permanecer en contacto con su trabajo, 25% lleva una computadora portátil consigo, 23% revisa su correo electrónico de trabajo y utiliza un teléfono celular para permanecer en contacto, y 51% no revisa su correo electrónico de trabajo ni utiliza un teléfono celular para permanecer en contacto y tampoco lleva consigo una computadora portátil. Además, 88 de cada 100 que llevan una computadora portátil también revisan su correo electrónico de trabajo y 70 de cada 100 que utilizan un teléfono celular para permanecer en contacto también llevan una computadora portátil.

a. ¿Cuál es la probabilidad de que un vacacionista seleccionado al azar que revisa su correo electrónico de trabajo también utilice un teléfono celular para permanecer en contacto?

b. ¿Cuál es la probabilidad de que alguien que lleva una computadora portátil también utilice un teléfono celular para permanecer en contacto?

c. Si el vacacionista seleccionado al azar revisa su correo electrónico de trabajo y lleva una computadora portátil, ¿cuál es la probabilidad de que utilice un teléfono celular para permanecer en contacto?

67. Ha habido gran controversia durante los últimos años respecto a qué tipos de vigilancia son apropiados para prevenir el terrorismo. Suponga que un sistema de vigilancia particular tiene 99% de probabilidades de identificar correctamente a un futuro terrorista y 99.9% de probabilidades de identificar correctamente a alguien que no es un futuro terrorista. Si existen 1000 futuros terroristas en una población de 300 millones y se selecciona al azar un individuo de estos 300 millones, que es examinado por el sistema e identificado como futuro terrorista, ¿cuál es la probabilidad de que en realidad sea un futuro terrorista? ¿Le inquieta el valor de esta probabilidad sobre el uso del sistema de vigilancia? Explique.

68. Una amiga que vive en Los Ángeles hace frecuentes viajes de consultoría a Washington, D.C.; 50% del tiempo viaja con la línea aérea #1, 30% del tiempo con la aerolínea #2 y el restante 20% con la aerolínea #3. Los vuelos de la aerolínea #1 llegan demorados a D.C. 30% del tiempo y 10% del tiempo llegan demorados a L.A. Para la aerolínea #2 estos porcentajes son 25 y 20%, respectivamente; en tanto que para la aerolínea #3 los porcentajes son 40 y 25% respectivamente. Si se sabe que en un viaje particular nuestra amiga llegó demorada a exactamente uno de los dos destinos, ¿cuáles son las probabilidades posteriores de que haya volado con las aerolíneas #1, #2 y #3? Suponga que la probabilidad de arribar con demora a L.A. no se ve afectada por lo que suceda en el vuelo a D.C. [*Sugerencia*: Desde la punta de cada rama de primera generación en un diagrama de árbol, trace tres ramas de segunda generación identificadas, respectivamente, como 0 demorado, 1 demorado y 2 demorado.]

69. En el ejercicio 59 considere la siguiente información adicional sobre el uso de tarjetas de crédito:

70% de todos los clientes que utilizan gasolina regular y que llenan el tanque usa una tarjeta de crédito.

50% de todos los clientes que utilizan gasolina regular y que no llenan el tanque usa una tarjeta de crédito.

60% de todos los clientes que llenan el tanque con gasolina plus usa una tarjeta de crédito.

50% de todos los clientes que utilizan gasolina plus y que no llenan el tanque usa una tarjeta de crédito.

50% de todos los clientes que utilizan gasolina premium y que llenan el tanque usan una tarjeta de crédito.

40% de todos los clientes que utilizan gasolina premium y que no llenan el tanque usa una tarjeta de crédito.

Calcule la probabilidad de cada uno de los siguientes eventos para el siguiente cliente que llegue (un diagrama de árbol podría ayudar).

a. {plus, tanque lleno y tarjeta de crédito}

b. {premium, tanque no lleno y tarjeta de crédito}

c. {premium y tarjeta de crédito}

d. {tanque lleno y tarjeta de crédito}

e. {tarjeta de crédito}

f. Si el siguiente cliente utiliza una tarjeta de crédito, ¿cuál es la probabilidad de que pida premium?

2.5 Independencia

La definición de probabilidad condicional permite revisar la probabilidad $P(A)$ originalmente asignada a A cuando después se nos informa que otro evento B ha ocurrido; la nueva probabilidad de A es $P(A \mid B)$. En nuestros ejemplos, con frecuencia se dio el caso de que $P(A \mid B)$ difería de la probabilidad no condicional $P(A)$. Luego la información "B ha ocurrido" cambia la probabilidad de que ocurra A. A menudo la probabilidad de que ocurra o haya ocurrido A no se ve afectada por el conocimiento de que B ha ocurrido, así que $P(A \mid B) = P(A)$. Es entonces natural considerar a A y B como eventos independientes, es decir que la ocurrencia o no ocurrencia de un evento no afecta la probabilidad de que el otro evento ocurra.

DEFINICIÓN

Los eventos A y B son **independientes** si $P(A \mid B) = P(A)$; y **dependientes** en caso contrario.

La definición de independencia podría parecer "no simétrica" porque no pedimos que $P(B|A) = P(B)$. Sin embargo, utilizando la definición de probabilidad condicional y la regla de multiplicación,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{2.7}$$

El lado derecho de la ecuación (2.7) es $P(B)$ si y sólo si $P(A|B) = P(A)$ (independencia), así que la igualdad en la definición implica la otra igualdad (y viceversa). También es fácil demostrar que si A y B son independientes, entonces también lo son los pares de eventos: 1) A' y B, 2) A y B' y 3) A' y B'.

EJEMPLO 2.32 Considere una gasolinera con seis bombas numeradas 1, 2,…, 6, y sea E_i el evento simple en que un cliente seleccionado al azar utiliza la bomba i ($i = 1,\ldots, 6$). Suponga que

$$P(E_1) = P(E_6) = 0.10, \quad P(E_2) = P(E_5) = 0.15, \quad P(E_3) = P(E_4) = 0.25$$

Defina los eventos A, B, C como

$$A = \{2, 4, 6\}, B = \{1, 2, 3\}, C = \{2, 3, 4, 5\}.$$

Luego se tiene $P(A) = 0.50$, $P(A \mid B) = 0.30$ y $P(A \mid C) = 0.50$. Es decir, los eventos A y B son dependientes, en tanto que los eventos A y C son independientes. Intuitivamente, A y C son independientes porque la división de probabilidad relativa entre las bombas pares e impares es la misma entre las bombas 2, 3, 4, 5, como lo es entre todas las seis bombas. ■

EJEMPLO 2.33 Sean A y B dos eventos mutuamente excluyentes cualesquiera con $P(A) > 0$. Por ejemplo, para un automóvil seleccionado al azar, sea $A = \{$el carro es de cuatro cilindros$\}$ y $B = \{$el carro es de seis cilindros$\}$. Como los eventos son mutuamente excluyentes, si ocurre B, entonces A quizá puede no haber ocurrido, así que $P(A \mid B) = 0 \neq P(A)$. El mensaje aquí es que *si dos eventos son mutuamente excluyentes, no pueden ser independientes*. Cuando A y B son mutuamente excluyentes la información de que ocurrió A dice algo sobre B (no puede haber ocurrido), así que se impide la independencia. ■

Regla de multiplicación para $P(A \cap B)$

Con frecuencia la naturaleza de un experimento sugiere que dos eventos A y B deben ser supuestos independientes. Este es el caso, por ejemplo, si un fabricante recibe una tarjeta de circuito de cada uno de dos proveedores diferentes, cada tarjeta se somete a prueba al llegar y $A = \{$la primera está defectuosa$\}$ y $B = \{$la segunda está defectuosa$\}$. Si $P(A) = 0.1$, también deberá ser el caso de que $P(A \mid B) = 0.1$; sabiendo la condición de la segunda tarjeta no informa sobre la condición de la primera. La probabilidad de que ambos eventos ocurran se calcula fácilmente a partir de la probabilidad individual de los eventos cuando estos son independientes.

PROPOSICIÓN

A y B son independientes si y sólo si

$$P(A \cap B) = P(A) \cdot P(B) \tag{2.8}$$

La verificación de esta regla de multiplicación es como sigue:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B) \tag{2.9}$$

donde la segunda igualdad en la ecuación (2.9) es válida si y sólo si A y B son independientes. Debido a la equivalencia de independencia y a la ecuación (2.8), la segunda puede ser utilizada como definición de independencia.

EJEMPLO 2.34 Se sabe que 30% de las lavadoras de cierta compañía requieren servicio mientras se encuentran dentro del plazo de garantía, en tanto que sólo 10% de sus secadoras necesita dicho servicio. Si alguien adquiere una lavadora y una secadora fabricadas por esta compañía, ¿cuál es la probabilidad de que ambas máquinas requieran servicio de garantía?

Sea A el evento en que la lavadora necesita servicio mientras se encuentra dentro del plazo de garantía, y defina B de igual forma para la secadora. Entonces $P(A) = 0.30$ y $P(B) = 0.10$. Suponiendo que las dos máquinas funcionan independientemente una de otra, la probabilidad deseada es

$$P(A \cap B) = P(A) \cdot P(B) = (0.30)(0.10) = 0.03$$ ■

Es fácil demostrar que A y B son independientes si y sólo si A' y B son independientes, A y B' son independientes y A' y B' son independientes. Por tanto, en el ejemplo 2.34 la probabilidad de que ninguna máquina necesite servicio es

$$P(A' \cap B') = P(A') \cdot P(B') = (0.70)(0.90) = 0.63$$

EJEMPLO 2.35 Cada día, de lunes a viernes, un lote de componentes enviado por un primer proveedor llega a una instalación de inspección. Dos días a la semana, también llega un lote de un segundo proveedor. Ochenta por ciento de todos los lotes del proveedor 1 son inspeccionados y 90% de los del proveedor 2 también. ¿Cuál es la probabilidad de que, en un día seleccionado al azar, dos lotes sean inspeccionados? Esta pregunta se responderá suponiendo que en los días en que se inspeccionan dos lotes, si el primer lote pasa es independiente de si también pasa el segundo. La figura 2.13 muestra la información relevante.

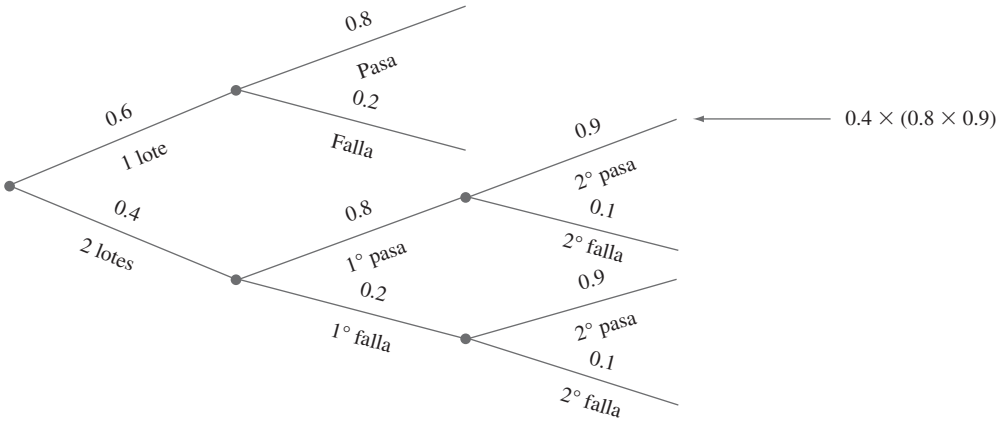


Figura 2.13 Diagrama de árbol para el ejemplo 2.35

$$\begin{aligned} P(\text{dos pasan}) &= P(\text{dos recibidos} \cap \text{ambos pasan}) \\ &= P(\text{ambos pasan} \mid \text{dos recibidos}) \cdot P(\text{dos recibidos}) \\ &= [(0.8)(0.9)](0.4) = 0.288 \end{aligned}$$ ■

Independencia de más de dos eventos

El concepto de independencia de dos eventos puede ser ampliada a conjuntos de más de dos eventos. Aunque es posible ampliar la definición para dos eventos independientes trabajando

en función de probabilidades condicionales y no condicionales, es más directo y menos tedioso seguir las líneas de la última proposición.

DEFINICIÓN

Los eventos $A_1,\ldots,A_n$ son **mutuamente independientes** si por cada k ($k = 2, 3,\ldots, n$) y cada subconjunto de índices $i_1, i_2,\ldots,i_k$,

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_1}) \cdot \cdots \cdot P(A_{i_k})$$

Parafraseando la definición, los eventos son mutuamente independientes si la probabilidad de la intersección de cualquier subconjunto de los n eventos es igual al producto de las probabilidades individuales. Al utilizar la propiedad de multiplicación para más de dos eventos independientes es legítimo reemplazar una o más de las A_i por sus complementos (p. ej., si A_1, A_2 y A_3 son eventos independientes, también lo son A_1', A_2' y A_3'). Tal como sucedió con dos eventos, con frecuencia al principio de un problema se especifica la independencia de ciertos eventos. La probabilidad de una intersección puede entonces ser calculada mediante multiplicación.

EJEMPLO 2.36 El artículo **"Reliability Evaluation of Solar Photovoltaic Arrays" (*Solar Energy*, 2002: 129–141)** presenta varias configuraciones de redes fotovoltaicas solares compuestas de celdas solares de silicio cristalino. Considere primero el sistema que se ilustra en la figura 2.14(a).

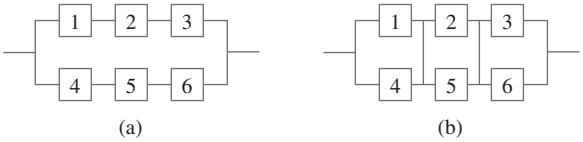


Figura 2.14 Configuraciones de sistema para el ejemplo 2.36: (a) serie-paralelo; (b) matriz interconectada

Existen dos subsistemas conectados en paralelo, y cada uno contiene tres celdas. Para que el sistema funcione, al menos uno de los dos subsistemas en paralelo debe funcionar. Dentro de cada subsistema las tres celdas están conectadas en serie, así que un subsistema funcionará sólo si todas sus celdas funcionan. Considere un valor de duración particular t_0 y suponga que desea determinar la probabilidad de que la duración del sistema exceda de t_0. Sea A_i el evento en que la duración de la celda i excede t_0 ($i = 1, 2,\ldots, 6$). Se supone que las A_i son eventos independientes (si alguna celda particular dura más de t_0 nada tiene que ver con si cualquier otra celda hace o no) y que $P(A_i) = 0.9$ para cada i, ya que las celdas son idénticas. Entonces

$$\begin{aligned} P(\text{la duración del sistema excede } t_0) &= P[(A_1 \cap A_2 \cap A_3) \cup (A_4 \cap A_5 \cap A_6)] \\ &= P(A_1 \cap A_2 \cap A_3) + P(A_4 \cap A_5 \cap A_6) \\ &\quad - P[(A_1 \cap A_2 \cap A_3) \cap (A_4 \cap A_5 \cap A_6)] \\ &= (0.9)(0.9)(0.9) + (0.9)(0.9)(0.9) \cdot (0.9)(0.9)(0.9)(0.9)(0.9)(0.9) = 0.927 \end{aligned}$$

Alternativamente,

$$\begin{aligned} P(\text{la duración del sistema excede } t_0) &= 1 - P(\text{ambas duraciones del subsistema son} \le t_0) \\ &= 1 - [P(\text{la duración del subsistema es} \le t_0)]^2 \\ &= 1 - [1 - P(\text{la duración del subsistema es} > t_0)]^2 \\ &= 1 - [1 - (0.9)^3]^2 = 0.927 \end{aligned}$$

Considere a continuación el sistema matriz interconectado que se muestra en la figura 2.14(b), que se obtuvo a partir de la red conectada en serie-paralela mediante la conexión

de enlaces a través de cada columna de uniones. Ahora, el sistema falla cuando toda una columna falla, y la duración del sistema excede t_0 sólo si la duración de cada columna lo hace. Para esta configuración,

$$\begin{aligned} P(\text{la duración del sistema es de al menos } t_0) &= [P(\text{la duración de la columna excede } t_0)]^3 \\ &= [1 - P(\text{la duración de la columna es} \le t_0)]^3 \\ &= [1 - P(\text{la duración de ambas celdas en una columna es} \le t_0)]^3 \\ &= [1 - (1 - 0.9)^2]^3 = 0.970 \end{aligned}$$

■

EJERCICIOS Sección 2.5 (70–89)

70. Reconsidere el escenario de la tarjeta de crédito del ejercicio 47 (sección 2.4) y demuestre que A y B son dependientes utilizando primero la definición de independencia y luego verificando que no prevalezca la propiedad de multiplicación.

71. En la actualidad una compañía de exploración petrolera tiene dos proyectos activos: uno en Asia y el otro en Europa. Sea A el evento en que el proyecto asiático tiene éxito y B el evento en que el proyecto europeo tiene éxito. Suponga que A y B son eventos independientes con $P(A) = 0.4$ y $P(B) = 0.7$.

- **a.** Si el proyecto asiático no tiene éxito, ¿cuál es la probabilidad de que el europeo tampoco lo tenga? Explique su razonamiento.
- **b.** ¿Cuál es la probabilidad de que al menos uno de los dos proyectos tenga éxito?
- **c.** Dado que al menos uno de los dos proyectos tiene éxito, ¿cuál es la probabilidad de que sólo el proyecto asiático lo tenga?

72. En el ejercicio 13, ¿es cualquier A_i independiente de cualquier otro A_j? Responda utilizando la propiedad de multiplicación para eventos independientes.

73. Si A y B son eventos independientes, demuestre que A' y B también son independientes. [*Sugerencia*: Primero establezca una relación entre $P(A' \cap B)$, $P(B)$ y $P(A \cap B)$.]

74. Suponga que las proporciones de fenotipos sanguíneos en una población son las siguientes:

A	B	AB	O
0.40	0.11	0.04	0.45

Suponiendo que los fenotipos de dos individuos seleccionados al azar son independientes uno de otro, ¿cuál es la probabilidad de que ambos fenotipos sean O? ¿Cuál es la probabilidad de que los fenotipos de dos individuos seleccionados al azar coincidan?

75. Una de las suposiciones que sustentan la teoría de las gráficas de control (véase el capítulo 16) es que los puntos graficados sucesivos son independientes entre sí. Cada punto puede señalar que un proceso de producción está funcionando correctamente o que existe algún funcionamiento defectuoso. Aun cuando un proceso esté funcionando de manera correcta, existe una pequeña probabilidad de que un punto particular señale algún problema con el proceso. Suponga que esta probabilidad es de 0.05. ¿Cuál es la probabilidad de que al menos uno de 10 puntos sucesivos indique un problema cuando de hecho el proceso está operando correctamente? Responda esta pregunta para 25 puntos sucesivos.

76. En octubre de 1994 se descubrió un defecto en un determinado chip Pentium instalado en las computadoras el cual podía dar lugar a una respuesta equivocada al realizar una división. El fabricante sostuvo inicialmente que la posibilidad de que cualquier división particular fuera incorrecta era sólo 1 de cada 9000 millones, así que tomaría miles de años antes de que un usuario típico detectara un error. Sin embargo, los estadísticos no son usuarios típicos, algunas técnicas estadísticas modernas son tan computacionalmente intensivas que mil millones de divisiones en un corto periodo de tiempo no están fuera del reino de la posibilidad. Suponiendo que la cifra de 1 en 9000 millones es correcta y que los resultados de las distintas divisiones son independientes uno del otro, ¿cuál es la probabilidad de que al menos haya un error en mil millones de divisiones con este chip?

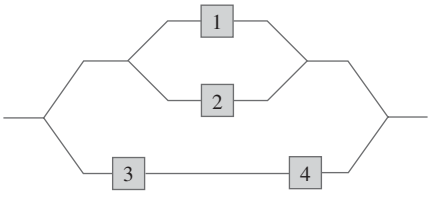
77. La unión de piel de un avión requiere 25 remaches. La unión de piel tendrá que volver a trabajarse si alguno de los remaches está defectuoso. Suponga que los remaches están defectuosos independientemente uno de otro, cada uno con la misma probabilidad.

- **a.** Si 15% de todas las costuras tiene que volver a trabajarse, ¿cuál es la probabilidad de que un remache esté defectuoso?
- **b.** ¿Qué tan pequeña deberá ser la probabilidad de un remache defectuoso para garantizar que sólo 10% de las costuras tengan que volver a trabajarse?

78. Una caldera tiene cinco válvulas de alivio idénticas. La probabilidad de que cualquier válvula particular se abra en un momento de demanda es de 0.96. Suponiendo que operan independientemente, calcule P(al menos una válvula se abre) y P(al menos una válvula no se abre).

79. Dos bombas conectadas en paralelo fallan independientemente una de la otra cualquier día dado. La probabilidad de que falle sólo la bomba más vieja es de 0.10 y la probabilidad de que sólo la bomba más nueva falle es de 0.05. ¿Cuál es la probabilidad de que el sistema de bombeo falle en cualquier día dado (lo cual sucede si ambas bombas fallan)?

80. Considere el sistema de componentes conectados tal como en la siguiente figura. Los componentes 1 y 2 están conectados en paralelo, de modo que el subsistema trabaja si y sólo si 1 o 2 trabajan; puesto que 3 y 4 están conectados en serie ¿el subsistema trabaja si y sólo si 3 y 4 trabajan? Si los componentes funcionan independientemente uno de otro y P(el componente i trabaja) $= 0.9$ para $i = 1,2$ y $= 0.8$ para $i = 3,4$, calcule P(el sistema trabaja).



81. Remítase otra vez al sistema en serie-paralelo introducido en el ejemplo 2.36 y suponga que existen sólo dos celdas en lugar de tres en cada subsistema en paralelo [en la figura 2.14(a), elimine las celdas 3 y 6 y vuelva a numerar las celdas 4 y 5 como 3 y 4]. Utilizando $P(A_i) = 0.9$ es fácil ver que la probabilidad de que la duración del sistema exceda t_0 es de 0.9639. ¿A qué valor tendría que cambiar 0.9 para incrementar la confiabilidad y la duración del sistema de 0.9639 a 0.99? [*Sugerencia*: Sea $P(A_i) = p$, exprese la confiabilidad del sistema en función de p y luego haga $x = p^2$.]

82. Considere lanzar en forma independiente dos dados imparciales, uno rojo y otro verde. Sea A el evento en que el dado rojo muestra 3 puntos, B el evento en que el dado verde muestra 4 puntos y C el evento en que el número total de puntos que muestran los dos dados es 7. ¿Son estos eventos independientes por pares (es decir, ¿son A y B eventos independientes, son A y C independientes, y son B y C independientes? ¿Son los tres eventos mutuamente independientes?

83. Los componentes enviados a un distribuidor son revisados en cuanto a defectos por dos inspectores diferentes (cada componente es revisado por ambos inspectores). El primer inspector detecta 90% de todos los defectuosos que están presentes y el segundo hace lo mismo. Al menos un inspector no detecta un defecto en 20% de todos los componentes defectuosos. ¿Cuál es la probabilidad de que ocurra lo siguiente?

- **a.** ¿Que un componente defectuoso sea detectado sólo por el primer inspector? ¿Y por exactamente uno de los dos inspectores?
- **b.** ¿Que los tres componentes defectuosos en un lote no sean detectados por ninguno de los dos inspectores (suponiendo que las inspecciones de los diferentes componentes son independientes unas de otras)?

84. Considere la compra de un sistema de componentes de audio que consta de un receptor, un par de altavoces y un reproductor de CD. Que A_1 sea el evento en que las funciones del receptor trabajen correctamente a lo largo del periodo de garantía, que A_2 sea el evento en que los altavoces funcionen correctamente durante el periodo de garantía, y A_3 el evento en que las funciones del reproductor de CD trabajan correctamente durante el periodo de garantía. Supongamos que estos eventos son (mutuamente) independientes con $P(A_1) = 0.95$, $P(A_2) = 0.98$, y $P(A_3) = 0.80$.

- **a.** ¿Cuál es la probabilidad de que los tres componentes funcionen adecuadamente durante todo el periodo de garantía?
- **b.** ¿Cuál es la probabilidad de que al menos un componente necesite servicio durante el periodo de garantía?
- **c.** ¿Cuál es la probabilidad de que los tres componentes necesiten servicio durante el periodo de garantía?
- **d.** ¿Cuál es la probabilidad de que sólo el receptor necesite servicio durante el periodo de garantía?
- **e.** ¿Cuál es la probabilidad de que exactamente uno de los tres componentes necesite servicio durante el periodo de garantía?
- **f.** ¿Cuál es la probabilidad de que los tres componentes funcionen adecuadamente durante todo el periodo de garantía, pero que, al menos, uno falle un mes después de que expire la garantía?

85. Un inspector de control de calidad verifica artículos recién producidos en busca de fallas. El inspector examina un artículo en busca de fallas en una serie de observaciones independientes, cada una de duración fija. Dado que en realidad hay una imperfección, sea p la probabilidad de que la imperfección sea detectada durante cualquier observación (este modelo se analiza en **"Human Performance in Sampling Inspection", *Human Factors*, 1979: 99–105)**.

- **a.** Suponiendo que un artículo tiene una imperfección, ¿cuál es la probabilidad de que sea detectada al final de la segunda observación (una vez que una imperfección ha sido detectada, la secuencia de observaciones termina)?
- **b.** Ofrezca una expresión para la probabilidad de que una imperfección sea detectada al final de la n-ésima observación.
- **c.** Si luego de tres observaciones no ha sido detectada una imperfección el artículo es aprobado, ¿cuál es la probabilidad de que un artículo imperfecto pase la inspección?
- **d.** Suponga que 10% de todos los artículos contiene una imperfección [P(artículo seleccionado al azar muestra una imperfección) $= 0.1$]. Con la suposición del inciso c), ¿cuál es la probabilidad de que un artículo seleccionado al azar pase la inspección (pasará automáticamente si no muestra una imperfección, pero también podría pasar si muestra una imperfección)?
- **e.** Dado que un artículo ha pasado la inspección (ninguna imperfección en tres observaciones), ¿cuál es la probabilidad de que sí tenga una imperfección? Calcule para $p = 0.5$.

86. **a.** Una compañía maderera acaba de recibir un lote de 10 000 tablas de 2×4. Suponga que 20% de estas tablas (2000) en realidad están demasiado tiernas o verdes para ser utilizadas en construcción de primera calidad. Se eligen dos tablas al azar, una después de la otra. Sea $A = \{$la primera tabla está verde$\}$ y $B = \{$la segunda tabla está verde$\}$. Calcule $P(A)$, $P(B)$ y $P(A \cap B)$ (un diagrama de árbol podría ayudar). ¿Son A y B independientes?

- **b.** Con A y B independientes y $P(A) = P(B) = 0.2$, ¿cuál es $P(A \cap B)$? ¿Cuánta diferencia existe entre esta respuesta y $P(A \cap B)$ en el inciso a)? Para propósitos de cálculo $P(A \cap B)$, ¿se puede suponer que A y B del inciso a) son independientes para obtener en esencia la probabilidad correcta?

c. Suponga que un lote consta de 10 tablas, de las cuales dos están verdes. ¿Produce ahora la suposición de independencia aproximadamente la respuesta correcta para $P(A \cap B)$? ¿Cuál es la diferencia crítica entre la situación en este caso y la del inciso a)? ¿Cuándo cree que una suposición de independencia sería válida al obtener una respuesta aproximadamente correcta para $P(A \cap B)$?

87. Considere la posibilidad de seleccionar al azar a una sola persona y que esta pruebe tres vehículos diferentes. Defina los eventos A_1, A_2 y A_3 por

A_1 = como el vehículo #1 A_2 = como el vehículo #2

A_3 = como el vehículo #3

Suponga que $P(A_1) = 0.55$, $P(A_2) = 0.65$, $P(A_3) = 0.70$, $P(A_1 \cup A_2) = 0.80$, $P(A_2 \cap A_3) = 0.40$, y $P(A_1 \cup A_2 \cup A_3) = 0.88$.

a. ¿Cuál es la probabilidad de que a la persona le gusten tanto el vehículo #1 como el vehículo #2?

b. Determine e interprete $P(A_2 \mid A_3)$.

c. ¿Son eventos independientes A_2 y A_3? Responda de dos maneras diferentes.

d. Si usted sabe que a la persona no le gusta el vehículo #1, ¿cuál es ahora la probabilidad de que le guste al menos uno de los otros dos vehículos?

88. La probabilidad de que una persona seleccionada aleatoriamente de una población particular tenga una determinada enfermedad es 0.5. Una prueba de diagnóstico detecta correctamente la presencia de la enfermedad 98% del tiempo y detecta correctamente la ausencia de la enfermedad 99% del tiempo. Si la prueba se aplica dos veces, los dos resultados son independientes y ambos son positivos ¿cuál es la probabilidad (posterior) de que la persona seleccionada padezca la enfermedad? [*Sugerencia*: El diagrama de árbol con ramas primera generación correspondiente a la enfermedad y a la no enfermedad; y las ramas de segunda y tercera generación corresponden a los resultados de los dos exámenes.]

89. Suponga que se colocan etiquetas idénticas en las orejas de un zorro. El zorro es liberado durante un lapso de tiempo. Considere los dos eventos C_1 = {se pierde la etiqueta de la oreja izquierda} y C_2 = {se pierde la etiqueta de la oreja derecha}. Sea $\pi = P(C_1) = P(C_2)$, y suponga que C_1 y C_2 son eventos independientes. Deduzca una expresión (que implique π) para la probabilidad de que exactamente una etiqueta se pierda, dado que cuando mucho una se pierde (**"Ear Tag Loss in Red Foxes", *J. Wildlife Mgmt*., 1976: 164–167**). [*Sugerencia*: Trace un diagrama de árbol en el cual las dos ramas iniciales se refieren a si la etiqueta de la oreja izquierda se pierde.]

EJERCICIOS SUPLEMENTARIOS (90–97)

90. Un cierto comité legislativo consta de 10 senadores. Se seleccionará al azar un subcomité de 3 senadores.

a. ¿Cuántos diferentes subcomités de estos hay?

b. Si los senadores se clasifican 1, 2,…, 10 por orden de antigüedad, ¿cuántos subcomités diferentes incluirían al senador más veterano?

c. ¿Cuál es la probabilidad de que el subcomité seleccionado tenga al menos 1 de los 5 senadores más veteranos?

d. ¿Cuál es la probabilidad de que el subcomité no incluya a ninguno de los dos senadores más veteranos?

91. Una fábrica utiliza tres líneas de producción para fabricar latas de cierto tipo. La siguiente tabla proporciona los porcentajes de latas que no cumplen con las especificaciones, clasificadas por tipo de incumplimiento de las especificaciones para cada una de las tres líneas durante un lapso de tiempo particular.

	Línea 1	Línea 2	Línea 3
Manchas	15	12	20
Grietas	50	44	40
Problemas con la argolla	21	28	24
Defecto superficial	10	8	15
Otro	4	8	2

Durante este periodo la línea 1 produjo 500 latas fuera de especificación, la 2 produjo 400 latas como esas y la 3 fue responsable de 600 latas fuera de especificación. Suponga que se selecciona al azar una de estas 1500 latas.

a. ¿Cuál es la probabilidad de que la lata provenga de la línea 1? ¿Y de que la razón del incumplimiento de la especificación sea una grieta?

b. Si la lata seleccionada provino de la línea 1, ¿cuál es la probabilidad de que tenga una mancha?

c. Dado que la lata seleccionada mostró un defecto superficial, ¿cuál es la probabilidad de que provenga de la línea 1?

92. Un empleado de la oficina de inscripciones en una universidad tiene en su escritorio en este momento diez formas en espera de ser procesadas. Seis de estas son peticiones de baja y las otras cuatro son solicitudes de sustitución de curso.

a. Si selecciona al azar seis de estas formas para dárselas a un subordinado, ¿cuál es la probabilidad de que sólo uno de los dos tipos de formas permanezca en su escritorio?

b. Suponga que tiene tiempo para procesar sólo cuatro de estas formas antes de salir del trabajo. Si estas cuatro se seleccionan al azar, una por una, ¿cuál es la probabilidad de que cada forma subsiguiente sea de un tipo diferente de la anterior?

93. Un satélite está programado para ser lanzado desde Cabo Cañaveral, en Florida, y otro lanzamiento está programado para la Base de la Fuerza Aérea Vandenberg en California. Sea A el evento en que el lanzamiento en Vandenberg se hace a la hora programada y B el evento en que el lanzamiento en Cabo Cañaveral se hace a la hora programada. Si A y B son eventos independientes con $P(A) > P(B)$, $P(A \cup B) = 0.626$ y $P(A \cap B) = 0.144$, determine los valores de $P(A)$ y $P(B)$.

94. Un transmisor envía un mensaje utilizando un código binario, esto es, una secuencia de ceros y unos. Cada bit transmitido (0 o 1) debe pasar a través de tres relevadores para llegar al receptor. En cada relevador la probabilidad de que el bit enviado sea diferente del bit recibido (una inversión) es 0.20. Suponga que los relevadores operan independientemente uno de otro.

Transmisor → Relevador 1 → Relevador 2 → Relevador 3 → Receptor

- **a.** Si el transmisor envía un 1, ¿cuál es la probabilidad de que los tres relevadores envíen un 1?
- **b.** Si el transmisor envía un 1, ¿cuál es la probabilidad de que el receptor reciba un 1? [*Sugerencia:* Los ocho resultados experimentales pueden ser mostrados en un diagrama de árbol con tres ramas de generación, una por cada relevador.]
- **c.** Suponga que 70% de todos los bits enviados por el transmisor son unos. Si el receptor recibe un 1, ¿cuál es la probabilidad de que un 1 haya sido enviado?

95. El individuo A tiene un círculo de cinco amigos cercanos (B, C, D, E y F). A escuchó cierto rumor originado fuera del círculo e invitó a sus cinco amigos a una fiesta para contarles el rumor. Para empezar, A escoge al azar a uno de sus cinco y se lo cuenta. Dicho individuo escoge entonces al azar a uno de los cuatro individuos restantes y repite el rumor. Después, de aquellos que ya oyeron el rumor uno se lo cuenta a otro nuevo individuo y así hasta que todos han oído el rumor.

- **a.** ¿Cuál es la probabilidad de que el rumor se repita en el orden B, C, D, E y F?
- **b.** ¿Cuál es la probabilidad de que F sea la tercera persona en la reunión a la que se le contará el rumor?
- **c.** ¿Cuál es la probabilidad de que F sea la última persona en oír el rumor?
- **d.** Si en cada etapa la persona que en ese momento "tiene" el rumor no sabe quién ya lo ha escuchado y selecciona al siguiente destinatario aleatoriamente de entre cinco individuos posibles, ¿cuál es la probabilidad de que F no haya escuchado todavía el rumor después de haber sido dicho 10 veces en la fiesta?

96. De acuerdo con el artículo **"Optimization of Distribution Parameters for Estimating Probability of Crack Detection" (*J. of Aircraft*, 2009: 2090–2097)**, la siguiente ecuación de "Palmberg" se usa comúnmente para determinar la probabilidad $P_d(c)$ de la detección de una grieta de tamaño c en la estructura de la aeronave:

$$P_d(c) = \frac{(c/c^*)^\beta}{1 + (c/c^*)^\beta}$$

donde c^* es el tamaño de la grieta que corresponde a una probabilidad de detección de 0.5 (y, por tanto, es una evaluación de la calidad del proceso de inspección).

- **a.** Compruebe que $P_d(c^*) = 0.5$
- **b.** ¿Qué es $P_d(2c^*)$ cuando $\beta = 4$?
- **c.** Supongamos que un inspector revisa dos paneles diferentes, uno con un tamaño de grieta c^* y otro con un tamaño de grieta $2c^*$. Una vez más, suponiendo que $\beta = 4$ y también que los resultados de las dos inspecciones son independientes uno del otro, ¿cuál es la probabilidad de que exactamente se detecte una de las dos grietas?
- **d.** ¿Qué le sucede a $P_d(c)$ mientras $\beta \to \infty$?

97. Un ingeniero químico está interesado en determinar si cierta impureza está presente en un producto. Un experimento tiene una probabilidad de 0.80 de detectarla, si está presente. La probabilidad de no detectarla, si está ausente, es de 0.90. Las probabilidades previas de que la impureza esté presente o ausente son de 0.40 y 0.60, respectivamente. Tres experimentos distintos producen sólo dos detecciones. ¿Cuál es la probabilidad posterior de que la impureza esté presente?

BIBLIOGRAFÍA

Carlton, Matthew y Jay Devore, *Probability with Applications in Engineering, Science, and Technology*, Springer, Nueva York, 2014. Una completa introducción a la probabilidad, escrita en un nivel matemático ligeramente superior a este texto, pero con muy buenos ejemplos.

Durrett, Richard, *Elementary Probability for Applications*, Cambridge Univ. Press, Londres, Inglaterra, 2009. Una presentación concisa a un nivel un poco más alto que el presente texto.

Mosteller, Frederick, Robert Rourke y George Thomas, *Probability with Statistical Applications* (2a. ed.), Addison-Wesley, Reading, MA, 1970. Una muy buena introducción a la probabilidad, con muchos ejemplos entretenidos, especialmente buenos respecto a las reglas de conteo y su aplicación.

Ross, Sheldon, *A First Course in Probability* (8a. ed.), Macmillan, Nueva York, 2009. Algo concisamente escrito y más matemáticamente complejo que este texto, pero contiene gran cantidad de ejemplos y ejercicios interesantes.

Winkler, Robert, *Introduction to Bayesian Inference and Decision*, Holt, Rinehart & Winston, Nueva York, 1972. Una muy buena introducción a la probabilidad subjetiva.

Capítulo

3

Variables aleatorias discretas y distribuciones de probabilidad

INTRODUCCIÓN

Ya sea que un experimento produzca resultados cualitativos o cuantitativos los métodos de análisis estadístico requieren enfocarse en ciertos aspectos numéricos de los datos (como la proporción de una muestra x/n, la media $\bar{x}$ o la desviación estándar s). El concepto de variable aleatoria permite pasar de los resultados experimentales a la función numérica de los resultados. Existen fundamentalmente dos tipos diferentes de variables aleatorias: las variables aleatorias discretas y las variables aleatorias continuas. En este capítulo se examinan las propiedades básicas y se discuten los ejemplos más importantes de variables discretas. El capítulo 4 se enfoca en las variables aleatorias continuas.

3.1 Variables aleatorias

En cualquier experimento existen numerosas características que pueden ser observadas o medidas, pero en la mayoría de los casos un experimentador se centra en algún aspecto específico o aspectos de una muestra. Por ejemplo, en un estudio sobre los patrones de viaje entre los suburbios y la ciudad en un área metropolitana, a cada individuo de una muestra se le podría preguntar sobre la distancia que recorre para ir de su casa al trabajo, y viceversa, así como el número de personas que viajan en el mismo vehículo, pero no sobre su coeficiente intelectual, ingreso, tamaño de su familia y otras características. Por otra parte, un investigador puede probar una muestra de componentes y anotar sólo el número de los que han fallado en un periodo de 1000 horas, en lugar de anotar los tiempos en que ha fallado cada uno individualmente.

En general, el resultado de cada experimento puede ser vinculado con un número especificando una regla de asociación (p. ej., entre una muestra de diez componentes el número de estos que no duran 1000 horas o el peso total del equipaje en una muestra de 25 pasajeros de una aerolínea). Esta regla de asociación se llama **variable aleatoria**, variable porque diferentes valores numéricos son posibles y aleatoria porque el valor observado depende de cuál de los posibles resultados experimentales se obtenga (figura 3.1).

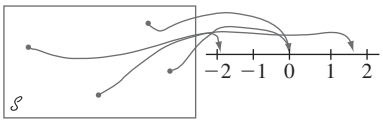


Figura 3.1 Una variable aleatoria

DEFINICIÓN

> Para un espacio muestral dado $\mathcal{S}$ de algún experimento, una **variable aleatoria (va)** es cualquier regla que asocia un número con cada resultado en $\mathcal{S}$. En lenguaje matemático una variable aleatoria es una función cuyo dominio es el espacio muestral y cuyo rango es el conjunto de los números reales.

Se acostumbra denotar las variables aleatorias con letras mayúsculas, tales como X y Y, que son las más cercanas al final de nuestro alfabeto. En contraste al uso previo de una letra minúscula como es la x para denotar una variable, ahora utilizaremos letras minúsculas para representar algún valor particular de la variable aleatoria correspondiente. La notación $X(\omega) = x$ significa que x es el valor asociado con el resultado ω por medio de la variable aleatoria X.

EJEMPLO 3.1 Cuando un estudiante llama por teléfono a un servicio de asistencia universitaria para apoyo técnico, inmediatamente podrá hablar con alguien (S) o será puesto en espera (F). Con $\mathcal{S} = \{S, F\}$, la variable aleatoria X se define como

$$X(S) = 1 \qquad X(F) = 0$$

La variable aleatoria X indica si el estudiante puede hablar inmediatamente con alguien (1) o si no puede hablar de inmediato con alguien (0). ■

En el ejemplo 3.1 se especificó la variable aleatoria X al poner en lista explícitamente cada elemento de $\mathcal{S}$ y el número asociado. Una lista como esa es tediosa si $\mathcal{S}$ contiene más de algunos cuantos resultados, pero con frecuencia puede ser evitada.

EJEMPLO 3.2 Considere el experimento en el cual se marca un número telefónico en cierto código de área con un marcador de números aleatorio (estos dispositivos se utilizan extensivamente en las organizaciones encuestadoras) y defina una variable aleatoria Y como

$$Y = \begin{cases} 1 & \text{si el número seleccionado no aparece en el directorio} \\ 0 & \text{si el número seleccionado sí aparece en el directorio} \end{cases}$$

Por ejemplo, si 5282966 aparece en el directorio telefónico, entonces $Y(5282966) = 0$ en tanto que $Y(7727350) = 1$ significa que el número 7727350 no aparece en el directorio telefónico. Es más eficaz una descripción verbal de este estilo que una lista completa, por lo que se utilizará tal descripción siempre que sea posible. ■

En los ejemplos 3.1 y 3.2 los únicos valores posibles de la variable aleatoria fueron 0 y 1. Tal variable aleatoria se presenta con tal frecuencia que le ha dado un nombre especial, en honor de la primera persona que la estudió.

DEFINICIÓN

Cualquier variable aleatoria cuyos únicos valores posibles son 0 y 1 se llama **variable aleatoria de Bernoulli.**

En ocasiones se deseará definir y estudiar diferentes variables del mismo espacio muestral.

EJEMPLO 3.3 El ejemplo 2.3 describe un experimento en el cual se determinó el número de bombas en uso en cada una de dos gasolineras. Defina las variables aleatorias X, Y y U como

X = el número total de bombas en uso en las dos gasolineras

Y = la diferencia entre el número de bombas en uso en la gasolinera 1 y el número de bombas en uso en la gasolinera 2

U = el máximo del número de bombas en uso en ambas gasolineras

Si se realiza este experimento y se obtiene $\omega = (2, 3)$, entonces $X((2, 3)) = 2 + 3 = 5$, por lo que se dice que el valor observado de X fue $x = 5$. Asimismo, el valor observado de Y sería $y = 2 - 3 = -1$ y el de U sería $u = \text{máx}\,(2, 3) = 3$. ■

Cada una de las variables aleatorias de los ejemplos 3.1–3.3 puede asumir sólo un número finito de posibles valores. Este no tiene que ser el caso.

EJEMPLO 3.4 Se considera un experimento en que se examinaron baterías de 9 volts hasta que se obtuvo una con un voltaje aceptable (S). El espacio muestral es $\mathcal{S} = \{S, FS, FFS, \ldots\}$. Defina una variable aleatoria X como

X = número de baterías examinadas antes de que se termine el experimento

Así, $X(S) = 1$, $X(FS) = 2$, $X(FFS) = 3$, . . . , $X(FFFFFFS) = 7$, y así sucesivamente. Cualquier entero positivo es un posible valor de X, así que el conjunto de valores posibles es infinito. ■

EJEMPLO 3.5 Suponga que del mismo modo aleatorio se selecciona un lugar (latitud y longitud) en el territorio continental de los Estados Unidos. Defina una variable aleatoria Y como

Y = la altura sobre el nivel del mar en el lugar seleccionado

Por ejemplo, si el lugar seleccionado fuera (39°50′N, 98°35′W), entonces se podría tener $Y((39°50'\text{N}, 98°35'\text{O})) = 1748.26$ pies. El valor más grande posible de Y es 14 494 (Monte Whitney) y el valor más pequeño posible es -282 (Valle de la Muerte). El conjunto de todos los valores posibles de Y es el conjunto de todos los números en el intervalo entre –282 y 14 494, es decir,

$$\{y : y \text{ es un número}, -282 \leq y \leq 14\,494\}$$

y existe un número infinito de números en este intervalo. ■

Dos tipos de variables aleatorias

En la sección 1.2 se distinguió entre los datos que resultan de las observaciones de una variable de conteo y los datos obtenidos observando valores de una variable de medición. Una distinción un poco más formal caracteriza dos tipos diferentes de variables aleatorias.

DEFINICIÓN

Una variable aleatoria **discreta** es una variable aleatoria cuyos valores posibles constituyen un conjunto finito, o bien pueden ser puestos en lista en una secuencia infinita en la cual existen un primer elemento, un segundo elemento y así sucesivamente ("contablemente" infinita).

Una variable aleatoria es **continua** si las siguientes condiciones dos se cumplen:

1. Su conjunto de valores posibles se compone de todos los números que hay en un solo intervalo sobre la línea de numeración (posiblemente de extensión infinita, es decir, desde $-\infty$ hasta ∞) o todos los números en una unión disjunta de dichos intervalos (por ejemplo, $[0,10] \cup [20,30]$).
2. Ningún valor posible de la variable tiene probabilidad positiva, esto es, $P(X = c) = 0$ con cualquier valor posible de c.

Aunque cualquier intervalo sobre la línea de numeración contiene un número infinito de números, se puede demostrar que no hay ninguna forma de crear una lista infinita de todos estos valores, pues son demasiados. La segunda condición que describe una variable aleatoria continua es tal vez contraintuitiva, puesto que parecería que implica una probabilidad total de cero para todos los valores posibles. Pero en el capítulo 4 se verá que los *intervalos* de valores tienen probabilidad positiva; la probabilidad de un intervalo se reducirá a cero a medida que su ancho tienda a cero.

EJEMPLO 3.6 Todas las variables aleatorias de los ejemplos 3.1–3.4 son discretas. En otro ejemplo, suponga que se eligen al azar parejas de casados y que a cada persona se le hace una prueba de sangre hasta encontrar a un esposo y su esposa con el mismo factor Rh. Con X = número de pruebas de sangre que serán realizadas, los posibles valores de X son $D = \{2, 4, 6, 8, \ldots\}$. Puesto que los posibles valores se dieron en secuencia, X es una variable aleatoria discreta. ■

Para estudiar las propiedades básicas de las variables aleatorias discretas sólo se requieren las herramientas de las matemáticas discretas: sumas y diferencias. El estudio de variables continuas requiere matemáticas continuas de cálculo: integrales y derivadas.

EJERCICIOS Sección 3.1 (1–10)

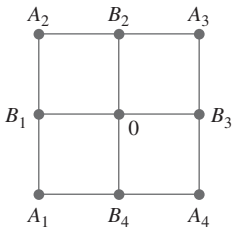
1. Una viga de concreto puede fallar por esfuerzo cortante (S) o por flexión (F). Suponga que se seleccionan al azar tres vigas que fallaron y se determina el tipo de falla de cada una. Sea X = el número de vigas de entre las tres seleccionadas que fallaron por esfuerzo cortante. Enumere cada resultado en el espacio muestral junto con el valor asociado de X.
2. Dé tres ejemplos de variables aleatorias de Bernoulli (aparte de los que aparecen en el texto).
3. Con el experimento del ejemplo 3.3 defina dos variables aleatorias más y mencione los valores posibles de cada una.
4. Sea X = el número de dígitos diferentes de cero en un código postal seleccionado al azar. ¿Cuáles son los posibles valores de X? Dé tres posibles resultados y sus valores X asociados.
5. Si el espacio muestral $\mathcal{S}$ es un conjunto infinito, ¿ello implica necesariamente que cualquier variable aleatoria X definida a partir de $\mathcal{S}$ tendrá un conjunto infinito de posibles valores? Si la respuesta es sí, ¿por qué? Si no, dé un ejemplo.
6. A partir de una hora fija, cada automóvil que entra a una intersección es observado para ver si vira a la izquierda (I) o a la derecha (D), o si sigue de frente (F). El experimento termina en cuanto se observa que un auto vira a la izquierda. Sea X = el número de autos observados. ¿Cuáles son los posibles valores de X? Dé cinco resultados y sus valores X asociados.

7. Para cada variable aleatoria definida aquí, describa el conjunto de posibles valores de la variable y diga si la variable es discreta.

- **a.** X = el número de huevos no quebrados en una caja estándar de huevos seleccionada al azar.
- **b.** Y = el número de estudiantes en una lista de clase de un curso particular que no asisten el primer día de clases.
- **c.** U = el número de veces que un aprendiz tiene que hacerle *swing* a una pelota de golf antes de golpearla.
- **d.** X = la longitud de una serpiente de cascabel seleccionada en forma aleatoria.
- **e.** Z = el porcentaje de impuestos derivado de las ventas de una compra en el portal de amazon.com seleccionada al azar.
- **f.** Y = el pH de una muestra de suelo elegida al azar.
- **g.** X = la tensión (lb/pulg^2) a la cual ha sido encordada una raqueta de tenis seleccionada al azar.
- **h.** X = el número total de veces que se requiere lanzar al aire una moneda para que tres individuos obtengan una coincidencia (*AAA* o *SSS*).

8. Cada vez que un componente se somete a prueba, el ensayo resulta en un éxito (S) o en un fracaso (F). Suponga que el componente se prueba repetidamente hasta que ocurre un éxito en tres pruebas *consecutivas*. Sea Y el número necesario de pruebas para lograrlo. Haga una lista de todos los resultados correspondientes a los cinco posibles valores más pequeños de Y y señale qué valor de Y está asociado con cada uno.

9. Un individuo de nombre Claudio se encuentra en el punto 0 del siguiente diagrama.



Con un dispositivo de aleatorización apropiado (tal como un dado tetraédrico, es decir, de cuatro lados), Claudio primero se mueve a uno de los cuatro lugares B_1, B_2, B_3, B_4. Una vez que está en uno de estos lugares, se utiliza otro dispositivo de aleatorización para decidir si Claudio regresa a 0 o si visita uno de los otros dos lugares adyacentes. Este proceso continúa: después de cada movimiento se determina otro movimiento a uno de los (nuevos) puntos adyacentes lanzando al aire un dado o una moneda apropiada.

- **a.** Sea X = el número de movimientos que Claudio realiza antes de regresar a 0. ¿Cuáles son los posibles valores de X? ¿Es X discreta o continua?
- **b.** Si también se permiten movimientos a lo largo de los trayectos diagonales que conectan 0 con A_1, A_2, A_3 y A_4, respectivamente, responda las preguntas del inciso a).

10. Se determinará el número de bombas en uso tanto en la gasolinera de seis bombas como en la de cuatro bombas. Dé los posibles valores de cada una de las siguientes variables aleatorias:

- **a.** T = el número total de bombas en uso.
- **b.** X = la diferencia entre el número de bombas en uso en las gasolineras 1 y 2.
- **c.** U = el número máximo de bombas en uso en una u otra gasolinera.
- **d.** Z = el número de gasolineras que tienen exactamente dos bombas en uso.

3.2 Distribuciones de probabilidad para variables aleatorias discretas

Las probabilidades asignadas a varios resultados en $\mathcal{S}$ determinan a su vez las probabilidades asociadas con los valores de cualquier variable aleatoria X particular. La *distribución de probabilidad* de X dice cómo está distribuida (asignada) la probabilidad total de 1 entre los varios posibles valores de X. Suponga, por ejemplo, que una empresa acaba de adquirir cuatro impresoras láser y sea X el número de estas que requieren servicio durante el periodo de garantía. Los posibles valores de X son entonces 0, 1, 2, 3 y 4. La distribución de probabilidad dirá cómo está subdividida la probabilidad de 1 entre estos cinco posibles valores: cuánta probabilidad está asociada con el valor 0 de X, cuánta está adjudicada al valor 1 de X, y así sucesivamente. Se utilizará la siguiente notación para las probabilidades en la distribución:

$$p(0) = \text{la probabilidad del valor 0 de } X = P(X = 0)$$
$$p(1) = \text{la probabilidad del valor 1 de } X = P(X = 1)$$

y así sucesivamente. En general, $p(x)$ denotará la probabilidad asignada al valor de x.

EJEMPLO 3.7 El Departamento de Estadística de Cal Poly tiene un laboratorio con seis computadoras reservadas para los estudiantes de estadística. Sea X el número de computadoras que están en servicio a una hora particular del día. Suponga que la distribución de probabilidad de X es como se da en la tabla siguiente; la primera fila de la tabla contiene los posibles valores de X y la segunda da la probabilidad de dicho valor.

x	0	1	2	3	4	5	6
$p(x)$	0.05	0.10	0.15	0.25	0.20	0.15	0.10

Ahora se pueden usar propiedades de probabilidad elemental para calcular otras probabilidades de interés. Por ejemplo, la probabilidad de que cuando mucho 2 computadoras estén en servicio es

$$P(X \leq 2) = P(X = 0 \text{ o } 1 \text{ o } 2) = p(0) + p(1) + p(2) = 0.05 + 0.10 + 0.15 = 0.30$$

Puesto que el evento *de que al menos 3 computadoras estén en servicio* es complementario a *cuando mucho dos computadoras están en servicio*,

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.30 = 0.70$$

la cual, desde luego, también se obtiene sumando las probabilidades de los valores 3, 4, 5 y 6. La probabilidad de que entre 2 y 5 computadoras inclusive estén en servicio es

$$P(2 \leq X \leq 5) = P(X = 2, 3, 4 \text{ o } 5) = 0.15 + 0.25 + 0.20 + 0.15 = 0.75$$

en tanto que la probabilidad de que el número de computadoras en servicio esté estrictamente entre 2 y 5 es

$$P(2 < X < 5) = P(X = 3 \text{ o } 4) = 0.25 + 0.20 = 0.45$$ ■

DEFINICIÓN

La **distribución de probabilidad** o **función de masa de probabilidad** (fmp) de una variable discreta se define para cada número x como $p(x) = P(X = x) = P(\text{todos } \omega \in \mathcal{S}.\ X(\omega) = x)$.

Es decir, para cada valor posible x de la variable aleatoria la función de masa de probabilidad especifica la probabilidad de observar dicho valor cuando se realiza el experimento. Se requieren las condiciones $p(x) \geq 0$ y $\Sigma_{\text{todas las } x \text{ posibles}}\, p(x) = 1$ de cualquier función de masa de probabilidad.

La función de masa de probabilidad de X en el ejemplo previo se dio simplemente en la descripción del problema. A continuación se consideran varios ejemplos en los cuales se explotan varias propiedades de probabilidad para obtener la distribución deseada.

EJEMPLO 3.8 Seis lotes de componentes están listos para ser enviados por un proveedor. El número de componentes defectuosos en cada lote es como sigue:

Caja	1	2	3	4	5	6
Número de componentes defectuosos	0	2	0	1	2	0

Uno de estos lotes tiene que ser seleccionado al azar para ser enviado a un cliente particular. Sea X el número de componentes defectuosos en el lote seleccionado. Los tres posibles valores de X son 0, 1 y 2. De los seis eventos simples igualmente probables, tres dan por resultado $X = 0$, uno $X = 1$ y los otros dos $X = 2$. Entonces

$$p(0) = P(X = 0) = P(\text{el lote 1 o 3 o 6 es enviado}) = \frac{3}{6} = 0.500$$

$$p(1) = P(X = 1) = P(\text{el lote 4 es enviado}) = \frac{1}{6} = 0.167$$

$$p(2) = P(X = 2) = P(\text{el lote 2 o 5 es enviado}) = \frac{2}{6} = 0.333$$

Es decir, se asigna una probabilidad de 0.500 al valor 0 de X, una probabilidad de 0.167 al valor 1 de X y la probabilidad restante, 0.333 se asocia con el valor 2 de X. Los valores de X junto con sus probabilidades especifican la función de masa de probabilidad. Si este experimento se repitiera una y otra vez a la larga $X = 0$ ocurriría la mitad del tiempo, $X = 1$ un sexto del tiempo y $X = 2$ un tercio del tiempo. ■

EJEMPLO 3.9 Considere si la siguiente persona que compre una computadora en cierta tienda de electrónicos elegirá un modelo portátil o uno de escritorio. Sea

$$X = \begin{cases} 1 & \text{si el cliente compra una computadora de escritorio} \\ 0 & \text{si el cliente compra una computadora portátil} \end{cases}$$

Si 20% de todos los compradores durante esa semana seleccionan una computadora de escritorio, la función de masa de probabilidad de X es

$$p(0) = P(X = 0) = P(\text{el siguiente cliente compra un modelo portátil}) = 0.8$$
$$p(1) = P(X = 1) = P(\text{el siguiente cliente compra un modelo de escritorio}) = 0.2$$
$$p(x) = P(X = x) = 0 \text{ para } x \neq 0 \text{ o } 1$$

Una descripción equivalente es

$$p(x) = \begin{cases} 0.8 & \text{si } x = 0 \\ 0.2 & \text{si } x = 1 \\ 0 & \text{si } x \neq 0 \text{ o } 1 \end{cases}$$

La figura 3.2 es una ilustración de esta función de masa de probabilidad, llamada *gráfica lineal*. X es, desde luego, una variable aleatoria de Bernoulli y $p(x)$ es una función de masa de probabilidad de Bernoulli.

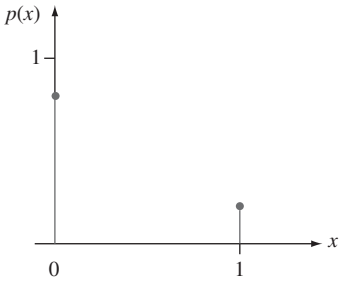


Figura 3.2 Gráfica lineal para la función de masa de probabilidad de Bernoulli en el ejemplo 3.9 ■

EJEMPLO 3.10 Considere un grupo de cinco donadores de sangre potenciales, *a, b, c, d* y *e*, de los cuales sólo *a* y *b* tienen sangre tipo O+. Se determinará en orden aleatorio el tipo de sangre con cinco muestras, una de cada individuo, hasta que se identifique un individuo O+. Sea la variable aleatoria Y = número de exámenes de sangre para identificar un individuo O+. Entonces la función de masa de probabilidad de Y es

$$p(1) = P(Y = 1) = P(a \text{ o } b \text{ examinados primero}) = \frac{2}{5} = 0.4$$
$$p(2) = P(Y = 2) = P(c, d \text{ o } e \text{ primero y luego } a \text{ o } b)$$
$$= P(c, d \text{ o } e \text{ primero}) \cdot P(a \text{ o } b \text{ en seguida} \mid c, d \text{ o } e \text{ primero}) = \frac{3}{5} \cdot \frac{2}{4} = 0.3$$
$$p(3) = P(Y = 3) = P(c, d \text{ o } e \text{ primero y segundo, y luego } a \text{ o } b)$$
$$= \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{2}{3}\right) = 0.2$$
$$p(4) = P(Y = 4) = P(c, d \text{ y } e \text{ todos primero}) = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right) = 0.1$$
$$p(y) = 0 \text{ si } y \neq 1, 2, 3, 4$$

En forma tabular, la función de masa de probabilidad es

y	1	2	3	4
$p(y)$	0.4	0.3	0.2	0.1

donde cualquier valor de y que no aparece en la tabla recibe cero probabilidad. La figura 3.3 muestra una gráfica lineal de la función de masa de probabilidad.

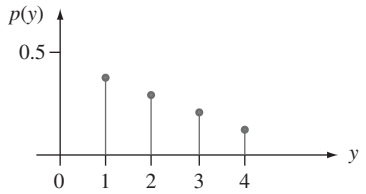


Figura 3.3 Gráfica lineal para la función de masa de probabilidad de Bernoulli del ejemplo 3.10 ■

Un modelo utilizado en física para un sistema de "masas puntuales" sugirió el nombre "función de masa de probabilidad". En este modelo las masas están distribuidas en varios lugares x a lo largo de un eje unidimensional. La función de masa de probabilidad describe cómo está distribuida la masa de probabilidad total de 1 en varios puntos a lo largo del eje de posibles valores de la variable aleatoria (dónde y cuánta masa hay en cada x).

Otra útil representación pictórica de una función de masa de probabilidad, llamada **histograma de probabilidad**, es similar a los histogramas discutidos en el capítulo 1. Sobre cada y con $p(y) > 0$ se construye un rectángulo con su centro en y. La altura de cada rectángulo es proporcional a $p(y)$ y la base es la misma para todos los rectángulos. Cuando los valores posibles están equidistantes con frecuencia se selecciona la base como la distancia entre valores y sucesivos (aunque podría ser más pequeña). La figura 3.4 muestra dos histogramas de probabilidad.

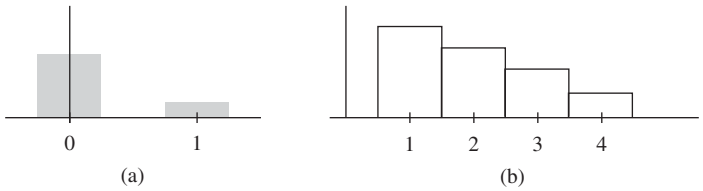


Figura 3.4 Histogramas de probabilidad: (a) Ejemplo 3.9; (b) Ejemplo 3.10

A menudo es útil pensar en una función de masa de probabilidad como un modelo matemático de una población discreta.

EJEMPLO 3.11 Considere seleccionar al azar una familia en una cierta región y suponga que X = el número de individuos en la familia seleccionada. Suponga que X tiene la siguiente función de masa de probabilidad:

x	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.140	0.175	0.220	0.260	0.155	0.025	0.015	0.005	0.004	0.001

[este ejemplo es muy parecido a la distribución del tamaño de las familias en la región rural de Tailandia que se muestra en el artículo **"The Probability of Containment for Multitype Branching Process Models for Emerging Epidemics" (*J. of Applied Probability*, 2011: 173–188)**, en el cual se hizo un modelo de la transmisión de influenza].

Suponga que esta está basada en 1 millón de familias. Una forma de ver esta situación es pensar que la población está compuesta de un millón de familias, cada una con su propio valor X; la proporción con cada valor X está dada por $p(x)$. Un punto de vista alternativo es olvidarse de las familias y pensar en la población como compuesta de los

valores X: 14% de estos valores son 1, 17.5% son dos, y así sucesivamente. La función de masa de probabilidad describe entonces la distribución de los valores posibles de la población 1, 2,…, 10. ■

Una vez que se tiene el modelo de la población se utilizará para calcular valores de las características de la población (p. ej., la media μ) y para hacer inferencias sobre tales características.

Parámetro de una distribución de probabilidad

La función de masa de probabilidad de Bernoulli de la variable aleatoria X en el ejemplo 3.9 fue $p(0) = 0.8$ y $p(1) = 0.2$ porque 20% de todos los compradores seleccionó una computadora de escritorio. En otro almacén puede ser el caso que $p(0) = 0.9$ y $p(1) = 0.1$. Más generalmente, la función de masa de probabilidad de cualquier variable aleatoria de Bernoulli puede ser expresada en la forma $p(1) = \alpha$ y $p(0) = 1-\alpha$, donde $0 < \alpha < 1$. Puesto que la función de masa de probabilidad depende del valor particular de α, con frecuencia se escribe $p(x; \alpha)$ en lugar de sólo $p(x)$:

$$p(x;\alpha) = \begin{cases} 1-\alpha & \text{si } x = 0 \\ \alpha & \text{si } x = 1 \\ 0 & \text{de lo contrario} \end{cases} \tag{3.1}$$

Entonces cada opción de α en la expresión (3.1) da una función de masa de probabilidad diferente.

DEFINICIÓN

> Suponga que $p(x)$ depende de la cantidad que puede ser asignada a cualquiera de un número de valores posibles, y cada valor determina una distribución de probabilidad diferente. Tal cantidad se llama un **parámetro** de la distribución. El conjunto de todas las distribuciones de probabilidad para diferentes valores del parámetro se llama **familia** de distribuciones de probabilidad.

La cantidad α en la expresión (3.1) es un parámetro. Cada número diferente α entre 0 y 1 determina un miembro diferente de la familia de distribuciones de Bernoulli.

EJEMPLO 3.12 A partir de cierto tiempo se observan los nacimientos en un hospital hasta que nace un varón (B). Sea $p = P(B)$, suponga que los nacimientos sucesivos son independientes y defina la variable aleatoria X como $x =$ número de nacimientos observados. Entonces

$$p(1) = P(X = 1) = P(B) = p$$

$$p(2) = P(X = 2) = P(GB) = P(G) \cdot P(B) = (1-p)p$$

y

$$p(3) = P(X = 3) = P(GGB) = P(G) \cdot P(G) \cdot P(B) = (1-p)^2 p$$

Continuando de esta manera, emerge una fórmula general:

$$p(x) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, 3, \ldots \\ 0 & \text{de lo contrario} \end{cases} \tag{3.2}$$

El parámetro p puede asumir cualquier valor entre 0 y 1. La expresión (3.2) describe la familia de distribuciones *geométricas*. En el ejemplo del género, $p = 0.51$ podría ser apropiado, pero si estábamos buscando el primer hijo con sangre Rh positivo, entonces podríamos tener $p = 0.85$. ■

Función de distribución acumulada

Para algún valor fijo x a menudo se desea calcular la probabilidad de que el valor observado de X sea, cuando mucho, x. Por ejemplo, sea X el número de camas ocupadas en la sala de emergencia de un hospital en un momento dado del día; suponga que la función de masa de probabilidad de X está dada por

x	0	1	2	3	4
$p(x)$	0.20	0.25	0.30	0.15	0.10

Entonces la probabilidad de que máximo dos camas estén ocupadas es

$$P(X \leq 2) = p(0) + p(1) + p(2) = 0.75$$

Más aún, dado que $X \leq 2.7$ si y sólo si $X \leq 2$, también tenemos que $P(X \leq 2.7) = 0.75$, y de manera similar $P(X \leq 2.999) = 0.75$. Como 0 es el valor más pequeño posible $P(X \leq -1.5) = 0$, $P(X \leq -10) = 0$ y de hecho para cualquier valor negativo de x, $P(X \leq x) = 0$. Y dado que 4 es el máximo valor posible de X, $P(X \leq 4) = 1$, $P(X \leq 9.8) = 1$, y así sucesivamente.

De manera importante

$$P(X < 2) = p(0) + p(1) = 0.45 < 0.75 = P(X \leq 2)$$

porque la segunda parte de la desigualdad incluye la probabilidad del valor 2 de x, en tanto que la primera no. Más generalmente, $P(X < x) < P(X \leq x)$ siempre que x sea un valor posible de X. Más aún, $P(X \leq x)$ es una probabilidad calculable y bien definida para *cualquier* valor de x.

DEFINICIÓN

La **función de distribución acumulada** (fda) $F(x)$ de una variable aleatoria discreta X con función de masa de probabilidad $p(x)$ se define para cada número x como

$$F(x) = P(X \leq x) = \sum_{y:\, y \leq x} p(y) \qquad \textbf{(3.3)}$$

Para cualquier número x, $F(x)$ es la probabilidad de que el valor observado de X sea cuando mucho x.

EJEMPLO 3.13 Una tienda vende unidades de memoria flash, ya sea con 1 GB, 2 GB, 4 GB, 8 GB o 16 GB de memoria. La siguiente tabla muestra la distribución de Y = la cantidad de memoria en un disco comprado:

y	1	2	4	8	16
$p(y)$	0.05	0.10	0.35	0.40	0.10

Primero se determina $F(y)$ para cada uno de los cinco valores posibles de Y:

$$\begin{aligned} F(1) &= P(Y \leq 1) = P(Y = 1) = p(1) = 0.05 \\ F(2) &= P(Y \leq 2) = P(Y = 1 \text{ o } 2) = p(1) + p(2) = 0.15 \\ F(4) &= P(Y \leq 4) = P(Y = 1 \text{ o } 2 \text{ o } 4) = p(1) + p(2) + p(4) = 0.50 \\ F(8) &= P(Y \leq 8) = p(1) + p(2) + p(4) + p(8) = 0.90 \\ F(16) &= P(Y \leq 16) = 1 \end{aligned}$$

Ahora con cualquier otro número y, $F(y)$ será igual al valor de F en el valor más próximo posible de Y a la izquierda de y. Por ejemplo,

$$F(2.7) = P(Y \leq 2.7) = P(Y \leq 2) = F(2) = 0.15$$

$$F(7.999) = P(Y \leq 7.999) = P(Y \leq 4) = F(4) = 0.50$$

Si y es menor que 1, $F(y) = 0$ [por ejemplo, $F(0.58) = 0$], y si y es al menos 16, $F(y) = 1$ [por ejemplo, $F(25) = 1$]. La fda es, pues,

$$F(y) = \begin{cases} 0 & y < 1 \\ 0.05 & 1 \le y < 2 \\ 0.15 & 2 \le y < 4 \\ 0.50 & 4 \le y < 8 \\ 0.90 & 8 \le y < 16 \\ 1 & 16 \le y \end{cases}$$

En la figura 3.5 se muestra una gráfica de esta fda.

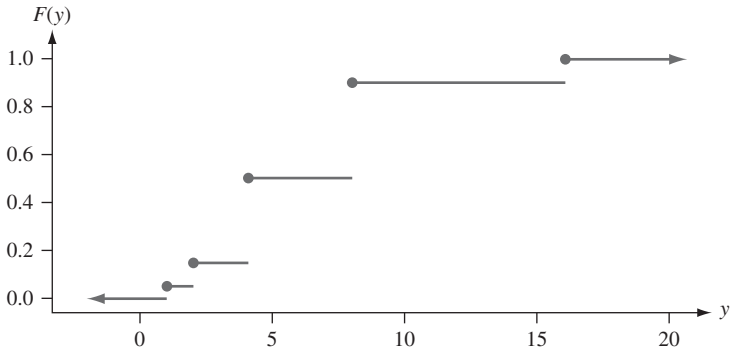


Figura 3.5 Gráfica de la función de distribución acumulada del ejemplo 3.13 ■

Para una variable aleatoria discreta X, la gráfica de $F(x)$ mostrará un salto con cada valor posible de X, y será plana entre los valores posibles. Tal gráfica se conoce como **función escalón**.

EJEMPLO 3.14 (Continuación del ejemplo 3.12)

La fmp de $X =$ *el número de nacimientos observados incluyendo el nacimiento del primer varón* tenía la forma

$$p(x) = \begin{cases} (1-p)^{x-1}p & x = 1, 2, 3, \ldots \\ 0 & \text{de lo contrario} \end{cases}$$

Para cualquier entero positivo x,

$$F(x) = \sum_{y \le x} p(y) = \sum_{y=1}^{x} (1-p)^{y-1}p = p\sum_{y=0}^{x-1} (1-p)^{y} \tag{3.4}$$

Para evaluar esta suma se utiliza el hecho de que la suma parcial de una serie geométrica es

$$\sum_{y=0}^{k} a^y = \frac{1-a^{k+1}}{1-a}$$

Utilizando esto en la ecuación (3.4), con $a = 1 - p$ y $k = x - 1$, se obtiene

$$F(x) = p \cdot \frac{1-(1-p)^x}{1-(1-p)} = 1-(1-p)^x \quad x \text{ es un entero positivo}$$

Puesto que F es una constante entre enteros positivos,

$$F(x) = \begin{cases} 0 & x < 1 \\ 1-(1-p)^{[x]} & x \ge 1 \end{cases} \tag{3.5}$$

donde $[x]$ es el entero más grande $\leq x$ (p. ej., $[2.7] = 2$). Así pues, si $p = 0.51$ como en el ejemplo de los nacimientos, entonces la probabilidad de tener que examinar cuando mucho cinco nacimientos para ver el primer varón es $F(5) = 1 - (0.49)^5 = 1 - 0.0282 = 0.9718$, mientras que $F(10) \approx 1.0000$. Esta función de distribución acumulada se ilustra en la figura 3.6.

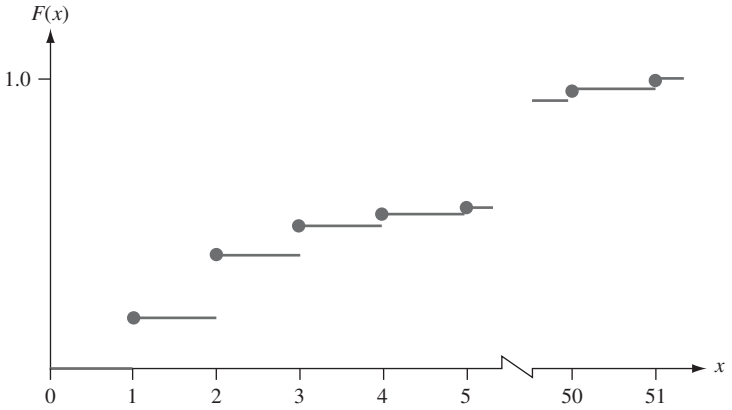


Figura 3.6 Gráfica de $F(x)$ para el ejemplo 3.14

En los ejemplos presentados hasta ahora, la función de distribución acumulada se dedujo de la función de masa de probabilidad. Este proceso puede ser invertido para obtener la función de masa de probabilidad a partir de la función de distribución acumulada, siempre que esta esté disponible. Por ejemplo, considere otra vez la variable aleatoria del ejemplo 3.7 (el número de computadoras usadas en un laboratorio); los valores posibles de X son 0, 1,…, 6. Entonces

$$\begin{aligned} p(3) &= P(X = 3) \\ &= [p(0) + p(1) + p(2) + p(3)] - [p(0) + p(1) + p(2)] \\ &= P(X \leq 3) - P(X \leq 2) \\ &= F(3) - F(2) \end{aligned}$$

Más generalmente, la probabilidad de que X quede dentro de un intervalo específico es fácil de obtener a partir de la función de distribución acumulada. Por ejemplo,

$$\begin{aligned} P(2 \leq X \leq 4) &= p(2) + p(3) + p(4) \\ &= [p(0) + \cdots + p(4)] - [p(0) + p(1)] \\ &= P(X \leq 4) - P(X \leq 1) \\ &= F(4) - F(1) \end{aligned}$$

Observe que $P(2 \leq X \leq 4) \neq F(4) - F(2)$. Esto es porque el valor 2 de X está incluido en $2 \leq X \leq 4$, así que no se desea restar su probabilidad. Sin embargo, $P(2 < X \leq 4) = F(4) - F(2)$ porque $X = 2$ no está incluido en el intervalo $2 < X \leq 4$.

DEFINICIÓN

Para dos números cualesquiera a y b con $a \leq b$,

$$P(a \leq X \leq b) = F(b) - F(a-)$$

donde "$a-$" representa el valor posible de X más grande que es estrictamente menor que a. En particular, si los únicos valores posibles son enteros, y si a y a son enteros, entonces

$$\begin{aligned} P(a \leq X \leq b) &= P(X = a \text{ o } a + 1 \text{ o} \ldots \text{ o } b) \\ &= F(b) - F(a - 1) \end{aligned}$$

Con $a = a$ se obtiene $P(X = a) = F(a) - F(a - 1)$ en este caso.

La razón de restar $F(a-)$ en lugar de $F(a)$ es que se desea incluir $P(X = a)$; $F(b) - F(a)$ da $P(a < X \leq b)$. Esta proposición se utilizará extensamente cuando se calculen las probabilidades binomial y de Poisson en las secciones 3.4 y 3.6.

EJEMPLO 3.15 Sea X = los días de ausencia por enfermedad de un empleado seleccionado al azar de una gran compañía durante un año particular. Si el número máximo de días de ausencia por enfermedad permisibles al año es de 14, los valores posibles de X son 0, 1,…, 14. Con $F(0) = 0.58$, $F(1) = 0.72$, $F(2) = 0.76$, $F(3) = 0.81$, $F(4) = 0.88$ y $F(5) = 0.94$,

$$P(2 \leq X \leq 5) = P(X = 2, 3, 4 \text{ o } 5) = F(5) - F(1) = 0.22$$

y

$$P(X = 3) = F(3) - F(2) = 0.05$$ ■

EJERCICIOS Sección 3.2 (11–28)

11. Sea X el número de estudiantes que se presentan en el horario de oficina de un profesor en un día en particular. Suponga que la función de masa de probabilidad de X es $p(0) = 0.20$, $p(1) = 0.25$, $p(2) = 0.30$, $p(3) = 0.15$ y $p(4) = 0.10$.

- **a.** Dibuje el histograma de probabilidad correspondiente.
- **b.** ¿Cuál es la probabilidad de que al menos dos estudiantes se presenten? ¿Y de que se presenten más de dos estudiantes?
- **c.** ¿Cuál es la probabilidad de que entre uno y tres estudiantes, inclusive, se presenten?
- **d.** ¿Cuál es la probabilidad de que se presente el profesor?

12. Las aerolíneas sobrevenden sus vuelos en algunas ocasiones. Suponga que para un avión de 50 asientos hay 55 pasajeros con boleto. Defina la variable aleatoria Y como el número de pasajeros con boleto comprado que llegan a tiempo para tomar el vuelo. La función de masa de probabilidad de Y aparece en la siguiente tabla.

y	45	46	47	48	49	50	51	52	53	54	55
$p(y)$	0.05	0.10	0.12	0.14	0.25	0.17	0.06	0.05	0.03	0.02	0.01

- **a.** ¿Cuál es la probabilidad de que la aerolínea acomode en el vuelo a todos los pasajeros que se presentan con boleto?
- **b.** ¿Cuál es la probabilidad de que no sean acomodados todos los pasajeros con un boleto?
- **c.** Si es la primera persona en la lista de espera (lo que significa que será la primera en subirse al vuelo si hay asientos disponibles después de que todos los pasajeros con boleto hayan sido acomodados), ¿cuál es la probabilidad de que tome el vuelo? ¿Cuál es la probabilidad si es la tercera persona en la lista de espera?

13. Una empresa de ventas por internet dispone de seis líneas telefónicas. Sea X el número de líneas en uso en un tiempo especificado. Suponga que la función de masa de probabilidad de X es la que se da en la siguiente tabla.

x	0	1	2	3	4	5	6
$p(x)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

Calcule la probabilidad de cada uno de los siguientes eventos.

- a. {cuando mucho tres líneas están en uso}
- b. {menos de tres líneas están en uso}
- c. {al menos tres líneas están en uso}
- d. {entre dos y cinco líneas, inclusive, están en uso}
- e. {entre dos y cuatro líneas, inclusive, no están en uso}
- f. {al menos cuatro líneas no están en uso}

14. El departamento de planeación de un condado requiere que un contratista presente una, dos, tres, cuatro o cinco formas (según la naturaleza del proyecto) para solicitar un permiso de construcción. Sea Y = número de formas requeridas del siguiente solicitante. Se sabe que la probabilidad de que se requieran y formas es proporcional a y, es decir, $p(y) = ky$ con $y = 1,\ldots, 5$.

- **a.** ¿Cuál es el valor de k? [*Sugerencia:* $\sum_{y=1}^{5} p(y) = 1$]
- **b.** ¿Cuál es la probabilidad de que cuando mucho se requieran tres formas?
- **c.** ¿Cuál es la probabilidad de que se requieran entre dos y cuatro formas (inclusive)?
- **d.** ¿Podría ser $p(y) = y^2/50$ con $y = 1,\ldots, 5$ como la función de masa de probabilidad de Y?

15. Muchos fabricantes cuentan con programas de control de calidad que incluyen la inspección de los materiales recibidos en busca de defectos. Suponga que un fabricante de computadoras recibe tarjetas madre en lotes de cinco. Se seleccionan dos tarjetas de cada lote para inspeccionarlas. Se pueden representar los posibles resultados del proceso de selección por pares. Por ejemplo, el par (1, 2) representa la selección de las tarjetas 1 y 2 para inspección.

- **a.** Mencione los diez posibles resultados diferentes.
- **b.** Suponga que las tarjetas 1 y 2 son las únicas defectuosas en un lote de cinco. Dos tarjetas tienen que ser selecciona-

das al azar. Defina X como el número de tarjetas defectuosas observadas entre las inspeccionadas. Encuentre la distribución de probabilidad de X.

c. Sea $F(x)$ la función de distribución acumulada de X. Primero determine $F(0) = P(X \leq 0)$, $F(1)$ y $F(2)$; luego obtenga $F(x)$ para todas las demás x.

16. Algunas partes de California son particularmente propensas a los temblores. Suponga que en un área metropolitana, 25% de todos los dueños de una casa están asegurados contra daños provocados por terremotos. Se seleccionan al azar cuatro de ellos; sea X el número, entre estos cuatro, que están asegurados contra terremotos.

a. Encuentre la distribución de probabilidad de X. [*Sugerencia*: Sea S un propietario de casa asegurado y F uno no asegurado. Entonces un posible resultado es $SFSS$, con probabilidad $(0.25)(0.75)(0.25)(0.25)$ y el valor 3 de X asociado. [Existen otros 15 resultados.]

b. Trace el histograma de probabilidad correspondiente.

c. ¿Cuál es el valor más probable de X?

d. ¿Cuál es la probabilidad de que al menos dos de los cuatro seleccionados estén asegurados contra terremotos?

17. El voltaje de una batería nueva puede ser aceptable (A) o inaceptable (I). Una linterna requiere dos baterías, así que las baterías serán seleccionadas independientemente y probadas hasta encontrar dos aceptables. Suponga que 90% de todas las baterías tiene voltajes aceptables. Sea Y el número de baterías que deben ser probadas.

a. ¿Cuál es $p(2)$, es decir, $P(Y = 2)$?

b. ¿Cuál es $p(3)$? [*Sugerencia:* Existen dos resultados diferentes que producen $Y = 3$.]

c. Para tener $Y = 5$, ¿qué debe ser cierto de la quinta batería seleccionada? Mencione los cuatro resultados con los cuales $Y = 5$ y luego determine $p(5)$.

d. Use el patrón de sus respuestas en los incisos a)−c) para obtener una fórmula general para $p(y)$.

18. Dos dados de seis caras son lanzados al aire en forma independiente. Sea M = el máximo de los dos lanzamientos (por tanto $M(1,5) = 5$, $M(3,3) = 3$, etc.).

a. ¿Cuál es la función de masa de probabilidad de M? [*Sugerencia:* Primero determine $p(1)$, luego $p(2)$ y así sucesivamente.]

b. Determine la función de distribución acumulada de M y grafíquela.

19. Una biblioteca se suscribe a dos diferentes revistas de noticias semanales, cada una de las cuales se supone que llega en el correo de los miércoles. En realidad, cada una puede llegar en miércoles, jueves, viernes o sábado. Suponga que las dos llegan independientemente una de otra y para cada una $P(\text{mié}) = 0.3$, $P(\text{jue}) = 0.4$, $P(\text{vie}) = 0.2$ y $P(\text{sáb}) = 0.1$. Sea Y = el número de días que se suceden, posteriores al miércoles, para que ambas revistas lleguen (por tanto los posibles valores de Y son 0, 1, 2 o 3). Calcule la función de masa de probabilidad de Y. [*Sugerencia:* Hay 16 posibles resultados: $Y(M, M) = 0$, $Y(V, J) = 2$, y así sucesivamente.]

20. Tres parejas y dos individuos solteros han sido invitados a un seminario de inversión y han aceptado asistir. Suponga que la probabilidad de que cualquier pareja o individuo particular llegue tarde es de 0.4 (una pareja viajará en el mismo vehículo, así que ambos llegarán a tiempo, o bien ambos llegarán tarde). Suponga que diferentes parejas e individuos llegan puntuales o tarde, independientemente unos de otros. Sea X = el número de personas que llegan tarde al seminario.

a. Determine la función de masa de probabilidad de X. [Sugerencia: designe las tres parejas #1, #2 y #3 y los dos individuos #4 y #5.]

b. Obtenga la función de distribución acumulada de X y úsela para calcular $P(2 \leq X \leq 6)$.

21. Suponga que lee los números de este año del *New York Times* y que anota cada número en el que aparece un artículo sobre el ingreso de un oficial ejecutivo en jefe, el número de cajas de vino producidas por una compañía vinícola, la contribución caritativa total de un político durante el año fiscal previo, la edad de una celebridad y así sucesivamente. Ahora enfóquese en el primer dígito de cada número, el cual podría ser 1, 2,..., 8 o 9. Usted podría pensar que es igual de probable que el primer dígito X de un número seleccionado al azar sea cualquiera de las nueve posibilidades (una distribución uniforme discreta). Sin embargo, mucha evidencia empírica, así como también algunos argumentos teóricos sugieren una distribución de probabilidad alternativa llamada *ley de Benford*:

$$p(x) = P(\text{el primer dígito es } x) = \log_{10}\left(\frac{x+1}{x}\right) \quad x = 1, 2, \ldots, 9$$

a. Sin calcular probabilidades individuales de esta fórmula, demuestre que esta especifica una función legítima de masa de probabilidad.

b. Ahora calcule las probabilidades individuales y compare con la correspondiente distribución uniforme discreta.

c. Obtenga la función de distribución acumulada de X.

d. Utilizando la función de distribución acumulada, ¿cuál es la probabilidad de que el primer dígito sea cuando mucho 3? ¿Al menos 5?

[*Nota*: La ley de Benford es la base de algunos procedimientos de auditoría utilizados para detectar fraudes en reportes financieros, por ejemplo, por el Servicio de Ingresos Internos.]

22. Remítase al ejercicio 13 y calcule y trace la gráfica de la función de distribución acumulada $F(x)$. Luego utilícela para calcular las probabilidades de los eventos dados en los incisos a)−d) de dicho problema.

23. Una organización de protección al consumidor que habitualmente evalúa automóviles nuevos reporta el número de defectos importantes encontrados en cada automóvil examinado. Sea X el número de defectos importantes en un auto seleccionado al azar de cierto tipo. La función de distribución acumulada de X es la siguiente:

$$F(x) = \begin{cases} 0 & x < 0 \\ 0.06 & 0 \leq x < 1 \\ 0.19 & 1 \leq x < 2 \\ 0.39 & 2 \leq x < 3 \\ 0.67 & 3 \leq x < 4 \\ 0.92 & 4 \leq x < 5 \\ 0.97 & 5 \leq x < 6 \\ 1 & 6 \leq x \end{cases}$$

Calcule las siguientes probabilidades directamente con la función de distribución acumulada:

a. $p(2)$, es decir, $P(X = 2)$ **b.** $P(X > 3)$
c. $P(2 \le X \le 5)$ **d.** $P(2 < X < 5)$

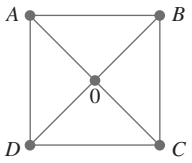
24. Una compañía de seguros ofrece a sus asegurados varias opciones diferentes de pago de primas. Para un asegurado seleccionado al azar, sea X = el número de meses entre pagos sucesivos. La función de distribución acumulada es la siguiente:

$$F(x) = \begin{cases} 0 & x < 1 \\ 0.30 & 1 \le x < 3 \\ 0.40 & 3 \le x < 4 \\ 0.45 & 4 \le x < 6 \\ 0.60 & 6 \le x < 12 \\ 1 & 12 \le x \end{cases}$$

a. ¿Cuál es la función de masa de probabilidad de X?
b. Con sólo la función de distribución acumulada, calcule $P(3 \le X \le 6)$ y $P(4 \le X)$.

25. En el ejemplo 3.12, sea Y = el número de niñas nacidas antes de que termine el experimento. Con $p = P(B)$ y $1 - p = P(G)$, ¿cuál es la función de masa de probabilidad de Y? [*Sugerencia*: Primero haga una lista con los posibles valores de Y, inicie con el más pequeño y continúe hasta que encuentre una fórmula general.]

26. Álvaro Singer vive en 0 en el diagrama adjunto y sus cuatro amigos viven en *A, B, C* y *D*. Un día Álvaro decide visitarlos, así que lanza una moneda al aire dos veces para decidir a cuál de los cuatro visitar. Luego de estar en la casa de uno de sus amigos, regresará a su propia casa o bien irá a una de las dos casas adyacentes (tales como 0, A o C, cuando está en B); la probabilidad de las tres posibilidades es 1/3. De este modo, Álvaro continúa con las visitas a sus amigos hasta que regresa a casa.



a. Sea X = el número de veces que Álvaro visita a un amigo. Obtenga la función de masa de probabilidad de X.
b. Sea Y = el número de segmentos de línea recta que Álvaro recorre (incluidos los que conducen a 0, o que parten de ahí). ¿Cuál es la función de masa de probabilidad de Y?
c. Suponga que sus amigas viven en A y C y sus amigos en B y D. Si Z = el número de visitas a sus amigas, ¿cuál es la función de masa de probabilidad de Z?

27. Después de que todos los estudiantes salieron del salón de clases, el profesor de estadística nota que cuatro ejemplares del texto fueron olvidados bajo los escritorios. Al inicio de la siguiente clase el profesor distribuye los cuatro libros al azar a cada uno de los cuatro estudiantes (1, 2, 3 y 4) que dicen haber olvidado sus libros. Un posible resultado es que 1 reciba el libro de 2, que 2 reciba el libro de 4, que 3 reciba su propio libro y que 4 reciba el libro de 1. Este resultado puede ser abreviado como (2, 4, 3, 1).
a. Mencione los otros 23 resultados posibles.
b. Si X es el número de estudiantes que reciben su propio libro, determine la función de masa de probabilidad de X.

28. Demuestre que la función de distribución acumulada de $F(x)$ es no decreciente; es decir, $x_1 < x_2$ implica que $F(x_1) \le F(x_2)$. ¿En qué condición será $F(x_1) = F(x_2)$?

3.3 Valores esperados

Considere una universidad que tiene 15 000 estudiantes y sea X = el número de cursos en los cuales está inscrito un estudiante seleccionado al azar. La función de masa de probabilidad de X se determina como sigue. Como $p(1) = 0.01$, se sabe que $(0.01) \cdot (15\,000) = 150$ de los estudiantes están inscritos en un curso y lo mismo para los otros valores de x.

x	1	2	3	4	5	6	7
$p(x)$	0.01	0.03	0.13	0.25	0.39	0.17	0.02
Número registrado	150	450	1950	3750	5850	2550	300

(3.6)

El número promedio de cursos por estudiante o el valor promedio de X en la población se obtiene al calcular el número total de cursos tomados por todos los estudiantes y al dividir entre el número total de estudiantes. Como cada uno de los 150 estudiantes está tomando un curso, estos 150 contribuyen con 150 cursos al total. Asimismo, 450 estudiantes contribuyen con 2(450) cursos y así sucesivamente. El valor promedio de la población X es entonces

$$\frac{1(150) + 2(450) + 3(1950) + \cdots + 7(300)}{15\,000} = 4.57 \qquad \textbf{(3.7)}$$

Puesto que 150/15 000 = 0.01 = $p(1)$, 450/15 000 = 0.03 = $p(2)$ y así sucesivamente, una expresión alternativa para (3.7) es

$$1 \cdot p(1) + 2 \cdot p(2) + \cdots + 7 \cdot p(7) \tag{3.8}$$

La expresión (3.8) muestra que para calcular el valor promedio de la población X sólo se necesitan los valores posibles de X junto con sus probabilidades (proporciones). En particular el tamaño de la población no viene al caso mientras la función de masa de probabilidad esté dada por (3.6). El valor promedio o medio de X es entonces el promedio *ponderado* de los posibles valores 1,…, 7, donde las ponderaciones son las probabilidades de esos valores.

Valor esperado de X

DEFINICIÓN

Sea X una variable aleatoria discreta con un conjunto de valores posibles D y una función de masa de probabilidad $p(x)$. El **valor esperado** o **valor medio** de X, denotado por $E(X)$ o μ_X o sólo μ, es

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

EJEMPLO 3.16 Para la función de masa de probabilidad de X = *número de cursos* en (3.6),

$$\begin{aligned}\mu &= 1 \cdot p(1) + 2 \cdot p(2) + \cdots + 7 \cdot p(7)\\ &= (1)(0.01) + 2(0.03) + \cdots + (7)(0.02)\\ &= 0.01 + 0.06 + 0.39 + 1.00 + 1.95 + 1.02 + 0.14 = 4.57\end{aligned}$$

Si se piensa en la población como compuesta de los valores 1, 2,…, 7 de X, entonces $\mu = 4.57$ es la *media de la población*. En consecuencia, a menudo se hará referencia a μ como la media de la población en lugar de la media de X en la población. Tenga en cuenta que aquí μ no es 4, el promedio normal de 1,…, 7, porque la distribución pone más peso en 4, 5 y 6 que en otros valores de X. ■

En el ejemplo 3.16 el valor esperado μ fue 4.57, el cual no es un valor posible de X. La palabra *esperado* deberá interpretarse con precaución porque no se esperaría ver un valor X de 4.57 cuando se selecciona un solo estudiante.

EJEMPLO 3.17 Exactamente después de nacer cada recién nacido es evaluado en una escala llamada escala de Apgar. Las evaluaciones posibles son 0, 1,…, 10; la evaluación del niño es determinada por color, tono muscular, esfuerzo para respirar, ritmo cardiaco e irritabilidad refleja (la mejor evaluación posible es 10). Sea X la evaluación Apgar de un niño seleccionado al azar que nacerá en cierto hospital durante el siguiente año y suponga que la función de masa de probabilidad de X es

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.002	0.001	0.002	0.005	0.02	0.04	0.18	0.37	0.25	0.12	0.01

Entonces el valor medio de X es

$$\begin{aligned}E(X) = \mu &= 0(0.002) + 1(0.001) + 2(0.002)\\ &\quad + \cdots + 8(0.25) + 9(0.12) + 10(0.01)\\ &= 7.15\end{aligned}$$

De nuevo, μ no es un valor posible de la variable X. Además, debido a que la variable se refiere a un niño que aún no ha nacido, no existe ninguna población concreta a la cual podría referirse μ. En cambio, la función de masa de probabilidad se considera un modelo de una población conceptual compuesta por los valores 0, 1, 2,…, 10. El valor medio de esta población conceptual es entonces $\mu = 7.15$. ■

EJEMPLO 3.18 Sea $X = 1$ si un vehículo seleccionado al azar aprueba un diagnóstico de emisiones y $X = 0$ si no. Entonces X es una variable aleatoria de Bernoulli con función de masa de probabilidad $p(1) = p$ y $p(0) = 1 - p$, a partir de la cual $E(X) = 0 \cdot p(0) + 1 \cdot p(1) = 0(1 - p) + 1(p) = p$. Es decir, el valor esperado de X es exactamente la probabilidad de que X tome el valor de 1. Si se conceptualiza una población compuesta de ceros en la proporción $1 - p$ y números 1 en la proporción p, entonces el promedio de la población es $\mu = p$. ■

EJEMPLO 3.19 La forma general de la función de masa de probabilidad de $X =$ *número de bebés nacidos hasta el primer varón incluido* es

$$p(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, 3, \ldots \\ 0 & \text{de lo contrario} \end{cases}$$

De acuerdo con la definición,

$$E(X) = \sum_{D} x \cdot p(x) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = p\sum_{x=1}^{\infty}\left[-\frac{d}{dp}(1-p)^{x}\right] \tag{3.9}$$

Si se intercambia el orden en que se evalúan la derivada y la suma, la suma es la de una serie geométrica. Una vez que se calcula la suma, se saca la derivada y el resultado final es $E(X) = 1/p$. Si p se aproxima a 1, se espera ver que nazca un varón muy pronto, mientras que si p se aproxima a 0, se esperan muchos nacimientos antes del primer varón. Con $p = 0.5$, $E(X) = 2$. ■

Hay otra interpretación frecuentemente utilizada de μ. Considere la posibilidad de observar un primer valor x_1 de X, un segundo valor x_2, un tercer valor x_3, y así sucesivamente. Después de hacer esto un gran número de veces se calcula el promedio de la muestra de las x_i observadas. Este promedio usualmente será muy cercano a μ. Es decir, μ puede interpretarse como el promedio a largo plazo del valor observado de X cuando el experimento se realiza en varias ocasiones. En el ejemplo 3.17 el promedio a largo plazo de Apgar es $\mu = 7.15$.

EJEMPLO 3.20 Sea X el número de entrevistas a las que se presenta un estudiante antes de conseguir un trabajo, con la función de masa de probabilidad

$$p(x) = \begin{cases} k/x^2 & x = 1, 2, 3, \ldots \\ 0 & \text{de lo contrario} \end{cases}$$

donde $k = \pi^2/6$ asegura que se elige de modo que $\Sigma p(x) = 1$ (el valor de k es el resultado de una serie de Fourier). El valor esperado de X es:

$$\mu = E(X) = \sum_{x=1}^{\infty} x \cdot \frac{k}{x^2} = k\sum_{x=1}^{\infty}\frac{1}{x} \tag{3.10}$$

La suma del lado derecho de la ecuación (3.10) es la famosa serie armónica de matemáticas y se puede demostrar que es igual a ∞. En este caso $E(X)$ no es finita porque $p(x)$ no disminuye suficientemente rápido a medida que x se incrementa; los estadísticos dicen que la distribución de probabilidad de X tiene "una cola gruesa". Si se selecciona una secuencia de valores X utilizando esta distribución el promedio muestral no se establecerá en un número finito, sino que tenderá a crecer sin límite.

Los estadísticos utilizan la frase "colas gruesas" en conexión con cualquier distribución con una gran cantidad de probabilidad alejada de μ (por tanto, las colas gruesas no requieren $\mu = \infty$). Tales colas gruesas vuelven difícil hacer inferencias respecto a μ. ■

Valor esperado de una función

A menudo interesará poner atención en el valor esperado de alguna función $h(X)$ en lugar de sólo en $E(X)$.

EJEMPLO 3.21 Suponga que una librería adquiere diez ejemplares de un libro a \$6.00 cada uno para venderlos a \$12.00 en el entendimiento de que al final de un periodo de 3 meses cualquier ejemplar no vendido puede ser canjeado por \$2.00. Si X = el número de ejemplares vendidos, entonces el ingreso neto $= h(X) = 12X + 2(10 - X) - 60 = 10X - 40$. En esta situación podríamos estar interesados no sólo en el número esperado de copias vendidas [por ejemplo, $E(X)$], sino también en la ganancia neta esperada, esto es, el valor esperado de una función particular de X. ■

El siguiente ejemplo sugiere una forma fácil de calcular el valor esperado de $h(X)$.

EJEMPLO 3.22 El costo de cierta prueba de diagnóstico de un vehículo depende del número de cilindros X en el motor. Suponga que la función de costo está dada por $h(X) = 20 + 3X + 0.5X^2$. Puesto que X es una variable aleatoria, también lo es $Y = h(X)$. Las funciones de masa de probabilidad de X y Y son las siguientes

x	4	6	8	$\Rightarrow$	y	40	56	76
$p(x)$	0.5	0.3	0.2		$p(y)$	0.5	0.3	0.2

Con D^* denotando posibles valores de Y,

$$\begin{aligned} E(Y) = E[h(X)] &= \sum_{D^*} y \cdot p(y) \\ &= (40)(0.5) + (56)(0.3) + (76)(0.2) \\ &= h(4) \cdot (0.5) + h(6) \cdot (0.3) + h(8) \cdot (0.2) \\ &= \sum_{D} h(x) \cdot p(x) \end{aligned} \tag{3.11}$$

De acuerdo con la ecuación (3.11) no fue necesario determinar la función de masa de probabilidad de Y para obtener $E(Y)$; en su lugar, el valor esperado deseado es un promedio ponderado de los posibles valores de $h(x)$ (en lugar de x). ■

PROPOSICIÓN

Si la variable aleatoria X tiene un conjunto de posibles valores D y una función de masa de probabilidad $p(x)$, entonces el valor esperado de cualquier función $h(X)$, denotada por $E[h(X)]$ o $\mu_{h(X)}$, se calcula con

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

Esto es, $E[h(X)]$ se calcula del mismo modo que $E(X)$, excepto que $h(x)$ sustituye a x.

EJEMPLO 3.23 Una tienda de computadoras adquirió tres computadoras de un cierto tipo en \$500 cada una. Las venderá a \$1000 cada una. El fabricante se comprometió a comprar a su vez en \$200 cualquier computadora que no se haya vendido después de un periodo especificado. Sea X el número de computadoras vendidas y suponga que $p(0) = 0.1$, $p(1) = 0.2$, $p(2) = 0.3$ y $p(3) = 0.4$. Con $h(X)$ denotando la utilidad asociada con la venta de X unidades,

la información dada implica que $h(X)$ = ingreso costo = $1000X + 200(3 - X) - 1500 = 800X - 900$. La utilidad esperada es entonces

$$\begin{aligned} E[h(X)] &= h(0) \cdot p(0) + h(1) \cdot p(1) + h(2) \cdot p(2) + h(3) \cdot p(3) \\ &= (-900)(0.1) + (-100)(0.2) + (700)(0.3) + (1500)(0.4) \\ &= \$700 \end{aligned}$$

■

Reglas de valor esperado

La función de interés $h(X)$ es con bastante frecuencia una función lineal $aX + b$. En este caso, $E[h(X)]$ es fácil de calcular a partir de $E(X)$.

DEFINICIÓN

$$E(aX + b) = a \cdot E(X) + b$$

(O, utilizando una notación alternativa, $\mu_{aX+b} = a \cdot \mu_X + b$)

Parafraseando, el valor esperado de una función lineal es igual a la función lineal evaluada con el valor esperado $E(X)$. Puesto que $h(X)$ en el ejemplo 3.23 es lineal y $E(X) = 2$, $E[h(x)] = 800(2) - 900 = \$700$, como antes.

Comprobación

$$\begin{aligned} E(aX + b) = \sum_D (ax + b) \cdot p(x) &= a\sum_D x \cdot p(x) + b\sum_D p(x) \\ &= aE(X) + b \end{aligned}$$

■

Dos casos especiales de la proposición producen dos reglas importantes de valor esperado.

1. Con cualquier constante a, $E(aX) = a \cdot \mathrm{E}(X)$ (considere $b = 0$).
2. Con cualquier constante b, $E(X + b) = E(X) + b$ (considere $a = 1$). **(3.12)**

Multiplicar X por una constante a por lo general cambia la unidad de medición, por ejemplo, de pulgadas a centímetros, donde $a = 2.54$. La regla 1 dice que el valor esperado en las nuevas unidades es igual al valor esperado en las viejas unidades multiplicado por el factor de conversión a. Asimismo, si se agrega una constante b a cada valor posible de X, entonces el valor esperado se desplazará en esa misma cantidad constante.

Varianza de X

El valor esperado de X describe dónde está centrada la distribución de probabilidad. Mediante la analogía física de colocar una masa puntual $p(x)$ en el valor x sobre un eje unidimensional que estuviera soportado por un fulcro colocado en μ, el eje no tendería a ladearse. Esto se ilustra para dos distribuciones diferentes en la figura 3.7.

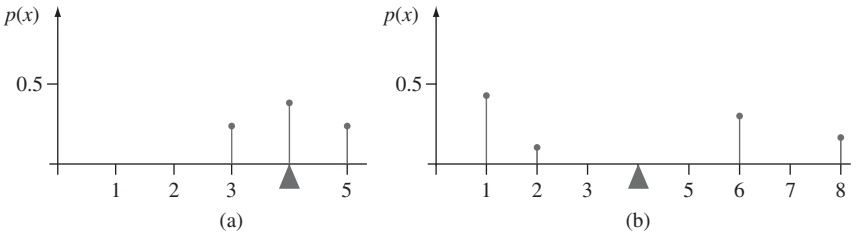


Figura 3.7 Dos diferentes distribuciones de probabilidad con $\mu = 4$

Aunque ambas distribuciones ilustradas en la figura 3.7 tienen el mismo centro μ, la distribución de la figura 3.7(b) tiene una mayor dispersión o variabilidad que la de la figura 3.7(a). Se utilizará la varianza de X para evaluar la cantidad de variabilidad en (la distribución de) X, del mismo modo que se utilizó s^2 en el capítulo 1 para medir la variabilidad en una muestra.

DEFINICIÓN

Sea $p(x)$ la función de masa de probabilidad de X y μ su valor esperado. En ese caso la varianza de X, denotada por $V(X)$ o σ_X^2, o simplemente σ^2, es

$$V(X) = \sum_D (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

La **desviación estándar** (DE) de X es

$$\sigma_X = \sqrt{\sigma_X^2}$$

La cantidad $h(X) = (X - \mu)^2$ es la desviación al cuadrado de X respecto a su media y σ^2 es la desviación al cuadrado esperada, es decir, el promedio ponderado de las desviaciones al cuadrado, donde las ponderaciones son probabilidades de la distribución. Si la mayor parte de la distribución de probabilidad está cerca de μ, entonces σ^2 será relativamente pequeña. Sin embargo, si existen valores x alejados de μ con una gran $p(x)$, en ese caso σ^2 será bastante grande. A grandes rasgos, σ se puede interpretar como el tamaño de una desviación representativa del valor medio μ. Así que si $\sigma = 10$, entonces en una larga secuencia de valores observados X, algunos se apartarán de μ por más de 10, mientras que otros estarán más cerca de la media que eso, una desviación típica de la media será del orden de 10.

EJEMPLO 3.24 En una biblioteca la cantidad de DVD que puede sacar una persona a la vez es de 6. Tenga en cuenta a quienes solamente echan un vistazo a los DVD y sea X el número de cintas que pide prestadas una persona seleccionada al azar. La función de masa de probabilidad de X es la siguiente:

x	1	2	3	4	5	6
$p(x)$	0.30	0.25	0.15	0.05	0.10	0.15

Es fácil ver que el valor esperado de X es $\mu = 2.85$. La varianza de X es entonces

$$\begin{aligned} V(X) = \sigma^2 &= \sum_{x=1}^{6} (x - 2.85)^2 \cdot p(x) \\ &= (1 - 2.85)^2(0.30) + (2 - 2.85)^2(0.25) + \cdots + (6 - 2.85)^2(0.15) = 3.2275 \end{aligned}$$

La desviación estándar de X es $\sigma = \sqrt{3.2275} = 1.800$. ■

Cuando la función de masa de probabilidad $p(x)$ especifica un modelo matemático para la distribución de los valores de la población, tanto σ^2 como σ miden la dispersión de los valores en la población; σ^2 es la varianza de la población y σ es su desviación estándar.

Fórmula abreviada para σ^2

El número de operaciones aritméticas necesarias para calcular σ^2 puede reducirse si se utiliza una fórmula alternativa.

PROPOSICIÓN

$$V(X) = \sigma^2 = \left[\sum_D x^2 \cdot p(x)\right] - \mu^2 = E(X^2) - [E(X)]^2$$

Al utilizar esta fórmula, $E(X^2)$ se calcula primero sin ninguna sustracción; luego se calcula $E(X)$, se eleva al cuadrado y se resta (una vez) de $E(X^2)$.

EJEMPLO 3.25 (Continuación del ejemplo 3.24)

La función de masa de probabilidad de la cantidad X de DVD prestados se dio como $p(1) = 0.30$, $p(2) = 0.25$, $p(3) = 0.15$, $p(4) = 0.05$, $p(5) = 0.10$ y $p(6) = 0.15$, a partir de las cuales $\mu = 2.85$ y

$$E(X^2) = \sum_{x=1}^{6} x^2 \cdot p(x) = (1^2)(0.30) + (2^2)(0.25) + \cdots + (6^2)(0.15) = 11.35$$

Por tanto, $\sigma^2 = 11.35 - (2.85)^2 = 3.2275$, como se obtuvo previamente de la definición. ■

Demostración de la fórmula abreviada Desarrolle $(x - \mu)^2$ en la definición de σ^2 para obtener $x^2 - 2\mu x + \mu^2$, y luego lleve Σ a cada uno de los tres términos:

$$\begin{aligned}\sigma^2 &= \sum_D x^2 \cdot p(x) - 2\mu \cdot \sum_D x \cdot p(x) + \mu^2 \sum_D p(x) \\ &= E(X^2) - 2\mu \cdot \mu + \mu^2 = E(X^2) - \mu^2\end{aligned}$$ ■

Varianza de una función lineal

La varianza de $h(X)$ es el valor esperado de la diferencia al cuadrado entre $h(X)$ y su valor esperado:

$$V[h(X)] = \sigma^2_{h(X)} = \sum_D \{h(x) - E[h(X)]\}^2 \cdot p(x) \qquad \textbf{(3.13)}$$

Cuando $h(X) = aX + b$, una función lineal,

$$h(x) - E[h(X)] = ax + b - (a\mu + b) = a(x - \mu)$$

Al sustituir esto en la ecuación (3.13) se obtiene una relación simple entre $V[h(X)]$ y $V(X)$:

PROPOSICIÓN

$$V(aX + b) = \sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \text{ y } \quad \sigma_{aX+b} = |a| \cdot \sigma_X$$

En particular,

$$\sigma_{aX} = |a| \cdot \sigma_X, \quad \sigma_{X+b} = \sigma_X \qquad (3.14)$$

El valor absoluto es necesario porque a podría ser negativa, no obstante una desviación estándar no puede serlo. Casi siempre multiplicar por a corresponde a un cambio en la unidad de medición (p. ej., kg a lb o dólares a euros). De acuerdo con la primera relación en (3.14) la desviación estándar en la nueva unidad es la desviación estándar original multiplicada por el factor de conversión. La segunda relación dice que la adición o sustracción de una constante no impacta la variabilidad, simplemente desplaza la distribución a la derecha o a la izquierda.

EJEMPLO 3.26 En el problema de ventas de computadoras del ejemplo 3.23, $E(X) = 2$ y

$$E(X^2) = (0)^2(0.1) + (1)2(0.2) + (2)^2(0.3) + (3)^2(0.4) = 5$$

así que $V(X) = 5 - (2)^2 = 1$. La función de utilidad $h(X) = 800X - 900$ tiene entonces varianza $(800)^2 \cdot V(X) = (640\,000)(1) = 640\,000$ y desviación estándar 800. ■

EJERCICIOS **Sección 3.3 (29–45)**

29. La función de masa de probabilidad de la cantidad de memoria X (GB) en una unidad flash comprada se dio en el ejemplo 3.13 como

x	1	2	4	8	16
$p(x)$	0.05	0.10	0.35	0.40	0.10

Calcule lo siguiente:

a. $E(X)$.

b. $V(X)$ directamente a partir de la definición.

c. La desviación estándar de X.

d. $V(X)$ mediante la fórmula abreviada.

30. Se selecciona al azar a un individuo que tiene asegurado su automóvil con una compañía. Sea Y el número de infracciones de tránsito por las que el individuo fue citado durante los últimos tres años. La función de masa de probabilidad de Y es

y	0	1	2	3
$p(y)$	0.60	0.25	0.10	0.05

a. Calcule $E(Y)$.

b. Suponga que un individuo con Y infracciones incurre en un recargo de $\$100Y^2$. Calcule el monto esperado del recargo.

31. Remítase al ejercicio 12 y calcule $V(Y)$ y σ_Y. Determine entonces la probabilidad de que Y esté dentro de una desviación estándar 1 de su valor medio.

32. Un distribuidor de enseres para el hogar vende tres modelos de congeladores verticales con una capacidad de 16, 18 y 20 pies cúbicos, respectivamente. Sea X = la cantidad de espacio de almacenamiento adquirido por el siguiente cliente que compre un congelador. Suponga que X tiene la función de masa de probabilidad

x	16	18	20
$p(x)$	0.2	0.5	0.3

a. Calcule $E(X)$, $E(X^2)$ y $V(X)$.

b. Si el precio de un congelador de X pies cúbicos de capacidad es $70X - 650$, ¿cuál es el precio que se espera que pague el siguiente cliente que compre un congelador?

c. ¿Cuál es la varianza del precio pagado por el siguiente cliente?

d. Suponga que aunque la capacidad nominal de un congelador es X, la real es $h(X) = X - 0.008X^2$. ¿Cuál es la capacidad real esperada del congelador adquirido por el siguiente cliente?

33. Sea X una variable aleatoria de Bernoulli con función de masa de probabilidad como la del ejemplo 3.18.

a. Calcule $E(X^2)$.

b. Demuestre que $V(X) = p(1 - p)$.

c. Calcule $E(X^{79})$.

34. Suponga que el número de plantas de un tipo particular encontradas en una región rectangular de muestreo (llamada cuadrado por los ecologistas) en cierta área geográfica es una variable aleatoria X con función de masa de probabilidad

$$p(x) = \begin{cases} c/x^3 & x = 1, 2, 3, \ldots \\ 0 & \text{de lo contrario} \end{cases}$$

¿Es $E(X)$ finita? Justifique su respuesta (esta es otra distribución que los estadísticos llamarían de cola gruesa).

35. Un pequeño mercado ordena ejemplares de cierta revista para su exhibidor de revistas cada semana. Sea X = demanda de la revista, con función de masa de probabilidad

x	1	2	3	4	5	6
$p(x)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{2}{15}$

Suponga que el propietario de la tienda paga $2.00 por cada ejemplar de la revista y el precio para los consumidores es de $4.00. Si las revistas que se quedan al final de la semana no tienen valor de recuperación, ¿es mejor solicitar tres o cuatro ejemplares de la revista? [*Sugerencia*: Para una orden de tres o cuatro ejemplares, exprese el ingreso neto como una función de la demanda X y luego calcule el ingreso esperado.]

36. Sea X el daño incurrido (en dólares) en un tipo de accidente durante un año dado. Valores posibles de X son 0, 1000, 5000 y 10 000 con probabilidades de 0.8, 0.1, 0.08 y 0.02, respectivamente. Una compañía particular ofrece una póliza con deducible de $500. Si la compañía desea que su utilidad esperada sea de $100, ¿qué cantidad de prima deberá cobrar?

37. Los n candidatos para un trabajo fueron clasificados como 1, 2, 3,…, n. Sea X = la clasificación de un candidato seleccionado al azar, de modo que X tenga la función de masa de probabilidad

$$p(x) = \begin{cases} 1/n & x = 1, 2, 3, \ldots, n \\ 0 & \text{de lo contrario} \end{cases}$$

(esta se llama *distribución uniforme discreta*). Calcule $E(X)$ y $V(X)$ mediante la fórmula abreviada. [*Sugerencia*: La suma de los primeros n enteros positivos es $n(n + 1)/2$, mientras que la suma de sus cuadrados es $n(n + 1)(2n + 1)/6$.]

38. Los posibles valores de X, el número de componentes en un sistema sometido a reparación que deben ser reemplazados son 1, 2, 3 y 4 con las probabilidades correspondientes de 0.15, 0.35, 0.35 y 0.15, respectivamente.

a. Calcule $E(X)$ y después $E(5 - X)$.

b. ¿Sería más redituable para el taller de reparación cobrar una cuota fija de \$75 o la cantidad de \$[150/(5 − X)]? [*Nota:* No siempre $E(c/Y) = c/E(Y)$.]

39. Una compañía de productos químicos tiene en existencia 100 lb de un producto químico, el cual se vende a sus clientes en lotes de 5 lb. Sea X = el número de lotes solicitados por un cliente seleccionado al azar y suponga que X tiene la función de masa de probabilidad

x	1	2	3	4
$p(x)$	0.2	0.4	0.3	0.1

Calcule $E(X)$ y $V(X)$. Luego calcule el número esperado de libras que quedan una vez que se envía el pedido del siguiente cliente y la varianza del número de libras que quedan. [*Sugerencia*: La cantidad de libras que quedan es una función lineal de X.]

40. **a.** Trace una gráfica lineal de la función de masa de probabilidad de X en el ejercicio 35. Enseguida determine la función de masa de probabilidad de $-X$ y trace su gráfica lineal. Con base en estas dos figuras, ¿qué se puede decir sobre $V(X)$ y $V(-X)$?

b. Use la proposición que implica $V(aX + b)$ para establecer una relación general entre $V(X)$ y $V(-X)$.

41. Use la definición en la expresión (3.13) para comprobar que $V(aX + b) = a^2 \cdot \sigma_X^2$. [*Sugerencia*: Con $h(X) = aX + b$, $E[h(X)] = a\mu + b$, donde $\mu = E(X)$.]

42. Suponga $E(X) = 5$ y $E[X(X - 1)] = 27.5$.

a. ¿Cuál es $E(X^2)$? [*Sugerencia:* Primero verifique que $E[X(X - 1)] = E(X^2) - E(X)$.]

b. ¿Cuál es $V(X)$?

c. ¿Cuál es la relación general entre las cantidades $E(X)$, $E[X(X - 1)]$ y $V(X)$?

43. Escriba una regla general para $E(X - c)$, donde c es una constante. ¿Qué sucede cuando $c = \mu$, el valor esperado de X?

44. Un resultado llamado **desigualdad de Chebyshev** establece que para cualquier distribución de probabilidad de una variable aleatoria X y cualquier número k que al menos sea 1, $P(X - \mu| \geq k\sigma) \leq 1/k^2$. Es decir, la posibilidad de que el valor de X quede al menos a k desviaciones estándar de su media es cuando mucho $1/k^2$.

a. ¿Cuál es el valor del límite superior con $k = 2$? ¿$k = 3$? ¿$k = 4$? ¿$k = 5$? ¿$k = 10$?

b. Calcule μ y σ para la distribución del ejercicio 13. Evalúe enseguida $P(|X - \mu| \geq k\sigma)$ con los valores de k dados en el inciso a). ¿Qué sugiere esto sobre el límite superior respecto a la probabilidad correspondiente?

c. Sea que X tenga los valores posibles -1, 0 y 1, con las probabilidades $\frac{1}{18}$, $\frac{8}{9}$ y $\frac{1}{18}$, respectivamente. ¿Cuál es $P(|X - \mu| \geq 3\sigma)$ y cómo se compara con el límite correspondiente?

d. Dé una distribución para la cual $P(|X - \mu| \geq 5\sigma) = 0.04$.

45. Si $a \leq X \leq b$ demuestre que $a \leq E(X) \leq b$.

3.4 Distribución de probabilidad binomial

Existen muchos experimentos que se ajustan exacta o aproximadamente a la siguiente lista de requerimientos.

1. El experimento consta de una secuencia de n experimentos más pequeños llamados *ensayos*, donde n se fija antes del experimento.

2. Cada ensayo puede dar por resultado uno de los mismos dos resultados posibles (ensayos dicotómicos) los cuales se denotan como éxito (S) y como falla (F).

3. Los ensayos son independientes, de modo que el resultado en un ensayo particular no influye en el resultado de cualquier otro ensayo.

4. La probabilidad de éxito $P(S)$ es constante de un ensayo a otro; esta probabilidad se denota por p.

DEFINICIÓN

> Un experimento para el que se satisfacen las condiciones 1–4 (un número fijo de ensayos homogéneos, independientes y dicotómicos) se llama **experimento binomial**.

EJEMPLO 3.27 Considere que cada uno de los siguientes n vehículos está pasando por una prueba de emisiones, S denota un vehículo que pasa la prueba y F denota uno que no la pasa. Entonces este experimento satisface las condiciones 1–4. El lanzamiento de una tachuela n veces, con S = la punta hacia arriba y F = la punta hacia abajo, también es un experimento binomial, como lo sería el experimento en el que el género (S para femenino y F para masculino) se determina para cada uno de los siguiente n niños nacidos en un hospital en particular. ■

Muchos experimentos implican una secuencia de ensayos independientes para los cuales existen más de dos resultados posibles en cualquier ensayo. Entonces, un experimento binomial puede crearse dividiendo los posibles resultados en dos grupos.

EJEMPLO 3.28 El color de las semillas de chícharo lo determina un solo locus genético. Si los dos alelos en este locus genético son AA o Aa (el genotipo), entonces el chícharo será amarillo (el fenotipo) y si el alelo es aa, el chícharo será verde. Suponga que se aparean 20 semillas Aa y se cruzan las dos semillas en cada uno de los diez pares para obtener diez nuevos genotipos. Designe cada nuevo genotipo como éxito (S) si es aa y falla (F) si es lo contrario. Entonces con esta identificación de S y F, el experimento es binomial con $n = 10$ y $p = P$ (genotipo aa). Si es igualmente probable que cada miembro del par contribuya con a o A, entonces $p = P(a) \cdot P(a) = (0.5)(0.5) = 0.25$. ■

EJEMPLO 3.29 El conjunto de posibles jurados para un cierto caso consiste en 50 individuos de los cuales 35 son contratados. Suponga que 6 de estos individuos son seleccionados uno por uno al azar para sentarse en la tribuna del jurado durante el interrogatorio inicial por parte de los abogados de la defensa y de la fiscalía. Nombre a la i-ésima persona seleccionada (el i-ésimo intento) como un éxito si fue contratada y como un fracaso si no. Por tanto:

$$P(S \text{ en el primer ensayo}) = \frac{35}{50} = 0.70$$

y

$$\begin{aligned} P(S \text{ en el segundo ensayo}) &= P(SS) + P(FS) \\ &= P(\text{segundo } S \mid \text{primer } S)P(\text{primer } S) \\ &\quad + P(\text{segundo } S \mid \text{primer } F)P(\text{primer } F) \\ &= \frac{34}{49} \cdot \frac{35}{50} + \frac{35}{49} \cdot \frac{15}{50} = \frac{35}{50}\left(\frac{34}{49} + \frac{15}{49}\right) = \frac{35}{50} = 0.70 \end{aligned}$$

De manera similar se puede demostrar que $P(S \text{ en el ensayo } i\text{-ésimo}) = 0.70$ con $i = 3, 4, 5$. Sin embargo, si los primeros cinco individuos seleccionados son todos S, entonces sólo restan 30 Ss para la sexta selección. Por tanto

$$P(S \text{ sexto intento} \mid \text{SSSS}) = 30/45 = 0.667$$

mientras que

$$P(S \text{ en el sexto intento} \mid FFFF) = 35/45 = 0.778$$

El experimento no es binomial porque los ensayos no son independientes. En general, si se muestrea sin reemplazo, el experimento no producirá ensayos independientes.

Ahora considere que un estado tiene 500 000 conductores con licencia, de los cuales 400 000 están asegurados. Se selecciona una muestra de 10 conductores sin reemplazo. El ensayo i-ésimo se denota con S si el conductor i-ésimo seleccionado está asegurado. La diferencia importante es que el tamaño de la población muestreada es muy grande respecto al tamaño de la muestra. En este caso

$$P(S \text{ en } 2 \mid S \text{ en } 1) = \frac{399\,999}{499\,999} = 0.8000$$

y

$$P(S \text{ en } 10 \mid S \text{ en los primeros } 9) = \frac{399\,991}{499\,991} = 0.799996 \approx 0.8000$$

$$P(S \text{ en } 10 \mid F \text{ en los primeros } 9) = \frac{400\,000}{499\,991} = 0.800014 \approx 0.8000$$

Estos cálculos sugieren que aunque los ensayos no son exactamente independientes, las probabilidades condicionales difieren tan poco una de otra que, para propósitos prácticos, los ensayos se consideran independientes con la constante $P(S) = 0.8$. Por tanto, para una muy buena aproximación, el experimento es binomial con $n = 10$ y $p = 0.8$. ■

Se utilizará la siguiente regla empírica para decidir si un experimento "sin reemplazo" puede ser tratado como un experimento binomial.

REGLA

Considere el muestreo sin reemplazo de una población dicotómica de tamaño N. Si el tamaño de la muestra (número de ensayos) n es cuando mucho 5% del tamaño de la población, el experimento puede ser analizado como si fuera exactamente un experimento binomial.

Por "analizado" se entiende que las probabilidades basadas en suposiciones de experimento binomial serán bastante cercanas a las probabilidades reales "sin reemplazo", las cuales generalmente son más difíciles de calcular. En el ejemplo 3.29, $n/N = 6/50 = 0.12 > 0.05$, de modo que el experimento binomial no es una buena aproximación, pero en el segundo escenario, $n/N = 10/500\,000 \ll 0.05$.

Variable y distribución aleatoria binomial

En la mayoría de los experimentos binomiales lo que interesa es el número total de éxitos (S), en lugar de qué ensayos dieron los éxitos.

DEFINICIÓN

La **variable aleatoria binomial** X asociada con un experimento binomial que consiste en n ensayos se define como

$$X = \text{el número de } S \text{ entre los } n \text{ ensayos}$$

Suponga, por ejemplo, que $n = 3$. Entonces existen ocho posibles resultados para el experimento:

$$SSS \quad SSF \quad SFS \quad SFF \quad FSS \quad FSF \quad FFS \quad FFF$$

Por la definición de X, $X(SSF) = 2$, $X(SFF) = 1$ y así sucesivamente. Valores posibles de X en un experimento de n ensayos son $x = 0, 1, 2, \ldots, n$. A menudo se escribirá $X \sim \text{Bin}(n, p)$ para indicar que X es una variable aleatoria binomial basada en n ensayos con probabilidad de éxito p.

NOTACIÓN

Puesto que la función de masa de probabilidad de una variable aleatoria binomial X depende de los dos parámetros n y p, la función de masa de probabilidad se denota por $b(x; n, p)$.

Considere primero el caso $n = 4$ para el cual cada resultado, su probabilidad y su valor x correspondiente se dan en la tabla 3.1. Por ejemplo,

$$\begin{aligned} P(SSFS) &= P(S) \cdot P(S) \cdot P(F) \cdot P(S) \text{ (ensayos independientes)} \\ &= p \cdot p \cdot (1 - p) \cdot p \text{ (constante } P(S)) \\ &= p^3 \cdot (1 - p) \end{aligned}$$

Tabla 3.1 Resultados y probabilidades para un experimento binomial con cuatro intentos

Resultado	x	Probabilidad	Resultado	x	Probabilidad
SSSS	4	p^4	*FSSS*	3	$p^3(1-p)$
SSSF	3	$p^3(1-p)$	*FSSF*	2	$p^2(1-p)^2$
SSFS	3	$p^3(1-p)$	*FSFS*	2	$p^2(1-p)^2$
SSFF	2	$p^2(1-p)^2$	*FSFF*	1	$p(1-p)^3$
SFSS	3	$p^3(1-p)$	*FFSS*	2	$p^2(1-p)^2$
SFSF	2	$p^2(1-p)^2$	*FFSF*	1	$p(1-p)^3$
SFFS	2	$p^2(1-p)^2$	*FFFS*	1	$p(1-p)^3$
SFFF	1	$p(1-p)^3$	*FFFF*	0	$(1-p)^4$

En este caso especial, se desea $b(x; 4, p)$ con $x = 0, 1, 2, 3$ y 4. Para $b(3; 4, p)$, identifique cuáles de los 16 resultados dan un valor x de 3 y sume las probabilidades asociadas con cada resultado:

$$b(3; 4, p) = P(FSSS) + P(SFSS) + P(SSFS) + P(SSSF) = 4p^3(1 - p)$$

Existen cuatro resultados con $X = 3$ y la probabilidad de cada uno es $p^3(1 - p)$(el orden de los S y las F no es importante, sino sólo el número de S), por tanto

$$b(3; 4, p) = \left\{\begin{matrix}\text{número de resultados} \\ \text{con } X = 3\end{matrix}\right\} \cdot \left\{\begin{matrix}\text{probabilidad de cualquier} \\ \text{resultado con } X = 3\end{matrix}\right\}$$

Asimismo, $b(2; 4, p) = 6p^2(1 - p)^2$, la cual también es el producto del número de resultados con $X = 2$ y la probabilidad de cualquier resultado como ese.

En general,

$$b(x; n, p) = \left\{\begin{matrix}\text{número de secuencias de longitud} \\ n \text{ compuestas de } x \text{ éxitos}\end{matrix}\right\} \cdot \left\{\begin{matrix}\text{probabilidad de cualquier} \\ \text{secuencia como esa}\end{matrix}\right\}$$

Puesto que el orden de los S y las F no es importante, el segundo factor en la ecuación previa es $p^x(1 - p)^{n-x}$ (p. ej., los primeros x ensayos producen S y los últimos $n - x$ producen F). El primer factor es el número de formas de escoger x de los n ensayos para que sean los S, es decir, el número de combinaciones de tamaño x que pueden ser construidas con n objetos distintos (ensayos en este caso).

TEOREMA

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x(1 - p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{de lo contrario} \end{cases}$$

EJEMPLO 3.30 A cada uno de seis bebedores de refrescos de cola seleccionados al azar se le sirve un vaso de refresco de cola *S* y uno de refresco de cola *F*. Los vasos son idénticos en apariencia excepto por un código que se muestra en el fondo para identificar cada refresco. Suponga que en realidad no existe una tendencia entre los bebedores de refresco de cola de preferir un refresco de cola en lugar del otro. Entonces $p = P(\text{un individuo seleccionado prefiere } S) = 0.5$, así que con $X =$ el número entre los seis que prefieren *S*, $X \sim \text{Bin}(6, 0.5)$.

Por tanto,

$$P(X = 3) = b(3; 6, 0.5) = \binom{6}{3}(0.5)^3(0.5)^3 = 20(0.5)^6 = 0.313$$

La probabilidad de que al menos tres prefieran *S* es

$$P(3 \le X) = \sum_{x=3}^{6} b(x; 6, 0.5) = \sum_{x=3}^{6} \binom{6}{x}(0.5)^x(0.5)^{6-x} = 0.656$$

y la probabilidad de que cuando mucho uno prefiera *S* es

$$P(X \le 1) = \sum_{x=0}^{1} b(x; 6, 0.5) = 0.109$$ ■

Utilización de tablas binomiales

Incluso con un valor relativamente pequeño de *n* el cálculo de probabilidades binomiales es tedioso. La tabla A.1 del apéndice tabula la función de distribución acumulada $F(x) = P(X \le x)$ con $n = 5, 10, 15, 20, 25$ en combinación con valores seleccionados de *p*. Varias otras probabilidades pueden entonces ser calculadas mediante la proposición sobre funciones de distribución acumulada de la sección 3.2. Una anotación de 0 en la tabla significa únicamente que la probabilidad es 0 a tres dígitos significativos puesto que todos los valores ingresados en la tabla en realidad son positivos.

NOTACIÓN

Para $X \sim \text{Bin}(n, p)$, la función de distribución acumulada será denotada por

$$B(x; n, p) = P(X \le x) = \sum_{y=0}^{x} b(y; n, p) \quad x = 0, 1, \ldots, n$$

EJEMPLO 3.31 Suponga que 20% de todos los ejemplares de un libro de texto particular no pasan una prueba de resistencia de encuadernación. Sea *X* el número entre 15 ejemplares seleccionados al azar que no pasan la prueba. Entonces *X* tiene una distribución binomial con $n = 15$ y $p = 0.2$.

1. La probabilidad de que cuando mucho 8 no pasen la prueba es

$$P(X \le 8) = \sum_{y=0}^{8} b(y; 15, 0.2) = B(8; 15, 0.2)$$

la cual es el dato en el renglón $x = 8$ y la columna $p = 0.2$ de la tabla binomial $n = 15$. Según la tabla A.1 del apéndice, la probabilidad es $B(8; 15, 0.2) = 0.999$.

2. La probabilidad de que exactamente 8 fallen es

$$P(X = 8) = P(X \le 8) - P(X \le 7) = B(8; 15, 0.2) - B(7; 15, 0.2)$$

que es la diferencia entre dos datos consecutivos en la columna $p = 0.2$. El resultado es $0.999 - 0.996 = 0.003$.

* Los paquetes de programas estadísticos tales como Minitab y R proporcionan la función de masa de probabilidad o la función de distribución acumulada en forma casi instantánea al solicitarla para cualquier valor de *p* y *n* desde 2 hasta millones. También existe un comando R para calcular la probabilidad de que *X* quede en algún intervalo.

3. La probabilidad de que al menos 8 fallen es

$$\begin{aligned} P(X \geq 8) &= 1 - P(X \leq 7) = 1 - B(7; 15, 0.2) \\ &= 1 - \begin{pmatrix} \text{dato en } x = 7 \\ \text{renglón de la columna } p = 0.2 \end{pmatrix} \\ &= 1 - 0.996 = 0.004 \end{aligned}$$

4. Finalmente, la probabilidad de que entre 4 y 7, inclusive, fallen es

$$\begin{aligned} P(4 \leq X \leq 7) &= P(X = 4, 5, 6 \text{ o } 7) = P(X \leq 7) - P(X \leq 3) \\ &= B(7; 15, 0.2) - B(3; 15, 0.2) = 0.996 - 0.648 = 0.348 \end{aligned}$$

Observe que esta última probabilidad es la diferencia entre los datos en los renglones $x = 7$ y $x = 3$, *no* en los renglones $x = 7$ y $x = 4$. ■

EJEMPLO 3.32 Un fabricante de aparatos electrónicos afirma que cuando mucho 10% de sus unidades de suministro de potencia necesitan servicio durante el periodo de garantía. Para investigar esta afirmación, los técnicos de un laboratorio de prueba adquieren 20 unidades y someten cada una a una prueba acelerada para simular el uso durante el periodo de garantía. Sea p la probabilidad de que una unidad de suministro de potencia necesite reparación durante dicho periodo (proporción de unidades que requieren reparación). Los técnicos de laboratorio deben decidir si los datos obtenidos con el experimento respaldan la afirmación de que $p \leq 0.10$. Sea X el número entre las 20 muestreadas que necesitan reparación, por lo que $X \sim \text{Bin}(20, p)$. Considere la regla de decisión:

> Rechazar la afirmación de que $p \leq 0.10$ en favor de la conclusión de que $p > 0.10$ si $x \geq 5$ (donde x es el valor observado de X) y considerar recomendable la afirmación si $x \leq 4$.

La probabilidad de que la afirmación sea rechazada cuando $p = 0.10$ (una conclusión incorrecta) es

$$P(X \geq 5 \text{ donde } p = 0.10) = 1 - B(4; 20, 0.1) = 1 - 0.957 = 0.043$$

La probabilidad de que la afirmación no sea rechazada cuando $p = 0.20$ (un tipo diferente de conclusión incorrecta) es

$$P(X \leq 4 \text{ cuando } p = 0.2) = B(4; 20, 0.2) = 0.630$$

La primera probabilidad es algo pequeña, pero la segunda es intolerablemente grande. Cuando $p = 0.20$, significa que el fabricante subestimó de manera excesiva el porcentaje de unidades que necesitan servicio, y si se utiliza la regla de decisión establecida, ¡63% de todas las muestras resultaron recomendables!

Se podría pensar que la probabilidad de este segundo tipo de conclusión errónea podría hacerse más pequeña al cambiar el valor de corte 5 en la regla de decisión a algo más. Sin embargo, aunque el reemplazo de 5 por un número más pequeño daría una probabilidad más pequeña que 0.630, la otra probabilidad se incrementaría. La única forma de hacer pequeñas ambas "probabilidades de error" es basar la regla de decisión en un experimento que implique muchas más unidades. ■

La media y la varianza de X

Con $n = 1$, la distribución binomial llega a ser la distribución de Bernoulli. De acuerdo con el ejemplo 3.18 la media de una variable de Bernoulli es $\mu = p$, así que el número esperado de las S en cualquier ensayo único es p. Puesto que un experimento binomial se compone de n ensayos, la intuición sugiere que para $X \sim \text{Bin}(n, p)$, $E(X) = np$, el producto del número de ensayos y la probabilidad de éxito en un solo ensayo. La expresión para $V(X)$ no es tan intuitiva.

PROPOSICIÓN

Si $X \sim \text{Bin}(n, p)$, entonces $E(X) = np$, $V(X) = np(1 - p) = npq$, y $\sigma_X = \sqrt{npq}$ (donde $q = 1 - p$).

Por tanto, para calcular la media y la varianza de una variable aleatoria binomial no se requiere evaluar las sumas. La comprobación del resultado para $E(X)$ se ilustra en el ejercicio 64.

EJEMPLO 3.33 Si 75% de todas las compras en una tienda se hacen con tarjeta de crédito y X es el número entre diez compras seleccionadas al azar realizadas con tarjeta de crédito, entonces $X \sim \text{Bin}(10, 0.75)$. Por tanto, $E(X) = np = (10)(0.75) = 7.5$, $V(X) = npq = 10(0.75)(0.25) = 1.875$, y $\sigma = \sqrt{1.875} = 1.37$. Otra vez, aun cuando X puede tomar sólo valores enteros, $E(X)$ no tiene que ser un entero. Si se realiza un gran número de experimentos binomiales independientes, cada uno con $n = 10$ ensayos y $p = 0.75$, entonces el número promedio de las S por experimento se acercará a 7.5.

La probabilidad de que X se encuentre dentro de una desviación estándar de su valor medio es $P(7.5 - 1.37 \le X \le 7.5 + 1.37) = P(6.13 \le X \le 8.87) = P(X = 7 \text{ u } 8) = 0.532$. ■

EJERCICIOS Sección 3.4 (46-67)

46. Calcule las siguientes probabilidades binomiales directamente con la fórmula para $b(x; n, p)$:
- **a.** $b(3; 8, 0.35)$.
- **b.** $b(5; 8, 0.6)$.
- **c.** $P(3 \le X \le 5)$ cuando $n = 7$ y $p = 0.6$.
- **d.** $P(1 \le X)$ cuando $n = 9$ y $p = 0.1$.

47. El artículo **"Should You Report That Fender-Bender?" (*Consumer Reports*, Sept. 2013: 15)** reportó que 7 de cada 10 accidentes de automóvil involucran un solo vehículo (el artículo recomienda siempre reportar a la compañía de seguros cualquier accidente que involucre varios vehículos). Suponga que se seleccionan 15 accidentes al azar. Utilice la tabla A.1 del apéndice para contestar las siguientes preguntas:
- **a.** ¿Cuál es la probabilidad de que máximo en 4 accidentes esté involucrado un solo vehículo?
- **b.** ¿Cuál es la probabilidad de que exactamente en 4 accidentes esté involucrado un solo vehículo?
- **c.** ¿Cuál es la probabilidad de que exactamente en 6 accidentes estén involucrados varios vehículos?
- **d.** ¿Cuál es la probabilidad de que en 2 a 4 accidentes inclusive esté involucrado un solo vehículo?
- **e.** ¿Cuál es la probabilidad de que al menos en 2 accidentes esté involucrado un solo vehículo?
- **f.** ¿Cuál es la probabilidad de que exactamente en 4 accidentes esté involucrado un solo vehículo y que en los otros 11 estén involucrados varios vehículos?

48. Noticias NBC reportó el 2 de mayo de 2013 que 1 de cada 20 niños en los Estados Unidos presenta una alergia alimentaria de algún tipo. Considere que se selecciona una muestra aleatoria de 25 niños y sea X la cantidad de niños en la muestra que tienen una alergia alimentaria. Así, $X \sim \text{Bin}(25, 0.05)$.
- **a.** Determine $P(X \le 3)$ y $P(X < 3)$.
- **b.** Determine $P(X \ge 4)$.
- **c.** Determine $P(1 \le X \le 3)$.
- **d.** ¿Cuáles son los valores de $E(X)$ y σ_X?
- **e.** En una muestra de 50 niños, ¿cuál es la probabilidad de que ninguno tenga una alergia alimentaria?

49. Una compañía que produce cristal fino sabe por experiencia que 10% de sus copas de mesa tienen imperfecciones cosméticas y deben ser clasificadas como "de segunda".
- **a.** Entre seis copas seleccionadas al azar, ¿qué tan probable es que sólo una sea de segunda?
- **b.** Entre seis copas seleccionadas al azar, ¿qué tan probable es que al menos dos sean de segunda?
- **c.** Si las copas se examinan una por una, ¿cuál es la probabilidad de que cuando mucho cinco deban ser seleccionadas para encontrar cuatro que no sean de segunda?

50. Se utiliza un número telefónico particular para recibir tanto llamadas de voz como faxes. Suponga que 25% de las llamadas entrantes son faxes y considere una muestra de 25 llamadas entrantes. Cuál es la probabilidad de que:

a. ¿Cuando mucho 6 de las llamadas sean faxes?
b. ¿Exactamente 6 de las llamadas sean faxes?
c. ¿Al menos 6 de las llamadas sean faxes?
d. ¿Más de 6 de las llamadas sean faxes?

51. Remítase al ejercicio previo.
a. ¿Cuál es el número de llamadas entre las 25 que se espera que sean faxes?
b. ¿Cuál es la desviación estándar del número entre las 25 llamadas que se espera que sean faxes?
c. ¿Cuál es la probabilidad de que el número de llamadas entre las 25, que se espera que sean una transmisión de fax sobrepase el número esperado por más de 2 desviaciones estándar?

52. Suponga que 30% de todos los estudiantes que deben comprar un texto para un curso particular desean un ejemplar nuevo (¡los exitosos!), mientras que el restante 70% desea comprar un ejemplar usado. Considere seleccionar 25 compradores al azar.
a. ¿Cuáles son la media y la desviación estándar del número de estudiantes que desean un ejemplar nuevo?
b. ¿Cuál es la probabilidad de que el número de estudiantes que desea ejemplares nuevos esté a más de dos desviaciones estándar del valor medio?
c. La librería tiene 15 ejemplares nuevos y 15 usados en existencia. Si 25 personas llegan una por una a comprar el texto, ¿cuál es la probabilidad de que las 25 obtengan el tipo de libro que desean entre las existencias actuales? [*Sugerencia*: Sea X = el número de estudiantes que desean un ejemplar nuevo. ¿Con qué valores de X obtendrán las 25 personas el libro que desean?]
d. Suponga que los ejemplares nuevos cuestan \$100 y los usados, \$70. Suponga en la actualidad que la librería tiene 50 ejemplares nuevos y 50 usados. ¿Cuál es el valor esperado del ingreso total por la venta de los siguientes 25 ejemplares comprados? Asegúrese de indicar qué regla de valor esperado está utilizando. [*Sugerencia*: Sea $h(X)$ = el ingreso cuando X de los 25 compradores desean ejemplares nuevos. Exprese esto como una función lineal.]

53. El ejercicio 30 (sección 3.3) dio la función de masa de probabilidad de Y, el número de infracciones de tránsito de un individuo asegurado por una compañía particular seleccionado al azar. Cuál es la probabilidad de que entre 15 individuos seleccionados al azar:
a. ¿Al menos 10 no tengan infracciones?
b. ¿Menos de la mitad tengan mínimo una infracción?
c. ¿Entre 5 y 10 inclusive tengan al menos una infracción?*

54. Un tipo particular de raqueta de tenis viene en tamaño mediano y en tamaño extra grande. Sesenta por ciento de todos los clientes en una tienda desea la versión extra grande.
a. Entre diez clientes seleccionados al azar que desean este tipo de raqueta, ¿cuál es la probabilidad de que al menos seis deseen la versión extra grande?
b. Entre diez clientes seleccionados al azar, ¿cuál es la probabilidad de que el número que desea la versión extra grande esté dentro de 1 desviación estándar de la media?
c. La tienda dispone actualmente de siete raquetas de cada versión. ¿Cuál es la probabilidad de que los siguientes diez clientes que desean esta raqueta puedan obtener la versión que desean de las existencias actuales?

55. Veinte por ciento de todos los teléfonos de cierto tipo son llevados a servicio mientras se encuentran dentro del periodo de garantía. De estos, 60% puede ser reparado, mientras el restante 40% debe ser reemplazado con unidades nuevas. Si una compañía adquiere diez de estos teléfonos, ¿cuál es la probabilidad de que exactamente dos sean reemplazados dentro de la vigencia de su garantía?

56. La Junta de Educación reporta que 2% de los dos millones de estudiantes de preparatoria que presentan el examen de aptitud escolar cada año reciben un trato especial debido a discapacidades documentadas (*Los Ángeles Times*, 16 de julio de 2002). Considere una muestra aleatoria de 25 estudiantes que recientemente presentaron el examen.
a. ¿Cuál es la probabilidad de que exactamente 1 reciba un trato especial?
b. ¿Cuál es la probabilidad de que al menos 1 reciba un trato especial?
c. ¿Cuál es la probabilidad de que al menos 2 reciban un trato especial?
d. ¿Cuál es la probabilidad de que el número entre los 25 que recibieron un trato especial esté dentro de 2 desviaciones estándar del número que esperaría reciba un trato especial?
e. Suponga que a un estudiante que no recibe un trato especial se le permiten 3 horas para el examen, mientras que a un estudiante que recibió un trato especial se le permiten 4.5 horas. ¿Qué tiempo promedio piensa que le sería permitido a los 25 estudiantes seleccionados?

57. Un tipo de linterna requiere que sus dos baterías sean tipo D y funcionará sólo si ambas baterías tienen voltajes aceptables. Suponga que 90% de todas las baterías de cierto proveedor tiene voltaje aceptable. Entre diez linternas seleccionadas al azar, ¿cuál es la probabilidad de que al menos nueve funcionen? ¿Qué suposiciones hizo para responder la pregunta planteada?

58. Un distribuidor recibe un lote muy grande de componentes. El lote sólo puede ser caracterizado como aceptable si la proporción de componentes defectuosos es cuando mucho de 0.10. El distribuidor decide seleccionar 10 componentes al azar y aceptar el lote sólo si el número de componentes defectuosos presentes en la muestra es cuando mucho de 2.
a. ¿Cuál es la probabilidad de que el lote sea aceptado cuando la proporción real de componentes defectuosos es de 0.01, 0.05, 0.10, 0.20, 0.25?
b. Sea p la proporción real de componentes defectuosos presentes en el lote. Una gráfica de P(se acepta el lote) en función de p, con p sobre el eje horizontal y P(se acepta

* "Entre a y b, inclusive" equivale a ($a \le X \le b$).

el lote) sobre el eje vertical, se llama *curva característica de operación* del plan de muestreo de aceptación. Use los resultados del inciso a) para trazar esta curva con $0 \leq p \leq 1$.

c. Repita los incisos a) y b) con "1" reemplazando a "2" en el plan de muestreo de aceptación.

d. Repita los incisos a) y b) con "15" reemplazando a "10" en el plan de muestreo de aceptación.

e. ¿Cuál de los planes de muestreo, los de los incisos a), c) o d), parece más satisfactorio y por qué?

59. Un reglamento que requiere que se instale un detector de humo en todas las casas ya construidas ha estado en vigor en una ciudad particular durante 1 año. Al departamento de bomberos le preocupa que muchas casas permanezcan sin detectores. Sea p = la proporción verdadera de las casas que tienen detectores y suponga que se inspecciona una muestra aleatoria de 25 casas. Si esta indica marcadamente que menos de 80% de todas las casas tiene un detector, el departamento de bomberos lanzará una campaña para poner en ejecución un programa de inspección obligatorio. Debido a lo caro del programa, el departamento prefiere no requerir tales inspecciones a menos que una evidencia muestral indique que sí se requieren. Sea X el número de casas con detectores entre las 25 muestreadas. Considere rechazar el requerimiento de que $p \geq 0.8$ si $x \leq 15$.

a. ¿Cuál es la probabilidad de que el requerimiento sea rechazado cuando el valor real de p es 0.8?

b. ¿Cuál es la probabilidad de no rechazar el requerimiento cuando $p = 0.7$? ¿Y cuando $p = 0.6$?

c. ¿Cómo cambian las "probabilidades de error" de los incisos a) y b) si el valor 15 en la regla de decisión es reemplazado por 14?

60. Un puente de peaje cobra \$1.00 para los vehículos de pasajeros y \$2.50 para los demás tipos de vehículos. Supongamos que durante las horas del día, 60% de todos los vehículos son de pasajeros. Si 25 vehículos cruzan el puente durante un periodo determinado durante el día, ¿cuál es el resultado de los ingresos por peaje previstos? [*Sugerencia*: Sea X = el número de vehículos de pasajeros, entonces, los ingresos por peaje $h(X)$ son una función lineal de X.]

61. Un estudiante que está tratando de escribir un ensayo para un curso debe optar entre dos temas, A y B. Si selecciona el tema A, el estudiante pedirá dos libros mediante préstamo interbibliotecas, mientras que si selecciona el tema B, pedirá cuatro libros. Él cree que para escribir un buen ensayo sobre cualquiera de los temas necesita recibir y utilizar al menos la mitad de los libros solicitados. Si la probabilidad de que un libro pedido mediante préstamo interbibliotecas llegue a tiempo es de 0.9 y los libros llegan independientemente uno de otro, ¿qué tema deberá seleccionar el estudiante para incrementar al máximo la probabilidad de escribir un buen ensayo? ¿Qué pasa si la probabilidad de que lleguen los libros es de sólo 0.5 en lugar de 0.9?

62. **a.** Con n fijo, ¿hay valores de p ($0 \leq p \leq 1$) para los cuales $V(X) = 0$? Explique por qué esto es así.

b. ¿Con qué valor de p se incrementa al máximo $V(X)$? [*Sugerencia*: Grafique $V(X)$ en función de p o bien saque una derivada.]

63. **a.** Demuestre que $b(x; n, 1 - p) = b(n - x; n, p)$.

b. Demuestre que $B(x; n, 1 - p) = 1 - B(n - x - 1; n, p)$. [*Sugerencia*: Cuando mucho el número x de los S equivale al menos a $(n - x)$ de las F.]

c. ¿Qué implican los incisos a) y b) sobre la necesidad de incluir valores de p más grandes que 0.5 en la tabla A.1 del apéndice?

64. Demuestre que $E(X) = np$ cuando X es una variable aleatoria binomial. [*Sugerencia*: Primero exprese $E(X)$ como una suma con límite inferior $x = 1$. Luego saque a np como factor, sea $y = x - 1$ de modo que la suma sea de $y = 0$ a $y = n - 1$ y demuestre que la suma es igual a 1.]

65. Los clientes en una gasolinera pagan con tarjeta de crédito (A), con tarjeta de débito (B) o en efectivo (C). Suponga que clientes sucesivos toman decisiones independientes con $P(A) = 0.5$, $P(B) = 0.2$ y $P(C) = 0.3$.

a. Entre los siguientes 100 clientes, ¿cuáles son la media y la varianza del número de clientes que pagan con tarjeta de débito? Explique su razonamiento.

b. Conteste el inciso a) para el número entre los 100 que no pagan en efectivo.

66. Una limusina del aeropuerto puede transportar hasta cuatro pasajeros en cualquier viaje. La compañía aceptará un máximo de seis reservaciones por viaje y cada pasajero debe tener reservación. Según registros previos, 20% de quienes reservan no se presentan para el viaje. Responda las siguientes preguntas, suponiendo independencia en los casos en que sea apropiado.

a. Si se hacen seis reservaciones, ¿cuál es la probabilidad de que al menos un individuo con reservación no pueda ser acomodado en el viaje?

b. Si se hacen seis reservaciones, ¿cuál es el número esperado de lugares sin ocupar cuando la limusina parte?

c. Suponga que en la siguiente tabla se da la distribución de probabilidad del número de reservaciones.

Número de reservaciones	3	4	5	6
Probabilidad	0.1	0.2	0.3	0.4

Sea X el número de pasajeros en un viaje seleccionado al azar. Obtenga la función de masa de probabilidad de X.

67. Remítase a la desigualdad de Chebyshev dada en el ejercicio 44. Calcule $P(|X - \mu| \geq k\sigma)$ para $k = 2$ y $k = 3$ cuando $X \sim$ Bin(20, 0.5), y compare con el límite superior correspondiente. Repita para $X \sim$ Bin(20, 0.75).

3.5 Distribuciones hipergeométrica y binomial negativa

Las distribuciones hipergeométrica y binomial negativa están relacionadas con la distribución binomial. La distribución binomial es el modelo de probabilidad aproximada de muestreo sin reemplazo de una población dicotómica finita ($S-F$). Si el tamaño n de la muestra es pequeño respecto al tamaño N de la población, la distribución hipergeométrica es el modelo de probabilidad exacta del número de éxitos (S) en la muestra. La variable aleatoria binomial X es el número de S cuando el número n de ensayos es fijo, mientras que la distribución binomial surge de fijar el número deseado de éxitos y de permitir que el número de ensayos sea aleatorio.

Distribución hipergeométrica

Las suposiciones que conducen a la distribución hipergeométrica son las siguientes:

1. La población o el conjunto que se va a muestrear se compone de N individuos, objetos o elementos (una población *finita*).
2. Cada individuo puede ser caracterizado como éxito (S) o falla (F) y hay M éxitos en la población.
3. Se selecciona una muestra de n individuos sin reemplazo, de tal modo que cada subconjunto de tamaño n tenga la misma probabilidad de ser seleccionado.

La variable aleatoria de interés es $X =$ el número de S en la muestra. La distribución de probabilidad de X depende de los parámetros n, M y N, así que se desea obtener $P(X = x) = h(x; n, M, N)$.

EJEMPLO 3.34 Durante un periodo particular una oficina de tecnología de la información de una universidad recibió 20 solicitudes de servicio por problemas con las impresoras, de las cuales 8 eran impresoras láser y 12 eran modelos de inyección de tinta. Se tiene que seleccionar una muestra de 5 de estas solicitudes de servicio para incluirla en una encuesta sobre satisfacción del cliente. Suponga que las 5 son seleccionadas completamente al azar, de modo que cualquier subconjunto de tamaño 5 tenga la misma probabilidad de ser seleccionado como cualquier otro subconjunto. ¿Cuál es entonces la probabilidad de que exactamente $x (x = 0, 1, 2, 3, 4 \text{ o } 5)$ de las solicitudes de servicio seleccionadas sean para impresoras de inyección de tinta?

En este caso el tamaño de la población es $N = 20$, el tamaño de la muestra es $n = 5$ y el número de éxitos (inyección de tinta $= S$) y las fallas (F) en la población son $M = 12$ y $N - M = 8$, respectivamente. Considere el valor $x = 2$. Ya que todos los resultados (cada uno de los cuales consta de 5 solicitudes particulares) son igualmente probables,

$$P(X = 2) = h(2; 5, 12, 20) = \frac{\text{número de resultados con } X = 2}{\text{número de resultados posibles}}$$

El número de resultados posibles en el experimento es la cantidad de formas de seleccionar 5 de entre los 20 objetos sin importar el orden, es decir, $\binom{20}{5}$. Para contar el número de resultados con $X = 2$, observe que existen $\binom{12}{2}$ formas de seleccionar 2 de las solicitudes para impresoras de inyección de tinta, y por cada forma existen $\binom{8}{3}$ formas de seleccionar las 3 solicitudes para impresoras láser a fin de completar la muestra. La regla de producto del capítulo 2 da entonces $\binom{12}{2}\binom{8}{3}$ como el número de resultados con $X = 2$, por tanto,

$$h(2; 5, 12, 20) = \frac{\binom{12}{2}\binom{8}{3}}{\binom{20}{5}} = \frac{77}{323} = 0.238$$

■

En general, si el tamaño de la muestra n es más pequeño que el número de éxitos en la población (M), entonces el valor de X más grande posible es n. Sin embargo, si $M < n$ (p. ej., un tamaño de muestra de 25 y sólo hay 15 éxitos en la población), entonces X puede ser, cuando mucho, M. Asimismo, siempre que el número de fallas en la población ($N - M$) sobrepase el tamaño de la muestra, el valor más pequeño posible de X es 0 (puesto que todos los individuos muestreados podrían entonces ser fallas). Sin embargo, si $N - M < n$, el valor más pequeño posible de X es $n - (N - M)$. Por tanto, los posibles valores de X satisfacen la restricción máx $(0, n - (N - M)) \le x \le$ mín (n, M). Un argumento paralelo al del ejemplo previo da la función de masa de probabilidad de X.

PROPOSICIÓN

Si X es el número de éxitos (S) en una muestra completamente aleatoria de tamaño n extraída de la población compuesta de M éxitos y $(N - M)$ fallas, entonces la distribución de probabilidad de X, llamada **distribución hipergeométrica**, está dada por

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (3.15)$$

con x, un entero, que satisface máx $(0, n - N + M) \le x \le$ mín (n, M).

En el ejemplo 3.34, $n = 5$, $M = 12$ y $N = 20$, por tanto, $h(x; 5, 12, 20)$ con $x = 0, 1, 2, 3, 4, 5$ se obtiene sustituyendo estos números en la ecuación (3.15).

EJEMPLO 3.35 Se capturaron, etiquetaron y liberaron cinco individuos de una población de animales que se piensa están al borde la extinción en cierta región para que se mezclen con la población. Luego de que han tenido la oportunidad de mezclarse se selecciona una muestra aleatoria de 10 de estos animales. Sea $X =$ el número de animales etiquetados en la segunda muestra. Suponga que hay 25 animales de este tipo en la región.

Los valores de los parámetros son $n = 10$, $M = 5$ (5 animales etiquetados en la población) y $N = 25$, por tanto la función de masa de probabilidad de X es

$$h(x; 10, 5, 25) = \frac{\binom{5}{x}\binom{20}{10-x}}{\binom{25}{10}} \quad x = 0, 1, 2, 3, 4, 5$$

La probabilidad de que exactamente dos de los animales en la segunda muestra estén etiquetados es

$$P(X = 2) = h(2; 10, 5, 25) = \frac{\binom{5}{2}\binom{20}{8}}{\binom{25}{10}} = 0.385$$

La probabilidad de que a lo más dos de los animales en la muestra de recaptura estén etiquetados es

$$P(X \le 2) = P(X = 0, 1, \text{o } 2) = \sum_{x=0}^{2} h(x; 10, 5, 25)$$
$$= 0.057 + 0.257 + 0.385 = 0.699$$

■

Varios paquetes de software estadístico generan fácilmente probabilidades hipergeométricas (tabular es enfadoso debido a los tres parámetros).

Como en el caso binomial, existen expresiones simples para $E(X)$ y $V(X)$ para variables aleatorias hipergeométricas.

PROPOSICIÓN

La media y la varianza de la variable aleatoria hipergeométrica X cuya función de masa de probabilidad es $h(x;n,M,N)$ son

$$E(X) = n \cdot \frac{M}{N} \qquad V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

El cociente M/N es la proporción de éxitos en la población. Si se reemplaza M/N por p en $E(X)$ y $V(X)$, se obtiene

$$E(X) = np$$

$$V(X) = \left(\frac{N-n}{N-1}\right) \cdot np(1-p) \tag{3.16}$$

La expresión (3.16) muestra que las medias de las variables aleatorias binomiales e hipergeométricas son iguales, en tanto que las varianzas de las dos variables aleatorias difieren por el factor $(N - n)/(N - 1)$, a menudo llamado **factor de corrección de población finita**. Este factor es menor que 1, así que la variable hipergeométrica tiene una varianza más pequeña que la variable aleatoria binomial. El factor de corrección puede escribirse como $(1 - n/N)/(1 - 1/N)$, el cual es aproximadamente 1 cuando n es pequeño respecto a N.

EJEMPLO 3.36 (Continuación del ejemplo 3.35)

En el ejemplo de etiquetado de animales, $n = 10$, $M = 5$ y $N = 25$, por tanto, $p = 5/25 = 0.2$ y

$$E(X) = 10(0.2) = 2$$

$$V(X) = \frac{15}{24}(10)(0.2)(0.8) = (0.625)(1.6) = 1$$

Si el muestreo se realizó con reemplazo, $V(X) = 1.6$.

Suponga que en realidad no se conoce el tamaño de la población N, así que se observa el valor x y se desea estimar N. Es razonable igualar la proporción muestral observada de éxitos x/n, con la proporción de la población, M/N, que da la estimación

$$\hat{N} = \frac{M \cdot n}{x}$$

Si $M = 100$, $n = 40$ y $x = 16$, entonces $\hat{N} = 250$.

La regla general empírica dada en la sección 3.4 plantea que si el muestreo se realizó sin reemplazo pero n/N era cuando mucho de 0.05, entonces la distribución binomial podría ser utilizada para calcular probabilidades aproximadas que implican el número de éxitos en la muestra. Un enunciado más preciso es el siguiente: permita que el tamaño de la población N y el número de M éxitos presentes en la población se hagan más grandes a medida que la razón M/N tiende a p. Entonces $h(x; n, M, N)$ tiende a $b(x; n, p)$; por tanto, con n/N pequeña, las dos son aproximadamente iguales, siempre y cuando p no esté muy cerca de 0 o 1. Esta es la razón de ser de la regla empírica.

Distribución binomial negativa

La variable aleatoria binomial y la distribución binomial negativa se basan en un experimento que satisface las siguientes condiciones:

1. El experimento consiste en una secuencia de ensayos independientes.
2. Cada ensayo puede dar por resultado un éxito (S) o una falla (F).

3. La probabilidad de éxito es constante de un ensayo a otro, por tanto, $P(S \text{ en el ensayo } i) = p$ con $i = 1, 2, 3, \ldots$
4. El experimento continúa (se realizan ensayos) hasta que un total de éxitos r haya sido observado, donde r es un entero positivo especificado.

La variable aleatoria de interés es $X =$ el número de fallas que preceden al éxito r-ésimo; X se llama **variable aleatoria binomial negativa** porque, en contraste con la variable aleatoria binomial, el número de éxitos es fijo y el número de ensayos es aleatorio.

Posibles valores de X son 0, 1, 2,… Sea $nb(x; r, p)$ la función de masa de probabilidad de X. Considere $nb(7, 3, p) = P(X = 7)$, la probabilidad de que ocurran exactamente 7F antes de la 3ª S. Para que esto suceda el décimo ensayo debe ser S y debe haber exactamente 2 S entre los 9 primeros ensayos. Por tanto,

$$nb(7; 3, p) = \left\{\binom{9}{2} \cdot p^2(1-p)^7\right\} \cdot p = \binom{9}{2} \cdot p^3(1-p)^7$$

La generalización de esta línea de razonamiento da la siguiente fórmula para la función de masa de probabilidad binomial negativa.

PROPOSICIÓN

La función de masa de probabilidad de la variable aleatoria binomial negativa X con los parámetros $r =$ número de éxitos (S) y $p = P(S)$ es

$$nb(x; r, p) = \binom{x + r - 1}{r - 1} p^r(1-p)^x \quad x = 0, 1, 2, \ldots$$

EJEMPLO 3.37 Un pediatra desea reclutar 5 parejas, cada una de las cuales espera a su primer hijo, para participar en un nuevo programa de parto natural. Sea $p = P(\text{una pareja seleccionada al azar está de acuerdo en participar})$. Si $p = 0.2$, ¿cuál es la probabilidad de que 15 parejas deban ser entrevistadas antes de encontrar 5 que estén de acuerdo en participar? Es decir, con $S = \{\text{está de acuerdo en participar}\}$, ¿cuál es la probabilidad de que ocurran 10 fallas antes del quinto éxito? Al sustituir $r = 5$, $p = 0.2$ y $x = 10$ en $nb(x; r, p)$ se obtiene

$$nb(10; 5, 0.2) = \binom{14}{4}(0.2)^5(0.8)^{10} = 0.034$$

La probabilidad de que a lo más se observen 10 fallas (cuando mucho con 15 parejas entrevistadas) es

$$P(X \le 10) = \sum_{x=0}^{10} nb(x; 5, 0.2) = (0.2)^5 \sum_{x=0}^{10} \binom{x + 4}{4}(0.8)^x = 0.164$$ ■

En algunas fuentes la variable aleatoria binomial negativa es el número de ensayos $X + r$ en lugar del número de fallas.

En el caso especial $r = 1$, la función de masa de probabilidad es

$$nb(x; 1, p) = (1-p)^x p \quad x = 0, 1, 2, \ldots \qquad \textbf{(3.17)}$$

En el ejemplo 3.12 se dedujo la función de masa de probabilidad para el número de ensayos necesarios para obtener el primer *éxito* (S), y ahí la función de masa de probabilidad es similar a la expresión (3.17). En la literatura se hace referencia tanto a $X =$ número de *fallas* (F) como a $Y =$ número de ensayos $(= 1 + X)$ como **variables aleatorias geométricas**, y la función de masa de probabilidad en la expresión (3.17) se llama **distribución geométrica**.

En el ejemplo 3.19 se demostró que el número esperado de ensayos hasta que aparece el primer éxito es $1/p$, así que el número esperado de fallas hasta que aparece el primer éxito es $(1/p) - 1 = (1 - p)/p$. Intuitivamente, se esperaría ver $r \cdot (1 - p)/pF$ antes del éxito r-ésimo y este en realidad es $E(X)$. También existe una fórmula simple para $V(X)$.

PROPOSICIÓN

Si X es una variable aleatoria binomial negativa con función de masa de probabilidad $nb(x; r, p)$, entonces

$$E(X) = \frac{r(1-p)}{p} \qquad V(X) = \frac{r(1-p)}{p^2}$$

Por último, al expandir el coeficiente binomial enfrente de $p^r(1 - p)^x$ y haciendo alguna cancelación se ve que $nb(x; r, p)$ está bien definido, incluso cuando r no es un entero. Se ha encontrado que la *distribución binomial negativa generalizada* para ajustar los datos observados se adapta verdaderamente bien en una amplia variedad de aplicaciones.

EJERCICIOS Sección 3.5 (68-78)

68. Dieciocho individuos tienen una cita para someterse a una prueba de manejo en una oficina de tránsito en particular un cierto día, ocho de los cuales realizarán la prueba por primera vez. Suponga que seis de estos individuos son asignados de manera aleatoria a un examinador en particular, y sea X el número de personas, entre estos seis, que están realizando la prueba por primera vez.

- **a.** Cuál es el tipo de distribución que tiene X (nombre y valores de todos los parámetros)?
- **b.** Calcule $P(X = 2)$, $P(X \leq 2)$, y $P(X \geq 2)$.
- **c.** Calcule la media y la desviación estándar para X.

69. Cada uno de 12 refrigeradores de un tipo ha sido devuelto a un distribuidor debido a que se escucha un sonido agudo cuando el refrigerador está funcionando. Suponga que 7 de estos refrigeradores tienen un compresor defectuoso y que los otros 5 tienen problemas menos serios. Si los refrigeradores se examinan en orden aleatorio, sea X el número entre los primeros 6 examinados que tienen un compresor defectuoso.

- **a.** Calcule $P(X = 4)$ y $P(X \leq 4)$
- **b.** Determine la probabilidad de que X exceda su valor medio por más de 1 desviación estándar.
- **c.** Considere un gran envío de 400 refrigeradores, 40 de los cuales tienen compresores defectuosos. Si X es el número de refrigeradores que tienen compresores defectuosos de entre 15 seleccionados al azar, describa una forma menos tediosa de calcular (al menos de forma aproximada) $P(X \leq 5)$ en lugar de utilizar la función de masa de probabilidad hipergeométrica.

70. Un instructor que impartió dos secciones de estadística para ingeniería el semestre pasado, la primera con 20 estudiantes y la segunda con 30, decidió asignar un proyecto semestral. Una vez que todos los proyectos le fueron entregados, el instructor los ordenó al azar antes de calificarlos. Considere los primeros 15 proyectos calificados.

- **a.** ¿Cuál es la probabilidad de que exactamente 10 de estos sean de la segunda sección?
- **b.** ¿Cuál es la probabilidad de que al menos 10 de estos sean de la segunda sección?
- **c.** ¿Cuál es la probabilidad de que al menos 10 de estos sean de la misma sección?
- **d.** ¿Cuáles son la media y la desviación estándar del número de proyectos entre estos 15 que son de la segunda sección?
- **e.** ¿Cuáles son la media y la desviación estándar del número de proyectos que no están entre estos primeros 15 que son de la segunda sección?

71. Un geólogo recolectó 10 especímenes de roca basáltica y 10 de granito. Le pide a su ayudante de laboratorio que seleccione al azar 15 de estos especímenes para analizarlos.

- **a.** ¿Cuál es la función de masa de probabilidad del número de especímenes de granito seleccionados para su análisis?
- **b.** ¿Cuál es la probabilidad de que todos los especímenes de uno de los dos tipos de roca sean seleccionados para su análisis?
- **c.** ¿Cuál es la probabilidad de que el número de especímenes de granito seleccionados para analizarlos esté dentro de 1 desviación estándar de su valor medio?

72. Un director de personal que va a entrevistar a 11 ingenieros para cuatro vacantes de trabajo ha programado seis entrevistas para el primer día y cinco para el segundo. Suponga que los candidatos son entrevistados en orden aleatorio.

a. ¿Cuál es la probabilidad de que x de los cuatro mejores candidatos sean entrevistados el primer día?

b. ¿Cuántos de los mejores cuatro candidatos se espera que puedan ser entrevistados el primer día?

73. Veinte parejas de individuos que participan en un torneo de bridge han sido sembrados del 1,…, 20. En esta primera parte del torneo, los 20 son divididos al azar en 10 parejas este-oeste y 10 parejas norte-sur.

a. ¿Cuál es la probabilidad de que x de las 10 mejores parejas terminen jugando este-oeste?

b. ¿Cuál es la probabilidad de que las cinco mejores parejas terminen jugando en la misma dirección?

c. Si existen $2n$ parejas, ¿cuál es la función de masa de probabilidad de X = el número entre las mejores n parejas que terminan jugando este-oeste? ¿Cuáles son $E(X)$ y $V(X)$?

74. Una alerta contra esmog de segunda etapa ha sido emitida en un área del condado de Los Ángeles en la cual hay 50 empresas industriales. Un inspector visitará 10 empresas seleccionadas al azar para verificar si no han violado los reglamentos.

a. Si 15 de las empresas están violando al menos un reglamento, ¿cuál es la función de masa de probabilidad del número de empresas visitadas por el inspector que violan al menos un reglamento?

b. Si existen 500 empresas en el área, 150 de las cuales violan algún reglamento, represente de forma aproximada la función de masa de probabilidad del inciso a) con una función de masa de probabilidad más simple.

c. Con X = el número de empresas que violan algún reglamento entre las 10 visitadas calcule $E(X)$ y $V(X)$, ambas para la función de masa de probabilidad exacta y la función de masa de probabilidad aproximada del inciso b).

75. La probabilidad de que una caja de un cierto tipo de cereal seleccionada al azar tenga un precio particular es de 0.2. Suponga que compra una caja tras otra hasta obtener dos con dicho precio.

a. ¿Cuál es la probabilidad de que compre x cajas que no tienen el precio deseado?

b. ¿Cuál es la probabilidad de que compre cuatro cajas?

c. ¿Cuál es la probabilidad de que compre cuando más cuatro cajas?

d. ¿Cuántas cajas sin el precio deseado esperaría comprar? ¿Cuántas cajas espera comprar?

76. Una familia decide tener hijos hasta que tengan tres niños del mismo sexo. Suponiendo que $P(B) = P(G) = 0.5$, ¿cuál es la función de masa de probabilidad de X = al número de niños en la familia?

77. Tres hermanos y sus esposas deciden tener hijos hasta que cada familia tenga dos niñas. ¿Cuál es la función de masa de probabilidad de X = el número total de varones procreados por las tres parejas? ¿Cuál es $E(X)$ y cómo se compara con el número esperado de varones procreados por cada pareja?

78. De acuerdo con el artículo **"Characterizing the Severity and Risk of Drought in the Poudre River, Colorado" (*J. of Water Res. Planning and Mgmnt*., 2005: 383–393)**, la longitud de la sequía Y es el número de intervalos de tiempo consecutivos en los que el suministro de agua se mantiene por debajo de un valor crítico y_0 (un déficit), precedido y seguido por periodos en los que el suministro supera este valor crítico (un excedente). El documento citado propone una distribución geométrica con $p = 0.409$ para esta variable aleatoria.

a. ¿Cuál es la probabilidad de que una sequía perdure exactamente 3 intervalos? ¿Y a lo más 3 intervalos?

b. ¿Cuál es la probabilidad de que la duración de una sequía exceda su valor medio por al menos una desviación estándar?

3.6 Distribución de probabilidad de Poisson

Las distribuciones binomial, hipergeométrica y binomial negativa se dedujeron partiendo de un experimento compuesto de ensayos o sorteos y aplicando las leyes de probabilidad a varios resultados del experimento. No existe un experimento simple en el cual esté basada la distribución de Poisson, no obstante, más adelante se describirá cómo puede ser obtenida mediante ciertas operaciones restrictivas.

DEFINICIÓN

Se dice que una variable aleatoria discreta X tiene una **distribución de Poisson** con parámetro μ ($\mu > 0$) si la función de masa de probabilidad de X es

$$p(x; \mu) = \frac{e^{-\mu} \cdot \mu^x}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

No es casualidad que se esté usando el símbolo μ para el parámetro de Poisson, más adelante se verá que μ es en realidad el valor esperado de X. La letra e en la función de masa de probabilidad representa la base del sistema de logaritmos naturales; su valor numérico es aproximadamente 2.71828. A diferencia de las distribuciones binomial e hipergeométrica, la distribución de probabilidad de Poisson se extiende a *todos* los números enteros no negativos, un número infinito de posibilidades.

No es evidente por inspección que $p(x; \mu)$ especifica una función de masa de probabilidad legítima, por no hablar de que esta distribución es útil. En primer lugar, $p(x; \mu) > 0$ para cada valor x posible, debido a la exigencia de que $\mu > 0$. El hecho de que $\Sigma p(x; \mu) = 1$ es una consecuencia de la expansión en series de Maclaurin de e^{μ} (consulte su libro de cálculo para este resultado):

$$e^{\mu} = 1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{\mu^x}{x!} \tag{3.18}$$

Si los dos términos extremos de la expresión (3.18) se multiplican por $e^{-\mu}$ y luego esta cantidad se coloca dentro de la suma en el lado derecho, el resultado es

$$1 = \sum_{x=0}^{\infty} \frac{e^{-\mu} \cdot \mu^x}{x!}$$

La tabla A.2 del Apéndice contiene la función de distribución acumulada de Poisson $F(x; \mu)$ para $\mu = 0.1, 0.2, \ldots, 1, 2, \ldots, 10, 15$ y 20. Por otra parte, si se les pide muchos paquetes de software proporcionarán $F(x; \mu)$ y $p(x; \mu)$.

EJEMPLO 3.38 Sea X el número de trampas (cierto tipo de defectos) en un tipo particular de transistor metal-óxido semiconductor, y suponga que tiene una distribución de Poisson con $\mu = 2$ (El modelo de Poisson se muestra en el artículo **"Analysis of Random Telegraph Noise in 45-nm CMOS Using On-Chip Characterization System "(*IEEE Trans. on Electron Devices*, 2013: 171621722)**; se cambió el valor del parámetro para facilitar el cálculo computacional.

La probabilidad de que haya exactamente tres trampas es

$$P(X = 3) = p(3; 2) = \frac{e^{-2}2^3}{3!} = 0.180,$$

Y la probabilidad de que haya cuando más tres trampas es

$$P(X \leq 3) = F(3; 2) = \sum_{x\,=\,0}^{3} \frac{e^{-2}2^x}{x!} = 0.135 + 0.271 + 0.271 + 0.180 = 0.857$$

Esta última probabilidad acumulada se encuentra en la intersección de la columna de $\mu = 2$ y en la fila de $x = 3$ de la tabla A.2. del apéndice, mientras que $p(3; 2) = F(3; 2) - F(2; 2) = 0.857 - 0.677 = 0.180$, la diferencia entre dos entradas consecutivas se encuentra en la columna $\mu = 2$ de la tabla acumulada de Poisson. ■

La distribución de Poisson como límite

La siguiente proposición revela la razón de ser del uso de la distribución de Poisson en muchas situaciones.

PROPOSICIÓN

> Suponga que en la función de masa de probabilidad binomial $b(x; n, p)$, $n \to \infty$ y $p \to 0$ de tal modo que np tienda a un valor $\mu > 0$. Entonces $b(x; n, p) \to p(x; \mu)$.

De acuerdo con esta proposición, *en cualquier experimento binomial en el cual n es grande y p pequeña*, $b(x; n, p) \approx p(x; \mu)$, donde $\mu = np$. Como regla empírica esta aproximación puede ser aplicada con seguridad si $n > 50$ y $np < 5$.

EJEMPLO 3.39 Si un editor de libros no técnicos hace todo lo posible porque sus libros estén libres de errores tipográficos, de modo que la probabilidad de que cualquier página dada contenga al menos un error de ese tipo es de 0.005 y los errores son independientes de una página a otra, ¿cuál es la probabilidad de que una de sus novelas de 400 páginas contenga exactamente una página con errores? ¿Y cuando mucho tres páginas con errores?

Con S denotando una página que contiene al menos un error y F una página libre de errores, el número X de páginas que contienen al menos un error es una variable aleatoria binomial con $n = 600$ y $p = 0.005$, así que $np = 3$. Se desea

$$P(X = 1) = b(1; 600, 0.005) \approx p(1; 3) = \frac{e^{-3}(3)^1}{1!} = 0.14936$$

El valor binomial es $b(1; 600, 0.005) = 0.14899$, así que la aproximación es muy buena.

Del mismo modo,

$$P(X \leq 3) \approx \sum_{x=0}^{3} p(x; 3) = F(3; 3) = 0.647$$

que con una precisión de tres decimales es idéntico a $B(3; 600, 0.005)$. ■

La tabla 3.2 muestra la distribución de Poisson con $\mu = 3$ junto con tres distribuciones binomiales con $np = 3$ y la figura 3.8 (generada por S-Plus) ilustra una gráfica de la distribución de Poisson junto con las dos primeras distribuciones binomiales. La aproximación es de uso limitado con $n = 30$, pero desde luego la precisión es mejor con $n = 100$ y mucho mejor con $n = 300$.

Tabla 3.2 Comparación de la distribución de Poisson y tres distribuciones binomiales

x	$n = 30, p = 0.1$	$n = 100, p = 0.3$	$n = 300, p = 0.1$	Poisson, $\mu = 3$
0	0.042391	0.047553	0.049041	0.049787
1	0.141304	0.147070	0.148609	0.149361
2	0.227656	0.225153	0.224414	0.224042
3	0.236088	0.227474	0.225170	0.224042
4	0.177066	0.170606	0.168877	0.168031
5	0.102305	0.101308	0.100985	0.100819
6	0.047363	0.049610	0.050153	0.050409
7	0.018043	0.020604	0.021277	0.021604
8	0.005764	0.007408	0.007871	0.008102
9	0.001565	0.002342	0.002580	0.002701
10	0.000365	0.000659	0.000758	0.000810

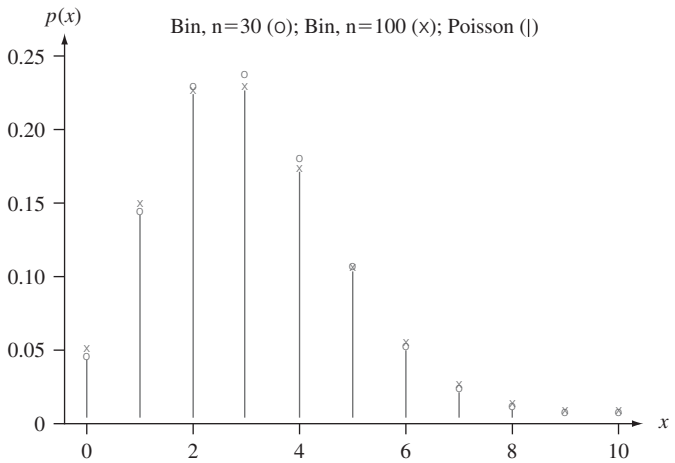


Figura 3.8 Comparación entre una distribución de Poisson y dos distribuciones binomiales

Media y varianza de X

Puesto que $b(x; n, p) \to p(x; \mu)$ a medida que $n \to \infty$, $p \to 0$, $np \to \mu$, la media y la varianza de una variable binomial deberán aproximarse a las de una variable de Poisson. Estos límites son $np \to \mu$ y $np(1 - p) \to \mu$.

PROPOSICIÓN

Si X tiene una distribución de Poisson con parámetro μ, entonces $E(X) = V(X) = \mu$.

Estos resultados también pueden obtenerse directamente de las definiciones de media y de varianza.

EJEMPLO 3.40 (Continuación del ejemplo 3.38)

Tanto el número esperado de criaturas atrapadas como la varianza de este son iguales a 2, y $\sigma_X = \sqrt{\mu} = \sqrt{2} = 1.414$. ■

Proceso de Poisson

Una aplicación muy importante de la distribución de Poisson surge en conexión con la ocurrencia de eventos de algún tipo en el transcurso del tiempo. Algunos eventos de interés podrían ser las visitas a un sitio web particular, asistencias a una cierta clase registradas mediante un contador, los mensajes de correo electrónico enviados a una dirección particular, los accidentes en una instalación industrial o las lluvias de rayos cósmicos observados por los astrónomos en un observatorio particular. Se hace la siguiente suposición sobre la forma en que los eventos de interés ocurren:

1. Existe un parámetro $\alpha > 0$ tal que durante cualquier intervalo de tiempo corto Δt, la probabilidad de que ocurra exactamente un evento es $\alpha \cdot \Delta t + o(\Delta t)$.
2. La probabilidad de que ocurra más de un evento durante Δt es $o(\Delta t)$ [la que junto con la suposición 1 implica que la probabilidad de ningún evento durante Δt es $[1 - \alpha \cdot \Delta t - o(\Delta t)]$.
3. El número de eventos ocurridos durante este intervalo de tiempo Δt es independiente del número ocurrido antes de dicho intervalo.

De manera informal la suposición 1 dice que durante un corto intervalo de tiempo la probabilidad de que ocurra un solo evento es aproximadamente proporcional a la duración del intervalo, donde α es la constante de proporcionalidad. Ahora sea $Pk(t)$ la probabilidad de que k eventos serán observados durante cualquier intervalo de tiempo particular de duración t.

PROPOSICIÓN

$P_k(t) = e^{-\alpha t} \cdot (\alpha t)^k / k!$, de modo que el número de eventos durante un intervalo de tiempo de duración t es una variable aleatoria de Poisson con parámetro $\mu = \alpha t$. El número esperado de eventos durante cualquier intervalo de tiempo es entonces αt, así que el número esperado durante un intervalo de tiempo unitario es α.

La ocurrencia de eventos en el transcurso del tiempo, como se describió, se llama *proceso de Poisson*; el parámetro α especifica *la rapidez* del proceso.

EJEMPLO 3.39

Suponga que llegan a un contador pulsaciones a un ritmo promedio de seis por minuto, por lo que $\alpha = 6$. Para determinar la probabilidad de que en un intervalo de 0.5 minutos se reciba al menos una pulsación, observe que el número de pulsaciones en ese intervalo tiene una distribución de Poisson con parámetro $\alpha t = 6(0.5) = 3$ (se utiliza 0.5 minutos porque

Una cantidad es $o(\Delta t)$ (léase "o pequeña de delta t") si, a medida que Δt tiende a 0, también lo hace $o(\Delta t)/\Delta t$. Es decir, $o(\Delta t)$ es incluso más insignificante (tiende a 0 más rápido) que Δt mismo. La cantidad $(\Delta t)^2$ tiene esta propiedad, pero no así $\text{sen}(\Delta t)$.

α está expresada como rapidez por minuto). Entonces con X = el número de pulsaciones recibidos en el intervalo de 30 segundos,

$$P(1 \le X) = 1 - P(X = 0) = 1 - \frac{e^{-3}(3)^0}{0!} = 0.950$$

■

En lugar de observar eventos en el transcurso del tiempo, considere observar eventos de algún tipo que ocurren en una región de dos o tres dimensiones. Por ejemplo, podría seleccionar un mapa de la región R de un bosque, ir a dicha región y contar el número de árboles. Cada árbol representaría un evento que ocurre en un punto particular del espacio. Conforme a suposiciones similares a 1−3, se puede demostrar que el número de eventos que ocurren en una región R tiene una distribución de Poisson con parámetro $\alpha \cdot a(R)$, donde $a(R)$ es el área de R. La cantidad α es el número esperado de eventos por unidad de área o volumen.

EJERCICIOS Sección 3.6 (79-93)

79. El artículo **"Expectation Analysis of the Probability of Failure for Water Supply Pipes" (*J. of Pipeline Systems Engr. and Practice*, mayo de 2012: 36-46)** propone utilizar la distribución de Poisson para modelar el número de fallas en las tuberías de varios tipos. Suponga que para una tubería de hierro forjado de una longitud particular, el número de fallas esperado es de 1 (muy cercano a uno de los casos considerados en el artículo). Entonces *X, el número de fallas, tiene una distribución de Poisson con* $\mu = 1$.

a. Obtenga $P(X \le 5)$ utilizando la tabla A.2. en el apéndice.

b. Determine $P(X = 2)$ primero a partir de la fórmula de la función de masa de probabilidad y luego a partir de la tabla A.2.

c. Determine $P(2 \le X \le 4)$.

d. ¿Cuál es la probabilidad de X que exceda su media por más de una desviación estándar?

80. Sea X el número de anomalías que ocurren en el material de una región particular de un disco de turbina de gas en los aviones. El artículo **"Methodology for Probabilistic Life Prediction of Multiple-Anomaly Materials" (*Amer. Inst. of Aeronautics and Astronautics J.*, 2006: 787-793)** propone una distribución de Poisson para X. Supongamos que $\mu = 4$.

a. Calcule $P(X \le 4)$ y $P(X < 4)$.

b. Calcule $P(4 \le X \le 8)$.

c. Calcule $P(8 \le X)$.

d. ¿Cuál es la probabilidad de que el número observado de anomalías sobrepase su media por no más de una desviación estándar?

81. Suponga que el número de conductores que viajan entre un origen y un destino particulares, durante un lapso de tiempo designado, tiene una distribución de Poisson con parámetro $\mu = 20$ (sugerido en el artículo **"Dynamic Ride Sharing: Theory and Practice", *J. of Transp. Engr.*, 1997:308-312).** Cuál es la probabilidad de que el número de conductores:

a. Sea cuando mucho de 10.

b. Sea de más de 20.

c. ¿Será de entre 10 y 20, inclusive? Será estrictamente de entre 10 y 20?

d. ¿Estará dentro de 2 desviaciones estándar de su media?

82. Considere escribir en un disco de computadora y luego enviarlo a través de un certificador que cuenta el número de impulsos magnéticos faltantes. Suponga que este número X tiene una distribución de Poisson con parámetro $\mu = 0.2$. (Sugerido en **"Average Sample Number for Semi-Curtailed Sampling Using the Poisson Distribution", *J. Quality Technology*, 1983: 126-129.)**

a. ¿Cuál es la probabilidad de que un disco tenga exactamente un impulso faltante?

b. ¿Cuál es la probabilidad de que un disco tenga al menos dos impulsos faltantes?

c. Si se seleccionan dos discos independientemente, ¿cuál es la probabilidad de que ninguno contenga un impulso faltante?

83. Un artículo en ***Los Angeles Times* (3 de diciembre de 1993)** reporta que 1 de cada 200 personas porta el gen defectuoso que provoca cáncer de colon hereditario. En una muestra de 1000 individuos, ¿cuál es la distribución aproximada del número que porta este gen? Use esta distribución para calcular la probabilidad aproximada de que:

a. Entre 5 y 8 (inclusive) porten el gen.

b. Al menos 8 porten el gen.

84. El **Centro para la Prevención y Control de Enfermedades** reportó en 2012 que 1 de cada 88 niños estadounidenses había sido diagnosticado con un trastorno del espectro autista (ASD).

a. Si se selecciona una muestra de 200 niños estadounidenses, ¿cuál es el valor esperado y la desviación estándar de la cantidad de niños que han sido diagnosticados con ASD?

b. En referencia a a) calcule la probabilidad aproximada de que al menos 2 niños en la muestra sean diagnosticados con ASD.

c. Si el tamaño de la muestra es de 352, ¿cuál es la probabilidad aproximada de que menos de 5 de los niños seleccionados hayan sido diagnosticados con ASD?

85. Suponga que una pequeña aeronave aterriza en un aeropuerto de acuerdo con un proceso de Poisson con razón $\alpha = 8$ por hora, de modo que el número de aterrizajes durante un lapso de tiempo de t horas es una variable aleatoria de Poisson con parámetro $\mu = 8t$.

a. ¿Cuál es la probabilidad de que exactamente 6 aeronaves pequeñas aterricen durante un intervalo de 1 hora? ¿Al menos 6? ¿Al menos 10?

b. ¿Cuáles son el valor esperado y la desviación estándar del número de aeronaves pequeñas que aterrizan durante un lapso de 90 minutos?

c. ¿Cuál es la probabilidad de que al menos 20 aeronaves pequeñas aterricen durante un lapso de 2.5 horas? ¿Y de que cuando mucho aterricen 10 durante este periodo?

86. En el agua de lastre que es descargada de un barco hay organismos una concentración de 10 organismos/m^3 de acuerdo con proceso de Poisson [el artículo **Counting at Low Concentrations: The Statistical Challenges of Verifying Ballast Water Discharge Standards" (*Ecological Applications*, 2013: 339-351)** considera utilizar el proceso de Poisson para este propósito].

a. ¿Cuál es la probabilidad de que un metro cúbico de descarga tenga al menos 8 organismos?

b. ¿Cuál es la probabilidad de que el número de organismos en 1.5 m^3 de agua de descarga exceda su valor medio por más de una desviación estándar?

c. Para qué cantidad de descarga la probabilidad de que haya menos de un organismo sería igual a 0.999?

87. El número de solicitudes de ayuda recibidas por un servicio de grúas es un proceso de Poisson con razón $\alpha = 4$ por hora.

a. Calcule la probabilidad de que exactamente diez solicitudes sean recibidas durante un periodo particular de 2 horas.

b. Si los operadores del servicio de grúas hacen una pausa de 30 minutos para el almuerzo, ¿cuál es la probabilidad de que no dejen de atender las llamadas de auxilio?

c. ¿Cuántas llamadas esperaría durante esta pausa?

88. Al someter a prueba tarjetas de circuito, la probabilidad de que cualquier diodo particular falle es de 0.01. Suponga que una tarjeta de circuito contiene 200 diodos.

a. ¿Cuántos diodos esperaría que fallen y cuál es la desviación estándar de la cantidad que se espera que falle?

b. ¿Cuál es la probabilidad (aproximada) de que al menos cuatro diodos fallen en una tarjeta seleccionada al azar?

c. Si se envían cinco tarjetas a un cliente particular, ¿qué tan probable es que al menos cuatro de ellas funcionen apropiadamente? (Una tarjeta funciona apropiadamente sólo si todos sus diodos funcionan.)

89. El artículo **"Reliability-Based Service-Life Assessment of Aging Concrete Structures" (*J. Structural Engr.*, 1993: 1600-1621)** sugiere que un proceso de Poisson puede ser utilizado para representar la ocurrencia de cargas estructurales en el transcurso del tiempo. Suponga que el tiempo medio entre ocurrencias de cargas es de 0.5 al año.

a. ¿Cuántas cargas se espera que ocurran durante un periodo de 2 años?

b. ¿Cuál es la probabilidad de que ocurran más de cinco cargas durante un periodo de 2 años?

c. ¿Qué tan largo debe ser un periodo de modo que la probabilidad de que no ocurran cargas durante dicho periodo sea cuando mucho de 0.1?

90. Sea X que tiene una distribución de Poisson con parámetro μ. Demuestre que $E(X) = \mu$ derivada directamente de la definición de valor esperado. [*Sugerencia*: El primer término en la suma es igual a 0 y luego x puede ser eliminada. Ahora saque como factor a μ y demuestre que lo que queda suma 1.]

91. Suponga que hay árboles distribuidos en un bosque de acuerdo con un proceso de Poisson bidimensional con parámetro α, el número esperado de árboles por acre es de 80.

a. ¿Cuál es la probabilidad de que en un terreno de un cuarto de acre, haya cuando mucho 16 árboles?

b. Si el bosque abarca 85 000 acres, ¿cuál es el número esperado de árboles en el mismo?

c. Suponga que selecciona un punto en el bosque y traza un círculo de 0.1 millas de radio. Sea X = el número de árboles dentro de esa región circular. ¿Cuál es la función de masa de probabilidad de X? [*Sugerencia*: 1 milla cuadrada = 640 acres.]

92. A una estación de inspección de equipo vehicular llegan automóviles de acuerdo con un proceso de Poisson con razón $\alpha = 10$ por hora. Suponga que un vehículo que llega con probabilidad de 0.5 no tendrá alteraciones en el equipo.

a. ¿Cuál es la probabilidad de que exactamente diez lleguen durante la hora y que ninguno tenga alteraciones?

b. Con cualquier $y \geq 10$ fija, ¿cuál es la probabilidad de que y llegue durante la hora y que diez no tengan alteraciones?

c. ¿Cuál es la probabilidad de que lleguen diez autos "sin alteraciones" durante la siguiente hora? [*Sugerencia*: Sume las probabilidades en el inciso b) desde $y = 10$ hasta ∞.]

93. **a.** En un proceso de Poisson, ¿qué tiene que suceder tanto en el intervalo de tiempo $(0, t)$ como en el intervalo $(t, t + \Delta t)$ de modo que no ocurran eventos en todo el intervalo $(0, t + \Delta t)$? Use esto y las suposiciones 1−3 para escribir una relación entre $P_0(t + \Delta t)$ y $P_0(t)$.

b. Use el resultado del inciso a) para escribir una expresión para la diferencia $P_0(t + \Delta t) - P_0(t)$. Divida entonces entre Δt y permita que $\Delta t \to 0$ para obtener una ecuación que implique $(d/dt)P_0(t)$, la derivada de $P_0(t)$ respecto a t.

c. Verifique que $P_0(t) = e^{-\alpha t}$ satisface la ecuación del inciso b).

d. Se puede demostrar de manera similar a los incisos a) y b) que los $P_k(t)$ deben satisfacer el sistema de ecuaciones diferenciales

$$\frac{d}{dt}P_k(t) = \alpha P_{k-1}(t) - \alpha P_k(t) \qquad k = 1, 2, 3, \ldots$$

Verifique que $P_k(t) = e^{-\alpha t}(\alpha t)^k/k!$ satisface el sistema. (En realidad esta es la única solución.)

EJERCICIOS SUPLEMENTARIOS (94-107)

94. Considere un mazo compuesto de siete cartas, marcadas 1, 2,…, 7. Se seleccionan al azar tres de ellas. Defina una variable aleatoria W como W = la suma de los números resultantes y calcule la función de masa de probabilidad de W. Calcule entonces μ y σ^2. [*Sugerencia*: Considere los resultados sin orden, de modo que (1, 3, 7) y (3, 1, 7) no sean resultados diferentes. Entonces existen 35 resultados y pueden ser puestos en lista. (Este tipo de variable aleatoria en realidad se presenta en conexión con una prueba de hipótesis llamada prueba de suma de renglones de Wilcoxon, en la cual hay una muestra x y una muestra y y W es la suma de los renglones de x en la muestra combinada.

95. Después de barajar un mazo de 52 cartas, un tallador reparte 5. Sea X = el número de palos representados en la mano de 5 cartas.

a. Demuestre que la función de masa de probabilidad de X es

x	1	2	3	4
$p(x)$	0.002	0.146	0.588	0.264

[*Sugerencia*: $p(1) = 4P$(todas son espadas), $p(2) = 6P$(sólo espadas y corazones con al menos una de cada palo) y $p(4) = 4P$(2 espadas ∩ una de cada uno de los otros palos).]

b. Calcule μ, σ^2 y σ.

96. La variable aleatoria binomial negativa X se definió como el número de fallas (F) que preceden al éxito (S) r-ésimo. Sea Y = el número de ensayos necesarios para obtener el *éxito* (S) r-ésimo. Del mismo modo en que fue obtenida la función de masa de probabilidad de X deduzca la función de masa de probabilidad de Y.

97. De todos los clientes que adquieren puertas de cochera automáticas, 75% adquiere el modelo de transmisión por cadena. Sea X = el número entre los siguientes 15 compradores que seleccionan el modelo de transmisión por cadena.

a. ¿Cuál es la función de masa de probabilidad de X?

b. Calcule $P(X > 10)$.

c. Calcule $P(6 \le X \le 10)$.

d. Calcule μ y σ^2.

e. Si actualmente la tienda tiene en existencia 10 modelos de transmisión por cadena y 8 modelos de transmisión por flecha, ¿cuál es la probabilidad de que las solicitudes de estos 15 clientes sean satisfechas con las existencias actuales?

98. En algunas aplicaciones la distribución de una variable aleatoria discreta X se asemeja a la distribución de Poisson con excepción de que cero no es un valor posible para X. Por ejemplo, sea X = el número de tatuajes que un individuo quiere remover cuando llega a una clínica especializada. Suponga que la función de masa de probabilidad de X es

$$p(x) = k\frac{e^{-\theta}\theta^x}{x} \quad x = 1, 2, 3,\ldots$$

a. Determine el valor de k. [*Sugerencia:* la suma de todas las probabilidades en la función de masa de probabilidad de Poisson es 1, y esta función de masa de probabilidad también debe sumar 1.]

b. Si el valor de la media de X es 2.313035, ¿cuál es la probabilidad de que un individuo quiera remover a lo más 5 tatuajes?

c. Determine la desviación estándar de X cuando el valor de la media es igual al que se menciona en b).

[*Nota*: El artículo **"An Exploratory Investigation of Identity Negotiation and Tattoo Removal" (*Academy of Marketing Science Review*, vol. 12, núm. 6, 2008)** da una muestra de 22 observaciones en el número de tatuajes que la gente quiere remover; los *estimados* de μ y σ calculados a partir de los datos fueron 2.318182 y 1.249242, respectivamente.]

99. Un *sistema k de n* es uno que funcionará si y sólo si al menos k de los n componentes individuales en el sistema funcionan. Si los componentes individuales funcionan independientemente uno de otro, cada uno con probabilidad de 0.9, ¿cuál es la probabilidad de que un sistema 3 de 5 funcione?

100. Un fabricante de chips de circuitos integrados desea controlar la calidad de sus productos rechazando cualquier lote en el que la proporción de chips sea demasiado alta. Con esta finalidad de cada lote de 10 000 chips se seleccionarán y probarán 25. Si al menos 5 de estos están defectuosos, todo el lote será rechazado.

a. ¿Cuál es la probabilidad de que un lote sea rechazado si 5% de los chips en el lote están, de hecho, defectuosos?

b. Responda la pregunta del inciso a) si el porcentaje de chips defectuosos es 10%.

c. Responda la pregunta del inciso a) si el porcentaje de chips defectuosos es 20%.

d. ¿Qué les sucedería a las probabilidades en los incisos a)−c) si el número de rechazo crítico se incrementara de 5 a 6?

101. De las personas que pasan a través de un detector de metales en un aeropuerto, 0.5% lo activan; sea X = el número de personas, entre un grupo de 500 seleccionado al azar, que activan el detector.

a. ¿Cuál es la función de masa de probabilidad (aproximada) de X?

b. Calcule $P(X = 5)$.

c. Calcule $P(5 \le X)$.

102. Una consultora educativa está tratando de decidir si los estudiantes de preparatoria que nunca antes han utilizado una calculadora de mano pueden resolver cierto tipo de problema más fácilmente con una calculadora que utiliza lógica polaca inversa o con una que no utiliza esta lógica. Se selecciona una muestra de 25 estudiantes y se les permite practicar con ambas calculadoras. Luego a cada estudiante se le pide que resuelva un problema con la calculadora polaca inversa y un problema similar con la otra. Sea $p = P(S)$, donde S indica que un estudiante resolvió el problema más rápido con la lógica polaca inversa que sin ella y sea X = número de éxitos.

a. Si $p = 0.5$, ¿cuáles $P(7 \le X \le 18)$?

b. Si $p = 0.8$, ¿cuáles $P(7 \le X \le 18)$?

c. Si la pretensión de que $p = 0.5$ tiene que ser rechazada cuando $x \leq 7$ o cuando $x \geq 18$, ¿cuál es la probabilidad de rechazar la pretensión cuando en realidad es correcta?

d. Si la decisión de rechazar la pretensión $p = 0.5$ se hace como en el inciso c), ¿cuál es la probabilidad de que la pretensión no sea rechazada cuando $p = 0.6$? ¿Y cuándo $p = 0.8$?

e. ¿Qué regla de decisión escogería para rechazar la pretensión de que $p = 0.5$, si desea que la probabilidad en el inciso c) sea cuando mucho de 0.01?

103. Considere una enfermedad cuya presencia puede ser identificada mediante un análisis de sangre. Sea p la probabilidad de que un individuo seleccionado al azar tenga la enfermedad. Suponga que se seleccionan independientemente n individuos para analizarlos. Una forma de proceder es analizar cada una de las n muestras de sangre. Un procedimiento potencialmente más económico, de análisis en grupo, se introdujo durante la segunda Guerra Mundial para identificar a sifilíticos entre los reclutas. En primer lugar, se toma una parte de cada muestra de sangre, se combinan estos especímenes y se realiza un solo análisis. Si ninguno tiene la enfermedad, el resultado será negativo y sólo se requiere un análisis. Si al menos un individuo está enfermo, el análisis de la muestra combinada dará un resultado positivo, en cuyo caso se realizan los análisis de los n individuos. Si $p = 0.1$ y $n = 3$, ¿cuál es el número esperado de análisis si se utiliza este procedimiento? ¿Cuál es el número esperado cuando $n = 5$? [El artículo **"Random Multiple-Access Communication and Group Testing" (*IEEE Trans. on Commun.*, 1984: 769-774)** aplicó estas ideas a un sistema de comunicación en el cual la dicotomía fue usuario ocioso/activo en lugar de enfermo/no enfermo.]

104. Sea p_1 la probabilidad de que cualquier símbolo de código particular sea erróneamente transmitido a través de un sistema de comunicación. Suponga que en diferentes símbolos ocurren errores de manera independiente uno de otro. Suponga también que con probabilidad p_2 un símbolo erróneo es corregido al ser recibido. Sea X el número de símbolos correctos en un bloque de mensajes compuesto de n símbolos (una vez que el proceso de corrección ha terminado). ¿Cuál es la distribución de probabilidad de X?

105. El comprador de una unidad generadora de potencia requiere c arranques consecutivos exitosos antes de aceptar la unidad. Suponga que los resultados de arranques individuales son independientes entre sí. Sea p la probabilidad de que cualquier arranque particular sea exitoso. La variable aleatoria de interés es $X =$ el número de arranques que deben hacerse antes de la aceptación. Dé la función de masa de probabilidad de X en el caso $c = 2$. Si $p = 0.9$, ¿cuál es $P(X \leq 8)$? [*Sugerencia*: Con $x \geq 5$, exprese $p(x)$ "recursivamente" en términos de la función de masa de probabilidad evaluada con los valores más pequeños $x - 3$, $x - 4$,…, 2.] (Este problema fue sugerido por el artículo **"Evaluation of a Start-Up Demonstration Test", *J. Quality Technology*, 1983: 103-106.)**

106. Una aerolínea ha desarrollado un plan para un club de viajeros ejecutivos sobre la premisa de que 10% de sus clientes actuales calificaría para la membresía.

a. Suponiendo la validez de esta premisa, de 25 clientes actuales seleccionados al azar, ¿cuál es la probabilidad de que entre 2 y 6 (inclusive) califiquen para la membresía?

b. De nuevo, suponiendo la validez de la premisa ¿cuál es el número esperado de clientes que califican y la desviación estándar del número que califica en una muestra aleatoria de 100 clientes actuales?

c. Sea X el número en una muestra al azar de 25 clientes actuales que califican para la membresía. Considere rechazar la premisa de la compañía en favor de la pretensión de que $p > 0.10$ si $x \geq 7$. ¿Cuál es la probabilidad de que la premisa de la compañía sea rechazada cuando en realidad es válida?

d. Remítase a la regla de decisión introducida en el inciso c). ¿Cuál es la probabilidad de que la premisa de la compañía no sea rechazada aun cuando $p = 0.20$ (es decir, 20% califica)?

107. Cuarenta por ciento de las semillas de las mazorcas de maíz (maíz moderno) portan sólo una espiga y el restante 60% porta dos espigas. Una semilla con una espiga producirá una mazorca con espigas únicas 29% del tiempo, en tanto que una semilla con dos espigas producirá una mazorca con espigas únicas 26% del tiempo. Considere seleccionar al azar diez semillas.

a. ¿Cuál es la probabilidad de que exactamente cinco de estas semillas porten una sola espiga y produzcan una mazorca con una sola espiga?

b. ¿Cuál es la probabilidad de que exactamente cinco de estas mazorcas producidas por estas semillas tengan espigas únicas? ¿Cuál es la probabilidad de que a lo más cinco mazorcas tengan espigas únicas?

BIBLIOGRAFÍA

Johnson, Norman, Samuel Kotz y Adrienne Kemp, *Discrete Univariate Distributions*, Wiley, Nueva York, 1992. Una enciclopedia de información sobre distribuciones discretas.

Olkin, Ingram, Cyrus Derman y Leon Gleser, *Probability Models and Applications* (2a. ed.), Macmillan, Nueva York, 1994. Contiene una discusión a fondo tanto de las propiedades generales de las distribuciones discretas y continuas como los resultados para las distribuciones específicas.

Ross, Sheldon, *Introduction to Probability Models* (9a. ed.), Academic Press, Nueva York, 2007. Una buena fuente de material sobre el proceso de Poisson y generalizaciones, y una amena introducción a otros temas de probabilidad aplicada.

Capítulo

4

Variables aleatorias continuas y distribuciones de probabilidad

INTRODUCCIÓN

El capítulo 3 se concentró en el desarrollo de distribuciones de probabilidad de variables aleatorias discretas. En este capítulo se estudia el segundo tipo general de variable aleatoria, que se presenta en muchos problemas aplicados. Las secciones 4.1 y 4.2 presentan las definiciones y propiedades básicas de las variables aleatorias continuas y sus distribuciones de probabilidad. En la sección 4.3 se estudia con detalle la variable aleatoria normal y su distribución, sin duda la más importante y útil en la probabilidad y la estadística. La sección 4.4 se ocupa de otras distribuciones continuas utilizadas con frecuencia en el trabajo aplicado.

4.1 Funciones de densidad de probabilidad

Una variable aleatoria discreta es aquella cuyos valores posibles constituyen un conjunto finito, o bien pueden ser puestos en lista en una secuencia infinita (una lista en la cual existe un primer elemento, un segundo elemento, etc.). Una variable aleatoria cuyo conjunto de valores posibles es un intervalo completo de números no es discreta.

De acuerdo con el capítulo 3 recuerde que una variable aleatoria X es continua si 1) sus valores posibles comprenden un solo intervalo sobre la recta numérica (para alguna $A < B$, cualquier número x entre A y B es un valor posible) o una unión de intervalos disjuntos y 2) $P(X = c) = 0$ para cualquier número c que sea un valor posible de X.

EJEMPLO 4.1 En el estudio de la ecología de un lago se mide la profundidad en lugares seleccionados, entonces $X =$ la profundidad en ese lugar es una variable aleatoria continua. En este caso A es la profundidad mínima en la región muestreada y B es la profundidad máxima. ■

EJEMPLO 4.2 Si se selecciona al azar un compuesto químico y se determina su pH X, entonces X es una variable aleatoria continua porque cualquier valor pH entre 0 y 14 es posible. Si se conoce más sobre el compuesto seleccionado para su análisis, entonces el conjunto de posibles valores podría ser un subintervalo de [0, 14], tal como $5.5 \leq x \leq 6.5$, pero X seguiría siendo continua. ■

EJEMPLO 4.3 Sea X el tiempo que espera un cliente seleccionado al azar antes de que comience su corte de cabello. El primer pensamiento podría ser que X es una variable aleatoria continua, puesto que es necesario medirla para determinar su valor. Sin embargo, existen clientes suficientemente afortunados que no tienen que esperar antes de sentarse en el sillón de la peluquería. Así que el caso debe ser $P(X = 0) > 0$. Aunque, en caso de que no haya sillones vacíos, el tiempo de espera será continuo puesto que X podría asumir entonces cualquier valor entre un tiempo mínimo posible A y un tiempo máximo posible B. Esta variable aleatoria no es ni puramente discreta ni puramente continua, sino que es una mezcla de los dos tipos. ■

Se podría argumentar que aunque en principio las variables como altura, peso y temperatura son continuas, en la práctica las limitaciones de los instrumentos de medición nos restringen a un mundo discreto (aunque en ocasiones muy sutilmente subdividido). Sin embargo, los modelos continuos a menudo representan muy bien de forma aproximada situaciones del mundo real y con frecuencia es más fácil trabajar con matemáticas continuas (cálculo) que con matemáticas de variables discretas y distribuciones.

Distribuciones de probabilidad de variables continuas

Suponga que la variable X que nos interesa es la profundidad de un lago en un punto sobre la superficie seleccionado al azar. Sea $M =$ la profundidad máxima (en metros), así que cualquier número en el intervalo $[0, M]$ es un valor posible de X. Si se "discretiza" X midiendo la profundidad al metro más cercano, entonces los valores posibles son enteros no negativos menores o iguales a M. La distribución discreta de la profundidad resultante se ilustra con un histograma de probabilidad. Si se traza el histograma de modo que el área del rectángulo sobre cualquier entero posible k sea la proporción del lago cuya profundidad es (al metro más cercano) k, entonces el área total de todos los rectángulos es 1. En la figura 4.1(a) aparece un posible histograma.

Si se mide la profundidad con mucha más precisión y se utiliza el mismo eje de medición de la figura 4.1(a), cada rectángulo en el histograma de probabilidad resultante es

mucho más angosto, aun cuando el área total de todos los rectángulos sigue siendo 1. En la figura 4.1(b) se ilustra un posible histograma; tiene una apariencia mucho más regular que el histograma de la figura 4.1(a). Si se continúa de esta manera, midiendo la profundidad más y más finamente, la secuencia resultante de histogramas se aproxima a una curva más regular, tal como la que se ilustra en la figura 4.1(c). Puesto que en cada histograma el área total de todos los rectángulos es igual a 1, el área total bajo la curva regular también es 1. La probabilidad de que la profundidad en un punto seleccionado al azar se encuentre entre a y b es simplemente el área bajo la curva regular entre a y b. Es de manera exacta una curva suave del tipo ilustrado en la figura 4.1(c) la que especifica una distribución de probabilidad continua.

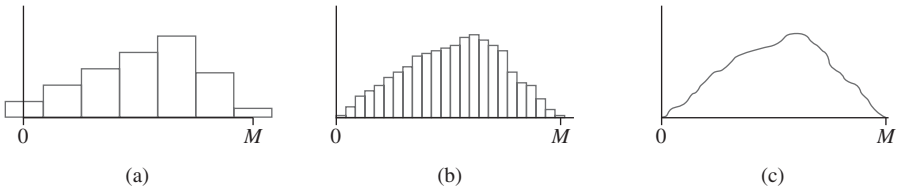


Figura 4.1 (a) Histograma de probabilidad de profundidad medida al metro más cercano; (b) histograma de probabilidad de profundidad medida al centímetro más cercano; (c) un límite de una secuencia de histogramas discretos

DEFINICIÓN

Sea X una variable aleatoria continua. Entonces, una **distribución de probabilidad** o **función de densidad de probabilidad** (pdf) de X es una función $f(x)$ de modo tal que para dos números cualesquiera a y b con $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)dx$$

Es decir, la probabilidad de que X asuma un valor en el intervalo $[a, b]$ es el área sobre este intervalo y bajo la gráfica de la función de densidad, como se ilustra en la figura 4.2. La gráfica de $f(x)$ a menudo se conoce como *curva de densidad*.

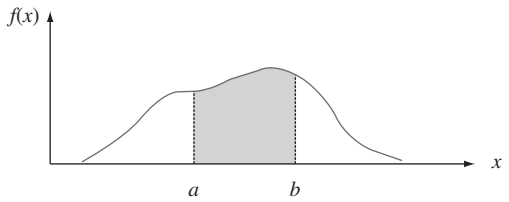


Figura 4.2 $p(a \le X \le b)$ = área bajo la curva de densidad entre a y b

Para que $f(x)$ sea una función de densidad de probabilidad legítima debe satisfacer las dos siguientes condiciones:

1. $f(x) \ge 0$ con todas las x

2. $\int_{-\infty}^{\infty} f(x)\,dx$ = área bajo toda la gráfica de $f(x)$
$= 1$

EJEMPLO 4.4 La dirección de una imperfección respecto a una línea de referencia sobre un objeto circular como un neumático, un rotor de freno o un volante está, en general, sujeta a incertidumbre. Considere la línea de referencia que conecta el vástago de la válvula de un neumático con el punto central, y sea X el ángulo medido en el sentido de las manecillas

del reloj respecto a la ubicación de una imperfección. Una posible función de densidad de probabilidad de X es

$$f(x) = \begin{cases} \dfrac{1}{360} & 0 \le x < 360 \\ 0 & \text{de lo contrario} \end{cases}$$

La función de densidad de probabilidad aparece graficada en la figura 4.3. Claramente $f(x) \ge 0$. El área bajo la curva de densidad es simplemente el área de un rectángulo: (altura)(base) = (1/360)(360) = 1. La probabilidad de que el ángulo sea de entre 90° y 180° es

$$P(90 \le X \le 180) = \int_{90}^{180} \frac{1}{360}\,dx = \frac{x}{360}\Big|_{x=90}^{x=180} = \frac{1}{4} = 0.25$$

La probabilidad de que el ángulo de ocurrencia esté dentro de 90° de la línea de referencia es

$$P(0 \le X \le 90) + P(270 \le X < 360) = 0.25 + 0.25 = 0.50$$

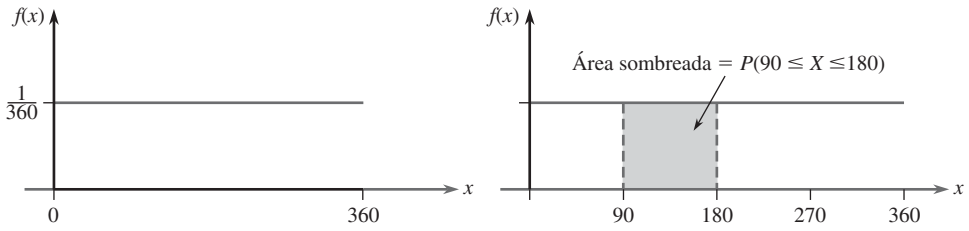


Figura 4.3 Función de densidad de probabilidad del ejemplo 4.4

Debido a que siempre que $0 \le a \le b \le 360$ en el ejemplo 4.4, $P(a \le X \le b)$ depende sólo del ancho $b - a$ del intervalo, se dice que X tiene una distribución uniforme.

DEFINICIÓN

Se dice que una variable aleatoria continua X tiene una **distribución uniforme** en el intervalo $[A, B]$ si la función de densidad de probabilidad de X es

$$f(x; A, B) = \begin{cases} \dfrac{1}{B - A} & A \le x \le B \\ 0 & \text{de lo contrario} \end{cases}$$

La gráfica de cualquier función de densidad de probabilidad uniforme es como la de la figura 4.3, excepto que el intervalo de densidad positiva es $[A, B]$ en lugar de [0, 360].

En el caso discreto, una función de masa de probabilidad (pmf) dice cuántas pequeñas "burbujas" de masa de probabilidad de varias magnitudes están distribuidas a lo largo del eje de medición. En el caso continuo la densidad de probabilidad está "repartida" en forma continua a lo largo del intervalo de posibles valores. Cuando la densidad está distribuida uniformemente a lo largo del intervalo se obtiene una función de densidad de probabilidad uniforme tal como en la figura 4.3.

Cuando X es una variable aleatoria discreta a cada valor posible se le asigna una probabilidad positiva. Esto no es así en el caso de una variable aleatoria continua (es decir, se

satisface la segunda condición de la definición) porque el área bajo una curva de densidad situada sobre cualquier valor único es cero:

$$P(X = c) = \int_c^c f(x)\,dx = \lim_{\varepsilon \to 0} \int_{c-\varepsilon}^{c+\varepsilon} f(x)\,dx = 0$$

El hecho de que $P(X = c) = 0$ cuando X es continua tiene una importante consecuencia práctica: la probabilidad de que X quede en algún intervalo entre a y b no depende de si el límite inferior a o el límite superior b están incluido en el cálculo de probabilidad:

$$P(a \le X \le b) = P(a < X < b) = P(a < X \le b) = P(a \le X < b) \qquad \textbf{(4.1)}$$

Si X es discreta y tanto a como b son valores posibles (p. ej., X es binomial con $n = 20$ y $a = 5$, $b = 10$), entonces todas las cuatro probabilidades en (4.1) son diferentes.

La condición de probabilidad cero tiene un análogo físico. Considere una barra circular sólida con área de sección transversal = 1 pulg^2. Coloque la barra a lo largo de un eje de medición y suponga que la densidad de la barra en cualquier punto x está dada por el valor $f(x)$ de una función de densidad. Entonces si se corta un segmento en los puntos a y b de la barra y este segmento se retira, la cantidad de masa eliminada es $\int_a^b f(x)\,dx$; si la barra se corta exactamente en el punto c no se elimina masa. Se asigna masa a segmentos de intervalo de la barra pero no a puntos individuales.

EJEMPLO 4.5 "Intervalo de tiempo" en el flujo de tránsito es el lapso transcurrido entre el tiempo en que un automóvil termina de pasar por un punto fijo y el instante en que el siguiente vehículo comienza a pasar por ese punto. Sea X = el intervalo de tiempo para dos automóviles consecutivos seleccionados al azar en una autopista durante un periodo de tráfico intenso. La siguiente función de densidad de probabilidad de X es en esencia la sugerida en **"The Statistical Properties of Freeway Traffic" (*Transp. Res*. vol. 11: 221-228)**:

$$f(x) = \begin{cases} 0.15e^{-15(x-0.5).} & x \ge 0.5 \\ 0 & \text{de lo contrario} \end{cases}$$

La gráfica de $f(x)$ se da en la figura 4.4; no hay ninguna densidad asociada con intervalos de tiempo de menos de 0.5 y la densidad del intervalo de tiempo decrece con rapidez (exponencialmente rápido) a medida que x se incrementa a partir de 0.5. Claramente $f(x) \ge 0$; para demostrar que $\int_{-\infty}^{\infty} f(x)\,dx = 1$, se utiliza el resultado obtenido con cálculo integral $\int_a^{\infty} e^{-kx}\,dx = (1/k)e^{-k\cdot a}$. Entonces

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0.5}^{\infty} 0.15e^{-0.15(x-0.5)}\,dx = 0.15e^{0.075}\int_{0.5}^{\infty} e^{-0.15x}\,dx$$

$$= 0.15e^{0.075} \cdot \frac{1}{0.15}\, e^{-(0.15)(0.5)} = 1$$

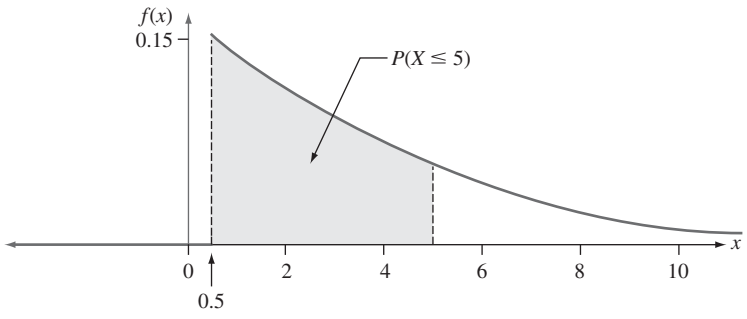


Figura 4.4 Curva de densidad del intervalo de tiempo entre vehículos en el ejemplo 4.5

La probabilidad de que el intervalo de tiempo sea cuando mucho de 5 segundos es

$$\begin{aligned} P(X \le 5) &= \int_{-\infty}^{5} f(x)\,dx = \int_{0.5}^{5} 0.15e^{-0.15(x-0.5)}\,dx \\ &= 0.15e^{0.075}\int_{0.5}^{5} e^{-0.15x}\,dx = 0.15e^{0.075} \cdot \left(-\frac{1}{0.15}\, e^{-0.15x} \Big|_{x=0.5}^{x=5} \right) \\ &= e^{0.075}(-e^{-0.75} + e^{-0.075}) = 1.078(-0.472 + 0.928) = 0.491 \\ &= P(\text{menos de 5 s}) = P(X < 5) \end{aligned}$$

■

A diferencia de las distribuciones discretas como la binomial, la hipergeométrica y la binomial negativa, la distribución de cualquier variable aleatoria continua dada no puede, en general, ser obtenida mediante argumentos probabilísticos. En cambio, se debe hacer una selección juiciosa de la función de densidad de probabilidad basada en conocimientos previos y en los datos disponibles. Afortunadamente, existen algunas familias generales de funciones de densidad de probabilidad que se ajustan bien a una amplia variedad de situaciones experimentales; varias de éstas se discuten más adelante en el capítulo.

Exactamente como en el caso discreto, a menudo es útil pensar en la población de interés como compuesta de valores X en lugar de individuos u objetos. La función de densidad de probabilidad es entonces un modelo de la distribución de valores en esta población numérica y, con base en este modelo, se pueden calcular varias características de la población (tal como la media).

EJERCICIOS **Sección 4.1 (1-10)**

1. La corriente en un circuito determinado medido por un amperímetro es una variable aleatoria continua X con la función de densidad siguiente:

$$f(x) = \begin{cases} 0.075x + 0.2 & 3 \le x \le 5 \\ 0 & \text{de lo contrario} \end{cases}$$

- **a.** Grafique la función de densidad de probabilidad para verificar que el área total bajo la curva de densidad es, de hecho, 1.
- **b.** Calcule $P(X \le 4)$. ¿Cómo se compara esta probabilidad con $P(X < 4)$?
- **c.** Calcule $P(3.5 \le X \le 4.5)$ y $P(4.5 < X)$.

2. Suponga que la temperatura de reacción X (en °C) en cierto proceso químico tiene una distribución uniforme con $A = -5$ y $B = 5$.

- **a.** Calcule $P(X < 0)$.
- **b.** Calcule $P(-2.5 < X < 2.5)$.
- **c.** Calcule $P(-2 \le X \le 3)$.
- **d.** Para que k satisfaga $-5 < k < k + 4 < 5$, calcule $P(k < X < k + 4)$.

3. El error implicado al hacer una medición es una variable aleatoria continua X con función de densidad de probabilidad

$$f(x) = \begin{cases} 0.09375(4 - x^2) & -2 \le x \le 2 \\ 0 & \text{de lo contrario} \end{cases}$$

- **a.** Trace la gráfica de $f(x)$.
- **b.** Calcule $P(X > 0)$.
- **c.** Calcule $P(-1 < X < 1)$.
- **d.** Calcule $P(X < -0.5 \text{ o } X > 0.5)$.

4. Sea X el esfuerzo vibratorio (lb/pulg^2) en el aspa de una turbina de viento a una velocidad del viento particular en un túnel aerodinámico. El artículo **"Blade Fatigue Life Assessment with Application to VAWTS"** (***J. of Solar Energy Engr.*, 1982: 107–111**) propone la distribución de Rayleigh, con función de densidad de probabilidad

$$f(x;\theta) = \begin{cases} \dfrac{x}{\theta^2} \cdot e^{-x^2/(2\theta^2)} & x > 0 \\ 0 & \text{de lo contrario} \end{cases}$$

como modelo de la distribución X.

- **a.** Verifique que $f(x;\ \theta)$ es una función de densidad de probabilidad legítima.
- **b.** Suponga que $\theta = 100$ (un valor sugerido por una gráfica en el artículo). ¿Cuál es la probabilidad de que X sea cuando mucho 200? ¿Y menos de 200? ¿Por lo menos 200?
- **c.** ¿Cuál es la probabilidad de que X esté entre 100 y 200 (de nuevo suponiendo $\theta = 100$)?
- **d.** Dé una expresión para $P(X \le x)$.

5. Un profesor universitario nunca termina su disertación antes de que concluya la hora y siempre termina cerca de dos minutos después de la hora. Sea X = el tiempo que transcurre entre el final de la hora y el final de la disertación, y suponga que la función de densidad de probabilidad de X es

$$f(x) = \begin{cases} kx^2 & 0 \le x \le 2 \\ 0 & \text{de lo contrario} \end{cases}$$

a. Determine el valor de k y trace la curva de densidad correspondiente. [*Sugerencia:* El área total bajo la gráfica de $f(x)$ es 1.]

b. ¿Cuál es la probabilidad de que la disertación termine dentro del minuto final de la hora?

c. ¿Cuál es la probabilidad de que la disertación continúe después de la hora entre 60 y 90 segundos?

d. ¿Cuál es la probabilidad de que la disertación continúe durante al menos los primeros 90 segundos después de la hora?

6. El peso real de lectura de la pastilla de un estéreo ajustado a 3 gramos en un tocadiscos particular puede ser considerado como una variable aleatoria continua X con función de densidad de probabilidad

$$f(x) = \begin{cases} k[1 - (x - 3)^2] & 2 \le x \le 4 \\ 0 & \text{de lo contrario} \end{cases}$$

a. Trace la gráfica de $f(x)$.

b. Determine el valor de k.

c. ¿Cuál es la probabilidad de que el peso real de lectura sea mayor que el peso prescrito?

d. ¿Cuál es la probabilidad de que el peso real de lectura esté dentro de 0.25 gramos del peso prescrito?

e. ¿Cuál es la probabilidad de que el peso real difiera del peso prescrito por más de 0.5 gramos?

7. El artículo **"Second Moment Reliability Evaluation vs. Monte Carlo Simulations for Weld Fatigue Strength" (*Quality and Reliability Engr. Intl.*, 2012: 887–896)** considera el uso de una distribución uniforme con $A = 0.20$ y $B = 4.25$ para el diámetro X de un determinado tipo de soldadura (mm).

a. Determine la función de densidad de probabilidad de X y trace la gráfica.

b. ¿Cuál es la probabilidad de que el diámetro supere los 3 mm?

c. ¿Cuál es la probabilidad de que el diámetro esté dentro de 1 mm del diámetro promedio?

d. Para cualquier valor a satisfactorio $0.20 < a < a + 1 < 4.25$, ¿qué es $P(a < X < a + 1)$?

8. Para llegar a su trabajo un profesor primero debe abordar un autobús cerca de su casa y luego un segundo autobús. Si el tiempo de espera (en minutos) en cada parada tiene una distribución uniforme con $A = 0$ y $B = 5$, entonces se puede demostrar que el tiempo de espera total Y tiene la función de densidad de probabilidad

$$f(y) = \begin{cases} \frac{1}{25}y & 0 \le y < 5 \\ \frac{2}{5} - \frac{1}{25}y & 5 \le y \le 10 \\ 0 & y < 0 \text{ o } y > 10 \end{cases}$$

a. Trace la gráfica de la función de densidad de probabilidad de Y.

b. Verifique que $\int_{-\infty}^{\infty} f(y)\,dy = 1$.

c. ¿Cuál es la probabilidad de que el tiempo de espera total sea cuando mucho de 3 min?

d. ¿Cuál es la probabilidad de que el tiempo de espera total sea cuando mucho de 8 min?

e. ¿Cuál es la probabilidad de que el tiempo de espera total esté entre 3 y 8 min?

f. ¿Cuál es la probabilidad de que el tiempo de espera total sea de menos de 2 min o de más de 6 min?

9. Con base en un análisis de datos muestrales, el artículo **"Pedestrians′ Crossing Behaviors and Safety at unmarked Roadways in China" (*Accident Analysis and Prevention*, 2011: 1927–1936)** propone la función de densidad de probabilidad $f(x) = 0.15e^{-.15(x-1)}$ cuando $x \ge 1$ como un modelo para la distribución de $X =$ tiempo (s) en la línea media.

a. ¿Cuál es la probabilidad de que ese tiempo de espera sea a lo sumo de 5 s? De más de 5 s?

b. ¿Cuál es la probabilidad de que ese tiempo de espera esté entre 2 y 5 s?

10. Una familia de funciones de densidad de probabilidad que ha sido utilizada para aproximar la distribución del ingreso, el tamaño de la población de una ciudad y el tamaño de las empresas es la familia Pareto. La familia tiene dos parámetros, k y θ, ambos > 0, y la función de densidad de probabilidad es

$$f(x; k, \theta) = \begin{cases} \dfrac{k \cdot \theta^k}{x^{k+1}} & x \ge \theta \\ 0 & x < \theta \end{cases}$$

a. Trace la gráfica de $f(x; k, \theta)$.

b. Verifique que el área total bajo la gráfica es igual a 1.

c. Si la variable aleatoria X tiene una función de densidad de probabilidad $f(x; k, \theta)$, con cualquier $b > \theta$ fija, obtenga una expresión para $P(X \le b)$.

d. Para $\theta < a < b$, obtenga una expresión para la probabilidad $P(a \le X \le b)$.

4.2 Funciones de distribución acumulada y valores esperados

Varios de los más importantes conceptos introducidos en el estudio de las distribuciones discretas también desempeñan un importante papel en las distribuciones continuas. Las definiciones análogas a las del capítulo 3 implican reemplazar suma por integración.

Función de distribución acumulada

La función de distribución acumulada $F(x)$ de una variable aleatoria discreta X, con cualquier número especificado, da x; la probabilidad $P(X \leq x)$. Esta se obtiene al sumar la función de masa de probabilidad $p(y)$ a lo largo de todos los valores posibles y que satisfagan $y \leq x$. La función de distribución acumulada de una variable aleatoria continua da las mismas probabilidades $P(X \leq x)$ y se obtiene integrando la función de densidad de probabilidad $f(y)$ entre los límites $-\infty$ y x.

DEFINICIÓN

La **función de distribución acumulada** $F(x)$ de una variable aleatoria continua X se define para todo número x como

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy$$

Para cada x, $F(x)$ es el área bajo la curva de densidad a la izquierda de x. Esto se ilustra en la figura 4.5 donde $F(x)$ se incrementa con suavidad a medida que x se incrementa.

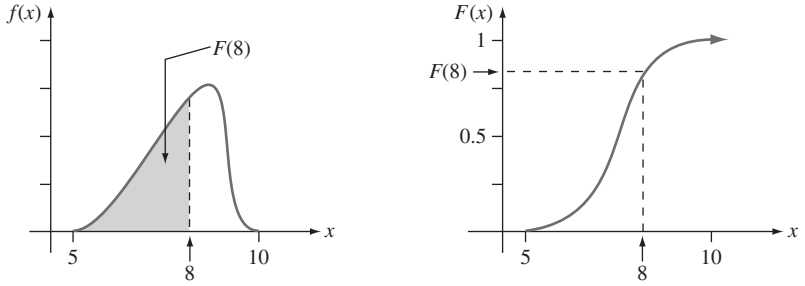


Figura 4.5 Una función de densidad de probabilidad y una función de distribución acumulada asociada

EJEMPLO 4.6 Sea X el espesor de una cierta lámina de metal con distribución uniforme en $[A, B]$. La función de densidad se muestra en la figura 4.6. Para $x < A$, $F(x) = 0$, dado que no hay área bajo la gráfica de la función de densidad a la izquierda de la x. Con $x \geq B$, $F(x) = 1$, puesto que toda el área está acumulada a la izquierda de la x. Finalmente para $A \leq x \leq B$,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{A}^{x} \frac{1}{B-A}\,dy = \frac{1}{B-A} \cdot y \,\Big|_{y=A}^{y=x} = \frac{x-A}{B-A}$$

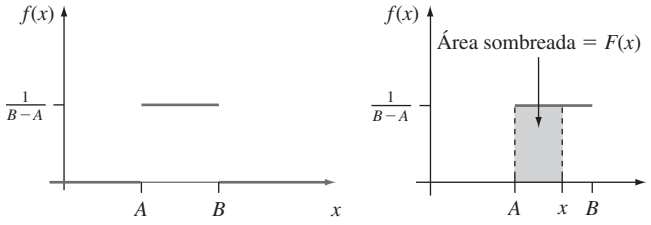


Figura 4.6 Función de densidad de probabilidad de una distribución uniforme

La función de distribución acumulada es

$$F(x) = \begin{cases} 0 & x < A \\ \dfrac{x-A}{B-A} & A \leq x < B \\ 1 & x \geq B \end{cases}$$

La gráfica de esta función de distribución acumulada aparece en la figura 4.7.

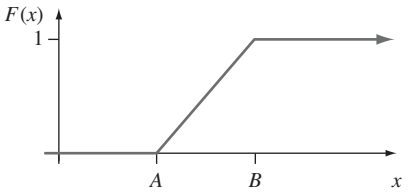


Figura 4.7 Función de distribución acumulada de una distribución uniforme ■

Utilizar $F(x)$ para calcular probabilidades

La importancia de la función de distribución acumulada en este caso, lo mismo que para variables aleatorias discretas, es que las probabilidades de varios intervalos pueden ser calculadas con una fórmula o tabla de $F(x)$.

PROPOSICIÓN

Sea X una variable aleatoria continua con función de densidad de probabilidad $f(x)$ y función de distribución acumulada $F(x)$. Entonces para cualquier número a,

$$P(X > a) = 1 - F(a)$$

y para dos números cualesquiera a y b con $a < b$,

$$P(a \le X \le b) = F(b) - F(a)$$

La figura 4.8 ilustra la segunda parte de esta proposición; la probabilidad deseada es el área sombreada bajo la curva de densidad entre a y b, y es igual a la diferencia entre las dos áreas acumuladas sombreadas. Esto difiere de lo que es apropiado para una variable aleatoria discreta de valor entero (p. ej., binomial o Poisson): $P(a \le X \le b) = F(b) - F(a - 1)$ cuando a y b son enteros.

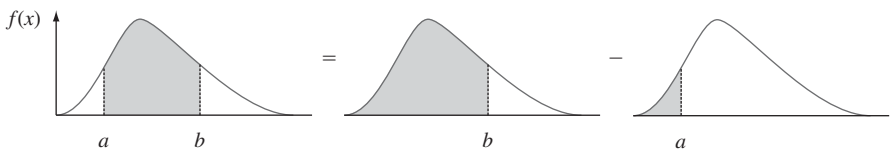


Figura 4.8 Cálculo de $P(a \le X \le b)$ a partir de probabilidades acumuladas

EJEMPLO 4.7 Suponga que la función de densidad de probabilidad de la magnitud X de una carga dinámica sobre un puente (en newtons) está dada por

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \le x \le 2 \\ 0 & \text{de lo contrario} \end{cases}$$

Para cualquier número x entre 0 y 2,

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{0}^{x} \left(\frac{1}{8} + \frac{3}{8}y\right) dy = \frac{x}{8} + \frac{3}{16}x^2$$

Por tanto,

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{8} + \frac{3}{16}x^2 & 0 \le x \le 2 \\ 1 & 2 < x \end{cases}$$

Las gráficas de $f(x)$ y $F(x)$ se muestran en la figura 4.9. La probabilidad de que la carga esté entre 1 y 1.5 es

$$\begin{aligned} P(1 \le X \le 1.5) &= F(1.5) - F(1) \\ &= \left[\frac{1}{8}(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] \\ &= \frac{19}{64} = 0.297 \end{aligned}$$

La probabilidad de que la carga sea de más de 1 es

$$\begin{aligned} P(X > 1) &= 1 - P(X \le 1) = 1 - F(1) = 1 - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] \\ &= \frac{11}{16} = 0.688 \end{aligned}$$

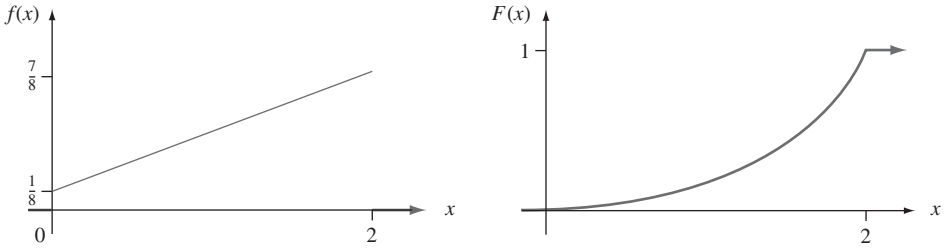


Figura 4.9 Función de densidad de probabilidad y función de distribución acumulada del ejemplo 4.7

Una vez que se obtiene la función de distribución acumulada, cualquier probabilidad que implique X es fácil de calcular sin ninguna integración adicional.

Obtención de $f(x)$ a partir de $F(x)$

Para X discreta la función de masa de probabilidad se obtiene a partir de la función de distribución acumulada considerando la diferencia entre dos valores $F(x)$. El análogo continuo de una diferencia es una derivada. El siguiente resultado es una consecuencia del teorema fundamental del cálculo.

PROPOSICIÓN

Si X es una variable aleatoria continua con función de densidad de probabilidad $f(x)$ y función de distribución acumulada $F(x)$, entonces en cada x que hace posible que la derivada $F'(x)$ exista, $F'(x) = f(x)$.

EJEMPLO 4.8 (Continuación del ejemplo 4.6)

Cuando X tiene una distribución uniforme, $F(x)$ es derivable excepto en $x = A$ y $x = B$, donde la gráfica de $F(x)$ tiene esquinas afiladas. Como $F(x) = 0$ para $x < A$ y $F(x) = 1$ para $x > B$, $F'(x) = 0 = f(x)$ con dicha x. Para $A < x < B$,

$$F'(x) = \frac{d}{dx}\left(\frac{x - A}{B - A}\right) = \frac{1}{B - A} = f(x)$$

Percentiles de una distribución continua

Cuando se dice que la calificación de un individuo en una prueba estaba en el 85º percentil de la población, significa que 85% de todas las calificaciones de la población estuvieron por debajo de dicha calificación y que 15% estuvo por encima. Asimismo, el 40º percentil

es la calificación que sobrepasa a 40% de todas las calificaciones y es superado por 60% de estas (tener un valor correspondiente a un percentil elevado no es necesariamente bueno; por ejemplo, usted no querría estar en el percentil 99 por el contenido de alcohol en la sangre).

DEFINICIÓN

Sea p un número entre 0 y 1. El **$(100p)^\circ$ percentil** de la distribución de una variable aleatoria continua X, denotada por $\eta(p)$, se define como

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)\,dy \qquad \textbf{(4.2)}$$

De acuerdo con la expresión (4.2), $\eta(p)$ es ese valor sobre el eje de medición, de tal suerte que $100p\%$ del área bajo la gráfica de $f(x)$ queda a la izquierda de $\eta(p)$ y $100(1 - p)\%$ queda a la derecha. Por tanto, $\eta(0.75)$, el 75º percentil, es tal que el área bajo la gráfica de $f(x)$ a la izquierda de $\eta(0.75)$ es 0.75. La figura 4.10 ilustra la definición.

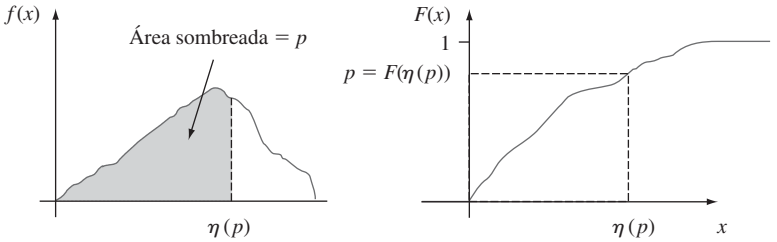


Figura 4.10 El $(100p)^\circ$ percentil de una distribución continua

EJEMPLO 4.9 La distribución de la cantidad de grava (en toneladas) vendida por una compañía de materiales para la construcción particular en una semana dada es una variable aleatoria continua X con función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{de lo contrario} \end{cases}$$

La función de distribución acumulada de las ventas para cualquier x entre 0 y 1 es

$$F(x) = \int_0^x \frac{3}{2}(1 - y^2)\,dy = \frac{3}{2}\left(y - \frac{y^3}{3}\right)\Big|_{y=0}^{y=x} = \frac{3}{2}\left(x - \frac{x^3}{3}\right)$$

Las gráficas de $f(x)$ y de $F(x)$ aparecen en la figura 4.11. El $(100p)^\circ$ percentil de esta distribución satisface la ecuación

$$p = F(\eta(p)) = \frac{3}{2}\left[\eta(p) - \frac{(\eta(p))^3}{3}\right]$$

es decir,

$$(\eta(p))^3 - 3\eta(p) + 2p = 0$$

Para el 50º percentil, $p = 0.5$ y la ecuación que se tiene que resolver es $\eta^3 - 3\eta + 1 = 0$; la solución es $\eta = \eta(0.5) = 0.347$. Si la distribución no cambia de una semana a otra, a la larga 50% de todas las semanas se realizarán ventas de menos de 0.347 ton y 50% de más de 0.347 ton.

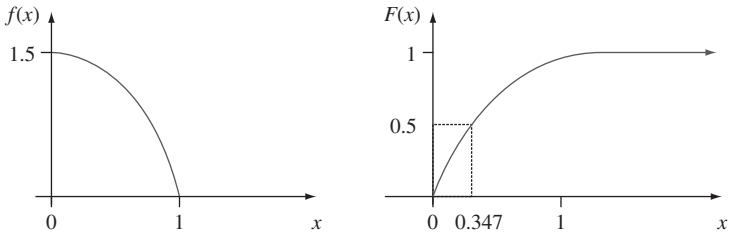


Figura 4.11 Función de densidad de probabilidad y función de distribución acumulativa del ejemplo 4.9 ■

DEFINICIÓN

La **mediana** de una distribución continua, denotada por $\tilde{\mu}$, es el 50º percentil, así que $\tilde{\mu}$ satisface $0.5 = F(\tilde{\mu})$. Es decir, la mitad del área bajo la curva de densidad se encuentra a la izquierda de $\tilde{\mu}$ y la mitad a la derecha de $\tilde{\mu}$.

Una distribución continua cuya función de densidad de probabilidad es **simétrica** —es decir, la gráfica de la función de densidad de probabilidad a la izquierda de algún punto es una imagen en espejo de la gráfica a la derecha de dicho punto—, tiene una mediana $\tilde{\mu}$ igual al punto de simetría, puesto que la mitad del área bajo la curva queda a uno u otro lado de este punto. La figura 4.12 da varios ejemplos. A menudo se supone que el error en la medición de una cantidad física tiene una distribución simétrica.

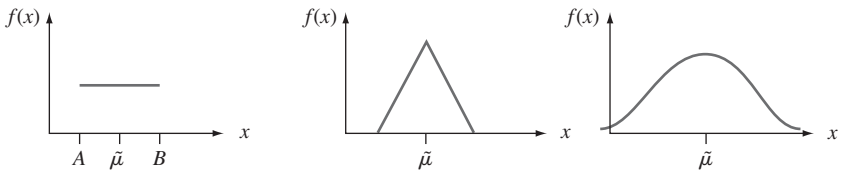


Figura 4.12 Medianas de distribuciones simétricas

Valores esperados

Para una variable aleatoria discreta X, $E(X)$ se obtuvo sumando $x \cdot p(x)$ a lo largo de posibles valores de X. Aquí se reemplaza la suma por la integración y la función de masa de probabilidad por la función de densidad de probabilidad para obtener un promedio ponderado continuo.

DEFINICIÓN

El **valor esperado** o **valor medio** de una variable aleatoria continua X con función de densidad de probabilidad $f(x)$ es

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

EJEMPLO 4.10 (Continuación del ejemplo 4.9)

La función de densidad de probabilidad de las ventas semanales de grava X fue

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2) & 0 \le x \le 1 \\ 0 & \text{de lo contrario} \end{cases}$$

por tanto,

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_{0}^{1} x \cdot \frac{3}{2}(1 - x^2)\,dx$$

$$= \frac{3}{2}\int_{0}^{1} (x - x^3)\,dx = \frac{3}{2}\left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Big|_{x=0}^{x=1} = \frac{3}{8}$$ ■

Cuando la función de densidad de probabilidad $f(x)$ especifica un modelo para la distribución de valores en una población numérica, μ es la media de la población que es la medida de ubicación o centralización de la población que se utiliza con más frecuencia.

Con frecuencia se desea calcular el valor esperado de alguna función $h(X)$ de la variable aleatoria X. Si se piensa en $h(X)$ como una nueva variable aleatoria Y, se utilizan técnicas de estadística matemática para obtener la función de densidad de probabilidad de Y, y $E(Y)$ se calcula a partir de la definición. Afortunadamente, como en el caso discreto, existe una forma más fácil de calcular $E[h(X)]$.

PROPOSICIÓN

Si X es una variable aleatoria continua con función de densidad de probabilidad $f(x)$ y $h(X)$ es cualquier función de X, entonces

$$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx$$

Es decir, así como $E(X)$ es un promedio ponderado de posibles X valores, donde la función de ponderación es la función de densidad de probabilidad $f(x)$, $E[h(X)]$ es un promedio ponderado de los valores h(X).

EJEMPLO 4.11 Dos especies compiten en una región por el control de una cantidad limitada de cierto recurso. Sea X la proporción del recurso controlado por la especie 1 y suponga que la función de densidad de probabilidad de X es

$$f(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \text{de lo contrario} \end{cases}$$

la cual es una distribución uniforme en [0, 1]. (En su libro ***Ecological Diversity***, E. C. Pielou llama a esto el modelo del "palo roto" para la asignación de recursos, puesto que es análogo a la ruptura de un palo en un lugar seleccionado al azar.) Entonces la especie que controla la mayor parte de este recurso controla la cantidad

$$h(X) = \text{máx}\,(X, 1 - X) = \begin{cases} 1 - X & \text{si } 0 \le X < \dfrac{1}{2} \\ X & \text{si } \dfrac{1}{2} \le X \le 1 \end{cases}$$

La cantidad que se espera que controle la especie que tiene el control mayoritario es entonces

$$E[h(X)] = \int_{-\infty}^{\infty} \text{máx}(x, 1 - x) \cdot f(x)\,dx = \int_{0}^{1} \text{máx}(x, 1 - x) \cdot 1\,dx$$

$$= \int_{0}^{1/2} (1 - x) \cdot 1\,dx + \int_{1/2}^{1} x \cdot 1\,dx = \frac{3}{4}$$ ■

En el caso discreto la varianza de X se definió como la desviación al cuadrado esperada respecto a μ, y se calculó sumando. De nuevo, en este caso, la integración reemplaza a la suma.

DEFINICIÓN

La **varianza** de una variable aleatoria continua X con función de densidad de probabilidad $f(x)$ y valor medio μ es

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx = E[(X - \mu)^2]$$

La **desviación estándar** (DE) de X es $\sigma_X = \sqrt{V(X)}$.

La varianza y la desviación estándar dan medidas cuantitativas de qué tanta dispersión hay en la distribución o población de valores x. Una vez más σ es aproximadamente del tamaño de una desviación típica de μ. El cálculo de σ^2 se facilita mediante el uso de la fórmula abreviada similar a la que se utiliza en el caso discreto.

PROPOSICIÓN

$$V(X) = E(X^2) - [E(X)]^2$$

EJEMPLO 4.12 (Continuación del ejemplo 4.10)

Para $X =$ ventas semanales de grava se calculó $E(X) = \frac{3}{8}$. Puesto que

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx = \int_0^1 x^2 \cdot \frac{3}{2}(1 - x^2)\,dx$$

$$= \int_0^1 \frac{3}{2}(x^2 - x^4)\,dx = \frac{1}{5}$$

$$V(X) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = \frac{19}{320} = 0.059 \quad \text{y} \quad \sigma_X = 0.244$$ ■

Cuando $h(X) = aX + b$, el valor esperado y la varianza de $h(X)$ satisfacen las mismas propiedades que en el caso discreto: $E[h(X)] = a\mu + b$ y $V[h(X)] = a^2 \cdot \sigma^2$.

EJERCICIOS Sección 4.2 (11–27)

11. Sea X la cantidad de tiempo que un libro en reserva de dos horas permanece realmente prestado y supongamos que la función de distribución acumulada es

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{4} & 0 \le x < 2 \\ 1 & 2 \le x \end{cases}$$

a. Calcule $P(X \le 1)$.
b. Calcule $P(0.5 \le X \le 1)$.
c. Calcule $P(X > 1.5)$.
d. ¿Cuál es la mediana del tiempo de préstamo $\tilde{\mu}$? [resuelva $0.5 = F(\tilde{\mu})$].
e. Obtenga la función de densidad $f(x)$.
f. Calcule $E(X)$
g. Calcule $V(X)$ y σ_X.
h. Si al prestatario se le cobra una cantidad $h(X) = X^2$ cuando el tiempo de préstamo es X, calcule el cobro esperado $E[h(X)]$.

12. La función de distribución acumulada de X (= error de medición) del ejercicio 3 es

$$F(x) = \begin{cases} 0 & x < -2 \\ \frac{1}{2} + \frac{3}{32}\left(4x - \frac{x^3}{3}\right) & -2 \le x < 2 \\ 1 & 2 \le x \end{cases}$$

a. Calcule $P(X < 0)$.
b. Calcule $P(-1 < X < 1)$.
c. Calcule $P(0.5 < X)$.

d. Verifique que $f(x)$ sea la dada en el ejercicio 3 al obtener $F'(x)$.

e. Verifique que $\tilde{\mu} = 0$.

13. El ejemplo 4.5 introdujo el concepto de intervalo de tiempo en el flujo de tránsito y propuso una distribución particular para $X =$ el intervalo de tiempo entre dos vehículos consecutivos seleccionados al azar (s). Suponga que en un entorno de tránsito diferente la distribución del intervalo de tiempo tiene la forma

$$f(x) = \begin{cases} \dfrac{k}{x^4} & x > 1 \\ 0 & x \le 1 \end{cases}$$

a. Determine el valor de k con el cual $f(x)$ es una función de densidad de probabilidad legítima.

b. Obtenga la función de distribución acumulada.

c. Use la función de distribución acumulada de (b) para determinar la probabilidad de que el intervalo de tiempo exceda 2 segundos y también la probabilidad de que el intervalo sea de entre 2 y 3 segundos.

d. Obtenga un valor medio del intervalo de tiempo y su desviación estándar.

e. ¿Cuál es la probabilidad de que el intervalo de tiempo quede dentro de 1 desviación estándar de la media?

14. El artículo **"Modeling Sediment and Water Column Interactions for Hidrophobic Pollutants" (*Water Research*, 1984: 1169–1174)** sugiere la distribución uniforme en el intervalo (7.5, 20) como modelo de profundidad (cm) de la capa de bioturbación en el sedimento en una cierta región.

a. ¿Cuáles son la media y la varianza de la profundidad?

b. ¿Cuál es la función de distribución acumulada de la profundidad?

c. ¿Cuál es la probabilidad de que la profundidad observada sea cuando mucho de 10? ¿Y entre 10 y 15?

d. ¿Cuál es la probabilidad de que la profundidad observada esté dentro de 1 desviación estándar del valor medio? ¿Y dentro de 2 desviaciones estándar?

15. Sea X la cantidad de espacio ocupado por un artículo colocado en un contenedor de 1 pie^3. La función de densidad de probabilidad de X es

$$f(x) = \begin{cases} 90x^8(1-x) & 0 < x < 1 \\ 0 & \text{de lo contrario} \end{cases}$$

a. Grafique la función de densidad de probabilidad. Luego obtenga la función de distribución acumulada de X y grafíquela.

b. ¿Cuál es $P(X \le 0.5)$ [es decir, $F(0.5)$]?

c. Con la función de distribución acumulada de a), ¿cuál es $P(.25 < X \le 0.5)$? ¿Cuál es $P(0.25 \le X \le 0.5)$?

d. ¿Cuál es el percentil 75 de la distribución?

e. Calcule $E(X)$ y σ_X.

f. ¿Cuál es la probabilidad de que X esté a más de 1 desviación estándar de su media?

16. El artículo **"A Model of Pedestrians' Waiting Times for Street Crossings at Signalized Intersections" (*Transportation Research*, 2013: 17–28)** sugirió que, bajo ciertas circunstancias, la distribución del tiempo de espera X podría ser modelada con la siguiente función de densidad de probabilidad:

$$f(x;\theta,\tau) = \begin{cases} \dfrac{\theta}{\tau}(1 - x/\tau)^{\theta-1} & 0 \le x < \tau \\ 0 & \text{de lo contrario} \end{cases}$$

a. Trace la gráfica de $f(x; \theta, 80)$ para los tres casos $\theta = 4$ y 1.5 (estas gráficas aparecen en el artículo citado) y comente sus formas.

b. Obtenga la función de distribución acumulada de X.

c. Obtenga una expresión para la mediana de la distribución del tiempo de espera.

d. Para el caso $\theta = 4$, $\tau = 80$, calcule $P(50 \le X \le 70)$ sin hacer por el momento ninguna integración adicional.

17. Si la distribución de X en el intervalo $[A, B]$ es uniforme.

a. Obtenga una expresión para el $(100p)^{\text{o}}$ percentil.

b. Calcule $E(X)$, $V(X)$ y σ_X.

c. Con n, un entero positivo, calcule $E(X^n)$.

18. Sea X el voltaje de salida de un micrófono y suponga que X tiene una distribución uniforme en el intervalo de -1 a 1. El voltaje es procesado por un "limitador duro" con valores de corte de -0.5 y 0.5, de modo que la salida del limitador es una variable aleatoria Y relacionada con X mediante $Y = X$ si $|X| \le 0.5$, $Y = 0.5$ si $X > 0.5$ y $Y = -0.5$ si $X < -0.5$.

a. ¿Cuál es $P(Y = 0.5)$?

b. Obtenga la función de distribución acumulada de Y y grafíquela.

19. Sea X una variable aleatoria continua con función de distribución acumulada

$$F(x) = \begin{cases} 0 & x \le 0 \\ \dfrac{x}{4}\left[1 + \ln\left(\dfrac{4}{x}\right)\right] & 0 < x \le 4 \\ 1 & x > 4 \end{cases}$$

[Este tipo de función de distribución acumulada es sugerido en el artículo **"Variability in Measured Bedload-Transport Rates" (*Water Resources Bull.*, 1985: 39–48)** como modelo de cierta variable hidrológica.]

a. ¿Cuál es $P(X \le 1)$?

b. ¿Cuál es $P(1 \le X \le 3)$?

c. ¿Cuál es la función de densidad de probabilidad de X?

20. Considere la función de densidad de probabilidad del tiempo de espera total Y de dos autobuses

$$f(y) = \begin{cases} \dfrac{1}{25}y & 0 \le y < 5 \\ \dfrac{2}{5} - \dfrac{1}{25}y & 5 \le y \le 10 \\ 0 & \text{de lo contrario} \end{cases}$$

introducida en el ejercicio 8.

a. Calcule y trace la función de distribución acumulada de Y. [*Sugerencia:* Considere por separado $0 \le y \le 5$ y $5 \le y \le 10$ al calcular $F(y)$. Sería útil una gráfica de la función de densidad.]

b. Obtenga una expresión para el $(100p)^{\text{o}}$ percentil. [*Sugerencia:* Considere por separado $0 < p < 0.5$ y $0.5 < p < 1$.]

c. Calcule $E(Y)$ y $V(Y)$. ¿Cómo se comparan estos valores con el tiempo de espera probable y la varianza de un solo camión cuando el tiempo está uniformemente distribuido en [0, 5]?

21. Un ecólogo desea marcar una región de muestreo circular de 10 m de radio. Sin embargo, el radio de la región resultante en realidad es una variable aleatoria R con función de densidad de probabilidad

$$f(r) = \begin{cases} \frac{3}{4}[1 - (10 - r)^2] & 9 \le r \le 11 \\ 0 & \text{de lo contrario} \end{cases}$$

¿Cuál es el área esperada de la región circular resultante?

22. La demanda semanal de gas propano (en miles de galones) de una instalación particular es una variable aleatoria X con función de densidad con probabilidad

$$f(x) = \begin{cases} 2\left(1 - \frac{1}{x^2}\right) & 1 \le x \le 2 \\ 0 & \text{de lo contrario} \end{cases}$$

a. Calcule la función de distribución acumulada de X.

b. Obtenga una expresión para el $(100p)^{\text{o}}$ percentil. ¿Cuál es el valor de $\tilde{\mu}$?

c. Calcule $E(X)$ y $V(X)$.

d. Si al principio de la semana hay 1500 galones en existencia y no se espera ningún nuevo suministro durante la semana, ¿cuántos de los 1500 galones se espera que queden al final de la semana? [*Sugerencia:* sea $h(x)$ = la cantidad que queda cuando la demanda = x.]

23. Si la temperatura a la cual cierto compuesto se funde es una variable aleatoria con valor medio de 120°C y desviación estándar de 2°C, ¿cuáles son la temperatura media y la desviación estándar medidas en °F? [*Sugerencia:* °F = 1.8°C + 32.]

24. La función de densidad de probabilidad de Pareto de X es

$$f(x; k, \theta) = \begin{cases} \frac{k \cdot \theta^k}{x^{k+1}} & x \ge \theta \\ 0 & x < \theta \end{cases}$$

introducida en el ejercicio 10.

a. Si $k > 1$, calcule $E(X)$.

b. ¿Qué se puede decir sobre $E(X)$ si $k = 1$?

c. Si $k > 2$, demuestre que $V(X) = k\theta^2(k - 1)^{-2}(k - 2)^{-1}$.

d. Si $k = 2$, ¿qué se puede decir sobre $V(X)$?

e. ¿Qué condiciones en cuanto a k son necesarias para garantizar que $E(X^n)$ sea finito?

25. Sea X la temperatura en °C a la cual ocurre una reacción química y sea Y la temperatura en °F (por lo que $Y = 1.8X + 32$).

a. Si la mediana de la distribución X es $\tilde{\mu}$, demuestre que $1.8\tilde{\mu} + 32$ es la mediana de la distribución Y.

b. ¿Cómo está relacionado el percentil 90 de la distribución Y con el percentil 90 de la distribución X? Verifique su conjetura.

c. Más generalmente, si $Y = aX + b$, ¿cómo está relacionado cualquier percentil de la distribución Y con el percentil correspondiente de la distribución X?

26. Sean X los gastos médicos totales (en miles de dólares) en que incurre un individuo particular durante un cierto año. Aunque X es una variable aleatoria discreta, suponga que su distribución se aproxima bastante bien mediante una distribución continua con función de densidad de probabilidad $f(x) = k(1 - x/2.5)^{-7}$ para $x \ge 0$.

a. ¿Cuál es el valor de k?

b. Grafique la función de densidad de probabilidad de X.

c. ¿Cuáles son el valor esperado y la desviación estándar de los gastos médicos totales?

d. Este individuo está cubierto por un plan de aseguramiento que le impone una provisión deducible de \$500 (así que los primeros \$500 de gastos son pagados por el individuo). El plan pagará 80% de cualquier gasto adicional que exceda de \$500 y el pago máximo por parte del individuo (incluida la cantidad deducible) es de \$2500. Sea Y la cantidad de gastos médicos de este individuo que paga la compañía de seguros. ¿Cuál es el valor esperado de Y?

[*Sugerencia:* primero indague qué valor de X corresponde al gasto máximo que sale del bolsillo de \$2500. Luego escriba una expresión para Y como una función de X (la cual implique varios precios diferentes) y calcule el valor esperado de la función.]

27. Cuando se lanza un dardo a un blanco circular considere la ubicación del punto de aterrizaje respecto al centro del blanco. Sea X el ángulo en grados medido respecto a la horizontal y suponga que X está uniformemente distribuida en [0, 360]. Defina Y como la variable transformada $Y = h(X) = (2\pi/360)X - \pi$; por tanto, Y es el ángulo medido en radianes y Y está entre $-\pi$ y π. Obtenga $E(Y)$ y σ_Y obteniendo primero $E(X)$ y σ_X y luego utilizando el hecho de que $h(X)$ es una función lineal de X.

4.3 Distribución normal

La **distribución normal** es la más importante tanto en la probabilidad y en la estadística. Muchas poblaciones numéricas tienen distribuciones que pueden ser representadas muy fielmente mediante una curva normal apropiada. Los ejemplos incluyen estaturas, pesos y otras características físicas (el famoso artículo **1903 *Biometrika* "On the Laws of Inheritance in Man"** discutió muchos ejemplos de esta clase), errores de medición en experimentos científicos, mediciones antropométricas en fósiles, tiempos de reacción

en experimentos psicológicos, mediciones de inteligencia y aptitud, calificaciones en varios exámenes y numerosas medidas e indicadores económicos. Además, aun cuando las variables individuales no estén normalmente distribuidas, las sumas y los promedios de las variables en condiciones adecuadas tendrán de manera aproximada una distribución normal; este es el contenido del teorema del límite central que se aborda en el siguiente capítulo.

DEFINICIÓN

Se dice que una variable aleatoria continua X tiene una **distribución normal** con parámetros μ y σ (o μ y σ^2), donde $-\infty < \mu < \infty$ y $0 < \sigma$, si la función de densidad de probabilidad de X es

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty \qquad \textbf{(4.3)}$$

De nuevo e denota la base del sistema de logaritmos naturales y es aproximadamente igual a 2.71828 y π representa la conocida constante matemática con un valor aproximado de 3.14159. El enunciado de que X está normalmente distribuida con los parámetros μ y σ^2 a menudo se abrevia como $X \sim N(\mu,\sigma^2)$.

Claramente $f(x; \mu, \sigma) \geq 0$ aunque se tiene que utilizar un argumento de cálculo un tanto complicado para verificar que $\int_{-\infty}^{\infty} f(x; \mu, \sigma)\,dx = 1$. Se puede demostrar que $E(X) = \mu$ y $V(X) = \sigma^2$, de modo que los parámetros son la media y la desviación estándar de X. La figura 4.13 representa gráficas de $f(x; \mu, \sigma)$ de varios pares diferentes (μ, σ). Cada curva de densidad es simétrica respecto a μ y acampanada, de modo que el centro de la campana (punto de simetría) es tanto la media de la distribución como la mediana. La media μ es un *parámetro de ubicación*, ya que al cambiar su valor de forma rígida desplaza la curva de densidad hacia uno u otro lado; σ se conoce como un *parámetro de escala* porque al cambiar su valor estira o comprime la curva horizontal sin cambiar la forma básica. Los puntos de inflexión de una curva normal (puntos donde la curva cambia de girar hacia abajo a girar hacia arriba) se producen en $\mu - \sigma$ y $\mu + \sigma$. Así el valor de σ se puede visualizar como la distancia de la media de estos puntos de inflexión. Un valor grande de σ corresponde a una curva de densidad que se extiende bastante sobre μ, mientras que un valor pequeño produce una curva altamente concentrada. Entre más grande sea el valor de σ, es más probable que se pueda observar un valor de X lejos de la media.

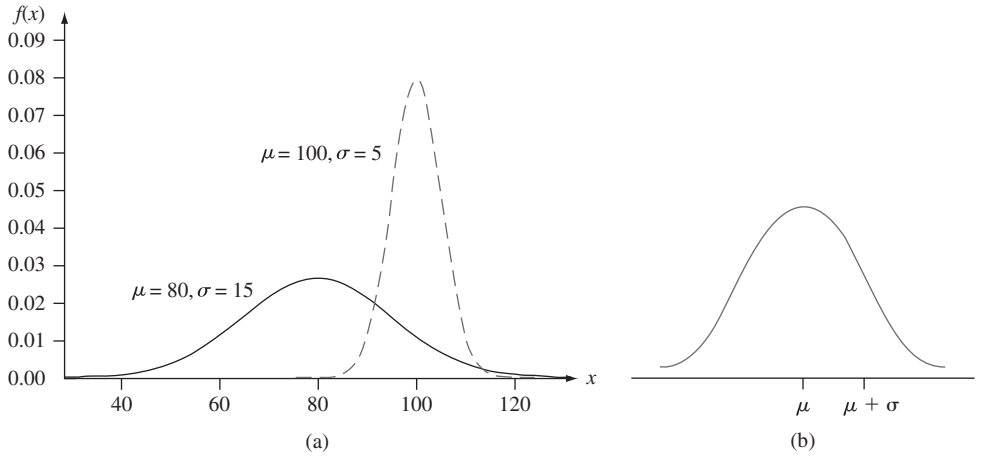


Figura 4.13 (a) Dos curvas diferentes de densidad normal (b) Visualización de μ y σ para una distribución normal

Distribución normal estándar

El cálculo de $P(a \leq X \leq b)$ cuando X es una variable aleatoria normal con parámetros μ y σ, requiere determinar

$$\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx \tag{4.4}$$

Ninguna de las técnicas estándar de integración puede ser utilizada para lograr esto. En cambio, con $\mu = 0$ y $\sigma = 1$ se calculó la expresión (4.4) mediante técnicas numéricas y se tabuló para ciertos valores de a y b. Esta tabla también puede ser utilizada para calcular probabilidades con cualesquiera otros valores de μ y σ considerados.

DEFINICIÓN

La distribución normal con valores de parámetro $\mu = 0$ y $\sigma = 1$ se llama **distribución normal estándar**. Una variable aleatoria que tiene una distribución normal estándar se llama **variable aleatoria normal estándar** y se denotará por Z. La función de densidad de probabilidad de Z es

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty$$

La gráfica de $f(z; 0, 1)$ se llama curva *normal estándar* (o z). Sus puntos de inflexión están en 1 y -1. La función de distribución acumulada de Z es $P(Z \leq z) = \int_{-\infty}^{z} f(y; 0, 1)\, dy$ la cual será denotada por $\Phi(z)$.

La distribución normal estándar no siempre sirve como modelo de una población que surge naturalmente. En cambio, es una distribución de referencia de la que se obtiene información sobre otra distribución normal. La tabla A.3 del apéndice, da $\Phi(z) = P(Z \leq z)$, el área bajo la curva de densidad normal estándar a la izquierda de z con $z = -3.49$, -3.48,…, 3.48, 3.49. La figura 4.14 ilustra el tipo de área acumulada (probabilidad) tabulada en la tabla A.3. Con esta tabla pueden ser calculadas varias otras probabilidades que implican Z.

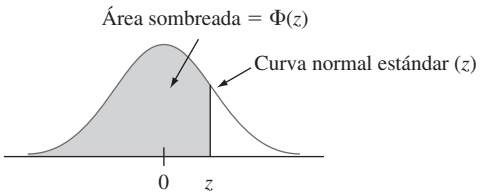


Figura 4.14 Áreas acumuladas normales estándar tabuladas en la tabla A.3 del apéndice

EJEMPLO 4.13 Determine las siguientes probabilidades normales estándar: a) $P(Z \leq 1.25)$, b) $P(Z > 1.25)$, c) $P(Z \leq -1.25)$, d) $P(-0.38 \leq Z \leq 1.25)$ y e) $P(Z \leq 5)$.

a. $P(Z \leq 1.25) = \Phi(1.25)$, una probabilidad tabulada en la tabla A.3 del apéndice en la intersección de la fila 1.2 y la columna 0.05. El número allí es 0.8944, por lo que $P(Z \leq 1.25) = 0.8944$. La figura 4.15(a) ilustra esta probabilidad.

b. $P(Z > 1.25) = 1 - P(Z \leq 1.25) = 1 - \Phi(1.25)$ el área bajo la curva z a la derecha de 1.25 (un área de cola superior). En ese caso $\Phi(1.25) = 0.8944$ implica que $P(Z > 1.25) = 0.1056$. Puesto que Z es una variable aleatoria continua, $P(Z \geq 1.25) = 0.1056$. Véase la figura 4.15(b).

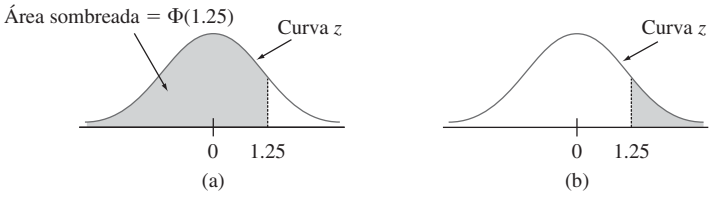


Figura 4.15 Áreas (probabilidades) de curvas normales del ejemplo 4.13

c. $P(Z \leq -1.25) = \Phi(-1.25)$, un área de cola inferior. Directamente de la tabla A.3 del apéndice, $\Phi(-1.25) = 0.1056$. Por simetría de la curva z, esta es la misma respuesta del inciso b).

d. $P(-0.38 \leq Z \leq 1.25)$ es el área bajo la curva normal estándar sobre el intervalo cuyo punto extremo izquierdo es -0.38 y cuyo punto extremo derecho es 1.25. Según la sección 4.2, si X es una variable aleatoria continua con función de distribución acumulada $F(x)$, entonces $P(a \leq X \leq b) = F(b) - F(a)$. Por tanto, $P(-0.38 \leq Z \leq 1.25) = \Phi(1.25) - \Phi(-0.38) = 0.8944 - 0.3520 = 0.5424$. (Véase la figura 4.16.)

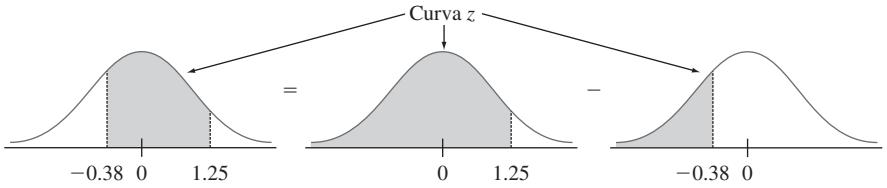


Figura 4.16 $P(-0.38 \leq Z \leq 1.25)$ como la diferencia entre dos áreas acumuladas

e. $P(Z \leq 5) = \Phi(5)$, el área acumulada bajo la curva z a la izquierda del 5. Esta probabilidad no aparece en la tabla ya que la última fila está etiquetada como 3.4. Sin embargo, la última entrada en esta fila es $\Phi(3.49) = 0.9998$. Es decir, esencialmente toda el área bajo la curva se encuentra a la izquierda de 3.49 (a lo más 3.49 desviaciones estándar a la derecha de la media). Por tanto, concluimos que $P(Z \leq 5) \approx 1$. ■

Percentiles de la distribución normal estándar

Con cualquier p entre 0 y 1 se puede utilizar la tabla A.3 del apéndice para obtener el $(100p)^{\text{o}}$ percentil de la distribución normal estándar.

El percentil 99 de la distribución normal estándar es ese valor sobre el eje horizontal tal que el área bajo la curva z a la izquierda de dicho valor es 0.9900. La tabla A.3 del apéndice da para z fija el área bajo la curva normal estándar a la izquierda de z, mientras que aquí se tiene el área y se desea el valor de z. Este es el problema "inverso" a $P(Z \leq z) = ?$ por lo que la tabla se utiliza a la inversa: encuentre en la mitad de la tabla 0.9900; la fila y la columna en la que se encuentra identifica el 99º percentil z. En este caso 0.9901 queda en la intersección de la fila 2.3 y la columna 0.03, así que el percentil 99 es (aproximadamente) $z = 2.33$. (Véase la figura 4.17.) Por simetría, el primer percentil está tan debajo de 0, como el 99º está sobre 0, así que es igual a -2.33 (1% queda debajo del primero y también sobre el 99º). (Véase la figura 4.18.)

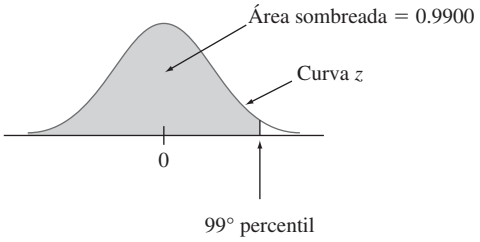


Figura 4.17 Ubicación del 99º percentil

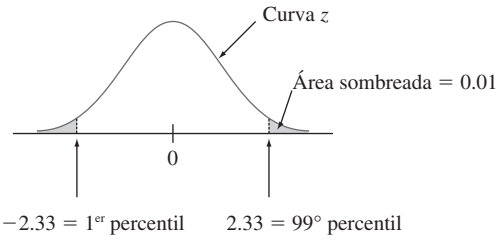


Figura 4.18 Relación entre los percentiles 1 y 99

En general, la fila y la columna de la tabla A.3 del apéndice, donde la entrada p está localizada, identifican el $(100p)$º percentil (p. ej., el percentil 67 se obtiene localizando 0.6700 en el cuerpo de la tabla, la cual da $z = 0.44$). Si p no aparece a menudo se utiliza el número más cercano a él, aunque la interpolación lineal da una respuesta más precisa. Por ejemplo, para encontrar el percentil 95 se busca 0.9500 adentro de la tabla. Aunque 0.9500 no aparece tanto 0.9495 como 0.9505 sí aparecen, correspondiendo a $z = 1.64$ y 1.65, respectivamente. Puesto que 0.9500 está a la mitad entre las dos probabilidades que sí aparecen, se utilizará 1.645 como el percentil 95 y -1.645 como el 5º percentil.

Notación z_α para valores z críticos

En inferencia estadística se necesitan valores sobre el eje horizontal z que capturen ciertas áreas de cola pequeña bajo la curva normal estándar.

Notación

z_α denotará el valor sobre el eje z para el cual α del área bajo la curva z queda a la derecha de z_α. (Véase la figura 4.19.)

Por ejemplo, $z_{0.10}$ captura el área de cola superior 0.10, y $z_{0.01}$ captura el área de cola superior

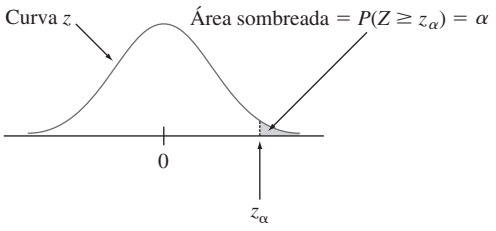


Figura 4.19 Notación z_α ilustrada

Puesto que α del área bajo la curva z queda a la derecha de z_α, $1 - \alpha$ del área queda a su izquierda. Por tanto, z_α *es el* $100(1 - \alpha)$º *percentil de la distribución normal estándar*. Por simetría el área bajo la curva normal estándar a la izquierda de $-z_\alpha$ también es α. Los valores z_α en general se conocen como **valores críticos** z. La tabla 4.1 incluye los percentiles z y los valores z_α más útiles.

Tabla 4.1 Percentiles de la distribución normal estándar y valores críticos

Percentil	90	95	97.5	99	99.5	99.9	99.95
α (área de la cola)	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
$z_\alpha = 100(1-\alpha)°$ percentil	1.28	1.645	1.96	2.33	2.58	3.08	3.27

EJEMPLO 4.15 $z_{0.05}$ es el $100(1 - 0.05)° = 95°$ percentil de la distribución normal estándar, por tanto $z_{0.05} = 1.645$. El área bajo la curva normal estándar a la izquierda de $-z_{0.05}$ también es 0.05. (Véase la figura 4.20.)

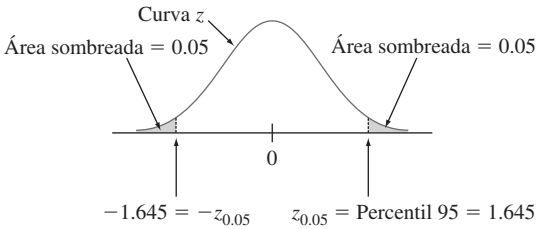


Figura 4.20 Determinación de $z_{0.05}$ ■

Distribuciones normales no estándar

Cuando $X \sim N(\mu, \sigma^2)$, las probabilidades que implican X se calculan "estandarizando". La **variable estandarizada** es $(X - \mu)/\sigma$. Al restar la media μ cambia de μ a cero y luego, al dividir entre σ cambian las escalas de la variable de modo que la desviación estándar es 1 en lugar de σ.

PROPOSICIÓN

Si X tiene una distribución normal con media μ y desviación estándar σ, entonces

$$Z = \frac{X - \mu}{\sigma}$$

tiene una distribución normal estándar. Por tanto

$$P(a \le X \le b) = P\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$P(X \le a) = \Phi\left(\frac{a-\mu}{\sigma}\right) \qquad P(X \ge b) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$$

De acuerdo con la primera parte de la proposición, el área debajo de la curva normal (μ, σ^2) que se encuentra en el intervalo $[a, b]$ es idéntica al área bajo la curva normal estándar que se encuentra en el intervalo desde el límite inferior estandarizado $(a - \mu)/\sigma$ para el límite superior estandarizado $(b - \mu)/\sigma$. En la figura 4.21 se ilustra la segunda parte. La idea clave de la proposición es que, estandarizando, cualquier probabilidad que implique X puede ser expresada como una probabilidad que implica una variable aleatoria normal estándar Z, de modo que se pueda utilizar la tabla A.3 del apéndice. La proposición se comprueba escribiendo la función de distribución acumulativa $Z = (X - \mu)/\sigma$ como

$$P(Z \le z) = P(X \le \sigma z + \mu) = \int_{-\infty}^{\sigma z + \mu} f(x; \mu, \sigma)\, dx$$

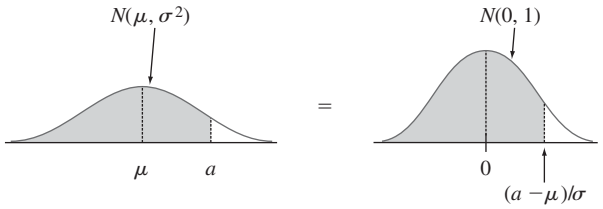


Figura 4.21 Igualdad de áreas de curvas normales estándar y no estándar

Al utilizar un resultado del cálculo, esta integral puede ser diferenciada respecto a z para que dé la función de densidad de probabilidad deseada $f(z; 0, 1)$.

EJEMPLO 4.16 El tiempo que requiere un conductor para reaccionar a las luces de freno de un vehículo que está desacelerando es crítico para evitar colisiones por alcance. El artículo **"Fast-Rise Brake Lamp as a Collision-Prevention Device" (*Ergonomics*, 1993: 391–395)** sugiere que el tiempo de reacción de respuesta en tráfico a una señal de luces de freno estándar puede ser modelado con una distribución normal que tiene un valor medio de 1.25 s y desviación estándar de 0.46 s. ¿Cuál es la probabilidad de que el tiempo de reacción sea de entre 1.00 y 1.75 segundos? Si X denota el tiempo de reacción, estandarizando se obtiene

$$1.00 \le X \le 1.75$$

si y sólo si

$$\frac{1.00 - 1.25}{0.46} \le \frac{X - 1.25}{0.46} \le \frac{1.75 - 1.25}{0.46}$$

Por tanto

$$\begin{aligned} P(1.00 \le X \le 1.75) &= P\left(\frac{1.00 - 1.25}{0.46} \le Z \le \frac{1.75 - 1.25}{0.46}\right) \\ &= P(-0.54 \le Z \le 1.09) = \Phi(1.09) - \Phi(-0.54) \\ &= 0.8621 - \ 0.2946 = 0.5675 \end{aligned}$$

Esto se ilustra en la figura 4.22. Asimismo, si se ven los 2 segundos como un tiempo de reacción críticamente largo, la probabilidad de que el tiempo de reacción real exceda este valor es

$$P(X > 2) = P\left(Z > \frac{2 - 1.25}{0.46}\right) = P(Z > 1.63) = 1 - \Phi(1.63) = 0.0516$$

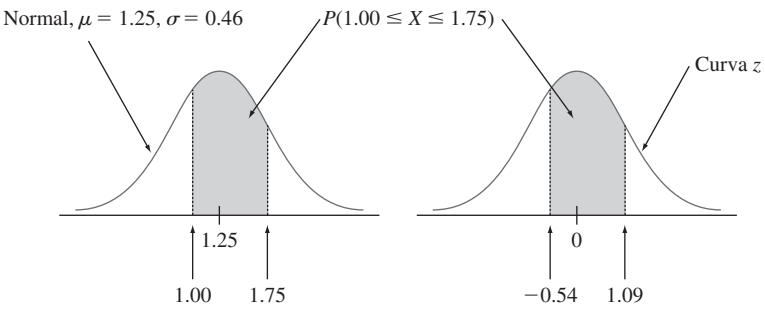


Figura 4.22 Curvas normales del ejemplo 4.16 ■

Estandarizar cantidades no lleva más que a calcular una distancia media y luego reexpresarla como algún número de desviaciones estándar. Por tanto, si $\mu = 100$ y $\sigma = 15$, entonces $x = 130$ corresponde a $z = (130 - 100)/15 = 30/15 = 2.00$. Es decir, 130 está a 2 desviaciones estándar sobre (a la derecha de) la media. Asimismo, estandarizando 85 se obtiene $(85 - 100)/15 = -1.00$, por tanto, 85 está a 1 desviación estándar por debajo de la media. La tabla z se aplica a *cualquier* distribución normal siempre que se piense en función del número de desviaciones estándar respecto al valor medio.

EJEMPLO 4.17 Se sabe que el voltaje de ruptura de un diodo de un tipo particular seleccionado al azar está normalmente distribuido. ¿Cuál es la probabilidad de que el voltaje de ruptura de un diodo esté dentro de 1 desviación estándar de su valor medio? Esta pregunta puede contestarse sin conocer μ o σ, en tanto se sepa que la distribución es normal; la respuesta es la misma para *cualquier* distribución normal:

$$\begin{aligned} P(X \text{ está dentro de 1 desviación estándar de su media}) &= P(\mu - \sigma \le X \le \mu + \sigma) \\ &= P\left(\frac{\mu - \sigma - \mu}{\sigma} \le Z \le \frac{\mu + \sigma - \mu}{\sigma}\right) \\ &= P(-1.00 \le Z \le 1.00) \\ &= \Phi(1.00) - \Phi(-1.00) = 0.6826 \end{aligned}$$

La probabilidad de que X esté dentro de 2 desviaciones estándar de su media es $P(-2.00 \le Z \le 2.00) = 0.9544$ y dentro de 3 desviaciones estándar de su media es $P(-3.00 \le Z \le 3.00) = 0.9974$. ■

Los resultados del ejemplo 4.17 a menudo se reportan en forma de porcentaje y se les conoce como la *regla empírica* (porque la evidencia empírica ha demostrado que los histogramas de datos reales con frecuencia pueden ser aproximados por curvas normales).

> Si la distribución de la población de una variable es (aproximadamente) normal, entonces
>
> **1.** Alrededor de 68% de los valores está dentro de 1 DE de la media.
> **2.** Alrededor de 95% de los valores está dentro de 2 DE de la media.
> **3.** Alrededor de 99.7% de los valores está dentro de 3 DE de la media

En realidad es inusual observar un valor de una población normal que esté mucho más lejos de 2 desviaciones estándar de μ. Estos resultados serán importantes en el desarrollo de procedimientos de prueba de hipótesis en capítulos posteriores.

Percentiles de una distribución normal arbitraria

El $(100p)^{\circ}$ percentil de una distribución normal con media μ y desviación estándar σ es fácil de relacionar con el $(100p)^{\circ}$ percentil de la distribución normal estándar.

PROPOSICIÓN

$$\begin{pmatrix} (100p)^{\circ} \text{ percentil} \\ \text{para } (\mu, \sigma) \text{ normal} \end{pmatrix} = \mu + \left[\begin{pmatrix} (100p)^{\circ} \text{ para} \\ \text{normal estándar} \end{pmatrix}\right] \cdot \sigma$$

Otra forma de explicar esto es que si z es el percentil deseado de la distribución normal estándar, entonces el percentil deseado de la distribución normal (μ, σ) está a z desviaciones estándar de μ.

EJEMPLO 4.18 Los autores de **"Assessment of Lifetime of Railway Axle" (*Intl. J. of Fatigue*, 2013: 40–46)** utilizan los datos recolectados de un experimento con una longitud de grieta inicial específica y un número de ciclos para proponer una distribución normal con valor promedio de carga 5.496 mm y desviación estándar 0.067 mm para la variable aleatoria X = profundidad final de la grieta. Para este modelo, ¿qué valor de profundidad final de la grieta podría ser superado por sólo 0.5% de todas las grietas bajo estas circunstancias? Que c denote el valor requerido. Entonces la condición deseada es que $P(X > c) = 0.005$ o, lo que es lo mismo, que $P(X \leq c) = 0.995$. Por tanto c es el percentil 99.5 de la distribución normal con $\mu = 5.496$ y $\sigma = 0.067$. El percentil 99.5 de la distribución normal estándar es 2.58, por lo que

$$c = \eta(0.995) = 5.496 + (2.58)(0.067) + 5.496 + 0.173 = 5.669 \text{ mm}$$

Esto se ilustra en la figura 4.23.

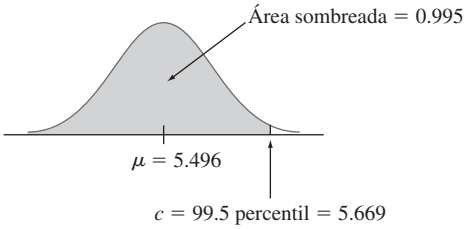


Figura 4.23 Distribución de la profundidad final de la grieta en el ejemplo 4.18 ■

Distribución normal y poblaciones discretas

La distribución normal a menudo se utiliza como una aproximación a la distribución de valores en una población discreta. En semejantes situaciones se debe tener cuidado especial para asegurarse de que las probabilidades se calculen con precisión.

EJEMPLO 4.19 Se sabe que el coeficiente intelectual en una población particular (medido con una prueba estándar) está más o menos normalmente distribuido con $\mu = 100$ y $\sigma = 15$. ¿Cuál es la probabilidad de que un individuo seleccionado al azar tenga un CI de al menos 125? Con X = el CI de una persona seleccionada al azar, se desea $P(X \geq 125)$. La tentación en este caso es estandarizar tal como en ejemplos previos. Sin embargo, la distribución de la población de coeficientes intelectuales en realidad es discreta, puesto que estos son valores enteros. Por tanto, la curva normal es una aproximación a un histograma de probabilidad discreto tal como se ilustra en la figura 4.24.

Los rectángulos del histograma están *centrados* en enteros, por lo que los coeficientes intelectuales de al menos 125 corresponden a rectángulos que comienzan en 124.5, la zona sombreada en la figura 4.24. Por tanto, en realidad se desea a la derecha de 124.5 el área bajo la curva aproximadamente normal. Si se estandariza este valor se obtiene $P(Z \geq 1.63) = 0.0516$, en tanto que si se estandariza 125 se obtiene $P(Z \geq 1.67) = 0.0475$. La diferencia no es grande, pero la respuesta 0.0516 es más precisa. Asimismo, $P(X = 125)$ sería aproximada por el área entre 124.5 y 125.5, puesto que el área bajo la curva normal sobre el valor único de 125 es cero.

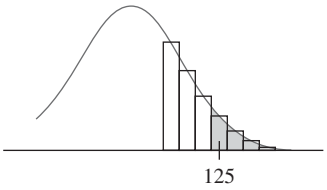


Figura 4.24 Aproximación normal a una distribución discreta ■

La corrección en cuanto a discrecionalidad de la distribución subyacente en el ejemplo 4.19 con frecuencia se llama **corrección de continuidad**. Esta es útil en la siguiente aplicación de la distribución normal al cálculo de probabilidades binomiales.

Aproximación de la distribución binomial

Recuerde que la media y la desviación estándar de una variable aleatoria binomial X son $\mu_X = np$ y $\sigma_X = \sqrt{npq}$, respectivamente. La figura 4.25 muestra un histograma de probabilidad binomial de la distribución binomial con $n = 25$, $p = 0.6$ con el cual $\mu = 25(0.6) = 15$ y $\sigma = \sqrt{25(0.6)(0.4)} = 2.449$. Sobre el histograma de probabilidad se superpuso una curva normal con estas μ y σ. Aunque el histograma de probabilidad es un poco asimétrico (debido a que $p \neq 0.5$) la curva normal da una muy buena aproximación, sobre todo en la parte media de la figura. El área de cualquier rectángulo (probabilidad de cualquier valor X particular), excepto las de los se ubican en las colas extremas, puede aproximada con precisión mediante el área de la curva normal correspondiente. Por ejemplo, $P(X = 10) = B(10; 25, 0.6) - B(9; 25, 0.6) = 0.021$, mientras que el área bajo la curva normal entre 9.5 y 10.5 es $P(-2.25 \leq Z \leq -1.84) = 0.0207$.

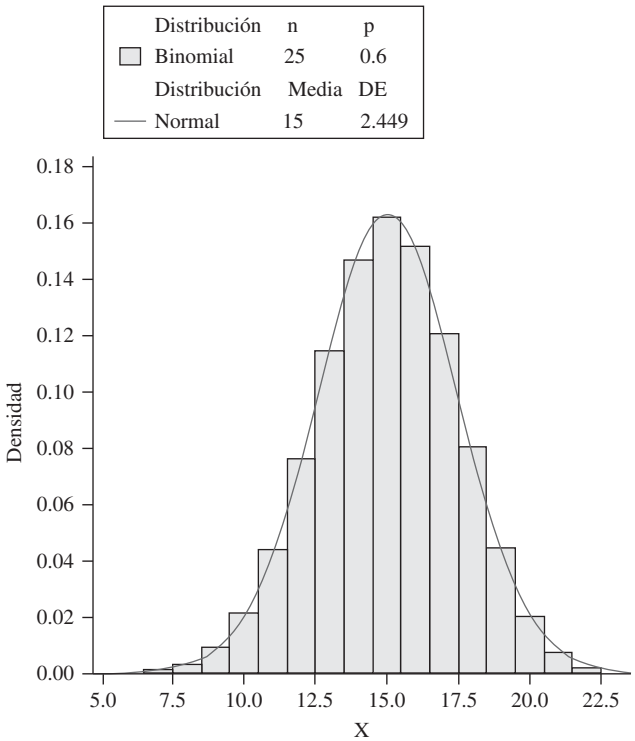


Figura 4.25 Histograma de probabilidad binomial para $n = 25$, $p = 0.6$ con curva de aproximación normal sobrepuesta

De modo más general, en tanto que el histograma de probabilidad binomial no sea demasiado asimétrico las probabilidades binomiales pueden ser aproximadas muy bien por áreas de curva normal. Suele decirse entonces que X tiene aproximadamente una distribución normal

PROPOSICIÓN

Sea X una variable aleatoria binomial basada en n ensayos con probabilidad de éxito p. Entonces, si el histograma de probabilidad binomial no es demasiado asimétrico, X tiene aproximadamente una distribución normal con $\mu = np$ y $\sigma = \sqrt{npq}$. En particular, con $x =$ un valor posible de X,

$$P(X \le x) = B(x, n, p) \approx \left(\begin{array}{c} \text{(área bajo la curva normal} \\ \text{a la izquierda de } x + 0.5) \end{array} \right)$$

$$= \Phi\left(\frac{x + 0.5 - np}{\sqrt{npq}}\right)$$

En la práctica la aproximación es adecuada siempre que $np \ge 10$ y $nq \ge 10$ (es decir, el número esperado de éxitos y el número esperado de fallas son ambos al menos 10), ya que en ese caso existe bastante simetría en la distribución binomial subyacente.

Una comprobación directa de este resultado es bastante difícil. En el siguiente capítulo se verá que esta es consecuencia de un resultado más general llamado teorema del límite central. Con toda honestidad, esta aproximación no es tan importante en el cálculo de probabilidad como alguna vez lo fue. Esto se debe a que los programas de computadora ahora son capaces de calcular con exactitud probabilidades binomiales para valores bastante grandes de n.

EJEMPLO 4.20 Suponga que 25% de todos los estudiantes en una gran universidad pública recibe ayuda financiera. Sea X el número de estudiantes que reciben esta ayuda en una muestra aleatoria de tamaño 50, de modo que $p = 0.25$. Entonces $\mu = 12.5$ y $\sigma = 3.06$. Puesto que $np = 50(0.25) = 12.5 \ge 10$ y $nq = 37.5 \ge 10$, la aproximación puede ser aplicada con seguridad. La probabilidad de que a lo más 10 estudiantes reciban ayuda es

$$P(X \le 10) = B(10; 50, 0.25) \approx \Phi\left(\frac{10 + 0.5 - 12.5}{3.06}\right)$$

$$= \Phi(-0.65) = 0.2578$$

Asimismo, la probabilidad de que entre 5 y 15 (inclusive) de los estudiantes seleccionados reciban ayuda es

$$P(5 \le X \le 15) = B(15; 50, 0.25) - B(4; 50, 0.25)$$

$$\approx \Phi\left(\frac{15.5 - 12.5}{3.06}\right) - \Phi\left(\frac{4.5 - 12.5}{3.06}\right) = 0.8320$$

Las probabilidades exactas son 0.2622 y 0.8348, respectivamente, así que las aproximaciones son bastante buenas. En el último cálculo la probabilidad está siendo aproximada por el área bajo la curva normal entre 4.5 y 15.5; se utiliza la corrección de continuidad tanto para el límite superior como para el inferior. ■

Cuando el objetivo de la investigación es hacer una inferencia sobre una proporción de población p, el interés se centrará en la proporción muestral de X/n éxitos y no en X. Debido a que esta proporción es exactamente X multiplicada por la constante $1/n$, también tendrá aproximadamente una distribución normal (con media $\mu = p$ y desviación estándar $\sigma = \sqrt{pq/n}$, siempre que $np \ge 10$ y $nq \ge 10$. Esta aproximación normal es la base de varios procedimientos inferenciales que se abordarán en capítulos posteriores.

EJERCICIOS Sección 4.3 (28–58)

28. Sea Z una variable aleatoria normal estándar y calcule las siguientes probabilidades. Trace las figuras siempre que sea apropiado.

a. $P(0 \le Z \le 2.17)$ **b.** $P(0 \le Z \le 1)$
c. $P(-2.50 \le Z \le 0)$ **d.** $P(-2.50 \le Z \le 2.50)$
e. $P(Z \le 1.37)$ **f.** $P(-1.75 \le Z)$
g. $P(-1.50 \le Z \le 2.00)$ **h.** $P(1.37 \le Z \le 2.50)$
i. $P(1.50 \le Z)$ **j.** $P(|Z| \le 2.50)$

29. En cada caso determine el valor de la constante c que hace que el enunciado de probabilidad sea correcto.

a. $\Phi(c) = 0.9838$ **b.** $P(0 \le Z \le c) = 0.291$
c. $P(c \le Z) = 0.121$ **d.** $P(-c \le Z \le c) = 0.668$
e. $P(c \le |Z|) = 0.016$

30. Encuentre los siguientes percentiles de la distribución normal estándar. Interpole en los casos en que sea apropiado.

a. 91° **b.** 9° **c.** 75°
d. 25° **e.** 6°

31. Determine z_α para los siguientes valores de α:

a. $\alpha = 0.0055$ **b.** $\alpha = 0.09$
c. $\alpha = 0.663$

32. Suponga que la fuerza que actúa en una columna que ayuda a soportar un edificio es una variable aleatoria X normalmente distribuida con media de 15.0 kips y desviación estándar de 1.25 kips. Calcule las siguientes probabilidades por estandarización y luego use la tabla A.3

a. $P(X \le 15)$ **b.** $P(X \le 17.5)$
c. $P(X \ge 10)$ **d.** $P(14 \le X \le 18)$
e. $P(|X - 15| \le 3)$

33. Las mopeds (motos pequeñas con una cilindrada inferior a 50 cm^3) son muy populares en Europa debido a su movilidad, facilidad de uso y bajo costo. El artículo **"Procedure to Verify the Maximum Speed of Automatic Transmission Mopeds in Periodic Motor Vehicle Inspections" (*J. of Automobile Engr.*, 2008: 1615–1623)** describe un banco de pruebas rodante para determinar la velocidad máxima del vehículo. Se propone una distribución normal con valor medio de 46.8 km/h y desviación estándar de 1.75 km/h. Considere la posibilidad de seleccionar al azar una sola de esas mopeds.

a. ¿Cuál es la probabilidad de que la velocidad máxima sea a lo más 50 km/h?
b. ¿Cuál es la probabilidad de que la velocidad máxima sea al menos de 48 km/h?
c. ¿Cuál es la probabilidad de que la velocidad máxima difiera de la media por más de 1.5 desviaciones estándar?

34. El artículo **"Reliability of Domestic-Waste Biofilm Reactors" (*J. of Envir. Engr.*, 1995: 785–790)** sugiere que la concentración de sustrato (mg/cm^3) del afluente que llega a un reactor está normalmente distribuida con $\mu = 0.30$ y $\sigma = 0.06$.

a. ¿Cuál es la probabilidad de que la concentración exceda de 0.50?
b. ¿Cuál es la probabilidad de que la concentración sea cuando mucho de 0.20?
c. ¿Cómo caracterizaría el más grande 5% de todos los valores de concentración?

35. En un proceso de pavimentación de una carretera la mezcla bituminosa se entrega en la tolva de la pavimentadora mediante camiones que transportan el material desde la planta de concreto. El artículo **"Modeling of Simultaneously Continuous and Stochastic Construction Activities for Simulation" (*J. of Construction Engr. and Mgmnt.*, 2013: 1037–1045)** propone una distribución normal con media de 8.46 min y desviación estándar de 0.913 min para la variable aleatoria X = tiempo del camión transportador.

a. ¿Cuál es la probabilidad de que el tiempo de entrega del camión transportador sea al menos de 10 minutos? ¿Superará los 10 min?
b. ¿Cuál es la probabilidad de que el tiempo de entrega del camión transportador supere los 15 min?
c. ¿Cuál es la probabilidad de que el tiempo del camión transportador esté entre 8 y 10 min?
d. ¿Qué valor c es tal que 98% de todos los tiempos del camión transportador están en el intervalo de $8.46 - c$ a $8.46 + c$?
e. Si se seleccionan en forma independiente cuatro tiempos de entrega del camión transportador, ¿cuál es la probabilidad de que al menos uno de ellos exceda los 10 min?

36. La dispersión de las atomizaciones de los pesticidas es una preocupación constante de los fumigadores y los productores agrícolas. La relación inversa entre el tamaño de la gota y el potencial de deriva es bien conocida. El artículo **"Effects of 2,4-D Formulation and Quinclorac on Spray Droplet Size and Deposition" (*Weed Technology*, 2005: 1030–1036)** es una investigación sobre los efectos de las formulaciones de herbicidas en atomizaciones. Una figura en el artículo sugiere que la distribución normal con media de 1050 μm y desviación estándar de 150 μm es un modelo razonable del tamaño de las gotas de agua (el "tratamiento de control") rociadas a través de una boquilla de 760 ml/min.

a. ¿Cuál es la probabilidad de que el tamaño de una sola gota sea de menos de 1500 μm? ¿Y al menos de 1000 μm?
b. ¿Cuál es la probabilidad de que el tamaño de una sola gota este entre 1000 y 1500 μm?
c. ¿Cómo caracterizaría el más pequeño 2% de todas las gotas?
d. Si se mide el tamaño de cinco gotas, seleccionadas de manera independiente, ¿cuál es la probabilidad de que al menos una exceda de 1500 μm?

37. Suponga que la concentración de cloruro en sangre (mmol/L) tiene una distribución normal con media de 104 y desviación estándar de 5 (la información en el artículo **"Mathematical Model of Chloride Concentration in Human Blood", *J. of Med. Engr. and Tech.*, 2006; 25–30**, que incluye una gráfica

de probabilidad normal, como se describe en la sección 4.6, apoya esta suposición).

a. ¿Cuál es la probabilidad de que la concentración de cloruro sea igual a 105? ¿Y de sea menor de 105? ¿Y de que sea cuando mucho de 105?

b. ¿Cuál es la probabilidad de que la concentración de cloruro difiera de la media por más de 1 desviación estándar? ¿Depende esta probabilidad de los valores de μ y σ?

c. ¿Cómo caracterizaría el más extremo 0.1% de los valores de concentración de cloruro?

38. Hay dos máquinas disponibles que cortan los corchos para tapar las botellas de vino. La primera produce corchos con diámetros que están normalmente distribuidos con media de 3 cm y desviación estándar de 0.1 cm. La segunda máquina produce corchos con diámetros que tienen una distribución normal con media de 3.04 cm y desviación estándar de 0.02 cm. Los corchos aceptables tienen diámetros de entre 2.9 y 3.1 cm. ¿Cuál máquina es más probable que produzca un corcho aceptable?

39. La longitud del defecto de un defecto de corrosión en un tubo de acero presurizado se distribuye normalmente con una media de 30 mm y desviación estándar de 7.8 mm [sugeridos en el artículo **"Reliability Evaluation of Corroding Pipelines Considering Multiple Failure Modes and Time-Dependent Internal Pressure" (*J. of Infrastructure Systems*, 2011: 216–224)**].

a. ¿Cuál es la probabilidad de que el defecto de longitud sea a lo más de 20 mm? ¿Y de menos de 20 mm?

b. ¿Cuál es el percentil 75 de la distribución de longitud de defecto, es decir, el valor que separa el más mínimo 75% de todas las longitudes del 25% más grande?

c. ¿Cuál es el percentil 15 de la distribución de longitud de defecto?

d. ¿Qué valores separan 80% del medio de la distribución de longitud de defecto del menor 10% y del 10% más grande?

40. El artículo **"Monte Carlo Simulation–Tool for Better Understanding of LRFD" (*J. of Structural Engr.*, 1993: 1586–1599)** sugiere que la resistencia a ceder (kg/pulg^2) de un acero grado A36 normalmente está distribuida con $\mu = 43$ y $\sigma = 4.5$.

a. ¿Cuál es la probabilidad de que la resistencia a ceder sea cuando mucho de 40? ¿Y de más de 60?

b. ¿Qué valor de resistencia a ceder separa al más resistente 75% del resto?

41. El dispositivo de apertura automática de un paracaídas de carga militar está diseñado para abrirse 200 m del suelo. Suponga que la altitud de abertura en realidad tiene una distribución normal con valor medio de 200 m y desviación estándar de 30 m. La carga útil se dañará si el paracaídas se abre a una altitud de menos de 100 m. ¿Cuál es la probabilidad de que se dañe la carga útil de al menos uno de cinco paracaídas lanzados en forma independiente?

42. La lectura de temperatura tomada con un termopar colocado en un medio a temperatura constante normalmente se distribuye con media μ, la temperatura real del medio, y desviación estándar σ. ¿Qué valor debe tener σ para asegurarse de que 95% de todas las lecturas están dentro de 0.1° de μ?

43. La velocidad del vehículo sobre un puente en particular de China se puede modelar como una distribución normal **("Fatigue Reliability Assessment for Long-Span Bridges under Combined Dynamic Loads from Winds and Vehicles", *J. of Bridge Engr.*, 2013: 735–747).**

a. Si 5% de todos los vehículos viaja a menos de 39.12 m/h y 10% viaja a más de 73.24 m/h, ¿cuáles son la media μ y la desviación estándar σ de la velocidad de los vehículos? [*Nota*: Los valores resultantes deben concordar con los datos que se indican en el artículo citado.]

b. ¿Cuál es la probabilidad de que la velocidad de un vehículo seleccionado al azar se halle entre 50 y 65 m/h?

c. ¿Cuál es la probabilidad de que la velocidad de un vehículo seleccionado al azar exceda el límite de velocidad de 70 m/h?

44. Si la longitud de rosca de un perno está normalmente distribuida, cuál es la probabilidad de que la longitud de rosca de un perno seleccionado al azar esté:

a. ¿Dentro de 1.5 desviaciones estándar de su media?

b. ¿A más de 2.5 desviaciones estándar de su media?

c. ¿Entre 1 y 2 desviaciones estándar de su media?

45. Una máquina que produce cojinetes de bolas inicialmente se ajustó de manera que el diámetro promedio real de los cojinetes que produce sea de 0.500 pulg. Un cojinete es aceptable si su diámetro está dentro de 0.004 pulg de su valor objetivo. Suponga, sin embargo, que el ajuste cambia durante el curso de la producción, de modo que los cojinetes tengan diámetros normalmente distribuidos con media de 0.499 pulg y desviación estándar de 0.002 pulg. ¿Qué porcentaje de los cojinetes producidos no será aceptable?

46. La dureza Rockwell de un metal se determina al hincar una punta endurecida en la superficie del metal y luego medir la profundidad de penetración. Suponga que la dureza Rockwell de una aleación particular está normalmente distribuida con media de 70 y desviación estándar de 3.

a. Si una muestra es aceptable sólo si su dureza oscila entre 67 y 75, ¿cuál es la probabilidad de que una muestra seleccionada al azar tenga una dureza aceptable?

b. Si el rango de dureza aceptable es $(70 - c, 70 + c)$, ¿con qué valor de c 95% de todas las muestras se tendría una dureza aceptable?

c. Si el rango de dureza aceptable es como el del inciso a) y la dureza de cada una de diez muestras seleccionadas al azar se determina de forma independiente, ¿cuál es el valor esperado de las muestras aceptables entre las diez?

d. ¿Cuál es la probabilidad de que cuando mucho ocho de diez muestras seleccionadas de manera independiente tengan una dureza de menos de 73.84? [*Sugerencia:* $Y =$ el número de entre las diez muestras con dureza de menos de 73.84 es una variable binomial; ¿cuál es p?]

47. La distribución del peso de paquetes enviados de cierta manera es normal con valor medio de 12 lb y desviación estándar de 3.5 lb. El servicio de paquetería desea establecer un valor de peso c más allá del cual se hará un cargo extra. ¿Qué valor de c es tal que 99% de todos los paquetes están al menos 1 lb por debajo del peso de cargo extra?

48. Suponga que la tabla A.3 del apéndice contiene $\Phi(z)$ sólo para $z \geq 0$. Explique cómo aun así podría calcular:

a. $P(-1.72 \leq Z \leq -0.55)$

b. $P(-1.72 \leq Z \leq 0.55)$

¿Es necesario tabular $\Phi(z)$ para z negativo? ¿Qué propiedad de la curva normal estándar justifica su respuesta?

49. Considere a los bebés nacidos en el rango "normal" de 37–43 semanas de gestación. Datos extensos sustentan la suposición de que al nacer estos bebés, nacidos en Estados Unidos, su peso está normalmente distribuido con media de 3432 g y desviación estándar de 482 g. [El artículo **"Are Babies Normal?" (*The American Statistician* (1999): 298–302)** analiza datos de un año particular; para una selección sensible de intervalos de clase, un histograma no parecía del todo normal pero después de una investigación se determinó que esto se debía a que en algunos hospitales medían el peso en gramos y en otros lo medían a la onza más cercana y luego lo convertían a gramos. Una selección modificada de intervalos de clase que permitía esto produjo un histograma que se describía muy bien mediante una distribución normal.]

a. ¿Cuál es la probabilidad de que el peso al nacer de un bebé de este tipo, seleccionado al azar, exceda los 4000 gramos? ¿Y de que esté entre 3000 y 4000 gramos?

b. ¿Cuál es la probabilidad de que el peso al nacer de un bebé de este tipo, seleccionado al azar, sea de menos de 2000 o de más de 5000 gramos?

c. ¿Cuál es la probabilidad de que el peso al nacer de un bebé de este tipo, seleccionado al azar, exceda las 7 libras?

d. ¿Cómo caracterizaría el más extremo 0.1% de todos los pesos al nacer?

e. Si X es una variable aleatoria con una distribución normal y a es una constante numérica ($a \neq 0$), entonces $Y = aX$ también tiene una distribución normal. Use esto para determinar la distribución de pesos al nacer expresados en libras (forma, media y desviación estándar) y luego calcule otra vez la probabilidad del inciso c). ¿Cómo se compara ésta con su respuesta previa?

50. En respuesta a las preocupaciones sobre el contenido nutricional de las comidas rápidas McDonald′s ha anunciado que utilizará un nuevo aceite de cocinar para sus papas a la francesa que reducirá sustancialmente los niveles de ácidos grasos e incrementará la cantidad de grasa poliinsaturada más benéfica. La compañía afirma que 97 de cada 100 personas son incapaces de detectar una diferencia de sabor entre los nuevos aceites y los anteriores. Suponiendo que esta cifra es correcta (como proporción de largo plazo), cuál es la probabilidad aproximada de que en una muestra aleatoria de 1000 individuos que han comprado papas a la francesa en McDonald′s:

a. ¿Al menos 40 puedan notar la diferencia de sabor entre los dos aceites?

b. ¿Cuando mucho 5% pueda notar la diferencia de sabor entre los dos aceites?

51. La desigualdad de Chebyshev (véase el ejercicio 44 del capítulo 3), es válida para distribuciones continuas y discretas. Estipula que para cualquier número k que satisfaga $k \geq 1$, $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ (para una interpretación véase el ejercicio 44 del capítulo 3). Obtenga esta probabilidad en el caso de una distribución normal con $k = 1, 2, 3$ y compare con el límite superior.

52. Sea X el número de defectos en un carrete de cinta magnética de 100 m (una variable de valor entero). Suponga que X tiene aproximadamente una distribución normal con $\mu = 25$ y $\sigma = 5$. Use la corrección de continuidad para calcular la probabilidad de que el número de defectos sea:

a. Entre 20 y 30, inclusive

b. Cuando mucho 30. Menos de 30.

53. Si X tiene una distribución binomial con parámetros $n = 25$ y p, calcule cada una de las siguientes probabilidades mediante la aproximación normal (con la corrección de continuidad) en los casos $p = 0.5$, 0.6 y 0.8 y compare con las probabilidades exactas calculadas con la tabla A.1 del apéndice.

a. $P(15 \leq X \leq 20)$

b. $P(X \leq 15)$

c. $P(20 \leq X)$

54. Suponga que 10% de todas las flechas de acero producidas mediante un proceso no cumplen con las especificaciones pero pueden volver a trabajarse (en lugar de ser desechadas). Considere una muestra aleatoria de 200 flechas y sea X el número de flechas entre estas que no cumplen con las especificaciones y pueden volver a trabajarse. Cuál es la probabilidad aproximada de que X sea:

a. ¿Cuando mucho 30?

b. ¿Menos que 30?

c. ¿Entre 15 y 25 (inclusive)?

55. Suponga que en un estado sólo 75% de todos los conductores usan con regularidad el cinturón de seguridad. Se selecciona una muestra aleatoria de 500 conductores. Cuál es la probabilidad de que:

a. ¿Entre 360 y 400 (inclusive) de los conductores en la muestra usen con regularidad el cinturón de seguridad?

b. ¿Menos de 400 de aquellos en la muestra usen con regularidad el cinturón de seguridad?

56. Demuestre que la relación entre un percentil normal general y el percentil z correspondiente es como se estipuló en esta sección.

57. **a.** Demuestre que si X tiene una distribución normal con parámetros μ y σ, entonces $Y = aX + b$ (una función lineal de X) también tiene una distribución normal. ¿Cuáles son los parámetros de la distribución de Y [es decir, $E(Y)$ y $V(Y)$]? [*Sugerencia:* Escriba la función de distribución acumulada de Y, $P(Y \leq y)$ como una integral que implique la función de densidad de probabilidad de X y luego derive respecto a y para obtener la función de densidad de probabilidad de Y.]

b. Si cuando se mide en °C la temperatura está normalmente distribuida con media de 115 y desviación estándar de 2, ¿qué se puede decir sobre la distribución de temperatura medida en °F?

58. No existe una fórmula exacta para la función de distribución acumulada normal estándar $\Phi(z)$, aunque se han publicado varias buenas aproximaciones en artículos. La siguiente se tomó de **"Approximations for Hand Calculators Using Small Integer Coefficients" (*Mathematics of Computation*, 1977: 214–222)**. Con $0 < z \leq 5.5$,

$$P(Z \geq z) = 1 - \Phi(z)$$
$$\approx 0.5 \exp\left\{-\left[\frac{(83z + 351)z + 562}{703/z + 165}\right]\right\}$$

El error relativo de esta aproximación es de menos de 0.042%. Úselo para calcular aproximaciones a las siguientes probabilidades y compare siempre que sea posible con las probabilidades obtenidas de la tabla A.3 del apéndice.

a. $P(Z \geq 1)$ **b.** $P(Z < -3)$

c. $P(-4 < Z < 4)$ **d.** $P(Z > 5)$

4.4 Distribuciones exponencial y gamma

La curva de densidad correspondiente a cualquier distribución normal tiene forma de campana y, por consiguiente, es simétrica. Existen muchas situaciones prácticas en las cuales la variable de interés para un investigador podría tener una distribución asimétrica. Una familia de distribuciones que tiene esta propiedad es la familia gamma. Primero consideraremos un caso especial, la distribución exponencial, y luego generalizaremos más adelante en esta sección.

Distribución exponencial

La familia de distribuciones exponenciales proporciona modelos de probabilidad que se utilizan mucho en disciplinas de ingeniería y ciencias.

DEFINICIÓN

Se dice que X tiene una **distribución exponencial** con parámetro λ ($\lambda > 0$) si la función de densidad de probabilidad de X es

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{de lo contrario} \end{cases} \qquad \textbf{(4.5)}$$

Algunas fuentes escriben la función de densidad de probabilidad exponencial en la forma $(1/\beta)e^{-x/\beta}$, de modo que $\beta = 1/\lambda$. El valor esperado de una variable aleatoria X exponencialmente distribuida es

$$\mu = E(X) = \int_0^\infty x\lambda e^{-\lambda x}\, dx$$

Para obtener este valor esperado se requiere integrar por partes. La varianza de X se calcula utilizando el hecho de que $V(X) = E(X^2) - [E(X)]^2$. La determinación de $E(X^2)$ requiere integrar por partes dos veces en sucesión. Los resultados de estas integraciones son los siguientes:

$$\mu = \frac{1}{\lambda} \qquad \sigma^2 = \frac{1}{\lambda^2}$$

Tanto la media como la desviación estándar de la distribución exponencial son iguales a $1/\lambda$. En la figura 4.26 aparecen varias gráficas de distintas funciones de densidad de probabilidad exponenciales.

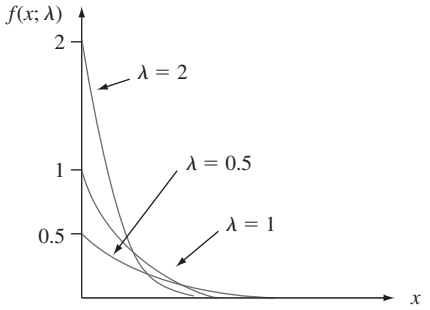


Figura 4.26 Curvas de densidad exponencial

La función de densidad de probabilidad exponencial es fácil de integrar para obtener la función de densidad acumulativa.

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

EJEMPLO 4.21 El artículo **"Probabilistic Fatigue Evaluation of Riveted Railway Bridges" (*J. of Bridge Engr.*, 2008: 237–244)** sugiere la distribución exponencial con valor medio de 6 MPa como modelo para la distribución del rango de esfuerzos en las conexiones de determinados puentes. Supongamos que este es en realidad el verdadero modelo. Entonces $E(X) = 1/\lambda = 6$ implica que $\lambda = 0.1667$. La probabilidad de que el rango de esfuerzos a lo más sea de 10 MPa es

$$P(X \leq 10) = F(10; 0.1667) = 1 - e^{-(0.1667)(10)} = 1 - 0.189 = 0.811$$

La probabilidad de que el rango de esfuerzo esté entre 5 y 10 MPa es

$$P(5 \leq X \leq 10) = F(10; 0.1667) - F(5; 0.1667) = (1 - e^{-1.667}) - (1 - e^{-0.8335}) = 0.246 \quad \blacksquare$$

La distribución exponencial se utiliza con frecuencia como modelo de la distribución de tiempos entre la ocurrencia de eventos sucesivos, tales como los clientes que llegan a una instalación de servicio o las llamadas que entran a través de un conmutador. La razón de esto es que la distribución exponencial está estrechamente relacionada con el proceso de Poisson que se abordó en el capítulo 3.

PROPOSICIÓN

Suponga que el número de eventos que ocurren en cualquier intervalo de tiempo de duración t tiene una distribución de Poisson con parámetro αt (donde α, la tasa del proceso de eventos, es el número esperado de eventos que ocurren en 1 unidad de tiempo) y que la cantidad de ocurrencias en intervalos no traslapantes es independiente entre uno y otro. Entonces la distribución del tiempo transcurrido entre la ocurrencia de dos eventos sucesivos es exponencial con parámetro $\lambda = \alpha$.

Aunque una comprobación completa queda fuera del alcance de este libro, el resultado es fácil de verificar para el tiempo X_1 hasta que ocurre el primer evento:

$$\begin{aligned} P(X_1 \leq t) &= 1 - P(X_1 > t) = 1 - P[\text{no hay eventos en } (0,\ t)] \\ &= 1 - \frac{e^{-\alpha t} \cdot (\alpha t)^0}{0!} = 1 - e^{-\alpha t} \end{aligned}$$

la cual es exactamente la función de distribución acumulada de la distribución exponencial.

EJEMPLO 4.22 Suponga que de acuerdo con un proceso de Poisson se reciben llamadas durante 24 horas en un "Centro de prevención de ayuda a víctimas de violación" a razón de $\alpha = 5$ llamadas por día. Entonces el número de días X entre llamadas sucesivas tiene una distribución exponencial con valor de parámetro 0.5, así que la probabilidad de que transcurran más de dos días entre llamadas es

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2; 0.5) = e^{-(0.5)(2)} = 0.368$$

El tiempo esperado entre llamadas sucesivas es $1/0.5 = 2$ días. ■

Otra aplicación importante de la distribución exponencial es modelar la distribución de la duración de un componente. Una razón parcial de la popularidad de tales aplicaciones es la **propiedad "de no memoria"** de la distribución exponencial. Suponga que la duración de un componente está exponencialmente distribuida con parámetro λ. Después de poner el componente en servicio se deja que pase un periodo de t_0 horas y luego se ve si el componente sigue trabajando; ¿cuál es ahora la probabilidad de que dure al menos t horas más? En símbolos, se desea $P(X \geq t + t_0 | X \geq t_0)$. Mediante la definición de probabilidad condicional,

$$P(X \geq t + t_0 | X \geq t_0) = \frac{P[(X \geq t + t_0) \cap (X \geq t_0)]}{P(X \geq t_0)}$$

Pero el evento $X \geq t_0$ en el numerador es redundante, puesto que ambos eventos pueden ocurrir si y sólo si $X \geq t + t_0$. Por consiguiente,

$$P(X \geq t + t_0 | X \geq t_0) = \frac{P(X \geq t + t_0)}{P(X \geq t_0)} = \frac{1 - F(t + t_0; \lambda)}{1 - F(t_0; \lambda)} = e^{-\lambda t}$$

Esta probabilidad condicional es idéntica a la probabilidad original $P(X \geq t)$ de que el componente dure t horas. Por tanto, *la distribución de duración adicional es exactamente la misma que la distribución de duración original*, así que en cada punto en el tiempo el componente no muestra ningún efecto de desgaste. En otras palabras, la distribución de la duración restante es independiente de la antigüedad actual.

Aunque la propiedad de no memoria se justifica al menos en forma aproximada en muchos problemas de aplicación, en otras situaciones los componentes se deterioran con el tiempo, o de vez en cuando mejoran con él (al menos hasta cierto punto). Las distribuciones gamma, de Weibull y lognormal proporcionan modelos de duración más generales (las dos últimas se abordan en la siguiente sección).

La función gamma

Para definir a la familia de distribuciones gamma, primero se tiene que introducir una función que desempeña un importante papel en muchas ramas de las matemáticas.

DEFINICIÓN

Con $\alpha > 0$, la **función gamma** $\Gamma(\alpha)$ se define como

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \qquad \textbf{(4.6)}$$

Las propiedades más importantes de la función gamma son las siguientes:

1. Con cualquier $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1) \cdot \Gamma(\alpha - 1)$ [a través de la integración por partes]

2. Con cualquier entero positivo, n, $\Gamma(n) = (n - 1)!$

3. $\Gamma(1/2) = \sqrt{\pi}$

Ahora, sea

$$f(x;\alpha)=\begin{cases}\dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & x\geq 0\\ 0 & \text{de lo contrario}\end{cases} \tag{4.7}$$

Entonces $f(x;\alpha)\geq 0$. La expresión (4.6), implica que $\int_0^\infty f(x;\alpha)\,dx=\Gamma(\alpha)/\Gamma(\alpha)=1$. Por tanto $f(x;\alpha)$ satisface las dos propiedades básicas de una función de densidad de probabilidad.

La distribución gamma

DEFINICIÓN

Se dice que una variable aleatoria continua X tiene una **distribución gamma** si la función de densidad de probabilidad de X es

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & x\geq 0\\ 0 & \text{de lo contrario}\end{cases} \tag{4.8}$$

donde los parámetros α y β satisfacen $\alpha>0$, $\beta>0$. La **distribución gamma estándar** tiene $\beta=1$, así que (4.7) da la función de densidad de probabilidad de una variable aleatoria gamma estándar.

La distribución exponencial es resultado de considerar $\alpha=1$ y $\beta=1/\lambda$.

La figura 4.27(a) ilustra las gráficas de la función de densidad de probabilidad gamma $f(x;\alpha,\beta)$(4.8) para varios pares (α,β), en tanto que la figura 4.27(b) presenta gráficas de la función de densidad de probabilidad gamma estándar. Para la función de densidad de probabilidad estándar cuando $\alpha\leq 1$, $f(x;\alpha)$ es estrictamente decreciente a medida que x se incrementa desde 0; cuando $\alpha>1$, $f(x;\alpha)$ se eleva desde 0 en $x=0$ hasta un máximo y luego decrece. El parámetro β en (4.8) se llama parámetro de escala, y a α se le conoce como un *parámetro de forma* porque al cambiar su valor altera la forma básica de la curva de densidad.

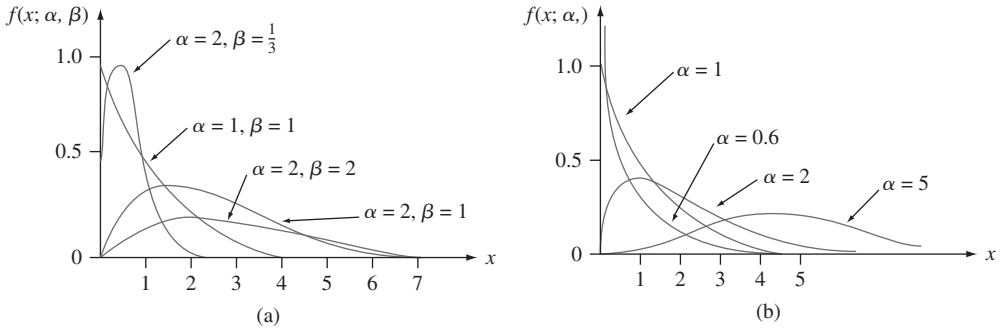


Figura 4.27 (a) Curvas de densidad gamma; (b) curvas de densidad gamma estándar

La media y la varianza de una variable aleatoria X que tiene la distribución gamma $f(x;\alpha,\beta)$ son

$$E(X)=\mu=\alpha\beta \qquad V(X)=\sigma^2=\alpha\beta^2$$

Cuando X es una variable aleatoria gamma estándar, la función de distribución acumulada de X,

$$F(x;\alpha)=\int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}\,dy \qquad x>0 \tag{4.9}$$

se llama **función gamma incompleta** [en ocasiones la función gamma incompleta se refiere a la expresión (4.9) sin el denominador $\Gamma(\alpha)$ en el integrando]. Existen tablas extensas de

$F(x; \alpha)$ disponibles; en la tabla A.4 del apéndice se presenta una pequeña tabulación para $\alpha = 1, 2, \ldots, 10$ y $x = 1, 2, \ldots, 15$.

EJEMPLO 4.23 El artículo **"The Probability Distribution of Maintenance Cost of a System Affected by the Gamma Process of Degradation" (*Reliability Engr. and System Safety*, 2012: 65–76)** observa que la distribución gamma es ampliamente utilizada para modelar el grado de degradación como la corrosión, la fluencia o el desgaste. X representa la cantidad de degradación de un tipo determinado y supongamos que tiene una distribución gamma estándar con $\alpha = 2$. Ya que

$$P(a \le X \le b) = F(b) - F(a)$$

cuando X es continua,

$$P(3 \le X \le 5) = F(5; 2) - F(3; 2) = 0.960 - 0.801 = 0.159$$

La probabilidad de que el tiempo de degradación sea de más de 4 s es

$$P(X > 4) = 1 - P(X \le 4) = 1 - F(4; 2) = 1 - 0.908 = 0.092$$ ■

La función gamma incompleta también se utiliza para calcular probabilidades que implican distribuciones gamma no estándar. Estas probabilidades también se obtienen casi instantáneamente con varios paquetes de software.

PROPOSICIÓN

Si X tiene una distribución gamma con parámetros α y β, entonces con cualquier $x > 0$, la función de distribución acumulada de X es

$$P(X \le x) = F(x; \alpha, \beta) = F\left(\frac{x}{\beta}; \alpha\right)$$

donde $F(\cdot\,; \alpha)$ es la función gamma incompleta.

EJEMPLO 4.24 Suponga que el tiempo de sobrevivencia X en semanas de un ratón macho seleccionado al azar y expuesto a 240 rads de radiación gamma tiene una distribución gamma con $\alpha = 8$ y $\beta = 15$. (La información en ***Survival Distributions: Reliability Applications in the Biomedical Services*, de A. J. Gross y V. Clark**, sugiere $\alpha \approx 8.5$ y $\beta \approx 13.3$.) El tiempo de sobrevivencia esperado es $E(X) = (8)(15) = 120$ semanas, en tanto que $V(X) = (8)(15)^2 = 1800$ y $\sigma_X = \sqrt{1800} = 42.43$ semanas. La probabilidad de que un ratón sobreviva entre 60 y 120 semanas es

$$\begin{aligned} P(60 \le X \le 120) &= P(X \le 120) - P(X \le 60) \\ &= F(120/15; 8) - F(60/15; 8) \\ &= F(8; 8) - F(4; 8) = 0.547 - 0.051 = 0.496 \end{aligned}$$

La probabilidad de que un ratón sobreviva por lo menos 30 semanas es

$$\begin{aligned} P(X \ge 30) &= 1 - P(X < 30) = 1 - P(X \le 30) \\ &= 1 - F(30/15; 8) = 0.999 \end{aligned}$$ ■

Distribución ji-cuadrada

La distribución ji-cuadrada es importante porque es la base de varios procedimientos de inferencia estadística. El papel central desempeñado por la distribución ji-cuadrada en inferencia se deriva de su relación con distribuciones normales (véase el ejercicio 71). Esta distribución se abordará con más detalle en capítulos posteriores.

DEFINICIÓN

Sea ν un entero positivo. Se dice, por tanto, que una variable aleatoria X tiene una **distribución chi o ji-cuadrada** con parámetro ν si la función de densidad de probabilidad de X es la densidad gamma con $\alpha = \nu/2$ y $\beta = 2$. La función de densidad de probabilidad de una variable aleatoria ji-cuadrada es entonces

$$f(x; \nu) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad \textbf{(4.10)}$$

El parámetro ν se llama **número de grados de libertad** (gl) de X. A menudo se utiliza el símbolo χ^2 en lugar de "ji-cuadrada".

EJERCICIOS Sección 4.4 (59–71)

59. Sea X = el tiempo entre dos llegadas sucesivas a la ventanilla de autopago de un banco local. Si X tiene una distribución exponencial con $\lambda = 1$ (la cual es idéntica a una distribución gamma estándar con $\alpha = 1$) calcule lo siguiente:

a. El tiempo esperado entre dos llegadas sucesivas.

b. La desviación estándar del tiempo entre llegadas sucesivas

c. $P(X \leq 4)$ **d.** $P(2 \leq X \leq 5)$

60. Sea X la distancia (m) que un animal recorre desde el sitio de su nacimiento hasta el primer territorio disponible que encuentra. Suponga que para ratas canguro cola de bandera, X tiene una distribución exponencial con parámetro $\lambda = 0.01386$ (como lo sugiere el artículo **"Competition and Dispersal from Multiple Nests", *Ecology*, 1997: 873–883**).

a. ¿Cuál es la probabilidad de que la distancia sea cuando mucho de 100 m? ¿Y cuándo mucho de 200 m? ¿Y entre 100 y 200 m?

b. ¿Cuál es la probabilidad de que la distancia exceda la distancia media por más de 2 desviaciones estándar?

c. ¿Cuál es el valor de la distancia mediana?

61. Los datos recogidos en el Aeropuerto Internacional Toronto Pearson sugieren que una distribución exponencial con valor medio de 2.725 horas es un buen modelo para la duración de la lluvia (***Urban Stormwater Management Planning with Analytical Probabilistic Models*, 2000, p. 69**).

a. ¿Cuál es la probabilidad de que la duración de un evento de lluvia en este lugar particular sea al menos de 2 horas? ¿A lo más de 3 horas? ¿Entre 2 y 3 horas?

b. ¿Cuál es la probabilidad de que la duración de la lluvia supere la media por más de dos desviaciones estándar? ¿Cuál es la probabilidad de que sea menor que la media en más de una desviación estándar?

62. El artículo **"Microwave Observations of Daily Antarctic Sea-Ice Edge Expansion and Contribution Rates" (*IEEE Geosci. and Remote Sensing Letters*, 2006: 54–58)** establece que "la distribución del avance-retroceso diario del hielo marino respecto a cada sensor es similar y es aproximadamente una exponencial doble". La distribución exponencial doble propuesta tiene una función de densidad con $f(x) = 0.5\lambda e^{-\lambda|x|}$ para $-\infty < x < \infty$. La desviación estándar se da como 40.9 km.

a. ¿Cuál es el valor del parámetro λ?

b. ¿Cuál es la probabilidad de que la extensión del cambio del hielo marino esté dentro de 1 desviación estándar de la media?

63. Un consumidor está tratando de decidir entre dos planes de llamadas de larga distancia. El primero aplica una sola tarifa de 10¢ por minuto, en tanto que el segundo cobra una tarifa de 99¢ por llamadas hasta de 20 minutos y luego 10¢ por cada minuto adicional que exceda de 20 (suponga que las llamadas que duran un número no entero de minutos se cobran en proporcionalmente a un cargo por minuto entero). Suponga que la distribución de la duración de llamadas del consumidor es exponencial con parámetro λ.

a. Explique intuitivamente cómo la selección del plan de llamadas deberá depender de cuál sea la duración de las llamadas.

b. ¿Cuál plan es mejor si la duración esperada de las llamadas es de 10 minutos? ¿Y si es de 15 minutos? [*Sugerencia:* Sea $h_1(x)$ el costo del primer plan cuando la duración de las llamadas es de x minutos y sea $h_2(x)$ la función de costo del segundo plan. Dé expresiones para estas dos funciones de costo y luego determine el costo esperado de cada plan.]

64. Evalúe lo siguiente:

a. $\Gamma(6)$ **b.** $\Gamma(5/2)$

c. $F(4; 5)$ (la función gamma incompleta) y $F(5; 4)$

d. $P(X \leq 5)$ cuando X tiene una distribución gamma estándar con $\alpha = 7$.

e. $P(3 < X < 8)$ cuando X tiene la distribución especificada en (d).

65. X indica el tiempo de transferencia de datos (ms) en un sistema de red informática (el tiempo requerido para la transferencia de

datos entre una computadora "trabajador" y una computadora "maestro". Suponga que X tiene una distribución gamma con media de 37.5 ms y desviación estándar 21.6 (sugeridos por el artículo **"Computation Time of Grid Computing with Data Transfer Times that Follow a Gamma Distribution,"** ***Proceedings of the First International Conference on Semantics, Knowledge, and Grid*****, 2005).**

a. ¿Cuáles son los valores de α y β?

b. ¿Cuál es la probabilidad de que el tiempo de la transferencia de datos sea mayor de 50 ms?

c. ¿Cuál es la probabilidad de que tiempo de transferencia de datos este entre 50 y 75 ms?

66. La distribución gamma de dos parámetros se puede generalizar al introducir un tercer parámetro γ, llamado parámetro *umbral* o de *ubicación*: sustituya x en (4.8) por $x - \gamma$ y $x \geq 0$ por $x \geq \gamma$. Estas cantidades recorren las curvas de densidad en la figura 4.27 de tal manera que empiezan a ascender o descender en γ y no en 0. El artículo **"Bivariate Flood Frequency Analysis with Historical Information Based on Copulas" (*J. of Hydrologic Engr*., 2013: 1018–1030)** emplea esta distribución para modelar $X = 3$ días de volumen de inundación (10^8 m^3). Supongamos que los valores de los parámetros son $\alpha = 12$, $\beta = 7$, $\gamma = 40$ (muy cerca de las estimaciones en el citado artículo con base en datos del pasado).

a. ¿Cuáles son la media y la desviación estándar de X?

b. ¿Cuál es la probabilidad de que el volumen de inundación esté entre 100 y 150?

c. ¿Cuál es la probabilidad que el volumen de inundación exceda su media por más de una desviación estándar?

d. ¿Cuál es el percentil 95 de la distribución del volumen de inundación?

67. Suponga que cuando un transistor de cierto tipo se somete a una prueba de duración acelerada, la duración X (en semanas) tiene una distribución gamma con media de 24 semanas y desviación estándar de 12 semanas.

a. ¿Cuál es la probabilidad de que un transistor dure entre 12 y 24 semanas?

b. ¿Cuál es la probabilidad de que un transistor dure cuando mucho 24 semanas? ¿Es la mediana de la distribución de duración menor que 24? ¿Por qué sí o por qué no?

c. ¿Cuál es el percentil 99 de la distribución de duración?

d. Suponga que en realidad la prueba termina después de t semanas. ¿Qué valor de t es tal que sólo 0.5% de todos los transistores continuarán funcionando al término?

68. El caso especial de la distribución gamma en la cual α es un entero positivo n se llama distribución de Erlang. Si se reemplaza β por $1/\lambda$ en la expresión (4.8), la función de densidad de probabilidad de Erlang es

$$f(x; \lambda, n) = \begin{cases} \dfrac{\lambda(\lambda x)^{n-1} e^{-\lambda x}}{(n-1)!} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Se puede demostrar que si los tiempos entre eventos sucesivos son independientes, cada uno con distribución exponencial con parámetro λ, entonces el tiempo total X que transcurre antes de que ocurran los siguientes n eventos tiene una función de densidad de probabilidad $f(x; \lambda, n)$.

a. ¿Cuál es el valor esperado de X? Si el tiempo (en minutos) entre llegadas de clientes sucesivos está exponencialmente distribuido con $\lambda = 0.5$, ¿cuánto tiempo se puede esperar que transcurra antes de que llegue el décimo cliente?

b. Si el tiempo entre llegadas de los clientes está exponencialmente distribuido con $\lambda = 0.5$, ¿cuál es la probabilidad de que el décimo cliente (después del que recién ha llegado) llegue dentro de los siguientes 30 min?

c. El evento $\{X \leq t\}$ ocurre si al menos ocurren n eventos en las siguientes t unidades de tiempo. Use el hecho de que el número de eventos que ocurren en un intervalo de duración t tiene una distribución de Poisson con parámetro λt para escribir una expresión (que implique probabilidades de Poisson) para la función de distribución acumulada $f(t; \lambda, n) = P(X \leq t)$.

69. Un sistema consta de cinco componentes idénticos conectados en serie tal como se muestra:

En cuanto un componente falla, todo el sistema lo hace. Suponga que cada componente tiene una duración que está exponencialmente distribuida con $\lambda = 0.01$ y que los componentes fallan de manera independiente uno de otro. Defina los eventos $A_i = \{$el componente i-ésimo dura al menos t horas$\}$, $i = 1, \ldots, 5$, de modo que los A_i son eventos independientes. Sea $X =$ el tiempo en el cual el sistema falla; es decir, la duración más corta (mínima) entre los cinco componentes.

a. ¿A qué evento equivale el evento $\{X \geq t\}$ que implique $A_1, \ldots A_5$?

b. Utilizando la independencia de los eventos A_i calcule $P(X \leq t)$. Luego obtenga $F(t) = P(X \leq t)$ y la función de densidad de probabilidad de X. ¿Qué tipo de distribución tiene X?

c. Suponga que existen n componentes y que cada uno tiene una duración exponencial con parámetro λ. ¿Qué tipo de distribución tiene X?

70. Si X tiene una distribución exponencial con parámetro λ, deduzca una expresión general para el $(100p)^{\text{o}}$ percentil de la distribución. Luego especialícela para obtener la mediana.

71. **a.** ¿A qué evento equivale el evento $\{X^2 \leq y\}$ que implica a la X misma?

b. Si X tiene una distribución normal estándar use el inciso a) para escribir la integral que es igual a $P(X^2 \leq y)$. Luego derívela respecto a y para obtener la función de densidad de probabilidad de X^2 [el cuadrado de una variable $N(0, 1)$]. Por último, demuestre que tiene una distribución ji-cuadrada con $\nu = 1$ [véase (4.10)]. [*Sugerencia:* Use la siguiente identidad.]

$$\frac{d}{dy}\left\{\int_{a(y)}^{b(y)} f(x)\,dx\right\} = f[b(y)] \cdot b'(y) - f[a(y)] \cdot a'(y)$$

EJERCICIOS SUPLEMENTARIOS (72–86)

72. Sea X = el tiempo que requiere una cabeza de lectura-escritura para localizar un registro deseado en un dispositivo de memoria de disco de una computadora luego que la cabeza se ha colocado sobre la pista correcta. Si los discos giran una vez cada 25 milisegundos, una suposición razonable es que X está uniformemente distribuida en el intervalo [0, 25].

a. Calcule $P(10 \le X \le 20)$.

b. Calcule $P(X \ge 10)$.

c. Obtenga la función de distribución acumulada $F(X)$.

d. Calcule $E(X)$ y σ_X.

73. Una barra de 12 pulg que está sujeta por ambos extremos se somete a una cantidad creciente de esfuerzo hasta que se rompe. Sea Y = la distancia del extremo izquierdo al punto donde ocurre la ruptura. Suponga que Y tiene la función de densidad de probabilidad

$$f(y) = \begin{cases} \left(\frac{1}{24}\right)y\left(1 - \frac{y}{12}\right) & 0 \le y \le 12 \\ 0 & \text{de lo contrario} \end{cases}$$

Calcule lo siguiente:

a. La función de distribución acumulada de Y y grafíquela.

b. $P(Y \le 4)$, $P(Y > 6)$ y $P(4 \le Y \le 6)$

c. $E(Y)$, $E(Y^2)$ y $V(Y)$.

d. La probabilidad de que el punto de ruptura ocurra a más de 2 pulg del punto de ruptura esperado.

e. La longitud esperada del segmento más corto cuando ocurre la ruptura.

74. Sea X el tiempo hasta la falla (en años) de cierto componente hidráulico. Suponga que la función de densidad de probabilidad de X es $f(x) = 32/(x + 4)^3$ con $x < 0$.

a. Verifique que $f(x)$ es una función de densidad de probabilidad legítima.

b. Determine la función de distribución acumulada.

c. Use el resultado del inciso b) para calcular la probabilidad de que el tiempo hasta la falla sea de entre 2 y 5 años.

d. ¿Cuál es el tiempo esperado hasta la falla?

e. Si el componente tiene un valor de recuperación igual a $100/(4 + x)$, cuando su tiempo hasta la falla es x, ¿cuál es el valor de recuperación esperado?

75. El tiempo X para terminar cierta tarea tiene una función de distribución acumulada $F(x)$ dada por

$$\begin{cases} 0 & x < 0 \\ \frac{x^3}{3} & 0 \le x < 1 \\ 1 - \frac{1}{2}\left(\frac{7}{3} - x\right)\left(\frac{7}{4} - \frac{3}{4}x\right) & 1 \le x \le \frac{7}{3} \\ 1 & x > \frac{7}{3} \end{cases}$$

a. Obtenga la función de densidad de probabilidad $f(x)$ y trace su gráfica.

b. Calcule $P(0.5 \le X \le 2)$.

c. Calcule $E(X)$.

76. X representa el número de individuos que responden a una oferta particular de cupones en línea. Suponga que X tiene aproximadamente una distribución Weibull con $\alpha = 10$ y $\beta = 20$. Calcule la mejor aproximación posible a la probabilidad de que X esté entre 15 y 20, inclusive.

77. El artículo **"Computer Assisted Net Weight Control"** (***Quality Progress*, 1983: 22–25**) sugiere una distribución normal con media de 137.2 oz y desviación estándar de 1.6 oz del contenido real de cierto tipo de frascos. El contenido declarado fue de 135 oz.

a. ¿Cuál es la probabilidad de que un solo frasco contenga más que el contenido declarado?

b. Entre diez frascos seleccionados al azar, ¿cuál es la probabilidad de que al menos ocho contengan más que el contenido declarado?

c. Suponiendo que la media permanece en 137.2, ¿a qué valor se tendría que cambiar la desviación estándar de modo que 95% de todos los frascos contengan más que el contenido declarado?

78. Cuando se someten a prueba las tarjetas de circuito utilizadas en la fabricación de reproductores de discos compactos el porcentaje de tarjetas defectuosas a largo plazo es de 5%. Suponga que se recibió un lote de 250 tarjetas y que la condición de cualquier tarjeta particular es independiente de la de cualquier otra.

a. ¿Cuál es la probabilidad aproximada de que al menos 10% de las tarjetas en el lote resulte defectuoso?

b. ¿Cuál es la probabilidad aproximada de que en el lote haya exactamente 10 tarjetas defectuosas?

79. El ejercicio 38 introdujo dos máquinas que producen corchos de vino, la primera de ellas con una distribución normal de diámetro con media de 3 cm y desviación típica de 0.1 cm; y la segunda con una distribución de diámetro normal con media de 3.04 cm y desviación estándar de 0.02 cm. Los corchos aceptables tienen diámetros entre 2.9 y 3.1 cm. Si 60% de todos los corchos utilizados proviene de la primera máquina y se encuentra aceptable un corcho seleccionado al azar, ¿cuál es la probabilidad de que se haya producido en la primera máquina?

80. El tiempo de reacción (en segundos) a un estímulo es una variable aleatoria continua con función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{3}{2} \cdot \frac{1}{x^2} & 1 \le x \le 3 \\ 0 & \text{de lo contrario} \end{cases}$$

a. Obtenga la función de distribución acumulada.

b. ¿Cuál es la probabilidad de que el tiempo de reacción sea cuando mucho de 2.5 s? ¿Y de que esté entre 1.5 y 2.5 s?

c. Calcule el tiempo de reacción esperado.

d. Calcule la desviación estándar del tiempo de reacción.

e. Si un individuo requiere más de 1.5 s para reaccionar, una luz se enciende y permanece así hasta que transcurre un segundo más o hasta que la persona reacciona (lo que suceda primero). Determine la cantidad de tiempo que se espera que la luz permanezca encendida. [*Sugerencia:* Sea $h(X)$ = el tiempo que la luz está encendida como una función del tiempo de reacción X.]

81. Sea X la temperatura a la cual ocurre una reacción química. Suponga que X tiene una función de densidad de probabilidad

$$f(x) = \begin{cases} \frac{1}{9}(4 - x^2) & -1 \le x \le 2 \\ 0 & \text{de lo contrario} \end{cases}$$

a. Trace la gráfica $f(x)$.

b. Determine la función de distribución acumulada y grafíquela.

c. ¿Es 0 la temperatura mediana a la cual ocurre la reacción? Si no, ¿es la temperatura mediana más pequeña o más grande que 0?

d. Suponga que esta reacción se realiza de manera independiente una vez en cada uno de diez laboratorios diferentes y que la función de densidad de probabilidad del tiempo de reacción en cada laboratorio es como se da. Sea Y = el número entre los diez laboratorios en los cuales la temperatura excede de 1. ¿Qué clase de distribución tiene Y? (Dé los nombres y valores de los parámetros.)

82. Un ovocito es una célula germinal femenina relacionado con la reproducción. Con base en el análisis de una muestra grande, el artículo **"Reproductive Traits of Pioneer Gastropod Species Colonizing Deep-Sea Hydrothermal Vents After an Eruption" (*Marine Biology*, 2011: 181–192)** propone la siguiente mezcla de distribuciones normales como un modelo para la distribución de X = diámetro del ovocito (μm):

$$f(x) = pf_1(x; \mu_1, \sigma) + (1 - p) f_2(x; \mu_2, \sigma)$$

donde f_1 y f_2 son funciones de densidad de probabilidad normal. Los valores de los parámetros sugeridos fueron $p = 0.35$, $\mu_1 = 4.4$, $\mu_2 = 5.0$ y $\sigma = 0.27$.

a. ¿Cuál es el valor esperado (es decir la media) del diámetro del ovocito?

b. ¿Cuál es la probabilidad de que el diámetro del ovocito esté entre 4.4 μm y 5.0 μm? [*Sugerencia:* Escriba una expresión para la integral correspondiente, integre las dos componentes y después use el hecho de que cada componente es una función de densidad de probabilidad normal].

c. ¿Cuál es la probabilidad de que el diámetro del ovocito sea menor que su media? ¿Qué implica esto respecto a la forma de la curva densidad?

83. El artículo **"The Prediction of Corrosion by Statistical Analysis of Corrosion Profiles" (*Corrosion Science*, 1985: 305–315)** sugiere la siguiente función de distribución acumulada de la profundidad X del pozo más profundo en un experimento que implica exponer acero de manganeso de carbono al agua de mar acidificada.

$$F(x; \alpha, \beta) = e^{-e^{-(x-\alpha)/\beta}} \quad -\infty < x < \infty$$

Los autores proponen los valores $\alpha = 150$ y $\beta = 90$. Suponga que este es el modelo correcto.

a. ¿Cuál es la probabilidad de que la profundidad del pozo más profundo sea cuando mucho de 150? ¿Y cuando mucho de 300? ¿De que esté entre 150 y 300?

b. ¿Por debajo de qué valor estará la profundidad del pozo máximo en 90% de todos los experimentos?

c. ¿Cuál es la función de densidad de X?

d. Se puede demostrar que la función de densidad es unimodal (una sola cresta). ¿Por encima de qué valor sobre el eje de medición ocurre esta cresta? (Este valor es la moda.)

e. Se puede demostrar que $E(X) \approx 0.5772\beta + \alpha$. ¿Cuál es la media de los valores dados de α y β y cómo se compara esto con la mediana y la moda? Trace la gráfica de la función de densidad. [*Nota*: Esta se conoce como *distribución de valor extremo más grande*.]

84. Sea t = la cantidad del impuesto sobre las ventas que un minorista debe al gobierno por un periodo determinado. En el artículo **"Statistical Sampling in Tax Audits" (*Statistics and the Law*, 2008: 320–343)** se propone modelar la incertidumbre en t, considerándola como una variable aleatoria distribuida normalmente con media μ y desviación estándar σ (en el artículo, estos dos parámetros se estiman a partir de los resultados de una inspección fiscal que implican n operaciones de muestreo). Si a representa el monto con el que minorista es evaluado, entonces resulta una subevaluación si $t > a$ y una sobreevaluación de resultados si $a > t$. La función de sanción propuesta (es decir, la pérdida) para la sobreevaluación o la subevaluación es $L(a, t) = t - a$ si $t > a$ y $= k(a - t)$ si $t \le a$ (se sugiere $k > 1$ para incorporar la idea de que la sobreevaluación es más grave que una subevaluación).

a. Demuestre que $a^* = \mu + \sigma\Phi^{-1}(1/(k + 1))$ es el valor de a que minimiza la pérdida esperada, donde Φ^{-1} es la función inversa de la función de distribución acumulativa normal estándar.

b. Si $k = 2$ (como se sugiere en el artículo), $\mu = \$100\,000$ y $\sigma = \$10\,000$, ¿cuál es el valor óptimo de a y cuál es la probabilidad resultante de la sobreevaluación?

85. La *moda* de una distribución continua es el valor x^* que incrementa al máximo $f(x)$.

a. ¿Cuál es la moda de una distribución normal con parámetros μ y σ?

b. ¿Tiene una sola moda la distribución uniforme con parámetros A y B? ¿Por qué sí o por qué no?

c. ¿Cuál es la moda de una distribución exponencial con parámetro λ? (Trace una gráfica.)

d. Si X tiene una distribución gamma con parámetros α y β, y $\alpha > 1$, halle la moda. [*Sugerencia:* $\ln[f(x)]$ se incrementará al máximo si y sólo si $f(x)$ es, y será más simple sacar la derivada de $\ln[f(x)]$.

e. ¿Cuál es la moda de una distribución ji-cuadrada con ν grados de libertad?

86. El artículo **"Error Distribution in Navigation" (*J. of the Institute of Navigation*, 1971: 429–442)** sugiere que una distribución de frecuencia de errores positivos (magnitudes de errores) es mejor aproximada por una distribución exponencial.

Sea X = el error de posición lateral (millas náuticas) el cual puede ser positivo o negativo. Suponga que la función de densidad de probabilidad de X es

$$f(x) = (0.1)e^{-0.2|x|} \quad -\infty < x < \infty$$

a. Trace una gráfica $f(x)$ y compruebe que $f(x)$ es una función de densidad de probabilidad legítima (demuestre que esta se integra a 1).

b. Obtenga la función de distribución acumulada de X y trácela.

c. Calcule $P(X \leq 0)$, $P(X \leq 2)$, $P(-1 \leq X \leq 2)$, y la probabilidad de que se cometa un error de más de 2 millas.

BIBLIOGRAFÍA

Bury, Karl, *Statistical Distributions in Engineering*, Cambridge Univ. Press, Cambridge, Inglaterra, 1999. Un estudio informativo y fácil de leer sobre las distribuciones y sus propiedades.

Johnson, Norman, Samuel Kotz y N. Balakrishnan, *Continuous Univariate Distributions*, vols. 1–2, Wiley, Nueva York, 1994. Estos dos volúmenes juntos presentan un estudio exhaustivo de varias distribuciones continuas.

Nelson, Wayne, *Applied Data Analysis*, Wiley, Nueva York, 1982. Aborda ampliamente las distribuciones y los métodos que se utilizan en el análisis de datos de vida útil.

Olkin, Ingram, Cyrus Derman y Leon Gleser, *Probability Models and Applications* (2a. ed.), Macmillan, Nueva York, 1994. Una buena cobertura de las propiedades generales y las distribuciones específicas.

Capítulo

5

Estimación puntual

INTRODUCCIÓN

Dado un parámetro de interés, tal como la media μ o la proporción p de una población, el objetivo de la estimación puntual es utilizar una muestra para calcular un número que represente en cierto sentido una buena suposición del valor verdadero del parámetro. El número resultante se llama *estimación puntual*. En la sección 5.1 se presentan algunos conceptos generales de estimación puntual. En la sección 5.2 se describen e ilustran dos métodos importantes para obtener estimaciones puntuales: el método de momentos y el método de máxima probabilidad.

5.1 Algunos conceptos generales de la estimación puntual

El objetivo de la inferencia estadística casi siempre es sacar algún tipo de conclusión sobre uno o más parámetros (características de la población). Para hacer eso, un investigador tiene que obtener datos muestrales de cada una de las poblaciones estudiadas. Las conclusiones pueden entonces basarse en los valores calculados de varias cantidades muestrales. Por ejemplo, sea μ (un parámetro) la resistencia a la ruptura promedio verdadera de conexiones alámbricas utilizadas en la unión de láminas semiconductoras. Se podría tomar una muestra aleatoria de $n = 10$ conexiones y determinar la resistencia a la ruptura de cada una y se tendrían las resistencias observadas $x_1, x_2, \ldots, x_{10}$. La resistencia a la ruptura media muestral $\bar{x}$ se utilizaría entonces para sacar una conclusión respecto al valor de μ. Asimismo, si σ^2 es la varianza de la distribución de la resistencia a la ruptura (varianza de la población, otro parámetro), el valor de la varianza muestral s^2 se utiliza para inferir algo sobre σ^2.

Cuando se discuten los métodos y conceptos generales de inferencia, es conveniente disponer de un símbolo genérico para el parámetro de interés. Se utilizará la letra griega θ para este propósito. En muchas investigaciones, θ será una población con media μ, una diferencia $\mu_1 - \mu_2$, entre dos medias de población, o una proporción de la población de "éxitos" p. El objetivo de la estimación puntual es seleccionar un solo número, con base en los datos muestrales, que represente un valor sensible de θ. Suponga, por ejemplo, que el parámetro de interés es μ, la vida útil promedio verdadera de las baterías de un cierto tipo. Una muestra aleatoria de $n = 3$ baterías podría dar las vidas útiles (horas) observadas $x_1 = 5.0$, $x_2 = 6.4$, $x_3 = 5.9$. El valor calculado de la media muestral de la vida útil es $\bar{x} = 5.77$ y es razonable considerar 5.77 como un valor muy factible de μ, la "mejor suposición" del valor de μ basado en la información muestral disponible.

Suponga que se desea estimar un parámetro de una sola población (p. ej., μ o σ) con una muestra aleatoria de tamaño n. Recuerde por el capítulo previo que antes de que los datos estén disponibles, las observaciones muestrales deben ser consideradas variables aleatorias $X_1, X_2, \ldots, X_n$. Se deduce que cualquier función de las X_i, es decir, cualquier estadístico, tal como la media muestral $\overline{X}$ o la desviación estándar muestral S también es una variable aleatoria. Puede decirse lo mismo si los datos disponibles se componen de más de una muestra. Por ejemplo, se pueden representar las resistencias a la tensión de m especímenes de tipo 1 y de n especímenes de tipo 2 por $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_n$, respectivamente. La diferencia entre las dos medias muestrales de las resistencias es $\overline{X} - \overline{Y}$; este es el estadístico natural para inferir sobre $\mu_1 - \mu_2$, la diferencia entre las resistencias medias de la población.

DEFINICIÓN

La **estimación puntual** de un parámetro θ es un número único que puede ser considerado un valor sensible de θ. Se obtiene una estimación puntual al seleccionar un estadístico apropiado y calcular su valor con los datos muestrales dados. El estadístico seleccionado se llama **estimador puntual** de θ.

En el ejemplo de las baterías que hemos visto el estimador utilizado para obtener la estimación puntual de μ fue $\overline{X}$, y la estimación puntual de μ fue 5.77. Si las tres vidas útiles hubieran sido $x_1 = 5.6$, $x_2 = 4.5$ y $x_3 = 6.1$, el uso del estimador $\overline{X}$ habría dado por resultado la estimación $\bar{x} = (5.6 + 4.5 + 6.1)/3 = 5.40$. El símbolo $\hat{\theta}$ ("teta testada") se utiliza comúnmente para denotar tanto la estimación de θ como la estimación puntual que resulta de una muestra dada.* Por tanto, $\hat{\mu} = \overline{X}$ se lee como "el estimador puntual de μ

*Siguiendo la primera notación, se podría utilizar $\hat{\Theta}$ (una teta mayúscula) para el estimador, pero es difícil de escribir.

es la media muestral $\overline{X}$". El enunciado "la estimación puntual de μ es 5.77" se escribe concisamente como $\hat{\mu} = 5.77$. Observe que cuando se escribe $\hat{\theta} = 72.5$, no hay ninguna indicación de cómo se obtuvo esta estimación puntual (qué estadístico se utilizó). Se recomienda reportar tanto el estimador como la estimación resultante.

EJEMPLO 5.1 Un fabricante automotriz ha producido un nuevo tipo de parachoques, del cual se dice que absorbe impactos con menos daño que diseños anteriores. El fabricante lo ha utilizado en una secuencia de 25 choques controlados contra un muro, cada uno a 10 mph, utilizando uno de sus modelos de automóvil compacto. Sea X = el número de choques que no provocaron daños visibles al automóvil. El parámetro que deberá ser estimado es p = la proporción de todos los choques que no provocaron daños [alternativamente, $p = P$(ningún daño en un choque)]. Si se observa que X es $x = 15$, el estimador y la estimación más razonables son

$$\text{estimador } \hat{p} = \frac{X}{n} \qquad \text{estimación } = \frac{x}{n} = \frac{15}{25} = 0.60$$

■

Si cada parámetro de interés tuviera sólo un estimador puntual razonable, no tendría mucho caso la estimación puntual. En la mayoría de los problemas, sin embargo, habrá más de un estimador razonable.

EJEMPLO 5.2 Considere 20 observaciones adjuntas de voltaje de ruptura dieléctrica de piezas de resina epóxica.

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

El patrón en la gráfica de probabilidad normal dado ahí es bastante recto, así que ahora se supone que la distribución de voltaje de ruptura es normal con valor de la media μ. Puesto que las distribuciones normales son simétricas, μ también es la vida útil mediana de la distribución. Se supone entonces que las observaciones dadas son el resultado de una muestra aleatoria $X_1, X_2, \ldots, X_{20}$ de esta distribución normal. Considere los siguientes estimadores y las estimaciones resultantes para μ:

a. Estimador = $\overline{X}$, estimación = $\bar{x} = \Sigma x_i/n = 555.86/20 = 27.793$

b. Estimador = $\widetilde{X}$, estimación = $\tilde{x} = (27.94 + 27.98)/2 = 27.960$

c. Estimador = $[\text{mín}(X_i) + \text{máx}(X_i)]/2$ = el promedio de las dos vidas útiles extremas, estimación = $[\text{mín}(x_i) + \text{máx}(x_i)]/2 = (24.46 + 30.88)/2 = 27.670$

d. Estimador = $\overline{X}_{\text{rec}(10)}$, la media recortada 10% (desechar el más pequeño 10% y el más grande de la muestra y luego promediar),

$$\begin{aligned}\text{estimación} &= \bar{x}_{\text{tr}(10)} \\ &= \frac{555.86 - 24.46 - 25.61 - 29.50 - 30.88}{16} \\ &= 27.838\end{aligned}$$

Cada uno de los estimadores (a)–(d) utiliza una medición diferente del centro de la muestra para estimar μ. ¿Cuál de las estimaciones se acerca más al valor verdadero? No se puede responder esta pregunta sin conocer el valor verdadero. Una pregunta que puede responderse es: "¿Cuándo se utiliza en otras muestras de X_i, cuál estimador tiende a producir estimaciones cercanas al valor verdadero? Más adelante se considerará este tipo de pregunta. ■

EJEMPLO 5.3 El artículo **"Is a Normal Distribution the Most Appropriate Statistical Distribution for Volumetric Properties in Asphalt Mixtures?"** ya citado en el ejemplo 4.26, informa de las siguientes observaciones sobre X = *vacíos llenos de asfalto* (%) de 52 muestras de un cierto tipo de mezcla caliente de asfalto:

74.33	71.07	73.82	77.42	79.35	82.27	77.75	78.65	77.19
74.69	77.25	74.84	60.90	60.75	74.09	65.36	67.84	69.97
68.83	75.09	62.54	67.47	72.00	66.51	68.21	64.46	64.34
64.93	67.33	66.08	67.31	74.87	69.40	70.83	81.73	82.50
79.87	81.96	79.51	84.12	80.61	79.89	79.70	78.74	77.28
79.97	75.09	74.38	77.67	83.73	80.39	76.90		

Estimemos la varianza σ^2 de la distribución de la población. Un estimador natural es la varianza de la muestra:

$$\hat{\sigma}^2 = S^2 = \frac{\sum(X_i - \overline{X})^2}{n - 1}$$

Minitab arroja el siguiente resultado a una petición para mostrar los estadísticos descriptivos:

Variable	Count	Mean	SE Mean	StDev	Variance	Q1	Median	Q3
VFA(B)	52	73.880	0.889	6.413	41.126	67.933	74.855	79.470

Por tanto, la estimación puntual de la varianza de la población es

$$\hat{\sigma}^2 = s^2 = \frac{\sum(x_i - \bar{x})^2}{52 - 1} = 41.126$$

[de forma alternativa, la fórmula de cálculo para el numerador de s^2 da

$$S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2/n = 285\,929.5964 - (3841.78)^2/52 = 2097.4124].$$

Una estimación puntual de la desviación estándar de la población es entonces $\hat{\sigma} = s = \sqrt{41.126} = 6.413$.

Un estimador alternativo resulta de usar el divisor n en lugar de $n - 1$:

$$\hat{\sigma}^2 = \frac{\sum(X_i - \overline{X})^2}{n}, \qquad \text{estimación} = \frac{2097.4124}{52} = 40.335$$

En breve se indicará por qué muchos estadísticos prefieren S^2 en lugar de este último estimador.

El citado artículo considera ajustar cuatro distribuciones diferentes a los datos: normal, lognormal, de dos parámetros de Weibull y de tres parámetros de Weibull. Se utilizaron diferentes técnicas para concluir que los dos parámetros de Weibull proporcionan el mejor ajuste (el gráfico de probabilidad normal de los datos muestra alguna desviación de un patrón lineal). De la sección 4.5 la varianza de una variable aleatoria de Weibull es

$$\sigma^2 = \beta^2\{\Gamma(1 + 2/\alpha) - [\Gamma(1 + 1/\alpha)]^2\}$$

donde α y β son los parámetros de forma y escala de la distribución. Los autores del artículo utilizaron el método de máxima verosimilitud (probabilidad) (véase la sección 5.2) para estimar estos parámetros. Las estimaciones resultantes son $\hat{\alpha} = 11.9731$, $\hat{\beta} = 77.0153$. Una estimación razonable de la varianza de la población ahora se puede obtener al sustituir las estimaciones de los dos parámetros en la expresión para σ^2; el resultado es $\hat{\sigma}^2 = 56.035$. Esta última estimación es obviamente muy diferente de la varianza de la muestra. Su validez depende de que la distribución de la población sea Weibull, mientras que la varianza de la muestra es una manera sensata para estimar σ^2 cuando hay incertidumbre en cuanto a la forma específica de la distribución de la población. ■

En el mejor de todos los mundos posibles, se podría hallar un estimador $\hat{\theta}$ con el cual siempre $\hat{\theta} = \theta$. Sin embargo, $\hat{\theta}$ es una función de las X_i muestrales, así que es una variable aleatoria. Con algunas muestras, $\hat{\theta}$ dará un valor más grande que θ, mientras que con otras muestras $\hat{\theta}$ subestimará θ. Si escribimos

$$\hat{\theta} = \theta + \text{error de estimación}$$

entonces un estimador preciso sería uno que produjera errores de estimación pequeños, de manera que los valores estimados estén cerca del valor verdadero.

Una forma sensible de cuantificar la idea de $\hat{\theta}$ cercano a θ es considerar el error cuadrático $(\hat{\theta} - \theta)^2$. Con algunas muestras, $\hat{\theta}$ se acercará bastante a θ y el error cuadrático resultante se aproximará a 0. Otras muestras pueden dar valores de $\hat{\theta}$ alejados de θ, correspondientes a errores cuadráticos muy grandes. Una medida general de precisión es la esperanza o *error cuadrático medio* (MSE por sus siglas en inglés) $\text{MSE} = E[(\hat{\theta} - \theta)^2]$. Si un primer estimador tiene una MSE más pequeña que un segundo, es natural decir que el primer estimador es el mejor. Sin embargo, la MSE en general dependerá del valor de θ. Lo que a menudo sucede es que un estimador tiene un MSE más pequeño con algunos valores de θ y un MSE más grande con otros valores. En general no es posible determinar un estimador con el MSE más pequeño.

Una forma de librarse de este dilema es limitar la atención sólo en estimadores que tengan una propiedad deseable específica y luego determinar el mejor estimador en este grupo limitado. Una propiedad popular de esta clase en la comunidad estadística es la *ausencia de sesgo*.

Estimadores insesgados

Suponga que se tienen dos instrumentos de medición: uno ha sido calibrado con precisión, pero el otro sistemáticamente da lecturas más pequeñas que las reales de lo que se está midiendo. Cuando cada uno de los instrumentos se utiliza repetidamente en el mismo objeto, debido al error de medición, las mediciones observadas no serán idénticas. Sin embargo, las mediciones producidas por el primer instrumento se distribuirán en torno al valor verdadero, de tal modo que en promedio este instrumento medirá lo que se proponga medir, por lo que se le conoce como instrumento insesgado. El segundo instrumento proporciona observaciones que tienen un componente de error o sesgo sistemático.

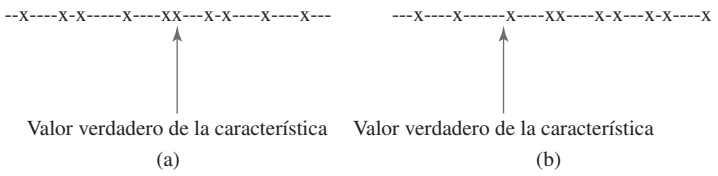


Figura 5.1 Medidas de (a) un estimador sesgado y (b) un estimador insesgado

DEFINICIÓN

Se dice que un estimador puntual $\hat{\theta}$ es un **estimador insesgado** de θ si $E(\hat{\theta}) = \theta$ para todo valor posible de θ. Si $\hat{\theta}$ es insesgado, la diferencia $E(\hat{\theta}) - \theta$ se conoce como el **sesgo** de $\hat{\theta}$.

Es decir, $\hat{\theta}$ es sesgado si su distribución de probabilidad (es decir, su muestreo) siempre está "centrada" en el valor verdadero del parámetro. Suponga que $\hat{\theta}$ es un estimador insesgado; entonces si $\theta = 100$ la distribución muestral $\hat{\theta}$ está centrada en 100; si $\theta = 27.5$ la distribución muestral $\hat{\theta}$ está centrada en 27.5, y así sucesivamente. La figura 5.2 ilustra la distribución de varios estimadores sesgados e insesgados. Observe que "centrada" en este caso significa que el valor esperado, no la mediana, de la distribución de $\hat{\theta}$ es igual a θ.

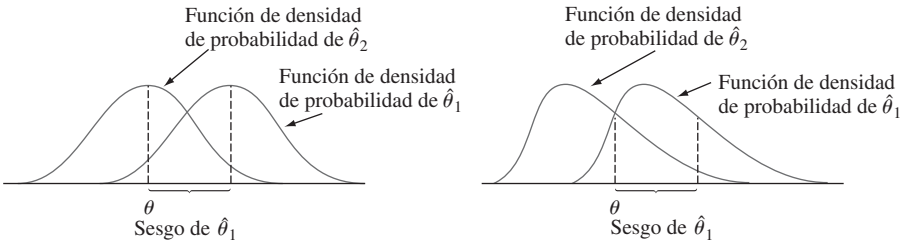


Figura 5.2 Funciones de densidad de probabilidad de un estimador sesgado $\hat{\theta}_1$ y un estimador insesgado $\hat{\theta}_2$, de un parámetro θ

Parece como si fuera necesario conocer el valor de θ (en cuyo caso la estimación es innecesaria) para ver si $\hat{\theta}$ es insesgado. Este no suele ser el caso, sin embargo, porque la ausencia de sesgo es una propiedad general de la distribución muestral donde el muestreo del estimador está centrado lo que por lo general no depende de ningún valor de parámetro en particular.

En el ejemplo 5.1 se utilizó la proporción muestral X/n como estimador de p, donde X, el número de éxitos muestrales, tenía una distribución binomial con parámetros n y p. Por tanto,

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

PROPOSICIÓN

Cuando X es una variable aleatoria binomial con parámetros n y p, la proporción muestral $\hat{p} = X/n$ es un estimador sesgado de p.

No importa cuál sea el valor verdadero de p, la distribución del estimador $\hat{p}$ estará centrada en el valor verdadero.

EJEMPLO 5.4 Suponga que X, el tiempo de reacción a un estímulo, tiene una distribución uniforme en el intervalo desde 0 hasta un límite superior desconocido θ (por tanto, la función de densidad de X es de forma rectangular con una altura de $1/\theta$ en el intervalo $0 \le x \le \theta$). Se desea estimar θ con base en una muestra aleatoria $X_1, X_2, \ldots, X_n$ de los tiempos de reacción. Como θ es el mayor tiempo posible en toda la población de tiempos de reacción, considere como un primer estimador el mayor tiempo de reacción muestral $\hat{\theta}_1 = \text{máx}(X_1, \ldots, X_n)$. Si $n = 5$ y $x_1 = 4.2$, $x_2 = 1.7$, $x_3 = 2.4$, $x_4 = 3.9$ y $x_5 = 1.3$, la estimación puntual de θ es $\hat{\theta}_1 = \text{máx}(4.2, 1.7, 2.4, 3.9, 1.3) = 4.2$.

La ausencia de sesgo implica que algunas muestras arrojarán estimaciones que exceden θ y otras que darán estimaciones más pequeñas que θ, de lo contrario posiblemente θ no podría ser el centro (punto de equilibrio) de la distribución de $\hat{\theta}_1$. Sin embargo, el estimador propuesto nunca sobreestimará θ (el valor muestral más grande no puede exceder el valor de la población más grande) y subestimará θ a menos que el valor muestral más grande sea igual a θ. Este argumento intuitivo demuestra que $\hat{\theta}_1$ es un estimador insesgado. Más precisamente (véase el ejercicio 32) se puede demostrar que

$$E(\hat{\theta}_1) = \frac{n}{n+1} \cdot \theta < \theta \qquad \left(\text{puesto que } \frac{n}{n+1} < 1\right)$$

El sesgo de $\hat{\theta}$ está dado por $n\theta/(n+1) - \theta = -\theta/(n+1)$, el cual tiende a cero a medida que n se hace grande.

Es fácil modificar $\hat{\theta}_1$ para obtener un estimador insesgado de θ. Considere el estimador

$$\hat{\theta}_2 = \frac{n+1}{n} \cdot \text{ máx } (X_1, \ldots, X_n)$$

Si se utiliza este estimador en los datos se obtiene la estimación $(6/5)(4.2) = 5.04$. El hecho de que $(n+1)/n > 1$ implica que $\hat{\theta}_2$ sobreestimará θ para algunas muestras y la subestimará en otras. La media de este estimador es

$$E(\hat{\theta}_2) = E\left[\frac{n+1}{n} \text{máx}(X_1, \ldots, X_n)\right] = \frac{n+1}{n} \cdot E\left[\text{máx}(X_1, \ldots, X_n)\right]$$
$$= \frac{n+1}{n} \cdot \frac{n}{n+1}\theta = \theta$$

Si $\hat{\theta}_2$ se utiliza repetidamente en diferentes muestras para estimar θ, algunas estimaciones serán demasiado grandes y otras demasiado pequeñas, pero a la larga no habrá ninguna tendencia simétrica para subestimar o sobreestimar θ. ■

Principio de estimación insesgada

Cuando se elige entre varios estimadores diferentes de θ, se elige uno insesgado.

De acuerdo con este principio, el estimador insesgado $\hat{\theta}_2$ en el ejemplo 5.4 deberá ser preferido al estimador sesgado $\hat{\theta}_1$. Considere ahora el problema de estimar σ^2.

PROPOSICIÓN

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con media μ y varianza σ^2. Entonces el estimador

$$\hat{\sigma}^2 = S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$$

es un estimador insesgado de σ^2.

Demostración Para cualquier variable aleatoria Y, $V(Y) = E(Y^2) - [E(Y)]^2$, por tanto, $E(Y^2) = V(Y) - [E(Y)]^2$. Aplicando esto a

$$S^2 = \frac{1}{n-1}\left[\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}\right]$$

se obtiene

$$E(S^2) = \frac{1}{n-1}\left\{\sum E\left(X_i^2\right) - \frac{1}{n}E\left[\left(\sum X_i\right)^2\right]\right\}$$
$$= \frac{1}{n-1}\left\{\sum(\sigma^2 + \mu^2) - \frac{1}{n}\left\{V\left(\sum X_i\right) + \left[E\left(\sum X_i\right)\right]^2\right\}\right\}$$
$$= \frac{1}{n-1}\left\{n\sigma^2 + n\mu^2 - \frac{1}{n}n\sigma^2 - \frac{1}{n}(n\mu)^2\right\}$$
$$= \frac{1}{n-1}\{n\sigma^2 - \sigma^2\} = \sigma^2 \quad \text{(como se desea)}$$ ■

El estimador que utiliza el divisor n se expresa como $(n-1)S^2/n$, por tanto,

$$E\left[\frac{(n-1)S^2}{n}\right] = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}\sigma^2$$

Este estimador es, por consiguiente, sesgado. El sesgo es $(n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n$. Debido a que el sesgo es negativo, el estimador con divisor n tiende a subestimar σ^2 y por eso muchos estadísticos prefieren el divisor $n-1$ (aunque cuando n es grande, el sesgo es pequeño y hay poca diferencia entre ambos).

Lamentablemente, el hecho de que S^2 sea insesgado para la estimación de σ^2 no implica que S sea insesgado para la estimación de σ. Sacar la raíz cuadrada estropea la propiedad de ausencia de sesgo (el valor esperado de la raíz cuadrada no es la raíz cuadrada del valor esperado). Afortunadamente, el sesgo de S es pequeño a menos que n sea muy pequeño. Hay otras buenas razones para utilizar S como estimador, especialmente cuando la distribución de la población es normal. Esto se volverá más aparente cuando se aborden los intervalos de confianza y la prueba de hipótesis en los siguientes capítulos.

En el ejemplo 5.2 se propusieron varios estimadores diferentes de la media μ de una distribución normal. Si hubiera un estimador insesgado único para μ, el problema de la estimación se resolvería utilizando dicho estimador. Lamentablemente, éste no es el caso.

PROPOSICIÓN

Si $X_1, X_2, \ldots, X_n$ son una muestra aleatoria tomada de una distribución con media μ, entonces $\overline{X}$ es un estimador sesgado de μ. Si además la distribución es continua y simétrica, entonces $\widetilde{X}$ y cualquier media recortada también son estimadores insesgados de μ.

El hecho de que $\overline{X}$ sea insesgado es simplemente el replanteamiento de una de las reglas de valor esperado: $E(\overline{X}) = \mu$ para cada valor posible de μ (para distribuciones discretas y continuas). La ausencia de sesgo de los demás estimadores es más difícil de verificar.

El siguiente ejemplo introduce otra situación en la cual existen varios estimadores insesgados para un parámetro particular.

EJEMPLO 5.5 En ciertas circunstancias los contaminantes orgánicos se adhieren con facilidad a la superficie de las láminas y deterioran los dispositivos de fabricación de los semiconductores. El artículo **"Ceramic Chemical Filter for Removal of Organic Contaminants" (*J. of the Institute of Environmental Sciences and Technology*, 2003: 59–65)** menciona una alternativa, recientemente desarrollada, de filtros de carbón convencionales para eliminar la contaminación molecular del aire en aplicaciones de salas limpias. Un aspecto de la investigación del desempeño de los filtros implicó estudiar cómo se relaciona la concentración de contaminantes en el aire con la concentración en la superficie de una lámina luego de una exposición prolongada. Considere los siguientes datos representativos de x = concentración de DBP en el aire y y = concentración de DBP en la superficie de las láminas luego de 4 horas de exposición (ambas en $\mu g/m^3$, donde DBP = ftalato de dibutilo).

Láminas	i:	1	2	3	4	5	6
	x:	0.8	1.3	1.5	3.0	11.6	26.6
	y:	0.6	1.1	4.5	3.5	14.4	29.1

Los autores argumentan que la "adhesión de DBP en la superficie de láminas fue aproximadamente proporcional a la concentración de DBP en el aire". La figura 5.3 muestra una gráfica de y contra x, es decir, de pares (x, y).

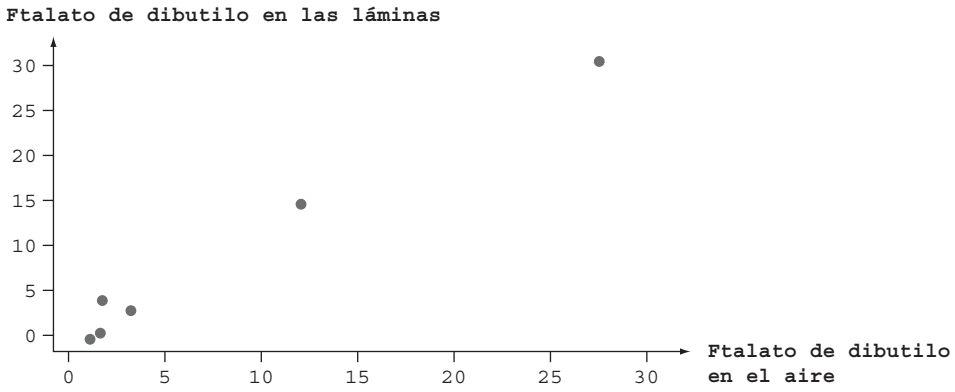


Figura 5.3 Gráfica de los datos de ftalato de dibutilo del ejemplo 5.5

Si y fuera exactamente proporcional a x, se tendría $y = \beta x$ para algún valor β lo cual expresa que los puntos (x, y) en la gráfica quedarían exactamente sobre una línea recta con pendiente β que pasa por (0, 0). Pero esto sólo es aproximadamente el caso. Así que a continuación se supone que para cualquier x fija la concentración de DBP en las láminas es una variable aleatoria Y con media βx. Es decir, se postula que la *media* de Y está relacionada con x mediante una recta que pasa por (0, 0) pero que el valor observado de Y en general se desviará de esta recta (esto se conoce en la literatura estadística como "regresión a través del origen").

Considere los siguientes tres estimadores para el parámetro de la pendiente β:

$$\#1\colon \hat{\beta} = \frac{1}{n}\sum\frac{Y_i}{x_i} \quad \#2\colon \hat{\beta} = \frac{\sum Y_i}{\sum x_i} \quad \#3\colon \hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$$

Las estimaciones resultantes basadas en los datos dados son 1.3497, 1.1875 y 1.1222, respectivamente. Así que de manera definitiva la estimación depende de qué estimador se utilice. Si uno de estos tres estimadores fuera insesgado y los otros dos fueran sesgados habría un buen motivo para utilizar el insesgado. Pero los tres son insesgados; el argumento se apoya en el hecho de que cada uno es una función lineal de las Y_i (aquí se supone que las x_i son fijas, no aleatorias):

$$E\left(\frac{1}{n}\sum\frac{Y_i}{x_i}\right) = \frac{1}{n}\sum\frac{E(Y_i)}{x_i} = \frac{1}{n}\sum\frac{\beta x_i}{x_i} = \frac{1}{n}\sum\beta = \frac{n\beta}{n} = \beta$$

$$E\left(\frac{\sum Y_i}{\sum x_i}\right) = \frac{1}{\sum x_i}E\left(\sum Y_i\right) = \frac{1}{\sum x_i}\left(\sum\beta x_i\right) = \frac{1}{\sum x_i}\beta\left(\sum x_i\right) = \beta$$

$$E\left(\frac{\sum x_i Y_i}{\sum x_i^2}\right) = \frac{1}{\sum x_i^2}E\left(\sum x_i Y_i\right) = \frac{1}{\sum x_i^2}\left(\sum x_i\beta x_i\right) = \frac{1}{\sum x_i^2}\beta\left(\sum x_i^2\right) = \beta$$

■

Tanto en el ejemplo anterior como en la situación que implica estimar una media de población normal, no puede ser invocado el principio de ausencia de sesgo (preferir un estimador insesgado a uno sesgado) para seleccionar un estimador. Lo que ahora se requiere es un criterio para elegir entre estimadores sesgados.

Estimadores con varianza mínima

Suponga que $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos estimadores de θ insesgados. Entonces, aunque la distribución de cada estimador esté centrada en el valor verdadero de θ, las dispersiones de las distribuciones en torno al valor verdadero pueden ser diferentes.

Principio de estimación insesgada con varianza mínima

Entre todos los estimadores de θ insesgados se selecciona el de varianza mínima. El $\hat{\theta}$ resultante se llama **estimador insesgado con varianza mínima** (**MVUE** por sus siglas en inglés) de θ.

La figura 5.4(a) muestra las distribuciones de dos diferentes estimadores insesgados. Es más probable el uso del estimador con la distribución más concentrada en lugar de otros para producir una estimación más cercana a θ. La figura 5.4(b) muestra las estimaciones de dos estimadores basados en 10 muestras diferentes. El MVUE es, en cierto sentido, el que tiene más probabilidades entre todos los estimadores insesgados para producir una estimación cercana al θ verdadero.

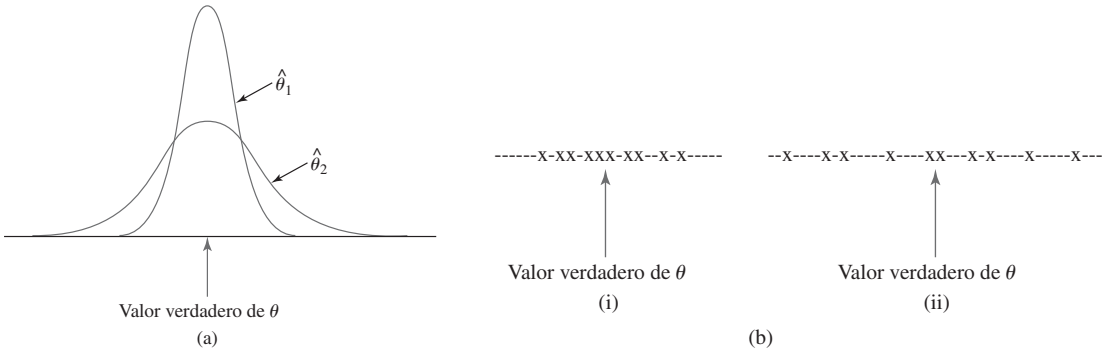


Figura 5.4 (a) Distribuciones de dos estimadores insesgados diferentes (b) Estimaciones basadas en 10 muestras diferentes.

En el ejemplo 5.5, suponga que cada Y_i está normalmente distribuida con media βx_i y varianza σ^2 (se asume una varianza constante). Entonces se puede demostrar que el tercer estimador $\hat{\beta} = \Sigma x_i Y_i / \Sigma x_i^2$ no sólo tiene una varianza más pequeña que cualquiera de los otros dos estimadores insesgados, sino que de hecho es el estimador insesgado con varianza mínima; tiene una varianza más pequeña que *cualquier* otro estimador insesgado de β.

EJEMPLO 5.6 En el ejemplo 5.4 se argumentó que cuando $X_1, X_2, \ldots, X_n$ es una muestra aleatoria tomada de una distribución uniforme en el intervalo $[0, \theta]$, el estimador

$$\hat{\theta}_1 = \frac{n+1}{n} \cdot \text{máx}\,(X_1, \ldots, X_n)$$

es insesgado para θ (previamente este estimador se denotó con $\hat{\theta}_2$). Este no es el único estimador insesgado de θ. El valor esperado de una variable aleatoria uniformemente distribuida es simplemente el punto medio del intervalo de densidad positiva, por tanto, $E(X_i) = \theta/2$. Esto implica que $E(\overline{X}) = \theta/2$, a partir de la cual $E(2\overline{X}) = \theta$. Es decir, el estimador $\hat{\theta}_2 = 2\overline{X}$ es insesgado para θ.

Si X está uniformemente distribuida en el intervalo de A a B, entonces $V(X) = \sigma^2 = (B - A)^2/12$. Así pues, en esta situación, $V(X_i) = \theta^2/12$, $V(\overline{X}) = \sigma^2/n = \theta^2/(12n)$ y $V(\hat{\theta}_2) = V(2\overline{X}) = 4V(\overline{X}) = \theta^2/(3n)$. Se pueden utilizar los resultados del ejercicio 32 para demostrar que $V(\hat{\theta}_1) = \theta^2/[n(n+2)]$. El estimador $\hat{\theta}_1$ tiene una varianza más pequeña que $\hat{\theta}_2$ si $3n < n(n+2)$; es decir, si $0 < n^2 - n = n(n-1)$. Mientras que $n > 1$, $V(\hat{\theta}_1) < V(\hat{\theta}_2)$, así que $\hat{\theta}_1$ es mejor estimador que $\hat{\theta}_2$. Se pueden utilizar métodos más avanzados para demostrar que $\hat{\theta}_1$ es el estimador insesgado con varianza mínima de θ; cualquier otro estimador insesgado de θ tiene una varianza que excede a $\theta^2/[n(n+2)]$. ■

Uno de los triunfos de la estadística matemática ha sido el desarrollo de una metodología para identificar el estimador insesgado con varianza mínima en una amplia variedad de situaciones. Para nuestros propósitos el resultado más importante de este tipo tiene que ver con la estimación de la media μ de una distribución normal.

TEOREMA

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria tomada de una distribución normal con parámetros μ y σ. Entonces el estimador $\hat{\mu} = \overline{X}$ es el estimador insesgado con varianza mínima para μ.

Siempre que exista la seguridad de que la población que se está muestreando es normal, el teorema dice que debe usarse $\overline{x}$ para estimar μ. Entonces, en el ejemplo 5.2 la estimación sería $\overline{x} = 27.793$.

En algunas situaciones es posible obtener un estimador con sesgo pequeño que se preferirá al mejor estimador insesgado. Esto se ilustra en la figura 5.5. Sin embargo, los estimadores insesgados con varianza mínima a menudo son más fáciles de obtener que el tipo de estimador sesgado cuya distribución se ilustra.

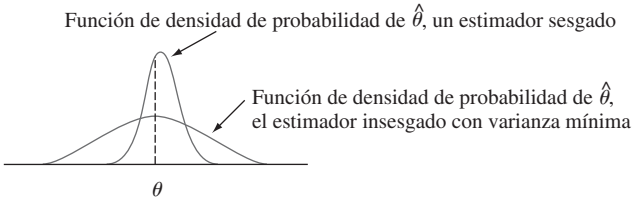


Figura 5.5 Un estimador sesgado que es preferible al estimador insesgado con varianza mínima

Algunas complicaciones

El último teorema no dice que al estimar la media μ de una población se debe utilizar el estimador $\overline{X}$, independientemente de la distribución que se está muestreando.

EJEMPLO 5.7 Suponga que se desea estimar la conductividad térmica μ de un material. Con técnicas de medición estándar se obtendrá una muestra aleatoria $X_1, \ldots, X_n$ de n mediciones de conductividad térmica. Suponga que la distribución de la población es miembro de una de las siguientes tres familias:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty \tag{5.1}$$

$$f(x) = \frac{1}{\pi\beta[1 + ((x-\mu)/\beta)^2]} \quad -\infty < x < \infty \tag{5.2}$$

$$f(x) = \begin{cases} \dfrac{1}{2c} & \mu - c \le x \le \mu + c \\ 0 & \text{de lo contrario} \end{cases} \tag{5.3}$$

La función de densidad de probabilidad (5.1) es la distribución normal, a (5.2) se le llama distribución de Cauchy y (5.3) es una distribución uniforme. Las tres distribuciones son simétricas respecto a μ y, de hecho, la distribución de Cauchy tiene forma de campana pero con colas mucho más gruesas (más probabilidad hacia fuera) que la curva normal. La distribución uniforme no tiene colas. Los cuatro estimadores de μ considerados con anterioridad son $\overline{X}$, $\widetilde{X}$, $\overline{X}_e$ (el promedio de las dos observaciones extremas) y $\overline{X}_{\text{rec}(10)}$, una media recortada.

La muy importante moraleja en este caso es que el mejor estimador de μ depende crucialmente de qué distribución está siendo muestreada. En particular,

1. Si la muestra aleatoria proviene de una distribución normal, $\overline{X}$ es el mejor de los cuatro estimadores, puesto que tiene una varianza mínima entre todos los estimadores insesgados.
2. Si la muestra aleatoria proviene de una distribución de Cauchy, entonces $\overline{X}$ y $\overline{X}_e$ son estimadores terribles de μ, en tanto que $\widetilde{X}$ es bastante bueno (el estimador insesgado con varianza mínima no es conocido); $\overline{X}$ es malo porque es muy sensible a las observaciones subyacentes y las colas gruesas de la distribución de Cauchy hacen que sea improbable que aparezcan tales observaciones en cualquier muestra.
3. Si la distribución subyacente es uniforme, el mejor estimador es $\overline{X}_e$; este estimador está influido en gran medida por las observaciones subyacentes, pero la carencia de colas hace que tales observaciones sean imposibles.
4. *En ninguna de estas tres situaciones es mejor la media recortada pero funciona razonablemente bien en las tres*. Es decir, $\overline{X}_{\text{rec}(10)}$ no sufre demasiado en comparación con el mejor procedimiento en cualquiera de las tres situaciones. ■

Más generalmente, investigaciones recientes en estadística han establecido que cuando se estima un punto de simetría μ de una distribución de probabilidad continua, una media recortada con proporción de recorte de 10 o 20% (para cada extremo de la muestra) produce estimaciones razonablemente comportadas dentro de un rango muy amplio de posibles modelos. Por esta razón se dice que una media recortada con poco porcentaje de recorte es un **estimador robusto**.

En algunas situaciones, la selección no es entre dos estimadores diferentes construidos con la misma muestra, sino entre estimadores basados en dos experimentos distintos.

EJEMPLO 5.8 Suponga que cierto tipo de componente tiene una distribución de vida útil exponencial con parámetro λ, de modo que la vida útil esperada es $\mu = 1/\lambda$. Se selecciona una muestra n de esos componentes y cada uno es puesto en operación. Si el experimento continúa hasta que todas las n vidas útiles $X_1, \ldots, X_n$ han sido observadas, $\overline{X}$ es un estimador insesgado de μ.

En algunos experimentos, sin embargo, los componentes se dejan en operación sólo hasta el tiempo de la r-ésima falla donde $r < n$. Este procedimiento se conoce como **censura**. Sea Y_1 el tiempo de la primera falla (la vida útil mínima entre los n componentes) y Y_2 el tiempo en el cual ocurre la segunda falla (la segunda vida útil más pequeña) y así sucesivamente. Puesto que el experimento termina en el tiempo Y_r, la vida útil acumulada al final es

$$T_r = \sum_{i=1}^{r} Y_i + (n - r)Y_r$$

A continuación se demuestra que $\hat{\mu} = T_r/r$ es un estimador insesgado de μ. Para hacerlo se requieren dos propiedades de las variables exponenciales:

1. La propiedad de no memoria (véase la sección 4.4), la cual dice que en cualquier punto de tiempo, la vida útil restante tiene la misma distribución exponencial que la vida útil original.
2. Si $X_1, \ldots, X_k$ son independientes, cada mín$(X_1, \ldots, X_k)$ exponencialmente distribuida con parámetro λ, es exponencial con parámetro $k\lambda$.

A partir de que los n componentes duran hasta Y_1, $n - 1$ duran una cantidad de tiempo $Y_2 - Y_1$, $n - 2$ adicional duran una cantidad de tiempo $Y_3 - Y_2$ adicional, y así sucesivamente, otra expresión para T_r es

$$\begin{aligned} T_r = {} & nY_1 + (n-1)(Y_2 - Y_1) + (n-2)(Y_3 - Y_2) + \cdots \\ & + (n - r + 1)(Y_r - Y_{r-1}) \end{aligned}$$

Pero Y_1 es la mínima de n variables exponenciales, por tanto $E(Y_1) = 1/(n\lambda)$. Asimismo, $Y_2 - Y_1$ es la más pequeña de las $n - 1$ vidas útiles restantes, cada una exponencial

con parámetro λ (por la propiedad de no memoria), así que $E(Y_2 - Y_1) = 1/[(n-1)\lambda]$. Continuando, $E(Y_{i+1} - Y_i) = 1/[(n-i)\lambda]$ por lo que

$$\begin{aligned} E(T_r) &= nE(Y_1) + (n-1)E(Y_2 - Y_1) + \cdots + (n-r+1)E(Y_r - Y_{r-1}) \\ &= n \cdot \frac{1}{n\lambda} + (n-1) \cdot \frac{1}{(n-1)\lambda} + \cdots + (n-r+1) \cdot \frac{1}{(n-r+1)\lambda} \\ &= \frac{r}{\lambda} \end{aligned}$$

Por consiguiente, $E(T_r/r) = (1/r)E(T_r) = (1/r) \cdot (r/\lambda) = 1/\lambda = \mu$ como se proponía.

Por ejemplo, suponga que se prueban 20 componentes y $r = 10$. Si los primeros diez tiempos de falla son 11, 15, 29, 33, 35, 40, 47, 55, 58 y 72 la estimación de μ es

$$\hat{\mu} = \frac{11 + 15 + \cdots + 72 + (10)(72)}{10} = 111.5$$

La ventaja del experimento con censura es que termina más rápido que el experimento sin censura. Sin embargo, se puede demostrar que $V(T_r/r) = 1/(\lambda^2 r)$, que es más grande que $1/(\lambda^2 n)$, la varianza de $\overline{X}$ en el experimento sin censura. ■

Reporte de una estimación puntual: El error estándar

Además de reportar el valor de una estimación puntual se debe dar alguna indicación de su precisión. La medición usual de precisión es el error estándar del estimador usado.

DEFINICIÓN

El **error estándar** de un estimador $\hat{\theta}$ es su desviación estándar $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. Esta es la magnitud de una desviación típica o representativa entre una estimación y el valor de θ. Si el error estándar implica parámetros desconocidos cuyos valores pueden ser estimados, la sustitución de estas estimaciones en $\sigma_{\hat{\theta}}$ da el **error estándar estimado** (desviación estándar estimada) del estimador. El error estándar estimado puede ser denotado por $\hat{\sigma}_{\hat{\theta}}$ (el ˆ sobre σ recalca que $\sigma_{\hat{\theta}}$ está siendo estimada) o por $s_{\hat{\theta}}$.

EJEMPLO 5.9 (Continuación del ejemplo 5.2)

Suponiendo que el voltaje de ruptura está normalmente distribuido, $\hat{\mu} = \overline{X}$ es la mejor estimación de μ. Si se sabe que el valor de σ es 1.5, el error estándar de $\overline{X}$ es $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 1.5/\sqrt{20} = 0.335$. Si, como casi siempre es el caso, el valor de σ es desconocido, la estimación $\hat{\sigma} = s = 1.462$ se sustituye en $\sigma_{\overline{X}}$ para obtener el error estándar estimado $\hat{\sigma}_{\overline{X}} = s_{\overline{X}} = s/\sqrt{n} = 1.462/\sqrt{20} = 0.327$. ■

EJEMPLO 5.10 (Continuación del ejemplo 5.1)

El error estándar de $\hat{p} = X/n$ es

$$\sigma_{\hat{p}} = \sqrt{V(X/n)} = \sqrt{\frac{V(X)}{n^2}} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}$$

A partir de que p y $q = 1 - p$ son desconocidas (¿qué otra razón para estimarlas?), se sustituyen $\hat{p} = x/n$ y $\hat{q} = 1 - x/n$ en $\sigma_{\hat{p}}$, para obtener el error estándar estimado $\hat{\sigma}_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n} = \sqrt{(0.6)(0.4)/25} = 0.098$. Alternativamente, puesto que el valor más grande de pq se obtiene cuando $p = q = 0.5$, un límite superior en el error estándar es $\sqrt{1/(4n)} = 0.10$. ■

Cuando la distribución del estimador puntual $\hat{\theta}$ es aproximadamente normal, lo cual será frecuente cuando n sea grande, entonces se puede confiar de manera razonable en que

el valor verdadero de θ queda dentro de aproximadamente 2 errores estándar (desviaciones estándar) de $\hat{\theta}$. De este modo, si una muestra de $n = 36$ vidas útiles de componentes da $\hat{\mu} = \bar{x} = 28.50$ y $s = 3.60$, entonces $s/\sqrt{n} = 0.60$, por tanto, dentro de dos errores estándar estimados, $\hat{\mu}$ se traslada al intervalo $28.50 \pm (2)(.60) = (27.30, 29.70)$

Si $\hat{\theta}$ no necesariamente es aproximadamente normal, pero es insesgado, entonces se puede demostrar que la estimación se desviará de θ hasta por 4 errores estándar cuando mucho 6% del tiempo. Se esperaría entonces que el valor verdadero quedara dentro de 4 errores estándar de $\hat{\theta}$ (y esta es una afirmación muy conservadora, puesto que se aplica a cualquier $\hat{\theta}$ insesgado). En resumen, el error estándar indica aproximadamente a qué distancia de $\hat{\theta}$ se puede esperar que quede el valor verdadero de θ.

La forma del estimador $\hat{\theta}$ puede ser suficientemente complicada, de modo que la teoría estadística estándar no pueda aplicarse para obtener una expresión para $\sigma_{\hat{\theta}}$. Esto es así, por ejemplo, en el caso $\theta = \sigma$, $\hat{\theta} = S$; la desviación estándar del estadístico S, σ_S, en general no puede ser determinada. No hace mucho se introdujo un método de computadora intensivo llamado ***bootstrap*** para abordar este problema. Suponga que la función de densidad de probabilidad de la población es $f(x; \theta)$, un miembro de una familia paramétrica particular, y que los datos $x_1, x_2, \ldots, x_n$ dan $\hat{\theta} = 21.7$. Ahora se utiliza la computadora para obtener "muestras *bootstrap*" tomadas de la función de densidad de probabilidad $f(x; 21.7)$, y por cada muestra se calcula una "estimación *bootstrap*" $\hat{\theta}^*$:

Primera muestra "*bootstrap*": $x_1^*, x_2^*, \ldots, x_n^*$; estimación $= \hat{\theta}_1^*$
Segunda muestra "*bootstrap*": $x_1^*, x_2^*, \ldots, x_n^*$; estimación $= \hat{\theta}_2^*$
$\vdots$
B-ésima muestra *bootstrap*: $x_1^*, x_2^*, \ldots, x_n^*$; estimación $= \hat{\theta}_B^*$

A menudo se utiliza $B = 100$ o 200. Ahora sea $\bar{\theta}^* = \Sigma\hat{\theta}_i^*/B$, la media muestral de las estimaciones *bootstrap*. La **estimación *bootstrap*** del error estándar de $\hat{\theta}$ ahora es simplemente la desviación estándar muestral de los $\hat{\theta}_i^*$:

$$s_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum(\hat{\theta}_i^* - \bar{\theta}^*)^2}$$

(En la literatura sobre *bootstrap* a menudo se utiliza B en lugar de $B - 1$; para valores típicos de B casi siempre hay poca diferencia entre las estimaciones resultantes.)

EJEMPLO 5.11 Un modelo teórico sugiere que X, el tiempo para la ruptura de un fluido aislante entre electrodos a un voltaje particular, tiene $f(x; \lambda) = \lambda e^{-\lambda x}$, una distribución exponencial. Una muestra aleatoria de $n = 10$ tiempos de ruptura (minutos) da los datos siguientes:

41.53 18.73 2.99 30.34 12.33 117.52 73.02 223.63 4.00 26.78

Puesto que $E(X) = 1/\lambda$, $E(\bar{X}) = 1/\lambda$, una estimación razonable de λ es $\hat{\lambda} = 1/\bar{x} = 1/55.087 = 0.018153$. Se utiliza entonces un paquete informático de estadística para obtener $B = 100$ muestras *bootstrap*, cada una de tamaño 10, provenientes de $f(x; 0.018153)$. La primera muestra fue 41.00, 109.70, 16.78, 6.31, 6.76, 5.62, 60.96, 78.81, 192.25, 27.61, con la cual $\Sigma x_i^* = 545.8$ y $\hat{\lambda}_i^* = 1/54.58 = 0.01832$. El promedio de las 100 estimaciones *bootstrap* es $\bar{\lambda}^* = 0.02153$, y la desviación estándar muestral de estas 100 estimaciones es $s_{\hat{\lambda}} = 0.0091$, la estimación *bootstrap* del error estándar de $\hat{\lambda}$. Un histograma de los $100\hat{\lambda}_i^*$ resultó un tanto positivamente asimétrico, lo que sugiere que la distribución muestral de $\hat{\lambda}$ también tiene esta propiedad. ■

En ocasiones un investigador desea estimar una característica poblacional sin suponer que la distribución de la población pertenece a una familia paramétrica particular. Una instancia de esto ocurrió en el ejemplo 5.7, cuando una media recortada 10% fue propuesta

para estimar el centro θ de una distribución de población simétrica. Los datos del ejemplo 5.2 dieron $\hat{\theta} = \bar{x}_{\text{rec}(10)} = 27.838$, pero ahora no hay ninguna $f(x; \theta)$ supuesta, por consiguiente, ¿cómo se puede obtener una muestra *bootstrap*? La respuesta es considerar que la muestra constituye la población (las $n = 20$ observaciones en el ejemplo 5.2) y tome B muestras diferentes, cada una de tamaño n, *con* reemplazo, de esta población. Varios de los libros listados en la bibliografía de este capítulo proporcionan más información acerca del *bootstrap*.

EJERCICIOS **Sección 5.1 (1–19)**

1. Los siguientes datos sobre resistencia a la flexión (MPa) de un tipo de vigas de concreto se mencionan en el ejemplo 1.2.

5.9	7.2	7.3	6.3	8.1	6.8	7.0
7.6	6.8	6.5	7.0	6.3	7.9	9.0
8.2	8.7	7.8	9.7	7.4	7.7	9.7
7.8	7.7	11.6	11.3	11.8	10.7	

a. Calcule una estimación puntual de la media de resistencia de la población conceptual de todas las vigas fabricadas de esta manera y mencione qué estimador utilizó. [*Sugerencia*: $\Sigma x_i = 219.8$.]

b. Calcule una estimación puntual del valor de resistencia que separa el 50% más débil de dichas vigas del 50% más resistente, y diga qué estimador utilizó.

c. Calcule e interprete una estimación puntual de la desviación estándar de la población σ. ¿Qué estimador utilizó? [*Sugerencia*: $\Sigma x_i^2 = 1860.94$.]

d. Calcule una estimación puntual de la proporción de las vigas cuya resistencia a la flexión exceda de 10 MPa. [*Sugerencia*: Considere una observación como "éxito" si excede de 10.]

e. Calcule una estimación puntual del coeficiente de variación de la población σ/μ y mencione qué estimador utilizó.

2. El **National Health and Nutrition Examination Survey (NHANES)** recopila información demográfica, socioeconómica, dietética y relacionada con la salud sobre una base anual. Aquí tenemos una muestra de 20 observaciones en el nivel de colesterol HDL (mg/dl) obtenidos de la encuesta 2009-2010 (HDL es el colesterol "bueno"; cuanto mayor sea su valor menor es el riesgo de enfermedad cardiaca):

35	49	52	54	65	51	51
47	86	36	46	33	39	45
39	63	95	35	30	48	

a. Calcule un punto de estimación de la media poblacional del colesterol HDL.

b. Sin hacer suposiciones acerca de la forma de la distribución de la población, calcule un punto de estimación de punto del valor que separa el 50% más grande de los niveles de HDL en el 50% más pequeño.

c. Calcule un punto de estimación de la desviación estándar de la población.

d. Un nivel de HDL de menos de 60 se considera deseable ya que corresponde a un menor riesgo de enfermedad cardiaca. Sin hacer ninguna hipótesis acerca de la forma de la distribución de la población, estime la proporción p de la población que tiene un nivel de HDL de menos de 60.

3. Considere la siguiente muestra de observaciones sobre el espesor del recubrimiento de pintura de baja viscosidad (**"Achieving a Target Value for a Manufacturing Process: A Case Study", *J. of Quality Technology*, 1992: 22–26**):

0.83	0.88	0.88	1.04	1.09	1.12	1.29	1.31
1.48	1.49	1.59	1.62	1.65	1.71	1.76	1.83

Suponga que la distribución del espesor del recubrimiento es normal (una gráfica de probabilidad normal avala firmemente esta suposición).

a. Calcule la estimación puntual de la media del espesor de recubrimiento y diga qué estimador utilizó.

b. Calcule una estimación puntual de la mediana de la distribución del espesor de recubrimiento y diga qué estimador utilizó.

c. Calcule la estimación puntual del valor que separa el 10% más grande de todos los valores de la distribución del espesor del restante 90% y diga qué estimador utilizó. [*Sugerencia*: Exprese lo que está tratando de estimar en función de μ y σ.]

d. Estime $P(X < 1.5)$; es decir, la proporción de todos los valores de espesor menor de 1.5. [*Sugerencia*: Si conociera los valores de μ y σ podría calcular esta probabilidad. Estos valores no están disponibles, pero pueden ser estimados.]

e. ¿Cuál es el error estándar estimado del estimador que utilizó en el inciso b)?

4. El artículo del cual se tomaron los datos en el ejercicio 1 también dio las siguientes observaciones de las resistencias de los cilindros:

6.1	5.8	7.8	7.1	7.2	9.2	6.6	8.3	7.0	8.3
7.8	8.1	7.4	8.5	8.9	9.8	9.7	14.1	12.6	11.2

Antes de obtener los datos denote las resistencias de las vigas mediante $X_1, \ldots, X_m$ y las resistencias de los cilindros $Y_1, \ldots, Y_m$. Suponga que las X_i constituyen una muestra

aleatoria de una distribución con media μ_1 y desviación estándar σ_1 y que las Y_i forman una muestra aleatoria (independiente de las X_i) de otra distribución con media μ_2 y desviación estándar σ_2.

a. Use las reglas de valor esperado para demostrar que $\overline{X} - \overline{Y}$ es un estimador insesgado de $\mu_1 - \mu_2$. Calcule el estimador para los datos dados.

b. Use las reglas de varianza del capítulo 5 para obtener una expresión para la varianza y la desviación estándar (error estándar) del estimador del inciso a) y luego calcule el error estándar estimado.

c. Calcule una estimación puntual de la razón σ_1/σ_2 de las dos desviaciones estándar.

d. Suponga que se seleccionan al azar una sola viga y un solo cilindro. Calcule una estimación puntual de la varianza de la diferencia $\overline{X} - \overline{Y}$ entre la resistencia de las vigas y la resistencia de los cilindros.

5. Como ejemplo de una situación en la que distintos estadísticos podrían ser razonablemente utilizados para calcular una estimación puntual considere una población de N facturas. Asociado a cada factura se encuentra su "valor en libros", el importe registrado de dicha factura. Sea T el valor total en libros, una cantidad conocida. Algunos de estos valores en libros son erróneos. Se realizará una auditoría, se seleccionarán al azar n facturas y se determinará el valor auditado (correcto) para cada una. Suponga que la muestra aporta los siguientes resultados (en dólares).

	Factura				
	1	**2**	**3**	**4**	**5**
Valor en libros	300	720	526	200	127
Valor auditado	300	520	526	200	157
Error	0	200	0	0	−30

Sea

$\overline{Y}$ = media muestral de libros

$\overline{X}$ = media muestral auditada

$\overline{D}$ = error medio muestral

Proponga tres estadísticos diferentes para estimar el valor total (correcto) auditado: uno que implique exactamente N y $\overline{X}$, otro que implique T, N y $\overline{D}$ y el último que implique T y $\overline{X}/\overline{Y}$. Si $N = 5000$ y $T = 1\,761\,300$, calcule las tres estimaciones puntuales correspondientes. (El artículo **"Statistical Models and Analysis in Auditing", *Statistical Science*, 1989: 2-33**, aborda las propiedades de estos tres estimadores.)

6. El nivel urinario angiotensinógeno (AGT) es un indicador cuantitativo de la función renal. El artículo **"Urinary Angiotensinogen as a Potential Biomarker of Chronic Kidney Diseases" (*J. of the Amer. Society of Hypertension*, 2008: 349–354)** describe un estudio en el que se determinó el nivel urinario AGT (μg) para una muestra de adultos con enfermedad renal crónica. Aquí le damos datos representativos (coherentes con las cantidades y las descripciones resumidas en el citado artículo):

2.6	6.2	7.4	9.6	11.5	13.5	14.5	17.0
20.0	28.8	29.5	29.5	41.7	45.7	56.2	56.2
66.1	66.1	67.6	74.1	97.7	141.3	147.9	177.8
186.2	186.2	190.6	208.9	229.1	229.1	288.4	288.4
346.7	407.4	426.6	575.4	616.6	724.4	812.8	1122.0

Una gráfica de probabilidad adecuada apoya el uso de la distribución lognormal (véase la sección 4.5) como un modelo razonable del nivel urinario AGT (esto es lo que hicieron los investigadores).

a. Estime los parámetros de la distribución. [*Sugerencia*: Recuerde que X tiene una distribución lognormal con parámetros μ y σ^2 si $\ln(X)$ se distribuye normalmente con media μ y varianza σ^2.]

b. Utilice las estimaciones del inciso a) para calcular una estimación del valor esperado del nivel AGT. [*Sugerencia*: ¿Qué es $E(X)$?]

7. **a.** Se selecciona una muestra aleatoria de 10 casas en un área particular, cada una de las cuales se calienta con gas natural, y se determina la cantidad de gas (*therms*) que utiliza cada casa durante el mes de enero. Las observaciones que resultan son: 103, 156, 118, 89, 125, 147, 122, 109, 138, 99. Sea μ el consumo promedio de gas de todas las casas del área durante enero. Calcule una estimación puntual de μ.

b. Suponga que 10 000 casas en esta área utilizan gas natural para calefacción. Sea τ la cantidad total del gas que consumieron todas estas casas durante enero. Calcule τ con los datos del inciso a). ¿Qué estimador utilizó para calcular su estimación?

c. Use los datos del inciso a) para estimar p, la proporción de todas las casas que usaron al menos 100 *terms*.

d. Proporcione una estimación puntual de la mediana de la población usada (el valor intermedio en la población de todas las casas) con base en la muestra del inciso a). ¿Qué estimador utilizó?

8. En una muestra aleatoria de 80 componentes de un tipo se encontraron 12 defectuosos.

a. Dé una estimación puntual de la proporción de todos los componentes que *no* están defectuosos.

b. Se tiene que construir un sistema mediante la selección al azar de dos de estos componentes para luego conectarlos en serie, como se muestra a continuación.

La conexión en serie implica que el sistema funcionará si y sólo si ningún componente está defectuoso (es decir, que ambos componentes funcionen apropiadamente). Estime la proporción de todos los sistemas que funcionan adecuadamente. [*Sugerencia:* Si p denota la probabilidad de que el componente funcione de manera apropiada, ¿cómo puede expresarse P(el sistema funciona) en función de p?]

9. Se examina cada uno de 150 artículos recién fabricados y se anota el número de rayones por artículo (se supone que los

artículos están libres de rayones) y se obtienen los siguientes datos:

Número de rayones por artículo	0	1	2	3	4	5	6	7
Frecuencia observada	18	37	42	30	13	7	2	1

Sea X = el número de rayones en un artículo seleccionado al azar y suponga que X tiene una distribución de Poisson con parámetro μ.

a. Determine un estimador insesgado de μ y calcule la estimación de los datos. [*Sugerencia*: $E(X) = \mu$ para una distribución de Poisson de X, por tanto, $E(\overline{X}) =$?]

b. ¿Cuál es la desviación estándar (error estándar) de su estimador? Calcule el error estándar estimado. [*Sugerencia*: $\sigma_X^2 = \mu$ con distribución de Poisson X.]

10. Con una larga varilla de longitud μ se va a trazar una gráfica cuadrada en la cual la longitud de cada lado será μ. Por consiguiente, el área de la curva será μ^2. Sin embargo, se desconoce el valor de μ por lo cual decide hacer n mediciones independientes $X_1, X_2, \ldots, X_n$ de la longitud. Suponga que cada X_i tiene una media μ (mediciones insesgadas) y varianza σ^2.

a. Demuestre que $\overline{X}^2$ no es un estimador insesgado de μ^2. [*Sugerencia*: Con cualquier variable aleatoria Y, $E(Y^2) = V(Y) + [E(Y)]^2$. Aplique esta con $Y = \overline{X}$.]

b. ¿Para qué valor de k el estimador $\overline{X}^2 - kS^2$ es insesgado para μ^2? [*Sugerencia*: Calcule $E(\overline{X}^2 - kS^2)$.]

11. De n_1 varones fumadores seleccionados al azar, X_1 fuman cigarrillos con filtro, mientras que de n_2 fumadoras seleccionadas, x_2 fuman cigarrillos con filtro. Sean p_1 y p_2 las probabilidades de que un varón y una mujer seleccionados al azar fumen, ambos, cigarrillos con filtro.

a. Demuestre que $(X_1/n_1) - (X_2/n_2)$ es un estimador insesgado de $p_1 - p_2$. [*Sugerencia*: $E(X_i) = n_i p_i$ con $i = 1, 2$.]

b. ¿Cuál es el error estándar del estimador en el inciso a)?

c. ¿Cómo utilizaría los valores observados x_1 y x_2 para estimar el error estándar de su estimador?

d. Si $n_1 = n_2 = 200$, $x_1 = 127$ y $x_2 = 176$, use el estimador del inciso a) para obtener una estimación de $p_1 - p_2$.

e. Use el resultado del inciso c) y los datos del inciso d) para estimar el error estándar del estimador.

12. Suponga que un tipo de fertilizante rinde por acre μ_1 con varianza σ^2, mientras que el rendimiento esperado de un segundo tipo de fertilizante es μ_2, con la misma varianza σ^2. Sean S_1^2 y S_2^2 las varianzas muestrales de los rendimientos basadas en tamaños muestrales n_1 y n_2, respectivamente, de los dos fertilizantes. Demuestre que el estimador combinado

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

es un estimador insesgado de σ^2.

13. Considere una muestra aleatoria $X_1, \ldots, X_n$ de la función de densidad de probabilidad

$$f(x; \theta) = 0.5(1 + \theta x) \quad -1 \le x \le 1$$

donde $-1 \le \theta \le 1$ (esta distribución se presenta en la física de partículas). Demuestre que es un estimador insesgado de θ. [*Sugerencia*: Primero determine $\mu = E(X) = E(\overline{X})$.]

14. Una muestra de n aviones de combate *Pandemonium* capturados tienen los números de serie $x_1, x_2, x_3, \ldots, x_n$. La Agencia Central de Inteligencia (CIA) sabe que los aviones fueron numerados consecutivamente en la fábrica comenzando con α y terminando con β, por lo que el número total de aviones fabricados es $\beta - \alpha + 1$ (p. ej., si $\alpha = 17$ y $\beta = 29$, entonces fueron fabricados $29 - 17 + 1 = 13$ aviones con números de serie 17, 18, 19, . . . , 28, 29). Sin embargo, la CIA no conoce los valores de α ni β. Un estadístico de la CIA sugiere utilizar el estimador máx(X_i) − mín(X_i) + 1 para estimar el número total de aviones fabricados.

a. Si $n = 5$, $x_1 = 237$, $x_2 = 375$, $x_3 = 202$, $x_4 = 525$ y $x_5 = 418$, ¿cuál es la estimación correspondiente?

b. ¿En qué condiciones de la muestra será el valor de la estimación exactamente igual al número total verdadero de aviones? ¿Alguna vez será más grande la estimación que el total verdadero? ¿Piensa que el estimador es insesgado para estimar $\beta - \alpha + 1$? Explique en uno o dos renglones.

15. Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria tomada de una distribución de Rayleigh con función de densidad de probabilidad

$$f(x; \theta) = \frac{x}{\theta} e^{-x^2/(2\theta)} \quad x > 0$$

a. Se puede demostrar que $E(X^2) = 2\theta$. Use este hecho para construir un estimador insesgado de θ basado en ΣX_i^2 (y use reglas de valor esperado para demostrar que es insesgado).

b. Calcule θ a partir de las siguientes $n = 10$ observaciones de esfuerzo vibratorio del aspa de una turbina en condiciones específicas:

16.88	10.23	4.59	6.66	13.68
14.23	19.87	9.40	6.51	10.95

16. Suponga que el crecimiento promedio verdadero μ de un tipo de planta durante un periodo de 1 año es idéntico al de un segundo tipo, aunque la varianza del crecimiento del primer tipo es σ^2, en tanto que para el segundo tipo la varianza es $4\sigma^2$. Sean $X_1, \ldots, X_m$, m observaciones de crecimiento independientes del primer tipo [por consiguiente $E(X_i) = \mu$, $V(X_i) = \sigma^2$], y sean $Y_1, \ldots, Y_n$, n observaciones de crecimiento independientes del segundo tipo [$E(Y_i) = \mu$, $V(Y_i) = 4\sigma^2$].

a. Demuestre que el estimador $\hat{\mu} = \delta\overline{X} = (1 - \delta)\overline{Y}$ es insesgado para μ (para $0 < \delta < 1$ el estimador es un promedio pesado de dos medias muestrales simples).

b. Con m y n fijas calcule $V(\hat{\mu})$y luego determine el valor de δ que reduzca al mínimo $V(\hat{\mu})$. [*Sugerencia*: Derive $V(\hat{\mu})$ respecto a δ.]

17. En el capítulo 3 se definió una variable aleatoria binomial negativa como el número de fallas que ocurren antes del r–ésimo éxito en una secuencia de ensayos con éxitos y fallas

independientes e idénticos. La función de masa de probabilidad (pmf) de X es

$$nb(x; r, p) = \binom{x + r - 1}{x} p^r(1 - p)^x \quad x = 0, 1, 2, \ldots$$

a. Suponga que $r \geq 2$. Demuestre que

$$\hat{p} = (r - 1)/(X + r - 1)$$

es un estimador insesgado de p. [*Sugerencia*: Escriba $E(\hat{p})$ y elimine $x + r - 1$ dentro de la suma.]

b. Un reportero desea entrevistar a cinco individuos que apoyan a un candidato y comienza preguntándoles si apoyan (S) o si no apoyan (F) al candidato. Si la secuencia de respuestas es *SFFSFFFSSS*, estime p = la proporción verdadera que apoya al candidato.

18. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una función de densidad de probabilidad $f(x)$ que es simétrica respecto a μ, de modo que $\widetilde{X}$ es un estimador insesgado de μ. Si n es grande se puede demostrar que $V(\widetilde{X}) \approx 1/(4n[f(\mu)]^2)$.

a. Compare $V(\widetilde{X})$ con $V(\overline{X})$ cuando la distribución subyacente es normal.

b. Cuando la función de densidad de probabilidad subyacente es de Cauchy (véase el ejemplo 5.7), $V(\overline{X}) = \infty$, por tanto, $\overline{X}$ es un estimador terrible. Cuando n es grande ¿cuál es $V(\widetilde{X})$ en este caso?

19. Una investigadora desea estimar la proporción de estudiantes en una universidad que han violado el código de honor. Habiendo obtenido una muestra aleatoria de n estudiantes, se da cuenta de que si a cada uno le pregunta "¿Has violado el código de honor?", probablemente reciba algunas respuestas poco veraces. Considere el siguiente esquema, conocido como técnica de **respuesta aleatorizada**. La investigadora forma un mazo de 100 cartas de las cuales 50 son de tipo I y 50 de tipo II.

Tipo I: ¿Has violado el código de honor (sí o no)?

Tipo II: ¿El último dígito de su número telefónico es un 0, 1 o 2 (sí o no)?

A cada estudiante en la muestra aleatoria se le pide que baraje el mazo, que saque una carta y que responda la pregunta con sinceridad. Debido a lo irrelevante de la pregunta en las cartas de tipo II, una respuesta afirmativa ya no estigmatiza al que responde, por tanto, se supone que es una respuesta sincera. Sea p la proporción de quienes violan el código de honor (es decir, la probabilidad de que un estudiante seleccionado al azar sea un infractor) y sea $\lambda = P$(respuesta afirmativa). Entonces λ y p están relacionados por $\lambda = 0.5p + (0.5)(0.3)$.

a. Sea Y el número de respuestas afirmativas, por consiguiente $Y \sim \text{Bin}(n, \lambda)$. Por tanto, Y/n es un estimador insesgado de λ. Deduzca un estimador de p basado en Y. Si $n = 80$ y $y = 20$, ¿cuál es su estimación? [*Sugerencia*: Despeje p de $\lambda = 0.5p + 0.15$ y luego sustituya Y/n en lugar de λ.]

b. Use el hecho de que $E(Y/n) = \lambda$ para demostrar que su estimador es insesgado.

c. Si hubiera 70 cartas de tipo I y 30 de tipo II, ¿cuál sería su estimador para p?

5.2 Métodos de estimación puntual

A continuación se discuten dos métodos "constructivos" para obtener estimadores puntuales: el método de momentos y el método de máxima probabilidad. Por *constructivo* se quiere dar a entender que la definición general de cada tipo de estimador sugiere explícitamente cómo obtener el estimador en cualquier problema específico. Aun cuando se prefieren los estimadores de máxima probabilidad a los estimadores de momento, debido a ciertas propiedades de eficiencia, a menudo requieren significativamente más cálculo que estos últimos. En ocasiones es el caso que estos métodos dan estimadores insesgados.

Método de momentos

La idea básica de este método es poder igualar ciertas características muestrales, tales como la media, a los valores esperados de la población correspondiente. Luego, al resolver estas ecuaciones para valores de parámetros desconocidos se obtienen los estimadores.

DEFINICIÓN

Sean $X_1, \ldots, X_n$ una muestra aleatoria proveniente de una función de masa de probabilidad o de una función de densidad de probabilidad $f(x)$. Con $k = 1, 2, 3, \ldots$, el **momento *k*-ésimo de la población** o el **momento *k*-ésimo de la distribución *f*(*x*)**, es $E(X^k)$. El **momento muestral *k*-ésimo** es $(1/n)\Sigma_{i=1}^{n} X_i^k$.

Por consiguiente, el primer momento de la población es $E(X) = \mu$ y el primer momento muestral es $\Sigma X_i/n = \overline{X}$. Los segundos momentos de la población y muestral son $E(X^2)$ y $\Sigma X_i^2/n$, respectivamente. Los momentos de la población serán funciones de cualesquiera parámetros desconocidos θ_1, θ_2, ….

DEFINICIÓN

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con función de masa de probabilidad o función de densidad de probabilidad $f(x; \theta_1, \ldots, \theta_m)$, donde $\theta_1, \ldots, \theta_m$ son parámetros cuyos valores son desconocidos. Entonces los **estimadores de momento** $\hat{\theta}_1, \ldots, \hat{\theta}_m$ se obtienen igualando los primeros m momentos muestrales con los primeros m momentos de la población correspondientes y resolviendo para $\theta_1, \ldots, \theta_m$.

Si, por ejemplo, $m = 2$, $E(X)$ y $E(X^2)$ serán funciones de θ_1 y θ_2. Con $E(X) = (1/n)\,\Sigma X_i$ $(= \overline{X})$ y $E(X^2) = (1/n)\,\Sigma X_i^2$ se obtienen dos ecuaciones con θ_1 y θ_2. La solución define entonces los estimadores.

EJEMPLO 5.12 Sean $X_1, X_2, \ldots, X_n$ la representación de una muestra aleatoria de tiempos de servicio de n clientes en una instalación, donde la distribución subyacente se supone exponencial con el parámetro λ. Como sólo hay un parámetro que tiene que ser estimado, el estimador se obtiene igualando $E(X)$ a $\overline{X}$. Puesto que $E(X) = 1/\lambda$ con una distribución exponencial, esta da $1/\lambda = \overline{X}$ o $\lambda = 1/\overline{X}$. El estimador de momento de λ es entonces $\hat{\lambda} = 1/\overline{X}$. ■

EJEMPLO 5.13 Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución gamma con parámetros α y β. De acuerdo con la sección 4.4, $E(X) = \alpha\beta$ y $E(X^2) = \beta^2\,\Gamma(\alpha + 2)/\Gamma(\alpha) = \beta^2(\alpha + 1)\alpha$. Los estimadores de momento de α y β se obtienen resolviendo

$$\overline{X} = \alpha\beta \qquad \frac{1}{n}\sum X_i^2 = \alpha(\alpha + 1)\beta^2$$

Puesto que $\alpha(\alpha + 1)\beta^2 = \alpha^2\beta^2 + \alpha\beta^2$ y la primera ecuación implica que $\alpha^2\beta^2 = \overline{X}^2$, la segunda ecuación se vuelve

$$\frac{1}{n}\sum X_i^2 = \overline{X}^2 + \alpha\beta^2$$

Ahora si se divide cada miembro de esta segunda ecuación entre el miembro correspondiente de la primera ecuación y se sustituye otra vez, se obtienen los estimadores

$$\hat{\alpha} = \frac{\overline{X}^2}{(1/n)\sum X_i^2 - \overline{X}^2} \qquad \hat{\beta} = \frac{(1/n)\sum X_i^2 - \overline{X}^2}{\overline{X}}$$

Para ilustrar, los datos de tiempo de sobrevivencia mencionados en el ejemplo 4.24 son

152	115	109	94	88	137	152	77	160	165
125	40	128	123	136	101	62	153	83	69

con $\overline{x} = 113.5$ y $(1/20)\Sigma x_i^2 = 14\,087.8$. Los estimadores son

$$\hat{\alpha} = \frac{(113.5)^2}{14\,087.8 - (113.5)^2} = 10.7 \qquad \hat{\beta} = \frac{14\,087.8 - (113.5)^2}{113.5} = 10.6$$

Estas estimaciones de α y β difieren de los valores sugeridos por Gross y Clark porque ellos utilizaron una técnica de estimación diferente. ■

EJEMPLO 5.14 Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución binomial negativa generalizada con parámetros r y p. Puesto que $E(X) = r(1 - p)/p$ y $V(X) = r(1 - p)/p^2$, $E(X^2) = V(X) + [E(X)]^2 = r\,(1 - p)(r - rp + 1)/p^2$. Si se iguala $E(X)$ a $\overline{X}$ y $E(X^2)$ a $(1/n)\Sigma X_i^2$ a la larga se obtiene

$$\hat{p} = \frac{\overline{X}}{(1/n)\sum X_i^2 - \overline{X}^2} \qquad \hat{r} = \frac{\overline{X}^2}{(1/n)\sum X_i^2 - \overline{X}^2 - \overline{X}}$$

A modo de ilustración, Reep, Pollard y Benjamin (**"Skill and Chance in Ball Games", *J. of Royal Stat. Soc*., 1971: 623–629**) consideran la distribución binomial negativa como modelo del número de goles por juego anotados por los equipos de la Liga Nacional de Jockey. Los datos de 1966–1967 son los siguientes (420 juegos):

Goles	0	1	2	3	4	5	6	7	8	9	10
Frecuencia	29	71	82	89	65	45	24	7	4	1	3

Por tanto,

$$\bar{x} = \sum x_i/420 = [(0)(29) + (1)(71) + \cdots + (10)(3)]/420 = 2.98$$

y

$$\sum x_i^2/420 = [(0)^2(29) + (1)^2(71) + \cdots + (10)^2(3)]/420 = 12.40$$

Por consiguiente,

$$\hat{p} = \frac{2.98}{12.40 - (2.98)^2} = 0.85 \qquad \hat{r} = \frac{(2.98)^2}{12.40 - (2.98)^2 - 2.98} = 16.5$$

Aunque por definición r debe ser positivo, el denominador de $\hat{r}$ podría ser negativo, lo que indica que la distribución binomial negativa no es apropiada (o que el estimador de momento es defectuoso). ■

Estimación de máxima probabilidad

El método de máxima probabilidad lo introdujo por primera vez R. A. Fisher, genetista y estadístico en la década de 1920. La mayoría de los estadísticos recomiendan este método, al menos cuando el tamaño de muestra es grande, puesto que los estimadores resultantes tienen ciertas propiedades de eficiencia deseables.

EJEMPLO 5.15 La mejor protección contra la piratería en una cuenta en línea es utilizar una contraseña que tenga al menos 8 caracteres que consisten en mayúsculas, minúsculas, números y caracteres especiales. [*Nota*: La edición de enero de 2012 de ***Consumer Reports*** informa que sólo 25% de las personas encuestadas utiliza una contraseña fuerte.] Supongamos que se seleccionan 10 personas con cuentas de correo electrónico con un determinado proveedor. Se encontró que el primer, tercer y décimo individuos tienen dicha protección fuerte, mientras que los otros no. Sea $p = P$(protección fuerte), es decir, p es la proporción de todos esos titulares de cuenta con protección fuerte. Definir las variables aleatorias (Bernoulli) $X_1, X_2, \ldots, X_{10}$ mediante

$$X_1 = \begin{cases} 1 \text{ si la primera persona tiene protección fuerte} \\ 0 \text{ si la primera persona no tiene protección fuerte} \end{cases} \ldots X_{10} = \begin{cases} 1 \text{ si la décima persona tiene protección fuerte} \\ 0 \text{ si la décima persona no tiene protección fuerte} \end{cases}$$

Entonces, para la muestra obtenida, $X_1 = X_3 = X_{10} = 1$ y las otras siete X_i son cero. La función de masa de probabilidad de cualquier X_i particular es $p^{x_i}(1 - p)^{1-x_i}$, que se convierte en p si $x_i = 1$ y $1 - p$ cuando $x_i = 0$. Suponga ahora que las condiciones de las contraseñas diferentes son independientes una de la otra. Esto implica que las X_i son independientes,

por lo que su función de masa de probabilidad conjunta es el producto de las funciones de masa de probabilidad individuales. Así que la función de masa de probabilidad conjunta de las X_i observadas es

$$f(x_1,\ldots,x_{10};p) = p(1-p)p\cdots p = p^3(1-p)^7 \tag{5.4}$$

Suponga que $p = 0.25$. Entonces la probabilidad de observar la muestra que en realidad se obtiene es $(0.25)^3(0.75)^7 = 0.002086$. Si, en cambio, $p = 0.50$, entonces esta probabilidad es $(0.50)^3(0.50)^7 = 0.000977$. ¿Para qué valor de p es más probable que la muestra observada haya ocurrido? Es decir, ¿para qué valor de p es la función de masa de probabilidad conjunta (5.4) tan grande como puede ser? ¿Qué valor de p maximiza (5.4)? La figura 5.6(a) muestra un gráfico de la *probabilidad* (5.4) en función de p. Parece que la gráfica alcanza su pico por encima de $p = 0.3 =$ la proporción de contraseñas fuertes en la muestra. La figura 5.6(b) muestra una gráfica del logaritmo natural de (5.4) ya que ln[g(u)] es una función estrictamente creciente de $g(u)$; encontrar u para maximizar la función $g(u)$ es lo mismo que encontrar u para maximizar ln[$g(u)$].

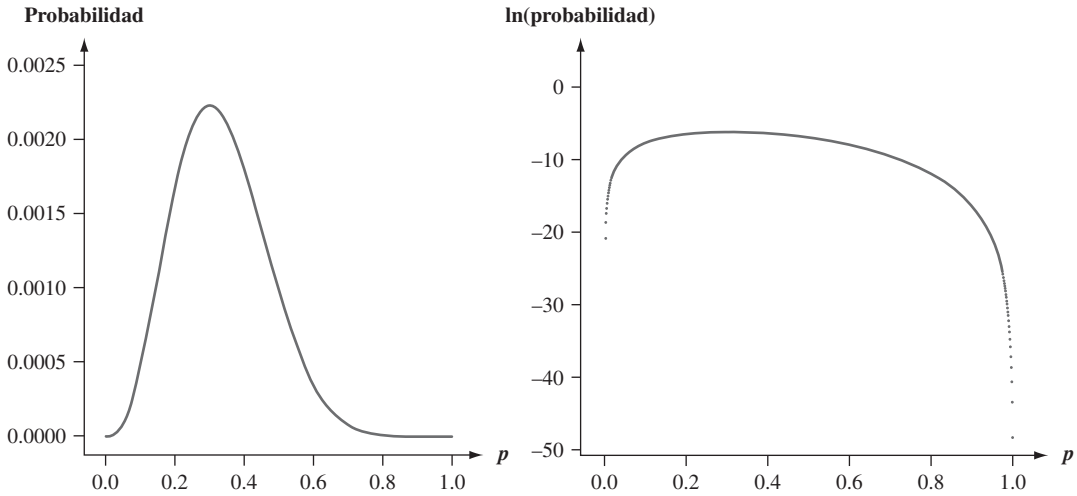


Figura 5.6 (a) Gráfica de la probabilidad (función de masa de probabilidad conjunta) (5.4) del ejemplo 5.15; (b) Gráfica del logaritmo natural de la probabilidad

Podemos comprobar nuestra impresión visual mediante el cálculo para hallar el valor de p que maximiza (5.4). Trabajar con el logaritmo natural de la función de masa de probabilidad conjunta suele ser más fácil que trabajar con la función de masa de probabilidad conjunta; puesto que esta última es típicamente un producto, su logaritmo será una suma. Aquí

$$\ln[f(x_1,\ldots,x_{10};p)] = \ln[p^3(1-p)^7] = 3\ln(p) + 7\ln(1-p) \tag{5.5}$$

Por tanto

$$\frac{d}{dp}\{\ln[f(x_1,\ldots,x_{10};p)]\} = \frac{d}{dp}\{3\ln(p) + 7\ln(1-p)\} = \frac{3}{p} + \frac{7}{1-p}(-1)$$
$$= \frac{3}{p} - \frac{7}{1-p}$$

[el (-1) viene de la regla de la cadena en cálculo]. Igualando esta derivada a 0 y despejando p da $3(1-p) = 7p$, de lo cual $3 = 10p$ y así $p = 3/10 = 0.30$ como hemos deducido. Es decir, nuestra estimación puntual es $\hat{p} = 0.30$. Se llama *estimación de máxima verosimilitud (probabilidad)*, ya que es el valor del parámetro que maximiza la probabilidad (pmf conjunta) de la muestra observada. En general, la segunda derivada debe ser examinada para asegurarse de que se ha obtenido un máximo, pero aquí esto es obvio en la figura 5.5.

Supongamos que en vez de conocer la condición de cada contraseña, sólo sabemos que tres de diez eran fuertes. Entonces tendríamos el valor observado de una variable aleatoria binomial $X =$ el número de contraseñas fuertes. La función de masa de probabilidad de X es $\binom{10}{x}p^x(1-p)^{10-x}$. Para $x = 3$, esto se convierte en $\binom{10}{3}p^3(1-p)^7$. El coeficiente binomial $\binom{10}{3}$ es irrelevante para la maximización, así que nuevamente $\hat{p} = 0.30$. ■

DEFINICIÓN

Sean $X_1, X_2, \ldots, X_n$ que tienen una función de masa de probabilidad o una función de densidad de probabilidad

$$f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \tag{5.6}$$

donde los parámetros $\theta_1, \ldots, \theta_m$ tienen valores desconocidos. Cuando $X_1, \ldots, X_n$ son los valores muestrales observados y (5.6) se considera una función de $\theta_1, \ldots, \theta_m$, esto se conoce como **función de probabilidad**. Las estimaciones de máxima probabilidad $\hat{\theta}, \ldots, \hat{\theta}_m$ son aquellos valores de las θ_i que incrementan al máximo la función de probabilidad, de modo que

$$f(x_1,\ldots, x_n; \hat{\theta}_1,\ldots, \hat{\theta}_m) \geq f(x_1,\ldots, x_n; \theta_1,\ldots, \theta_m) \text{ para todas las } \theta_1,\ldots, \theta_m$$

Cuando se sustituyen las X_i en lugar de las x_i se obtienen los **estimadores de máxima probabilidad**.

La función de probabilidad dice qué tan probable es que la muestra observada sea una función de los posibles valores de parámetro. Al incrementarse al máximo la probabilidad, se obtienen los valores de parámetro con los que es más probable que la muestra observada haya sido generada; es decir, los valores de parámetro que "más concuerdan" con los datos observados.

EJEMPLO 5.16 Suponga que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una distribución exponencial con parámetro λ. Debido a la independencia, la función de probabilidad es un producto de las funciones de densidad de probabilidad individuales:

$$f(x_1,\ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot\cdots\cdot (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \Sigma x_i}$$

El logaritmo natural de la función de probabilidad es

$$\ln[f(x_1,\ldots, x_n; \lambda)] = n \ln(\lambda) - \lambda \sum x_i$$

Si se iguala $(d/d\lambda)$[ln(probabilidad)] a cero se obtiene $n/\lambda - \Sigma x_i = 0$, o $\lambda = n/\Sigma x_i = 1/\overline{x}$. Por consiguiente, el estimador de máxima probabilidad es $\hat{\lambda} = 1/\overline{X}$; es idéntico al método de estimador de momentos [pero no es un estimador insesgado, puesto que $E(1/\overline{X}) \neq 1/E(\overline{X})$]. ■

EJEMPLO 5.17 Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución normal. La función de probabilidad es

$$\begin{aligned} f(x_1,\ldots, x_n; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)} \end{aligned}$$

por consiguiente

$$\ln[f(x_1,\ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

Para determinar los valores maximizantes de μ y σ^2 se deben sacar las derivadas parciales de $\ln(f)$ respecto a μ y σ^2, igualarlas a cero y resolver las dos ecuaciones resultantes. Omitiendo los detalles, los estimadores de máxima probabilidad resultantes son

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

El estimador de máxima probabilidad de σ^2 no es el estimador insesgado, por consiguiente dos principios diferentes de estimación (ausencia de sesgo y máxima probabilidad) dan dos estimadores diferentes. ■

EJEMPLO 5.18 En el capítulo 3 se mencionó el uso de la distribución de Poisson para modelar el número de "eventos" que ocurren en una región bidimensional. Suponga que cuando la región R se está muestreando tiene área $a(R)$, el número X de eventos que ocurren en R tiene una distribución de Poisson con parámetro $\lambda a(R)$ (donde λ es el número esperado de eventos por unidad de área) y que las regiones no traslapantes dan X independientes.

Suponga que un ecólogo selecciona n regiones no traslapantes $R_1, \ldots, R_n$ y cuenta el número de plantas de una especie en cada región. La función de masa de probabilidad (verosimilitud) conjunta es entonces

$$\begin{aligned} p(x_1, \ldots, x_n; \lambda) &= \frac{[\lambda \cdot a(R_1)]^{x_1} e^{-\lambda \cdot a(R_1)}}{x_1!} \cdot \cdots \cdot \frac{[\lambda \cdot a(R_n)]^{x_n} e^{-\lambda \cdot a(R_n)}}{x_n!} \\ &= \frac{[a(R_1)]^{x_1} \cdot \cdots \cdot [a(R_n)]^{x_n} \cdot \lambda^{\Sigma x_i} \cdot e^{-\lambda \Sigma\, a(R_i)}}{x_1! \cdot \cdots \cdot x_n!} \end{aligned}$$

El ln(probabilidad) es

$$\ln[p(x_1, \ldots, x_n; \lambda)] = \sum x_i \cdot \ln[a(R_i)] + \ln(\lambda) \cdot \sum x_i - \lambda \sum a(R_i) - \sum \ln(x_i!)$$

Tomando $d/d\lambda \ln(p)$ e igualándola a cero da

$$\frac{\sum x_i}{\lambda} - \sum a(R_i) = 0$$

de lo cual

$$\lambda = \frac{\sum x_i}{\sum a(R_i)}$$

El estimador de máxima probabilidad es entonces $\hat{\lambda}\ \Sigma X_i / \Sigma a(R_i)$. Esta es razonablemente intituitiva porque λ es la densidad verdadera (plantas por unidad de área), mientras que $\hat{\lambda}$ es la densidad muestral puesto que $\Sigma a(R_i)$ es tan sólo el área total muestreada. Debido a que $E(X_i) = \lambda \cdot a(R_i)$, el estimador es insesgado.

En ocasiones se utiliza un procedimiento de muestreo alternativo. En lugar de fijar las regiones que serán muestreadas, el ecólogo seleccionará n puntos en toda la región de interés, y sea y_i = la distancia del punto i-ésimo a la planta más cercana. La función de distribución acumulada de Y = distancia a la planta más cercana es

$$F_Y(y) = P(Y \le y) = 1 - P(Y > y) = 1 - P\left(\begin{array}{c}\text{ninguna planta en} \\ \text{un círculo de radio } y\end{array}\right)$$

$$= 1 - \frac{e^{-\lambda \pi y^2} (\lambda \pi y^2)^0}{0!} = 1 - e^{-\lambda \cdot \pi y^2}$$

Al sacar la derivada de $F_Y(y)$ respecto a y resulta

$$f_Y(y; \lambda) = \begin{cases} 2\pi\lambda y e^{-\lambda \pi y^2} & y \ge 0 \\ 0 & \text{de lo contrario} \end{cases}$$

Si ahora formamos la probabilidad $f_Y(y_1; \lambda) \cdot \ldots \cdot f_Y(y_n; \lambda)$, derivamos ln(probabilidad), etcétera, el estimador de máxima probabilidad resultante es

$$\hat{\lambda} = \frac{n}{\pi \sum Y_i^2} = \frac{\text{número de plantas observadas}}{\text{área total muestreada}}$$

la cual también es una densidad muestral. Se puede demostrar que un ambiente escaso (pequeño λ), el método de distancia es mejor en cierto sentido, en tanto que en un ambiente denso, el primer método de muestreo es mejor. ■

EJEMPLO 5.19 Sean $X_1, \ldots, X_n$ una muestra aleatoria de una función de densidad de probabilidad de Weibull

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{\alpha}{\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-(x/\beta)^\alpha} & x \geq 0 \\ 0 & \text{de lo contrario} \end{cases}$$

Si se escribe la probabilidad y el ln(probabilidad) y luego $(\partial/\partial\alpha)[\ln(f)] = 0$ y $(\partial/\partial\beta)[\ln(f)] = 0$ se obtienen las ecuaciones

$$\alpha = \left[\frac{\sum x_i^\alpha \cdot \ln(x_i)}{\sum x_i^\alpha} - \frac{\sum \ln(x_i)}{n} \right]^{-1} \qquad \beta = \left(\frac{\sum x_i^\alpha}{n} \right)^{1/\alpha}$$

Estas dos ecuaciones no pueden ser resueltas explícitamente para obtener fórmulas generales de los estimadores de máxima probabilidad $\hat{\alpha}$ y $\hat{\beta}$. En su lugar por cada muestra $x_1, \ldots, x_n$, las ecuaciones deben ser resueltas con un procedimiento numérico iterativo. Incluso los estimadores de momento de α y β son un tanto complicados. ■

Estimación de funciones de parámetros

Una vez que está disponible el estimador de máxima probabilidad para un parámetro θ, el estimador de máxima probabilidad para cualquier función de θ, tal como $1/\theta$ o $\sqrt{\theta}$, es fácil de obtener.

PROPOSICIÓN

Principio de invarianza

Sean $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ los estimadores de máxima probabilidad de los parámetros $\theta_1, \theta_2, \ldots, \theta_m$. Entonces el estimador de máxima probabilidad de cualquier función $h(\theta_1, \theta_2, \ldots, \theta_m)$ de estos parámetros es la función $h(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ de los estimadores de máxima probabilidad.

EJEMPLO 5.20 (Continuación del ejemplo 5.17) En el caso normal los estimadores de máxima probabilidad de μ y σ^2 son $\hat{\mu} = \overline{X}$ y $\hat{\sigma}^2 = \Sigma(X_i - X)^2/n$. Para obtener el estimador de máxima probabilidad de la función $h(\mu, \sigma^2) = \sqrt{\sigma^2}\ \sigma$, sustituya los estimadores de máxima probabilidad en la función:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[\frac{1}{n} \sum (X_i - \overline{X})^2 \right]^{1/2}$$

El estimador de máxima probabilidad de σ no es la desviación estándar muestral S, aunque se aproximan bastante a menos que n sea bastante pequeño. ■

EJEMPLO 5.21 (Continuación del ejemplo 5.19) La media de una variable aleatoria X que tiene una distribución de Weibull es

$$\mu = \beta \cdot \Gamma(1 + 1/\alpha)$$

El estimador de máxima probabilidad de μ es, por consiguiente, $\hat{\mu} = \hat{\beta}\Gamma\,(1 + 1/\hat{\alpha})$ donde $\hat{\alpha}$ y $\hat{\beta}$ son los estimadores de máxima probabilidad de α y β. En particular, $\overline{X}$ no es el estimador de máxima probabilidad de μ, aunque es un estimador insesgado. Por lo menos para n grande, $\hat{\mu}$ es un mejor estimador que $\overline{X}$.

Para los datos que figuran en el ejemplo 5.3, los estimadores de máxima probabilidad de los parámetros de Weibull son $\hat{\alpha} = 11.9731$ y $\hat{\beta} = 77.0153$, de los cuales $\hat{\mu} = 73.80$. Esta estimación está muy cerca de la media de la muestra 73.88. ■

Comportamiento del estimador de máxima probabilidad con muestra grande

Aunque el principio de la estimación de máxima probabilidad tiene un considerable atractivo intuitivo, la siguiente proposición brinda razones adicionales fundamentales para el uso de estimadores de máxima probabilidad.

PROPOSICIÓN

En condiciones muy generales en relación con la distribución conjunta de la muestra, cuando el tamaño n de la muestra es grande, el estimador de máxima probabilidad de cualquier parámetro θ es aproximadamente insesgado [$E(\hat{\theta}) \approx \theta$] y su varianza es casi tan pequeña como la que puede ser lograda por cualquier estimador. Expresado de otra manera, el estimador de máxima probabilidad $\hat{\theta}$ es aproximadamente el estimador insesgado con varianza mínima de θ.

Debido a este resultado y al hecho de que las técnicas basadas en el cálculo casi siempre pueden ser utilizadas para obtener los estimadores de máxima probabilidad (aunque a veces se requieren métodos numéricos, tales como el método de Newton), la estimación de máxima probabilidad es la técnica de estimación más ampliamente utilizada entre los estadísticos. Muchos de los estimadores que se emplean en lo que resta del libro son estimadores de máxima probabilidad. Sin embargo, la obtención de un estimador de máxima probabilidad requiere que se especifique la distribución subyacente.

Algunas complicaciones

En ocasiones no se puede utilizar el cálculo para obtener estimadores de máxima probabilidad.

EJEMPLO 5.22 Suponga que mi tiempo de espera del autobús está uniformemente distribuido en [0, θ] y que se observaron los resultados $x_1, \ldots, x_n$ de una muestra aleatoria tomada de esta distribución. Puesto que $f(x; \theta) = 1/\theta$ con $0 \le x \le \theta$ y de lo contrario 0,

$$f(x_1,\ldots,x_n;\theta) = \begin{cases} \dfrac{1}{\theta^n} & 0 \le x_1 \le \theta,\ldots, 0 \le x_n \le \theta \\ 0 & \text{de lo contrario} \end{cases}$$

En tanto máx$(x_i) \le \theta$, la probabilidad es $1/\theta^n$ la cual es positiva, pero en cuanto $\theta <$ máx(x_i), la probabilidad se reduce a 0. Esto se ilustra en la figura 5.7. El cálculo no funciona porque el máximo de la probabilidad ocurre en un punto de discontinuidad, pero la figura indica que $\hat{\theta} =$ máx(X_i). Por consiguiente, si mis tiempos de espera son 2.3, 3.7, 1.5, 0.4 y 3.2, entonces el estimador de máxima probabilidad es $\hat{\theta} = 3.7$. De acuerdo con el ejemplo 5.4 el estimador de máxima probabilidad no es insesgado.

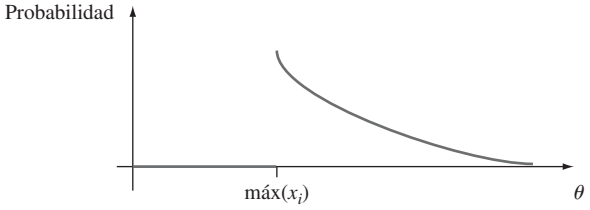


Figura 5.7 Función de probabilidad para el ejemplo 5.22 ■

EJEMPLO 5.23 Un método que a menudo se utiliza para estimar el tamaño de una población de vida silvestre implica realizar un experimento de captura-recaptura. En este experimento se captura una muestra inicial de M animales; cada uno los cuales se etiqueta y se regresa a la población. Luego de un tiempo suficiente para que los individuos etiquetados se mezclen con la población, se captura otra muestra de tamaño n. Con $X =$ el número de animales etiquetados en la segunda muestra, el objetivo es utilizar las x observadas para estimar la población de tamaño N.

El parámetro de interés es $\theta = N$ el cual puede asumir sólo valores enteros así que, incluso después de determinar la función de probabilidad (función de masa de probabilidad de X en este caso), el uso del cálculo para obtener N presentaría dificultades. Si se considera un éxito la recaptura de un animal previamente etiquetado, entonces el muestreo es sin reemplazo de una población que contiene M éxitos y $N - M$ fallas, de modo que X es una variable aleatoria hipergeométrica y la función de probabilidad es

$$p(x; N) = h(x; n, M, N) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

No obstante, la naturaleza de valor entero de N, sería difícil evaluar la derivada de $p(x; N)$. Sin embargo, si se considera la razón de $p(x; N)$ y $p(x; N - 1)$, se tiene

$$\frac{p(x; N)}{p(x; N-1)} = \frac{(N-M) \cdot (N-n)}{N(N-M-n+x)}$$

Esta razón es más grande que 1 si y sólo si $N < Mn/x$. El valor de N con el cual $p(x; N)$ se incrementa al máximo es, por consiguiente, el entero más grande menor que Mn/x. Si se utiliza la notación matemática estándar $[r]$ para el entero más grande menor o igual a r, el estimador de máxima probabilidad de N es $\hat{N} = [Mn/x]$. A manera de ilustración, si $M = 200$ peces se sacan de un lago y se etiquetan, posteriormente $n = 100$ son recapturados y entre los 100 hay $x = 11$ etiquetados; en ese caso $\hat{N} = [(200)(100)/11]\ [1818.18] = 1818$. La estimación es en realidad un tanto intuitiva; x/n es la proporción de la muestra recapturada etiquetada, mientras que M/N es la proporción de toda la población etiquetada. La estimación se obtiene igualando estas dos proporciones (estimando una proporción poblacional mediante una proporción muestral). ■

Suponga que $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una función de densidad de probabilidad $f(x; \theta)$ simétrica respecto a θ, aunque el investigador no está seguro de la forma de la función f. Es entonces deseable utilizar un estimador $\hat{\theta}$ *robusto*; es decir, uno que funcione bien con una amplia variedad de funciones de densidad de probabilidad subyacentes. Un estimador como ese es una media recortada. En años recientes los estadísticos han propuesto otro tipo de estimador llamado *estimador M*, basado en una generalización de la estimación de máxima probabilidad. En lugar de incrementar al máximo el logaritmo de la probabilidad $\Sigma \ln[f(x; \theta)]$ para una f específica, se incrementa al máximo $\Sigma \rho(x_i; \theta)$. Se selecciona la "función objetivo" ρ para que dé un estimador con buenas propiedades de robustez. El libro de David Hoaglin y colaboradores (véase la bibliografía) contiene una buena exposición de esta materia.

EJERCICIOS **Sección 5.2 (20–30)**

20. Se aplica una prueba de diagnóstico para una enfermedad determinada a n individuos de quienes se sabe que no tienen la enfermedad. Sea X el número uno de los n resultados de prueba que son positivos (lo que indica la presencia de la enfermedad, por lo que X es el número de falsos positivos) y p = probabilidad de que el resultado de un individuo de prueba libre de la enfermedad sea positivo (es decir, p es la verdadera proporción de resultados de las pruebas de individuos libres de enfermedades que son positivos). Supongamos que sólo X está disponible en lugar de la secuencia real de los resultados de la prueba.

- **a.** Deduzca el estimador de máxima probabilidad de p. Si $n = 20$ y $x = 3$, ¿cuál es la estimación?
- **b.** ¿Es insesgado el estimador del inciso a)?
- **c.** Si $n = 20$ y $x = 3$, ¿cuál es el estimador de máxima probabilidad $(1 - p)^5$ de que ninguna de las próximas cinco pruebas realizadas en los individuos libres de la enfermedad sea positiva?

21. Si X tiene una distribución de Weibull con parámetros α y β, entonces

$$E(X) = \beta \cdot \Gamma(1 + 1/\alpha)$$
$$V(X) = \beta^2\{\Gamma(1 + 2/\alpha) - [\Gamma(1 + 1/\alpha)]^2\}$$

- **a.** Con base en una muestra aleatoria $X_1, \ldots, X_n$, escriba ecuaciones para el método de estimadores de momentos de β y α. Demuestre que una vez que se obtiene la estimación de α, la estimación de β se puede hallar en una tabla de la función gamma, y que la estimación de α es la solución de una ecuación complicada que implica la función gamma.
- **b.** Si $n = 20$, $\bar{x} = 28.0$ y $\Sigma x_i^2 = 16\,500$, calcule las estimaciones. [*Sugerencia*: $[\Gamma(1.2)]^2/\Gamma(1.4) = 0.95$.]

22. Sea X la proporción de tiempo asignado que un estudiante seleccionado al azar pasa resolviendo cierta prueba de aptitud. Suponga que la función de densidad de probabilidad de X es

$$f(x; \theta) = \begin{cases} (\theta + 1)x^\theta & 0 \le x \le 1 \\ 0 & \text{de lo contrario} \end{cases}$$

donde $-1 < \theta$. Una muestra aleatoria de diez estudiantes produce los datos $x_1 = 0.92$, $x_2 = 0.79$, $x_3 = 0.90$, $x_4 = 0.65$, $x_5 = 0.86$, $x_6 = 0.47$, $x_7 = 0.73$, $x_8 = 0.97$, $x_9 = 0.94$, $x_{10} = 0.77$.

- **a.** Use el método de momentos para obtener un estimador de θ y luego calcule la estimación para estos datos.
- **b.** Obtenga el estimador de máxima probabilidad de θ y luego calcule la estimación para los datos dados.

23. Sea que X represente el error al hacer una medición de una característica o propiedad física (por ejemplo, el punto de ebullición de un líquido especial). A menudo es lógico suponer que $E(X) = 0$ y que X tiene una distribución normal. Así, la función de probabilidad de cualquier error de medición particular es

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta} \qquad -\infty < x < \infty$$

(donde hemos usado θ en lugar de σ^2). Ahora, suponga que se realizan n mediciones independientes, dando como resultado errores de medida $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. Obtenga el estimador de máxima probabilidad de θ.

24. Un vehículo con un defecto particular en su sistema de control de emisiones es llevado a una serie de mecánicos seleccionados al azar hasta que $r = 3$ de ellos han diagnosticado correctamente el problema. Supongamos que esto requiere los diagnósticos de 20 mecánicos diferentes (por lo que hubo 17 diagnósticos incorrectos). Sea $p = P$(diagnóstico correcto), por tanto, el estimador de máxima probabilidad es la proporción de todos los mecánicos que bien podrían diagnosticar el problema. ¿Cuál es el estimador de máxima probabilidad de p? ¿Es el mismo estimador de máxima probabilidad si una muestra aleatoria de 20 mecánicos resulta en tres diagnósticos correctos? Explique. ¿Cómo funciona el estimador de máxima probabilidad en comparación con la estimación resultante de utilizar el estimador imparcial dada en el ejercicio 17?

25. Se determina la resistencia al esfuerzo cortante de las soldaduras de 10 puntos de prueba y se obtienen los siguientes datos (lb/pulg^2):

392 376 401 367 389 362 409 415 358 375

- **a.** Suponiendo que la resistencia al esfuerzo cortante está normalmente distribuida, estime la resistencia al esfuerzo cortante promedio verdadera y la desviación estándar de la resistencia al esfuerzo cortante mediante el método de máxima probabilidad.
- **b.** De nuevo, suponiendo una distribución normal, calcule el valor de resistencia por debajo del cual estará 95% de las resistencias de todas las soldaduras. [*Sugerencia*: ¿Cuál es el 95° percentil en función de μ y σ? Utilice ahora el principio de invarianza.]
- **c**. Suponga que decidimos examinar otro punto de soldadura de prueba. Sea X = resistencia al corte de la soldadura. Utilice los datos para obtener el estimador de máxima probabilidad de $P(X \le 400)$. [*Sugerencia*: $P(X \le 400) = \Phi((400 - \mu)/\sigma)$.]

26. Considere seleccionar aleatoriamente n segmentos de tubo y determinar la pérdida por corrosión (mm) en el espesor de pared de cada uno. Denote estas pérdidas de corrosión por $Y_1, \ldots, Y_n$. El artículo **"A Probabilistic Model for a Gas Explosion Due to Leakages in the Grey Cast Iron Gas Mains" (*Reliability Engr. and System Safety*, 2013: 270–279)** propone un modelo lineal de la corrosión: $Y_i = t_i R$, donde t_i es la edad de la tubería y R la velocidad de corrosión, se distribuye exponencialmente con parámetro λ. Obtenga el estimador de máxima probabilidad del parámetro exponencial (el estimador de máxima probabilidad resultante aparece en el artículo citado). [*Sugerencia*: Si $c > 0$ y X tiene una distribución exponencial, también cX la tiene.]

27. Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución gamma con parámetros α y β.

- **a.** Deduzca las ecuaciones cuyas soluciones dan los estimadores de máxima probabilidad de α y β. ¿Piensa que pueden ser resueltos explícitamente?

b. Demuestre que el estimador de máxima probabilidad de $\mu = \alpha\beta$ es $\hat{\mu} = \overline{X}$.

28. Si $X_1, X_2, \ldots, X_n$ representan una muestra aleatoria de la distribución de Rayleigh con función de densidad dada en el ejercicio 15, determine:

a. El estimador de máxima probabilidad de θ y luego calcule la estimación para los datos de esfuerzo de vibración dados en ese ejercicio. ¿Es este estimador el mismo estimador insesgado que se sugiere en el ejercicio 15?

b. El estimador de máxima probabilidad de la mediana de la distribución del esfuerzo de vibración. [*Sugerencia*: Exprese primero la mediana en función de θ.]

29. Considere la muestra aleatoria $X_1, X_2, \ldots, X_n$ de la función de densidad de probabilidad exponencial desplazada

$$f(x; \lambda, \theta) = \begin{cases} \lambda e^{-\lambda(x-\theta)} & x \geq \theta \\ 0 & \text{de lo contrario} \end{cases}$$

Con $\theta = 0$ da la función de densidad de probabilidad de la distribución exponencial que se consideró previamente (con densidad positiva a la derecha de cero). Un ejemplo de la distribución exponencial desplazada apareció en el ejemplo 4.5, en el cual la variable de interés fue el tiempo entre vehículos en el flujo de tránsito y $\theta = 0.5$ fue el tiempo entre vehículos mínimo posible.

a. Obtenga los estimadores de máxima probabilidad de θ y λ.

b. Si se realizan entre vehículos $n = 10$ observaciones de tiempo y se obtienen los siguientes resultados 3.11, 0.64, 2.55, 2.20, 5.44, 3.42, 10.39, 8.93, 17.82 y 1.30, calcule las estimaciones de θ y λ.

30. En el instante $t = 0$ son puestos a prueba 20 componentes idénticos. La distribución de la vida útil de cada uno es exponencial con parámetro λ. El experimentador deja la instalación de prueba sin supervisar. A su regreso, 24 horas más tarde, el experimentador termina de inmediato la prueba después de notar que $y = 15$ de los 20 componentes aún están en operación (así que 5 han fallado). Obtenga el estimador de máxima probabilidad de l. [*Sugerencia*: Sea $Y =$ el número que sobrevive 24 horas. En ese caso $Y \sim \text{Bin}(n, p)$. ¿Cuál es el estimador de máxima probabilidad de p? Observe ahora que $p = P(X_i \geq 24)$ donde X_i está exponencialmente distribuida. Esto relaciona a λ con p, de modo que el primero puede ser estimado una vez que lo ha sido el segundo.]

EJERCICIOS SUPLEMENTARIOS (31–38)

31. Se dice que un estimador $\hat{\theta}$ es **consistente** si con cualquier $\epsilon > 0$, $P(|\hat{\theta} - \theta| \geq \epsilon) \to 0$ a medida que $n \to \infty$. Es decir, $\hat{\theta}$ es consistente si, a medida que el tamaño de la muestra se hace más grande, es menos y menos probable que $\hat{\theta}$ se aleje más que ϵ del valor verdadero de θ. Demuestre que $\overline{X}$ es un estimador consistente de μ cuando $\sigma^2 < \infty$ mediante la desigualdad de Chebyshev del ejercicio 44 del capítulo 3. [*Sugerencia*: La desigualdad puede ser reescrita en la forma

$$P(|Y - \mu_Y| \geq \epsilon) \leq \sigma_Y^2/\epsilon$$

Ahora identifique Y con $\overline{X}$.]

32. **a.** Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución uniforme en $[0, \theta]$. Entonces el estimador de máxima probabilidad de θ es $\hat{\theta} = Y = \text{máx}(X_i)$. Use el hecho de que $Y \leq y$ si y sólo si cada $X_i \leq y$ para deducir la función de distribución acumulada de Y. Luego demuestre que la función de densidad de probabilidad de $Y = \text{máx}(X_i)$ es

$$f_Y(y) = \begin{cases} \dfrac{ny^{n-1}}{\theta^n} & 0 \leq y \leq \theta \\ 0 & \text{de lo contrario} \end{cases}$$

b. Use el resultado del inciso a) para demostrar que el estimador de máxima probabilidad es sesgado, pero que $(n + 1)\,\text{máx}(X_i)/n$ es insesgado.

33. En el instante $t = 0$ hay un individuo vivo en una población. Un **proceso de nacimientos puro** se desarrolla como sigue. El tiempo que transcurre hasta que se da el primer nacimiento está exponencialmente distribuido con parámetro λ. Después del primer nacimiento hay dos individuos vivos. El tiempo que transcurre hasta que el primer individuo da a luz otra vez es exponencial con parámetro λ, y del mismo modo para el segundo individuo. Por consiguiente, el tiempo que pasa hasta el siguiente nacimiento es el mínimo de dos variables (λ) exponenciales, el cual es exponencial con parámetro 2λ. Asimismo, una vez que el segundo nacimiento ha ocurrido, hay tres individuos vivos, de modo que el tiempo que transcurre hasta el siguiente nacimiento es una variable aleatoria exponencial con parámetro 3λ, y así sucesivamente (aquí se está utilizando la propiedad de no memoria de la distribución exponencial). Suponga que se observa el proceso hasta que ha ocurrido el sexto nacimiento y los tiempos hasta los nacimientos sucesivos son 25.2, 41.7, 51.2, 55.5, 59.5, 61.8 (con los cuales deberá calcular los tiempos entre nacimientos sucesivos). Obtenga el estimador de máxima probabilidad de λ. [*Sugerencia*: La probabilidad es un producto de términos exponenciales.]

34. El **error cuadrático medio** de un estimador $\hat{\theta}$ es $\text{MSE}(\hat{\theta}) = E(\hat{\theta}^2 - \theta^2)$. Si $\hat{\theta}$ es insesgado, entonces $\text{MSE}(\hat{\theta}) = V(\hat{\theta})$pero en general $\text{MSE}(\hat{\theta}) = V(\hat{\theta}) + (\text{sesgo})^2$. Considere el estimador $\hat{\sigma}^2 = KS^2$, donde $S^2 =$ varianza muestral. ¿Qué valor de K reduce al mínimo el error cuadrático medio de este estimador cuando la distribución de la población es normal? [*Sugerencia*: Se puede demostrar que

$$E[(S^2)^2] = (n + 1)\sigma^4/(n - 1)$$

En general, es difícil determinar $\hat{\theta}$ para reducir al mínimo el $\text{MSE}(\hat{\theta})$ por lo cual se buscan sólo estimadores insesgados y se reduce al mínimo $V(\hat{\theta})$.]

35. Sean $X_1, \ldots, X_n$ una muestra aleatoria de una función de densidad de probabilidad simétrica respecto a μ. Un estimador de μ que evidentemente funciona bien con una amplia variedad de distribuciones subyacentes es el *estimador de Hodges-Lehmann*. Para definir esto primero calcule para cada $i \leq j$ y cada $j = 1, 2, \ldots, n$ el promedio por pares $\overline{X}_{ij} = (X_i + X_j)/2$. Entonces el estimador es $\hat{\mu} =$ la mediana de las $\overline{X}_{ij}$. Calcule el valor de esta estimación con los datos del ejercicio 44 del capítulo 1. [*Sugerencia*: Construya una tabla cuadrada con las x_i en el margen izquierdo y en la parte superior. Luego calcule los promedios en la diagonal y encima de ella.]

36. Cuando la distribución de la población es normal se puede utilizar la mediana estadística $\{|X_1 - \widetilde{X}|, \ldots, |X_n - \widetilde{X}|\}/0.6745$ para estimar σ. Este estimador es más resistente a los efectos de los valores apartados (observaciones alejadas del grueso de los datos) que la desviación estándar muestral. Calcule tanto la estimación puntual correspondiente como s para los datos del ejemplo 5.2.

37. Cuando la desviación estándar muestral S está basada en una muestra aleatoria de una distribución de población normal se puede demostrar que

$$E(S) = \sqrt{2/(n-1)}\,\Gamma(n/2)\sigma/\Gamma((n-1)/2)$$

Use esto para obtener un estimador insesgado de σ de la forma cS. ¿Cuál es c cuando $n = 20$?

38. Cada uno de los n especímenes tiene que ser pesado dos veces en la misma báscula. Sean X_i y Y_i los dos pesos observados del i-ésimo espécimen. Suponga que X_i y Y_i son independientes uno de otro, cada uno normalmente distribuido con media μ_i (el peso verdadero del espécimen i) y varianza σ^2.

a. Demuestre que el estimador de probabilidad máxima de σ^2 es $\hat{\sigma}^2 = \Sigma(X_i - Y_i)^2/(4n)$. [*Sugerencia*: Si $\overline{z} = (z_1 - z_2)/2$, entonces $\Sigma(z_i - \overline{z})^2 = (z_1 - z_2)^2/2$.]

b. ¿Es el estimador de máxima probabilidad $\hat{\sigma}^2$ un estimador de σ^2? Determine un estimador insesgado de σ^2. [*Sugerencia*: Para cualquier variable aleatoria Z, $E(Z^2) = V(Z) - [E(Z)]^2$. Aplique esto a $Z = X_i - Y_i$.]

BIBLIOGRAFÍA

DeGroot, Morris y Mark Schervish, *Probability and Statistics* (4a. ed.), Addison-Wesley, Boston, MA, 2012. Incluye un excelente análisis tanto de propiedades generales como de métodos de estimación puntual; de particular interés son los ejemplos que muestran cómo los principios y métodos generales pueden dar estimadores insatisfactorios en situaciones particulares.

Devore, Jay y Kenneth Berk, *Modern Mathematical Statistics with Applications* (2ª ed.) Springer, Nueva York, 2012. La exposición es un poco más completa y compleja que la de este libro.

Efron, Bradley y Robert Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, Nueva York, 1993. La Biblia del *bootstrap*.

Hoaglin, David, Frederick Mosteller y John Tukey, *Understanding Robust and Exploratory Data Analysis*, Wiley, Nueva York, 1983. Contiene varios buenos capítulos sobre estimación puntual robusta, incluido uno sobre estimación *M*.

Rice, John, *Mathematical Statistics and Data Analysis* (3ª ed.), Thomson-Brooks/Cole, Belmont, CA, 2007. Una agradable combinación de teoría y datos estadísticos.

Capítulo

6

Intervalos estadísticos basados en una sola muestra

INTRODUCCIÓN

Una estimación puntual, por el hecho de ser un solo número, no proporciona información sobre la precisión y confiabilidad de la estimación. Considere, por ejemplo, utilizar el estadístico $\overline{X}$ para calcular una estimación puntual de la resistencia a la ruptura promedio verdadera (g) de las toallas de papel de cierta marca, y suponga que $\overline{x} = 9322.7$. Debido a la variabilidad del muestreo, virtualmente nunca es el caso de que $\overline{x} = \mu$. La estimación puntual no dice nada sobre qué tan cerca pudiera estar de μ. Una alternativa para reportar un solo valor sensible del parámetro que se está estimando es calcular y reportar un intervalo completo de valores factibles: *una estimación de intervalo* o un *intervalo de confianza* (IC). Un intervalo de confianza siempre se calcula al seleccionar primero un nivel de confianza, el cual mide el grado de confiabilidad del intervalo. Un intervalo de confianza con 95% de nivel de confianza de la resistencia a la ruptura promedio verdadera podría tener un límite inferior de 9162.5 y un límite superior de 9482.9. Entonces al nivel de confianza de 95%, cualquier valor de μ entre 9162.5 y 9482.5 es factible. Un nivel de confianza de 95% implica que 95% de todas las muestras daría un intervalo que incluye μ o cualquier otro parámetro que se esté estimando, y sólo 5% de las muestras daría un intervalo erróneo. Los niveles de confianza más frecuentemente utilizados son 95, 99 y 90%. Mientras más alto es el nivel de confianza, más fuerte es la creencia de que el valor del parámetro que se está estimando queda dentro del intervalo (en breve se dará una interpretación de cualquier nivel de confianza particular).

El ancho del intervalo proporciona información sobre la precisión de una estimación de intervalo. Si el nivel de confianza es alto y el intervalo resultante es bastante angosto, el conocimiento del valor del parámetro es razonablemente preciso. Un muy amplio intervalo de confianza, sin embargo, transmite el mensaje de que existe gran cantidad de incertidumbre sobre el

valor de lo que se está estimando. La figura 6.1 muestra intervalos de confianza de 95% para las resistencias a la ruptura promedio verdaderas de dos marcas diferentes de toallas de papel. Uno de estos intervalos sugiere un conocimiento preciso de μ, mientras que el otro sugiere un rango muy amplio de valores factibles.

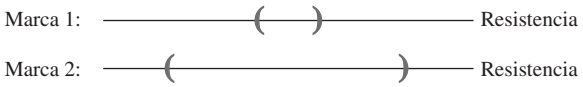


Figura 6.1 Intervalos de confianza que indican información precisa (marca 1) e imprecisa (marca 2) sobre μ

6.1 Propiedades básicas de los intervalos de confianza

Los conceptos y propiedades básicos de los intervalos de confianza son más fáciles de introducir si primero se presta atención a un problema simple, aunque un tanto irreal. Suponga que el parámetro de interés es una media poblacional μ y que

1. La distribución de la población es normal

2. El valor de la desviación estándar σ de la población es conocido

Con frecuencia es razonable suponer que la distribución de la población es normal. Sin embargo, si el valor de μ es desconocido, no es factible que el valor de σ esté disponible (el conocimiento del centro de una población en general precede a la información respecto a la dispersión). En las secciones 6.2 y 6.3 se desarrollarán métodos basados en suposiciones menos restrictivas.

EJEMPLO 6.1 Ingenieros industriales especialistas en ergonomía se ocupan del diseño de los espacios de trabajo y de los dispositivos operados por trabajadores con objeto de alcanzar una alta productividad y comodidad. El artículo **"Studies on Ergonomically Designed Alphanumeric Keyboards" (*Human Factors*, 1985: 175–187)** reporta sobre un estudio de la altura preferida de un teclado experimental con un gran soporte para el antebrazo y la muñeca. Se seleccionó una muestra de $n = 31$ mecanógrafos entrenados y se determinó la altura preferida del teclado de cada mecanógrafo. La altura preferida promedio muestral resultante fue de $\bar{x} = 80.0$ cm. Suponiendo que la altura preferida está normalmente distribuida con $\sigma = 2.0$ cm (un valor sugerido por los datos que aparecen en el artículo), obtenga un intervalo de confianza para μ, la altura promedio verdadera preferida por la población de todos los mecanógrafos experimentados. ■

Se supone que las observaciones muestrales reales $x_1, x_2, \ldots, x_n$ son el resultado de una muestra aleatoria $X_1, \ldots, X_n$ tomada de una distribución normal con media μ y desviación estándar σ. Los resultados del capítulo 5 implican entonces que independientemente del tamaño de muestra n, la media muestral $\overline{X}$ está normalmente distribuida con valor esperado μ y desviación estándar $\sigma/\sqrt{n}$. Si se estandariza $\overline{X}$ al restar primero su valor esperado y luego dividir entre su desviación estándar se obtiene la variable normal estándar

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \textbf{(6.1)}$$

Debido a que el área bajo la curva normal estándar entre −1.96 y 1.96 es 0.95,

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \tag{6.2}$$

A continuación manipule las desigualdades que están adentro del paréntesis en (6.2) de modo que aparezcan en la forma equivalente $l < \mu < u$, donde los puntos extremos l y u implican $\overline{X}$ y $\sigma/\sqrt{n}$. Esto se logra mediante la siguiente secuencia de operaciones, cada una de las cuales da desigualdades equivalentes a las originales.

1. Multiplíquese por $\sigma/\sqrt{n}$

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

2. Réstese $\overline{X}$ de cada término:

$$-\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

3. Multiplique por −1 para eliminar el signo menos enfrente de u (el cual invierte la dirección de cada desigualdad):

$$\overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} > \mu > \overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

es decir,

$$\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

La equivalencia de cada conjunto de desigualdades con el conjunto original implica que

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \tag{6.3}$$

El evento en el interior del paréntesis en (6.3) tiene una apariencia poco común; previamente, la cantidad aleatoria aparecía a la mitad con constantes en ambos extremos, como en $a \le Y \le b$. En (6.3) la cantidad aleatoria aparece en los dos extremos, mientras que la constante desconocida μ aparece a la mitad. Para interpretar (6.3), considere un **intervalo aleatorio** con el punto extremo izquierdo $\overline{X} - 1.96 \cdot \sigma/\sqrt{n}$ y punto extremo derecho $\overline{X} + 1.96 \cdot \sigma/\sqrt{n}$. En notación de intervalo, esto se transforma en

$$\left(\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}},\quad \overline{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{6.4}$$

El intervalo (6.4) es aleatorio porque sus dos puntos extremos implican una variable aleatoria. Está centrada en la media muestral $\overline{X}$ y se extiende $1.96\sigma/\sqrt{n}$ a cada lado de $\overline{X}$. Por consiguiente, el ancho del intervalo es $2 \cdot (1.96) \cdot \sigma/\sqrt{n}$, el cual no es aleatorio; sólo la localización del intervalo (su punto medio $\overline{X}$) lo es (figura 6.2). Ahora (6.3) se parafrasea como "*la probabilidad es 0.95 de que el intervalo aleatorio* (6.4) *incluya o abarque el valor verdadero de* μ". Antes de realizar cualquier experimento y de recolectar cualquier dato, es bastante probable que μ esté dentro del intervalo (6.4).

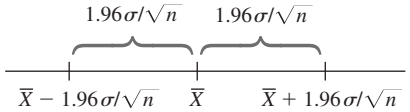


Figura 6.2 Intervalo aleatorio (6.4) con su centro en $\overline{X}$

DEFINICIÓN

Si después de observar $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, se calcula la media muestral observada $\bar{x}$ y luego se sustituye $\bar{x}$ en (6.4) en lugar de $\bar{X}$, al intervalo fijo resultante se le llama **intervalo de 95% de confianza para** μ. Este intervalo de confianza se expresa como

$$\left(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \text{ es un intervalo de confianza de 95\% para } \mu$$

o como

$$\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ con 95\% de confianza}$$

Una expresión concisa para el intervalo es $\bar{x} \pm 1.96 \cdot \sigma/\sqrt{n}$, donde $-$ da el punto extremo izquierdo (límite inferior) y $+$ da el punto extremo derecho (límite superior).

EJEMPLO 6.2 (Continuación del ejemplo 6.1)

Las cantidades requeridas para calcular el intervalo de confianza de 95% para la altura preferida promedio verdadera son $\sigma = 2.0$, $n = 31$ y $\bar{x} = 80.0$. El intervalo resultante es

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 80.0 \pm (1.96)\frac{2.0}{\sqrt{31}} = 80.0 \pm 0.7 = (79.3, 80.7)$$

Es decir, se puede estar totalmente confiado, en el nivel de confianza de 95%, de que $79.3 < \mu < 80.7$. Este intervalo es relativamente angosto, lo que indica que μ ha sido estimada con bastante precisión. ■

Interpretación de un intervalo de confianza

El nivel de confianza de 95% para el intervalo que se acaba de definir fue heredado del 0.95 de probabilidad para el intervalo aleatorio (6.4). Los intervalos con otros niveles de confianza serán introducidos más adelante. Por ahora, más bien, considere cómo se puede interpretar 95% de confianza.

Se inició con un evento cuya probabilidad era de 0.95 —que el intervalo aleatorio (6.4) capturaría el valor verdadero de μ— y luego se utilizaron los datos del ejemplo 6.1 para calcular el intervalo de confianza (79.3, 80.7). Es tentador concluir que μ está dentro de este intervalo fijo con probabilidad de 0.95. Pero al sustituir $\bar{x} = 80.0$ por $\bar{X}$, toda la aleatoriedad desaparece; el intervalo (79.3, 80.7) no es un intervalo aleatorio y μ es una constante (desafortunadamente desconocida). Por tanto, es *incorrecto* escribir la proposición $P(\mu$ queda en $(79.3, 80.7)) = 0.95$.

Una interpretación correcta de "95% de confianza" se basa en la interpretación de probabilidad de frecuencia relativa a largo plazo: decir que un evento A tiene una probabilidad de 0.95 es decir que si el experimento en el cual se definió A se realiza una y otra vez, a la larga A ocurrirá 95% del tiempo. Suponga que se obtiene otra muestra de alturas preferidas por los mecanógrafos y se calcula otro intervalo de 95%. Luego se considera repetir esto con una tercera muestra, una cuarta, una quinta, y así sucesivamente. Sea A el evento en que $\bar{X} - 1.96 \cdot \sigma/\sqrt{n} < \mu < \bar{X} + 1.96 \cdot \sigma/\sqrt{n}$. Ya que $P(A) = 0.95$, a la larga 95% de los intervalos de confianza calculados contendrán a μ. Esto se ilustra en la figura 6.3, donde la línea vertical corta el eje de medición en el valor verdadero (pero desconocido) de μ. Observe que 7 de los 100 intervalos mostrados fallan al contener a μ. A la larga, sólo 5% de los intervalos construidos así no contendrán a μ.

De acuerdo con esta interpretación, el nivel de confianza de 95% no es en sí una proposición sobre cualquier intervalo particular tal como (79.3, 80.7). En su lugar pertenece a lo que sucedería si se construyera un número muy grande de intervalos parecidos por medio de la misma fórmula de intervalo de confianza. Aunque esto puede parecer no satisfactorio, el origen de la dificultad yace en la interpretación de probabilidad, es válida

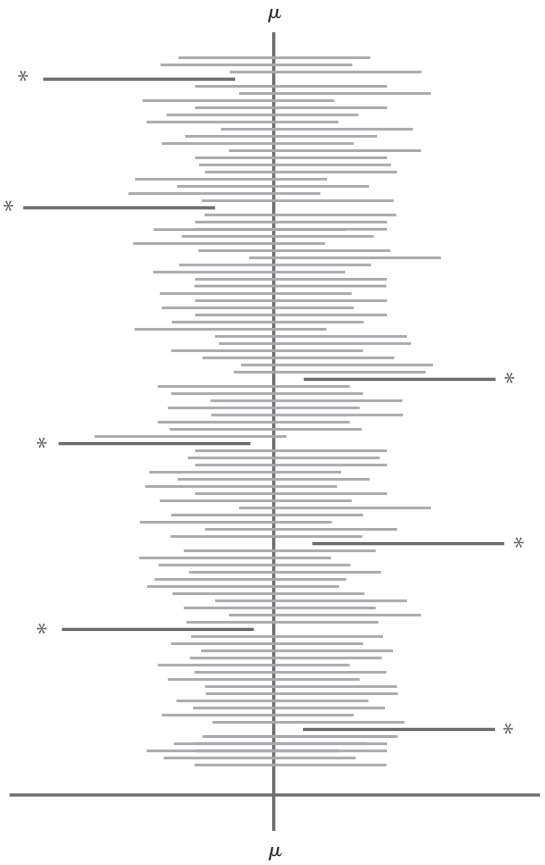


Figura 6.3 Gráfica de simulación aleatoria de cien niveles de confianza de 95% (los asteriscos identifican intervalos que no incluyen a μ).

para una larga secuencia de réplicas de un experimento en lugar de sólo para una. Existe otro método para abordar la construcción y la interpretación de intervalos de confianza que utiliza la noción de probabilidad subjetiva y el teorema de probabilidad de Bayes, aunque los detalles técnicos se salen del alcance de este libro; el libro de DeGroot y colaboradores (véase la bibliografía del capítulo 6) es una buena fuente. El intervalo presentado aquí (así como también cada intervalo presentado subsecuentemente) se llama intervalo de confianza "clásico" porque su interpretación se apoya en la noción clásica de probabilidad.

Otros niveles de confianza

El nivel de confianza de 95% fue heredado de la probabilidad de 0.95 de las desigualdades iniciales que aparecen en (6.2). Si se desea un nivel de confianza de 99% la probabilidad inicial de 0.95 debe ser reemplazada por 0.99, lo que implica cambiar el valor crítico z de 1.96 a 2.58. Un intervalo de confianza de 99% resulta entonces de utilizar el valor de 2.58 en lugar de 1.96 en la fórmula para el intervalo de confianza de 95%.

De hecho, cualquier nivel de confianza deseado se obtiene reemplazando 1.96 o 2.58 con el valor crítico normal estándar apropiado. Recuerde que en capítulo 4 se menciona la notación para un valor crítico z: valor crítico z_α es el número en la escala horizontal z que captura el área de la cola superior. Tal como lo muestra la figura 6.4, utilizando $z_{\alpha/2}$ en lugar de 1.96 se logra una probabilidad (es decir, el área central de la curva z) de $1 - \alpha$.

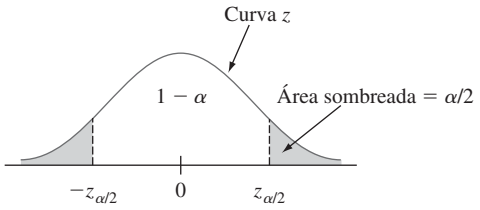


Figura 6.4 $P(-z_{\alpha/2} < Z, z_{\alpha/2}) = 1 - \alpha$

DEFINICIÓN

La siguiente expresión da un **intervalo de confianza de 100(1 − α)%** para la media μ de una población normal cuando se conoce el valor de σ

$$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \qquad \textbf{(6.5)}$$

o, de forma equivalente, por $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

La fórmula (6.5) para el intervalo de confianza también se puede expresar en palabras como

estimación puntual de $\mu \pm$ (valor crítico z) (error estándar de la media).

EJEMPLO 6.3 No hace mucho tiempo que el proceso de producción de una caja de control de un tipo particular para un motor fue modificado. Antes de esta modificación, los datos históricos sugirieron que la distribución de los diámetros de agujeros para bujes en las cajas era normal con desviación estándar de 0.100 mm. Se cree que la modificación no ha afectado la forma de la distribución o la desviación estándar, pero que el valor del diámetro medio pudo cambiar. Se selecciona una muestra de 40 cajas y se determina el diámetro de agujero para cada una, y el resultado es un diámetro medio muestral de 5.426 mm. Calcule un intervalo de confianza para el diámetro de agujero promedio verdadero utilizando un nivel de confianza de 90%. Esto requiere que $100(1 - \alpha) = 90$, de donde $\alpha = 0.10$ y $z_{\alpha/2} = z_{.05} = 1.645$ (correspondiente a un área de curva z acumulada de 0.9500). El intervalo deseado es entonces

$$5.426 \pm (1.645)\frac{0.100}{\sqrt{40}} = 5.426 \pm 0.026 = (5.400, 5.452)$$

Con un razonable alto grado de confianza se puede decir que $5.400 < \mu < 5.452$. Este intervalo es algo angosto debido a la pequeña cantidad de variabilidad del diámetro del agujero ($\sigma = 0.100$). ■

Nivel de confianza, precisión y tamaño de la muestra

¿Por qué decidirse por un nivel de confianza de 95% cuando un nivel de 99% es alcanzable? Porque el precio pagado por el nivel de confianza más alto es un intervalo más ancho. Ya que el intervalo de 95% se extiende $1.96 \cdot \sigma/\sqrt{n}$ a cada lado de $\bar{x}$, el ancho del intervalo es $2(1.96) \cdot \sigma/\sqrt{n} = 3.92 \cdot \sigma/\sqrt{n}$. Asimismo, el ancho del intervalo de 99% es $2(2.58) \cdot \sigma/\sqrt{n} = 5.16 \cdot \sigma/\sqrt{n}$. Es decir, se tiene más confianza en el intervalo de 99% precisamente porque es más ancho. Mientras más alto es el grado de confianza deseado, más ancho es el intervalo resultante.

Si se considera que el ancho del intervalo especifica su precisión, entonces el nivel de confianza (o confiabilidad) del intervalo está relacionado de manera inversa con su precisión. La estimación de un intervalo altamente confiable puede ser imprecisa por el hecho de que los puntos extremos del intervalo pueden estar muy alejados, mientras que un intervalo preciso puede acarrear una confiabilidad relativamente baja. Por consiguiente, no se puede

decir de modo inequívoco que se tiene que preferir un intervalo de 99% a uno de 95%; la ganancia de confiabilidad acarrea una pérdida de precisión.

Una estrategia atractiva es especificar tanto el nivel de confianza deseado como el ancho del intervalo y luego determinar el tamaño de muestra necesario.

EJEMPLO 6.4 Un monitoreo exhaustivo de un sistema de tiempo compartido de computadoras sugiere que el tiempo de respuesta a un comando de edición particular está normalmente distribuido con desviación estándar de 25 milisegundos. Se instaló un nuevo sistema operativo y se desea estimar el tiempo de respuesta promedio verdadero μ en el nuevo entorno. Suponiendo que los tiempos de respuesta siguen estando normalmente distribuidos con $\sigma = 25$, ¿qué tamaño de muestra es necesario para asegurarse de que el intervalo de confianza de 95% resultante tiene un ancho (a lo más) de 10? El tamaño de muestra n debe satisfacer

$$10 = 2 \cdot (1.96)(25/\sqrt{n})$$

Reordenando esta ecuación se obtiene

$$\sqrt{n} = 2 \cdot (1.96)(25)/10 = 9.80$$

por consiguiente

$$n = (9.80)^2 = 96.04$$

En vista de que n debe ser un entero, se requiere un tamaño de muestra de 97. ■

Una fórmula general para el tamaño de muestra n necesario para garantizar un ancho de intervalo w se obtiene igualando w a $2 \cdot z_{\alpha/2} \cdot \sigma/\sqrt{n}$ y despejando n.

El tamaño de muestra necesario para que el intervalo de confianza (6.5) dé un ancho w es

$$n = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

Mientras más pequeño es el ancho deseado w, más grande debe ser n. Además, n es una función creciente de σ (más variabilidad de la población requiere un tamaño de muestra más grande) y del nivel de confianza $100(1 - \alpha)$ (conforme α decrece, $z_{\alpha/2}$ se incrementa).

A la mitad del ancho $1.96/\sqrt{n}$ del intervalo de confianza de 95% en ocasiones se le llama **límite en el error de estimación** asociado con un nivel de confianza de 95%. Es decir, con 95% de confianza la estimación puntual x no estará a más de esta distancia de μ. Antes de obtener datos, es posible que un investigador desee determinar un tamaño de muestra con el cual se logre un valor particular del límite. Por ejemplo, si μ representa la eficiencia de combustible promedio (mpg) de todos los vehículos de cierto tipo, el objetivo de una investigación puede ser estimar μ dentro de 1 mpg con 95% de confianza. Más generalmente, si se desea estimar μ dentro de una cantidad β (el límite especificado en el error de estimación) con confianza de $100(1 - \alpha)\%$, el tamaño de muestra necesario se obtiene al reemplazar $2/w$ por $1/\beta$ en el recuadro precedente.

Deducción de un intervalo de confianza

Sean $X_1, X_2, \ldots, X_n$ la muestra en la cual se tiene que basar el intervalo de confianza para un parámetro θ. Suponga que se puede determinar una variable aleatoria que satisfaga las siguientes dos propiedades:

1. La variable depende funcionalmente tanto de $X_1, \ldots, X_n$ como de θ.
2. La distribución de probabilidad de la variable no depende de θ ni de cualesquiera otros parámetros desconocidos.

Sea $h(X_1, X_2, \ldots, X_n; \theta)$ esta variable aleatoria. Por ejemplo, si la distribución de la población es normal con σ y $\theta = \mu$ conocidos, la variable $h(X_2, \ldots, X_n; \mu) = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ satisface ambas propiedades; claramente depende funcionalmente de μ, no obstante su distribución de probabilidad normal estándar, la cual no depende de μ. En general, la forma de la función h casi siempre se pone de manifiesto al examinar la distribución de un estimador apropiado $\hat{\theta}$.

Con cualquier α entre 0 y 1, se ve que las constantes a y b satisfacen

$$P(\alpha < h(X_1,\ldots, X_n; \theta) < b) = 1 - \alpha \tag{6.6}$$

A causa de la segunda propiedad, a y b no dependen de θ. En el ejemplo normal, $a = -z_{\alpha/2}$ y $b = z_{\alpha/2}$. Ahora suponga que las desigualdades en (6.6) pueden ser manipuladas para aislar θ, y así obtener la proposición de probabilidad equivalente

$$P(l(X_1, X_2,\ldots, X_n) < \theta < u(X_1, X_2,\ldots, X_n)) = 1 - \alpha$$

Entonces $l(x_1, x_2, \ldots, x_n)$ y $u(x_1, \ldots, x_n)$ son los límites de confianza inferior y superior, respectivamente, para un intervalo de confianza de $100(1 - \alpha)\%$. En el ejemplo normal, se vio que $l(X_1,\ldots, X_n) = \overline{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n}$ y $u(X_1,\ldots, X_n) = \overline{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

EJEMPLO 6.5 Un modelo teórico sugiere que el tiempo hasta la ruptura de un fluido aislante entre electrodos a un voltaje particular tiene una distribución exponencial con parámetro λ (véase la sección 4.4). Una muestra aleatoria de $n = 10$ tiempos de ruptura proporciona los siguientes datos muestrales (en min): $x_1 = 41.53$, $x_2 = 18.73$, $x_3 = 2.99$, $x_4 = 30.34$, $x_5 = 12.33$, $x_6 = 117.52$, $x_7 = 73.02$, $x_8 = 223.63$, $x_9 = 4.00$, $x_{10} = 26.78$. Se desea un intervalo de confianza de 95% para λ y para el tiempo de ruptura promedio verdadero.

Sea $h(X_1, X_2, \ldots, X_n; \lambda) = 2\lambda\Sigma X_i$. Se puede demostrar que esta variable aleatoria tiene una distribución de probabilidad llamada distribución ji-cuadrada con $2n$ grados de libertad (gl) ($\nu = 2n$, donde ν es el parámetro de una distribución ji-cuadrada como se menciona en la sección 4.4). La tabla A.7 del apéndice ilustra una curva de densidad ji-cuadrada típica y tabula valores críticos que capturan áreas de colas específicas. El número pertinente de grados de libertad en este caso es 2(10)=20. La fila $\nu = 20$ de la tabla muestra que 34.170 captura un área de cola superior de 0.025, y que 9.591 captura un área de cola inferior de 0.025 (área de cola superior de 0.975). Por consiguiente, con $n = 10$,

$$P\left(9.591 < 2\lambda\sum X_i < 34.170\right) = 0.95$$

La división entre $2\Sigma X_i$ aísla λ y se obtiene

$$P\left(9.591/\left(2\sum X_i\right) < \lambda < \left(34.170/\left(2\sum X_i\right)\right) = 0.95\right.$$

El límite inferior del intervalo de confianza de 95% para λ es $9.591/(2\Sigma x_i)$, y el límite superior es $34.170/(2\Sigma x_i)$. Con los datos dados $\Sigma x_i = 550.87$ proporciona el intervalo (0.00871, 0.03101).

El valor esperado de una variable aleatoria exponencial es $\mu = 1/\lambda$. Puesto que

$$P\left(2\sum X_i/34.170 < 1/\lambda < 2\sum X_i/9.591\right) = 0.95$$

el intervalo de confianza de 95% para el tiempo de ruptura promedio verdadero es $(2\Sigma x_i/34.170, 2\Sigma x_i/9.591) = (32.24, 114.87)$. Obviamente este intervalo es bastante ancho, lo que refleja una variabilidad sustancial de los tiempos de ruptura y un pequeño tamaño de muestra. ■

En general, los límites de confianza superior e inferior resultan de reemplazar cada $<$ en (7.6) por $=$ y al resolver para θ. En el ejemplo del fluido aislante que se acaba de considerar, $2\lambda\Sigma x_i = 34.170$ da $\lambda = 34.170/(2\Sigma x_i)$ como límite de confianza superior y el

límite inferior se obtiene con la otra ecuación. Observe que los dos límites de intervalo no están equidistantes de la estimación puntual, en vista de que el intervalo no es de la forma $\hat{\theta} \pm c$.

Intervalos de confianza *bootstrap*

La técnica *bootstrap* es una forma de estimar $\sigma_{\hat{\theta}}$. También puede ser aplicada para obtener un intervalo de confianza para θ. Considere de nuevo la estimación de la media μ de una distribución normal cuando se conoce σ. Reemplace μ con θ y use $\hat{\theta} = \overline{X}$ como estimador puntual. Observe que $1.96\sigma/\sqrt{n}$ es el 97.5º percentil de la distribución de $\hat{\theta} - \theta$ [esto es, $P(\overline{X} - \mu < 1.96\sigma/\sqrt{n}) = P(Z < 1.96) = 0.9750$]. Del mismo modo, $-1.96\sigma/\sqrt{n}$ es el 2.5º percentil, por consiguiente

$$\begin{aligned} 0.95 &= P(2.5^{\text{o}} \text{ percentil} < \hat{\theta} = \theta < 97.5^{\text{o}} \text{ percentil}) \\ &= P(\hat{\theta} - 2.5^{\text{o}} \text{ percentil} > \theta > \hat{\theta} - 97.5^{\text{o}} \text{ percentil}) \end{aligned}$$

Es decir, con

$$\begin{aligned} l &= \hat{\theta} - 97.5^{\text{o}} \text{ percentil de } \hat{\theta} = \theta \\ u &= \hat{\theta} - 2.5^{\text{o}} \text{ percentil de } \hat{\theta} = \theta \end{aligned} \qquad \textbf{(6.7)}$$

el intervalo de confianza para θ es (l, u). En muchos casos, los percentiles en (6.7) no pueden ser calculados, pero sí *pueden* ser estimados con muestras *bootstrap*. Suponga que se obtienen $B = 1000$ muestras *bootstrap* y se calculan $\hat{\theta}_1^*, \ldots, \hat{\theta}_{1000}^*$ y $\overline{\theta}^*$ seguidos por las 1000 diferencias $\hat{\theta}_1^* - \hat{\theta}^*, \ldots, \hat{\theta}_{1000}^* - \hat{\theta}^*$. La 25ª más grande y la 25ª más pequeña de estas diferencias son estimaciones de los percentiles desconocidos en (6.7). Para más información consulte los libros de Devore y Berk o el de Efron citados en el capítulo 6.

EJERCICIOS Sección 6.1 (1–11)

1. Considere una distribución de población normal con el valor de σ conocido.
- **a.** ¿Cuál es el nivel de confianza para el intervalo $\overline{x} \pm 2.81\sigma/\sqrt{n}$?
- **b.** ¿Cuál es el nivel de confianza para el intervalo $\overline{x} \pm 1.44\sigma/\sqrt{n}$?
- **c.** ¿Qué valor de $z_{\alpha/2}$ en la fórmula de intervalo de confianza (6.5) da un nivel de confianza de 99.7%?
- **d.** Responda la pregunta del inciso c) para un nivel de confianza de 75%.

2. Cada uno de los siguientes es un intervalo de confianza para μ = frecuencia verdadera (es decir, media de la población) de resonancia promedio (Hz) para todas las raquetas de tenis de un tipo:

(114.4, 115.6) (114.1, 115.9)
- **a.** ¿Cuál es el valor de la frecuencia de resonancia media muestral?
- **b.** Ambos intervalos se calcularon con los mismos datos muestrales. El nivel de confianza para uno de estos intervalos es de 90% y para el otro es de 99%. ¿Cuál de los intervalos tiene el nivel de confianza de 90% y por qué?

3. Suponga que se selecciona una muestra aleatoria de 50 botellas de una marca particular de jarabe para la tos y se determina el contenido de alcohol de cada una. Sea μ el contenido promedio de alcohol de la población de todas las botellas de la marca estudiada. Suponga que el intervalo de confianza de 95% resultante es (7.8, 9.4).
- **a.** Un intervalo de confianza de 90%, calculado con esta muestra, ¿habría resultado más angosto o más ancho que el intervalo dado? Explique su razonamiento.
- **b.** Considere la siguiente proposición: existe 95% de probabilidades de que μ esté entre 7.8 y 9.4. ¿Es correcta esta proposición? ¿Por qué sí o por qué no?
- **c.** Considere la siguiente proposición: se puede tener la total confianza de que 95% de todas las botellas de este tipo de jarabe para la tos tienen un contenido de alcohol de entre 7.8 y 9.4. ¿Es correcta esta proposición? ¿Por qué sí o por qué no?
- **d.** Considere la siguiente proposición: si el proceso de selección de una muestra de tamaño 50 y el cálculo del intervalo de 95% correspondiente se repite 100 veces, 95 de los intervalos resultantes incluirán μ. ¿Es correcta esta proposición? ¿Por qué sí o por qué no?

4. Se desea un intervalo de confianza para la pérdida por carga parásita promedio verdadera μ (watts) de cierto tipo de motor de inducción cuando la corriente a través de la línea se mantiene en 10 amps a una velocidad de 1500 rpm. Suponga que la pérdida por carga parásita está normalmente distribuida con $\sigma = 3.0$.

- **a.** Calcule un intervalo de confianza de 95% para μ cuando $n = 25$ y $\bar{x} = 58.3$.
- **b.** Calcule un intervalo de confianza de 95% para μ cuando $n = 100$ y $\bar{x} = 58.3$.
- **c.** Calcule un intervalo de confianza de 99% para μ cuando $n = 100$ y $\bar{x} = 58.3$.
- **d.** Calcule un intervalo de confianza de 82% para μ cuando $n = 100$ y $\bar{x} = 58.3$.
- **e.** ¿Qué tan grande debe ser n si el ancho del intervalo de 99% para μ tiene que ser 1.0?

5. Suponga que la porosidad al helio (en porcentaje) de muestras de carbón tomadas de cualquier costura particular está normalmente distribuida con desviación estándar verdadera de 0.75.

- **a.** Calcule un intervalo de confianza de 95% para la porosidad promedio verdadera de una costura, si la porosidad promedio en 20 especímenes de la costura fue de 4.85.
- **b.** Calcule un intervalo de confianza de 98% para la porosidad promedio verdadera de otra costura basada en 16 especímenes con porosidad promedio muestral de 4.56.
- **c.** ¿Qué tan grande debe ser un tamaño de muestra si el ancho del intervalo de 95% tiene que ser de 0.40?
- **d.** ¿Qué tamaño de muestra se necesita para estimar la porosidad promedio verdadera dentro de 0.2 con confianza de 99%?

6. Con base en pruebas extensas se sabe que el límite de elasticidad de un tipo particular de varilla de refuerzo de acero suave está normalmente distribuido con $\sigma = 100$. La composición de la varilla se modificó un poco, pero no se cree que la alteración haya afectado la normalidad o el valor de σ.

- **a.** Suponiendo que es este el caso si una muestra de 25 varillas modificadas dio por resultado un límite de elasticidad promedio muestral de 8439 lb, calcule un intervalo de confianza de 90% para el punto de elasticidad promedio verdadero de la varilla modificada.
- **b.** ¿Cómo modificaría el intervalo del inciso a) para obtener un nivel de confianza de 92%?

7. ¿En cuánto se debe incrementar el tamaño de muestra n si el ancho del intervalo de confianza (6.5) tiene que ser reducido a la mitad? Si el tamaño de muestra se incrementa por un factor de 25, ¿qué efecto tendrá esto en el ancho del intervalo? Justifique sus aseveraciones.

8. Sea $\alpha_1 > 0$, $\alpha_2 > 0$, con $\alpha_1 + \alpha_2 = \alpha$. Entonces

$$P\left(-z_{\alpha_1} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha_2}\right) = 1 - \alpha$$

- **a.** Use esta ecuación para obtener una expresión más general para un intervalo de confianza $100(1 - \alpha)\%$ para μ del cual el intervalo (7.5) es un caso especial.
- **b.** Sea $\alpha = 0.05$ y $\alpha_1 = \alpha/4$, $\alpha_2 = 3\alpha/4$. ¿Resulta esto en un intervalo más angosto o más ancho que el intervalo (6.5)?

9. **a.** En las mismas condiciones que aquellas que conducen al intervalo (6.5), $P[(\overline{X} - \mu)/(\sigma/\sqrt{n}) < 1.645] = 0.95$. Use esta expresión para deducir un intervalo unilateral para μ de ancho infinito y que proporcione un límite de confianza inferior para μ. ¿Cuál es el intervalo para los datos del ejercicio 5(a)?

- **b.** Generalice el resultado del inciso a) para obtener un límite inferior con nivel de confianza de $100(1 - \alpha)\%$.
- **c.** ¿Cuál es un intervalo análogo al del inciso b) que proporcione un límite superior para μ? Calcule este intervalo de 99% para los datos del ejercicio 4(a).

10. Una muestra aleatoria de $n = 15$ bombas térmicas de cierto tipo produjo las siguientes observaciones de vida útil (en años):

2.0	1.3	6.0	1.9	5.1	0.4	1.0	5.3
15.7	0.7	4.8	0.9	12.2	5.3	0.6	

- **a.** Suponga que la distribución de la vida útil es exponencial y use un argumento paralelo al del ejemplo 6.5 para obtener un intervalo de confianza de 95% para la vida útil esperada (promedio verdadero).
- **b.** ¿Cómo debería modificarse el intervalo del inciso a) para obtener un nivel de confianza de 99%?
- **c.** ¿Cuál es un intervalo de confianza de 95% para la desviación estándar de la distribución de la vida útil? [*Sugerencia:* ¿Cuál es la desviación estándar de una variable aleatoria exponencial?]

11. Considere los siguientes 1000 intervalos de confianza de 95% para μ que un consultor estadístico obtendrá para varios clientes. Suponga que los conjuntos de datos en los cuales están basados los intervalos se seleccionan independientemente uno de otro. ¿Cuántos de estos 1000 intervalos espera que capturen el valor correspondiente de μ? ¿Cuál es la probabilidad de que entre 940 y 960 de estos intervalos contengan el valor correspondiente de μ? [*Sugerencia*: Sea Y = el número entre los 1000 intervalos que contienen a μ. ¿Qué clase de variable aleatoria es Y?]

6.2 Intervalos de confianza de muestra grande para una media y para una proporción de población

En el intervalo de confianza para μ, dado en la sección previa, se supuso que la distribución de la población es normal con el valor de σ conocido. A continuación se presenta un intervalo de confianza de muestra grande cuya validez no requiere estas suposiciones. Después de demostrar cómo el argumento que lleva a este intervalo se aplica en forma extensa para producir otros intervalos de muestra grande, habrá que enfocarse en un intervalo para una proporción p de población.

Intervalo de muestra grande para μ

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población con media μ y desviación estándar σ. Siempre que n es grande, el teorema del límite central implica que $\overline{X}$ tiene de manera aproximada una distribución normal, cualquiera que sea la naturaleza de la distribución de la población. Se deduce entonces que $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ tiene aproximadamente una distribución estándar normal, de modo que

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

Un argumento paralelo al dado en la sección 6.1 da $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ como intervalo de confianza de muestra grande para μ con un nivel de confianza de *aproximadamente* $100(1 - \alpha)\%$. Es decir, cuando n es grande el intervalo de confianza para μ dado antes permanece válido, cualquiera que sea la distribución de la población, siempre que el calificador esté insertado "aproximadamente" enfrente del nivel de confianza.

Una dificultad práctica con este desarrollo es que el cálculo del intervalo de confianza requiere el valor de σ, el cual rara vez es conocido. Considere reemplazar la desviación estándar de la población σ en Z por la desviación estándar de la muestra para obtener la variable estandarizada

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Previamente hubo aleatoriedad sólo en el numerador de Z gracias a $\overline{X}$. En la nueva variable estandarizada, tanto $\overline{X}$ *como* S cambian de valor de una muestra a otra. Así que aparentemente la distribución de la nueva variable deberá estar más dispersa que la curva z para reflejar la variación extra en el denominador. Esto en realidad es cierto cuando n es pequeño. Sin embargo, con n grande la sustitución de S en lugar de σ agrega un poco de variabilidad extra, así que esta variable también tiene aproximadamente una distribución normal estándar. La manipulación de la variable en la proposición de probabilidad, como en el caso de la σ conocida, da un intervalo de confianza de muestra grande general para μ.

PROPOSICIÓN

Si n es suficientemente grande, la variable estandarizada

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

tiene aproximadamente una distribución normal estándar. Esto implica que

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \qquad \textbf{(6.8)}$$

es un **intervalo de confianza de muestra grande para μ** con nivel de confianza aproximadamente de $100(1 - \alpha)\%$. Esta fórmula es válida sin importar la forma de la distribución de la población.

Es decir, el intervalo de confianza (6.8) es

la estimación puntual de $\mu \pm$ (z valor crítico) (error estándar estimado de la media).

En general, $n > 40$ será suficiente para justificar el uso de este intervalo. Esto es algo más conservador que la regla empírica del teorema del límite central, debido a la variabilidad adicional introducida por el uso de S en lugar de σ.

EJEMPLO 6.6 ¿Nunca ha deseado tener un Porsche? El autor piensa que tal vez podía permitirse un Boxster, el modelo más barato. Así que entró a www.cars.com el 18 de noviembre de 2009 y encontró un total de 1113 automóviles de este tipo en la lista. Los precios iban de $3499 a

$130 000 (el precio de este último fue uno de los dos que excedían los $70 000). Los precios lo deprimieron, por lo que mejor se centró en las lecturas del odómetro (millas). Aquí se presentan las lecturas de una muestra de 50 de estos Boxster:

2948	2996	7197	8338	8500	8759	12710	12925
15767	20000	23247	24863	26000	26210	30552	30600
35700	36466	40316	40596	41021	41234	43000	44607
45000	45027	45442	46963	47978	49518	52000	53334
54208	56062	57000	57365	60020	60265	60803	62851
64404	72140	74594	79308	79500	80000	80000	84000
113000	118634						

Una gráfica de caja de los datos (figura 6.5) muestra que a excepción de los dos valores límite en el extremo superior la distribución de valores es bastante simétrica (de hecho, una gráfica de probabilidad normal muestra un patrón bastante lineal, aunque los puntos correspondientes a las dos observaciones más pequeñas y a las dos mayores están un tanto alejadas de un ajuste lineal a través de los puntos restantes).

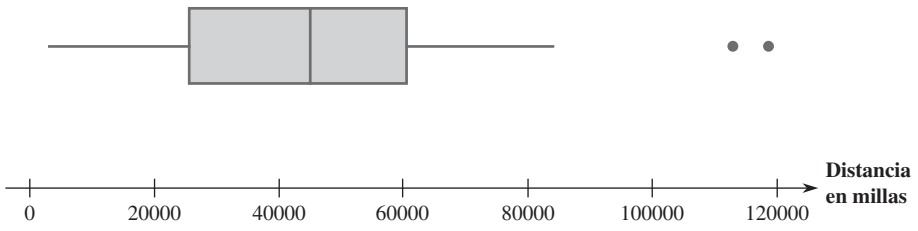


Figura 6.5 Gráfica de caja para las lecturas del odómetro del ejemplo 7.6

Las cantidades resumidas incluyen $n = 50$, $\bar{x} = 45\ 679.4$, $\tilde{x} = 45\ 013.5$, $s = 26\ 641.675$, $f_s = 34\ 265$. La media y la mediana están relativamente cerca (si los dos valores mayores fueran reducidos por 30 000, la media bajaría a 44 479.4, mientras que la mediana no se vería afectada). La gráfica de caja y las magnitudes de s y f_s respecto a la media y la mediana de ambos indican una cantidad considerable de variabilidad. El intervalo de confianza de 95% requiere que $z_{.025} = 1.96$, y el intervalo es entonces

$$45\ 679.4 \pm (1.96)\left(\frac{26\ 641.675}{\sqrt{50}}\right) = 45\ 679.4 \pm 7384.7$$
$$= (38\ 294.7, 53\ 064.1)$$

Es decir, $38\ 294.7 < \mu < 53\ 064.1$ con un nivel de confianza de aproximadamente 95%. Este intervalo es bastante amplio debido a un tamaño de muestra de 50, que aunque es grande por nuestra regla general, no es lo suficientemente grande como para superar la variabilidad en la muestra. No tenemos una estimación muy precisa de la población media de la lectura del odómetro.

¿El intervalo que hemos calculado es uno de este 95% que en el largo plazo incluye el parámetro calculado o es uno de los "malos" del 5% que no lo hace? Sin saber el valor de μ no podemos contestar. Recuerde que el nivel de confianza se refiere al porcentaje de captura a largo plazo cuando la fórmula se utiliza repetidamente en varias muestras; no se puede interpretar para una sola muestra y el intervalo resultante. ■

Desafortunadamente, la selección del tamaño de muestra para que dé un ancho de intervalo deseado no es simple en este caso, como lo fue en el caso de la σ conocida. Por eso el ancho de (6.8) es $2z_{\alpha/2}s/\sqrt{n}$. Ya que el valor de s no está disponible antes de que los datos hayan sido recopilados, el ancho del intervalo no puede ser determinado tan sólo con

la selección de n. La única opción para un investigador que desea especificar un ancho deseado es hacer una suposición fundamentada de cuál podría ser el valor de s. Siendo conservadores y suponiendo un valor más grande de s, será seleccionado un n más grande de lo necesario. El investigador puede ser capaz de especificar un valor razonablemente preciso del rango de población (la diferencia entre los valores más grande y más pequeño). Entonces si la distribución de la población no es demasiado asimétrica y si se divide el rango entre 4 se obtiene un valor aproximado de lo que s podría ser.

EJEMPLO 6.7 El tiempo de carga (minutos) para el acero de carbono en un tipo de horno de hogar abierto se determinará para cada calor en una muestra de tamaño n. Si el investigador cree que casi todos los tiempos en la distribución están entre 320 y 440, ¿qué tamaño de la muestra sería apropiado para estimar el tiempo promedio real a cuando mucho 5 minutos con un nivel de confianza de 95%?

Un valor razonable para s es $(440 - 320)/4 = 30$. Por tanto

$$n = \left[\frac{(1.96)(30)}{5}\right]^2 = 138.3$$

Dado que el tamaño de la muestra debe ser un número entero, $n = 139$ debe ser utilizado. Tenga en cuenta que la estimación está dentro de 5 minutos con el nivel de confianza especificado que es equivalente a un ancho de intervalo de confianza de 10 minutos. ■

Un intervalo de confianza de muestra grande general

Los intervalos de muestra grande $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ y $x \pm z_{\alpha/2} \cdot s/\sqrt{n}$ son casos especiales de un intervalo de confianza de muestra grande general para un parámetro θ. Suponga que $\hat{\theta}$ es un estimador que satisface las siguientes propiedades: 1) Tiene aproximadamente una distribución normal; 2) es insesgado (al menos aproximadamente); y 3) una expresión para $\sigma_{\hat{\theta}}$, la desviación estándar de $\hat{\theta}$, está disponible. Por ejemplo, en el caso de $\theta = \mu$, $\hat{\mu} = \bar{X}$ es un estimador insesgado cuya distribución es aproximadamente normal cuando n es grande y $\sigma_{\hat{\mu}} = \sigma_{\bar{X}} = \sigma/\sqrt{n}$. Estandarizando $\hat{\theta}$ se obtiene la variable aleatoria $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$, la cual tiene aproximadamente una distribución normal estándar. Esto justifica la proposición de probabilidad

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha \qquad \textbf{(6.9)}$$

Suponga primero que $\sigma_{\hat{\theta}}$ no involucra ningún parámetro desconocido (p. ej., σ conocida en el caso de $\theta = \mu$). Entonces si se reemplaza cada $<$ en (6.9) por $=$ se obtiene $\theta = \hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$, por consiguiente, los límites de confianza inferior y superior son $\hat{\theta} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$ y $\hat{\theta} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$, respectivamente. Suponga ahora que $\sigma_{\hat{\theta}}$ no implica a θ pero sí implica al menos otro parámetro desconocido. Sea $s_{\hat{\theta}}$ la estimación de $\sigma_{\hat{\theta}}$ obtenida utilizando estimaciones en lugar de los parámetros desconocidos (p. ej., $s/\sqrt{n}$ estima $\sigma/\sqrt{n}$). En condiciones generales (esencialmente que $s_{\hat{\theta}}$ se aproxime a $\sigma_{\hat{\theta}}$ con la mayoría de las muestras), un intervalo de confianza válido es $\hat{\theta} \pm z_{\alpha/2} \cdot s_{\hat{\theta}}$. El intervalo muestral grande $\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$ es un ejemplo.

Por último, suponga que $\sigma_{\hat{\theta}}$ implica el θ desconocido. Este es el caso, por ejemplo, cuando $\theta = p$, una proporción de la población. Entonces $(\hat{\theta} - \theta)/\sigma_{\hat{\theta}} = z_{\alpha/2}$ puede ser difícil de resolver. Con frecuencia se puede obtener una solución aproximada reemplazando θ en $\sigma_{\hat{\theta}}$ por su estimación $\hat{\theta}$. Esto da una desviación estándar estimada $s_{\hat{\theta}}$ y el intervalo correspondiente es de nuevo $\hat{\theta} \pm z_{\alpha/2} \cdot s_{\hat{\theta}}$.

Es decir, este intervalo de confianza es una estimación puntual de $\theta \pm$ (valor crítico z) (error estándar estimado del estimador).

Un intervalo de confianza para una proporción de población

Sea p la proporción de "éxitos" en una población, donde *éxito* identifica a un individuo u objeto que tiene una propiedad específica (p. ej., individuos que se graduaron en una universidad, computadoras que no requieren servicio de garantía, etc.). Una muestra aleatoria de n individuos tiene que ser seleccionada y X es el número de éxitos en la muestra. Siempre que n sea pequeño comparado con el tamaño de la población, X puede ser considerada una variable aleatoria binomial con $E(X) = np$ y $\sigma_X = \sqrt{np(1-p)}$. Además, si tanto $np \geq 10$ como $nq \geq 10$, $(q = 1 - p)$, X tiene una distribución normal.

El estimador natural de p es $\hat{p} = X/n$, la fracción muestral de éxitos. Puesto que $\hat{p}$ es simplemente X multiplicada por la constante $1/n$, $\hat{p}$ también tiene aproximadamente una distribución normal. Tal como se muestra en la sección 6.1, $E(\hat{p}) = p$ (insesgado) y $\sigma_{\hat{p}} = \sqrt{p(1-p)}$. La desviación estándar $\sigma_{\hat{p}}$ implica el parámetro desconocido p. Si se estandariza $\hat{p}$ restando p y dividiendo entre $\sigma_{\hat{p}}$ entonces se tiene

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

Procediendo como se sugiere en la subsección "Deducción de un intervalo de confianza" (sección 6.1), los límites de confianza se obtienen al reemplazar cada $<$ por $=$ y resolver la ecuación cuadrática resultante para p. Pero mientras las ecuaciones $(\bar{x} - \mu)/(s/\sqrt{n}) = \pm z_{\alpha/2}$ utilizadas para derivar el intervalo de confianza de una muestra grande para μ son lineares en μ, las ecuaciones aquí son cuadráticas (p^2 aparece en el numerador cuando ambos lados de cada ecuación se elevan al cuadrado para eliminar la raíz cuadrada). Esto da las dos raíces

$$p = \frac{\hat{p} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$$

$$= \tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$$

PROPOSICIÓN

Sea $\tilde{p} = [\hat{p} + z_{\alpha/2}^2/2n]/[1 + z_{\alpha/2}^2/n]$. Entonces, un **intervalo de confianza para una proporción de población *p*** con nivel de confianza aproximadamente de $100(1-\alpha)\%$ es

$$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}\hat{q}/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n} \qquad \textbf{(6.10)}$$

donde $\hat{q} = 1 - \hat{p}$ y, tal como antes, el signo $-$ en la ecuación 6.10 corresponde al límite de confianza inferior y el signo $+$ al límite de confianza superior.

Esto a menudo se denomina como la *puntuación del intervalo de confianza* para p.

Si el tamaño n de la muestra es bastante grande, entonces $z^2/2n$ suele ser insignificante (pequeño) comparado con $\hat{p}$ y z^2/n es insignificante comparado con 1, partiendo de que $\tilde{p} \approx \hat{p}$. En este caso $z^2/4n^2$ también es despreciable comparado con pq/n (n^2 es un divisor mucho más grande que n); como resultado, el término dominante en la expresión $\pm$ es $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$ y el intervalo de puntuación es aproximadamente

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} \qquad \textbf{(6.11)}$$

Este último intervalo tiene la forma general $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$ de un amplio intervalo de la muestra sugerido en la última subsección. La aproximación del intervalo de confianza (6.11) es la que durante décadas ha aparecido en los libros de texto de introducción a la estadística. Está claro que tiene una forma mucho más simple y más atractiva que la puntuación del intervalo de confianza. Así que, ¿por qué molestarse con este último?

Primero, suponga que se utiliza $z_{0.025} = 1.96$ en la fórmula (6.11). Así, nuestro nivel de confianza *nominal* (el que se cree que va a comprar utilizando este valor crítico z) es de aproximadamente 95%. Así que antes de seleccionar una muestra la probabilidad de que el intervalo aleatorio incluya el valor real de p (es decir, la *probabilidad de cobertura*) debe ser de 0.95. Pero, como lo muestra la figura 6.6 para el caso $n = 100$, la probabilidad de cobertura real de este intervalo puede variar considerablemente de la probabilidad nominal de 0.95, en particular cuando p no está cerca de 0.5 (la gráfica de probabilidad de cobertura frente a p es muy irregular debido a que la distribución subyacente de probabilidad binomial es discreta y no continua). Esto es en general una deficiencia del intervalo tradicional; el nivel de confianza real puede ser bastante diferente del nivel nominal, incluso para tamaños de muestra razonablemente grandes. Investigaciones recientes han demostrado que el intervalo de la puntuación rectifica este comportamiento; para prácticamente todos los tamaños de las muestras y los valores de p su nivel de confianza real será bastante cercano al nivel nominal especificado por la elección de $z_{\alpha/2}$. Esto se debe en gran parte al hecho de que el intervalo de la puntuación se desplaza un poco hacia el 0.5 en comparación con los intervalos tradicionales. En particular, el punto medio $\tilde{p}$ del intervalo de la puntuación es siempre un poco más cercano a 0.5 que el punto medio $\hat{p}$ del intervalo tradicional. Esto es especialmente importante cuando p está cerca de 0 o 1.

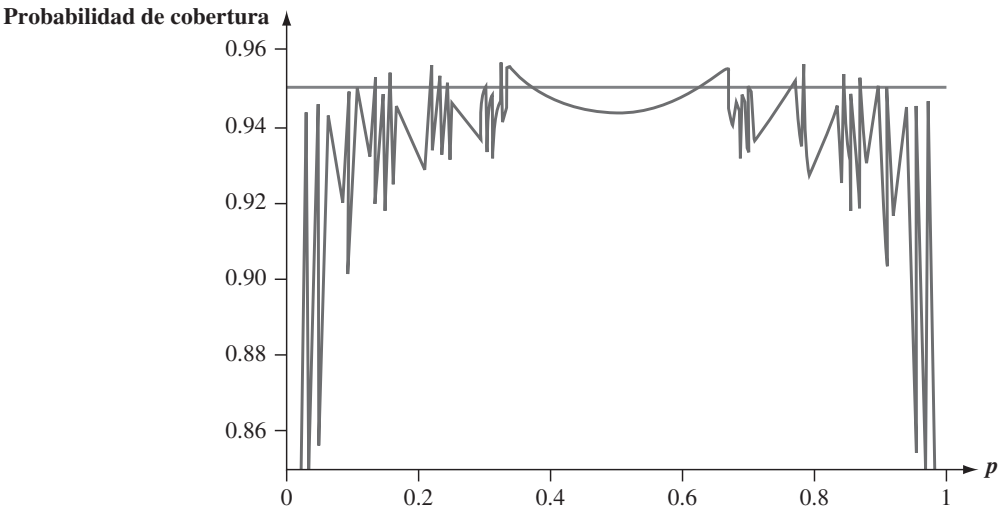


Figura 6.6 Probabilidad de cobertura real para el intervalo (6.11) para variaciones en los valores de p cuando $n = 100$

Además, el intervalo de la puntuación se puede utilizar con casi todos los tamaños de muestra y valores de los parámetros. No es necesario controlar las condiciones $n\hat{p} \geq 10$ y $n(1 - \hat{p}) \geq 10$ que se requerirían al emplear intervalos tradicionales. Así que en lugar de preguntar cuándo n es suficientemente grande para (6.11) para obtener una buena aproximación a (6.10), nuestra recomendación es que *siempre* debe usarse la puntuación del intervalo de confianza. El leve aburrimiento adicional de los cálculos se ve compensado por las propiedades deseables del intervalo.

EJEMPLO 6.8 El artículo **"Repeatability and Reproducibility for Pass/Fail Data" (*J. of Testing and Eval.*, 1997: 151–153)** reporta que en $n = 48$ ensayos en un laboratorio particular, 16 dieron como resultado la ignición de un tipo particular de sustrato por un cigarrillo encendido.

Sea p la proporción a largo plazo de todos los ensayos que producirían ignición. Una estimación puntual de p es $\hat{p} = 16/48 = 0.333$. Un intervalo de confianza para p con un nivel de confianza de aproximadamente 95% es

$$\frac{0.333 + (1.96)^2/96}{1 + (1.96)^2/48} \pm (1.96)\frac{\sqrt{(0.333)(0.667)/48 + (1.96)^2/9216}}{1 + (1.96)^2/48}$$
$$= 0.345 \pm 0.129 = (0.216, 0.474)$$

Este intervalo es bastante amplio ya que un tamaño de muestra de 48 no es tan grande al estimar una proporción.

El intervalo tradicional es

$$0.333 \pm 1.96\sqrt{(0.333)(0.667)/48} = 0.333 \pm 0.133 = (0.200,\ 0.466)$$

Estos dos intervalos concordarían mucho más si el tamaño de muestra fuera sustancialmente más grande. ■

Si se iguala el ancho del intervalo de confianza para p con el ancho preespecificado w se obtiene una ecuación cuadrática para el tamaño de muestra n necesario para dar un intervalo con un grado de precisión deseado. Si se suprime el subíndice en $z_{\alpha/2}$, la solución es

$$n = \frac{2z^2\hat{p}\hat{q} - z^2w^2 \pm \sqrt{4z^4\hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2z^4}}{w^2} \qquad \textbf{(6.12)}$$

Al omitir los términos en el numerador que implican w^2 se obtiene

$$n \approx \frac{4z^2\hat{p}\hat{q}}{w^2}$$

Esta última expresión es lo que resulta de igualar el ancho del intervalo tradicional con w.

Estas fórmulas desafortunadamente implican a la $\hat{p}$ desconocida. El método más conservador es aprovechar el hecho de que $\hat{p}\hat{q}[= \hat{p}(1 - \hat{p}]$ es un máximo cuando $\hat{p} = 5$. Por consiguiente, si se utiliza $\hat{p} = \hat{q} = 0.5$ en (7.12), el ancho será cuando mucho w haciendo caso omiso de qué valor de $\hat{p}$ resulte de la muestra. De manera alternativa, si el investigador cree firmemente, basado en información previa, que $p \leq p_0 \leq 0.5$, entonces se utiliza p_0 en lugar de $\hat{p}$. Una observación similar es válida cuando $p \geq p_0 \geq 0.5$.

EJEMPLO 6.9 El ancho del intervalo de confianza de 95% en el ejemplo 6.8 es 0.258. El valor de n necesario para garantizar un ancho de 0.10 independientemente del valor de $\hat{p}$ es

$$n = \frac{2(1.96)^2(0.25) - (1.96)^2(0.01) \pm \sqrt{4(1.96)^4(0.25)(0.25 - 0.01) + (0.01)(1.96)^4}}{0.01} = 380.3$$

Por consiguiente, se deberá utilizar un tamaño de muestra de 381. La expresión para n basada en el intervalo de confianza tradicional da un valor un poco más grande que 385. ■

Intervalos de confianza unilaterales (límites de confianza)

Los intervalos de confianza que se han abordado hasta ahora dan un límite de confianza inferior *así como* uno superior para el parámetro que se está estimando. En algunas circunstancias es posible que un investigador desee sólo uno de estos dos tipos de límites. Por ejemplo, es posible que un psicólogo desee calcular un límite de confianza superior de 95% para el tiempo de reacción promedio verdadero a un estímulo particular, o es posible que un ingeniero de seguridad desee sólo un límite de confianza inferior para la vida útil promedio

real de los componentes de un tipo. Debido a que el área acumulada bajo la curva normal estándar a la izquierda de 1.645 es de 0.95,

$$P\left(\frac{\overline{X}-\mu}{S/\sqrt{n}} < 1.645\right) \approx 0.95$$

Si se manipula la desigualdad entre paréntesis para aislar μ en un lado y se reemplazan las variables aleatorias con valores calculados, se obtiene la desigualdad $\mu > \bar{x} - 1.645s/\sqrt{n}$; la expresión a la derecha es el límite de confianza inferior deseado. Comenzando con $P(-1.645 < Z) \approx 0.95$ y manipulando la desigualdad se obtiene el límite de confianza superior. Un argumento similar proporciona un límite unilateral asociado con cualquier otro nivel de confianza.

PROPOSICIÓN

Un **límite de confianza superior muestral grande para μ** es

$$\mu < \bar{x} + z_{\alpha} \cdot \frac{s}{\sqrt{n}}$$

y un **límite de confianza inferior muestral grande para μ** es

$$\mu > \bar{x} - z_{\alpha} \cdot \frac{s}{\sqrt{n}}$$

Se obtiene un **límite de confianza unilateral para p** reemplazando $z_{\alpha/2}$ en lugar de z_{α} y $\pm$ en lugar de $+$ o $-$ en la fórmula para el intervalo de confianza (6.10) para p. En todos los casos, el nivel de confianza es aproximadamente de $100(1-\alpha)\%$.

EJEMPLO 6.10 El titanio y sus aleaciones se utilizan cada vez con mayor frecuencia para aplicaciones automotrices y aeroespaciales, debido a su durabilidad y a su elevada relación fuerza/peso. Sin embargo, el mecanizado puede ser difícil debido a la baja conductividad térmica. El artículo **"Modeling and Multi-Objective Optimization of Process Parameters of Wire Electrical Discharge Machining Using Non-Dominated Sorting Genetic Algorithm-II (*J. of Engr. Manuf.*, 2012: 1186–2001)** describe una investigación sobre los diferentes parámetros que afectan el mecanizado por descarga eléctrica del titanio 6-2-4-2. Una característica de interés era la rugosidad superficial (μg) del metal después del mecanizado. Una muestra de 54 observaciones de la rugosidad superficial dio como resultado una media de rugosidad de 1.9042 y una desviación estándar de la muestra de 0.1455. El límite de confianza superior para la rugosidad promedio verdadera μ con un nivel de confianza de 95% requiere $z_{0.05} = 1.645$ (no el valor $z_{.0.25} = 1.96$ necesario para un intervalo de confianza de dos extremos). El límite es

$$1.9042 + (1.645) \cdot \frac{(0.1455)}{\sqrt{54}} = 1.9042 + 0.0326 = 1.9368$$

Esto es, con un nivel de confianza de 95%, se puede decir que $\mu < 1.9368$. ■

EJERCICIOS Sección 6.2 (12-27)

12. Las siguientes observaciones representan el tiempo de vida (días) para individuos que han sido diagnosticados con cáncer de sangre (**"A Goodness of Fit Approach to the Class of Life Distributions with Unknown Age", *Quality and Reliability Engr. Intl.*, 2012: 761–766):**

115 181 255 418 441 461 516 739 743 789 807
865 924 983 1025 1062 1063 1165 1191 1222 1222 1251
1277 1290 1357 1369 1408 1455 1478 1519 1578 1578 1599
1603 1605 1696 1735 1799 1815 1852 1899 1925 1965

a. ¿Se puede calcular un intervalo de confianza para el tiempo de vida real sin asumir nada sobre la naturaleza de la

distribución del tiempo de vida? Explique su razonamiento. [*Nota*: Una gráfica de probabilidad normal de los datos exhibe un patrón lineal razonable.]

b. Calcule e interprete un intervalo de confianza con un nivel de confianza de 99% para el tiempo de vida verdadero. [*Sugerencia*: $\bar{x} = 1191.6$ y $s = 506.6$.]

13. El artículo **"Gas Cooking, Kitchen Ventilation, and Exposure to Combustion Products" (*Indoor Air*, 2006: 65–73)** reporta que para una muestra de 50 cocinas con estufas de gas, monitoreadas durante una semana, el nivel de CO_2 medio muestral (ppm) fue de 654.16 y la desviación estándar muestral fue de 164.43.

a. Calcule e interprete un intervalo de confianza de 95% (bilateral) para un nivel de CO_2 promedio verdadero en la población de todas las casas de la cual se seleccionó la muestra.

b. Suponga que el investigador había hecho una suposición preliminar de 175 para el valor de s antes de recopilar los datos. ¿Qué tamaño de muestra sería necesario para obtener un ancho de intervalo de 50 ppm para un nivel de confianza de 95%?

14. Se han estudiado bastante bien los efectos negativos sobre los pulmones de los niños de la contaminación atmosférica, pero se ha estudiado menos el impacto de la contaminación del aire en interiores. Los autores de **"Indoor Air Pollution and Lung Function Growth Among Children in Four Chinese Cities" (*Indoor Air*, 2012: 3–11)** investigaron la relación entre la medición de la contaminación del aire en interiores y el crecimiento de la función pulmonar en niños de entre 6 y 13 años que viven en cuatro ciudades chinas. Los autores midieron para cada sujeto del estudio un importante índice de capacidad pulmonar conocido como FEV_1, el volumen forzado de aire (en ml) que es exhalado en un segundo. Los valores de FEV_1 más altos están asociados a una capacidad pulmonar mayor. De los niños en el estudio 514 pertenecen a hogares que utilizan carbón para cocinar o calentarse, o ambos. Su FEV_1 medio fue de 1427 con una desviación estándar de 325. (Para demostrar que quemar carbón tenía un claro efecto negativo en los niveles medios de FEV_1 se utilizó un complejo procedimiento estadístico.)

a. Calcule e interprete un intervalo de confianza (bilateral) a 95% para el valor promedio verdadero del nivel de FEV_1 en la población de todos los niños a partir de la cuales se seleccionó la muestra. ¿Parece que el parámetro de interés fue estimado correctamente?

b. Suponga que los investigadores tienen un estimado grueso de que 320 es el valor de s antes de recolectar los datos. ¿Cuál es el tamaño de muestra que se necesitaría para obtener un intervalo de confianza al 95% de 50 ml?

15. Determine el nivel de confianza de cada uno de los siguientes límites de confianza unilaterales de muestras grandes:

a. Límite superior: $\bar{x} + 0.84s/\sqrt{n}$

b. Límite inferior: $\bar{x} - 2.05s/\sqrt{n}$

c. Límite superior: $\bar{x} + 0.67s/\sqrt{n}$

16. El voltaje de ruptura de corriente alterna (AC) de un líquido aislante indica su rigidez dieléctrica. El artículo **"Testing Practices for the AC Breakdown Voltage Testing of Insulation Liquids" (*IEEE Electrical Insulation Magazine*, 1995: 21-26)** dio las siguientes observaciones muestrales de voltaje de ruptura (kV) de un circuito particular, en ciertas condiciones.

```
62 50 53 57 41 53 55 61 59 64 50 53 64 62 50 68
54 55 57 50 55 50 56 55 46 55 53 54 52 47 47 55
57 48 63 57 57 55 53 59 53 52 50 55 60 50 56 58
```

a. Construya un diagrama de caja de los datos y comente sobre las características interesantes.

b. Calcule e interprete un intervalo de confianza de 95% para el promedio real del voltaje de ruptura μ. ¿Parece que μ ha sido estimada con precisión? Explique.

c. Supongamos que el investigador piensa que prácticamente todos los valores de voltaje de ruptura están entre 40 y 70. ¿Qué tamaño de la muestra sería conveniente para que el intervalo de confianza de 95% tuviera una anchura de 2 kV (de modo que μ se estime dentro de 1 kV con 95% de confianza)?

17. El ejercicio 1.13 dio una muestra de observaciones de resistencia última a la tensión ($kg/pulg^2$). Use los datos de salida estadísticos descriptivos adjuntos de Minitab para calcular un límite de confianza inferior de 99% para la resistencia a la tensión última promedio verdadera e interprete el resultado.

```
   N    Mean  Median  TrMean  StDev  SE Mean
 153  135.39  135.40  135.41   4.59     0.37

Minimum  Maximum      Q1      Q3
 122.20   147.70  132.95  138.25
```

18. La armada de los Estados Unidos comisionó un estudio para evaluar qué tan profundamente penetra una bala en una armadura de cerámica **("Testing Body Armor Materials for Use by the U.S. Army–Phase III", 2012)**. En la prueba estándar, debajo del chaleco de la armadura se coloca un modelo cilíndrico de arcilla en capas. Después se dispara un proyectil, y provoca una muesca en la arcilla. La impresión en la arcilla más profunda se mide como un indicador de sobrevivencia de cualquiera que use la armadura. Aquí se muestran los datos de una de las pruebas bajo ciertas condiciones experimentales; las mediciones (en mm) se hicieron utilizando un calibrador digital controlado manualmente:

```
22.4   23.6   24.0   24.9   25.5   25.6
25.8   26.1   26.4   26.7   27.4   27.6
28.3   29.0   29.1   29.6   29.7   29.8
29.9   30.0   30.4   30.5   30.7   30.7
31.0   31.0   31.4   31.6   31.7   31.9
31.9   32.0   32.1   32.4   32.5   32.5
32.6   32.9   33.1   33.3   33.5   33.5
33.5   33.5   33.6   33.6   33.8   33.9
34.1   34.2   34.6   34.6   35.0   35.2
35.2   35.4   35.4   35.4   35.5   35.7
35.8   36.0   36.0   36.0   36.1   36.1
36.2   36.4   36.6   37.0   37.4   37.5
37.5   38.0   38.7   38.8   39.8   41.0
42.0   42.1   44.6   48.3   55.0
```

a. Construya una gráfica de caja para los datos y comente sobre algunas características interesantes de la misma.

b. Construya una gráfica de probabilidad normal. ¿Es posible que la profundidad de impresión tenga una distribución normal? ¿Será necesario asumir una distribución normal con el fin de calcular un intervalo de confianza o uno obligado

por la profundidad promedio verdadera μ utilizando los datos anteriores? Explique.

c. Utilice los datos de Minitab que acompañan este texto como base para calcular e interpretar un límite de confianza superior para μ con un nivel de confianza de 99%.

```
Variable   Count    Mean    SE Mean    StDev
Depth         83  33.370      0.578    5.268

Q1       Median       Q3     IQR
30.400   33.500   36.000   5.600
```

19. El artículo **"Limited Yield Estimation for Visual Defect Sources" (*IEEE Trans. on Semiconductor Manuf.*, 1997: 17–23)** reportó que, en un estudio de un proceso particular de inspección de láminas, 356 troqueles fueron examinados por una sonda de inspección y 201 de éstos pasaron la prueba. Suponiendo un proceso estable, calcule un intervalo de confianza (bilateral) de 95% para la proporción de todos los troqueles que pasan la prueba.

20. Las agencias de publicidad para televisión se enfrentan a retos cada vez mayores para alcanzar a la audiencia, debido a que ver programas de televisión mediante descarga digital cada vez es más popular. La **encuesta Harris** reportó el de 13 de noviembre de 2012 que 53% de los 2343 adultos estadounidenses encuestados declararon haber visto programas de televisión descargados de manera digital en algún tipo de dispositivo.

a. Calcule e interprete un intervalo de confianza al 99% para la proporción de todos los adultos estadounidenses que vieron un programa descargado de manera digital hasta ese momento.

b. ¿Cuál es el tamaño de la muestra que se necesitaría para que un intervalo de confianza al 99% tuviera una amplitud de 0.5 sin tomar en cuenta el valor de $\hat{p}$?

21. En una muestra de 1000 consumidores seleccionados al azar que tuvieron la oportunidad de enviar un formulario de solicitud de reembolso después de comprar un producto, 250 negaron haberlo hecho **("Rebates: Get What You Deserve", *Consumer Reports*, mayo de 2009: 7)**. Las razones que citaron para ese tipo de comportamiento incluyen demasiados pasos en el proceso, cantidad muy pequeña, vencimiento del plazo, el temor de ser incluido en una lista de correo, pérdida de la nota de compra y las dudas en torno a recibir el dinero. Calcule un límite de confianza superior al nivel de confianza del 95% para la verdadera proporción de estos consumidores que nunca solicitaron un reembolso. Con base en este límite, ¿hay pruebas convincentes de que la verdadera proporción de estos consumidores es menor que 1/3? Explique su razonamiento.

22. La tecnología subyacente de reemplazo de cadera ha cambiado, ya que estas cirugías se han vuelto más populares (más de 250 000 en los Estados Unidos en el año 2008). A partir de 2003 se comercializan las caderas de cerámica de alta duración. Desafortunadamente, para muchos pacientes la mayor durabilidad ha sido compensada por un aumento en la incidencia de rechinidos. La edición del 11 de mayo de 2008 del *New York Times* informó que en un estudio de 143 individuos que recibieron las caderas de cerámica, entre 2003 y 2005, 10 reportaron que las caderas rechinaban.

a. Calcule un límite de confianza inferior en el nivel de confianza del 95% para la verdadera proporción de las caderas que rechinaban.

b. Interprete el nivel de confianza del 95% utilizado en el inciso a).

23. El **Pew Forum on Religion and Public Life** reportó el 9 de diciembre de 2009 que en una encuesta de 2003 adultos estadounidenses, 25% dijo que creía en la astrología.

a. Calcule e interprete un intervalo de confianza al nivel de confianza del 99% para la proporción de todos los adultos estadounidenses que creen en la astrología.

b. ¿Qué tamaño de muestra se requiere para que el ancho de un intervalo de confianza de 99% tenga un máximo de 0.05, independientemente del valor de $\hat{p}$?

24. Una muestra de 56 muestras de algodón produjo un porcentaje de alargamiento promedio muestral de 8.17 y una desviación estándar de 1.42 **("An Apparent Relation Between the Spiral Angle ϕ, the Percent Elongation E_1, and the Dimensions of the Cotton Fiber", *Textile Research J.*, 1978: 407–410)**. Calcule un intervalo de confianza de 95% de una muestra grande para el porcentaje de alargamiento promedio verdadero μ. ¿Qué suposiciones está haciendo sobre la distribución del porcentaje de alargamiento?

25. Una legisladora estatal desea encuestar a los residentes de su distrito para ver qué proporción del electorado está consciente de su posición sobre la utilización de fondos estatales para solventar abortos.

a. ¿Qué tamaño de la muestra es necesario si el intervalo de confianza de 95% para p debe tener un ancho de a lo más 0.10 independientemente de p?

b. Si la legisladora está firmemente convencida de que al menos 2/3 del electorado conoce su posición, ¿qué tamaño de muestra recomendaría?

26. El superintendente de un gran distrito escolar, que alguna vez tomó un curso de probabilidad y estadística, cree que el número de maestros ausentes en cualquier día dado tiene una distribución de Poisson con parámetro μ. Use los siguientes datos sobre ausencias durante 50 días para obtener un intervalo de confianza muestral grande para μ. [*Sugerencia*: La media y la varianza de una variable de Poisson son iguales a μ, por consiguiente,

$$Z = \frac{\overline{X} - \mu}{\sqrt{\mu/n}}$$

tiene aproximadamente una distribución normal estándar. Ahora prosiga, tal y como en la deducción del intervalo para p, haciendo una proposición de probabilidad (con probabilidad de $1 - \alpha$) y resuelva las desigualdades resultantes para μ (véase el argumento exactamente después de (6.10).]

Número de ausencias	0	1	2	3	4	5	6	7	8	9	10
Frecuencia	1	4	8	10	8	7	5	3	2	1	1

27. Reconsidere el intervalo de confianza (6.10) para p y enfóquese en un nivel de confianza de 95%. Demuestre que los límites de confianza concuerdan bastante bien con los del intervalo tradicional (6.11) una vez que dos éxitos y dos fallas se han anexado a la muestra [es decir, (6.11) basado en $x + 2$ S en $n = 4$ ensayos]. [*Sugerencia*: $1.96 \approx 2$. *Nota*: Agresti y Coull demostraron que este ajuste del intervalo tradicional también tiene un nivel de confianza próximo al nivel nominal.]

6.3 Intervalos basados en una distribución de población normal

El intervalo de confianza para μ en la sección 6.2 es válido siempre que n es grande. El intervalo resultante puede ser utilizado cualquiera que sea la naturaleza de la distribución de la población. Sin embargo, el teorema del límite central no puede ser invocado cuando n es pequeña. En este caso, una forma de proceder es hacer una suposición específica sobre la forma de la distribución de la población y luego deducir un intervalo de confianza adecuado a esa suposición. Por ejemplo, se podría desarrollar un intervalo de confianza para μ cuando una distribución gamma describe la población, otro para el caso de una población de Weibull, etcétera. Los estadísticos en realidad han realizado este programa para varias familias distribucionales diferentes. Puesto que frecuentemente la distribución normal es más apropiada como modelo de población que cualquier otro tipo de distribución, aquí nos concentraremos en un intervalo de confianza para esta situación.

SUPOSICIÓN

La población de interés es normal, de modo que $X_1, \ldots, X_n$ constituyen una muestra aleatoria tomada de una distribución normal con μ y σ desconocidas.

El resultado clave que subyace en el intervalo de la sección 6.2 fue que con un n grande, la variable aleatoria $Z = (\overline{X} - \mu)/(S/\sqrt{n})$ tiene aproximadamente una distribución normal estándar. Cuando n es pequeño, no es probable que S se aproxime a σ, de modo que la variabilidad de la distribución de Z surge de la aleatoriedad tanto en el numerador como en el denominador. Esto implica que la distribución de probabilidad de $(\overline{X} - \mu)/(S/\sqrt{n})$ se dispersará más que la distribución normal estándar. El resultado en el cual están basadas las inferencias introduce una nueva familia de distribuciones de probabilidad llamada *distribuciones t*.

TEOREMA

Cuando $\overline{X}$ es la media de una muestra aleatoria de tamaño n tomada de una distribución normal con media μ, la variable aleatoria

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{6.13}$$

tiene una distribución de probabilidad llamada distribución t con $n - 1$ grados de libertad (gl).

Propiedades de distribuciones *t*

Antes de aplicar este teorema se impone un análisis de las propiedades de distribuciones t. Aunque la variable de interés sigue siendo $(\overline{X} - \mu)/(S/\sqrt{n})$, ahora se denota por T para recalcar que no tiene una distribución normal estándar cuando n es pequeña. Recuerde que una distribución normal está regida por dos parámetros, cada elección diferente de μ en combinación con σ resulta en una distribución normal particular. Cualquier distribución t particular resulta de especificar el valor de sólo un parámetro, llamado **número de grados de libertad**, abreviado como gl. Este parámetro se denota con la letra griega ν. Los posibles valores de ν son los enteros positivos 1, 2, 3,… . Por lo que hay una distribución t con 1 gl, otra con 2 gl, aún otra con 3 gl y así sucesivamente.

Para cualquier valor fijo del parámetro ν, la función de densidad que especifica la curva t asociada es incluso más complicada que la función de densidad normal. Afortunadamente, sólo hay que ocuparse de algunas de las más importantes características de estas curvas.

Propiedades de distribuciones *t*

Sea que t_ν denote la distribución t con ν gl.

1. Cada curva t_ν tiene forma de campana y centrada en 0.
2. Cada curva t_ν está más extendida que la curva (z) normal estándar.
3. Conforme ν se incrementa, la extensión de la curva t_ν correspondiente disminuye.
4. A medida que $\nu \to \infty$, la secuencia de curvas t_ν tiende a la curva normal estándar (así que la curva z a menudo se llama curva t con grados de libertad $= \infty$).

La figura 6.7 ilustra varias de estas propiedades para valores seleccionados de ν.

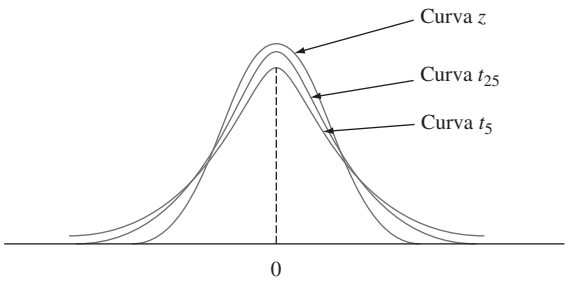


Figura 6.7 Curvas t_ν y z

El número de grados de libertad para T en (6.13) es $n - 1$ porque, aunque S está basada en las n desviaciones $X_1 - \overline{X}, \ldots, X_n - \overline{X}$, $\Sigma(X_1 - \overline{X}) = 0$ implica que sólo $n - 1$ de estas están "libremente determinadas". El número de grados de libertad para una variable t es el número de desviaciones libremente determinadas en las cuales está basada la desviación estándar estimada en el denominador de T.

El uso de la distribución t al hacer inferencias requiere notación para capturar áreas de cola de la curva t análogas a z_α de la curva z. Se podría pensar que t_α haría el truco. Sin embargo, el valor deseado depende no sólo del área de la cola capturada, sino también de gl.

NOTACIÓN

Sea $t_{\alpha,\nu}$ = el número sobre el eje de medición con el cual el área bajo la curva t con n grados de libertad a la derecha de $t_{\alpha,\nu}$ es α; $t_{\alpha,\nu}$ se llama **valor crítico** t.

Por ejemplo, $t_{0.05,6}$ es el valor crítico t que captura un área de cola superior de 0.05 bajo la curva t con 6 gl. La notación general se ilustra en la figura 6.8. Debido a que las curvas t son simétricas alrededor de cero, $-t_{\alpha,\nu}$ captura el área α de la cola inferior. La tabla A.5 del apéndice proporciona $t_{\alpha,\nu}$ para valores seleccionados de α y ν. Esta tabla también aparece al final del libro. Las columnas de la tabla corresponden a diferentes valores de α. Para obtener $t_{0.05,15}$ vaya a la columna $\alpha = 0.05$, mire hacia abajo al renglón $\nu = 15$, y lea $t_{0.05,15} = 1.753$. Del mismo modo, $t_{0.05,22} = 1.717$ (columna 0.05, renglón $\nu = 22$) y $t_{0.01,22} = 2.508$.

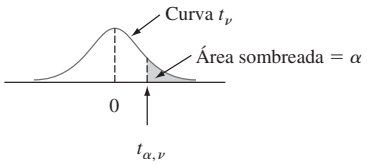


Figura 6.8 Ilustración de un valor crítico t

Los valores de $t_{\alpha,\nu}$ exhiben un comportamiento regular al recorrer una fila o al descender por una columna. Con ν fijo, $t_{\alpha,\nu}$ se incrementa a medida que α disminuye, puesto que hay que moverse más a la derecha de cero para capturar el área α en la cola. Con α fija, a medida que ν se incrementa (es decir, cuando cualquier columna particular de la tabla t se recorre hacia abajo) el valor de $t_{\alpha,\nu}$ disminuye. Esto es porque un valor más grande de ν implica una distribución t con extensión más pequeña, de modo que no es necesario ir más lejos de cero para capturar el área de cola α. Además, $t_{\alpha,\nu}$ disminuye más lentamente a medida que ν se incrementa. Por consiguiente, los valores que aparecen en la tabla se muestran en incrementos de 2 entre 30 y 40 grados de libertad y luego saltan a $\nu = 50$, 60, 120 y, por último, ∞. Puesto que t_∞ es la curva normal estándar, los valores z_α conocidos aparecen en la última fila de la tabla. La regla empírica sugería con anterioridad que el uso del intervalo de confianza muestral grande (si $n > 40$) proviene de la igualdad aproximada de las distribuciones normales estándar y t para $\nu \geq 40$.

Intervalo de confianza t para una muestra

La variable estandarizada T tiene una distribución t con $n - 1$ grados de libertad y el área bajo la curva de densidad t correspondiente entre $-t_{\alpha/2,n-1}$ y $t_{\alpha/2,n-1}$ es $1 - \alpha$ (el área $\alpha/2$ queda en cada cola), por consiguiente,

$$P\left(-t_{\alpha/2,\,n-1} < T < t_{\alpha/2,\,n-1}\right) = 1 - \alpha \tag{6.14}$$

La expresión (6.14) difiere de las expresiones que aparecen en secciones previas en las que T y $t_{\alpha/2,n-1}$ se utilizan en lugar de Z y $z_{\alpha/2}$, aunque pueden ser manipuladas de la misma manera para obtener un intervalo de confianza para μ.

PROPOSICIÓN

Sean $\bar{x}$ y s la media y la desviación estándar muestrales calculadas con los resultados de una muestra aleatoria tomada de una población normal con media μ. Entonces un **intervalo de confianza de 100(1 − α)% para μ** es

$$\left(\bar{x} - t_{\alpha/2,\,n-1} \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,\,n-1} \cdot \frac{s}{\sqrt{n}}\right) \tag{6.15}$$

o, más compactamente, $\bar{x} \pm t_{\alpha/2,\,n-1} \cdot s/\sqrt{n}$.

Un **límite de confianza superior para μ** es

$$\bar{x} + t_{\alpha,\,n-1} \cdot \frac{s}{\sqrt{n}}$$

y reemplazando + por − en la última expresión se obtiene un **límite de confianza inferior para μ**, ambos con nivel de confianza de 100(1 − α)%.

EJEMPLO 6.11 A pesar de que los mercados tradicionales de madera de ocozol han disminuido, las maderas sólidas de gran sección que usualmente se utiliza para la construcción de puentes y duelas se han vuelto cada vez más escasas. El artículo **"Development of Novel Industrial Laminated Planks from Sweetgum Lumber"** (***J. of Bridge Engr.*, 2008: 64–66**) describe la fabricación y el ensayo de las vigas compuestas, concebida para agregar valor a la

madera de ocozol de bajo grado. Estos son los datos sobre el módulo de ruptura (lb/pulg^2; el artículo contenía un resumen de los datos, expresados en MPa):

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7422.69	7886.87	6316.67	7713.65	7503.33	7674.99

La figura 6.9 muestra un diagrama de probabilidad normal obtenido con el software R. Lo recto del patrón en el diagrama apoya fuertemente la suposición de que la distribución de la población del módulo de ruptura es al menos aproximadamente normal.

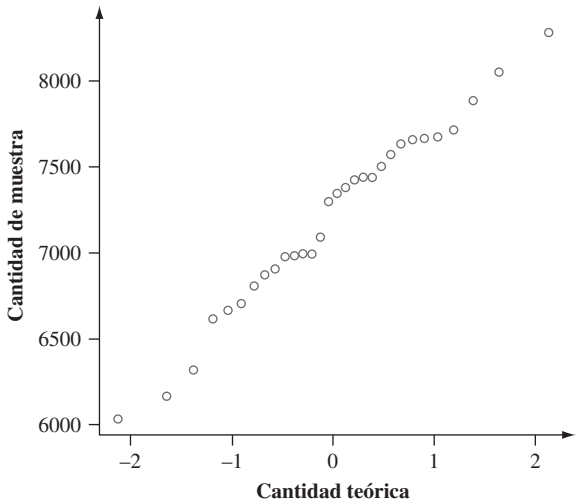


Figura 6.9 Diagrama de probabilidad normal de los datos del módulo de ruptura

La media muestral y la desviación estándar de la muestra son 7203.191 y 543.5400, respectivamente (para abatir cualquier realización de cálculos a mano, la carga de cálculos se alivia un poco al restar 6000 a cada valor de x para obtener $y_i = x_i - 6000$, entonces $\Sigma y_i = 36\ 095.72$ y $\Sigma y_i^2 = 51\ 997{,}668.77$, de la cual $\bar{y} = 1203.191$ y $s_y = s_x$ tal como se indica).

Ahora se calcula un intervalo de confianza para el promedio real del módulo de ruptura con un nivel de confianza de 95%. El intervalo de confianza se basa en $n - 1 = 29$ grados de libertad, por lo que el valor de t crítico necesario es $t_{0.025,29} = 2.045$. La estimación por intervalo es ahora

$$\bar{x} \pm t_{0.025,29} \cdot \frac{s}{\sqrt{n}} = 7203.191 \pm (2.045) \cdot \frac{543.5400}{\sqrt{30}}$$
$$= 7203.191 \pm 202.938 = (7000.253, 7406.129)$$

Se estima que $7000.253 < \mu < 7406.129$ con un 95% de confianza. Si se utiliza la misma fórmula muestra tras muestra, en el largo plazo 95% de los intervalos calculados contendrá a μ. Dado que el valor de μ no está disponible, no sabemos si el intervalo calculado es uno de los "buenos" del 95% o el "malo" del 5%. Incluso con el tamaño de la muestra moderadamente grande, el intervalo es bastante amplio. Esto es una consecuencia de la cantidad sustancial de variabilidad de la muestra en los valores del módulo de ruptura.

De conservar únicamente el límite de confianza inferior (el que tiene el signo $(-)$) y al reemplazar 2.045 con $t_{0.05,29} = 1.699$ resultaría un límite de confianza inferior a 95%. ■

Por desgracia no es fácil seleccionar n para controlar el ancho del intervalo t. Esto es porque el ancho implica la s desconocida (antes de recopilar los datos) y porque n se ingresa no sólo a través de $1/\sqrt{n}$, sino también a través de $t_{\alpha/2,n-1}$. Por consiguiente, se puede obtener una n apropiada sólo mediante ensayo y error.

Mas adelante, se analizará un intervalo de confianza de muestra pequeña para μ, que será válido siempre que la distribución de la población sea simétrica, una suposición más débil que la de normalidad. No obstante, cuando la distribución de la población es normal el intervalo t tiende a acortarse más de lo que lo haría *cualquier otro* intervalo con el mismo nivel de confianza.

Un intervalo de predicción para un solo valor futuro

En muchas aplicaciones el objetivo es *predecir* un solo valor de una variable que tiene que ser observada en un tiempo futuro, en lugar de *estimar* la media de dicha variable.

EJEMPLO 6.12 Considere la siguiente muestra de contenido de grasa (en porcentaje) de $n = 10$ *hot dogs* seleccionados al azar (**"Sensory and Mechanical Assessment of the Quality of Frankfurters", *J. of Texture Studies*, 1990: 395–409**):

25.2 21.3 22.8 17.0 29.8 21.0 25.5 16.0 20.9 19.5

Suponiendo que estas observaciones se seleccionaron de una distribución de población normal, un intervalo de confianza de 95% para (estimación del intervalo de) el contenido de grasa medio de la población es

$$\bar{x} \pm t_{0.025,9} \cdot \frac{s}{\sqrt{n}} = 21.90 \pm 2.262 \cdot \frac{4.134}{\sqrt{10}} = 21.90 \pm 2.96$$
$$= (18.94, 24.86)$$

Suponga, sin embargo, que se va a comer un solo *hot dog* de este tipo y desea *predecir* el contenido de grasa resultante. Una predicción *puntual*, análoga a una estimación *puntual*, es simplemente $\bar{x} = 21.90$. Esta predicción desafortunadamente no da información sobre confiabilidad o precisión. ■

El escenario general es como sigue: se dispone de una muestra aleatoria $X_1, X_2, \ldots, X_n$ tomada de una distribución de población normal y se desea predecir el valor de X_{n+1}, una sola observación futura (por ejemplo, la vida útil de un foco sencillo que se compra o la eficiencia del combustible de un automóvil rentado). Un predictor puntual es $\overline{X}$ y el error de predicción resultante es $\overline{X} - X_{n+1}$. El valor esperado del error de predicción es

$$E(\overline{X} - X_{n+1}) = E(\overline{X}) - E(X_{n+1}) = \mu - \mu = 0$$

Siendo que X_{n+1} es independiente de $X_1, \ldots, X_n$, también es independiente de $\overline{X}$, así que la varianza del error de predicción es

$$V(\overline{X} - X_{n+1}) = V(\overline{X}) + V(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

El error de predicción se distribuye normalmente, ya que es una combinación lineal de variables aleatorias independientes con distribución normal. Por consiguiente,

$$Z = \frac{(\overline{X} - X_{n+1}) - 0}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} = \frac{\overline{X} - X_{n+1}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}}$$

tiene una distribución normal estándar. Se puede demostrar que si se reemplaza σ con la desviación estándar muestral S (de $X_1, \ldots, X_n$) se obtiene

$$T = \frac{\overline{X} - X_{n+1}}{S\sqrt{1 + \frac{1}{n}}} \sim \text{distribución } t \text{ con } n - 1 \text{ grados de libertad}$$

Si se manipula esta variable T como se manipuló $T = (\overline{X} - \mu)/(S/\sqrt{n})$ en el desarrollo de un intervalo de confianza se obtiene el siguiente resultado.

PROPOSICIÓN

Un **intervalo de predicción** (**IP**) para una sola observación que tiene que ser seleccionado de una distribución de población normal es

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}} \qquad \textbf{(6.16)}$$

El *nivel de predicción* es $100(1 - \alpha)\%$. Una predicción del límite inferior resulta de la sustitución de $t_{\alpha/2}$ por t_α y desechar la parte + de (7.16), una modificación similar da una predicción del límite superior.

La interpretación de un nivel de predicción de 95% es similar a la de un nivel de confianza de 95%. Si se calcula el intervalo (6.16) muestra tras muestra, y si luego de cada cálculo se observa X_{n+1}, a la larga 95% de estos intervalos incluirá los valores futuros correspondientes.

EJEMPLO 6.13 (Continuación del ejemplo 6.12)

Con $n = 10, \bar{x} = 21.90, s = 4.134$ y $t_{0.025,9} = 2.262$, un intervalo de predicción de 95% para el contenido de grasa de un solo *hot dog* es

$$21.90 \pm (2.262)(4.134)\sqrt{1 + \frac{1}{10}} = 21.90 \pm 9.81$$
$$= (12.09, 31.71)$$

El intervalo es bastante ancho, lo que indica una incertidumbre sustancial en cuanto al contenido de grasa. Observe que el ancho del intervalo de predicción es más de tres veces el del intervalo de confianza. ■

El error de predicción es $\overline{X} - X_{n+1}$, la diferencia entre dos variables aleatorias, en tanto que el error de estimación es $\overline{X} - \mu$, la diferencia entre una variable aleatoria y un valor fijo (aunque desconocido). El intervalo de predicción es más ancho que el intervalo de confianza porque hay más variabilidad en el error de predicción (debido a X_{n+1}) que en el error de estimación. De hecho, a medida que n se hace arbitrariamente grande, el intervalo de confianza se contrae a un solo valor μ y el intervalo de predicción tiende a $\mu \pm z_{\alpha/2} \cdot \sigma$. Existe incertidumbre respecto a un solo valor X incluso cuando no hay necesidad de estimarlo.

Intervalos de tolerancia

Considere una población de automóviles de cierto tipo y suponga que en condiciones específicas, la eficiencia de combustible (mpg) tiene una distribución normal con $\mu = 30$ y $\sigma = 2$. Entonces como el intervalo de -1.645 a 1.645 captura 90% del área bajo la curva z, 90% de todos estos automóviles tendrán valores de eficiencia de combustible entre $\mu - 1.645s = 26.71$ y $\mu + 1.645\sigma = 33.29$. Pero, ¿qué sucederá si los valores de μ y σ no son conocidos? Se puede tomar una muestra de tamaño n, determinar las eficiencias de combustible,

$\bar{x}$ y s, y formar el intervalo cuyo límite inferior es $\bar{x} - 1.645s$ y cuyo límite superior es $\bar{x} + 1.645s$. Sin embargo, debido a la variabilidad de muestreo en las estimaciones de μ y σ existe una buena probabilidad de que el intervalo resultante incluya menos de 90% de los valores de la población. Intuitivamente, para tener *a priori* una probabilidad de que 95% del intervalo resultante incluya al menos 90% de los valores de la población, cuando $\bar{x}$ y s se utilizan en lugar de μ y σ también se deberá reemplazar 1.645 con un número más grande. Por ejemplo, cuando $n = 20$, el valor 2.310 es tal que se puede estar 95% confiado en que el intervalo $\bar{x} \pm 2.310s$ incluirá al menos 90% de los valores de eficiencia de combustible en la población.

Sea k un número entre 0 y 100. Un **intervalo de tolerancia** para capturar al menos k% de los valores en una distribución de población normal con nivel de confianza de 95% tiene la forma

$$\bar{x} \pm (\text{valor crítico de tolerancia}) \cdot s$$

En la tabla A.6 del apéndice aparecen valores críticos de tolerancia para $k = 90$, 95 y 99 en combinación con varios tamaños de muestra. Esta tabla también incluye valores críticos para un nivel de confianza de 99% (estos valores son más grandes que los valores correspondientes a 95%). Si se reemplaza $\pm$ con $+$ se obtiene un límite de tolerancia superior, y si se utiliza $-$ en lugar de $\pm$ se obtiene un límite de tolerancia inferior. En la tabla A.6 también aparecen valores críticos para obtener estos límites unilaterales.

EJEMPLO 6.14 Como parte de un proyecto más amplio para estudiar el comportamiento de los paneles de corteza comprimida, un componente estructural que se utiliza ampliamente en América del Norte, el artículo **"Time-Dependent Bending Properties of Lumber" (*J. of Testing and Eval.*, 1996: 187-193)** informa sobre varias propiedades mecánicas de muestras de madera de pino escocés. Considere las siguientes observaciones sobre el módulo de elasticidad (MPa) obtenido un minuto después de la carga en una cierta configuración:

10 490	16 620	17 300	15 480	12 970	17 260	13 400	13 900
13 630	13 260	14 370	11 700	15 470	17 840	14 070	14 760

Hay un patrón lineal pronunciado en el gráfico de probabilidad normal de los datos. Un resumen de las cantidades importantes es $n = 16$, $\bar{x} = 14\,532.5$, $s = 2055.67$. Para un nivel de confianza de 95%, un intervalo de tolerancia bilateral para la captura de al menos 95% de los valores de los módulos de elasticidad de las muestras de madera en la población de la muestra, utiliza el valor de tolerancia crítica de 2.903. El intervalo resultante es

$$14\,532.5 \pm (2.903)(2055.67) = 14\,532.5 \pm 5967.6 = (8\,564.9, 20\,500.1)$$

Se puede confiar totalmente en que al menos 95% de todos los especímenes de madera tienen valores de módulo de elasticidad de entre 8564.9 y 20 500.1.

El intervalo de confianza de 95% para μ fue (13 437.3, 15 627.7) y el intervalo de predicción de 95% para el módulo de elasticidad de un solo espécimen de madera es (10 017.0, 19 048.0). Tanto el intervalo de predicción como el intervalo de tolerancia son sustancialmente más anchos que el intervalo de confianza. ■

Intervalos basados en distribuciones de población no normales

El intervalo de confianza t para una muestra de μ es robusto en cuanto a alejamientos pequeños o incluso moderados de la normalidad a menos que n sea bastante pequeño. Con esto se quiere decir que si se utiliza un valor crítico para confianza de 95%, por ejemplo, al calcular el intervalo el nivel de confianza real se aproximará de manera razonable al

nivel nominal de 95%. Sin embargo, si n es pequeño y la distribución de la población es altamente no normal, entonces el nivel de confianza real puede ser diferente de forma considerable del que se utiliza cuando se obtiene un valor crítico particular de la tabla t. Ciertamente, ¡sería penoso creer que el nivel de confianza es de más o menos 95% cuando en realidad es como de 88%! Se ha visto que la técnica *bootstrap*, que se presentó en la sección 6.1 es bastante exitosa al estimar parámetros en una amplia variedad de situaciones no normales.

En contraste con el intervalo de confianza la validez de los intervalos de predicción y tolerancia descritos en esta sección está estrechamente vinculada a la suposición de normalidad. Estos últimos intervalos no deberán ser utilizados sin evidencia apremiante de normalidad. La excelente referencia *Statistical Intervals*, que se cita en la bibliografía al final de este capítulo, analiza procedimientos alternativos de esta clase en varias otras situaciones.

EJERCICIOS Sección 6.3 (28–41)

28. Determine los valores de las siguientes cantidades:
a. $t_{0.1,15}$ **b.** $t_{0.05,15}$ **c.** $t_{0.05,25}$ **d.** $t_{0.05,40}$ **e.** $t_{0.005,40}$

29. Determine el valor o los valores críticos t que capturará el área deseada de la curva t en cada uno de los siguientes casos:
a. Área central = 0.95, gl = 10
b. Área central = 0.95, gl = 20
c. Área central = 0.99, gl = 20
d. Área central = 0.99, gl = 50
e. Área de cola superior = 0.01, gl = 25
f. Área de cola inferior = 0.025, gl = 5

30. Determine el valor crítico t de un intervalo de confianza bilateral en cada una de las siguientes situaciones:
a. Nivel de confianza = 95%, gl = 10
b. Nivel de confianza = 95%, gl = 15
c. Nivel de confianza = 99%, gl = 15
d. Nivel de confianza = 99%, $n = 5$
e. Nivel de confianza = 98%, gl = 24
f. Nivel de confianza = 99%, $n = 38$

31. Determine el valor crítico t para un límite de confianza inferior o superior en cada una de las situaciones descritas en el ejercicio 30.

32. De acuerdo con el artículo **"Fatigue Testing of Condoms" (2009: 567-571)**, "las pruebas que se utilizan actualmente para los condones emulan los desafíos que enfrentan al ser utilizados", incluyendo perforaciones, inflado, se prueba el sello del paquete, las dimensiones del condón, incluso la calidad del lubricante (¡todo un territorio fértil para el uso de la metodología estadística!). Los investigadores desarrollaron una nueva prueba que agrega tensión cíclica a un nivel muy por debajo de la rotura y determina el número de ciclos hasta llegar a la ruptura. Una muestra de 20 condones de un tipo particular resultó en una media muestral de 1584 y una desviación estándar muestral de 607. Calcule e interprete un intervalo de confianza al nivel de confianza de 99% para el verdadero número promedio de ciclos de ruptura. [*Nota*: El artículo presenta los resultados de las pruebas de hipótesis basadas en la distribución t, la validez de éstas depende de suponer la distribución normal de la población.]

33. El artículo **"Measuring and Understanding the Aging of Kraft Insulating Paper in Power Transformers" (*IEEE Electrical Insul. Mag.*, 1996: 28–34)** contiene las siguientes observaciones del grado de polimerización de especímenes de papel para los cuales la concentración de tiempos de viscosidad cayeron en un rango medio:

418 421 421 422 425 427 431
434 437 439 446 447 448 453
454 463 465

a. Construya una gráfica de caja de los datos y comente sobre cualquier característica interesante.
b. ¿Es aceptable que las observaciones muestrales dadas se hayan seleccionado de una distribución normal?
c. Calcule un intervalo de confianza de 95% bilateral para un grado de polimerización promedio verdadero (como lo hicieron los autores del artículo). ¿Sugiere este intervalo que 440 es un valor factible del grado de polimerización promedio verdadero? ¿Qué hay en cuanto a 450?

34. Una muestra de 14 especímenes de juntas de un tipo particular produjo un esfuerzo límite proporcional medio muestral de 8.48 MPa y una desviación estándar muestral de 0.79 MPa **("Characterization of Bearing Strength Factors in Pegged Timber Connections", *J. of Structural Engr.*, 1997: 326–332)**.
a. Calcule e interprete un límite de confianza inferior de 95% para el esfuerzo límite proporcional promedio verdadero de todas las juntas. ¿Cuáles fueron sus suposiciones, si las hubo, sobre la distribución del esfuerzo límite proporcional?

b. Calcule e interprete un límite de predicción inferior de 95% para el esfuerzo límite proporcional de una sola unión de este tipo.

35. Para corregir deformidades nasales congénitas se utiliza rinoplastia de aumento mediante un implante de silicón. El éxito del procedimiento depende de varias propiedades biomecánicas del periostio y fascia nasales humanas. El artículo **"Biomechanics in Augmentation Rhinoplasty" (*J. of Med. Engr. and Tech*., 2005: 14-17)** reporta que para una muestra de 15 adultos (recién fallecidos) la deformación por falla media (en porcentaje) fue de 25.0 y la desviación estándar fue de 3.5.

a. Suponiendo una distribución normal para la deformación por falla, estime la deformación promedio verdadera de modo que transmita información acerca de la precisión y la confiabilidad.

b. Pronostique la deformación para un solo adulto de forma que transmita información sobre precisión y confiabilidad. ¿Cómo se compara la predicción con la estimación calculada en el inciso a)?

36. Las $n = 26$ observaciones de tiempo de escape dadas en el ejercicio 36 del capítulo 1 proporcionan una media y desviación estándar muestrales de 370.69 y 24.36, respectivamente.

a. Calcule un límite de confianza superior para el tiempo de escape medio de la población utilizando un nivel de confianza de 95%.

b. Calcule un límite de predicción superior para el tiempo de escape de un solo trabajador adicional utilizando un nivel de predicción de 95%. ¿Cómo se compara este límite con el límite de confianza del inciso a)?

c. Suponga que se escogerán dos trabajadores más para participar en el ejercicio de escape simulado. Denote sus tiempos de escape por X_{27} y X_{28} y sea $\overline{X}_{\text{nuevo}}$ el promedio de estos dos valores. Modifique la fórmula para un intervalo de predicción con un solo valor de x para obtener un intervalo de predicción para $\overline{X}_{\text{nuevo}}$ y calcule un intervalo bilateral de 95% basado en los datos de escape dados.

37. Un estudio de la capacidad de los individuos para caminar en línea recta **("Can We Really Walk Straight?" *Amer. J. of Physical Anthro*., 1992: 19-27)** reportó los siguientes datos sobre la cadencia (pasos por segundo) con una muestra de $n = 20$ hombres saludables seleccionados al azar.

0.95 0.85 0.92 0.95 0.93 0.86 1.00 0.92 0.85 0.81
0.78 0.93 0.93 1.05 0.93 1.06 1.06 0.96 0.81 0.96

Un diagrama de probabilidad normal apoya de manera sustancial la suposición de que la distribución de la población de cadencia es aproximadamente normal. A continuación se da un resumen descriptivo de los datos obtenidos con Minitab:

```
Variable N  Mean   Median  TrMean   StDev SEMean
cadence 20 0.9255 0.9300  0.9261  0.0809 0.0181

Variable     Min     Max      Q1      Q3
cadence    0.7800 1.0600  0.8525  0.9600
```

a. Calcule e interprete un intervalo de confianza de 95% para la cadencia media de la población.

b. Calcule e interprete un intervalo de predicción de 95% para la cadencia de un solo individuo seleccionado al azar de esta población.

c. Calcule un intervalo que incluya al menos 99% de las cadencias en la distribución de la población utilizando un nivel de confianza de 95%.

38. El concreto de ultraalto desempeño (UHPC, por sus siglas en inglés) es un material de construcción relativamente nuevo que se caracteriza por sus fuertes propiedades adhesivas con otros materiales. El artículo **"Adhesive Power of Ultra High Performance Concrete from a Thermodynamic Point of View" (*J. of Materials in Civil Engr*., 2012: 1050–1058)** describe una investigación de las fuerzas intermoleculares del UHPC unido a varios sutratos. En el artículo se presenta el siguiente trabajo de mediciones de adhesión (en mJ/m^2) para las especies de UHPC adheridas al acero:

107.1 109.5 107.4 106.8 108.1

a. ¿Es aceptable que las observaciones en la muestra dada se seleccionaran de una distribución normal?

b. Calcule un intervalo de confianza a 95% (bilateral) para el promedio verdadero del trabajo de adhesión para el UHPC adherido al acero. ¿Sugiere el intervalo que 107 es un valor posible para el trabajo de adhesión promedio verdadero para el UHPC adherido al acero? ¿Qué tal 110?

c. Prediga el valor del trabajo de adhesión resultante a partir de una sola réplica futura del experimento calculando una predicción del intervalo a 95% y compare la amplitud de este intervalo con la amplitud del intervalo de confianza en b).

d. Calcule un intervalo con el cual se puede tener un alto grado de confianza de que al menos 95% de que todas las muestras de UHPC adheridas al acero tienen valores de trabajo de adhesión entre los límites del intervalo.

39. El ejercicio 72 del capítulo 1 aportó las siguientes observaciones en una medición de afinidad de receptor (volumen de distribución ajustado) con una muestra de 13 individuos sanos: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72.

a. ¿Es aceptable que la distribución de la población de la cual se seleccionó esta muestra es normal?

b. Calcule un intervalo con el cual pueda tener 95% de confianza de que al menos 95% de todos los individuos saludables en la población tienen volúmenes de distribución ajustados que quedan entre los límites del intervalo.

c. Pronostique el volumen de distribución ajustado de un solo individuo saludable calculando un intervalo de predicción de 95%. ¿Cómo se compara el ancho de este intervalo con el ancho del intervalo calculado en el inciso b)?

40. El ejercicio 13 del capítulo 1 presenta una muestra de $n = 153$ observaciones de resistencia última a la tensión y el ejercicio 17 de la sección previa dio cantidades resumidas y solicitó un intervalo de confianza muestral grande. Puesto que el tamaño de la muestra es grande, no se requieren suposiciones sobre la distribución de la población en cuanto a la validez del intervalo de confianza.

a. ¿Se requiere alguna suposición sobre la distribución de la resistencia a la tensión, antes de calcular un límite de predicción inferior para la resistencia a la tensión del nuevo

espécimen seleccionado por medio del método descrito en esta sección? Explique.

b. Use un paquete de software estadístico para investigar la factibilidad de una distribución de población normal.

c. Calcule un límite de predicción inferior con un nivel de predicción de 95% para la resistencia última a la tensión del siguiente espécimen seleccionado.

41. Una tabulación más extensa de valores críticos t que la que aparece en este libro muestra que para la distribución t con 20 grados de libertad, las áreas a la derecha de los valores 0.687, 0.860 y 1.064 son 0.25, 0.20 y 0.15, respectivamente. ¿Cuál es el nivel de confianza de cada uno de los siguientes tres intervalos de confianza para la media μ de una distribución de población normal? ¿Cuál de los tres intervalos recomendaría utilizar y por qué?

a. $(\bar{x} - 0.687s/\sqrt{21}, \bar{x} + 1.725s/\sqrt{21})$

b. $(\bar{x} - 0.860s/\sqrt{21}, \bar{x} + 1.325s/\sqrt{21})$

c. $(\bar{x} - 1.064s/\sqrt{21}, \bar{x} + 1.064s/\sqrt{21})$

6.4 Intervalos de confianza para la varianza y desviación estándar de una población normal

Aun cuando las inferencias en lo que se refiere a la varianza σ^2 o a la desviación estándar de una población en general son de menor interés que aquellas respecto a una media o proporción, hay ocasiones en las que se requieren tales procedimientos. En el caso de una distribución de población normal, las inferencias están basadas en el siguiente resultado por lo que se refiere a la varianza muestral S^2.

TEOREMA

Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución normal con parámetros μ y σ^2. Entonces la variable aleatoria

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\Sigma(X_i - \bar{X})^2}{\sigma^2}$$

tiene una distribución de probabilidad ji-cuadrada (χ^2) con $n - 1$ grados de libertad.

Tal como se analizó en las secciones 4.4 y 6.1, la distribución ji-cuadrada es una distribución de probabilidad continua con un solo parámetro ν, llamado número de grados de libertad, con posibles valores de 1, 2, 3, . . . En la figura 6.10 se ilustran las gráficas de varias funciones de densidad de probabilidad χ^2. Cada función de densidad de probabilidad $f(x; \nu)$ es positiva sólo para $x > 0$, y cada una tiene asimetría positiva (una larga cola superior), aunque la distribución se mueve hacia la derecha y se vuelve más simétrica a medida que se incrementa ν. Para especificar procedimientos inferenciales que utilizan la distribución ji-cuadrada, se requiere una notación análoga a aquella para un valor t crítico $t_{\alpha,\nu}$.

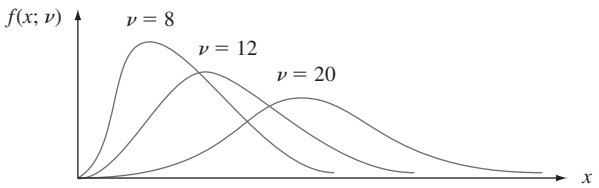


Figura 6.10 Gráficas de funciones de densidad ji-cuadrada

NOTACIÓN

Sea $\chi^2_{\alpha,\nu}$, llamado **valor crítico ji-cuadrada,** el número sobre el eje horizontal de modo que α del área bajo la curva ji-cuadrada con ν grados de libertad quede a la derecha de $X^2_{\alpha,\nu}$.

La simetría de las distribuciones t hizo que fuera necesario tabular sólo valores críticos t de cola superior ($t_{\alpha,\nu}$ con valores pequeños de α). La distribución ji-cuadrada no es simétrica, por lo que la tabla A.7 del apéndice contiene valores de $\chi^2_{\alpha,\nu}$ tanto para α cerca de 0 como cerca de 1, como se ilustra en la figura 6.11(b). Por ejemplo, $\chi^2_{0.025,14} = 26.119$, y $\chi^2_{0.95,20}$ (el 5º percentil) $= 10.851$.

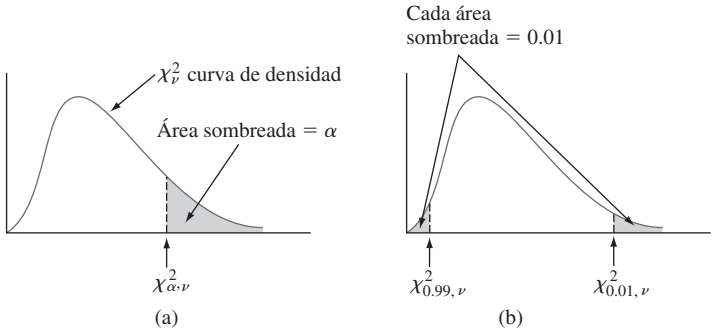


Figura 6.11 $\chi^2_{\alpha,\nu}$ Notación ilustrada

La variable aleatoria $(n-1)S^2/\sigma^2$ satisface las dos propiedades en las cuales está basado el método general para obtener un intervalo de confianza: es una función del parámetro de interés σ^2; no obstante, su distribución de probabilidad (ji-cuadrada) no depende de este parámetro. El área bajo una curva ji-cuadrada con n grados de libertad a la derecha de $\chi^2_{\alpha/2,\nu}$. es $\alpha/2$, lo mismo que a la izquierda de $\chi^2_{1-\alpha/2,\nu}$. De este modo el área capturada entre estos dos valores críticos es $1-\alpha$. Como una consecuencia de esto y del teorema que hemos formulado,

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha \tag{6.17}$$

Las desigualdades en (6.17) equivalen a

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

Al sustituir el valor calculado s^2 en los límites se obtiene un intervalo de confianza para σ^2, y sacar las raíces cuadradas se obtiene un intervalo para σ.

TEOREMA

Un **intervalo de confianza de $100(1-\alpha)\%$ para la varianza σ^2 de una población normal** tiene un límite inferior

$$(n-1)s^2/\chi^2_{\alpha/2,n-1}$$

y el límite superior

$$(n-1)s^2/\chi^2_{1-\alpha/2,n-1}$$

Un **intervalo de confianza para σ** tiene límites superior e inferior que son las raíces cuadradas de los límites correspondientes en el intervalo para σ^2. Un límite de confianza superior o inferior resulta de la sustitución de $\alpha/2$ con α en el límite correspondiente del intervalo de confianza.

EJEMPLO 6.15 Los datos adjuntos sobre voltaje de ruptura de circuitos eléctricamente sobrecargados se tomaron de un diagrama de probabilidad normal que aparece en el artículo **"Damage of Flexible Printed Wiring Boards Associated with Lightning-Induced Voltage Surges"**

(*IEEE Transactions on Components, Hybrids, and Manuf. Tech*., 1985: 214-220). Lo recto del diagrama apoyó de manera firme la suposición de que el voltaje de ruptura está aproximadamente distribuido en forma normal.

1470	1510	1690	1740	1900	2000	2030	2100	2190
2200	2290	2380	2390	2480	2500	2580	2700	

Sea σ^2 la varianza de la distribución del voltaje de ruptura. El valor calculado de la varianza muestral es $s^2 = 137\ 324.3$, la estimación puntual de σ^2. Con grados de libertad $= n - 1 = 16$, un intervalo de confianza de 95% requiere $\chi^2_{0.975,16} = 6.908$ y $\chi^2_{0.025,16} = 28.845$. El intervalo es

$$\left(\frac{16(137\ 324.3)}{28.845},\ \frac{16(137\ 324.3)}{6.908}\right) = (76\ 172.3,\ 318\ 064.4)$$

Al sacar la raíz cuadrada de cada punto extremo se obtiene (276.0, 564.0) como el intervalo de confianza de 95% para σ. Estos intervalos son bastante anchos, lo que refleja la variabilidad sustancial del voltaje de ruptura en combinación con un tamaño de muestra pequeño. ■

Los intervalos de confianza para σ^2 y σ cuando la distribución de la población no es normal pueden ser difíciles de obtener. En esos casos, consulte a un estadístico conocedor.

EJERCICIOS Sección 6.4 (42–46)

42. Determine los valores de las siguientes cantidades:

a. $\chi^2_{0.1,15}$ **b.** $\chi^2_{0.1,25}$
c. $\chi^2_{0.01,25}$ **d.** $\chi^2_{0.005,25}$
e. $\chi^2_{0.99,25}$ **f.** $\chi^2_{0.995,25}$

43. Determine lo siguiente:

a. El 95º percentil de la distribución ji-cuadrada con $\nu = 10$
b. El 5º percentil de la distribución ji-cuadrada con $\nu = 10$.
c. $P(10.98 \leq \chi^2 \leq 36.78)$, donde χ^2 es una variable aleatoria ji-cuadrada con $\nu = 22$
d. $P(\chi^2 < 14.611$ o $\chi^2 > 37.652)$, donde χ^2 es una variable aleatoria ji-cuadrada con $\nu = 25$

44. Se determinó la cantidad de expansión lateral (mils) para una muestra de $n = 9$ soldaduras de arco de gas metálico de energía pulsante utilizadas en tanques de almacenamiento de buques LNG. La desviación estándar muestral resultante fue $s = 2.81$ mils. Suponiendo normalidad, obtenga un intervalo de confianza de 95% para σ^2 y para σ.

45. El mecanizado por descarga eléctrica (WEDM) es un proceso que se utiliza para la fabricación de componentes conductores de metales duros. Utiliza un alambre que se mueve continuamente y que sirve como electrodo. La cobertura del electrodo de alambre permite que el núcleo del electrodo de alambre se enfríe y proporciona un desempeño de corte mejorado. El artículo **"High- Performance Wire Electrodes for Wire Electrical- Discharge Machining—A Review"** (*J. of Engr. Manuf.,* **2012: 1757–1773**) menciona las siguientes observaciones sobre el grosor total de la cobertura del electrodo (en μm) de ocho electrodos utilizados para WEDM:

21 16 29 35 42 24 24 25

Calcule el intervalo de confianza a 99% para la desviación estándar de la distribución del grosor de la cobertura del electrodo. ¿Es válido este intervalo sin importar cuál es la naturaleza de la distribución? Explique.

46. El artículo **"Concrete Pressure on Formwork"** (*Mag. of Concrete Res*., **2009: 407-417**) proporciona las siguientes observaciones sobre la presión máxima del concreto (kN/m^2):

33.2	41.8	37.3	40.2	36.7	39.1	36.2	41.8
36.0	35.2	36.7	38.9	35.8	35.2	40.1	

a. ¿Es factible que esta muestra haya sido seleccionada de una distribución de población normal?
b. Calcule un límite de confianza superior con nivel de confianza de 95% para la desviación estándar de la población de presión máxima.

EJERCICIOS SUPLEMENTARIOS (47–57)

47. El ejemplo 1.11 introdujo las siguientes observaciones sobre la fuerza de adhesión.

11.5	12.1	9.9	9.3	7.8	6.2	6.6	7.0
13.4	17.1	9.3	5.6	5.7	5.4	5.2	5.1
4.9	10.7	15.2	8.5	4.2	4.0	3.9	3.8
3.6	3.4	20.6	25.5	13.8	12.6	13.1	8.9
8.2	10.7	14.2	7.6	5.2	5.5	5.1	5.0
5.2	4.8	4.1	3.8	3.7	3.6	3.6	3.6

a. Calcule la fuerza de adhesión promedio verdadera de una manera que proporcione información sobre precisión y confiabilidad. [*Sugerencia*: $\Sigma x_i = 387.8$ y $\Sigma x_i^2 = 4247.08$.]

b. Calcule un intervalo de confianza de 95% para la proporción de todas las adhesiones cuyos valores de fuerza excederían de 10.

48. El artículo **"Distributions of Compressive Strength Obtained from Various Diameter Cores" (*ACI Materials J.*, 2012: 597–606)** describe un estudio en el que se determinaron las fuerzas de compresión en muestras de concreto de diferentes tipos, con diferentes diámetros de núcleo y con diferentes relaciones longitud-diámetro. Para un tipo, diámetro y relación l/d, en particular las 18 muestras probadas resultaron en una media muestral de fuerza de compresión de 64.41 MPa y una desviación estándar de muestra de 10.32 MPa. El comportamiento normal de la distribución de la fuerza de compresión se consideró bastante posible.

a. Calcule el intervalo de confianza con un nivel de confianza de 98% para el promedio de la fuerza de compresión real bajo estas circunstancias.

b. Calcule un límite inferior de predicción de 98% para la fuerza de compresión de una sola especie futura probada en las circunstancias dadas. [*Sugerencia*: $t_{0.02,17} = 2.224$.]

49. Para aquellos que no lo saben, las regatas Dragon Boat son un deporte acuático de competencia que involucra a 20 remeros que reman a lo largo de diversas distancias durante la carrera. Se ha hecho bastante popular en los últimos años. El artículo **"Physiological and Physical Characteristics of Elite Dragon Boat Paddlers" (*J. of Strength and Conditioning*, 2013: 137–145)** resume un extenso análisis estadístico de los datos obtenidos a partir de una muestra de 11 remeros. Reporta que un intervalo de confianza a 95% para la fuerza promedio real (N) durante una carrera simulada de 200 m fue (60.2, 70.6). Obtenga una predicción del intervalo de confianza a 95% para la fuerza de un solo remero, seleccionado al azar, que participa en la carrera simulada.

50. Un artículo publicado en un periódico reporta que se utilizó una muestra de tamaño 5 como base para calcular un intervalo de confianza de 95% para la frecuencia natural (Hz) promedio verdadera de vigas deslaminadas de cierto tipo. El intervalo resultante fue (229.764, 233.504). Usted decide que un nivel de confianza de 99% es más apropiado que el de 95% utilizado. ¿Cuáles son los límites del intervalo de 99%? [*Sugerencia*: Use el centro del intervalo y su ancho para determinar $\bar{x}$ y s.]

51. Los síntomas respiratorios inexplicables reportados por los atletas suelen ser considerados asma secundaria inducida por ejercicio. El artículo **"High Prevalence of Exercise-Induced Laryngeal Obstruction in Athletes" (*Medicine and Science in Sports and Exercise*, 2013: 2030–2035)** sugiere que muchos de esos casos podrían explicarse mejor por la obstrucción de la laringe. De una muestra de 88 atletas que ingresaron para realizarse un estudio de asma, 31 mostraron un asma secundaria inducida por ejercicio.

a. Calcule e interprete el intervalo de confianza utilizando un nivel de confianza a 95% para la proporción real de todos los atletas que presentaron asma secundaria inducida por ejercicio en estas circunstancias.

b. ¿Qué tamaño de muestra se requiere si se desea que el ancho del intervalo de confianza a 95% sea cuando más de 0.04, sin importar los resultados de la muestra?

c. ¿El límite superior del intervalo en el inciso a) especifica una fiabilidad de 95% en el límite superior para la proporción calculada? Explique.

52. La alta concentración del elemento tóxico arsénico es demasiado común en el agua subterránea. El artículo **"Evaluation of Treatment Systems for the Removal of Arsenic from Groundwater" (*Practice Periodical of Hazardous, Toxic, and Radioactive Waste Mgmt.*, 2005: 152–157)** reporta que para una muestra de $n = 5$ especímenes de agua seleccionada para tratamiento por coagulación, la concentración de arsénico media muestral fue de 24.3 μg/L y la desviación estándar muestral fue de 4.1. Los autores del artículo citado utilizaron métodos basados en t para analizar sus datos, así que venturosamente tuvieron razón al creer que la distribución de la concentración de arsénico era normal.

a. Calcule e interprete un intervalo de confianza de 95% para la concentración de arsénico promedio verdadera en todos los especímenes de agua.

b. Calcule un límite de confianza superior de 90% para la desviación estándar de la distribución de la concentración de arsénico.

c. Pronostique la concentración de arsénico de un solo espécimen de agua de modo que proporcione información sobre precisión y confiabilidad.

53. La infestación con pulgones de árboles frutales puede ser controlada rociando un pesticida o mediante el tratamiento con mariquitas. En un área particular se seleccionan cuatro diferentes huertas de árboles frutales para experimentación. Las primeras tres arboledas se rocían con los pesticidas 1, 2 y 3,

respectivamente, y la cuarta se trata con mariquitas con los siguientes resultados de cosecha:

Tratamiento	n_i = número de árboles	$\bar{x}_i$ arbustos/árbol	s_i
1	100	10.5	1.5
2	90	10.0	1.3
3	100	10.1	1.8
4	120	10.7	1.6

Sea μ_i = la cosecha promedio verdadera (arbustos/árbol) después de recibir el i-ésimo tratamiento. Entonces

$$\theta = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4$$

mide la diferencia de las cosechas promedio verdaderas entre el tratamiento con pesticidas y el tratamiento con mariquitas. Cuando n_1, n_2, n_3 y n_4 son grandes, el estimador $\hat{\theta}$ obtenido al reemplazar cada μ_i con $\overline{X}_i$ es aproximadamente normal. Use esto para deducir un intervalo de confianza muestral grande de $100(1 - \alpha)\%$ para θ y calcule el intervalo de 95% para los datos dados.

54. Es importante que las máscaras que utilizan los bomberos sean efectivas para soportar altas temperaturas porque los bomberos comúnmente trabajan a temperaturas de 200–500° F. En una prueba de un tipo de máscara, a 11 de 55 máscaras se les desprendió la mica a 250°. Construya un intervalo de confianza de 90% para la verdadera proporción de máscaras de este tipo cuya mica se desprendería a 250°.

55. Un fabricante de libros de texto universitarios está interesado en investigar la resistencia de las encuadernaciones producidas por una encuadernadora particular. La resistencia puede ser medida registrando la fuerza que se requiere para arrancar las páginas de un libro. Si esta fuerza se mide en libras, ¿cuántos libros deberán ser probados para calcular la fuerza promedio requerida para romper la encuadernación dentro de 0.1 lb con 95% de confianza? Suponga que se sabe que σ es de 0.8.

56. Los siguientes datos sobre la profundidad a la que inician las grietas (μm) fueron tomados a partir de una gráfica de probabilidad lognormal que aparece en el artículo **"Incorporating Small Fatigue Crack Growth in Probabilistic Life Prediction: Effect of Stress Ratio in Ti-6Al-2Sn-6Mo" (*Intl. J. of Fatigue,* 2013: 83–95)**. Aunque el patrón en la gráfica es bastante recto, una gráfica de probabilidad normal de los datos también muestra un patrón razonablemente lineal. Una gráfica de caja indica que la distribución es bastante simétrica cerca del 50% central de los datos y sólo ligeramente sesgada de manera global. Es, por tanto, razonable estimar y predecir utilizando intervalos t.

4.7 5.1 5.2 5.3 5.6 5.8 6.3 6.7
7.2 7.4 7.7 8.5 8.9 9.3 10.1 11.2

a. Estime el valor promedio real de la profundidad a la que inicia la grieta con un intervalo de confianza a 99% e interprete el intervalo resultante.

b. Prediga el valor de una sola medición de la profundidad a la que inicia la grieta construyendo un intervalo de predicción de 99%.

c. En este contexto interprete el significado de 99% en el inciso b).

57. En el ejemplo 5.8 se introdujo el concepto de experimento censurado en el cual se prueban n componentes y el experimento termina en cuanto fallan r de los componentes. Suponga que las vidas útiles de los componentes son independientes, cada una con distribución exponencial y parámetro λ. Sea Y_1 el tiempo en el cual ocurre la primera falla, Y_2 el tiempo en el cual ocurre la segunda falla, y así sucesivamente, de modo que $T_r = Y_1 + \cdots + Y_r + (n - r)Y_r$, es la vida útil total acumulada. En este caso se puede demostrar que $2\lambda T_r$ tiene una distribución ji-cuadrada con $2r$ grados de libertad. Use esto para desarrollar una fórmula para un intervalo de confianza $100(1 - \alpha)\%$ para una vida útil promedio verdadera $1/\lambda$. Calcule un intervalo de confianza de 95% con los datos del ejemplo 5.8.

BIBLIOGRAFÍA

DeGroot, Morris y Mark Schervish, *Probability and Statistics* (3a. ed.), Addison-Wesley, Reading, MA, 2002. Una muy buena exposición de los principios generales de inferencia estadística.

Devore, Jay y Kenneth Berk, *Modern Mathematical Statistics with Applications*, Cengage, Belmont, CA, 2007. La exposición es un poco más completa y sofisticada que la del presente libro e incluye más material sobre *bootstrapping*.

Hahn, Gerald y William Meeker, *Statistical Intervals*, Wiley, Nueva York, 1991. Todo lo que alguna vez quiso saber sobre intervalos estadísticos (de confianza, predicción, tolerancia y otros).

Pruebas de hipótesis basadas en una sola muestra

Capítulo 7

INTRODUCCIÓN

Un parámetro puede ser estimado a partir de datos muestrales con un solo número (una estimación puntual) o un intervalo completo de valores factibles (un intervalo de confianza). Con frecuencia, sin embargo, el objetivo de una investigación no es estimar un parámetro sino decidir cuál de dos afirmaciones contradictorias sobre el parámetro es la correcta. Los métodos para lograr esto comprenden la parte de la inferencia estadística llamada *prueba de hipótesis*. En este capítulo primero se analizan algunos conceptos y terminología básicos en la prueba de hipótesis y luego se desarrollan procedimientos para la toma de decisiones para los problemas de realización de pruebas con base en una muestra tomada de una sola población más frecuentemente encontrados.

7.1 Hipótesis y procedimientos de prueba

Una **hipótesis estadística** o simplemente una *hipótesis* es una afirmación o aseveración sobre el valor de un solo parámetro (característica de una población o característica de una distribución de probabilidad), sobre los valores de varios parámetros o sobre la forma de una distribución de probabilidad completa. Un ejemplo de una hipótesis es la pretensión de que $\mu = 0.75$, donde μ es el diámetro interno promedio verdadero de un cierto tipo de tubo de PVC. Otro ejemplo es la proposición $p < 0.10$, donde p es la proporción de tarjetas de circuito defectuosas entre todas las tarjetas de circuito producidas por un fabricante. Si μ_1 y μ_2 denotan las verdaderas resistencias a la ruptura promedio de dos tipos diferentes de cuerdas, una hipótesis es la aseveración de que $\mu_1 - \mu_2 = 0$ y otra es que $\mu_1 - \mu_2 > 5$. No obstante, otro ejemplo de una hipótesis es la aseveración de que la distancia de frenado en condiciones particulares tiene una distribución normal.

En cualquier problema de prueba de hipótesis hay dos hipótesis contradictorias en consideración. Una podría ser la pretensión de que $\mu = 0.75$ y la otra $\mu \neq 0.75$, o las dos proposiciones contradictorias podrían ser $p \geq 0.10$ y $p < 0.10$. El objetivo es decidir, con base en información muestral, cuál de las dos hipótesis es la correcta. Existe una conocida analogía de esto durante un juicio criminal. Una pretensión es la aseveración de que el individuo acusado es inocente. En el sistema judicial estadounidense esta es la pretensión que inicialmente se cree que es verdadera. Sólo de cara a una fuerte evidencia que diga lo contrario el jurado deberá rechazar esta pretensión en favor de la aseveración alternativa de que el acusado es culpable. En este sentido, la pretensión de inocencia es la hipótesis favorecida o protegida y la obligación de la comprobación recae en aquellos que creen en la pretensión alternativa.

Asimismo, al probar hipótesis estadísticas, el problema se formulará de modo que una de las pretensiones sea favorecida al inicio. Esta pretensión inicialmente favorecida no será rechazada en favor de la pretensión alternativa a menos que la evidencia de la muestra la contradiga y apoye con fuerza la aseveración alternativa.

DEFINICIÓN

La **hipótesis nula** denotada por H_0, es la pretensión que inicialmente se supone verdadera (la pretensión de "creencia previa"). La **hipótesis alternativa** denotada por H_a, es la aseveración contradictoria de H_0.

La hipótesis nula será rechazada en favor de la hipótesis alternativa sólo si la evidencia muestral sugiere que H_0 es falsa. Si la muestra no contradice fuertemente a H_0 se continuará creyendo en la factibilidad de la hipótesis nula. Las dos posibles conclusiones derivadas de un análisis de prueba de hipótesis son entonces *rechazar* H_0 o *no rechazar* H_0.

Una **prueba de hipótesis** es un método en el que se utilizan datos muestrales para decidir si la hipótesis nula debe ser rechazada. Por consiguiente, se podría probar H_0: $\mu = 0.75$ contra la H_a alternativa: $\mu \neq 0.75$. Sólo si los datos muestrales sugieren fuertemente que μ es otra diferente de 0.75 la hipótesis nula deberá ser rechazada. Sin semejante evidencia, H_0 no deberá ser rechazada, puesto que sigue siendo bastante factible.

En ocasiones un investigador no desea aceptar una aseveración particular a menos y hasta que los datos apoyen fuertemente la aseveración. Como ejemplo, suponga que una compañía está considerando aplicar un nuevo tipo de recubrimiento en los cojinetes que fabrica. Se sabe que la vida de desgaste promedio verdadera con el recubrimiento

actual es de 1000 horas. Si μ denota la vida promedio verdadera del nuevo recubrimiento, la compañía no desea cambiar a menos que la evidencia sugiera fuertemente que μ excede de 1000. Una formulación apropiada del problema implicaría probar H_0: $\mu = 1000$ contra H_a: $\mu > 1000$. La conclusión de que se justifica un cambio está identificada con H_a y se requeriría evidencia conclusiva para justificar el rechazo de H_0 y cambiar al nuevo recubrimiento.

La investigación científica a menudo implica tratar de decidir si una teoría actual debe ser reemplazada por una explicación más factible y satisfactoria del fenómeno investigado. Un método conservador es identificar la teoría actual con H_0 y la explicación alternativa del investigador con H_a. El rechazo de la teoría actual ocurrirá entonces sólo cuando la evidencia sea mucho más compatible con la nueva teoría. En muchas situaciones H_a se conoce como la "hipótesis del investigador", puesto que es la pretensión que al investigador en realidad le gustaría validar. La palabra *nula* significa "sin ningún valor, efecto o consecuencia", lo que sugiere que H_0 debería ser identificada con la hipótesis de ningún cambio (de la opinión actual), ninguna diferencia, ninguna mejora, y así sucesivamente. Suponga, por ejemplo, que 10% de todas las tarjetas de circuito producidas por un fabricante durante un periodo reciente están defectuosas. Un ingeniero ha sugerido un cambio en el proceso de producción en la creencia de que dará por resultado una proporción reducida de tarjetas defectuosas. Sea p la proporción verdadera de tarjetas defectuosas que resultan del cambio en el proceso. Entonces la hipótesis de investigación en la cual recae la obligación de la comprobación es la aseveración de que $p < 0.10$. Por consiguiente, la hipótesis alternativa es H_a: $p < 0.10$.

En el tratamiento de la prueba de hipótesis, H_0 generalmente será formulada como pretensión de igualdad. Si θ denota el parámetro de interés, la hipótesis nula tendrá la forma H_0: $\theta = \theta_0$ donde θ_0 es un número específico llamado **valor nulo** del parámetro (valor pretendido para θ por la hipótesis nula). Por ejemplo, considere la situación de la tarjeta de circuito que se acaba de comentar. La hipótesis alternativa sugerida fue H_a: $p < 0.10$, la pretensión de que la modificación del proceso reduce la proporción de tarjetas defectuosas. Una elección natural de H_0 en esta situación es la pretensión de que $p \geq 0.10$ de acuerdo con la cual el nuevo proceso no es mejor o peor que el actualmente utilizado. En su lugar se considerará H_0: $p = 0.10$ contra H_a: $p < 0.10$. El razonamiento para utilizar esta hipótesis nula simplificada es que cualquier procedimiento de decisión razonable para decidir entre H_0: $p = 0.10$ y H_a: $p < 0.10$ también será razonable para decidir entre la pretensión de que $p \geq 0.10$ y H_a. Se prefiere utilizar una H_0 simplificada porque tiene ciertos beneficios técnicos, los que en breve serán aparentes.

La alternativa a la hipótesis nula H_0: $\theta = \theta_0$ se verá como una de las siguientes tres aseveraciones:

1. H_a: $\theta > \theta_0$ (en cuyo caso la hipótesis nula implícita es $\theta \leq \theta_0$),
2. H_a: $\theta < \theta_0$ (en cuyo caso la hipótesis nula implícita es $\theta \geq \theta_0$), o
3. H_a: $\theta \neq \theta_0$

Por ejemplo, sea σ la desviación estándar de la distribución de diámetros internos (pulgadas) de cierto tipo de mango de metal. Si se decidió utilizar el mango a menos que la evidencia muestral demuestre conclusivamente que $\sigma > 0.001$, la hipótesis apropiada sería H_0: $\sigma = 0.001$ versus H_a: $\sigma > 0.001$. El número θ_0 que aparece tanto en H_0 como en H_a (separando la alternativa de la nula) se llama valor nulo.

Procedimientos de prueba y valores *P*

Un procedimiento de prueba es una regla basada en datos muestrales para decidir si se rechaza H_0. El punto clave será el siguiente. Supongamos que H_0 es verdadera. Entonces ¿cuál es la probabilidad de que al menos una muestra (aleatoria) contradiga esta hipótesis como lo haría nuestra muestra? Considere los siguientes dos escenarios:

1. Hay sólo 0.1% de posibilidad (probabilidad de 0.001) de obtener una muestra al menos tan contradictoria a H_0 como la que obtuvimos al suponer que H_0 es verdadera.
2. Hay 25% de posibilidades (probabilidad de 0.25) de obtener una muestra al menos tan contradictoria a H_0 como la que se obtiene cuando H_0 es verdadera.

En el primer escenario, algo tan extremo como nuestra muestra es muy poco probable que se produzca cuando H_0 es verdadera, a la larga sólo 1 en 1000 muestras sería al menos tan contradictoria con la hipótesis nula como la que hemos seleccionado. Por otro lado, para el segundo escenario, a largo plazo 25 de cada 100 muestras serían al menos tan contradictorias a H_0 como la que obtuvimos al suponer que la hipótesis nula es verdadera. Así, nuestra muestra es muy consistente con H_0 y no hay ninguna razón para rechazarla.

Debemos ahora profundizar en este razonamiento por ser más específicos en cuanto a lo que se entiende por "al menos tan contradictorio a H_0 como la muestra que se obtiene cuando H_0 es verdadera". Antes de hacerlo de una manera general, consideremos varios ejemplos.

EJEMPLO 7.1 La empresa que fabrica yogur estilo griego marca D desea aumentar su cuota de mercado y en particular de convencer a los que actualmente prefieren la marca C a cambiar de marca. Por tanto, el departamento de mercadeo ha ideado el siguiente experimento ciego de sabor. A cada uno de los 100 consumidores de la marca C se le pide que pruebe yogur de dos tazones, uno con la marca C y el otro con la marca D y luego se les pregunta cuál prefiere. Los tazones están marcados con un código para que los experimentadores sepan qué tazón contiene cuál yogur, pero los experimentadores no tienen esta información (Nota: Este experimento se realizó en realidad con cervezas hace varias décadas, con la ya desaparecida cerveza Schlitz representando el papel de la marca D y siendo la cerveza Michelob la marca objetivo).

Sea p la proporción de todos los consumidores de marca C que prefiere C en lugar de D en dichas circunstancias. Vamos a considerar la prueba de las hipótesis H_0: $p = 0.5$ contra H_a: $p < 0.5$. La hipótesis alternativa afirma que la mayoría de los consumidores de la marca C prefiere la marca D. Por supuesto, a la compañía de la marca D le gustaría que H_0 fuera rechazada, por tanto, considera H_a como la hipótesis más factible. Si la hipótesis nula es verdadera y un solo consumidor de la marca C, seleccionado al azar, prefiere C o D el resultado es el mismo que si hubiera lanzado una moneda al aire.

Los datos de la muestra consistirán en una secuencia de 100 preferencias, siendo cada una C o D. Sea X = el número entre los 100 individuos seleccionados que prefieren C a D. Esta variable aleatoria servirá como nuestro *estadístico de prueba*, la función de los datos de la muestra en la que basaremos nuestra conclusión. Ahora X es una variable aleatoria binomial (el número de éxitos en un experimento con un número fijo de ensayos independientes con una probabilidad de éxito constante p). Cuando H_0 es verdadera, este estadístico de prueba tiene una distribución binomial con $p = 0.5$, en cuyo caso $E(X) = np = 100(0.5) = 50$.

Intuitivamente, un valor de X "considerablemente" menor de 50 defiende el rechazo de H_0 en favor de H_a. Suponga que el valor observado de X es $x = 37$. ¿Cómo contradice este valor a la hipótesis nula? Para responder esta pregunta, vamos a identificar primero los valores de X que son aún más contradictorios a H_0 como el mismo 37. Evidentemente 35 es un valor y 30 es otro; de hecho, cualquier número menor que 37 es un valor de X más contradictorio a la hipótesis nula que es el valor que realmente observamos. Ahora considere la probabilidad, calculada suponiendo que la hipótesis nula es verdadera, para obtener un valor de X al menos tan contradictorio a H_0 como lo es nuestro valor observado:

$$\begin{aligned} P(X \le 37 \text{ cuando } H_0 \text{ es verdadera}) &= P(X \le 37 \text{ cuando } X \sim \text{Bin}(100, 0.5)) \\ &= B(37; 100, 0.5) = 0.006 \end{aligned}$$

(del software). Así si la hipótesis nula es verdadera, hay menos de 1% de posibilidades de ver 37 o menos éxitos entre 100 ensayos. Esto sugiere que $x = 37$ es mucho más consistente con la hipótesis alternativa que con el valor nulo, y que el rechazo de H_0 a favor de H_a es una conclusión sensata. Además, considere que $\sigma_X = \sqrt{npq} = \sqrt{100(0.5)(0.5)} = 5$ cuando H_0 es verdadera. Se deduce que 37 es más de 2.5 desviaciones estándar menores de lo que esperaríamos ver si H_0 fuera verdadera.

Ahora supongamos que 45 de los 100 individuos en el experimento prefieren C (45 éxitos). Nuevamente calculamos la probabilidad, suponiendo H_0 verdadera, para conseguir un valor del estadístico de prueba menos contradictorio a H_0 como este:

$$\begin{aligned} P(X \le 45 \text{ cuando } H_0 \text{ es verdadera}) &= P(X \le 45 \text{ cuando } X \sim \text{Bin}(100, 0.5)) \\ &= B(45; 100, 0.5) = 0.184 \end{aligned}$$

O sea que si, de hecho, $p = 0.5$, no sería sorprendente ver 45 o menos éxitos. Por esta razón el valor 45 no parece muy contradictorio a H_0 (es sólo una desviación estándar más pequeña de lo que se esperaba que H_0 fuera verdadera). Rechazar H_0 en este caso no parece razonable. ■

EJEMPLO 7.2 De acuerdo con el artículo **"Freshman 15: Fact or Fiction" (*Obesity*, 2006: 1438–1443)**, "una creencia común entre el público en general es que después de entrar a la universidad se aumenta de peso corporal y la frase 'novato 15' ha sido acuñada para describir las 15 libras que presumiblemente los estudiantes aumentan durante su primer año". Sea μ la ganancia de peso promedio verdadero de la mujer en el transcurso de su primer año en la universidad. ¿La cita anterior sugiere que debemos probar las hipótesis H_0: $\mu = 15$ versus H_a: $\mu \neq 15$. Para esto, suponga que se selecciona una muestra aleatoria de n de estos individuos y se determina el aumento de peso de cada uno, obteniendo como resultado una media muestral de ganancia de peso $\bar{x}$ y una desviación estándar de la muestra s (Nota: Estos datos realmente están *pareados*, con cada aumento de peso resultante se obtiene un par de pesos (inicio, final) y después se resta para determinar la diferencia; en la sección 9.3 se comentará más acerca de estos datos). Antes de que se obtuvieran los datos la media de la ganancia de peso de la muestra es una variable aleatoria $\overline{X}$ y la desviación estándar de la muestra también es una variable aleatoria S.

Un estadístico de prueba natural (función de datos en la que se basará la decisión) es la media muestral $\overline{X}$; si H_0 es verdadera, entonces $E(\overline{X}) = \mu = 15$, mientras que si μ difiere considerablemente de 15, entonces la media muestral del aumento de peso debería ser la misma. Pero hay un estadístico de prueba más conveniente que tiene propiedades atractivas intuitivas y técnicas: la media muestral se estandarizada suponiendo que H_0 es verdadera. Recuerde que la desviación estándar (error estándar) de $\overline{X}$ es $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Suponiendo que la distribución poblacional de las ganancias de peso es normal, se deduce que la distribución muestral de $\overline{X}$ es normal. Ahora, al estandarizar una variable normalmente distribuida se obtiene una variable que tiene una distribución normal estándar (la curva z):

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Si se conoce el valor de σ, es posible obtener un estadístico de prueba simplemente al sustituir μ por el valor nulo $\mu_0 = 15$:

$$Z = \frac{\overline{X} - 15}{\sigma/\sqrt{n}}$$

Si se sustituyen $\bar{x}$, σ y n se obtiene $z = 3$, la interpretación es que el valor observado de la media muestral es más grande por tres desviaciones estándar que lo que esperaríamos si la hipótesis nula fuera verdadera. Por supuesto, en "situaciones normales" este caso es extremadamente

raro. Por otra parte, si $z = -1$, entonces la media muestral es sólo una desviación estándar menor de lo que se esperaría bajo H_0, un resultado que no es lo suficientemente sorprendente para formular una duda sustancial en H_0.

Un error práctico en el desarrollo anterior es que el valor de σ casi nunca está disponible para un investigador. Sin embargo, como comentamos en el capítulo anterior, la sustitución de S por σ en Z normalmente presenta poca variabilidad extra cuando n es grande ($n > 40$ era la regla empírica anterior). En este caso la variable resultante aún tiene *aproximadamente* una distribución normal estándar. El estadístico de prueba de muestra grande para nuestro escenario de aumento de peso es

$$Z = \frac{\overline{X} - 15}{S/\sqrt{n}}$$

Así cuando H_0 es verdadera, Z tiene aproximadamente una distribución normal estándar.

Suponga que $\overline{x} = 13.7$ y al sustituirla junto con s y n se obtiene $z = -2.80$. ¿Cuáles son los valores del estadístico de prueba al menos tan contradictorios a H_0 como -2.80 mismo? Para responder esto se determinan primero los valores de $\overline{x}$ que son al menos tan contradictorios a H_0 como 13.7. Uno de estos valores es 13.5, otro es 13.0, y de hecho *cualquier* valor menor de 13.7 es más contradictorio a H_0 que 13.7.

Pero esto no es todo. Recuerde que la hipótesis alternativa afirma que el valor de μ es uno diferente de 15. En vista de lo anterior, el valor de 16.3 es sólo tan contradictorio a H_0 como lo es 13.7; ya que se encuentra a la misma distancia arriba del valor nulo de 15 como 13.7 está por debajo de 15, y el valor de z resultante es 3.0, igual de extremo que -3.0. Y cualquier $\overline{x}$ particular que exceda a 16.3 es igual de contradictorio a H_0, ya que es un valor que está a la misma distancia por debajo de 15, por ejemplo, 15.8 y 14.2 , 17.0 y 13.0 etcétera.

Igual que los valores de $\overline{x}$ que están a más 13.7 corresponden a $z \leq -2.80$, los valores de $\overline{x}$ que están al menos a 16.3 corresponden a $z \geq 2.80$. Así, los valores del estadístico de prueba que son al menos tan contradictorios a H_0 como el valor -2.80 realmente obtenido son $\{z\colon z \leq -2.80 \text{ o } z \geq 2.80\}$. Ahora podemos calcular la probabilidad, suponiendo que H_0 es verdadera, obteniendo un valor del estadístico de prueba al menos tan contradictorio a H_0 como nuestra muestra producida:

$$\begin{aligned} &P(Z \leq -22.80 \text{ o } Z \geq 2.80 \text{ suponiendo que } H_0 \text{ es verdadera}) \\ &\quad \approx 2 \cdot (\text{área bajo la curva de } z \text{ a la derecha de } 2.80) \\ &\quad = 2[1 - \Phi(2.80)] = 2[1 - 0.9974] = 0.0052 \end{aligned}$$

Es decir, si la hipótesis nula es, en efecto, verdadera, sólo cerca de la mitad del uno por ciento de todas las muestras daría lugar a un valor del estadístico de prueba al menos tan contradictorio a la hipótesis nula como lo es nuestro valor. Es evidente que -2.80 está entre los valores posibles de los estadísticos de prueba que son más contradictorios a H_0. Por tanto, tendría sentido rechazar H_0 en favor de H_a.

Suponga que, en cambio, hubiésemos obtenido el valor del estadístico de prueba $z = 0.89$, el cual es menor que una desviación estándar más grande que lo que nos esperaría si H_0 fuera verdadera. La probabilidad anterior sería entonces

$$\begin{aligned} &P(Z \leq -0.89 \text{ o } Z \geq 0.89 \text{ suponiendo que } H_0 \text{ es verdadera}) \\ &\quad \approx 2 \cdot (\text{área bajo la curva } z \text{ a la derecha de } 0.89) \\ &\quad = 2[1 - \Phi(0.89)] = 2[1 - 0.8133] = 0.3734 \end{aligned}$$

Más de $1/3$ de todas las muestras darían un valor de estadístico de prueba al menos tan contradictorio a H_0 como lo es 0.89 cuando H_0 es verdadera. Así los datos son bastante coherentes con la hipótesis nula; y sigue siendo aceptable que $\mu = 15$.

El artículo citado al inicio de este ejemplo informa que en una muestra de 137 estudiantes, la media del aumento de peso de la muestra era sólo de 2.42 lb con una desviación estándar de muestra de 5.72 lb (algunos alumnos pierden peso). ¡Esto da $z = (2.42 - 15)/$

$(5.72/\sqrt{137}) = -25.7$! La probabilidad de observar un valor de al menos este valor en cualquier dirección es en principio 0. Los datos contradicen muy fuertemente la hipótesis nula, y hay pruebas sustanciales de que un verdadero aumento de peso promedio está mucho más cerca de 0 que de 15. ■

El tipo de probabilidad calculada en los ejemplos 7.1 y 7.2 ahora servirá de base para la obtención de procedimientos generales de prueba.

DEFINICIONES

Un **estadístico de prueba** es una función de los datos de la muestra utilizada como base para decidir si H_0 debe ser rechazada. El estadístico de prueba seleccionado debe discriminar eficazmente entre las dos hipótesis. Es decir, los valores del estadístico que tienden a producirse cuando H_0 es verdadera deben ser absolutamente diferentes de los observados generalmente cuando H_0 no es verdadera.

El **valor *P*** es la probabilidad, calculada al suponer que la hipótesis nula es verdadera, de obtener un valor del estadístico de prueba al menos tan contradictorio a H_0 como el valor calculado de los datos de la muestra disponible. En una hipótesis de un análisis de prueba se llega a una conclusión seleccionando un número α, llamado **nivel significancia** (o *nivel de significancia*) de la prueba, que está razonablemente cercano a 0. Entonces, si el valor $P \leq \alpha$ se rechazará H_0 en favor de H_a, mientras que si el valor $P > \alpha$, H_0 no se rechazará (aun considerando la posibilidad). En la práctica, los niveles de significancia utilizados con más frecuencia son (en orden) $\alpha = 0.05$, 0.01, 0.001 y 0.10.

Por ejemplo, si se selecciona un nivel de significancia de 0.05 y después calculamos un valor $P = 0.0032$, H_0 se rechazaría porque $0.0032 \leq 0.05$. Con este mismo valor de P la hipótesis nula también se rechazaría en el menor nivel de significancia de 0.01, ya que $0.0032 \leq 0.01$. Sin embargo, en un nivel de significancia de 0.001 no se puede rechazar a H_0 ya que $0.0032 > 0.001$. La figura 7.1 ilustra la comparación del valor P con el nivel de significancia para obtener una conclusión.

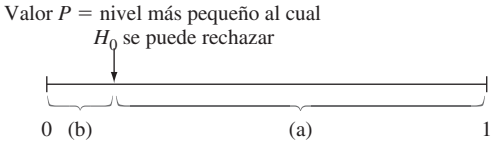


Figura 7.1 Comparación de α y el valor P: (a) H_0 se rechaza cuando α está aquí; (b) H_0 no se rechaza cuando α está aquí

Pronto se considerará en detalle las consecuencias de seleccionar un nivel de significancia más pequeño en lugar de uno más grande. Por el momento, observe que entre menor sea el nivel de significancia, más protección se le está dando a la hipótesis nula y más difícil resultará rechazar esa hipótesis.

La definición de un valor P obviamente es un poco complicada, y no se puede decir nada sin una buena dosis de práctica. De hecho, muchos usuarios de metodología estadística utilizan la regla de decisión dada repetidamente para probar la hipótesis, ¡pero les sería difícil decir lo que es un valor P! A continuación se presentan algunos puntos importantes:

- El valor P es una probabilidad.
- Esta probabilidad se calcula suponiendo que la hipótesis nula es verdadera.
- Para determinar el valor P primero se debe decidir qué valores del estadístico de prueba son al menos tan contradictorios a H_0 como el valor obtenido de nuestra muestra.

- Entre más pequeño sea el valor P, más fuerte es la evidencia contra H_0 y a favor de H_a.
- El valor P no es la probabilidad de que la hipótesis nula sea verdadera o falsa, ni la probabilidad de que se llegue a una conclusión errónea.

EJEMPLO 7.3 Las aguas pluviales urbanas pueden estar contaminadas por muchas fuentes, incluyendo las pilas desechadas. Cuando se rompen estas baterías liberan metales de importancia ambiental. El artículo **"Urban Battery Litter" (*J. Environ. Engr.*, 2009: 46–57)** presenta los datos que resumen las características de una variedad de baterías que se encuentran en las zonas urbanas de Cleveland. Una muestra aleatoria de 51 pilas Panasonic AAA dio una media muestral de masa de zinc de 2.06 g y una desviación estándar de la muestra de 0.141 g. ¿Estos datos proporcionarán pruebas convincentes para concluir que la media de la población de la masa de zinc es superior a 2.0 g? Se va a emplear un nivel de significancia de 0.01 para obtener una conclusión.

Con μ igual a la media de la masa de zinc verdadera de estas baterías, las hipótesis importantes son

$$H_0\text{: } \mu = 2.0 \text{ versus } H_a\text{: } \mu > 2.0.$$

Debido a que el tamaño de la muestra es razonablemente grande se puede utilizar el teorema del límite central, según el cual la media muestral $\overline{X}$ tiene aproximadamente una distribución normal. Además, la variable estandarizada $Z = (\overline{X} - \mu)/(S/\sqrt{n})$ tiene aproximadamente una distribución estándar normal (la curva z). El estadístico de prueba resultante de la estandarización de $\overline{X}$, suponiendo que H_0 es verdadera es:

$$\text{Estadístico de prueba: } Z = \frac{\overline{X} - 2.0}{S/\sqrt{n}}$$

Al sustituir $n = 51$, $\overline{x} = 2.06$ y $s = 0.141$ se obtiene $z = 0.06/0.0197 = 3.04$. Aquí la media muestral es aproximadamente tres errores estándar (estimada) más grande de lo que se esperaría si H_0 fuese verdadera (no parece ser mucho más grande de 2, pero hay sólo una pequeña variabilidad en las 51 observaciones de la muestra).

Cualquier valor de $\overline{x}$ mayor que 2.06 es más contradictorio a H_0 que 2.06 mismo, y los valores de $\overline{x}$ que superan a 2.06 corresponden a valores de z que superan a 3.04. Por lo que cualquier $z \geq 3.04$ es al menos contradictorio a H_0. Puesto que el estadístico de prueba tiene aproximadamente una distribución normal estándar, cuando H_0 es verdadera, se tiene

$$\begin{aligned}\text{valor } P &\approx P(\text{una variable aleatoria normal estándar es} \geq 3.04) = 1 - \Phi(3.04)\\ &= 1 - 0.9988 = 0.0012\end{aligned}$$

Como el valor $P = 0.0012 \leq 0.01 = \alpha$, la hipótesis nula debe ser rechazada en el nivel de significancia elegido. Se tiene que la masa de zinc promedio verdadera en efecto es mayor que 2. ■

Errores en la prueba de hipótesis

La base para elegir un nivel de significancia particular α radica en considerar los errores que se podrían presentar al sacar una conclusión. Recuérdese la situación judicial en la cual la hipótesis nula es que el individuo acusado de cometer un delito es inocente. Al emitir un veredicto, el jurado debe considerar la posibilidad de cometer uno de dos tipos diferentes de errores. Uno de ellos implica condenar a un inocente, y el otro consiste en dejar libre a un culpable. Del mismo modo hay dos tipos diferentes de errores que pueden hacerse durante un análisis de prueba de hipótesis estadística.

DEFINICIÓN

Un **error de tipo I** consiste en rechazar la hipótesis nula H_0 cuando es verdadera.

Un **error de tipo II** implica no rechazar H_0 cuando es falsa.

Por ejemplo, un fabricante de cereal afirma que una porción de una de sus marcas aporta 100 calorías (contenido calórico que suele determinarse mediante un método de pruebas destructivo, pero el requisito de presentar información nutricional en los paquetes ha dado lugar a técnicas más sencillas). Por supuesto, el contenido real de calorías varía un poco de porción a porción (de tamaño específico), así que 100 se debe interpretar como un promedio. Sería molesto para los consumidores de este cereal si el contenido de calorías promedio real excediera el valor que se informa. Por tanto, una adecuada formulación de hipótesis es probar H_0: $\mu = 100$ contra H_a: $\mu > 100$. La hipótesis alternativa afirma que los consumidores ingieren en promedio mayor cantidad de calorías de lo que afirma la empresa. Aquí un error tipo I consiste en rechazar la afirmación del fabricante en el sentido de que $\mu = 100$ cuando realmente es verdadera. Un error tipo II es no rechazar la afirmación del fabricante cuando es cierto que $\mu > 100$.

Suponga que μ_1 y μ_2 representan la vida promedio verdadera de dos marcas diferentes de plumas *rollerball* en condiciones experimentales controladas (utilizando una máquina que escribe continuamente hasta que una pluma falla). Es natural probar las hipótesis H_0: $\mu_1 - \mu_2 = 0$ (es decir, $\mu_1 = \mu_2$) contra H_a: $\mu_1 - \mu_2 \neq 0$ (es decir, $\mu_1 \neq \mu_2$). Un error tipo I sería concluir que las vidas medias verdaderas son diferentes cuando en realidad son idénticas. Un error tipo II consiste en decidir que las vidas promedio verdaderas pueden ser iguales cuando en efecto son en realidad diferentes entre sí.

En el mejor de los mundos posibles tenemos un sistema judicial que nunca condena a un inocente y nunca deja libre a un culpable. Esta norma de oro para las decisiones judiciales ha demostrado ser extremadamente difícil. Asimismo, nos gustaría encontrar procedimientos de prueba para los cuales sea poco probable que se cometa cualquier tipo de error. Sin embargo, se puede lograr este ideal sólo si la decisión se basa en el examen de toda la población. La dificultad con la utilización de un procedimiento basado en datos muestrales es que, debido a la variabilidad del muestreo, el resultado podría ser una muestra no representativa. En la situación del contenido de calorías, incluso si la afirmación del fabricante es correcta, un valor de $\overline{X}$ inusualmente grande puede dar como resultado un valor P menor que el nivel de significancia elegido y, por consiguiente, cometer un error tipo I. Por otra parte, el contenido de calorías promedio verdadero puede superar las afirmaciones del fabricante, sin embargo, una muestra de porciones puede producir un valor P relativamente grande para el cual la hipótesis nula no se puede rechazar.

En lugar de demandar procedimientos sin errores habrá que buscar procedimientos con los cuales sea improbable que ocurra cualquier tipo de error. Es decir, un buen procedimiento es aquel en que la probabilidad de cometer un error tipo I es pequeña y la probabilidad de cometer un error tipo II es pequeña.

EJEMPLO 7.4 Se sabe que cierto tipo de automóvil no sufre daños visibles 25% del tiempo en pruebas de choque a 10 mph. Se ha propuesto un diseño de parachoques modificado en un esfuerzo por incrementar este porcentaje. Sea p la proporción de todos los choques a 10 mph con este nuevo parachoques en los que no se producen daños visibles. Las hipótesis a probar son H_0: $p = 0.25$ (ninguna mejora) versus H_a: $p > 0.25$. La prueba se basará en un experimento que implica $n = 20$ choques independientes con prototipos del nuevo diseño. Aquí el estadístico de prueba natural es $X =$ el número de choques sin daño visible. Si H_0 es verdadera, $E(X) = np_0 = (20)(0.25) = 5$. La intuición sugiere que un valor observado x mucho mayor que esto proporcionaría una fuerte evidencia contra H_0 y apoyaría a H_a.

Considere utilizar un nivel de significancia de 0.10. El valor P es $P(X \geq x$ cuando X tiene una distribución binomial con $n = 20$ y $p = 0.25) = 1 - B(x - 1; 20, 0.25)$ para $x > 0$.

La tabla A.1 del apéndice muestra que, en este caso,

$$P(X \geq 7) = 1 - B(6; 20, 0.25) = 1 - 0.786 = 0.214$$

$$P(X \geq 8) = 1 - 0.898 = 0.102 \approx 0.10, P(X \geq 9) = 1 - 0.959 = 0.041$$

Así, rechazar H_0 cuando el valor $P \leq 0.10$ es equivalente a rechazar H_0 cuando $X \geq 8$. Por tanto,

$$\begin{aligned} P(\text{cometer un error tipo I}) &= P(\text{rechazar } H_0 \text{ cuando } H_0 \text{ es verdadera}) \\ &= P(\text{cuando X tiene una distribución binomial con } n = 20 \text{ y } p = 0.25) \\ &= 0.102 \\ &\approx 0.10 \end{aligned}$$

Es decir, *la probabilidad de un error tipo I es exactamente el nivel de significancia* α. Si la hipótesis nula es verdadera aquí y el procedimiento de prueba se utiliza una y otra vez, cada vez junto con un grupo de 20 accidentes, a largo plazo la hipótesis nula será incorrectamente rechazada en favor de la hipótesis alternativa, aproximadamente 10% de las veces. Por tanto, este procedimiento de prueba ofrece buena protección ante la posibilidad de cometer un error tipo I.

Hay solamente una probabilidad de error tipo I porque hay un solo valor del parámetro para el cual H_0 es verdadera (éste es uno de los beneficios de simplificar la hipótesis nula a una expresión de igualdad). Sea β la probabilidad de cometer un error tipo II. Lamentablemente no existe un único valor de β, ya que hay una multitud de maneras para las que H_0 podría ser falsa ya que $p = 0.30$, ya que $p = 0.37$, ya que $p = 0.5$ etcétera. De hecho hay un valor diferente de β para cada valor diferente de p superior a 0.25. En el nivel de significancia elegido 0.10, H_0 se rechazará si y sólo si $X \geq 8$, por lo que no se rechazará H_0 si y sólo si $X \leq 7$. Por tanto,

$$\begin{aligned} \beta.(3) &= P(\text{error tipo II cuando } p = 0.3) \\ &= P(H_0 \text{ no se rechaza cuando } p = 0.3) \\ &= P[X \leq 7 \text{ cuando } X \sim \text{Bin}(20, 0.3)] \\ &= B(7; 20, 0.3) = 0.772 \end{aligned}$$

Cuando p es en realidad 0.3 y no 0.25 (un "pequeño" alejamiento de H_0), ¡aproximadamente 77% de todos los experimentos de este tipo daría como resultado que H_0 no sea rechazada de manera incorrecta!

La siguiente tabla muestra β para los valores seleccionados de p (cada uno calculado como lo acabamos de hacer $\beta(0.3)$). Claramente, β disminuye conforme se aleja el valor de p hacia la derecha del valor nulo 0.25. De manera intuitiva, mientras más grande es el alejamiento de H_0, es más probable que se detecte dicho alejamiento.

p	0.3	0.4	0.5	0.6	0.7	0.8
$\beta(p)$	0.772	0.416	0.132	0.021	0.001	0.000

La probabilidad de cometer aquí un error tipo II es bastante grande cuando $p = 0.3$ o 0.4. Esto es porque esos valores están bastante cerca de lo que afirma H_0 y el tamaño de muestra 20 es demasiado pequeño para permitir la discriminación precisa entre 0.25 y los valores de p.

El procedimiento de prueba propuesto sigue siendo razonable para poner a prueba la hipótesis nula más realista de que $p \leq 0.25$. En este caso, ya no existe una sola probabilidad α de error tipo I, sino que hay una α por cada p que, cuando mucho, sea de 0.25: $\alpha(0.25)$, $\alpha(0.23)$, $\alpha(0.20)$, $\alpha(0.15)$ etcétera. No obstante, es fácil verificar que $\alpha(p) < \alpha(0.25) = 0.102$ si $p < 0.25$. Es decir, el valor más grande de α ocurre con el valor límite 0.25 entre H_0 y H_a. Por consiguiente, si α es pequeña para la hipótesis nula simplificada, también será igual o más pequeña para la H_0 más realista. ■

EJEMPLO 7.5 Se sabe que el tiempo de secado de un tipo de pintura en condiciones de prueba especificadas está normalmente distribuido con una media de 75 min y desviación estándar de 9 min. Algunos químicos propusieron un nuevo aditivo para reducir el promedio del tiempo de secado. Se cree que los tiempos de secado con este aditivo permanecerán distribuidos en forma normal con $\sigma = 9$. Debido al gasto asociado con el aditivo, la evidencia deberá

sugerir fuertemente una mejora en el tiempo de secado promedio antes de que se adopte semejante conclusión. Sea μ el tiempo de secado promedio verdadero cuando se utiliza el aditivo. Las hipótesis apropiadas son H_0: $\mu = 75$ versus H_a: $\mu < 75$. Sólo si H_0 puede ser rechazada el aditivo será declarado exitoso y utilizado.

Los datos experimentales tienen que estar compuestos de tiempos de secado de $n = 25$ especímenes de prueba. Sean $X_1, \ldots, X_{25}$ los 25 tiempos de secado, una muestra aleatoria de tamaño 25 de una distribución normal con media μ y desviación estándar $\sigma = 9$ (aunque elevar en la práctica un valor conocido de σ generalmente es poco realista, ello simplifica considerablemente el cálculo de probabilidades de error tipo II). La media muestral $\overline{X}$ del tiempo de secado tiene entonces una distribución normal con valor esperado $\mu_{\overline{X}} = \mu$ y desviación estándar $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 9/\sqrt{25} = 1.8$. Cuando H_0 es verdadera esperamos que $\overline{X}$ sea 75; una media muestral mucho menor contradiría fuertemente a H_0 y apoyaría a H_a.

Nuestro estadístico de prueba será $\overline{X}$ estandarizado, suponiendo que H_0 es verdadera:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - 75}{1.8}$$

La distribución muestral de $\overline{X}$ es normal porque la distribución de la población es normal, lo que implica que Z tiene una distribución normal estándar cuando H_0 es verdadera (en contraste con los ejemplos 7.2 y 7.3, aquí estamos suponiendo y utilizando un valor conocido de σ).

Considere realizar la prueba utilizando un nivel de significancia de 0.01, es decir, H_0 se rechaza si el valor $P \leq 01$. Para un valor dado $\overline{x}$ de la media muestral y el correspondiente valor calculado z, la forma de la hipótesis alternativa implica que los valores que más contradicen H_0 son valores inferiores a $\overline{x}$ y, en consecuencia, los valores del estadístico de prueba que son menores de z. Por tanto, el valor P es

$$\begin{aligned}\text{valor } P = {} & P(\text{la obtención de un valor de } Z \text{ al menos tan contradictorio} \\ & \text{a } H_0 \text{ como } z \text{ cuando } H_0 \text{ es verdadera}) \\ = {} & P(Z \leq z \text{ cuando } H_0 \text{ es verdadera}) \\ = {} & \text{área bajo la curva normal estándar a la izquierda de } z \\ = {} & \Phi(z)\end{aligned}$$

Así, el valor P será igual a 0.01 cuando z captura el área de cola inferior 0.01 bajo la curva z. De la tabla A.3 del apéndice se tiene que esto ocurre cuando $z = -2.33$ [verificar que $\Phi(-2.33) = 0.01$]. Como se muestra en la figura 7.2 el valor P, por tanto, será a lo más 0.01 cuando $z \leq -2.33$. Esto a su vez implica que

$$\begin{aligned}P(\text{error tipo I}) = {} & P(\text{se rechaza } H_0 \text{ cuando } H_0 \text{ es verdadera}) \\ = {} & P(\text{valor } P < 0.01 \text{ cuando } H_0 \text{ es verdadera}) \\ = {} & P(Z \leq -2.33 \text{ cuando } Z \text{ tiene una distribución estándar normal}) \\ = {} & 0.01\end{aligned}$$

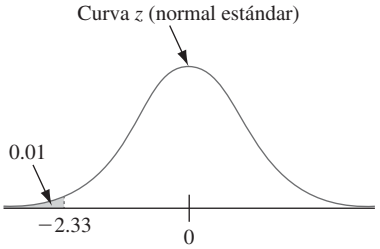


Figura 7.2 Valor $P \leq 0.01$ si y sólo si $z \leq -2.33$

Como en el ejemplo anterior, el nivel de significancia elegido α es, de hecho, la probabilidad de cometer un error tipo I. Si el procedimiento de prueba anterior [estadístico de prueba Z, rechaza H_0 si el valor $P \leq 0.01$] es utilizado en varias ocasiones, muestra tras muestra, a largo plazo la hipótesis nula será incorrectamente rechazada sólo 1% de las veces. Nuestra

propuesta de procedimiento de prueba ofrece una excelente protección ante la posibilidad de cometer un error tipo I. Observe que si se considera la hipótesis nula más realista H_0: $\mu \geq 75$, se puede demostrar que P(error tipo I) ≤ 0.01; el máximo se produce en el valor nulo 75, que es el límite entre H_0 y H_a.

El cálculo de P(error tipo I) en este ejemplo se basó en el hecho de que valor $P \leq 0.01$ es equivalente a $Z = (\overline{X} - 75)/1.8 \leq -2.33$. Al multiplicar ambos lados de esta última desigualdad por 1.8 y luego sumarle 75 a ambos lados se obtiene $\overline{X} \leq 70.8$. Por tanto, rechazar H_0 en el nivel de significancia 0.01 [si el valor $P \leq 0.01$] es equivalente a rechazar H_0 si $\overline{X} \leq 70.8$; H_0 no se rechazará si $\overline{X} > 70.8$. La probabilidad de cometer un error tipo II cuando $\mu = 72$ es ahora

$$\begin{aligned}\beta(72) &= P(\text{no se rechaza } H_0 \text{ cuando } \mu = 72)\\ &= P(\overline{X} > 70.8 \text{ cuando } \overline{X} \sim \text{normal con } \mu_{\overline{X}} = 72,\ \sigma_{\overline{X}} = 1.8)\\ &= 1 - \Phi[(70.8 - 72)/1.8] = 1 - \Phi(-0.67) = 1 - 0.2514 = 0.7486\end{aligned}$$

Esta es una probabilidad de error muy grande. Si la prueba con $\alpha = 0.01$ se utiliza repetidamente, muestra tras muestra, y el valor real de μ es 72, en casi el 75% de los casos no se rechazará la hipótesis nula. La dificultad es que 72 está demasiado cerca del valor nulo para una prueba con este tamaño de muestra y el valor α tiene una buena oportunidad de detectar este alejamiento de H_0.

Cálculos similares dan

$$\beta(70) = 1 - \Phi\,[(70.8 - 70)/1.8] = 0.3300,\ \beta(67) = 0.0174$$

Estas probabilidades de error tipo II son mucho menores que $\beta(72)$, ya que 70 y 67 están más alejados del valor nulo que 72. La figura 7.3 ilustra α y las primeras probabilidades de error tipo II.

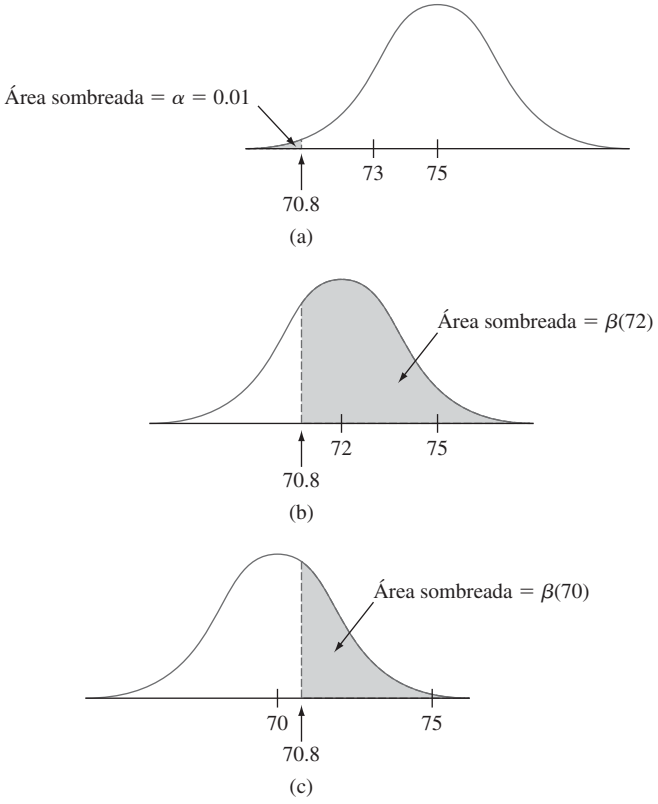


Figura 7.3 α y β ilustradas para el ejemplo 7.5: (a) la distribución de $\overline{X}$ cuando $\mu = 75$ (H_0 verdadera); (b) la distribución de $\overline{X}$ cuando $\mu = 72$ (H_0 falsa); (c) la distribución de $\overline{X}$ cuando $\mu = 70$ (H_0 falsa)

Los investigadores podrían considerar $\mu = 72$ como un importante alejamiento de la hipótesis nula, en cuyo caso $\beta(72) = 0.7486$ es intolerablemente grande. Considere cambiar el nivel de significancia (probabilidad de error tipo I de 0.01 a 0.05; es decir, ahora se propone rechazar H_0 si el valor $P \leq 0.05$. La tabla A.3 del apéndice muestra que el valor crítico de z, -1.645 captura un área de curva z de cola inferior de 0.05. Usando el mismo razonamiento que se ha aplicado antes cuando $\alpha = 0.01$, rechazar a H_0 cuando el valor $P \leq 0.05$ equivale a rechazar cuando $Z \leq -1.645$. Esto a su vez es equivalente a rechazar cuando $\overline{X} \leq 72$ (observe que aumentó el nivel de significancia y se hizo para facilitar que la hipótesis nula sea rechazada). Procediendo como en los cálculos anteriores, encontramos que

$$\beta(72) = 0.5, \quad \beta(70) = 0.1335, \quad \beta(67) = 0.0027$$

Todas estas probabilidades de error tipo II son más pequeñas que sus contrapartes para la prueba con $\alpha = 0.01$. Aquí el mensaje importante es que si se puede tolerar un nivel de significancia más grande (probabilidad de error tipo I), entonces la prueba resultante tendrá mejor capacidad de detección cuando la hipótesis nula es falsa. ■

No es casualidad que en los dos ejemplos anteriores, el nivel de significancia α resulte ser la probabilidad de un error tipo I.

PROPOSICIÓN

El procedimiento de prueba rechaza H_0 si el valor $P \leq \alpha$ y en caso contrario no rechaza H_0 si $P(\text{error tipo I}) = \alpha$. Es decir, el nivel de significancia empleado en el procedimiento de prueba es la probabilidad de que se cometa un error tipo I.

Al final de la sección se esboza una prueba parcial de esta proposición.

La relación inversa entre los niveles de significancia α y las probabilidades de error tipo II en el ejemplo 7.5 se pueden generalizar de la siguiente manera:

PROPOSICIÓN

Supóngase que se selecciona un experimento o procedimiento de muestreo, se especifica un tamaño de muestra y se elige un estadístico de prueba. Luego se aumenta el nivel de significancia α; es decir, al emplear una probabilidad más grande de error tipo I se obtiene un valor menor de β para cualquier valor particular del parámetro consistente con H_a.

Este resultado es intuitivamente obvio ya que cuando se aumenta α, es más probable que se tenga el valor $P \leq \alpha$ y, por tanto, es menos probable que el valor $P > \alpha$.

La proposición implica que una vez que se ha fijado el estadístico de prueba y n, no es posible hacer que α y cualquier valor de β que podrían ser de interés fueran arbitrariamente pequeños. Decidir sobre un nivel de significancia apropiado implica tener α y β pequeños. En el ejemplo 7.5 la probabilidad de error tipo II para una prueba con $\alpha = 0.01$ era absolutamente grande para un valor de μ cercano al valor de H_0. Una estrategia que a veces se utiliza en la práctica (pero quizás no lo suficiente) es especificar α y β para algún valor alternativo del parámetro que es de particular importancia para el investigador. Entonces se puede determinar el tamaño de muestra n que satisfaga estas dos condiciones. Por ejemplo, el artículo **"Cognitive Treatment of Illness Perceptions in Patients with Chronic Low Back Pain: A Randomized Controlled Trial" (*Physical Therapy*, 2013: 435–438)** contiene la siguiente frase: "Una disminución entre 18 y 24 en el puntaje del PSC (cuestionario de quejas especificas del paciente) se determinó como un cambio clínicamente relevante en pacientes con dolor de espalda baja. El tamaño de muestra se calculó con un cambio mínimo de 18, una α bilateral de 0.05, un $1 - \beta$ de 0.90

y una desviación estándar de 26.01... El cálculo de tamaño simple obtenido en un total de 135 participantes". Considere esas determinaciones de tamaño simple para las siguientes secciones y capítulos.

En la práctica, generalmente las hipótesis de interés se pueden formular de manera que un error tipo I sea más grave que un error tipo II. El método usado por la mayoría de los estadísticos es reflexionar acerca de la gravedad relativa de un error tipo I comparado con un error tipo II y luego usar el mayor valor de α que se pueda tolerar. Esto equivale a hacer lo mejor que se puede respecto a la probabilidad de error tipo II, asegurándose de que la probabilidad de error tipo I es suficientemente pequeña. Por ejemplo, si $\alpha = 0.05$ es el mayor nivel de significancia que se puede tolerar, sería mejor utilizarlo en vez de utilizar $\alpha = 0.01$, ya que todas las β para la última α serán más pequeñas que para esta. Como se mencionó anteriormente, los niveles de significancia más frecuentemente empleados son $\alpha = 0.05$, 0.01 y 0.001, 0.10. Sin embargo, hay excepciones. Por ejemplo en física de partículas: de acuerdo con el artículo **"Discovery or Fluke: Statistics in Particle Physics" (*Physics Today*, julio de 2012: 45–50)**, "la elección habitual de alfa es 3×10^{-7}, correspondiente a 5σ de una distribución gaussiana [es decir, normal] de H_0. […] ¿Por qué tan estrictos? Por un lado, la historia reciente ofrece muchos ejemplos que deben tratarse con precaución para las excitantes señales 3σ y 4σ que desaparecen cuando se obtienen más datos".

Si la distribución del estadístico de prueba es continua (por ejemplo, si el estadístico de prueba tiene la distribución normal estándar o una distribución particular t cuando H_0 es verdadera), entonces se puede emplear cualquier nivel de significancia α entre 0 y 1; por ejemplo, rechazar H_0 si el valor $P \leq 0.035$. Sin embargo, este no es necesariamente el caso si la distribución del estadístico de prueba es discreta. Como ejemplo, considere de nuevo el caso del diseño del parachoques del ejemplo 7.4 en el cual las hipótesis de interés fueron H_0: $p = 0.25$ versus H_a: $p > 0.25$. El estadístico de prueba X tenía una distribución binomial y

$$\text{valor } P = P(X \geq x \text{ cuando } n = 20 \text{ y } p = 0.25)$$

La tabla A.1 del apéndice muestra que en este caso $P(X \geq 8) = 0.102$ y $P(X \geq 9) = 0.041$. Por tanto, si se quiere que el nivel de significancia sea 0.05, el nivel más cercano posible es 0.041: H_0 se rechaza si el valor $P \leq 0.041$.

EJERCICIOS Sección 7.1 (1–14)

1. Por cada una de las siguientes aseveraciones exprese si es una hipótesis estadística legítima y por qué:

a. $H: \sigma > 100$ **b.** $H: \tilde{x} = 45$

c. $H: s \leq 0.20$ **d.** $H: \sigma_1/\sigma_2 < 1$

e. $H: \bar{X} - \bar{Y} = 5$

f. $H: \lambda \leq 0.01$, donde λ es el parámetro de una distribución exponencial utilizada para modelar la vida útil de un componente

2. Para los siguientes pares de aseveraciones, indique cuáles no satisfacen las reglas para establecer hipótesis y por qué (los subíndices 1 y 2 diferencian entre las cantidades para dos poblaciones o muestras diferentes).

a. $H_0: \mu = 100, H_a: \mu > 100$

b. $H_0: \sigma = 20, H_a: \sigma \leq 20$

c. $H_0: p \neq 0.25, H_a: p = 0.25$

d. $H_0: \mu_1 - \mu_2 = 25, H_a: \mu_1 - \mu_2 > 100$

e. $H_0: S_1^2 = S_2^2, H_a: S_1^2 \neq S_2^2$

f. $H_0: \mu = 120, H_a: \mu = 150$

g. $H_0: \sigma_1/\sigma_2 = 1, H_a: \sigma_1/\sigma_2 \neq 1$

h. $H_0: p_1 - p_2 = -0.1, H_a: p_1 - p_2 < -0.1$

3. ¿Para qué valores p dados sería rechazada la hipótesis nula al realizar una prueba de nivel 0.05?

a. 0.001 **b.** 0.021 **c.** 0.078

d. 0.047 **e.** 0.148

4. Se dan pares de valores P y niveles de significancia, α. Para cada par, indique si el valor P observado llevaría al rechazo de H_0 en el nivel de significancia dado.

a. *Valor* $P = 0.084$, $\alpha = 0.05$

b. *Valor* $P = 0.003$, $\alpha = 0.001$

c. *Valor* $P = 0.498$, $\alpha = 0.05$

d. *Valor* $P = 0.084$, $\alpha = 0.10$

e. *Valor P* = 0.039, $\alpha = 0.01$

f. *Valor P* = 0.218, $\alpha = 0.10$

5. Para determinar si las soldaduras de las tuberías en una planta de energía nuclear satisfacen las especificaciones, se selecciona una muestra aleatoria de soldaduras y se realizan pruebas en cada una de ellas. La resistencia de la soldadura se mide como la fuerza requerida para romperla. Suponga que las especificaciones indican que la resistencia media de las soldaduras debe exceder de 100 lb/pulg^2; el equipo de inspección decide probar H_0: $\mu = 100$ versus H_a: $\mu > 100$. Explique por qué podría ser preferible utilizar esta H_a en lugar de $\mu < 100$.

6. Sea μ el nivel de radiactividad promedio verdadero (*picocuries* por litro). Se considera que el valor 5 pCi/L es la línea divisoria entre agua segura e insegura. ¿Recomendaría probar H_0: $\mu = 5$ versus H_a: $\mu > 5$ o probar H_0: $\mu = 5$ versus H_a: $\mu < 5$? Explique su razonamiento [*Sugerencia*: Piense en las consecuencias de un error tipo I o un error tipo II con cada posibilidad.]

7. Antes de aprobar la compra de un gran pedido de fundas de polietileno para un tipo particular de cable de energía submarino relleno de aceite a alta presión, una compañía desea contar con evidencia concluyente de que la desviación estándar verdadera del espesor de la funda es de menos de 0.05 mm. ¿Qué hipótesis deberán ser probadas y por qué? En este contexto, ¿cuáles son los errores tipo I y tipo II?

8. Muchas casas viejas cuentan con sistemas eléctricos que utilizan fusibles en lugar de interruptores de circuito. Un fabricante de fusibles de 40 amp desea asegurarse de que el amperaje medio al cual se queman sus fusibles es en realidad de 40. Si el amperaje medio es menor que 40 los clientes se quejarán porque los fusibles tienen que ser reemplazados con demasiada frecuencia. Si el amperaje medio es de más de 40, el fabricante podría ser responsable de los daños que sufra un sistema eléctrico a causa del funcionamiento defectuoso de los fusibles. Para verificar el amperaje de los fusibles se selecciona e inspecciona una muestra de fusibles. Si tuviera que realizarse una prueba de hipótesis con los datos resultantes, ¿qué hipótesis nula y qué hipótesis alternativa serían de interés para el fabricante? Describa los errores tipos I y tipo II en el contexto de este problema.

9. Se toman muestras de agua utilizada para enfriamiento al momento de ser descargada por una planta de energía en un río. Se ha determinado que en tanto la temperatura media del agua descargada sea cuando mucho de 150°F no habrá efectos negativos en el ecosistema del río. Para investigar si la planta cumple con los reglamentos que prohíben una temperatura media por encima de 150° del agua de descarga, se tomarán al azar 50 muestras de agua y se registrará la temperatura de cada una. Los datos resultantes se utilizarán para probar la hipótesis H_0: $\mu = 150°$ versus H_a: $\mu > 150°$. En el contexto de esta situación, describa los errores de tipo I y de tipo II. ¿Qué tipo de error consideraría de mayor importancia? Explique.

10. Un tipo regular de laminado está siendo utilizado por un fabricante de tarjetas de circuito. Un laminado especial ha sido desarrollado para reducir la combadura. El laminado regular será utilizado en una muestra de especímenes y el laminado especial en otra muestra y se determinará entonces la cantidad de combadura en cada espécimen. Entonces el fabricante cambiará al laminado especial sólo si puede demostrar que la cantidad de combadura promedio verdadera de dicho laminado es menor que la del laminado regular. Formule las hipótesis pertinentes y describa los errores de tipo I y de tipo II en el contexto de esta situación.

11. Dos compañías diferentes han solicitado proporcionar el servicio de televisión por cable en una región. Sea p la proporción de todos los suscriptores potenciales que favorecen a la primera compañía sobre la segunda. Considere probar H_0: $p = 0.5$ contra H_a: $p \neq 0.5$ basado en una muestra aleatoria de 25 individuos. Sea el estadístico de la prueba X el número de suscriptores en la muestra que favorecen a la primera compañía y x el valor observado de X.

a. En el contexto de este problema describa cuáles son los errores de tipo I y de tipo II.

b. Suponga que $x = 6$. ¿Cuáles son los valores de X al menos tan contradictorios a H_0 como este?

c. ¿Cuál es la distribución de probabilidad del estadístico de prueba X cuando H_0 es verdadera? Úsela para calcular el valor P cuando $x = 6$.

d. Si H_0 debe rechazarse cuando el valor $P \leq 0.044$, calcule la probabilidad de un error tipo II cuando $p = 0.4$, otra vez cuando $p = 0.3$ y también cuando $p = 0.6$ y $p = 0.7$. [*Sugerencia*: ¿A qué desigualdades que implican x es equivalente que el valor $P > 0.044$? (véase el ejemplo 7.4)]

e. ¿Concluiría, mediante el procedimiento de prueba del inciso d), que 6 de los 25 suscriptores contratarían a la compañía favorecida 1?

12. Una mezcla de cenizas combustibles pulverizadas y cemento Portland utilizada para rellenar con lechada deberá tener una resistencia a la compresión de más de 1300 kN/m^2. La mezcla no será utilizada a menos que la evidencia experimental indique concluyentemente que la especificación de resistencia ha sido satisfecha. Suponga que la resistencia a la compresión de especímenes de esta muestra está normalmente distribuida con $\sigma = 60$. Sea μ la resistencia a la compresión promedio verdadera.

a. ¿Cuáles son la hipótesis nula y la hipótesis alternativa apropiadas?

b. Sea $\overline{X}$ la resistencia a la compresión promedio muestral de $n = 10$ especímenes seleccionados al azar. Considere el procedimiento de prueba con el mismo estadístico de prueba $\overline{X}$ (no estandarizado). Si $\overline{x} = 1340$, ¿se debe rechazar H_0 usando un nivel de significancia de 0.01? [*Sugerencia*: ¿Cuál es la distribución de probabilidad del estadístico de prueba cuando H_0 es verdadera?]

c. ¿Cuál es la distribución de probabilidad del estadístico de prueba cuando $\mu = 1350$? Para una prueba con $\alpha = 0.01$, ¿cuál es la probabilidad de que la mezcla será juzgada insatisfactoria cuando en realidad $\mu = 1350$ (un error tipo II)?

13. La calibración de una báscula tiene que ser verificada pesando 25 veces un espécimen de prueba de 10 kg. Suponga que los resultados de diferentes pesadas son independientes entre sí y que el peso en cada ensayo está normalmente distribuido con

$\sigma = 0.200$ kg. Sea μ la lectura de peso promedio verdadero en la báscula.

a. ¿Qué hipótesis deberá ponerse a prueba?

b. Con la media muestral μ como el estadístico de prueba, ¿cuál es el valor de P cuando $\bar{x} = 9.85$ y qué concluiría en el nivel de significancia 0.01?

c. Para una prueba con $\alpha = 0.01$, ¿cuál es la probabilidad de que la recalibración se considere innecesaria cuando en realidad $\mu = 10.1$? ¿Y cuando $\mu = 9.8$?

14. Se ha propuesto un nuevo diseño del sistema de frenos de un tipo de vehículo. Para el sistema actual, se sabe que la distancia de frenado promedio verdadera a 40 mph en condiciones específicas es de 120 pies. Se propone que el nuevo diseño sea instalado sólo si los datos muestrales indican fuertemente una reducción de la distancia de frenado promedio verdadera del nuevo diseño.

a. Defina el parámetro de interés y formule las hipótesis pertinentes.

b. Suponga que la distancia de frenado del nuevo sistema está normalmente distribuida con $\sigma = 10$. Sea $\overline{X}$ la distancia de frenado promedio de una muestra aleatoria de 36 observaciones. Cuales valores de $\bar{x}$ son más contradictorios a H_0 que 117.2, ¿cuál es el valor de P en este caso, y qué conclusión es apropiada si $\alpha = 0.10$?

c. Cuál es la probabilidad de que el nuevo diseño no sea implementado cuando su verdadera distancia media es en realidad 115 pies y se utiliza la prueba del inciso b)?

7.2 Pruebas de hipótesis *z* sobre una media de población

Recordemos de la sección anterior que se llegó a una conclusión en un análisis de prueba de hipótesis procediendo de la siguiente manera:

i. Calcular el valor de un estadístico de prueba apropiado.

ii. Luego determinar el valor P, la probabilidad, calculada suponiendo que la hipótesis nula H_0 es verdadera, observando un valor del estadístico de prueba al menos tan contradictorio a H_0 como el que resulta de los datos disponibles.

iii. Rechazar la hipótesis nula si el valor $P \leq \alpha$, donde α es el nivel de significancia especificado o solicitado, es decir, la probabilidad de que un error tipo I (rechazar H_0 cuando es verdadera); si el valor $P > \alpha$ no hay suficiente evidencia para justificar el rechazo de H_0 (esto todavía se considera factible).

La determinación del valor P depende de la distribución del estadístico de prueba cuando H_0 es verdadera. En esta sección describimos pruebas z para comprobación de hipótesis sobre una media μ de población. "Prueba z", significa que el estadístico de prueba tiene al menos aproximadamente una distribución normal estándar cuando H_0 es verdadera. Entonces el valor P será un área de la curva z que depende de si la desigualdad en H_a es $>$, $<$ o $\neq$.

En el desarrollo de intervalos de confianza para μ en el capítulo 6, primero se considera el caso en que la distribución de la población es normal con σ conocida, después se relaja a la normalidad y con σ supuestamente conocida, cuando el tamaño de muestra n es grande y finalmente se describe el intervalo de confianza de la muestra t de la media de una población normal. En esta sección estudiamos los dos primeros casos y luego se presenta la prueba t de una muestra en la sección 7.3.

Una distribución de población normal con σ conocida

Aun cuando la suposición de que el valor de σ es conocido rara vez se cumple en la práctica; este caso proporciona un buen punto de partida debido a la facilidad con la que se pueden desarrollar los procedimientos generales y sus propiedades. La hipótesis nula en los tres casos propondrá que μ tiene un valor numérico particular, el *valor nulo*, que será denotado por μ_0. Sean $X_1, \ldots, X_n$ una muestra aleatoria de tamaño n de la población nor-

mal. Entonces la media muestral $\overline{X}$ tiene una distribución normal con valor esperado $\mu_{\overline{X}} = \mu$ y desviación estándar $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Cuando H_0 es verdadera $\mu_{\overline{X}} = \mu_0$. Considere ahora el estadístico Z obtenido estandarizando $\overline{X}$, dada la suposición de que H_0 es verdadera:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Al sustituir la media muestral calculada $\overline{x}$ se obtiene z, la distancia entre $\overline{x}$ y μ_0 expresada en "unidades de desviación estándar". Por ejemplo, si la hipótesis nula es H_0: $\mu = 100$, $\sigma_{\overline{X}} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2.0$ y $\overline{x} = 103$, entonces el valor del estadístico de prueba es $z = (103 - 100)/2.0 = 1.5$. Es decir, el valor observado de $\overline{x}$ es 1.5 desviaciones estándar (de $\overline{X}$) más grande de lo que se espera que sea cuando H_0 es verdadera. El estadístico Z es una medida natural de la distancia entre $\overline{X}$, el estimador de μ, y su valor esperado cuando H_0 es verdadera. Si esta distancia es demasiado grande en una dirección consistente con H_a, hay pruebas sustanciales de que H_0 es falsa.

Suponga primero que la hipótesis alternativa tiene la forma H_a: $\mu > \mu_0$. Entonces un valor de $\overline{x}$ menor que μ_0 indudablemente no apoya a H_0. Tal $\overline{x}$ corresponde a un valor positivo grande de z. A su vez esto implica que cualquier valor calculado que *exceda* a z es más contradictorio a H_0 que z mismo. Se deduce que

$$\text{Valor } P = P(Z \geq z \text{ cuando } H_0 \text{ es verdadera})$$

Ahora, este es el punto clave: cuando H_0 es verdadera el estadístico de prueba Z tiene una distribución normal estándar, porque se creó Z estandarizando a $\overline{X}$ al suponer que H_0 es verdadera (es decir, restando μ_0). La implicación en este caso es que, el valor P es simplemente el área bajo la curva normal estándar a la derecha de z. Debido a esto la prueba es referida como de *cola superior*. Por ejemplo, en el párrafo anterior se calculó $z = 1.5$. Si en la hipótesis alternativa se tiene que H_a: $\mu > 100$, entonces el valor P = área debajo de la curva z a la derecha de $1.5 = 1 - \Phi(1.50) = 0.0668$. En el nivel de significancia 0.05 no podremos rechazar la hipótesis nula porque el valor P excede a α.

Ahora considere una hipótesis alternativa de la forma H_a: $\mu < \mu_0$. En este caso cualquier valor de la media muestral más pequeño que $\overline{x}$ es aún más contradictorio a la hipótesis nula. Así, cualquier valor del estadístico de prueba *más pequeño* que z calculado es más contradictorio a H_0 que z mismo. Por consiguiente,

$$\begin{aligned}\text{Valor } P &= P(Z \leq z \text{ cuando } H_0 \text{ es verdadera}) \\ &= \text{área bajo la curva normal a la izquierda de } z = \Phi(z)\end{aligned}$$

La prueba, en este caso, habitualmente se conoce como de *cola inferior*. Si, por ejemplo, la hipótesis alternativa es H_a: $\mu < 100$ y $z = -2.75$, entonces el valor $P = \Phi(-2.75) = 0.0030$. Esto es lo suficientemente pequeño como para justificar el rechazo de H_0 con un nivel de significancia de 0.05 o 0.01, pero no 0.001.

La tercera alternativa, H_a: $\mu \neq \mu_0$, requiere pensar con un poco más de cuidado. Supongamos, por ejemplo, que el valor nulo es 100 y que $\overline{x} = 103$, lo que resulta en $z = 1.5$. Entonces cualquier valor de $\overline{x}$ superior a 103 es más contradictorio a H_0 que 103. Así cualquier z superior a 1.5 es, asimismo, más contradictorio a H_0 que lo que es 1.5. Sin embargo, 97 es tan contradictorio a la hipótesis nula como lo es 103, ya que hay la misma distancia por debajo de 100 que lo que 103 está arriba de 100. Así $z = -1.5$ es tan contradictorio a H_0 como lo es $z = 1.5$. Por tanto, cualquier z menor que -1.5 es más contradictorio a H_0 que lo que es 1.5 o -1.5. Por consiguiente,

$$\begin{aligned}\text{Valor } P = {} & P(\text{sea } Z \geq 1.5 \text{ o} \leq -1.5 \text{ cuando } H_0 \text{ es verdadera}) \\ = {} & (\text{área bajo la curva } z \text{ a la derecha de } 1.5) \\ & + (\text{área bajo la curva } z \text{ a la izquierda de } -1.5) \\ = {} & 1 - \Phi(1.5) + \Phi(-1.5) = 2[1 - \Phi(1.5)] \\ = {} & 2(0.0668) = 0.1336\end{aligned}$$

Este sería el valor P si $\bar{x} = 97$ lo que da como resultado en $z = -1.5$. El punto importante es que debido a la desigualdad $\neq$ en H_a, el valor P es la suma de un área de cola superior y un área de cola inferior. Por simetría de la distribución normal estándar esto se convierte en dos veces el área capturada en la cola en la que se encuentra z. En forma equivalente, es dos veces el área capturada en la cola superior por $|z|$, es decir, $2[1 - \Phi(|z|)]$. Es natural referirse a esta prueba como de *dos colas*, ya que los valores de z se alejan hacia fuera en cualquier cola de la curva z para rechazar H_0.

El procedimiento de prueba se resume en el siguiente cuadro, y el valor P para cada una de las posibles hipótesis alternativas se ilustra en la figura 7.4.

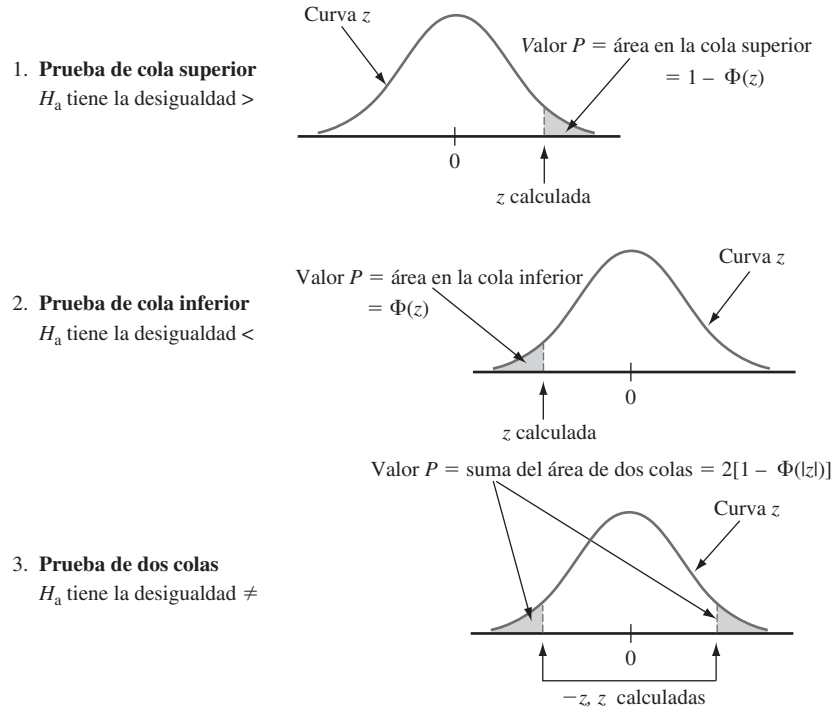


Figura 7.4 Determinación del valor P para una prueba z

Hipótesis nula H_0: $\mu = \mu_0$

Estadístico de prueba: $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$

Hipótesis alternativa	Determinación del valor P
H_a: $\mu > \mu_0$	Área bajo la curva normal estándar a la derecha de z
H_a: $\mu < \mu_0$	Área bajo la curva normal estándar a la izquierda de z
H_a: $\mu \neq \mu_0$	$2 \cdot$ (Área bajo la curva normal estándar a la derecha de $\|z\|$)

Suposiciones: Una distribución de población normal con σ de valor conocido.

Se recomienda utilizar la siguiente secuencia de pasos cuando se prueben hipótesis respecto a un parámetro. La factibilidad de cualquier suposición fundamenta el uso del procedimiento de prueba seleccionado que, por supuesto, se debe comprobar antes de realizar la prueba.

1. Identificar el parámetro de interés y describirlo en el contexto de la situación del problema.
2. Determinar el valor nulo y formular la hipótesis nula.

3. Formular la hipótesis alternativa apropiada
4. Dar la fórmula para el valor calculado del estadístico de prueba (sustituyendo el valor nulo y los valores conocidos de cualesquiera otros parámetros, pero *no* aquellos de cualesquiera cantidades basadas en una muestra).
5. Calcular cualquier cantidad muestral necesaria, sustituir en la fórmula para el valor estadístico de prueba y calcular dicho valor.
6. Determinar el valor P.
7. Comparar el nivel de significancia seleccionado o deseado con el valor P para decidir si H_0 debe ser rechazada y expresar esta conclusión en el contexto del problema.

La formulación de hipótesis (pasos 2 y 3) deberá ser realizada antes de examinar los datos, y se debe elegir el nivel de significancia α antes de determinar el valor P.

EJEMPLO 7.6 Un fabricante de sistemas rociadores utilizados como protección contra incendios en edificios de oficinas afirma que la temperatura de activación del sistema promedio verdadera es de 130°F. Una muestra de $n = 9$ sistemas, cuando se somete a prueba, da una temperatura de activación promedio muestral de 131.08°F. Si la distribución de los tiempos de activación es normal con desviación estándar de 1.5°F, ¿contradicen los datos la afirmación del fabricante a un nivel de significancia $\alpha = 0.01$?

1. Parámetro de interés: μ = temperatura de activación promedio verdadera.
2. Hipótesis nula: H_0: $\mu = 130$ (valor nulo = $\mu_0 = 130$).
3. Hipótesis alternativa: H_a: $\mu \neq 130$ (un alejamiento del valor declarado en *una u otra* dirección es de interés).
4. Valor del estadístico de prueba:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

5. Sustituyendo $n = 9$ y $\bar{x} = 131.08$,

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{0.5} = 2.16$$

Es decir, la media muestral observada está a un poco más de 2 desviaciones estándar arriba del valor que se esperaría si H_0 fuera verdadera.

6. La desigualdad en H_a implica que la prueba es de dos colas, por lo que el valor P da como resultado la duplicación del área capturada de la cola:

$$\text{Valor } P = 2[1 - \Phi(2.16)] = 2(0.0154) = 0.0308$$

7. Ya que el valor $P = 0.0308 > 0.01 = \alpha$, H_0 no se puede rechazar al nivel de significancia 0.01. Los datos no dan fuerte apoyo a la afirmación de que el promedio verdadero difiere del valor de diseño de 130. ■

β y determinación del tamaño de la muestra Las pruebas z con σ conocida se encuentran entre las pocas pruebas en estadística para las cuales existen fórmulas simples disponibles para β, la probabilidad de error de tipo II. Considere en primer lugar la alternativa H_a: $\mu > \mu_0$. La hipótesis nula se rechaza si el valor $P \leq \alpha$, y el valor P es el área bajo la curva normal estándar a la derecha de z. Supóngase que $\alpha = 0.05$. El valor crítico de z que captura un área de cola superior de 0.05 es $z_{0.05} = 1.645$ (busque un área acumulada de 0.95 en la tabla A.3). Así, si el valor del estadístico de prueba calculado z es menor que 1.645, el área a la derecha de z será mayor que 0.05 y entonces la hipótesis nula *no* se rechazará. Ahora, se sustituye $(\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ en lugar de z en la desigualdad, $z < 1.645$ y se maneja algebraicamente para despejar $\bar{x}$ a la izquierda (se multiplican ambos lados por $\sigma/\sqrt{n}$ y

luego se suma μ_0 a ambos lados). Esto da la desigualdad equivalente $\bar{x} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$. Ahora sea μ' un valor particular de μ que excede el valor nulo μ_0. Entonces,

$$\begin{aligned}\beta(\mu') &= P(H_0 \text{ no se rechaza cuando } \mu = \mu') \\ &= P(\bar{X} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n} \text{ cuando } \mu = \mu') \\ &= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}} \text{ cuando } \mu = \mu'\right) \\ &= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

Conforme μ' aumenta, $\mu_0 - \mu'$ se convierte más negativo, por lo que $\beta(\mu')$ será pequeño cuando μ' supere μ_0 (porque el valor en que se evalúa Φ será muy negativo). Las probabilidades de error para las pruebas de menor cola y dos colas se derivan de manera análoga.

Si σ es grande la probabilidad de un error tipo II puede ser grande con un valor alternativo de μ' que sea de interés particular para un investigador. Suponga que se fija σ y que también se especifica β para tal valor alternativo. En el ejemplo de los sistemas rociadores, los funcionarios de la compañía podrían considerar a $\mu' = 132$ como un alejamiento muy sustancial de H_0: $\mu = 130$ y desear, por consiguiente, $\beta(132) = 0.10$ además de $\alpha = 0.01$. Más generalmente, considere las dos restricciones $P(\text{error de tipo I}) = \alpha$ y $\beta(\mu') = \beta$ para α, μ' y β específicas. Entonces para una prueba de cola superior, el tamaño de la muestra n debe ser elegido para satisfacer

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) = \beta$$

Esto implica que

$$-z_\beta = \begin{matrix}\text{valor crítico } z \text{ que captura} \\ \text{el área de cola inferior } \beta\end{matrix} = z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}$$

Es fácil resolver esta ecuación para la n deseada. Un argumento paralelo da el tamaño de muestra necesario para las pruebas de cola inferior y de dos colas, tal como se resume en el siguiente recuadro.

Hipótesis alternativa	Probabilidad de error de tipo II para una prueba de nivel α
H_a: $\mu > \mu_0$	$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu < \mu_0$	$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
H_a: $\mu \neq \mu_0$	$\Phi\left(z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right) - \Phi\left(-z_{\alpha/2} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$

donde $\Phi(z) =$ función de distribución acumulada normal estándar.

El tamaño de la muestra n por el cual una prueba de nivel α también tiene $\beta(\mu') = \beta$ con el valor alternativo μ' es:

$$n = \begin{cases}\left[\dfrac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{prueba para una cola (superior o inferior)} \\ \left[\dfrac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2 & \text{para una prueba de dos colas (una solución aproximada)}\end{cases}$$

EJEMPLO 7.7 Sea μ la vida promedio verdadera de la banda de rodamiento de un cierto tipo de neumático. Considere poner a prueba H_0: $\mu = 30\,000$ versus H_a: $\mu > 30\,000$ basado en un tamaño de muestra $n = 16$ de una distribución de población normal con $\sigma = 1500$. Una prueba con $\alpha = 0.01$ requiere $z_\alpha = z_{0.01} = 2.33$. La probabilidad de cometer un error de tipo II cuando $\mu = 31\,000$ es

$$\beta(31\,000) = \Phi\left(2.33 + \frac{30\,000 - 31\,000}{1500/\sqrt{16}}\right) = \Phi(-0.34) = 0.3669$$

Puesto que $z_{.1} = 1.28$, la exigencia de que el nivel de prueba 0.01 también tenga $\beta(31\,000) = 0.1$ requiere

$$n = \left[\frac{1500(2.33 + 1.28)}{30\,000 - 31\,000}\right]^2 = (-5.42)^2 = 29.32$$

El tamaño de la muestra debe ser un entero, por tanto se deberán utilizar $n = 30$ neumáticos. ■

Pruebas con muestras grandes

Cuando el tamaño de la muestra es grande, las pruebas z anteriores son fáciles de modificar para dar procedimientos de prueba válidos sin requerir una distribución de población normal o una σ conocida. El resultado clave se utilizó en el capítulo 6 para justificar intervalos de confianza para una muestra grande: una n grande implica que la variable estandarizada

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

tiene *aproximadamente* una distribución normal estándar. La sustitución del valor nulo μ_0 en lugar de μ da el estadístico de prueba

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

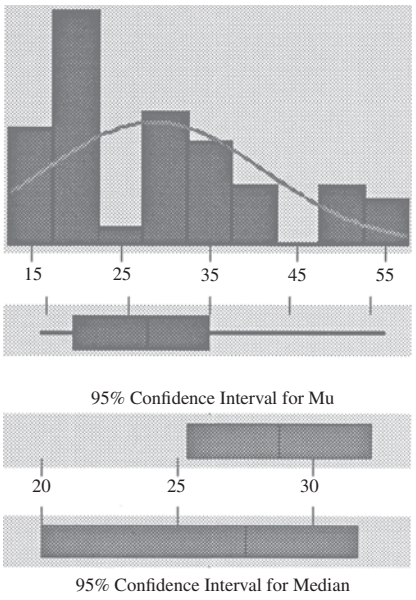
que tiene de manera aproximada una distribución normal estándar cuando H_0 es verdadera. El valor P se determina exactamente como se describió anteriormente en esta sección (por ejemplo, $\Phi(z)$ cuando la hipótesis alternativa H_a: $\mu < \mu_0$). Rechazar H_0 cuando el valor $P \leq \alpha$ da una prueba con nivel de significancia *aproximada* α. La regla empírica $n > 40$ nuevamente se utilizará para caracterizar una muestra de tamaño grande.

EJEMPLO 7.8 Se utiliza un penetrómetro cónico dinámico para medir la resistencia de un material a la penetración (mm/golpe), a medida que el cono es insertado en el pavimento o el subsuelo. Suponga que para una aplicación particular, se requiere que el valor de penetración cónica dinámica promedio verdadera para un cierto tipo de pavimento sea menor que 30. El pavimento no será utilizado a menos que exista evidencia concluyente de que la especificación se satisfizo. Formule y pruebe las hipótesis apropiadas utilizando los siguientes datos **("Probabilistic Model for the Analysis of Dynamic Cone Penetrometer Test Values in Pavement Structure Evaluation", *J. of Testing and Evaluation*, 1999: 7–14):**

14.1	14.5	15.5	16.0	16.0	16.7	16.9	17.1	17.5	17.8
17.8	18.1	18.2	18.3	18.3	19.0	19.2	19.4	20.0	20.0
20.8	20.8	21.0	21.5	23.5	27.5	27.5	28.0	28.3	30.0
30.0	31.6	31.7	31.7	32.5	33.5	33.9	35.0	35.0	35.0
36.7	40.0	40.0	41.3	41.7	47.5	50.0	51.0	51.8	54.4
55.0	57.0								

La figura 7.5 muestra un resumen descriptivo obtenido con Minitab. La penetración cónica dinámica media muestral es menor que 30. Sin embargo, existe una cantidad sustancial de

variación en los datos (coeficiente de variación muestral $= s/\bar{x} = 0.4265$), de modo que el hecho de que la media sea menor que el valor de corte de la especificación de diseño puede ser simplemente una consecuencia de la variabilidad muestral. Observe que el histograma no se asemeja en absoluto a una curva normal (y una gráfica de probabilidad normal no exhibe un patrón lineal), aunque las pruebas z con muestras grandes no requieren una distribución de población normal.



Estadística descriptiva

Variable: penetración cónica dinámica

Prueba de normalidad de Anderson-Darling

A-Squarect	1.902
P-Value:	0.000
Mean	28.7615
StDev	12.2647
Variance	150.423
Skewness	0.808264
Kurtosis	–3.9E–01
N	52
Minimum	14.1000
1st Quartile	18.2250
Median	27.5000
3rd Quartile	35.0000
Maximum	57.0000

95% Confidence Interval for Mu

25.3470	3.21761

95% Confidence Interval for Sigma

10.2784	15.2098

95% Confidence Interval for Median

20.0000	31.7000

Figura 7.5 Resumen descriptivo, generado por Minitab, para los datos de penetración cónica dinámica del ejemplo 7.8

1. μ = valor de penetración cónica dinámica promedio verdadero

2. H_0: $\mu = 30$

3. H_a: $\mu < 30$ (por consiguiente, el pavimento no será utilizado a menos que la hipótesis nula sea rechazada)

4. $z = \dfrac{\bar{x} - 30}{s/\sqrt{n}}$

5. Con $n = 52$, $\bar{x} = 28.76$ y $s = 12.2647$,

$$z = \frac{28.76 - 30}{12.2647/\sqrt{52}} = \frac{-1.24}{1.701} = -0.73$$

6. El valor P para esta prueba de cola larga z es $\Phi(-0.73) = 0.2327$.

7. Dado que $0.2327 > 0.05$, H_0 no puede ser rechazada. No se cuenta con evidencia precisa para concluir que $\mu < 30$; el uso del pavimento no se justifica. Considere que de no rechazar H_0, posiblemente cometeríamos un error tipo II. ■

La determinación de β y el tamaño de muestra necesario para estas pruebas con muestra grande pueden basarse en la especificación de un valor adecuado de σ y en el uso de las fórmulas anteriores (aun cuando se utilice s en la prueba) o en el uso de la metodología que se introducirá en conexión con las pruebas t de una muestra, que se analizarán en la sección 7.3.

EJERCICIOS Sección 7.2 (15–28)

15. Sea μ el verdadero tiempo promedio de reacción a un estímulo. Para una prueba z de H_0: $\mu = 5$ contra H_a: $\mu > 5$, determinar el valor P para cada uno de los siguientes valores del estadístico de prueba z.

a. 1.42 **b.** 0.90 **c.** 1.96 **d.** 2.48 **e.** −0.11

16. Se supone que los neumáticos de un tipo recién comprados están inflados a una presión de 30 lb/pulg^2. Sea μ la presión promedio verdadera. Se realiza una prueba para decidir si μ difiere del valor destino. Halle el valor P asociado a cada uno de los siguientes valores del estadístico de prueba z.

a. 2.10 **b.** −1.75 **c.** −0.55 **d.** 1.41 **e.** −5.3

17. Responda las siguientes preguntas en relación con el problema de los neumáticos en el ejemplo 7.7.

a. Si $\bar{x} = 30\,960$ y se utiliza una prueba de nivel $\alpha = 0.01$, ¿cuál es la decisión?

b. Si se utiliza una prueba de nivel 0.01, ¿cuál es $\beta(30\,500)$?

c. Si se utiliza una prueba de nivel 0.01 y también se requiere que $\beta(30\,500) = 0.05$, ¿qué tamaño de muestra n es necesario?

d. Si $\bar{x} = 30\,960$ ¿cuál es la α más pequeña con la cual H_0 puede ser rechazada (con base en $n = 16$)?

18. Reconsidere la situación de secado de pintura del ejemplo 7.5, en el cual el tiempo de secado para un espécimen de prueba está normalmente distribuido con $\sigma = 9$. Las hipótesis H_0: $\mu = 75$ contra H_a: $\mu < 75$ tienen que ser probadas con una muestra aleatoria de $n = 25$ observaciones.

a. ¿A cuántas desviaciones estándar (de $\overline{X}$) por debajo del valor nulo se encuentra $\bar{x} = 72.3$?

b. Si $\bar{x} = 72.3$, ¿cuál es la conclusión si utiliza $\alpha = 0.002$?

c. ¿Cuál es $\beta(70)$ para el procedimiento de prueba con $\alpha = 0.002$?

d. Si se utiliza el procedimiento de prueba con $\alpha = 0.002$, ¿qué n es necesaria para asegurar que $\beta(70) = 0.01$?

e. Si se utiliza una prueba de nivel 0.01 con $n = 100$, ¿cuál es la probabilidad de un error de tipo I cuando $\mu = 76$?

19. Se determinó el punto de fusión de cada una de las 16 muestras de una marca de aceite vegetal hidrogenado y el resultado fue $\bar{x} = 94.32$. Suponga que la distribución del punto de fusión es normal con $\sigma = 1.20$.

a. Pruebe H_0: $\mu = 95$ versus H_a: $\mu \neq 95$ mediante una prueba de dos colas de nivel 0.01.

b. Si se utiliza una prueba de nivel 0.01, ¿cuál es $\beta(94)$, la probabilidad de un error de tipo II cuando $\mu = 94$?

c. ¿Qué valor de n es necesario para garantizar que $\beta(94) = 0.1$ cuando $\alpha = 0.01$?

20. Se anuncia que los focos de un tipo duran un promedio de 750 horas. El precio de estos focos es muy favorable por lo que un cliente potencial ha decidido continuar con un convenio de compra hasta que en conclusión se demuestre que la duración promedio verdadera es menor que la anunciada. Se seleccionó una muestra aleatoria de 50 focos, se determinó la duración de cada uno, se probaron las hipótesis apropiadas con Minitab y se obtuvieron los siguientes resultados.

```
Variable   N    Mean  StDev  SE Mean      Z  P-Value
lifetime  50  738.44  38.20     5.40  −2.14    0.016
```

¿Qué conclusión sería apropiada para un nivel de significancia de 0.05? ¿Y un nivel de significancia de 0.01? ¿Qué nivel de significancia y qué conclusión recomendaría?

21. El porcentaje deseado de SiO_2 en cierto tipo de cemento aluminoso es de 5.5. Para comprobar si en una instalación de producción particular el porcentaje promedio verdadero es de 5.5 se analizaron 16 muestras obtenidas de manera independiente. Suponga que el porcentaje de SiO_2 en una muestra está normalmente distribuida con $\sigma = 0.3$ y que $\bar{x} = 5.25$.

a. ¿Indica esto concluyentemente que el porcentaje promedio verdadero difiere de 5.5?

b. Si el porcentaje promedio verdadero es $\mu = 5.6$ y se utiliza una prueba de nivel $\alpha = 0.01$ con $n = 16$, ¿cuál es la probabilidad de descubrir este alejamiento de H_0?

c. ¿Qué valor de n se requiere para satisfacer $\alpha = 0.01$ y $\beta(5.6) = 0.01$?

22. Con el fin de obtener información sobre las propiedades de resistencia a la corrosión de un tipo de conducto de acero, fueron enterrados en el suelo 45 especímenes durante 2 años. Se midió la penetración máxima (en mils) en cada espécimen y se obtuvo una penetración promedio muestral de $\bar{x} = 52.7$ y una desviación estándar muestral de $s = 4.8$. Los conductos se fabricaron con la especificación de que la penetración promedio verdadera fuera cuando mucho de 50 mils. Serán utilizados a menos que se pueda demostrar concluyentemente que la especificación no se satisface. ¿Qué concluiría?

23. La identificación automática de los límites de las estructuras significativas en una imagen médica es un área de investigación continua. El artículo **"Automatic Segmentation of Medical Images Using Image Registration: Diagnostic and Simulation Applications" (*J. of Medical Engr. and Tech.*, 2005: 53–63)** analiza una nueva técnica para realizar dicha identificación. Una medida de la precisión de la región automática es el desplazamiento lineal promedio. El artículo proporciona las siguientes observaciones de desplazamiento lineal promedio con una muestra de 49 riñones (unidades de dimensiones en pixeles).

1.38	0.44	1.09	0.75	0.66	1.28	0.51
0.39	0.70	0.46	0.54	0.83	0.58	0.64
1.30	0.57	0.43	0.62	1.00	1.05	0.82
1.10	0.65	0.99	0.56	0.56	0.64	0.45
0.82	1.06	0.41	0.58	0.66	0.54	0.83
0.59	0.51	1.04	0.85	0.45	0.52	0.58
1.11	0.34	1.25	0.38	1.44	1.28	0.51

a. Resuma y describa los datos.

b. ¿Es factible que el desplazamiento lineal promedio esté al menos normalmente distribuido en forma aproximada? ¿Se debe suponer normalidad antes de calcular un intervalo de confianza para el desplazamiento lineal promedio verdadero, o probar las hipótesis en cuanto a desplazamiento lineal promedio verdadero? Explique.

c. Los autores comentaron que en la mayoría de los casos el desplazamiento lineal promedio es del orden de 1.0 o mejor. ¿Proporcionan en realidad los datos una fuerte evidencia para concluir que el desplazamiento lineal promedio en estas circunstancias es menor que 1.0? Efectúe una prueba apropiada de hipótesis.

d. Calcule un límite de confianza superior para el desplazamiento lineal promedio verdadero utilizando un nivel de confianza de 95% e interprete este límite.

24. A diferencia de la mayoría de los alimentos envasados, las etiquetas en las bebidas alcohólicas no están obligadas a mostrar el contenido de nutrientes o calorías. El artículo **"What Am I Drinking? The Effects of Serving Facts Information on Alcohol Beverage Containers" (*J. of Consumer Affairs*, 2008: 81–99)** reporta acerca de un estudio piloto en el que se le pidió a cada una de las 58 personas de una muestra que estimaran el contenido de calorías de una lata de 12 onzas de cerveza sabiendo que contiene 153 calorías. La media resultante de la muestra estimada del nivel de calorías fue de 191 y la desviación estándar de la muestra fue de 89. ¿Sugerirán estos datos que la media verdadera estima que el contenido de calorías en la población encuestada excede el contenido real? Pruebe las hipótesis apropiadas para el nivel de significancia de 0.001.

25. Los chalecos antibalas proporcionan protección crítica para personal policial, pero afectan el equilibrio y la movilidad. El artículo **"Impact of Police Body Armour and Equipment on Mobility" (*Applied Ergonomics*, 2013: 957–961)** informa que en una muestra de 52 oficiales varones sometidos a una tarea de velocidad que simula descender de un vehículo con el chaleco antibalas puesto, la media de la muestra fue de 1.95 segundos, y la desviación estándar de la muestra fue 0.20 segundos. ¿Será que el tiempo de la verdadera tarea promedio es menor de 2 segundos? Realice una prueba de hipótesis apropiadas usando un nivel de significancia de 0.01.

26. La cantidad diaria recomendada de zinc en la dieta entre los varones mayores de 50 años de edad es de 15 mg/día. El artículo **"Nutrient Intakes and Dietary Patterns of Older Americans: A National Study" (*J. of Gerontology*, 1992: M145-150)** presenta el siguiente resumen de datos sobre el consumo de zinc en una muestra de varones con edades entre 65 y 74 años: $n = 115$, $\bar{x} = 11.3$ y $s = 6.43$. ¿Indicarán estos datos que la ingesta de zinc diaria promedio en la población de hombres de todas las edades de 65 a 74 años cae por debajo de la cantidad recomendada?

27. Demuestre que, cuando la distribución de la población es normal y cuando se conoce σ, para cualquier $\Delta > 0$ la prueba de dos colas satisface $\beta(\mu_0 - \Delta) = \beta(\mu_0 + \Delta)$, de modo que $\beta(\mu')$ es simétrica respecto a μ_0.

28. Para un valor μ' alternativo fijo, demuestre que $\beta(\mu') \to 0$ cuando $n \to \infty$, para una prueba z de una cola o de dos colas en el caso de una distribución normal de la población con σ conocida.

7.3 Prueba *t* de una sola muestra

Cuando n es pequeño, el teorema del límite central (TLC) ya no puede ser invocado para justificar el uso de una prueba con muestra grande. Esta misma dificultad se presentó al obtener un intervalo de confianza (IC) con muestra pequeña para μ en el capítulo 6. El método utilizado en este capítulo será el mismo que se empleó ahí: se supondrá que la distribución de población es al menos aproximadamente normal y se describirán los procedimientos de prueba cuya validez se fundamenta en esta suposición. Si un investigador tiene una buena razón para creer que la distribución de población es bastante no normal, puede utilizar una prueba libre de distribución. Alternativamente, un estadístico puede ser consultado en cuanto a procedimientos válidos para familias específicas de distribuciones de población, aparte de la familia normal. O puede desarrollarse un procedimiento *bootstrap*.

En el capítulo 6 se utilizó el resultado clave en el cual están basadas las pruebas con una media de población normal para obtener el intervalo de confianza t para una muestra: si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria de una distribución normal, la variable estandarizada

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

tiene una distribución *t* con $n - 1$ grados de libertad (gl). Considere poner a prueba H_0: $\mu = \mu_0$ utilizando el estadístico de prueba $T = (\overline{X} - \mu_0)/(S/\sqrt{n})$. Es decir, el estadístico de prueba resulta de estandarizar $\overline{X}$ conforme a la suposición de que H_0 es verdadera (utilizando $S/\sqrt{n}$ la desviación estándar estimada de $\overline{X}$, en lugar de $\sigma/\sqrt{n}$). Cuando H_0 es verdadera el estadístico de prueba tiene una distribución *t* con $n - 1$ grados de libertad. Conocer la distribución del estadístico de prueba cuando H_0 es verdadera (la "distribución nula") permite determinar el valor *P*.

El estadístico de prueba es en realidad el mismo del caso de muestra grande pero se denota con *T* para recalcar que su distribución de referencia para determinar el valor *P* es una distribución *t* con $n - 1$ grados de libertad en lugar de la distribución normal estándar (*z*). En lugar del área bajo la curva *z* como era el caso para las pruebas de muestras grandes, el valor *P* usará ahora un área bajo la curva t_{n-1} (véase la figura 7.6).

Prueba *t* con una muestra

Hipótesis nula: H_0: $\mu = \mu_0$

Valor del estadístico de prueba: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Hipótesis alternativa	Determinación del valor *P*
H_a: $\mu > \mu_0$	Área bajo la curva t_{n-1} a la derecha de *t*
H_a: $\mu < \mu_0$	Área bajo la curva t_{n-1} a la izquierda de *t*
H_a: $\mu \neq \mu_0$	$2 \cdot$ (Área bajo la curva t_{n-1} a la derecha de $\lvert t \rvert$)

Suposición: Los datos consisten en una muestra aleatoria de una distribución de población normal.

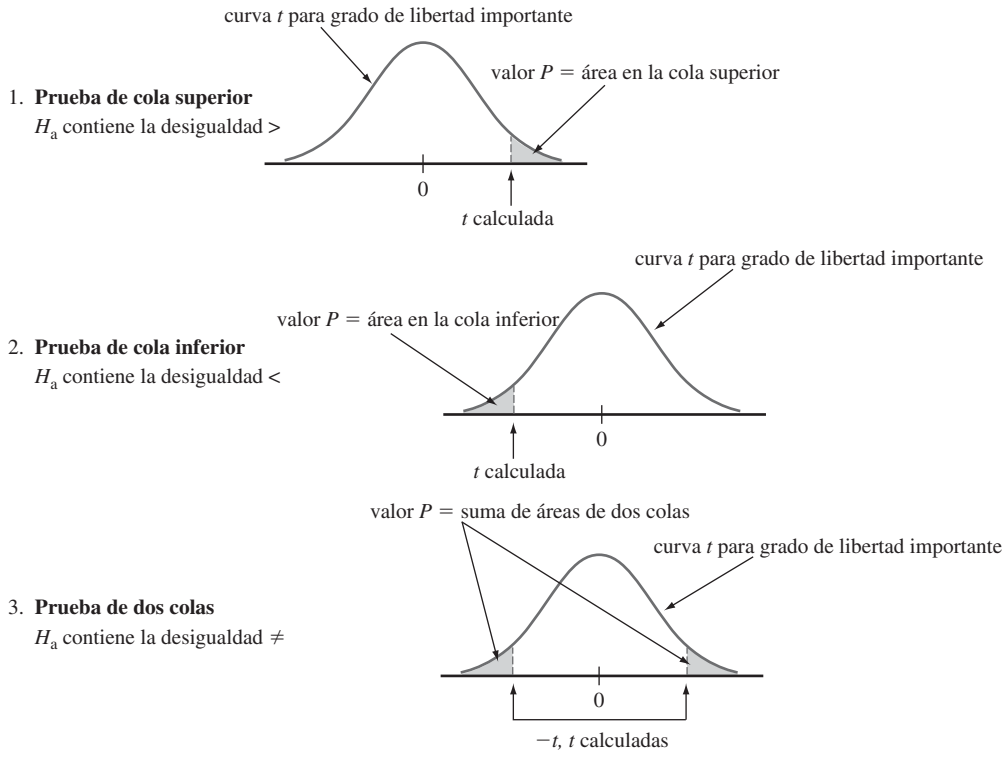


Figura 7.6 Valores *P* para las pruebas *t*

Lamentablemente la tabla de valores críticos de t que se utilizó para el cálculo del intervalo de confianza y la predicción en el capítulo 6 no proporciona mucha información acerca de las áreas de cola de la curva t. Esto es porque para cada distribución de t sólo hay siete valores para las áreas de cola que se usan más comúnmente: 0.10, 0.05, 0.025, 0.01, 0.005, 0.001 y 0.0005. La determinación del valor P sería fácil si tuviéramos una tabla de áreas de cola (o, alternativamente, áreas acumuladas) semejante a nuestra tabla z: para cada distribución t diferente, el área bajo la correspondiente curva a la derecha o la izquierda de valores 0.00, 0.01, 0.02, 0.03, . . . , 3.97, 3.98, 3.99 y finalmente 4.00. Pero para esto sería necesaria una página entera de texto para cada distribución diferente de t.

Por consiguiente, se ha incluido otra tabla t en la tabla A.8 del apéndice. Contiene una tabulación del área bajo la cola superior de la curva t pero con menos precisión decimal que la que proporciona la tabla z. Cada otra columna de la tabla es para un número distinto de grados de libertad, y las filas son para los valores calculados del estadístico de prueba t que van desde 0.0 a 4.0 en incrementos de 0.1. Por ejemplo, el número 0.074 aparece en la intersección de la fila 1.6 y la columna 8 de los grados de libertad. Por tanto, el área bajo la curva de 8 grados de libertad a la derecha de 1.6 (un área de cola superior) es 0.074. Puesto que las curvas t son simétricas sobre 0, 0.074 también es el área bajo la curva de 8 grados de libertad a la izquierda de -1.6.

Suponga, por ejemplo, que una prueba de H_0: $\mu = 100$ versus H_a: $\mu > 100$ se basa en la distribución t de 8 grados de libertad. Si el valor calculado del estadístico de prueba es $t = 1.6$, entonces el valor P para esta prueba con cola superior es 0.074. Debido a que 0.074 excede 0.05, no seríamos capaces de rechazar H_0 en un nivel de significancia de 0.05. Si la hipótesis alternativa es H_a: $\mu < 100$ y una prueba basada en 20 grados de libertad produce $t = -3.2$, entonces la tabla A.7 del apéndice muestra que el valor P es el área de cola inferior capturada 0.002. La hipótesis nula puede ser rechazada en cualquier nivel 0.05 o 0.01. En el capítulo siguiente presentamos una prueba t de hipótesis sobre la diferencia entre dos medias de población. Supongamos que las hipótesis relevantes son H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 \neq 0$; la hipótesis nula establece que las medias de las dos poblaciones son idénticas, mientras que la hipótesis alternativa establece que son diferentes sin especificar una dirección de alejamiento de H_0. Si una prueba t se basa en 20 grados de libertad y $t = 3.2$, entonces el valor P de esta prueba de dos colas es $2(0.002) = 0.004$. Este sería el valor P de $t = -3.2$. El área de la cola se duplica ya que valores mayores de 3.2 y más pequeños que -3.2 son más contradictorios a H_0 que lo que eran los valores calculados (valores más alejados de *cualquier* cola de la curva t).

EJEMPLO 7.9 Las nanofibras de carbono tienen la aplicación potencial de ser materiales que manejan el calor, que refuerzan compuestos y como componente en nanoelectrónica y en fotónica. Los siguientes datos de esfuerzo de falla (MPa) de especímenes de fibra se obtuvieron de una gráfica en el artículo **"Mechanical and Structural Characterization of Electrospun PAN2Derived Carbon Nanofibers" (*Carbon*, 2005: 2175–2185)**.

300	312	327	368	400	425	470	556	573	575
580	589	626	637	690	715	757	891	900	

El resumen de las cantidades es $n = 19$, $\bar{x} = 562.68$, $s = 180.874$, $s/\sqrt{n} = 41.495$. ¿Proporcionarán los datos evidencias convincentes para concluir que el esfuerzo de falla promedio verdadero excede los 500 MPa?

La figura 7.7 muestra un diagrama de probabilidad normal de los datos; el considerable patrón lineal indica que una distribución de la población normal del esfuerzo de falla es bastante verosímil, lo que nos permite emplear la prueba t de una muestra (el cuadro a la derecha de la gráfica da información sobre una prueba formal de la hipótesis de que la distribución de la población es normal).

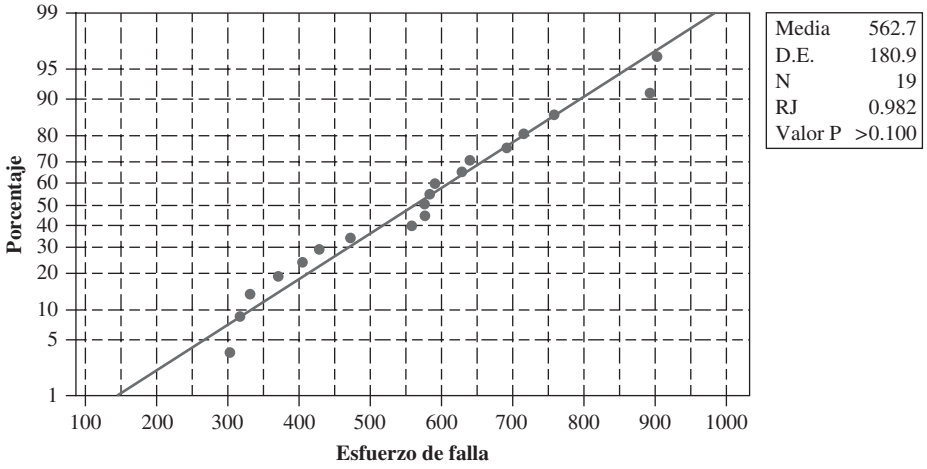


Figura 7.7 Gráfica de población normal de los datos del esfuerzo de falla

Realicemos una prueba de las hipótesis pertinentes con un nivel de significancia de 0.05.

1. El parámetro de interés es μ = esfuerzo de falla promedio verdadero
2. La hipótesis nula es H_0: $\mu = 500$
3. La hipótesis alternativa apropiada es H_a: $\mu > 500$ (así creeremos que el esfuerzo de falla promedio verdadera excede a 500 sólo si puede rechazarse la hipótesis).
4. El estadístico de prueba t de una muestra es $T = (\overline{X} - 500)/(S/\sqrt{n})$. Este valor t de los datos es resultado de sustituir $\overline{X}$ por $\overline{x}$ y S por s.
5. El valor del estadístico de prueba es $t = (562.68 - 500)/41.495 = 1.51$.
6. La prueba se basa en $19 - 1 = 18$ grados de libertad. La entrada en esa columna y la fila de 1.5 de la tabla A.8 del apéndice es 0.075. Puesto que la prueba es de cola superior (porque aparece $>$ en H_a), se deduce que el valor $P \approx 0.075$ (Minitab dice 0.074).
7. Debido a que $0.075 > 0.05$, no hay suficiente evidencia para justificar el rechazo de la hipótesis nula al nivel de significancia 0.05. En lugar de concluir que el esfuerzo de falla verdadero promedio supera los 500, parece que la variabilidad del muestreo proporciona una explicación factible para el hecho de que la media muestral exceda 500 por una cantidad bastante considerable. ■

EJEMPLO 7.10 Muchos de los efectos nocivos del tabaquismo sobre la salud han sido bien documentados. En el artículo **"Smoking Abstinence Impairs Time Estimation Accuracy in Cigarette Smokers" (*Psychopharmacology Bull*., 2003: 90–95)** se describe una investigación acerca de si la percepción del tiempo, que es un indicador de la capacidad de una persona para concentrarse, se deteriora durante los periodos de abstinencia de nicotina. Después de no fumar durante 24 horas, se le pidió a cada uno de 20 fumadores que estimara cuánto tiempo había transcurrido en un lapso de 45 segundos. Los siguientes datos de la percepción del tiempo transcurrido son consistentes con el resumen de cantidades que proporciona el artículo citado.

69	65	72	73	59	55	39	52	67	57
56	50	70	47	56	45	70	64	67	53

Una gráfica de probabilidad normal de los datos muestra un patrón lineal muy importante. Se va a realizar una prueba de hipótesis a nivel de significancia 0.05 para decidir si la percepción de cierto tiempo transcurrido promedio difiere del tiempo conocido de 45.

1. μ = percepción del tiempo transcurrido promedio verdadero para todos los fumadores expuestos al régimen experimental descrito
2. H_0: $\mu = 45$
3. H_0: $\mu \neq 45$
4. $t = (\bar{x} - 45)/(s/\sqrt{n})$
5. Con $\bar{x} = 59.30$ y $s/\sqrt{n} = 9.84/\sqrt{20} = 2.200$, el valor del estadístico de prueba es $t = 14.3/2.200 = 6.50$.
6. El valor P para una prueba de dos colas es dos veces el área bajo la curva de t con 19 grados de libertad a la derecha de 6.50. Puesto que la tabla A.8 muestra que el área bajo esta curva de t a la derecha de 4.0 es 0, el área a la derecha de 6.50 es 0. El valor P es entonces $2(0) = 0$ (0.00000 según el software).
7. Un valor P tan pequeño es un fuerte argumento para rechazar H_0 en cualquier nivel de significancia razonable y en particular en el nivel de significancia 0.05. La diferencia entre la media de la muestra y su valor esperado cuando H_0 es verdadera no puede explicarse simplemente por la variación de la probabilidad. La media verdadera de la percepción del tiempo transcurrido es evidentemente diferente de 45, por lo que la abstinencia de la nicotina parece afectar la percepción del tiempo. ■

β y determinación del tamaño de la muestra

El cálculo de β con el valor alternativo μ' para una distribución de población normal con σ conocida se realizó mediante la conversión de la desigualdad valor $P > \alpha$ en un enunciado de $\bar{x}$ (por ejemplo, $\bar{x} < \mu_0 + z_\alpha \cdot \sigma/\sqrt{n}$) y luego restar μ' para estandarizar correctamente. Un método implica observar que cuando $\mu = \mu'$, el estadístico de prueba $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ sigue teniendo una distribución normal con varianza 1, pero ahora la media de Z está dada por $(\mu' - \mu_0)/(\sigma/\sqrt{n})$. Es decir, cuando $\mu = \mu'$, el estadístico de prueba sigue teniendo una distribución normal pero no la distribución normal estándar. Debido a esto $\beta(\mu')$ es un área bajo la curva normal correspondiente a la media $(\mu' - \mu_0)/(\sigma/\sqrt{n})$ y varianza 1. Tanto α como β implican trabajar con variables normalmente distribuidas.

El cálculo de $\beta(\mu')$ para la prueba t es mucho menos directo. Esto es porque la distribución del estadístico de prueba $T = (\bar{X} - \mu_0)/(S/\sqrt{n})$ es bastante complicada cuando H_0 es falsa y H_a es verdadera. Por consiguiente, en una prueba de cola superior, determinar

$$\beta(\mu') = P(T < t_{\alpha, n-1} \text{ cuando } \mu = \mu' \text{ en lugar de } \mu_0)$$

implica integrar una desagradable función de densidad. Esto debe hacerse numéricamente. Los resultados se resumen en gráficas de β que aparecen en la tabla A.17 del apéndice. Existen cuatro juegos de gráficas, correspondientes a pruebas de una cola a nivel 0.05 y nivel 0.01, y pruebas de dos colas a los mismos niveles.

Para entender cómo se utilizan estas gráficas, obsérvese primero que tanto β como el tamaño de muestra necesario n son funciones no sólo de la diferencia absoluta $|\mu_0 - \mu'|$ sino de $d = |\mu_0 - \mu'|/\sigma$. Suponga, por ejemplo, que $|\mu_0 - \mu'| = 10$. Este alejamiento de H_0 será mucho más fácil de descubrir (β más pequeña) cuando $\sigma = 2$, en cuyo caso μ_0 y μ' están a 5 desviaciones estándar de la población, una de otra, que cuando $\sigma = 10$. El hecho de que β para la prueba t dependa de d y no sólo de $|\mu_0 - \mu'|$ es desafortunado, puesto que para utilizar las gráficas se debe tener alguna idea del valor verdadero de σ. Una suposición conservadora (grande) para σ dará por resultado un valor conservador (grande) de $\beta(\mu')$ y una estimación conservadora del tamaño de muestra necesario para α y $\beta(\mu')$ prescritas.

Una vez que se seleccionan la μ' alternativa y el valor de σ, se calcula d y su valor se localiza sobre el eje horizontal del conjunto de curvas pertinente. El valor de β es la altura de la curva con $n - 1$ grados de libertad por encima del valor de d (es necesaria una interpolación visual si $n - 1$ no es un valor para el cual aparezca la curva correspondiente) como se ilustra en la figura 7.8.

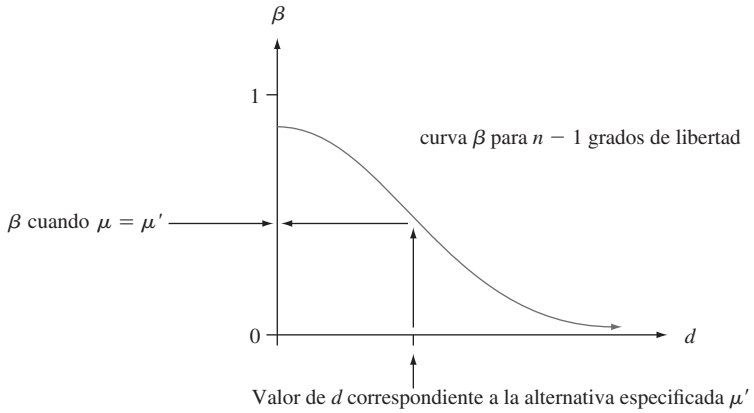


Figura 7.8 Curva β típica de la prueba t

En lugar de fijar n (es decir, $n - 1$) y, por consiguiente, la curva particular en donde se lee β) se podría prescribir tanto α (0.05 o 0.01 en este caso) y un valor de β para las μ' y σ seleccionadas. Después de calcular d se localiza el punto (d, β) en el conjunto de gráficas pertinentes. La curva debajo y más próxima a este punto da $n - 1$ y, por consiguiente, n (de nuevo con frecuencia se requiere interpolación).

EJEMPLO 7.11 Se supone que la caída de voltaje promedio verdadera entre el colector y el emisor de transistores bipolares de compuerta aislados de cierto tipo es cuando mucho de 2.5 volts. Un investigador selecciona una muestra de $n = 10$ de esos transistores y utiliza los voltajes resultantes para probar H_0: $\mu = 2.5$ contra H_a: $\mu > 2.5$ mediante una prueba t con nivel de significancia $\alpha = 0.05$. Si la desviación estándar de la distribución de voltaje es $\sigma = 0.100$ ¿qué tan probable es que H_0 no sea rechazada cuando en realidad $\mu = 2.6$? Con $d = |2.5 - 2.6|/0.100 = 1.0$, el punto sobre la curva β con 9 grados de libertad para una prueba de una cola con $\alpha = 0.05$ por encima de 1.0 tiene una altura aproximadamente de 0.1, por tanto $\beta \approx 0.1$. El investigador podría pensar que este es un valor de β demasiado grande con semejante alejamiento sustancial de H_0 y puede ser que desee tener $\beta = 0.05$ para este valor alternativo de μ. Puesto que $d = 1.0$, el punto $(d, \beta) = (1.0, 0.05)$ debe ser localizado. Este punto se aproxima mucho a la curva de 14 grados de libertad, por tanto, con $n = 15$ se obtendrá tanto $\alpha = 0.05$ como $\beta = 0.05$ cuando el valor de μ es 2.6 y $\sigma = 0.10$. Un valor más grande de σ daría una β más grande para esta alternativa, y un valor alternativo de μ más cercano a 2.5 también daría por resultado un valor incrementado de β. ■

La mayoría de los programas computacionales de estadística también calculará probabilidades de error tipo II. Por lo general, estos programas trabajan en términos de **potencia**, que es simplemente $1 - \beta$. Un valor pequeño de β (cerca de 0) es equivalente a una potencia grande (cerca de 1). Una prueba de *gran alcance* es la que tiene gran potencia y, por tanto, buena capacidad para detectar cuándo la hipótesis nula es falsa.

Por ejemplo, se le pide a Minitab que determine la potencia de la prueba de cola superior del ejemplo 7.11 para tres tamaños de muestra de 5, 10 y 15 cuando $\alpha = 0.05$, $\sigma = 0.10$ y el valor de μ es en realidad 2.6 en lugar del valor nulo de 2.5, una "diferencia" de $2.6 - 2.5 = 0.1$. También se le pidió al software determinar el tamaño necesario de la muestra para una potencia de 0.9 ($\beta = 0.1$) y 0.95. Aquí están los datos resultantes.

```
Power and Sample Size
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 0.1

            Sample
Difference    Size     Power
0.1             5     0.579737
0.1            10     0.897517
0.1            15     0.978916

                Sample Target
                                      Actual
Difference       Size       Power     Power
0.1               11        0.90      0.924489
0.1               13        0.95      0.959703
```

La potencia para el tamaño de muestra $n = 10$ es un poco menor que 0.9. Así que si se insiste en que la potencia sea al menos de 0.9, se requiere una muestra de tamaño 11 y la potencia real para n es aproximadamente 0.92. El software dice que para una potencia objetivo de 0.95, es necesario un tamaño de muestra de $n = 13$, mientras que al echar un vistazo a nuestras curvas β dio 15. Cuando está disponible, este tipo de software es más confiable que las curvas. Por último, ahora Minitab también proporciona curvas de potencia para los tamaños determinados de muestra, tal como se ilustra en la figura 7.9. Estas curvas muestran cómo aumenta la potencia para cada tamaño de muestra a medida que el valor real de μ se desplaza más allá y más lejos del valor nulo.

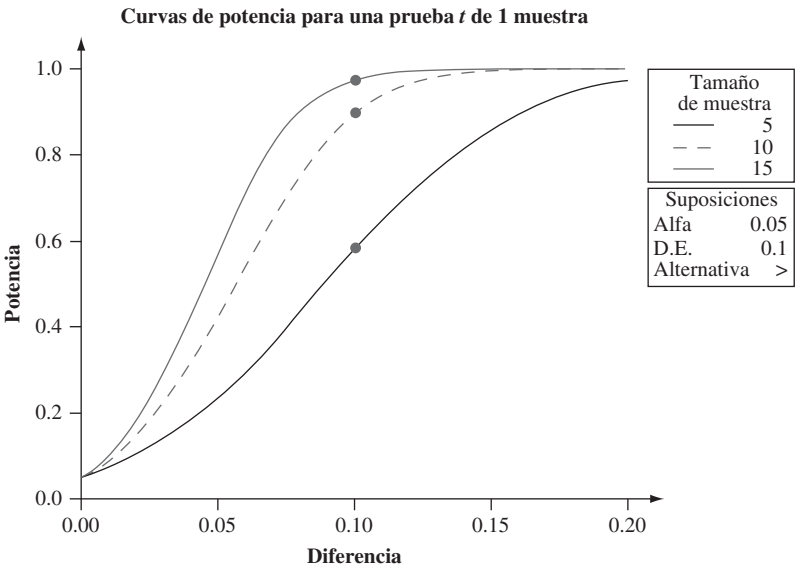


Figura 7.9 Curvas de potencia de Minitab para la prueba *t* del ejemplo 7.11

Variación en los valores *P*

El valor *P* que resulta al realizar una prueba en una muestra seleccionada *no* es la probabilidad de que H_0 sea verdadera, ni la probabilidad de rechazar la hipótesis nula. Una vez más, es la probabilidad, calculada suponiendo que H_0 es verdadera, de obtener un valor del estadístico de prueba al menos tan contradictorio a la hipótesis nula como el valor real

resultante. Por ejemplo, considere la prueba H_0: $\mu = 50$ versus H_0: $\mu < 50$, usando una prueba t de cola inferior con base en 20 grados de libertad. Si el valor calculado del estadístico de prueba es $t = -2.00$, entonces

$$\begin{aligned}\text{valor } P &= P(T < -2.00 \text{ cuando } \mu = 50)\\ &= \text{área bajo la } t_{20} \text{ a la izquierda de } -2.00 = 0.030\end{aligned}$$

Pero si se selecciona una segunda muestra, el valor resultante de t es casi seguro que será diferente de -2.00, por lo que el valor P correspondiente también es probable que difiera de 0.030. Ya que el valor del estadístico de prueba varía por sí mismo de una muestra a otra, el valor P también varía de una muestra a otra. Es decir, el estadístico de prueba es una variable aleatoria y, por tanto, el valor P será una variable aleatoria. Una primera muestra puede dar un valor P de 0.030, una segunda muestra puede dar como resultado un valor P de 0.117, una tercera muestra puede producir el valor P de 0.061 y así sucesivamente.

Si H_0 es falsa, esperamos que el valor P sea cercano a 0 por lo que se puede rechazar la hipótesis nula. Por otro lado, cuando H_0 es verdadera, nos gustaría que el valor P superara el nivel de significancia seleccionado, así la decisión correcta es no rechazar H_0. El siguiente ejemplo presenta simulaciones que muestran cómo se comporta el valor P cuando la hipótesis nula es verdadera así como cuando es falsa.

EJEMPLO 7.12 El consumo de combustible (mpg) de cualquier vehículo nuevo particular bajo condiciones de conducción puede no ser idéntico a la figura de EPA (Agencia de Protección Ambiental) que aparece en la etiqueta del vehículo. Supongamos se deben seleccionar cuatro diferentes vehículos de un tipo en particular y conducir en una ruta dada, después de lo cual se debe determinar la eficiencia de combustible de cada uno.

Sea μ la eficiencia de combustible promedio verdadero bajo estas condiciones. Considere la prueba H_0: $\mu = 20$ versus H_0: $\mu > 20$ utilizando la prueba t de una muestra basada en la muestra resultante. Puesto que la prueba se basa en $n - 1 = 3$ grados de libertad, el valor P para una prueba de cola superior es el área bajo la curva t con 3 grados de libertad a la derecha de la t calculada.

Suponga primero que la hipótesis nula es verdadera. Pedimos a Minitab que generara 10 000 muestras diferentes, cada una con 4 observaciones, de una distribución de la población normal con media $\mu = 20$ y desviación estándar $\sigma = 2$. La primera muestra y el resumen de cantidades resultante fueron

$$x_1 = 20.830,\ x_2 = 22.232,\ x_3 = 20.276,\ x_4 = 17.718$$
$$\bar{x} = 20.264 \quad s = 1.8864 \quad t = \frac{20.264 - 20}{1.8864/\sqrt{4}} = 0.2799$$

El valor P es el área bajo la curva de t con 3 grados de libertad a la derecha de 0.2799, que según Minitab es 0.3989. Utilizando un nivel de significancia de 0.05, por supuesto, la hipótesis nula no sería rechazada. Los valores t para las siguientes cuatro muestras fueron -1.7591, 0.6082, -0.7020 y 3.1053, con los valores P correspondientes 0.912, 0.293, 0.733 y 0.0265.

La figura 7.10(a) muestra un histograma de los 10 000 valores P de este experimento de simulación. Aproximadamente 4.5% de estos valores P están en el intervalo de primera de clase de 0 a 0.05. Así, cuando se utiliza un nivel de significancia de 0.05 se rechaza la hipótesis nula en aproximadamente 4.5% de estas 10 000 pruebas. Si continuamos generando muestras y realizando la prueba para cada muestra en el nivel de significancia 0.05, a largo plazo 5% de los valores P estarían en el intervalo de primera clase. Esto es

porque cuando H_0 es verdadera y se utiliza una prueba con nivel de significancia 0.05, por definición, la probabilidad de rechazar H_0 es 0.05.

Al observar el histograma, parece que la distribución de valores de P es relativamente plana. De hecho, puede demostrarse que cuando H_0 es verdadera, la distribución de probabilidad del valor P es una distribución uniforme en el intervalo de 0 a 1. Es decir, la curva de densidad es completamente plana en este intervalo y así debe tener una altura de 1 si el área total bajo la curva es 1. Puesto que el área bajo dicha curva a la izquierda de 0.05 es $(0.05)(1) = 0.05$, otra vez tenemos que la probabilidad de rechazar H_0 cuando es verdadera es 0.05, el nivel de significancia elegido.

Ahora considere lo que sucede cuando H_0 es falsa porque $\mu = 21$. Otra vez tuvimos que generar con Minitab 10 000 muestras de tamaño 4 (cada una a partir de una distribución normal con $\mu = 21$ y $\sigma = 2$), calcular $t = (\bar{x} - 20)/(s/\sqrt{n})$ para cada una y luego determinar el valor P. El primer resultado de dicha muestra en $\bar{x} = 20.6411$, $s = 0.49637$, $t = 2.5832$, valor $P = 0.0408$. La figura 7.10(b) muestra un histograma de los valores P obtenidos. La forma de este histograma es bastante diferente de la figura 7.10(a); hay una mayor tendencia a que el valor P sea pequeño (más cercano a 0) cuando $\mu = 21$ que cuando $\mu = 20$. Otra vez H_0 se rechaza al nivel de significancia 0.05 siempre que el valor P sea a lo más 0.05 (en el intervalo de primera clase). Lamentablemente, este es el caso de sólo 19% de los valores P. Tan sólo 19% de las 10 000 pruebas rechazan correctamente la hipótesis nula; para 81% se ha cometido un error tipo II. La dificultad es que el tamaño de muestra es muy pequeño y 21 no es muy diferente del valor afirmado por la hipótesis nula.

La figura 7.10(c) muestra lo que le sucede al valor P cuando H_0 es falsa porque $\mu = 22$ (aún con $n = 4$ y $\sigma = 2$). El histograma está aún más concentrado hacia valores cer-

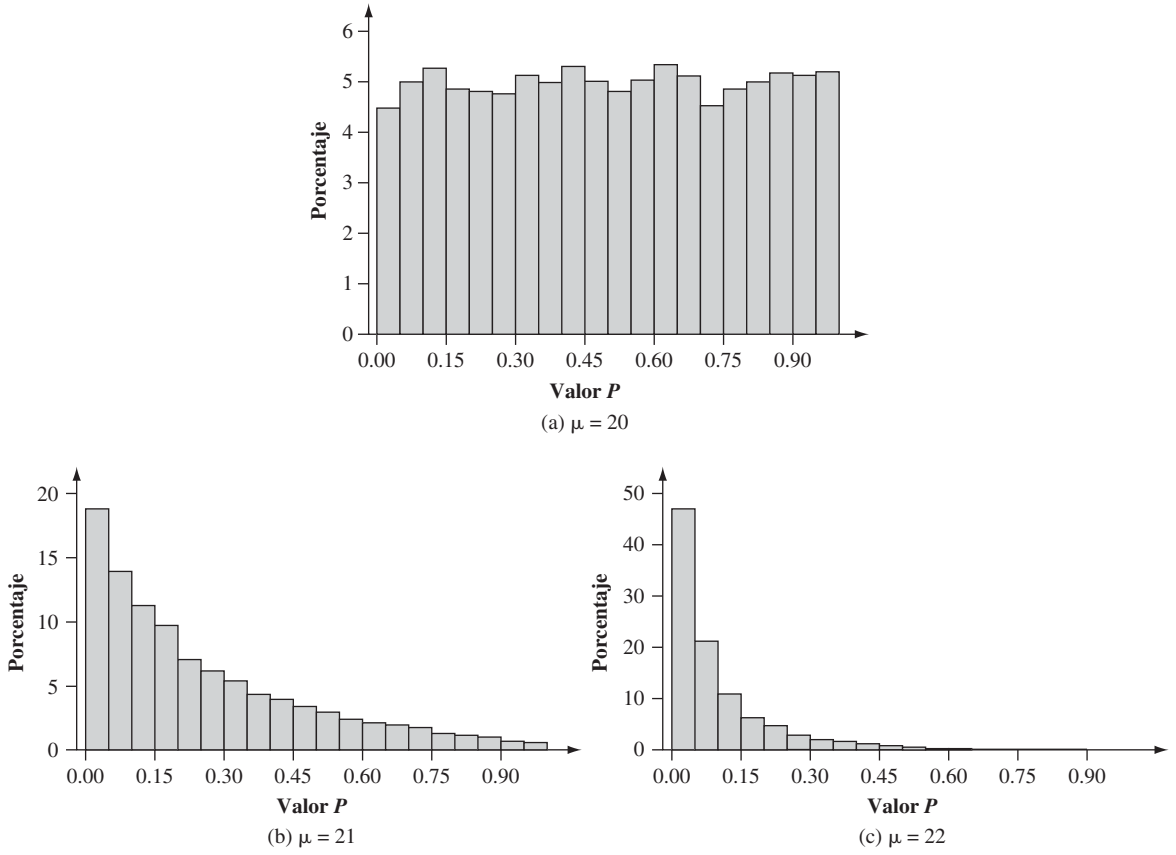


Figura 7.10 Resultados de la simulación del valor P del ejemplo 7.12

canos a 0, que fue el caso cuando $\mu = 21$. En general, puesto que μ se mueve alejándose a la derecha del valor nulo 20, la distribución de los valores P se vuelve cada vez más concentrada en valores cercanos a 0. Hasta aquí un poco menos de 50% de los valores P es menor que 0.05. Por tanto, aún es poco más probable que improbable que la hipótesis nula sea incorrectamente no rechazada. Sólo para valores μ mucho más grandes que 20 (p. ej., al menos 24 o 25) es muy probable que el valor P sea menor de 0.05 y así nos dará la conclusión correcta.

La gran idea de este ejemplo es que debido a que el valor de cualquier estadístico de prueba es aleatorio, también el valor P es una variable aleatoria y, por tanto, tienen una distribución. Cuanto más fuerte es el valor real del parámetro desde el valor especificado por la hipótesis nula, más se concentrará en valores cercanos a 0 la distribución de los valores P y mayor será la posibilidad de que la prueba rechace correctamente H_0 (correspondiente a β más pequeños). ■

Siempre que se reporta el valor observado de un estadístico como $\overline{X}$ o $\hat{p}$, es una buena práctica estadística incluir una medida cuantitativa de la precisión del estadístico, por ejemplo, que el error estándar estimado de $\overline{X}$ es $s/\sqrt{n}$. El valor P es en sí mismo un estadístico, su valor se puede calcular cuando se tienen los datos y en particular se selecciona un procedimiento de prueba, y antes que se tengan esos datos el valor P está sujeto a la aleatoriedad. Así que sería bueno contar con σ_P o con un cálculo de esta desviación estándar.

Lamentablemente la distribución de muestreo de un valor P es en general bastante complicada. Los resultados de la simulación del ejemplo 7.12 sugieren que la distribución muestral está muy sesgada cuando H_0 es falsa (está uniformemente distribuida en (0, 1) cuando H_0 es verdadera y el estadístico de prueba tiene una distribución continua, por ejemplo, una distribución t). Una desviación estándar no es tan fácil de interpretar y utilizar cuando hay una importante no normalidad. Los estadísticos Dennis Boos y Leonard Stefanski investigan el comportamiento aleatorio del valor P en su artículo **"*P*-Value Precision and Reproducibility" (*The American Statistician*, 2011: 213-221).** Para abordar la no normalidad se manejó la cantidad –log(valor P). El log-transformado del valor P hace que muchos procedimientos de prueba tengan aproximadamente una distribución normal cuando n es grande.

Suponga que se aplica un procedimiento de prueba en particular a los resultados de los datos de la muestra en un valor P de 0.001. Entonces H_0 se rechazaría utilizando un nivel de significancia de 0.05 o 0.01. Si entonces se selecciona una nueva muestra de la misma distribución de población, ¿qué tan probable es que el valor P para estos datos nuevos conduzca a rechazar H_0 con un nivel de significancia de 0.05 o 0.01? Esto es lo que los autores del artículo citado entienden por "reproducibilidad": ¿qué posibilidades hay de que una muestra nueva conduzca a la misma conclusión a la que se llegó usando la muestra original? La respuesta a esta pregunta depende de la distribución de la población, del tamaño de la muestra y del procedimiento de prueba utilizado. Sin embargo, con base en sus investigaciones, los autores sugieren las siguientes pautas generales:

> Si el valor P para los datos originales es 0.0001, entonces P(valor $P \leq 05) \approx 0.97$, mientras que esta probabilidad es aproximadamente 0.91 si el valor P original es 0.001 y aproximadamente 0.73 cuando el valor P original es 0.01.

Especialmente cuando el valor P original es alrededor de 0.01 hay razonablemente una buena probabilidad de que una muestra nueva no conducirá al rechazo de H_0 en el nivel de significancia 5%. Por tanto, a menos que el valor P original sea muy pequeño, no sería sorprendente que una nueva muestra contradijera la inferencia de los datos originales. Un valor P no mucho menor en un nivel de significancia elegido como 0.05 o 0.01 ¡debe verse con cierta cautela!

EJERCICIOS Sección 7.3 (29–40)

29. Se supone que el diámetro promedio verdadero de rodamientos de bolas de un tipo es de 0.5 pulg. Se realizará una prueba t con una muestra para ver si este es el caso. ¿Qué conclusión es apropiada en cada una de las siguientes situaciones?

a. $n = 13$, $t = 1.6$, $\alpha = 0.05$
b. $n = 13$, $t = -1.6$, $\alpha = 0.05$
c. $n = 25$, $t = -2.6$, $\alpha = 0.01$
d. $n = 25$, $t = -3.9$

30. Se selecciona una muestra de n ejemplares de lodo y se determina el pH de cada uno. La prueba t de una muestra se utilizará entonces para ver si hay pruebas contundentes con el fin de concluir que el pH promedio verdadero es menor de 7.0. ¿Qué conclusión es adecuada en cada una de las siguientes situaciones?

a. $n = 6$, $t = -2.3$, $\alpha = 0.05$
b. $n = 15$, $t = -3.1$, $\alpha = 0.01$
c. $n = 12$, $t = -1.3$, $\alpha = 0.05$
d. $n = 6$, $t = 0.7$, $\alpha = 0.05$
e. $n = 6$, $\bar{x} = 6.68$, $s/\sqrt{n} = 0.0820$

31. La pintura utilizada para trazar los señalamientos en las autopistas debe reflejar suficiente luz para que sean claramente visibles de noche. Sea μ la lectura promedio verdadera del *reflejómetro* de un nuevo tipo de pintura considerada. Una prueba de H_0: $\mu = 20$ versus H_a: $\mu > 20$ se basará en una muestra aleatoria de tamaño n de una distribución de población normal. ¿Qué conclusión es apropiada en cada una de las siguientes situaciones?

a. $n = 15$, $t = 3.2$, $\alpha = 0.05$
b. $n = 9$, $t = 1.8$, $\alpha = 0.01$
c. $n = 24$, $t = -0.2$

32. La conductividad relativa de un dispositivo semiconductor está determinado por la cantidad de impurezas "adicionadas" al mismo durante su fabricación. Un diodo de silicio usado para propósitos específicos requiere un voltaje de corte promedio de 0.60 V y si este no se alcanza se debe ajustar la cantidad de impurezas. Se seleccionó una muestra de diodos y se determinó el voltaje de corte. Los siguientes datos de salida obtenidos con el software SAS son el resultado de una solicitud para probar las hipótesis apropiadas.

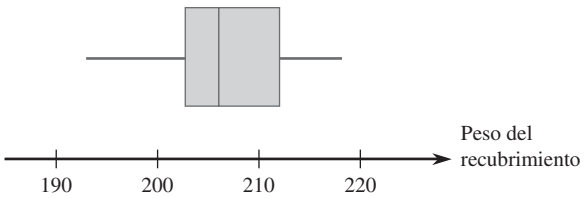
N	Mean	Std Dev	T	Prob. > |T|
15	0.0453333	0.0899100	1.9527887	0.0711

[*Nota*: SAS prueba explícitamente H_0: $\mu = 0$, así que para probar H_0: $\mu = 0.60$, el valor nulo 0.60 se debe restar de cada x_i; la media reportada es entonces el promedio de los valores $(x_i - 0.60)$. También, el valor P de SAS siempre es para una prueba de dos colas.] ¿Qué se concluiría con un nivel de significancia de 0.01?, ¿y de 0.05?, ¿y de 0.10?

33. El artículo **"The Foreman's View of Quality Control" (*Quality Engr.*, 1990: 257–280)** describe una investigación de los pesos del recubrimiento de grandes tuberías resultantes de un proceso de galvanizado. Los estándares de producción demandan un peso promedio verdadero de 200 lb por tubería. Los siguientes resumen y gráfica de caja descriptivos fueron producidos por Minitab.

Variable	N	Mean	Median	TrMean	StDev	SEMean
ctg wt	30	206.73	206.00	206.81	6.35	1.16

Variable	Min	Max	Q1	Q3
ctg wt	193.00	218.00	202.75	212.00



a. ¿Qué sugiere la gráfica de caja sobre el estado de la especificación del peso de recubrimiento promedio verdadero?
b. Una gráfica de probabilidad normal de los datos resultó bastante recta. Use los datos de salida descriptivos para probar las hipótesis apropiadas.

34. Las siguientes son observaciones en la distancia de frenado (pies) de un camión particular a 20 millas por hora en condiciones experimentales específicas **("Experimental Measurement of the Stopping Performance of a Tractor-Semitrailer from Multiple Speeds". NHTSA, DOT HS 811 488, junio de 2011):**

32.1 30.6 31.4 30.4 31.0 31.9

El trabajo citado señala que bajo estas condiciones, la distancia máxima permitida es 30. Una gráfica de probabilidad normal valida la suposición de que la distancia de frenado se distribuye normalmente**.**

a. ¿Sugerirán los datos que la media de la distancia de frenado verdadera supera este valor máximo? Pruebe las hipótesis apropiadas usando $\alpha = 0.01$.
b. Determine la probabilidad de un error tipo II cuando $\alpha = 0.01$, $\sigma = 5.65$ y el valor real de μ es 31. Repita este proceso para $\mu = 32$ (use software estadístico o la tabla A.17).
c. Repita el inciso b) usando $\sigma = 0.80$ y compare los resultados con los del inciso b).
d. ¿Qué tamaño muestral sería necesario para tener $\alpha = 0.01$ y $\beta = 0.10$ cuando $\mu = 31$ y $\sigma = 0.65$?

35. El artículo **"Uncertainty Estimation in Railway Track Life-Cycle Cost" (*J. of Rail and Rapid Transit*, 2009)** presenta los siguientes datos sobre el tiempo de reparación (minutos) de la rotura de un carril alto en una vía curva del tren de cierta línea de ferrocarril.

159 120 480 149 270 547 340 43 228 202 240 218

Una gráfica de probabilidad normal de los datos muestra un patrón bastante lineal, por lo que es factible que la distribu-

ción de la población del tiempo de reparación sea al menos aproximadamente normal. La desviación media y la desviación estándar de la muestra son 249.7 y 145.1, respectivamente.

a. ¿Habrá pruebas de peso para concluir que el tiempo medio verdadero de reparación es superior a 200 minutos? Lleve a cabo una prueba de hipótesis con un nivel de significancia de 0.05.

b. Usando $\sigma = 150$, ¿cuál es la probabilidad de error tipo II de la prueba utilizada en el inciso a) cuando el tiempo promedio verdadero de reparación es en realidad 300 minutos? Es decir, ¿cuál es $\beta(300)$?

36. ¿Alguna vez se ha visto frustrado porque no puede conseguir un contenedor de cierto tipo en el que se pueda liberar la última parte de su contenido? El artículo **"Shake, Rattle, and Squeeze: How Much Is Left in That Container?" (*Consumer Reports*, mayo de 2009: 8)** informa sobre una investigación de este tema para varios productos de consumo. Supongamos se seleccionan al azar cinco tubos de 6.0 onzas de pasta de dientes de una marca en particular y se les oprime hasta que ya no salga más pasta de dientes. Luego, se corta cada tubo y la cantidad restante se pesa, dando lugar a los siguientes datos (en consistencia con lo que informaba el artículo citado): 0.53, 0.65, 0.46, 0.50, 0.37. ¿Será que la cantidad promedio restante real es inferior a 10% del contenido neto anunciado?

a. Compruebe la validez de los supuestos necesarios para probar la hipótesis apropiada.

b. Lleve a cabo una prueba de las hipótesis adecuadas con un nivel de significancia de 0.05. ¿Cambiaría su conclusión si se ha utilizado un nivel de significancia de 0.01?

c. Describa en contexto los errores tipo I y II, y diga qué error podría haberse cometido para llegar a una conclusión.

37. Los datos que acompañan la fuerza del cubo de compresión (MPa) de probetas de hormigón apareció en el artículo **"Experimental Study of Recycled Rubber-Filled High-Strength Concrete" (*Magazine of Concrete Res*., 2009: 549-556)**:

112.3	97.0	92.7	86.0	102.0
99.2	95.8	103.5	89.0	86.7

a. ¿Es posible que la resistencia a la compresión para este tipo de concreto tenga una distribución normal?

b. Suponga que el concreto se utilizará para una aplicación particular, a menos que haya una fuerte evidencia de que la fuerza promedio real es inferior a 100 MPa. ¿Podría utilizarse el concreto? Realice una prueba de hipótesis adecuada.

38. Se obtuvo una muestra aleatoria de especímenes de suelo, se determinó la cantidad (%) de materia orgánica presente en el suelo por cada espécimen y se obtuvieron los siguientes datos **(tomados de "Engineering Properties of Soil", *Soil Science*, 1998: 93-102)**.

1.10	5.09	0.97	1.59	4.60	0.32	0.55	1.45
0.14	4.47	1.20	3.50	5.02	4.67	5.22	2.69
3.98	3.17	3.03	2.21	0.69	4.47	3.31	1.17
0.76	1.17	1.57	2.62	1.66	2.05		

Los valores de la media muestral, la desviación estándar muestral y el error estándar (estimado) de la media son 2.481, 1.616 y 0.295, respectivamente. ¿Sugieren estos datos que el porcentaje promedio verdadero de materia orgánica presente en el suelo es algún otro diferente de 3%? Realice una prueba de la hipótesis apropiada a un nivel de significancia de 0.10. ¿Sería diferente su conclusión si se hubiera usado $\alpha = 0.05$? [*Nota*: Una gráfica de probabilidad normal de los datos muestra un patrón aceptable a la luz del tamaño de la muestra razonablemente grande.]

39. Reconsidere los datos del ejemplo que muestran la proporción de gastos (%) de los fondos de crecimiento de gran capitalización mutua presentados por primera vez en el ejercicio 1.53.

0.52	1.06	1.26	2.17	1.55	0.99	1.10	1.07	1.81	2.05
0.91	0.79	1.39	0.62	1.52	1.02	1.10	1.78	1.01	1.15

Una gráfica de probabilidad normal muestra un patrón bastante lineal.

a. ¿Hay pruebas de peso para concluir que la media poblacional de la proporción de gastos excede 1%? Lleve a cabo una prueba de las hipótesis pertinentes con un nivel de significancia de 0.01.

b. Volviendo al inciso a), describa en contexto el tipo de errores I y II y diga qué error podría haberse cometido para llegar a su conclusión. La fuente de la cual se obtuvieron los datos informó que $\mu = 1.33$ para la población de los 762 fondos. Así que, ¿habrá cometido un error al llegar a su conclusión?

c. Suponiendo que $\sigma = 0.5$, determine e interprete la potencia de la prueba en el inciso (a) para el valor real de μ indicado en el inciso b).

40. Los materiales poliméricos compuestos han ganado popularidad porque tienen alta resistencia a las relaciones de peso y su fabricación es relativamente fácil y barata. Sin embargo, su naturaleza no degradable impulsó el desarrollo de materiales compuestos ecológicos con materiales naturales. El artículo **"Properties of Waste Silk Short Fiber/Cellulose Green Composite Films" (*J. of Composite Materials*, 2012: 123-127)** informa que para una muestra de 10 ejemplares con contenido de fibra de 2%, la media de la resistencia a la tracción (MPa) fue de 51.3 y la desviación estándar fue de 1.2. Suponga que se sabe que la fuerza promedio verdadera para 0% fibras (celulosa pura) es 48 MPa. ¿Proporcionarán los datos evidencias convincentes para concluir que la fuerza promedio verdadera para el compuesto de FSM/celulosa supera este valor?

EJERCICIOS SUPLEMENTARIOS (41–55)

41. Una muestra de 50 lentes utilizados en anteojos aporta un espesor medio muestral de 3.05 mm y una desviación estándar muestral de 0.34 mm. El espesor promedio verdadero deseado de los lentes es de 3.20 mm. ¿Sugerirán fuertemente los datos que el espesor promedio verdadero de los lentes es algún otro diferente del deseado? Haga la prueba con $\alpha = 0.05$.

42. En el ejercicio 57 suponga que el experimentador creía antes de recopilar los datos que el valor de σ era aproximadamente de 0.30. Si el experimentador deseó que la probabilidad de un error tipo II fuera de 0.05 cuando $\mu = 3.00$, ¿era innecesariamente grande un tamaño de muestra 50?

43. Se especificó que cierto tipo de hierro debía contener 0.85 g de silicio por cada 100 g de hierro (0.85%). Se determinó el contenido de silicio de cada uno de los 25 especímenes seleccionados al azar y se obtuvieron los siguientes resultados con Minitab, a partir de una prueba de las hipótesis apropiadas.

```
Variable   N    Mean   StDev  SE Mean     T     P
sil cont  25  0.8880  0.1807   0.0361  1.05  0.30
```

a. ¿Cuáles hipótesis se probaron?

b. ¿A qué conclusión llegaría con un nivel de significancia de 0.05 y por qué? Responda la misma pregunta para un nivel de significancia de 0.10.

44. Un método de enderezar alambre antes de enrollarlo para fabricar resortes se llama "enderezado con rodillos". El artículo **"The Effect of Roller and Spinner Wire Straightening on Coiling Performance and Wire Properties" (*Springs*, 1987: 27–28)** reporta sobre las propiedades de la tensión del alambre. Suponga que se selecciona una muestra de 16 alambres y cada uno se somete a prueba para determinar su resistencia a la tensión (N/mm^2). La media y la desviación estándar muestrales resultantes son 2160 y 30, respectivamente.

a. La resistencia media a la tensión de resortes hechos mediante una máquina enderezadora rotatoria es de 2150 N/mm^2. ¿Qué hipótesis deberán ser probadas para determinar si la resistencia media a la tensión del método de rodillos excede de 2150?

b. Suponiendo que la distribución de la resistencia a la tensión es aproximadamente normal, ¿qué estadístico de prueba utilizaría para probar la hipótesis del inciso a)?

c. ¿Cuál es el valor del estadístico de prueba con estos datos?

d. ¿Cuál es el valor P con el valor del estadístico de prueba calculado en el inciso c)?

e. Con una prueba de nivel 0.05, ¿a qué conclusión llegaría?

45. La contaminación de los suelos de minas en China es un grave problema ambiental. El artículo **"Heavy Metal Contamination in Soils and Phytoaccumulation in a Manganese Mine Wasteland, South China" (*Air, Soil, and Water Res.*, 2008: 31–41)** informa que, para una muestra de 3 especímenes de suelo de cierta área restaurada de minería, la media de la concentración total de la muestra de Cu fue 45.31 mg/kg, con un correspondiente error estándar (estimado) de la media de 5.26. Se menciona también que el valor histórico en China de esta concentración fue de 20. Los resultados de diversas pruebas estadísticas descritas en el artículo se basan en el supuesto de normalidad.

a. ¿Proporcionarán los datos una fuerte evidencia para concluir que la concentración real promedio en la región de la muestra supera el valor histórico planteado? Lleve a cabo una prueba al nivel de significancia 0.01 utilizando el método de valor P. ¿Le sorprende el resultado? Explique.

b. Volviendo a la prueba del inciso a), ¿qué tan probable es que el valor P sea al menos 0.01 cuando la concentración promedio real es de 50 y la verdadera desviación estándar de la concentración es de 10?

46. El artículo **"Orchard Floor Management Utilizing Soil Applied Coal Dust for Frost Protection" (*Agri. and Forest Meteorology*, 1988: 71–82)** reporta los siguientes valores de flujo de calor a través del suelo de ocho solares cubiertos con polvo de hulla.

34.7 35.4 34.7 37.7 32.5 28.0 18.4 24.9

El flujo de calor medio a través del suelo en solares cubiertos sólo con césped es de 29.0. Suponiendo que la distribución del flujo de calor es aproximadamente normal, ¿sugieren los datos que el polvo de hulla es eficaz para incrementar el flujo medio de calor sobre el del césped? Pruebe las hipótesis apropiadas con $\alpha = 0.05$.

47. El artículo **"Caffeine Knowledge, Attitudes, and Consumption in Adult Women" (*J. of Nutrition Educ.*, 1992: 179–184)** reporta los siguientes datos sobre consumo diario de cafeína para una muestra de mujeres adultas: $n = 47$, $\bar{x} = 215$ mg, $s = 235$ mg y rango 5–1176.

a. ¿Parece posible que la distribución de la población de consumo diario de cafeína sea normal? ¿Es necesario suponer una distribución de población normal para probar hipótesis acerca del valor del consumo medio de la población? Explique su razonamiento.

b. Suponga que antes se creía que el consumo medio era cuando mucho de 200 mg. ¿Contradicen los datos dados esta creencia previa? Pruebe las hipótesis apropiadas a nivel de significancia de 0.10.

48. El volumen de negocios anual de un fondo de inversión es el porcentaje de los activos de un fondo que se venden en un año determinado. En términos generales, un fondo con un valor bajo de volumen de negocios es más estable y contrario al riesgo, mientras que un valor alto de volumen de negocios indica una cantidad sustancial de compra y venta en un intento de tomar ventaja de las fluctuaciones del mercado a corto plazo. Los siguientes son los valores del volumen de negocios para una muestra de 20 fondos de gran capitalización combinados (véase el ejercicio 1.53 para un poco más de información), extraídos de **Morningstar.com**:

1.03 1.23 1.10 1.64 1.30 1.27 1.25 0.78 1.05 0.64
0.94 2.86 1.05 0.75 0.09 0.79 1.61 1.26 0.93 0.84

a. ¿Usaría la prueba t de una muestra para decidir si existen pruebas convincentes para concluir que la población media del volumen de negocios es inferior a 100%? Explique.

b. Una gráfica de probabilidad normal del 20 ln(el volumen de negocios) de los valores muestra un patrón lineal muy pronunciado, lo que sugiere que es razonable suponer que la distribución del volumen de negocios es lognormal. Recuerde que X tiene una distribución logarítmica normal si $\ln(X)$ es una distribución normal con media μ y varianza σ^2. Puesto que μ también es la mediana de la distribución $\ln(X)$, e^{μ} es la mediana de la distribución X. Utilice esta información para decidir si existen pruebas convincentes para concluir que la mediana de la distribución de la población del volumen de negocios es inferior a 100%.

49. Se supone que la resistencia a la ruptura promedio verdadera de aislantes de cerámica de cierto tipo es al menos de 10 lb/pulg^2. Se utilizará para una aplicación particular a menos que los datos muestrales indiquen concluyentemente que esta especificación no ha sido satisfecha. Una prueba de hipótesis con $\alpha = 0.01$ tiene que basarse en una muestra aleatoria de diez aislantes. Suponga que la distribución de resistencia a la ruptura es normal con desviación estándar desconocida.

a. Si la desviación estándar verdadera es de 0.80, ¿qué tan probable es que los aislantes sean juzgados satisfactorios cuando la resistencia a la ruptura promedio verdadera es en realidad de sólo 9.5? ¿Y cuando sólo es de 9.0?

b. ¿Qué tamaño de muestra sería necesario para tener 75% de posibilidad de detectar que la resistencia a la ruptura promedio verdadera es de 9.5 cuando la desviación estándar verdadera es de 0.80?

50. Las siguientes observaciones sobre el tiempo de permanencia del fuego (segundos) en tiras de ropa de dormir para niños tratada aparecieron en el artículo **"An Introduction to Some Precision and Accuracy of Measurement Problems" (*J. of Testing and Eval*., 1982: 132–140)**. Suponga que se asignó por encargo un tiempo de permanencia del fuego promedio verdadero de cuando mucho 9.75. ¿Sugieren los datos que esta condición no se ha cumplido? Realice una prueba apropiada después de investigar la factibilidad de las suposiciones que fundamentan su método de inferencia.

9.85	9.93	9.75	9.77	9.67	9.87	9.67
9.94	9.85	9.75	9.83	9.92	9.74	9.99
9.88	9.95	9.95	9.93	9.92	9.89	

51. Se cree que la incidencia de un tipo de cromosoma defectuoso en la población de varones adultos estadounidenses es de 1 en 75. Una muestra aleatoria de 800 individuos en instituciones penitenciarias estadounidenses revela que 16 tienen tales defectos. ¿Se puede concluir que la proporción de incidencia de este defecto entre los prisioneros difiere de la proporción supuesta para toda la población de varones adultos?

a. Formule y pruebe las hipótesis pertinentes con $\alpha = 0.05$. ¿Qué tipo de error podría haber cometido al llegar a una conclusión?

b. ¿Qué valor P está asociado con esta prueba? Basado en este valor P, ¿podría H_0 ser rechazada a un nivel de significancia de 0.20?

52. En una investigación de la toxina producida por una serpiente venenosa, un investigador preparó 26 frascos, cada uno con 1 g de la toxina y luego determinó la cantidad de antitoxina necesaria para neutralizar la toxina. Se encontró que la cantidad promedio muestral de antitoxina necesaria era de 1.89 mg y la desviación estándar muestral era de 0.42. Una investigación previa indicó que la cantidad neutralizante promedio verdadera fue de 1.75 mg/g de toxina. ¿Contradicen estos datos nuevos el valor sugerido por la investigación previa? Pruebe la hipótesis pertinente. ¿Dependerá la validez de su análisis de cualquier suposición sobre la distribución de la población de cantidad neutralizante? Explique.

53. La resistencia a la compresión no restringida promedio muestral de 45 especímenes de un tipo particular de ladrillos resultó ser de 3107 lb/pulg^2 y la desviación estándar muestral fue de 188. La distribución de la resistencia a la compresión no restringida puede ser un tanto asimétrica. ¿Indican los resultados fuertemente que la resistencia a la compresión no restringida promedio verdadera es menor que el valor de diseño de 3200? Haga la prueba con $\alpha = 0.001$.

54. El 30 de diciembre de 2009 el *New York Times* informó que en una encuesta de 948 adultos estadounidenses que señalaron estar al menos un poco interesados en el futbol universitario, 597 dijeron que el actual Bowl Championship System (B.C.S) debería ser sustituido por una eliminatoria similar a la utilizada en el baloncesto universitario. ¿Aportará esto pruebas convincentes para concluir que la mayoría de todos los individuos están a favor de la sustitución del B.C.S. con una eliminatoria? Pruebe la hipótesis apropiada utilizando el nivel de significancia de 0.001.

55. Cuando $X_1, X_2, \ldots, X_n$ son variables de Poisson independientes, cada una con parámetro μ, y la n es grande, la media muestral $\overline{X}$ tiene aproximadamente una distribución normal con $\mu = E(\overline{X})$ y $V(\overline{X}) = \mu/n$. Esto implica que

$$Z = \frac{\overline{X} - \mu}{\sqrt{\mu/n}}$$

tiene aproximadamente una distribución normal estándar. Para probar H_0: $\mu = \mu_0$, se puede reemplazar μ con μ_0 en la ecuación para Z a fin de obtener un estadístico de prueba. Normalmente se prefiere este estadístico al estadístico de muestra grande con denominador $S/\sqrt{n}$ (cuando las X_i son de Poisson) porque está explícitamente hecho a la medida de la suposición de Poisson. Si el número de solicitudes de consultoría recibidas por cierto estadístico durante una semana de trabajo de 5 días tiene una distribución de Poisson, y el número total de solicitudes de consultoría durante 36 semanas es de 160, ¿sugiere esto que el número promedio verdadero de solicitudes semanales excede de 4.0? Haga la prueba con $\alpha = 0.02$.

BIBLIOGRAFÍA

Véanse las bibliografías al final de los capítulos 5 y 6.

Capítulo

8

Análisis de la varianza

INTRODUCCIÓN

Al estudiar los métodos de análisis de datos cuantitativos primero se trataron problemas que implican una sola muestra de números y luego se abordó el análisis comparativo de dos muestras diferentes. En problemas de una muestra, los datos se componían de observaciones sobre los individuos u objetos experimentales seleccionados de una sola población o sus respuestas. En problemas de dos muestras, estas se tomaron de dos poblaciones diferentes y los parámetros de interés fueron las medias de población o bien se aplicaron dos tratamientos distintos a las unidades experimentales (individuos u objetos) seleccionadas de una sola población; en el último caso, los parámetros de interés fueron las medias de tratamiento verdaderas.

El **análisis de la varianza** o, más brevemente, **ANOVA** se refiere en general a un conjunto de situaciones experimentales y procedimientos estadísticos para el análisis de respuestas cuantitativas de unidades experimentales. El problema ANOVA más simple se conoce indistintamente como **unifactorial**, de **clasificación única** o **ANOVA unidireccional** e implica el análisis de los datos muestreados de más de dos poblaciones (distribuciones) numéricas o de datos de experimentos en los cuales se utilizaron más de dos tratamientos. La característica que diferencia los tratamientos de las poblaciones se llama **factor** en estudio y los distintos tratamientos o poblaciones se conocen como **niveles** del factor. Entre los ejemplos de tales situaciones se incluyen los siguientes:

1. Un experimento para estudiar los efectos de cinco marcas diferentes de gasolina respecto a la eficiencia de operación de un motor automotriz (mpg)
2. Un experimento para estudiar los efectos de la presencia de cuatro soluciones azucaradas diferentes (glucosa, sacarosa, fructosa y una mezcla de las tres) en cuanto a crecimiento de bacterias

3. Un experimento para investigar si la concentración de madera dura en la pulpa (%) afecta la resistencia a la tensión de las bolsas hechas de la misma pulpa a tres diferentes niveles de impacto

4. Un experimento para decidir si la densidad de color de un espécimen de tela depende de la cantidad de tinte utilizado

En el inciso 1) el factor de interés es la marca de la gasolina y existen cinco niveles diferentes del factor. En el inciso 2) el factor es el azúcar con cuatro niveles (o cinco, si se utiliza una solución de control que no contenga azúcar). Tanto en 1) como en 2) el factor es de naturaleza cualitativa y los niveles corresponden a posibles categorías del factor. En los incisos 3) y 4) los factores son la concentración de madera dura y la cantidad de tinte, respectivamente; estos dos factores son de naturaleza cuantitativa, por tanto, los niveles identifican diferentes ajustes del factor. Cuando el factor de interés es cuantitativo también se pueden utilizar técnicas estadísticas de análisis de regresión.

Este capítulo se enfoca en el ANOVA unifactorial. La sección 8.1 presenta la prueba F para demostrar la hipótesis nula de que las medias de población o tratamiento son idénticas. La sección 8.2 considera un análisis adicional de los datos cuando H_0 ha sido rechazada.

8.1 ANOVA unifactorial

ANOVA unifactorial se centra en la comparación de más de dos medias de población o tratamiento. Sean

I = el número de poblaciones o tratamientos que se están comparando

μ_1 = la media de población 1 o la respuesta promedio verdadera cuando se aplica el tratamiento 1

$\vdots$

μ_I = la media de población I o la respuesta promedio verdadera cuando se aplica el tratamiento I

Las hipótesis pertinentes son

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

contra

H_a: al menos dos de las μ_i son diferentes

Si $I = 4$, H_0 es verdadera sólo si las cuatro μ_i son idénticas. H_a sería verdadera, por ejemplo, si $\mu_1 = \mu_2 \neq \mu_3 = \mu_4$, si $\mu_1 = \mu_3 = \mu_4 \neq \mu_2$, o si las cuatro μ_i difirieran una de otra.

Una prueba de estas hipótesis requiere que se tenga disponible una muestra aleatoria de cada población o tratamiento.

EJEMPLO 8.1 El artículo **"Compression of Single–Wall Corrugated Shipping Containers Using Fixed and Floating Test Platens" (*J. Testing and Evaluation*, 1992: 318–320)** describe un experimento en el cual se comparan varios tipos diferentes de cajas respecto a la

resistencia a la compresión (lb). La tabla 8.1 presenta los resultados de un experimento ANOVA unifactorial que implica $I = 4$ tipos de cajas (las medias y las desviaciones estándar muestrales concuerdan con los valores dados en el artículo).

Tabla 8.1 Datos y cantidades resumidas para el ejemplo 8.1

Tipo de caja	Resistencia a la compresión (lb)	Media muestral	DE muestral
1	655.5 788.3 734.3 721.4 679.1 699.4	713.00	46.55
2	789.2 772.5 786.9 686.1 732.1 774.8	756.93	40.34
3	737.1 639.0 696.3 671.7 717.2 727.1	698.07	37.20
4	535.1 628.7 542.4 559.0 586.9 520.0	562.02	39.87
	Media grande =	682.50	

Con μ_i denotando la resistencia a la compresión promedio verdadera de las cajas de tipo i ($i = 1, 2, 3, 4$), la hipótesis nula es H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. La figura 8.1(a) muestra una gráfica de caja comparativa para las cuatro muestras. Existe una cantidad sustancial de traslape entre las observaciones de los primeros tres tipos de cajas, pero las resistencias a la compresión del cuarto tipo parecen considerablemente más pequeñas que para los demás tipos. Esto sugiere que H_0 no es verdadera. La gráfica de caja comparativa que aparece en la figura 8.1(b) está basada en agregar 120 a cada observación en la cuarta muestra (y así se obtiene una media de 682.02 y la misma desviación estándar) y las demás observaciones no cambian. Ya no es obvio si H_0 es verdadera o falsa. En situaciones como esta, se requiere un procedimiento de prueba formal.

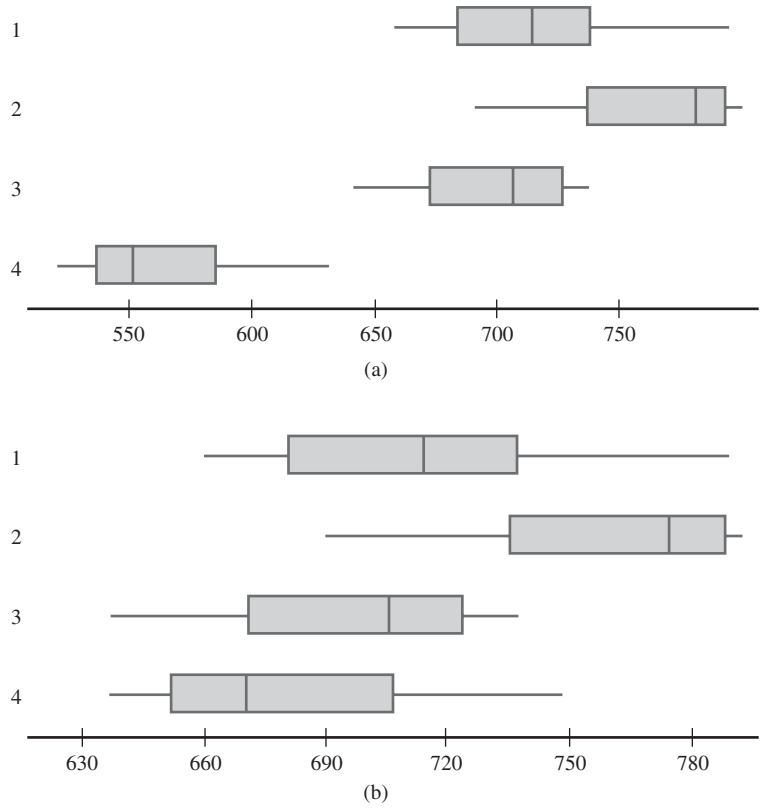


Figura 8.1 Gráficas de caja para el ejemplo 8.1: (a) datos originales; (b) datos modificados

Notación y suposiciones

En los problemas de dos muestras se utilizaron las letras X y Y para diferenciar las observaciones en una de ellas en la otra. Como esto es pesado con tres o más muestras, se acostumbra utilizar una sola letra con dos subíndices. El primero identifica el número de la muestra, correspondiente a la población o al tratamiento que se está muestreando y el segundo denota la posición de la observación dentro de dicha muestra. Sean

$X_{i,j}$ = la variable aleatoria (va) que denota la medición $j-$ésima tomada en la población $i-$ésima, o la medición tomada en la $j-$ésima unidad experimental que recibe el tratamiento $i-$ésimo.

$x_{i,j}$ = el valor observado de $X_{i,j}$ cuando se realiza el experimento

Los datos observados normalmente se muestran en una tabla rectangular, tal como la tabla 8.1. En ella las muestras de las diferentes poblaciones aparecen en filas distintas de la tabla, y $x_{i,j}$ es el número $j-$ésimo en la fila $i-$ésima. Por ejemplo, $x_{2,3} = 786.9$ (la tercera observación de la segunda población) y $x_{4,1} = 535.1$. Cuando no hay ambigüedad, se escribirá x_{ij} en lugar de $x_{i,j}$ (p. ej., si se realizaron 15 observaciones en cada uno de 12 tratamientos, x_{112} podría significar $x_{1,12}$ o $x_{11,2}$). Se supone que las X_{ij} dentro de cualquier muestra particular son independientes; una muestra aleatoria de la distribución de población o tratamiento $i-$ésima, y que las diferentes muestras son independientes entre sí.

En algunos experimentos diferentes muestras contienen distintos números de observaciones. Aquí se abordará el caso de tamaños de muestra iguales; la generalización en cuanto a tamaños de muestra desiguales requiere un estudio más profundo. Sea J el número de observaciones en cada muestra ($J = 6$ en el ejemplo 8.1). El conjunto de datos se compone de IJ observaciones. Las medias de muestra individual serán denotadas por $\overline{X}_{1\cdot}, \overline{X}_{2\cdot}, \ldots, \overline{X}_{I\cdot}$. Es decir,

$$\overline{X}_{i\cdot} = \frac{\sum_{j=1}^{J} X_{ij}}{J} \quad i = 1, 2, \ldots, I$$

El punto en lugar del segundo subíndice significa que se sumaron todos los valores de dicho subíndice al mismo tiempo que se mantuvo fijo el valor del otro subíndice; y la raya horizontal indica división entre J para obtener un promedio. Asimismo, el promedio de todas las observaciones IJ, llamada **media grande,** es

$$\overline{X}_{\cdot\cdot} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}}{IJ}$$

Con los datos en la tabla 8.1, $\overline{x}_{1\cdot} = 713.00$, $\overline{x}_{2\cdot} = 756.93$, $\overline{x}_{3\cdot} = 698.07$, $\overline{x}_{4\cdot} = 562.02$ y $\overline{x}_{\cdot\cdot} = 682.50$. Además, sean $S_1^2, S_2^2, \ldots, S_I^2$ las varianzas muestrales:

$$S_i^2 = \frac{\sum_{j=1}^{J} (X_{ij} - \overline{X}_{i\cdot})^2}{J - 1} \quad i = 1, 2, \ldots, I$$

De acuerdo con el ejemplo 8.1, $s_1 = 46.55$, $s_1^2 = 2166.90$, etcétera.

SUPOSICIONES

Las distribuciones de población o tratamiento I son normales con la misma varianza σ^2. Es decir, cada X_{ij} está normalmente distribuida con

$$E(X_{ij}) = \mu_i \quad V(X_{ij}) = \sigma^2$$

Las desviaciones estándar de la muestra I en general difieren un poco aun cuando las σ correspondientes sean idénticas. En el ejemplo 8.1, la más grande entre s_1, s_2, s_3 y s_4 es aproximadamente 1.25 veces la más pequeña. Una regla empírica preliminar es que si la s más grande no es mucho más de dos veces la más pequeña, es razonable suponer σ^2 iguales.

En capítulos previos se sugirió un diagrama de probabilidad normal para verificar la normalidad. Típicamente los tamaños de muestra individuales en ANOVA son demasiado pequeños como para que los distintos diagramas I sean informativos. Se puede construir un solo diagrama restando x_1 (escribir la ecuación de la media muestral 1) en la primera muestra, restando x_2 (escribir la ecuación de la media muestral 2) en la segunda y así sucesivamente, y luego graficando estas desviaciones IJ contra los percentiles z. La figura 8.2 muestra ese diagrama para los datos del ejemplo 8.1. La linealidad del patrón confirma fuertemente la suposición de normalidad.

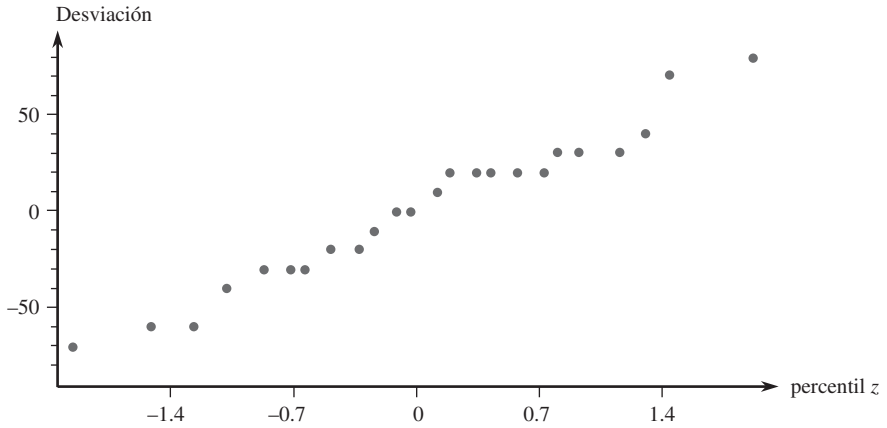


Figura 8.2 Diagrama de probabilidad normal con base en los datos del ejemplo 8.1

Si la suposición de normalidad o la suposición de varianzas iguales se juzgo no razonable, habrá que emplear un método de análisis distinto de la prueba F usual. Busque, por favor, asesoría experta en tales situaciones (en la sección 8.3 se sugiere una posibilidad, una transformación de datos, y en la sección 15.4 se desarrolla otra alternativa).

El estadístico de prueba

Si H_0 es verdadera, las J observaciones en cada muestra provienen de una distribución normal con la *misma* media μ, en cuyo caso las medias muestrales deberán ser razonablemente parecidas. El procedimiento de prueba se basa en comparar una medida de diferencias entre las $\overline{x}_{i\cdot}$ (variación "entre muestras") con una medida de variación calculada desde *dentro* de cada una de las muestras.

DEFINICIÓN

La **media de los cuadrados del tratamiento** (MSTr, por sus siglas en inglés) está dada, por

$$\text{MSTr} = \frac{J}{I-1}[(\overline{X}_{1\cdot} - \overline{X}_{\cdot\cdot})^2 + (\overline{X}_{2\cdot} - \overline{X}_{\cdot\cdot})^2 + \cdots + (\overline{X}_{I\cdot} - \overline{X}_{\cdot\cdot})^2]$$
$$= \frac{J}{I-1}\sum_i (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2$$

y la media de los cuadrados del error (MSE, por sus siglas en inglés) es

$$\text{MSE} = \frac{S_1^2 + S_2^2 + \cdots + S_I^2}{I}$$

El estadístico de prueba para ANOVA unifactorial es $F = \text{MSTr/MSE}$.

La terminología "media de los cuadrados" se explicará más adelante. Observe que se utilizan $\overline{X}$ y S^2 mayúsculas, de modo que MSTr y MSE se definen como estadísticos. Se seguirá la tradición y también se utilizarán MSTr y MSE (en lugar de mstr y mse) para denotar los valores calculados de estos estadísticos. Cada S_i^2 evalúa la variación dentro de una muestra particular, así que MSE es una medida de variación dentro de las muestras.

¿Qué clase de valor de F proporciona evidencia en pro o en contra de H_0? Si H_0 es verdadera (todas las μ_i son iguales), los valores de las medias muestrales individuales deberán estar próximos entre sí y, por consiguiente, próximos a la media grande con el resultado de un valor relativamente pequeño de MSTr. Sin embargo, si las μ_i son bastante diferentes, algunas difieren un poco de $\overline{x}\ldots$ De modo que el valor de MSTr es afectado por la condición de H_0 (verdadera o falsa). Este no es el caso con MSE, porque las S_i^2 dependen sólo del valor subyacente de σ^2 y no de en dónde están centradas las diversas distribuciones. El siguiente recuadro presenta una propiedad importante de E(MSTr) y E(MSE), los valores esperados de estos dos estadísticos.

PROPOSICIÓN

Cuando H_0 es verdadera,

$$E(\text{MSTr}) = E(\text{MSE}) = \sigma^2$$

mientras que cuando H_0 es falsa,

$$E(\text{MSTr}) > E(\text{MSE}) = \sigma^2$$

Es decir, ambos estadísticos son insesgados para estimar la varianza de población común σ^2 cuando H_0 es verdadera, pero MSTr tiende a sobrestimar σ^2 cuando H_0 es falsa.

La insesgadez de MSE es una consecuencia de $\text{E}(S_i^2) = \sigma^2$ si H_0 es verdadera o falsa. Cuando H_0 es verdadera, cada $\overline{X}_i$ tiene la misma media μ y varianza σ^2/J de modo que $\Sigma(\overline{X}_i - \overline{X}\cdot\cdot)^2/(I-1)$, la "varianza muestral" de las $\overline{X}_i$, estima σ^2/J insesgadamente; al multiplicar esta por J se obtiene MSTr como un estimador insesgado de σ^2 misma. Las $\overline{X}_i$ tienden a dispersarse más cuando H_0 es falsa que cuando es verdadera y, en este caso, tiende a inflar el valor de MSTr. Por consiguiente, un valor de F que excede en gran medida 1, correspondiente a un MSTr mucho más grande que MSE, provoca una duda considerable sobre H_0. Para determinar el valor P se requiere conocer la distribución de F cuando H_0 es verdadera.

Distribuciones *F* y la prueba *F*

Una distribución F surge en conexión con una proporción en la cual existe un número de grados de libertad (gl) asociado con el numerador, y otro número de grados de libertad asociado con el denominador. Sean ν_1 y ν_2 el número de grados de libertad asociados con el numerador y el denominador, respectivamente, para una variable con una distribución F. Tanto ν_1 como ν_2 son enteros positivos. La figura 8.3 ilustra una curva de densidad

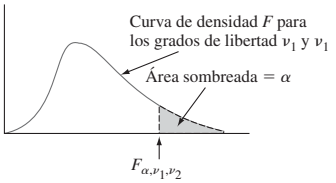


Figura 8.3 Curva de densidad F y valor crítico F_{α,ν_1,ν_2}.

F y el valor crítico de cola superior correspondiente F_{α,ν_1,ν_2}. La tabla A.9 del apéndice da estos valores críticos para $\alpha = 0.10, 0.05, 0.01$ y 0.001. Los valores de ν_1 están identificados con diferentes columnas de la tabla y las filas con varios valores de ν_2. Por ejemplo, el valor crítico F que captura un área de cola superior de 0.05 bajo la curva F con $\nu_1 = 4$ y $\nu_2 = 6$ es $F_{0.05,4,6} = 4.53$ en tanto que $F_{0.05,6,4} = 6.16$. El resultado teórico clave es que el estadístico de prueba F tiene una distribución F cuando H_0 es verdadera.

TEOREMA

Sea $F = \text{MSTr/MSE}$ el estadístico de prueba en un problema ANOVA unifactorial que implica poblaciones o tratamientos I con una muestra aleatoria de J observaciones de cada uno. Cuando H_0 es verdadera y las suposiciones básicas de esta sección se satisfacen, F tiene una distribución F con $\nu_1 = I - 1$ y $\nu_2 = I(J - 1)$. Puesto que un f más grande es más contradictorio a H_0 que un f más pequeño, la prueba es de cola superior:

$$\begin{aligned}\text{Valor } P &= P(F \geq f \text{ cuando } H_0 \text{ es verdadera})\\ &= \text{área bajo la curva } F_{I-1,I(J-1)} \text{ a la derecha de } f\end{aligned}$$

El software estadístico proporcionará un valor P exacto. Consulte la sección 9.5 para una descripción de cómo se puede utilizar la tabla de nuestro libro de valores críticos de F, tabla A.9, para obtener un límite superior o inferior (o ambos) sobre el valor P.

El razonamiento para $\nu_1 = I - 1$ es que aunque MSTr está basada en las desviaciones I, $\overline{X}_{1\cdot} - \overline{X}_{\cdot\cdot}$, ..., $\overline{X}_{I\cdot} - \overline{X}_{\cdot\cdot}$, $\Sigma(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}) = 0$, de modo que sólo $I - 1$ de estas son libremente determinadas. Debido a que cada muestra contribuye con $J - 1$ grados de libertad a MSE y estas muestras son independientes, $\nu_2 = (J - 1) + \cdots + (J - 1) = I(J - 1)$.

EJEMPLO 8.2 (Continuación del ejemplo 8.1)

Los valores de I y J con los datos de resistencia son 4 y 6, respectivamente, de modo que $\text{gl} = I - 1 = 3$ asociados con el numerador y $\text{gl} = I(J - 1) = 20$ asociados con el denominador. La media grande es $\overline{x}_{\cdot\cdot} = \Sigma\Sigma x_{ij}/(IJ) = 682.50$,

$$\begin{aligned}\text{MSTr} &= \frac{6}{4-1}[(713.00 - 682.50)^2 + (756.93 - 682.50)^2\\ &\qquad + (698.07 - 682.50)^2 + (562.02 - 682.50)^2] = 42\,455.86\end{aligned}$$

$$\text{MSE} = \frac{1}{4}[(46.55)^2 + (40.34)^2 + (37.20)^2 + (39.87)^2] = 1691.92$$

$$f = \text{MSTr/MSE} = 42\,455.86/1691.92 = 25.09$$

El mayor valor crítico de F en la tabla A.9 para $\nu_1 = 3$, $\nu_2 = 20$ es $F_{0.001,3,20} = 8.10$. Puesto que $f = 25.09 > 8.10$, el área debajo de la curva $F_{3,20}$ a la derecha de 25.09 es menor a 0.001. Por tanto, el valor $P \leq 0.05$, así que la hipótesis nula H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ es rechazada en el nivel de significancia de 0.05. La resistencia a la compresión promedio verdadera sí parece depender del tipo de caja. De hecho, ya que el valor P es tan pequeño H_0 sería rechazada a cualquier nivel de significancia razonable. ■

EJEMPLO 8.3

El artículo **"Influence of Contamination and Cleaning on Bond Strength to Modified Zirconia" (*Dental Materials*, 2009: 1541–1550)** informa sobre un experimento en el que 50 discos de óxido de circonio se dividieron en cinco grupos de 10 cada uno. A continuación, para cada grupo se utilizó un protocolo diferente de contaminación/limpieza. En el artículo aparece el siguiente resumen de datos sobre la resistencia adhesiva al corte (MPa):

Tratamiento:	1	2	3	4	5	
Media de la muestra	10.5	14.8	15.7	16.0	21.6	Media grande = 15.7
DE muestral	4.5	6.8	6.5	6.7	6.0	

Los autores del artículo citado utilizan la prueba F, por lo que esperamos que hayan examinado una gráfica de probabilidad normal de las desviaciones (o una gráfica independiente de cada muestra, ya que cada tamaño de muestra es de 10) para comprobar la factibilidad de suponer tratamiento normal de las distribuciones respuesta. Las desviaciones estándar de cinco muestras son, sin duda, lo suficientemente cercanas unas de otras para apoyar la hipótesis de σ iguales.

1. μ_i = resistencia adhesiva al corte promedio verdadera para el protocolo i (i = 1, 2, 3, 4, 5).
2. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (la resistencia adhesiva al corte promedio verdadera no depende de cuál protocolo se use)
3. H_a: al menos dos de las μ_i son diferentes
4. El valor del estadístico de prueba es f = MSTr/MSE
5. Los grados de libertad del numerador y del denominador son $I - 1 = 4$ e $I(J - 1) = 5(9) = 45$. Las medias de los cuadrados son

$$\begin{aligned}\text{MSTr} &= \frac{10}{5-1}[(10.5-15.7)^2+(14.8-15.7)^2+(15.7-15.7)^2\\ &\qquad +(16.0-15.7)^2+(21.6-15.7)^2]\\ &= 156.875\\ \text{MSE} &= [(4.5)^2+(6.8)^2+(6.5)^2+(6.7)^2+(6.0)^2]/5 = 37.926\end{aligned}$$

   Así el valor del estadístico de prueba es $f = 156.875/37.926 = 4.14$.
6. La tabla A.9 da $F_{0.01,4,40} = 3.83$, $F_{0.01,4,50} = 3.72$, $F_{0.001,4,40} = 5.70$ y $F_{0.001,4,50} = 5.46$. Por tanto, $F_{0.01,4,45} \approx 3.77$ y $F_{0.001,4,45} \approx 5.56$. Debido a que $f = 4.14$ está entre estos últimos dos valores críticos, el área debajo de la curva $F_{4,45}$ a la derecha de 4.14 (es decir, el valor P) está entre 0.001 y 0.01 (con el software se obtiene 0.0061).
7. Puesto que el valor $P < 0.01$, la hipótesis nula debe rechazarse en este nivel de significancia. La resistencia adhesiva al corte promedio verdadera parece depender del protocolo que se utiliza. ■

Cuando la hipótesis nula es rechazada por la prueba de F, como ocurrió en los ejemplos 8.2 y 8.3, el experimentador suele estar interesado en el análisis posterior de los datos para decidir cuáles μ_i difieren unas de las otras. Los métodos para hacer esto se llaman *procedimientos de comparación múltiple*, que es el tema de la sección 8.2. El artículo citado en el ejemplo 8.3 resume los resultados de dicho análisis.

Sumas de cuadrados

La introducción de las *sumas de cuadrados* facilita el desarrollo de una apreciación intuitiva para el razonamiento que fundamenta los ANOVA unifactoriales y multifactoriales. Sea $x_{i.}$ la *suma* (no el promedio, puesto que no tienen barra) de las x_{ij} con i fija (suma de los números en la $i-$ésima fila de la tabla) y sea que $x..$ denote la suma de *todas* las x_{ij} (el **gran total**).

DEFINICIÓN

La **suma total de cuadrados (SST, por sus siglas en inglés),** la **suma de los cuadrados del tratamiento (SSTr, por sus siglas en inglés)** y la **suma de los cuadrados del error (SSE, por sus siglas en inglés)** están dadas por

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-\bar{x}..)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2 - \frac{1}{IJ}x_{..}^2$$

$$\text{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{i.}-\bar{x}..)^2 = \frac{1}{J}\sum_{i=1}^{I}x_{i.}^2 - \frac{1}{IJ}x_{..}^2$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-\bar{x}_{i.})^2 \quad \text{donde } x_{i.} = \sum_{j=1}^{J}x_{ij} \quad x.. = \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}$$

La suma de los cuadrados SSTr aparece en el numerador de F y SSE lo hace en el denominador de F; la razón para definir la SST se pondrá de manifiesto más adelante.

Las expresiones a la extrema derecha de SST y SSTr son convenientes si los cálculos de ANOVA se realizan a mano, aunque la amplia disponibilidad de programas estadísticos hace que esto sea innecesario. Tanto SST como SSTr implican $x_{..}^2/(IJ)$ (el cuadrado del gran total dividido entre IJ), lo que normalmente se llama **factor de corrección para la media** (FC). Una vez que se calcula el factor de corrección, la SST se obtiene elevando al cuadrado cada número que aparece en la tabla, sumando los cuadrados y restando el factor de corrección. SSTr se obtiene al elevar al cuadrado cada total de fila, sumándolos, dividiendo entre J y restando el factor de corrección. SSE se calcula fácilmente como una consecuencia de la siguiente relación.

Identidad fundamental

$$\text{SST} = \text{SSTr} + \text{SSE} \tag{8.1}$$

Por consiguiente, si se calculan cualesquiera dos de las sumas de los cuadrados, la tercera se obtiene con (8.1); SST y SSTr son las más fáciles de calcular y en ese caso SSE = SST − SSTr. La comprobación se desprende de elevar al cuadrado ambos lados de la relación

$$x_{ij} - \bar{x}_{..} = (\bar{x}_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x}_{..}) \tag{8.2}$$

y sumando todas las i y j. Esto da la SST a la izquierda y SSTr y SSE como los dos términos extremos a la derecha. Es fácil ver que el término del producto cruzado es cero.

La interpretación de la identidad fundamental es una importante ayuda para entender el ANOVA. SST mide la variación total de los datos: la suma de todas las desviaciones al cuadrado respecto a la media grande. La identidad dice que esta variación total puede dividirse en dos partes. SSE mide la variación que estaría presente ((dentro de las filas) aun cuando H_0 fuera verdadera o falsa y es, por consiguiente, la parte de la variación total que *no es explicada* por la veracidad o la falsedad de H_0. SSTr es la cantidad de variación (entre filas) que *puede ser explicada* por las posibles diferencias en las μ_i. H_0 es rechazada si la variación explicada es grande respecto a la variación no explicada.

Una vez que SSTr y SSE son calculadas, cada una se divide por su número de grados de libertad asociado para obtener una media de cuadrados (*media* en el sentido de promedio). Entonces F es la proporción de los dos cuadrados de la media.

$$\text{MSTr} = \frac{\text{SSTr}}{I-1} \quad \text{MSE} = \frac{\text{SSE}}{I(J-1)} \quad F = \frac{\text{MSTr}}{\text{MSE}} \tag{8.3}$$

Con frecuencia, los cálculos se resumen en un formato tabular, llamado **tabla ANOVA,** como se ilustra en la tabla 8.2. Las tablas producidas por programas estadísticos comúnmente incluyen una columna de valor P a la derecha de f.

Tabla 8.2 Tabla ANOVA

Origen de la variación	Grados de libertad	Suma de cuadrados	Media de cuadrados	f
Tratamientos	$I - 1$	SSTr	MSTr = SSTr/($I - 1$)	MSTr/MSE
Error	$I(J - 1)$	SSE	MSE = SSE/[$I(J - 1)$]	
Total	$IJ - 1$	SST		

EJEMPLO 8.4 De acuerdo con el artículo **"Evaluating Fracture Behavior of Brittle Polymeric Materials Using an IASCB Specimen" (*J. of Engr. Manuf.*, 2013: 133–140)**, los investigadores han propuesto recientemente una prueba mayor para la investigación de la resistencia a la fractura de los materiales poliméricos frágiles. Esta nueva prueba de fractura se aplicó para el polímero frágil polimetacrilato de metilo (PMMA), más popularmente conocido como plexiglás, que es ampliamente utilizado en productos comerciales. La prueba se realizó aplicando cargas de flexión asimétricas de tres puntos en especímenes PMMA. Entonces se varió la ubicación de uno de los puntos de carga de los tres para determinar su efecto sobre la carga de fractura. En un experimento las ubicaciones de los 3 puntos de carga se basaron en diferentes distancias desde el centro de la base del espécimen, dando por resultado los siguientes datos de carga de fractura (kN):

						$x_{i\cdot}$
	42 mm:	2.62	2.99	3.39	2.86	11.86
Distancia	36 mm:	3.47	3.85	3.77	3.63	14.72
	31.2 mm:	4.78	4.41	4.91	5.06	19.16
						$x_{\cdot\cdot} = 45.74$

Sea μ_i la media verdadera de la carga de fractura cuando se usa la distancia i ($i = 1, 2, 3$). La hipótesis nula asegura que estas tres μ_i son idénticas, mientras que la hipótesis alternativa afirma que no todas las μ_i son iguales. Antes de usar la prueba de F en el nivel de significancia 0.01 debemos comprobar la factibilidad de las hipótesis subyacentes. Las tres desviaciones estándar de la muestra son 0.322, 0.167 y 0.278, respectivamente. Efectivamente, el mayor de estos tres es no más de dos veces el más pequeño. Por tanto, es factible la suposición de varianzas iguales. La figura 8.4 muestra una gráfica de probabilidad normal de los 12 residuos obtenidos restando la media de cada muestra de las cuatro observaciones de la misma. ¡No forman algo más recto que esto! Es razonable suponer que las distribuciones de carga de tres fracturas son normales.

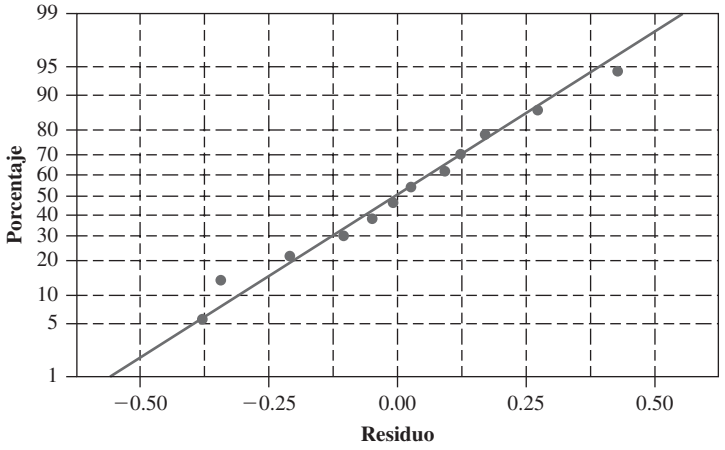


Figura 8.4 Gráfica de probabilidad normal de los residuos del ejemplo 8.4

Elevando al cuadrado cada una de las 12 observaciones y sumando se obtiene $\Sigma\Sigma x_{ij}^2 = (2.62)^2 + \cdots + (5.06)^2 = 181.7376$. Los valores de las tres sumas de cuadrados son

$$\text{SST} = 181.7376 - (45.74)^2/12 = 181.7376 - 174.3456 = 7.3920$$
$$\text{SSTr} = \tfrac{1}{4}[(11.86)^2 + (14.72)^2 + (19.16)^2] - 174.3456 = 6.7653$$
$$\text{SSE} = 7.3920 - 6.7653 = 0.6267$$

La siguiente tabla ANOVA de Minitab resume los cálculos. Con un valor P de 0.000, la hipótesis nula puede ser rechazada en cualquier nivel de significancia razonable y en particular en el nivel elegido 0.01. Existe evidencia convincente para concluir que la carga de fractura promedio verdadera no es la misma para todas las tres distancias.

Source	DF	SS	MS	F	P
Dist	2	6.7653	3.3826	48.58	0.000
Error	9	0.6267	0.0696		
Total	11	7.3920			

EJERCICIOS Sección 8.1 (1–10)

1. En un experimento para comparar las resistencias a la tensión de $I = 5$ tipos diferentes de alambre de cobre, se utilizaron $J = 4$ muestras de cada tipo. Las estimaciones entre las muestras y dentro de las muestras de σ^2 se calcularon como MSTr = 2673.3 y MSE = 1094.2, respectivamente. Use la prueba F a un nivel de 0.05 para probar H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ contra H_a: al menos dos μ_i son desiguales.

2. Suponga que las observaciones de resistencia a la compresión del cuarto tipo de caja del ejemplo 8.1 hubieran sido 655.1, 748.7, 662.4, 679.0, 706.9 y 640.0 (obtenidas sumando 120 a cada previa x_{4j}). Suponiendo que las observaciones restantes no cambian, realice una prueba F con $\alpha = 0.05$.

3. Se determinó el rendimiento en lúmenes de cada uno de los $I = 3$ marcas diferentes de focos de luz blanca de 60 watts, con $J = 8$ focos de cada marca probados. Las sumas de los cuadrados se calcularon como SSE = 4773.3 y SSTr = 591.2. Formule las hipótesis de interés (incluidas definiciones en palabras de los parámetros) y use la prueba F de ANOVA (con $\alpha = 0.05$) para decidir si existen diferencias en los rendimientos de lúmenes promedio verdaderos entre las tres marcas de este tipo de foco, obteniendo tanta información como sea posible sobre el valor P.

4. Es una práctica común en muchos países destruir (fragmentar) refrigeradores al final de su vida útil. En este proceso el material de espuma de aislamiento puede ser liberado a la atmósfera. El artículo **"Release of Fluorocarbons from Insulation Foam in Home Appliances during Shredding" (*J. of the Air and Waste Mgmt. Assoc*., 2007: 1452–1460)** dio los siguientes datos sobre la densidad de la espuma (g/L) para cada uno de dos refrigeradores producidos por cuatro distintos fabricantes:

1. 30.4, 29.2 **2.** 27.7, 27.1
3. 27.1, 24.8 **4.** 25.5, 28.8

¿Será que el promedio real de densidad de la espuma no es el mismo para todos estos fabricantes? Lleve a cabo una prueba adecuada de las hipótesis mediante la obtención de una mayor cantidad de información del valor P como sea posible y un resumen de su análisis en una tabla de ANOVA.

5. Considere los siguientes datos del módulo de elasticidad ($\times 10^6$ lb/pulg^2) de madera de tres grados diferentes (en concordancia con los valores que aparecen en el artículo **"Bending Strength and Stiffness of Second–Growth Douglas–Fir Dimension Lumber" (*Forest Products J*., 1991: 35–43)**, excepto que los tamaños de muestra allí eran más grandes):

Grado	J	$\bar{x}_{i\cdot}$	s_i
1	10	1.63	0.27
2	10	1.56	0.24
3	10	1.42	0.26

Use estos datos y un nivel de significancia de 0.01 para probar la hipótesis nula de no diferencia en el módulo medio de elasticidad para los tres grados.

6. El artículo **"Origin of Precambrian Iron Formations" (*Econ. Geology*, 1964: 1025–1057)** reporta los siguientes datos sobre Fe total para cuatro tipos de formación de hierro (1 = carbonato, 2 = silicato, 3 = magnetita, 4 = hematita).

1:	20.5	28.1	27.8	27.0	28.0
	25.2	25.3	27.1	20.5	31.3
2:	26.3	24.0	26.2	20.2	23.7
	34.0	17.1	26.8	23.7	24.9
3:	29.5	34.0	27.5	29.4	27.9
	26.2	29.9	29.5	30.0	35.6
4:	36.5	44.2	34.1	30.3	31.4
	33.1	34.1	32.9	36.3	25.5

Analice una prueba F de varianza a un nivel de significancia de 0.01 y resuma los resultados en una tabla ANOVA.

7. Se realizó un experimento para comparar la resistencia eléctrica de seis diferentes mezclas de hormigón de baja permeabilidad para la cubierta de un puente. Hubo 26 mediciones en cilindros de hormigón para cada mezcla, los cuales se obtuvieron 28 días después de la fundición. Las entradas de la siguiente tabla de ANOVA se basan en la información del artículo **"In–Place Resistivity of Bridge Deck Concrete Mixtures" (*ACI Materials J*., 2009: 114–122)**. Complete el resto de las entradas y la prueba de hipótesis adecuada.

Fuente	Grados de libertad	Suma de cuadrados	Media de cuadrados	f
Mezcla				
Error			13.929	
Total		5664.415		

8. Un estudio de las propiedades de las estructuras conectadas con placas metálicas para soportar techos (**"Modeling Joints Made with Light–GaugeMetal Connector Plates", *Forest Products J*., 1979: 39–44**) arroja las siguientes observaciones de índice de rigidez axial (libras/pulg^2) de tramos de placa de 4, 6, 8, 10 y 12 pulg:

4:	309.2	409.5	311.0	326.5	316.8	349.8	309.7
6:	402.1	347.2	361.0	404.5	331.0	348.9	381.7
8:	392.4	366.2	351.0	357.1	409.9	367.3	382.0
10:	346.7	452.9	461.4	433.1	410.6	384.2	362.6
12:	407.4	441.8	419.9	410.7	473.4	441.2	465.8

¿Tiene algún efecto la variación de la longitud de placas en la rigidez axial promedio verdadera? Formule y pruebe las hipótesis pertinentes mediante un análisis de varianza con $\alpha = 0.01$. Muestre sus resultados en una tabla ANOVA. [*Sugerencia*: $\Sigma\Sigma x_{ij}^2 = 5\ 241\ 420.79$.]

9. Se analizaron seis muestras de cada uno de los cuatro tipos de crecimiento de los granos de cereal en una región para determinar el contenido de tiamina y se obtuvieron los siguientes resultados (μg/g):

Trigo	5.2	4.5	6.0	6.1	6.7	5.8
Cebada	6.5	8.0	6.1	7.5	5.9	5.6
Maíz	5.8	4.7	6.4	4.9	6.0	5.2
Avena	8.3	6.1	7.8	7.0	5.5	7.2

¿Sugieren estos datos que al menos dos de los granos difieren respecto al contenido de tiamina promedio verdadero? Use una prueba con nivel $\alpha = 0.05$.

10. En ANOVA unifactorial con tratamientos I y observaciones J por cada tratamiento, sea $\mu = (1/I)\Sigma\mu_{i\cdot}$.

a. Exprese $E(\overline{X}_{\cdot\cdot})$ en función de μ. [*Sugerencia*: $\overline{X}_{\cdot\cdot} = (1/I)\ \Sigma\ \overline{X}_{i\cdot}$.]

b. Calcule $E(\overline{X}_{i\cdot}^2)$. [*Sugerencia*: con cualquier variable aleatoria Y, $E(Y^2) = V(Y) + [E(Y)]^2$.]

c. Calcule $E(\overline{X}_{\cdot\cdot}^2)$.

d. Calcule $E(\text{SSTr})$ y luego demuestre que

$$E(\text{MSTr}) = \sigma^2 + \frac{J}{I-1}\sum(\mu_i - \mu)^2$$

e. Con el resultado del inciso d), ¿cuál es la $E(\text{MSTr})$ cuando H_0 es verdadera? Cuando H_0 es falsa, ¿se compara la $E(\text{MSTr})$ con σ^2?

8.2 Comparaciones múltiples en ANOVA

Cuando el valor calculado del estadístico F en un ANOVA unifactorial no es significativo, el análisis se termina porque no se han identificado diferencias entre las μ_i. Pero cuando H_0 es rechazada, el investigador normalmente deseará saber cuáles de las μ_i son diferentes una de otra. Un método para realizar este análisis adicional se llama **procedimiento de comparaciones múltiples**.

Muchos de los procedimientos que se utilizan con más frecuencia están basados en la siguiente idea central. Primero se calcula un intervalo de confianza para cada diferencia $\mu_i - \mu_j$ con $i < j$. Por consiguiente, si $I = 4$, los seis intervalos de confianza requeridos serían para $\mu_1 - \mu_2$ (pero no para $\mu_2 - \mu_1$), $\mu_1 - \mu_3$, $\mu_1 - \mu_4$, $\mu_2 - \mu_3$, $\mu_2 - \mu_4$ y $\mu_3 - \mu_4$. Por tanto, si el intervalo para $\mu_1 - \mu_2$ no incluye 0, se concluye que μ_1 y μ_2 *difieren significativamente* una de otra; si el intervalo sí incluye 0, se considera que las dos μ no difieren de manera significativa. Si se sigue la misma línea de razonamiento para cada uno de los demás intervalos, finalmente se es capaz de juzgar si cada par de μ difiere o no en forma significativa uno de otro.

Los procedimientos basados en esta idea difieren en el método utilizado para calcular los diferentes intervalos de confianza. Aquí se presenta un método popular que controla el nivel de confianza *simultáneo* para todos los intervalos $I(I-1)/2$.

Procedimiento de Tukey (el método T)

El procedimiento de Tukey implica utilizar otra distribución de probabilidad llamada **distribución de rango estudentizado.** La distribución depende de dos parámetros: m grados de libertad asociados con el numerador y ν grados de libertad asociados con el denominador. Sea $Q_{\alpha,m,\nu}$ el valor crítico α de cola superior de la distribución de rango estudentizado con m grados de libertad asociados con el numerador y ν grados de libertad asociados con el denominador (análogo a F_{α,ν_1,ν_2}). En la tabla A.10 del apéndice se muestran los valores de $Q_{\alpha,m,\nu}$.

PROPOSICIÓN

Con la probabilidad $1 - \alpha$,

$$\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - Q_{\alpha,I,I(J-1)}\sqrt{\text{MSE}/J} \le \mu_i - \mu_j \le \overline{X}_{i\cdot} - \overline{X}_{j\cdot} + Q_{\alpha,I,I(J-1)}\sqrt{\text{MSE}/J} \qquad (8.4)$$

para *cada* i ($i = 1,..., I$ y $j = 1,..., I$) con $i < j$.

Observe que los grados de libertad asociados con el numerador para el valor crítico Q_α apropiado es I, el número de medias de la población o tratamiento que se están comparando, y no $I - 1$ como en la prueba F. Cuando las $\overline{x}_{i\cdot}$, $\overline{x}_{j\cdot}$ son calculadas y el MSE se sustituye en (8.4), el resultado es un conjunto de intervalos de confianza con nivel de confianza *simultáneo* de $100(1 - \alpha)\%$ para todas las diferencias de la forma $\mu_i - \mu_j$ con $i < j$. Cada intervalo que no incluye 0 da lugar a la conclusión de que los valores correspondientes de μ_i y μ_j difieren significativamente uno de otro.

Puesto que en realidad no interesan los límites inferior y superior de los diversos intervalos, sino únicamente cuál incluye 0 y cuál no, se puede evitar mucha de la aritmética asociada con (8.4). El siguiente recuadro proporciona detalles y describe cómo se pueden identificar las diferencias de modo visual con un "patrón de subrayado".

Método *T* para identificar μ_i significativamente diferentes

Se selecciona α, se extrae $Q_{\alpha\, I,I(J-1)}$ de la tabla A.10 del apéndice y se calcula $w = Q_{\alpha,\, I,\, I(J-1)} \cdot \sqrt{\text{MSE}/J}$. Luego se hace una lista de las medias muestrales en orden creciente y se subrayan los pares que menos difieren de w. Cualquier par de medias muestrales no subrayado por la misma raya corresponde a un par de medias de población o tratamiento juzgadas significativamente diferentes.

Suponga, por ejemplo, que $I = 5$ y que

$$\overline{x}_{2\cdot} < \overline{x}_{5\cdot} < \overline{x}_{4\cdot} < \overline{x}_{1\cdot} < \overline{x}_{3\cdot}$$

Por tanto,

1. Considere en primer lugar la media más pequeña $\overline{x}_{2\cdot}$. Si $\overline{x}_{5\cdot} - \overline{x}_{2\cdot} \ge w$, prosiga al paso 2. Sin embargo, si $\overline{x}_{5\cdot} - \overline{x}_{2\cdot} < w$, conecte estas primeras dos medias con un segmento de recta. Luego, si es posible, extienda este segmento de recta más a la derecha de la más grande $\overline{x}_{i\cdot}$ que difiera de $\overline{x}_{2\cdot}$ que difieren por menos del valor de w (de modo que la recta pueda conectar dos, tres, incluso más medias).
2. Ahora siga con $\overline{x}_{5\cdot}$ y otra vez extienda el segmento de línea hasta la derecha de la más grande que difiera de $\overline{x}_{5\cdot}$ que difieren por menos del valor de w (quizá no sea posible trazar esta línea o de manera alternativa pueden subrayarse sólo dos medias, o tres, o incluso las cuatro medias restantes).
3. Continúe con $\overline{x}_{4\cdot}$ y repita, y finalmente continúe con $\overline{x}_{1\cdot}$

Para resumir, al comenzar en cada media que aparece en la lista ordenada, un segmento de recta se extiende a la derecha tan lejos como sea posible, siempre y cuando la diferencia entre las medias sea menor que w. Es fácil verificar que un intervalo particular de la forma (8.4) contendrá 0 si y sólo si el par correspondiente de medias muestrales está subrayado por el mismo segmento de recta.

EJEMPLO 8.5 Se realizó un experimento para comparar cinco marcas diferentes de filtros de aceite para automóviles respecto a su capacidad de atrapar materia extraña. Sea μ_i la cantidad promedio verdadera de material atrapado por los filtros marca i ($i = 1,..., 5$) en condiciones

controladas. Se utilizó una muestra de nueve filtros de cada marca y se obtuvieron las siguientes cantidades medias muestrales: $\bar{x}_{1\cdot} = 14.5$, $\bar{x}_{2\cdot} = 13.8$, $\bar{x}_{3\cdot} = 13.3$, $\bar{x}_{4\cdot} = 14.3$ y $\bar{x}_{5\cdot} = 13.1$. La tabla 8.3 es una tabla ANOVA que resume la primera parte del análisis.

Tabla 8.3 Tabla ANOVA para el ejemplo 8.5

Origen de la variación	Grados de libertad	Suma de cuadrados	Media de cuadrados	f
Tratamientos (marcas)	4	13.32	3.33	37.84
Error	40	3.53	0.088	
Total	44	16.85		

Puesto que $F_{0.001,4,40} = 5.70$, el valor P es menor a 0.001. Por tanto, H_0 se rechaza (con decisión) al nivel 0.05. Ahora utilizamos el procedimiento de Tukey para buscar diferencias significativas entre las μ_i. De la tabla A.10 del apéndice, $Q_{0.05,5,40} = 4.04$ (el segundo subíndice en el Q es I no $I - 1$ como en F), por tanto, $w = 4.04\sqrt{0.088/9} = 0.4$. Después de arreglar las cinco medias de la muestra en orden creciente, se pueden conectar las dos más pequeñas mediante un segmento de recta debido a que difieren en menos de 0.4. Sin embargo, este segmento no puede extenderse más a la derecha, ya que 13.8 – 13.1 = $0.7 \geq 0.4$. Moviéndose una media a la derecha, el par $\bar{x}_{3\cdot}$ y $\bar{x}_{2\cdot}$ no puede ser subrayado porque estas medias difieren en más de 0.4. De nuevo, moviéndose a la derecha, la siguiente media, 13.8, no se puede conectar a cualquiera que esté más a la derecha. Las dos últimas medias pueden ser subrayadas con el mismo segmento de recta.

$\bar{x}_{5\cdot}$	$\bar{x}_{3\cdot}$	$\bar{x}_{2\cdot}$	$\bar{x}_{4\cdot}$	$\bar{x}_{1\cdot}$
13.1	13.3	13.8	14.3	14.5

Así pues las marcas 1 y 4 no son significativamente diferentes una de otra, pero sí son más altas de manera significativa que las otras tres marcas en sus contenidos promedio verdaderos. La marca 2 es significativamente mejor que la 3 y la 5, pero peor que la 1 y la 4, y las marcas 3 y 5 no difieren en modo significativo.

Si $\bar{x}_{2\cdot} = 14.15$ en lugar de 13.8 con el mismo valor w calculado, entonces la configuración de medias subrayadas sería

$\bar{x}_{5\cdot}$	$\bar{x}_{3\cdot}$	$\bar{x}_{2\cdot}$	$\bar{x}_{4\cdot}$	$\bar{x}_{1\cdot}$
13.1	13.3	14.15	14.3	14.5

■

EJEMPLO 8.6 Un biólogo deseaba estudiar los efectos del etanol en el periodo de sueño. Se seleccionó una muestra de 20 ratas coincidentes por edad y otras características y a cada rata se le administró una inyección oral con una concentración particular de etanol por peso corporal. Luego se registró el movimiento rápido de los ojos (REM, por sus siglas en inglés) durante el periodo de sueño de cada rata durante 24 horas, con los siguientes resultados:

Tratamiento (concentración de etanol)						$x_{i\cdot}$	$\bar{x}_{i\cdot}$
0 (control)	88.6	73.2	91.4	68.0	75.2	396.4	79.28
1 g/kg	63.0	53.9	69.2	50.1	71.5	307.7	61.54
2 g/kg	44.9	59.5	40.2	56.3	38.7	239.6	47.92
4 g/kg	31.0	39.6	45.3	25.2	22.7	163.8	32.76

$x_{\cdot\cdot} = 1107.5$ $\bar{x}_{\cdot\cdot} = 55.375$

¿Será que los datos indican datos que el promedio real del periodo de sueño REM depende de la concentración de etanol? (Este ejemplo está basado en el experimento reportado en **"Relationship of Ethanol Blood Level to REM and Non2REM Sleep Time and Distribution in the Rat", *Life Sciences*, 1978: 839–846.**)

Las $\bar{x}_{i.}$ difieren sustancialmente una de otra, aunque también existe una gran cantidad de variabilidad dentro de cada muestra, por lo que para responder la pregunta con precisión se debe realizar el ANOVA. Las más pequeña y la más grande de las desviaciones estándar de la muestra cuatro son 9.34 y 10.18, respectivamente, lo cual apoya la hipótesis de varianzas iguales. Una gráfica de probabilidad normal de los 20 residuos muestra un patrón razonablemente lineal, justificando la suposición de que las cuatro distribuciones de tiempo de sueño REM son normales. Por tanto, es legítimo emplear la prueba F.

La tabla 8.4 es una tabla ANOVA SAS. La última columna da el valor P como 0.0001. Con un nivel de significancia de 0.05 se rechaza la hipótesis nula H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, puesto que valor $P = 0.0001 < 0.05 = \alpha$. Al parecer el promedio real del periodo de sueño REM depende del nivel de concentración.

Tabla 8.4 Tabla ANOVA SAS

Analysis of Variance Procedure
Dependent Variable: TIEMPO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5882.35750	1960.78583	21.09	0.0001
Error	16	1487.40000	92.96250		
Corrected Total	19	7369.75750			

Existen $I = 4$ tratamientos y 16 grados de libertad asociados con el error, por tanto, $Q_{0.05,4,16} = 4.05$ y $w = 4.05\sqrt{93.0/5} = 17.47$. Ordenando las medias y subrayándolas se obtiene

$\bar{x}_{4.}$	$\bar{x}_{3.}$	$\bar{x}_{2.}$	$\bar{x}_{1.}$
32.76	47.92	61.54	79.28

La interpretación de este subrayado debe hacerse con cuidado, puesto que al parecer se ha concluido que los tratamientos 2 y 3 no difieren, 3 y 4 no difieren; no obstante, 2 y 4 sí difieren. Para expresar esto se sugiere decir que, aunque la evidencia permite concluir que los tratamientos 2 y 4 difieren uno de otro, no se ha demostrado que alguno sea significativamente diferente del 3. El tratamiento 1 tiene un periodo de sueño REM promedio verdadero más alto de manera significativa que cualquiera de los demás tratamientos.

La figura 8.5 muestra resultados obtenidos con SAS a partir de la aplicación del procedimiento de Tukey.

```
Alpha =0.05 df =16 MSE = 92.9625
Critical Value of Studentized Range = 4.046
Minimum Significant Difference = 17.446

Means with the same letter are not significantly different.

Tukey Grouping        Mean        N   TREATMENT
           A          79.280      5   0(control)

           B          61.540      5   1 gm/kg
           B
     C     B          47.920      5   2 gm/kg
     C
     C                32.760      5   4 gm/kg
```

Figura 8.5 Método de Turkey usando SAS

Interpretación de α en el método de Tukey

Previamente se manifestó que el método de Tukey controla el nivel de confianza *simultáneo*. Por tanto, ¿qué significa "simultáneo" en este caso? Calcule un intervalo de confianza de 95% para una media de población μ, basada en una muestra de dicha población, y

luego un intervalo de confianza de 95% para una proporción de población p basado en otra muestra seleccionada independientemente de la primera. Antes de obtener los datos, la probabilidad de que el primer intervalo incluya μ es de 0.95 y esta también es la probabilidad de que el segundo intervalo incluya p. Puesto que las dos muestras se seleccionan de manera independiente una de otra, la probabilidad de que ambos intervalos incluyan los valores de los parámetros respectivos es $(0.95)(0.95) = (0.95)^2 \approx 0.90$. Por consiguiente, el nivel de confianza *simultáneo* o *conjunto* para los dos intervalos es aproximadamente de 90%; si se calculan pares de intervalos una y otra vez con muestras independientes, a la larga aproximadamente 90% de las veces el primer intervalo capturará μ y el segundo incluirá p. Asimismo, si se calculan tres intervalos de confianza basados en muestras independientes, el nivel de confianza simultáneo será de $100(0.95)^3\% \approx 86\%$. Claramente, a medida que se incrementa el número de intervalos, se reducirá el nivel de confianza simultáneo de que todos los intervalos capturen sus respectivos parámetros.

Ahora suponga que se desea mantener el nivel de confianza simultáneo en 95%. Así, para dos muestras independientes, el nivel de confianza individual para cada una tendría que ser $100\sqrt{0.95}\,\% \approx 97.5\%$. Mientras más grande es el número de intervalos, más alto tendría que ser el nivel de confianza individual para mantener el nivel simultáneo en 95%.

El truco, en relación con los intervalos Tukey, es que no están basados en muestras independientes: MSE aparece en todos y varios intervalos comparten las mismas $\overline{x}_{i\cdot}$ (p. ej., en el caso $I = 4$, tres intervalos diferentes utilizan $\overline{x}_{1\cdot}$). Esto implica que no existe un argumento de probabilidad directo para discernir el nivel de confianza simultáneo de los niveles de confianza individuales. No obstante, se puede demostrar que, si se utiliza $Q_{0.05}$, el nivel de confianza simultáneo se controla a 95%; en tanto que si se utiliza $Q_{0.01}$ se obtiene un nivel simultáneo de 99%. Para obtener un nivel simultáneo de 95%, el nivel individual de cada intervalo debe ser considerablemente más grande que 95%. Expresado de forma un poco diferente, para obtener una proporción de error de 5% asociada con un *experimento* o *familia*, la proporción de error por comparación o individual para cada intervalo debe ser considerablemente más pequeña que 0.05. Minitab le pide al usuario que especifique la proporción de error asociado con la familia (p. ej., 5%) y luego incluye en los datos de salida la proporción de error individual (véase el ejercicio 16).

Intervalos de confianza para otras funciones paramétricas

En algunas situaciones se desea un intervalo de confianza para una función de las μ_i más complicada que una diferencia $\mu_i - \mu_j$. Sea $\theta = \Sigma c_i\mu_i$, donde las c_i son constantes. Una función como esa es $1/2(\mu_1 + \mu_2) - 1/3(\mu_3 + \mu_4 + \mu_5)$ la cual, en el contexto del ejemplo 8.5, mide la diferencia entre el grupo compuesto de las dos primeras marcas y la de las últimas tres. Puesto que las X_{ij} están normalmente distribuidas con $E(X_{ij}) = \mu_i$ y $V(X_{ij}) = \sigma^2$, $\hat{\theta} = \Sigma c_i\overline{X}_{i\cdot}$, está normalmente distribuida, insesgada para θ, y

$$V(\hat{\theta}) = V\Big(\sum_i c_i\overline{X}_{i\cdot}\Big) = \sum_i c_i^2 V(\overline{X}_{i\cdot}) = \frac{\sigma^2}{J}\sum_i c_i^2$$

La estimación de σ^2 mediante MSE y la formación de $\hat{\sigma}_{\hat{\theta}}$, da por resultado una variable $t(\hat{\theta} - \hat{\theta})/\hat{\sigma}_{\hat{\theta}}$, la cual puede ser manipulada para obtener el siguiente intervalo de confianza de $100(1 - \alpha)\%$ para $\Sigma c_i\mu_i$.

$$\sum c_i\,\overline{x}_{i\cdot} \pm t_{\alpha/2,J(J-1)}\sqrt{\frac{\text{MSE}\Sigma c_i^2}{J}} \tag{8.5}$$

EJEMPLO 8.7 La función paramétrica para comparar las primeras dos marcas de filtro de aceite (tienda) con las últimas tres marcas (nacionales) es $\theta = 1/2(\mu_1 + \mu_2) - 1/3(\mu_3 + \mu_4 + \mu_5)$ con la cual

$$\sum c_i^2 = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(-\frac{1}{3}\right)^2 = \frac{5}{6}$$

Con $\hat{\theta} = 1/2(\bar{x}_{1\cdot} + \bar{x}_{2\cdot}) - 1/3(\bar{x}_{3\cdot} + \bar{x}_{4\cdot} + \bar{x}_{5\cdot}) = 0.583$ y MSE = 0.088, un intervalo de 95% es

$$0.583 \pm 2.021\sqrt{5(0.088)/[(6)(9)]} = 0.583 \pm 0.182 = (0.401, 0.765)$$ ■

En ocasiones se realiza un experimento para comparar cada uno de los varios tratamientos "nuevos" con un tratamiento de control. En tales situaciones es apropiada una técnica de comparaciones múltiples llamada método de Dunnett.

EJERCICIOS Sección 8.2 (11–21)

11. En un experimento para comparar las proporciones de cobertura de cinco marcas diferentes de pintura amarilla de látex para interiores, disponibles en un área particular, se utilizaron 4 galones ($J = 4$) de cada pintura. Las proporciones de cobertura promedio de las muestras (pies²/gal) de las cinco marcas fueron $\bar{x}_{1\cdot} = 462.0, \bar{x}_{2\cdot} = 512.8, \bar{x}_{3\cdot} = 437.5, \bar{x}_{4\cdot} = 469.3$ y $\bar{x}_{5\cdot} = 532.1$. Se encontró que el valor calculado de F es significativo a un nivel de $\alpha = 0.05$. Con MSE = 272.8 use el procedimiento de Tukey para investigar diferencias significativas en las proporciones de cobertura promedio verdaderas entre marcas.

12. En el ejercicio 11 suponga $\bar{x}_{3\cdot} = 427.5$. Ahora, ¿cuáles proporciones de cobertura promedio verdaderas difieren significativamente una de otra? Asegúrese de utilizar el método de subrayar para ilustrar sus conclusiones y escriba un párrafo que resuma sus resultados.

13. Repita el ejercicio 12 suponiendo que $\bar{x}_{2\cdot} = 502.8$ además de $\bar{x}_{3\cdot} = 427.5$.

14. Use el procedimiento de Tukey con los datos del ejemplo 8.3 para identificar diferencias en la resistencia adhesiva real promedio entre los cinco protocolos.

15. El ejercicio 8.7 describe un experimento en el que se realizaron 26 observaciones de resistencia en cada una de seis diferentes mezclas de hormigón. El artículo citado dio las siguientes medias muestrales: 14.18, 17.94, 18.00, 18.00, 25.74, 27.67. Aplique el método de Tukey con un nivel de confianza simultáneo de 95% para identificar diferencias significativas, así como de sus resultados (use MSE = 13.929).

16. Reconsidere los datos de rigidez axial dados en el ejercicio 8. Los siguientes son datos ANOVA obtenidos con Minitab.

```
Analysis of Variance de rigidez axial
Source    DF       SS       MS        F       P
Length     4    43993    10998    10.48   0.000
Error     30    31475     1049
Total     34    75468

Level      N     Mean    StDev
4          7   333.21    36.59
6          7   368.06    28.57
8          7   375.13    20.83
10         7   407.36    44.51
12         7   437.17    26.00

Pooled StDev = 32.39

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.00693

Critical value = 4.10

Intervals for (column level mean) - (row level
mean)
 4         6           8           10
6        -85.0
          15.4
8        -92.1       -57.3
           8.3        43.1
10      -124.3       -89.5       -82.4
         -23.9        10.9        18.0
12      -154.2      -119.3      -112.2    -80.0
         -53.8       -18.9       -11.8     20.4
```

a. ¿Será factible que las varianzas de las cinco distribuciones de índices de rigidez axial sean idénticas? Explique.

b. Use los resultados (sin referencia a la tabla F) para probar las hipótesis pertinentes.

c. Use los intervalos de Tukey dados en los resultados para determinar cuáles medias difieren y construya el patrón de subrayado correspondiente.

17. Remítase al ejercicio 5. Calcule un intervalo de confianza t de 95% con $\theta = 1/2(\mu_1 + \mu_2) - \mu_3$.

18. Considere los datos adjuntos sobre el crecimiento de las plantas después de la aplicación de cinco diferentes tipos de la hormona del crecimiento

1:	13	17	7	14
2:	21	13	20	17
3:	18	15	20	17
4:	7	11	18	10
5:	6	11	15	8

a. Realice una prueba F al nivel $\alpha = 0.05$.

b. ¿Qué sucede cuando se aplica el procedimiento de Tukey?

19. Considere un experimento ANOVA unifactorial en el cual $I = 3, J = 5, \bar{x}_{1\cdot} = 10, \bar{x}_{2\cdot} = 12$ y $\bar{x}_{3\cdot} = 20$. Encuentre un valor de SSE con el cual $f > F_{0.05,2,12}$ de modo que H_0: $\mu_1 = \mu_2 = \mu_3$ sea rechazada, aunque cuando se aplica el procedimiento de Tukey se puede decir que ninguna de las μ_i difieren significativamente una de otra.

20. Remítase al ejercicio 19 y suponga $\bar{x}_{1\cdot} = 10$, $\bar{x}_{2\cdot} = 15$ y $\bar{x}_{3\cdot} = 20$. ¿Puede hallar ahora un valor de SSE que produzca semejante contradicción entre la prueba F y el procedimiento de Tukey?

21. El artículo **"The Effect of Enzyme Inducing Agents on the Survival Times of Rats Exposed to Lethal Levels of Nitrogen Dioxide" (*Toxicology and Applied Pharmacology*, 1978: 169–174)** reporta los siguientes datos sobre los tiempos de sobrevivencia de ratas expuestas a bióxido de nitrógeno (70 ppm) mediante diferentes regímenes de inyección. Hubo $J = 14$ ratas en cada grupo.

Régimen	$\bar{x}_{i\cdot}$ (mín)	s_i
1. Control	166	32
2. 3-Metilcolantreno	303	53
3. Alilisopropilacetamida	266	54
4. Fenobarbital	212	35
5. Cloropromazina	202	34
6. Ácido *p*-Aminobenzoico	184	31

a. Pruebe las hipótesis nulas de que el tiempo de sobrevivencia promedio verdadero no depende del régimen de inyección contra la alternativa de que existe alguna dependencia del régimen de inyección con $\alpha = 0.01$.

b. Suponga que se calculan intervalos de confianza de $100(1 - \alpha)\%$ para k funciones paramétricas diferentes con el mismo conjunto de datos ANOVA. Entonces es fácil verificar que el nivel de confianza simultáneo es al menos de $100(1 - k\alpha)\%$. Calcule los intervalos de confianza con nivel de confianza simultáneo de al menos 98% para
$\mu_1 - 1/5(\mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_6)$ y
$1/4\,(\mu_2 + \mu_3 + \mu_4 + \mu_5) - \mu_6$

EJERCICIOS SUPLEMENTARIOS (22–32)

22. El cortisol es una hormona que desempeña un papel importante en la mediación del estrés. Hay una creciente conciencia de que la exposición de los trabajadores al aire libre a los contaminantes puede afectar sus niveles de cortisol. El artículo **"Plasma Cortisol Concentration and Lifestyle in a Population of Outdoor Workers" (*Intl. J. of Envir. Health Res.*, 2011: 62–71)** reporta un estudio con tres grupos de agentes de la policía: 1) policía de tránsito (PT), 2) conductores (C) y 3) otras obligaciones (O). A continuación se presentan datos del resumen sobre la concentración de cortisol (ng/ml) para un subconjunto de oficiales que ni bebían ni fumaban.

Grupo	Tamaño muestral	Media	DE
TP	47	174.7	50.9
D	36	160.2	37.2
O	50	153.5	45.9

Asumiendo que se cumplen las suposiciones estándar para ANOVA unidireccional, efectúe una prueba en el nivel de significancia 0.05 para decidir si la concentración de cortisol promedio verdadero es diferente para los tres grupos. [*Nota*: Los investigadores utilizan métodos estadísticos más sofisticados (regresión múltiple) para evaluar el impacto de la edad, la duración del empleo, beber y fumar en el estado de concentración de cortisol; considerando estos factores, la concentración parece significativamente mayor en el grupo PT que en los otros dos grupos.]

23. Numerosos factores contribuyen al suave funcionamiento de un motor eléctrico **("Increasing Market Share Through Improved Product and Process Design: An Experimental Approach", *Quality Engineering*, 1991: 361–369)**. En particular, es deseable mantener el ruido del motor y las vibraciones a un mínimo. Para estudiar el efecto que la marca de los cojinetes ejerce en la vibración del motor, se examinaron cinco marcas diferentes de cojinetes, y se instaló cada tipo de cojinete en muestras aleatorias distintas de seis motores. Se registró la cantidad de vibración del motor (medida en micrones) cuando cada uno de los 30 motores estaba funcionando. Los datos de este estudio se dan a continuación. Formule y pruebe las hipótesis pertinentes a un nivel de significancia de 0.05 y luego realice un análisis de comparaciones múltiples, si es apropiado.

							Media
1:	13.1	15.0	14.0	14.4	14.0	11.6	13.68
2:	16.3	15.7	17.2	14.9	14.4	17.2	15.95
3:	13.7	13.9	12.4	13.8	14.9	13.3	13.67
4:	15.7	13.7	14.4	16.0	13.9	14.7	14.73
5:	13.5	13.4	13.2	12.7	13.4	12.3	13.08

24. Un artículo publicado en el diario científico británico *Nature* **("Sucrose Induction of Hepatic Hyperplasia in the Rat", 25 de agosto de 1972: 461)** reporta sobre un experimento en el cual cada uno de cinco grupos compuestos de seis ratas fue puesto a dieta con un carbohidrato diferente. Al final del

experimento se determinó el contenido de ADN del hígado de cada rata (mg/g hígado), con los siguientes resultados:

Carbohidrato	$\bar{x}_i$
Almidón	2.58
Sucrosa	2.63
Fructosa	2.13
Glucosa	2.41
Maltosa	2.49

Suponiendo también que $\Sigma\Sigma x_{ij}^2 = 183.4$, ¿indicarán estos datos que el tipo de carbohidrato presente en la dieta afecta el contenido de ADN promedio verdadero? Construya una tabla ANOVA y use un nivel de significancia de 0.05.

25. Remítase al ejercicio 24 y construya un intervalo de confianza *t* para

$$\theta = \mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

que mide la diferencia entre el contenido de ADN promedio para la dieta de almidón y el promedio combinado para las otras cuatro dietas. ¿Incluye cero el intervalo resultante?

26. Remítase al ejercicio 38. ¿Cuál es β para la prueba cuando el contenido de ADN promedio verdadero es idéntico para las tres dietas y queda a 1 desviación estándar (σ) por debajo de este valor común para las otras dos dietas?

27. Se seleccionan al azar cuatro laboratorios (1–4) de una población grande y a cada uno se le pide que realice tres determinaciones del porcentaje de alcohol metílico en especímenes de un compuesto tomado de un solo lote. Basado en los datos adjuntos, ¿serán las diferencias entre los laboratorios una causa de variación del porcentaje de alcohol metílico? Formule y pruebe las hipótesis pertinentes con un nivel de significancia de 0.05.

1: 85.06 85.25 84.87
2: 84.99 84.28 84.88
3: 84.48 84.72 85.10
4: 84.10 84.55 84.05

28. La frecuencia de parpadeo crítica (CFF, por sus siglas en inglés) es la frecuencia más alta (en ciclos/s) a la que una persona puede advertir el parpadeo en una fuente luminosa parpadeante. A frecuencias por encima de la frecuencia de parpadeo crítica, la fuente luminosa parece ser continua aun cuando en realidad parpadee. Una investigación para ver si la frecuencia de parpadeo crítica promedio verdadera depende del color del iris arrojó los siguientes datos (con base en el artículo **"The Effects of Iris Color in Critical Flicker Frequency", *J. of General Psych*., 1973: 91–95**):

	Color del iris		
	1. Café	**2. Verde**	**3. Azul**
	26.8	26.4	25.7
	27.9	24.2	27.2
	23.7	28.0	29.9
	25.0	26.9	28.5
	26.3	29.1	29.4
	24.8		28.3
	25.7		
	24.5		
J_i	8	5	6
$x_{i\cdot}$	204.7	134.6	169.0
$\bar{x}_{i\cdot}$	25.59	26.92	28.17

$n = 19 \quad x_{\cdot\cdot} = 508.3$

a. Formule y pruebe las hipótesis pertinentes a un nivel de significancia de 0.05. [*Sugerencia*: $\Sigma\Sigma x_{ij}^2 = 13659.67$ y CF = 13 598.36.]

b. Investigue las diferencias entre colores del iris respecto a la frecuencia de parpadeo crítica media.

29. Sean $c_1, c_2, \ldots, c_I$ los números que satisfacen la expresión $\Sigma c_i = 0$. Entonces $\Sigma c_i\mu_i = c_1\mu_1 + \cdots + c_I\mu_I$ se llama *contraste* en las μ_i. Observe que con $c_1 = 1$, $c_2 = -1$, $c_3 = \cdots = c_I = 0$, $\Sigma c_i\mu_i = \mu_1 - \mu_2$ la cual implica que toda diferencia tomada por pares entre las μ_i es un contraste (también lo es, p. ej., $\mu_1 - 0.5\mu_2 - 0.5\mu_3$). Un método atribuido a Scheffé da intervalos de confianza simultáneos con nivel de confianza simultáneo de $100(1 - \alpha)\%$ para *todos* los contrastes posibles (¡un número infinito de ellos!) El intervalo para $\Sigma c_i\mu_i$ es

$$\sum c_i \bar{x}_{i\cdot} \pm \left(\sum c_i^2/J_i\right)^{1/2} \cdot [(I-1) \cdot \text{MSE} \cdot F_{\alpha,I-1,n-I}]^{1/2}$$

Usando los datos del ejercicio 28, acerca de la frecuencia crítica del parpadeo, calcule los intervalos de Scheffé para los contrastes $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, $\mu_2 - \mu_3$ y $0.5\mu_1 + 0.5\mu_2 - \mu_3$ (este último contraste compara azul con el promedio de café y verde). ¿Cuál contraste parece diferir significativamente de 0 y por qué?

30. Cuatro tipos de morteros: de cemento ordinario (OCM), impregnado de polímero (PIM), con resina (RM) y mortero con cemento y polímero (PCM) se sometieron a una prueba de compresión para medir su resistencia (MPa). En el artículo **"Polymer Mortar Composite Matrices for Maintenance–Free Highly Durable Ferrocement" (*J. of Ferrocement*, 1984: 337–345)** se dan tres observaciones de resistencia de cada tipo de mortero y las reproducimos aquí. Construya una tabla ANOVA. Con un nivel de significancia de 0.05, determine si los datos sugieren que la resistencia media verdadera no es la misma para los cuatro tipos de mortero. Si determina que las resistencias medias

verdaderas no son iguales use el método de Tukey para identificar las diferencias significativas.

OCM	32.15	35.53	34.20
PIM	126.32	126.80	134.79
RM	117.91	115.02	114.58
PCM	29.09	30.87	29.80

31. Suponga que las x_{ij} están "codificadas" por $y_{ij} = cx_{ij} + d$. ¿Cómo se compara el valor del estadístico F calculado con las y_{ij} con el valor calculado con las x_{ij}? Justifique su argumentación.

32. En el ejemplo 10.11, reste $\bar{x}_i$ de cada observación en la i-ésima muestra ($i = 1,..., 6$) para obtener un conjunto de 18 residuos. Después construya una gráfica de probabilidad normal y comente sobre la factibilidad de la suposición de normalidad.

BIBLIOGRAFÍA

Miller, Rupert, *Beyond ANOVA*: *The Basics of Applied Statistics*, Wiley, Nueva York, 1986. Una excelente fuente de información sobre comprobación de suposiciones y métodos de análisis alternativos.

Montgomery, Douglas, *Design and Analysis of Experiments* (8a. ed.), Wiley, Nueva York, 2013. Una presentación muy al día de modelos y metodología ANOVA.

Neter, John, William Wasserman y Michael Kutner, *Applied Linear Statistical Models* (5a. ed.), Irwin, Homewood, IL., 2004. La segunda mitad de este libro contiene un estudio muy bien presentado de ANOVA; el nivel es comparable al del presente texto, aunque la discusión es más amplia, lo que hace del libro una excelente referencia.

Ott, R. Lyman y Michael Longnecker. *An Introduction to Statistical Methods and Data Analysis* (6a. ed.), Duxbury Press, Belmont, CA, 2010. Incluye varios capítulos sobre metodología ANOVA que pueden aprovechar los estudiantes que desean una exposición no muy matemática; incluye un capítulo muy bueno sobre varios métodos de comparaciones múltiples.

Capítulo

9

Regresión lineal simple y correlación

INTRODUCCIÓN

El **análisis de regresión** es la parte de la estadística que se ocupa de investigar la relación entre dos o más variables asociadas en una forma no determinística. En este capítulo, se generaliza la relación lineal determinística $y = \beta_0 + \beta_1 x$ a una relación probabilística lineal, se desarrollan procedimientos para hacer inferencias sobre del modelo y se obtiene una medida cuantitativa (el coeficiente de correlación) del grado al cual las dos variables están relacionadas.

9.1 Modelo de regresión lineal simple

La relación matemática determinística más simple entre dos variables x y y es una relación lineal $y = \beta_0 + \beta_1 x$. El conjunto de pares (x, y) para los cuales $y = \beta_0 + \beta_1 x$ determina una línea recta con pendiente β_1 e intersección en y β_0.* El objetivo de esta sección es desarrollar un modelo probabilístico lineal.

Si las dos variables no están determinísticamente relacionadas, entonces con un valor fijo de x el valor de la segunda variable es incierto. Por ejemplo, si se está investigando la relación entre la edad de un niño y el tamaño del vocabulario y se decide seleccionar un niño de $x = 5.0$ años de edad, entonces antes de hacer la selección, el tamaño del vocabulario es una variable aleatoria Y. Después de que un niño en particular de 5 años de edad ha sido seleccionado y sometido a prueba, el resultado puede ser un vocabulario de 2000 palabras. Se diría entonces que el valor observado de Y asociado con la fijación de $x = 5.0$ fue $y = 2000$.

De modo más general, la variable cuyo valor es fijado por el experimentador será denotada por x y se llamará **variable independiente, pronosticadora o explicativa**. Con x fija, la segunda variable será aleatoria; esta variable aleatoria y su valor observado se designan por Y y y, respectivamente, y se le conoce como **variable dependiente** o **de respuesta**.

Normalmente se realizarán observaciones para varios escenarios de la variable independiente. Sean $x_1, x_2, \ldots, x_n$ los valores de la variable independiente para la que se realizan las observaciones y sean Y_i y y_i, respectivamente, la variable aleatoria y el valor observado asociado con x_i. Los datos bivariantes disponibles se componen entonces de los n pares $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Una imagen de estos datos llamada **gráfica de dispersión** proporciona impresiones preliminares acerca de la naturaleza de cualquier relación. En una gráfica como esa, cada (x_i, y_i) se representa como un punto colocado en un sistema de coordenadas bidimensional.

EJEMPLO 9.1 Los problemas visuales y musculoesqueléticos asociados con el uso de terminales con pantalla de visualización (VDT, por sus siglas en inglés) se han vuelto un tanto comunes en años recientes. Algunos investigadores se han enfocado en la dirección vertical de la mirada fija como causa del cansancio e irritación de los ojos. Se sabe que esta relación está estrechamente relacionada con el área de la superficie ocular (OSA, por sus siglas en inglés), así que se requiere un método para medir el área de la superficie ocular. Los siguientes datos representativos sobre y = OSA (cm^2) y x = ancho de la fisura palprebal (es decir, el ancho horizontal de la apertura del ojo, en cm) fueron tomados del artículo **"Analysis of Ocular Surface Area for Comfortable VDT Workstation Layout" (*Ergonomics*, 1996: 877–884)**. No se proporciona el orden en el cual se obtuvieron las observaciones, así que por conveniencia se muestran en orden creciente de los valores x.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	0.40	0.42	0.48	0.51	0.57	0.60	0.70	0.75	0.75	0.78	0.84	0.95	0.99	1.03	1.12
y_i	1.02	1.21	0.88	0.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

* La pendiente de una recta es el cambio en y con un incremento de 1 unidad en x. Por ejemplo, si $y = -3x + 10$, entonces y se reduce en 3 cuando x se incrementa en 1, de modo que la pendiente es -3. La intersección en y es la altura a la cual la recta cruza el eje vertical y se obtiene estableciendo $x = 0$ en la ecuación.

Por consiguiente $(x_1, y_1) = (0.40, 1.02)$, $(x_5, y_5) = (0.57, 1.52)$, etcétera. En la figura 9.1 se muestra una gráfica de dispersión obtenida con Minitab; se utilizó una opción que produjo una gráfica de puntos tanto de valores x como de valores y individualmente a lo largo de los márgenes derecho y superior de la gráfica, lo que facilita visualizar las distribuciones de las variables individuales (los histogramas o gráficas de caja son opciones alternativas). He aquí algunas anotaciones que hay que tener en cuenta sobre los datos y la gráfica:

- Varias observaciones tienen valores x idénticos aunque valores y diferentes (p. ej., $x_8 = x_9 = 0.75$ pero $y_8 = 1.80$ y $y_9 = 1.74$). Por tanto, el valor de y *no* está determinado sólo por x, sino también por varios otros factores.
- Existe una fuerte tendencia a que y se incremente a medida que x lo hace. Es decir, los valores grandes de OSA tienden a asociarse con valores grandes de ancho de fisura, una relación positiva entre las variables.

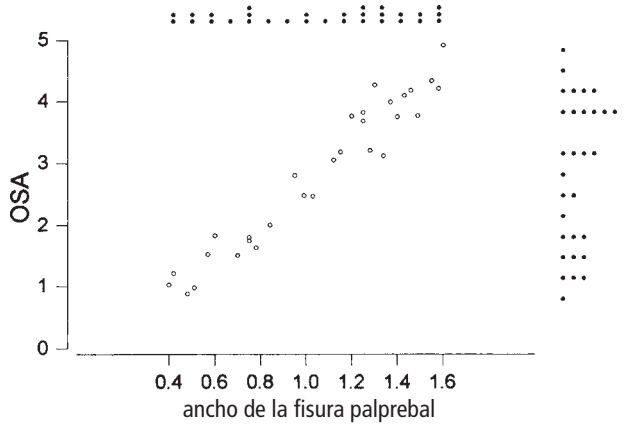


Figura 9.1 Gráfica de dispersión obtenida con Minitab con los datos del ejemplo 9.1, junto con gráficas de puntos de valores x y y

- Al parecer el valor de y podría ser pronosticado a partir de x al encontrar una recta que esté razonablemente cerca de los puntos presentes en la gráfica (los autores del artículo citado superponen tal recta en su gráfica). En otras palabras, existe evidencia de una relación lineal (aunque no perfecta) sustancial entre las dos variables. ■

Los ejes horizontal y vertical en la gráfica de dispersión de la figura 9.1 se cortan en el punto (0, 0). En muchos conjuntos de datos los valores de x o y o los valores de ambas variables difieren considerablemente de cero respecto a los rangos de los valores. Por ejemplo, un estudio de cómo se relaciona la eficiencia de un equipo de aire acondicionado con la temperatura diaria máxima a la intemperie podría implicar observaciones de temperaturas desde 80 hasta 100°F. Cuando es este el caso, una gráfica más informativa mostraría los ejes apropiadamente marcados que intersecan en algún punto diferente de (0, 0).

EJEMPLO 9.2 El arsénico está presente en muchas aguas subterráneas y en algunas aguas superficiales. Investigaciones recientes sobre sus efectos en la salud ha llevado a la Agencia de Protección Ambiental a reducir los niveles permisibles de arsénico en el agua potable, de modo que muchos sistemas de agua ya no son compatibles con las normas. Este interés ha estimulado el desarrollo de métodos para eliminar el arsénico. Los siguientes datos respecto a x = pH

y y = arsénico eliminado (%) mediante un proceso en particular fueron tomados de una gráfica de dispersión que aparece en el artículo **"Optimizing Arsenic Removal During Iron Removal: Theoretical and Practical Considerations" (*J. of Water Supply Res. and Tech*., 2005: 545–560)**.

x	7.01	7.11	7.12	7.24	7.94	7.94	8.04	8.05	8.07
y	60	67	66	52	50	45	52	48	40

x	8.90	8.94	8.95	8.97	8.98	9.85	9.86	9.86	9.87
y	23	20	40	31	26	9	22	13	7

La figura 9.2 muestra dos gráficas de dispersión de estos datos obtenidas con Minitab. En la figura 9.2(a) Minitab seleccionó la escala para ambos ejes. La figura 9.2(b) se obtuvo especificando una escala para los ejes, de modo que se intersecan aproximadamente en el punto (0, 0). La segunda gráfica está más amontonada que la primera; tal amontonamiento hace más difícil valorar la naturaleza general de cualquier relación. Por ejemplo, puede ser más difícil descubrir la curvatura en una gráfica amontonada.

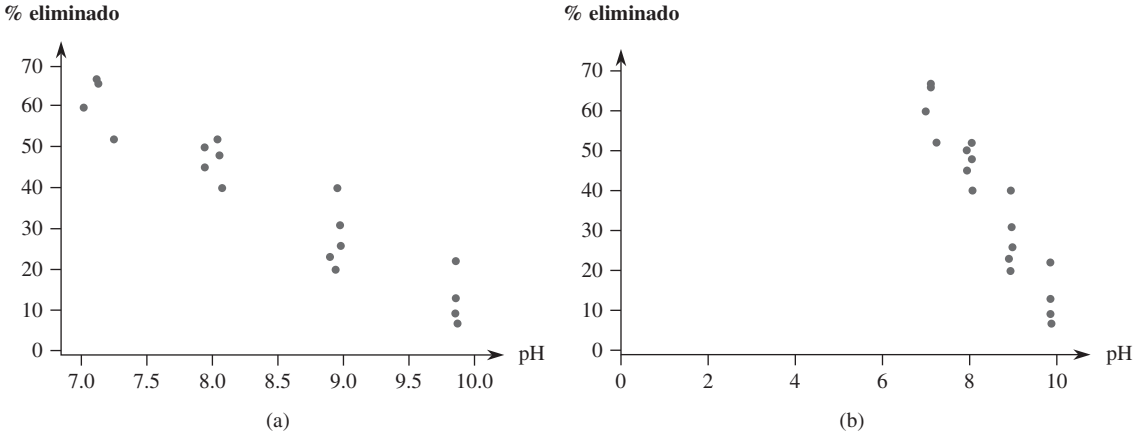


Figura 9.2 Gráficas de dispersión obtenidas con Minitab con los datos del ejemplo 9.2

Los grandes valores de arsénico eliminado tienden a asociarse con un bajo pH, una relación negativa o inversa. Además, las dos variables parecen estar al menos aproximadamente relacionadas de forma lineal, aunque los puntos en la gráfica se dispersarían en torno a cualquier línea recta sobrepuesta (dicha recta aparece en la gráfica en el artículo citado). ■

Modelo probabilístico lineal

Para el modelo determinístico $y = \beta_0 + \beta_1 x$, el valor observado real de y es una función lineal de x. La generalización apropiada de esto en un modelo probabilístico asume que *el valor esperado de Y es una función lineal de x*, pero para x fija la variable Y difiere de su valor esperado en una cantidad aleatoria.

DEFINICIÓN

Modelo de regresión lineal simple

Existen parámetros β_0, β_1 y σ^2 de tal suerte que con cualquier valor fijo de la variable independiente x, la variable dependiente es una variable aleatoria y está relacionada con x mediante la **ecuación modelo**

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{9.1}$$

La cantidad ϵ en la ecuación modelo es una variable aleatoria, la cual se supone que está normalmente distribuida con $E(\epsilon) = 0$ y $V(\epsilon) = \sigma^2$.

La variable ϵ se conoce como **término de error aleatorio** o **desviación aleatoria** en el modelo. Sin ϵ cualquier par observado (x, y) correspondería a un punto que queda exactamente sobre la recta $y = \beta_0 + \beta_1 x$, llamada **recta de regresión** (o de **población**) **verdadera**. La inclusión del término de error aleatorio permite a (x, y) quedar por encima de la recta de regresión (cuando $\epsilon > 0$) o por debajo de la recta (cuando $\epsilon < 0$). Los puntos (x_1, y_1),..., (x_n, y_n) provenientes de n observaciones independientes se dispersarán entonces en torno a la recta de regresión verdadera, como se ilustra en la figura 9.3. En ocasiones la conveniencia del modelo de regresión lineal simple puede sugerirse mediante consideraciones teóricas (p. ej., existe una relación lineal exacta entre las dos variables, con ϵ representando el error de medición). Con mucha más frecuencia, no obstante, la racionalidad del modelo se indica mediante una gráfica de dispersión que exhibe un patrón lineal sustancial (tal como en las figuras 9.1 y 9.2).

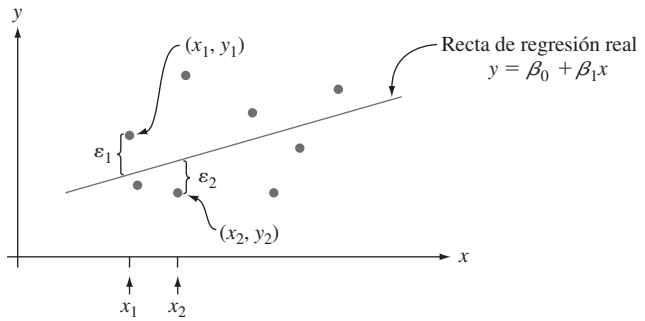


Figura 9.3 Puntos correspondientes a observaciones del modelo de regresión lineal simple

Las implicaciones de la ecuación modelo (9.1) se entienden mejor con la ayuda de la siguiente notación. Sea x^* un valor particular de la variable independiente x y

$$\mu_{Y \cdot x^*} = \text{el valor esperado (o media) de } Y \text{ cuando } x = x^*$$
$$\sigma^2_{Y \cdot x^*} = \text{la varianza de } Y \text{ cuando } x = x^*$$

La notación alternativa es $E(Y|x^*)$ y $V(Y|x^*)$. Por ejemplo, si x = esfuerzo aplicado (kg/mm)^2 y y = tiempo para la fractura (h), entonces $\mu_{Y \cdot 20}$ denotaría el valor esperado de tiempo para la fractura cuando se aplica un esfuerzo de 20 kg/mm^2. Si se piensa en una población completa de pares (x, y), entonces $\mu_{Y \cdot x^*}$ es la media de todos los valores y con los cuales $x = x^*$ y $\sigma^2_{Y \cdot x^*}$ es una medida de cuántos de estos valores de y se dispersan en torno a la media. Si

por ejemplo, x = edad de un niño y y = tamaño del vocabulario, entonces $\mu_{Y\cdot 5}$ es el tamaño de vocabulario promedio de todos los niños de 5 años que hay en la población y $\sigma^2_{Y\cdot 5}$ describe la cantidad de variabilidad del tamaño de vocabulario de esta parte de la población. Una vez que se fija x, la única aleatoriedad del lado derecho de la ecuación modelo (9.1) se encuentra en el error aleatorio $\boldsymbol{\epsilon}$ y su media y su varianza son 0 y σ^2, respectivamente, cualquiera que sea el valor de x. Esto implica que

$$\mu_{Y\cdot x^*} = E(\beta_0 + \beta_1 x^* + \boldsymbol{\epsilon}) = \beta_0 + \beta_1 x^* + E(\boldsymbol{\epsilon}) = \beta_0 + \beta_1 x^*$$

$$\sigma^2_{Y\cdot x^*} = V(\beta_0 + \beta_1 x^* + \boldsymbol{\epsilon}) = V(\beta_0 + \beta_1 x^*) + V(\boldsymbol{\epsilon}) = 0 + \sigma^2 = \sigma^2$$

Al reemplazar x^* en $\mu_{Y\cdot x^*}$ por x se obtiene la relación $\mu_{Y\cdot x} = \beta_0 + \beta_1 x$, la cual expresa que la *media de Y*, más que la Y misma, es una función lineal de x. La recta de regresión verdadera $y = \beta_0 + \beta_1 x$ es, por consiguiente, la *recta de las medias*; su altura por encima de cualquier valor x es el valor esperado de Y para ese valor de x. La pendiente β_1 de la recta de regresión verdadera se interpreta como el cambio *esperado* en Y asociado con el incremento de 1 unidad del valor de x. La segunda relación nos indica que la cantidad de variabilidad en la distribución de valores Y es la misma con cada valor diferente de x (homogeneidad de varianza). Si la variable independiente es el peso del vehículo y la variable dependiente es la eficiencia del combustible (mpg, millas por galón) entonces el modelo implica que el promedio de la eficiencia de combustible cambia linealmente con el peso (probablemente β_1 es negativa) y que la cantidad de variabilidad de la eficiencia para cualquier peso particular es la misma que para cualquier otro peso. Por último, con x fija, Y es la suma de una constante $\beta_0 + \beta_1 x$ y una variable aleatoria $\boldsymbol{\epsilon}$ normalmente distribuida, por tanto, tiene una distribución normal. Estas propiedades se ilustran en la figura 9.4. El parámetro de varianza σ^2 determina el grado al cual cada curva normal se extiende en torno a su media; en términos generales, el valor de σ es el tamaño de una desviación típica de la recta de regresión verdadera. Cuando σ es pequeña, un punto observado (x, y) normalmente quedará bastante cerca de la recta de regresión verdadera, en tanto que las observaciones pueden desviarse considerablemente de sus valores esperados (correspondientes a puntos alejados de la recta) cuando σ es grande.

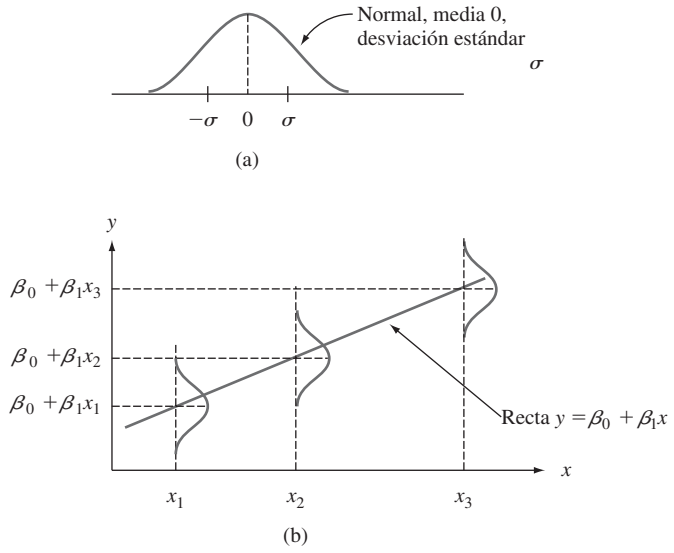


Figura 9.4 (a) Distribución de $\boldsymbol{\epsilon}$; (b) distribución de Y con diferentes valores de x

EJEMPLO 9.3 Suponga que el modelo de regresión lineal simple con una recta de regresión verdadera $y = 65 - 1.2x$ y $\sigma = 8$ describe la relación entre el esfuerzo aplicado x y el tiempo para la fractura y. Así, para cualquier valor fijo x^* de esfuerzo, el tiempo para la fractura tiene una distribución normal con media de $65 - 1.2x^*$ y desviación estándar 8. En la población de todos los puntos (x, y), la magnitud de una desviación típica de la recta de regresión verdadera es alrededor de 8. Con $x = 20$, Y tiene la media $\mu_{Y\cdot 20} = 65 - 1.2(20) = 41$, por tanto,

$$P(Y > 50 \text{ cuando } x = 20) = P\left(Z > \frac{50 - 41}{8}\right) = 1 - \Phi(1.13) = 0.1292$$

Debido a que $\mu_{Y\cdot 25} = 35$,

$$P(Y > 50 \text{ cuando } x = 25) = P\left(Z > \frac{50 - 35}{8}\right) = 1 - \Phi(1.88) = 0.0301$$

Estas probabilidades se ilustran como las áreas sombreadas en la figura 9.5.

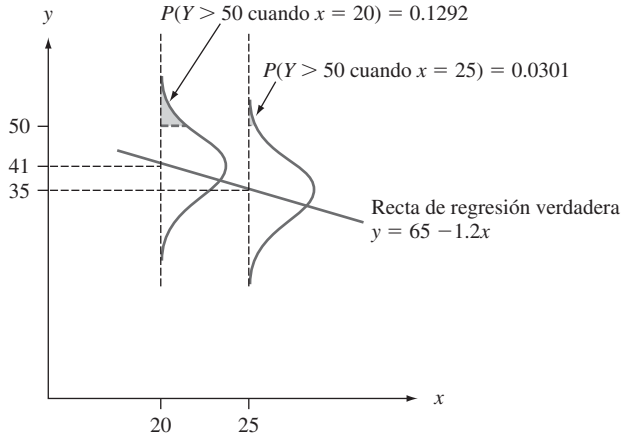


Figura 9.5 Probabilidades basadas en el modelo de regresión lineal simple

Suponga que Y_1 denota una observación del tiempo para la fractura realizada con $x = 25$ y que Y_2 denota una observación independiente realizada con $x = 24$. Entonces $Y_1 - Y_2$ está normalmente distribuida con media $E(Y_1 - Y_2) = \beta_1 = -1.2$, varianza $V(Y_1 - Y_2) = \sigma^2 + \sigma^2 = 128$ y desviación estándar $\sqrt{128} = 11.314$. La probabilidad de que Y_1 exceda Y_2 es

$$P(Y_1 - Y_2 > 0) = P\left(Z > \frac{0 - (-1.2)}{11.314}\right) = P(Z > 0.11) = 0.4562$$

Es decir, aunque se espera que Y disminuya a medida que x se incrementa en 1 unidad, no es improbable que la Y observada en $x + 1$ sea más grande que la Y observada en x. ■

EJERCICIOS **Sección 9.1 (1–11)**

1. La relación de eficiencia de un espécimen de acero sumergido en un tanque de fosfatado es el peso del recubrimiento de fosfato dividido entre la pérdida de metal (ambos en mg/pie^2). El artículo **"Statistical Process Control of a Phosphate Coating Line" (*Wire J. Intl*., mayo de 1997: 78–81)** aporta los siguientes datos sobre la temperatura del tanque (x) y relación de eficiencia (y):

Temp.	170	172	173	174	174	175	176
Relación	0.84	1.31	1.42	1.03	1.07	1.08	1.04
Temp.	177	180	180	180	180	180	181
Relación	1.80	1.45	1.60	1.61	2.13	2.15	0.84
Temp.	181	182	182	182	182	184	184
Relación	1.43	0.90	1.81	1.94	2.68	1.49	2.52
Temp.	185	186	188				
Relación	3.00	1.87	3.08				

a. Construya gráficas de tallo y hojas tanto de la relación de temperatura como de la relación de eficiencia, y comente las características interesantes.

b. El valor de la relación de eficiencia ¿estará determinado por completo y de forma única por la temperatura del tanque? Explique su razonamiento.

c. Construya una gráfica de dispersión de los datos. ¿Será que la relación de eficiencia bien podría pronosticarse mediante el valor de la temperatura? Explique su razonamiento.

2. El artículo **"Exhaust Emissions from Four-Stroke Lawn Mower Engines" (*J. of the Air and Water Mgmnt. Assoc*., 1997: 945–952)** reporta los datos de un estudio en el cual se utilizó una mezcla de gasolinas básicas y una gasolina reformulada. Considere las siguientes observaciones sobre edad (años) y emisiones de NO_x (g/kWh):

Motor	1	2	3	4	5
Edad	0	0	2	11	7
Línea de base	1.72	4.38	4.06	1.26	5.31
Reformulada	1.88	5.93	5.54	2.67	6.53
Motor	6	7	8	9	10
Edad	16	9	0	12	4
Línea de base	0.57	3.37	3.44	0.74	1.24
Reformulada	0.74	4.94	4.89	0.69	1.42

Construya gráficas de dispersión de emisiones de NO_x contra edad. ¿Cuál parece ser la naturaleza de la relación entre estas dos variables? [*Nota*: Los autores del artículo citado hacen comentarios sobre la relación.]

3. A menudo surgen datos bivariantes cuando se utilizan dos técnicas diferentes de medir la misma cantidad. Por ejemplo, las observaciones adjuntas de x = concentración de hidrógeno (ppm) mediante un método de cromatografía de gases y y = concentración mediante un nuevo método de sensor se leyeron en una gráfica que se muestra en el artículo **"A New Method to Measure the Diffusible Hydrogen Content in Steel Weldments Using a Polymer Electrolyte-Based Hydrogen Sensor" (*Welding Res*., julio de 1997: 251s–256s)**.

x	47	62	65	70	70	78	95	100	114	118
y	38	62	53	67	84	79	93	106	117	116
x	124	127	140	140	140	150	152	164	198	221
y	127	114	134	139	142	170	149	154	200	215

Construya una gráfica de dispersión. ¿Habrá una fuerte relación entre los dos tipos de mediciones de concentración? ¿Será que los dos métodos miden aproximadamente la misma cantidad? Explique su razonamiento.

4. Los datos adjuntos en y = concentración de amonio (mg/L) y x = transpiración (ml/h) se toman de una gráfica que aparece en el artículo **"Response of Ammonium Removal to Growth and Transpiration of Juncus effusus During the Treatment of Artificial Sewage in Laboratory-Scale Wetlands" (*Water Research*, 2013: 4265–4273)**. El resumen del artículo indica "una correlación lineal entre la concentración de amonio dentro de la rizosfera y la transpiración de las poblaciones de plantas implica que existe una influencia de actividad fisiológica de las plantas sobre la eficiencia de N remociones". (La rizosfera es la estrecha superficie en que las raíces de las plantas interaccionan con el suelo y la transpiración es el proceso de movimiento del agua a través de una planta y su evaporación). El artículo reporta las cantidades resumidas de un análisis de regresión lineal simple. Con base en un diagrama de dispersión, ¿cómo describiría la relación entre variables y será que la regresión lineal simple es una estrategia de modelado apropiada?

x	5.8	8.8	11.0	13.6	18.5	21.0	23.7
y	7.8	8.2	6.9	5.3	4.7	4.9	4.3
x	26.0	28.3	31.9	36.5	38.2	40.4	
y	2.7	2.8	1.8	1.9	1.1	0.4	

5. El artículo **"Objective Measurement of the Stretchability of Mozzarella Cheese" (*J. of Texture Studies*, 1992: 185–194)** reporta sobre un experimento para investigar la variación del comportamiento del queso mozzarella con la temperatura. Considere los siguientes datos sobre x = temperatura y y = alargamiento (%) en el momento de la falla del queso.

[*Nota*: Los investigadores eran italianos y utilizaron queso mozzarella *real*, no la imitación que se vende en los Estados Unidos.]

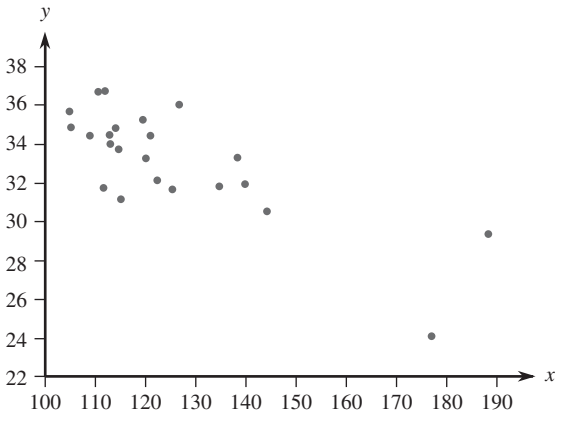
x	59	63	68	72	74	78	83
y	118	182	247	208	197	135	132

a. Construya una gráfica de dispersión en la cual los ejes se corten en (0, 0). Marque 0, 20, 40, 60, 80 y 100 en el eje horizontal y 0, 50, 100, 150, 200 y 250 en el eje vertical.

b. Construya una gráfica de dispersión en la cual los ejes se corten en (55, 100), como se hizo en el citado artículo. ¿Será preferible esta gráfica a la del inciso a)? Explique su razonamiento.

c. ¿Qué sugieren las gráficas de los incisos a) y b) sobre la naturaleza de la relación entre las dos variables?

6. Un factor en el desarrollo del codo de tenista, una dolencia que provoca profundo terror a los tenistas serios, es la vibración inducida por el impacto del sistema raqueta-brazo al contacto con la pelota. Es bien sabido que la probabilidad de sufrir codo de tenista depende de varias propiedades de la raqueta que se utilice. Considere la gráfica de dispersión de x = frecuencia de resonancia de la raqueta y y = suma de la aceleración pico a pico (una característica de la vibración del brazo, en m/s/s) de $n = 23$ raquetas diferentes (**"Transfer of Tennis Racket Vibrations into the Human Forearm", *Medicine and Science in Sports and Exercise*, 1992: 1134–1140**). Analice las características interesantes de los datos y la gráfica de dispersión.



7. El artículo **"Some Field Experience in the Use of an Accelerated Method in Estimating 28-Day Strength of Concrete" (*J. of Amer. Concrete Institute*, 1969: 895)** considera la regresión de y = resistencia estándar después de 28 días de curado (lb/pulg^2) contra x = resistencia acelerada (lb/pulg^2). Suponga que la ecuación de la recta de regresión verdadera es $y = 1800 + 1.3x$.

a. ¿Cuál es el valor esperado de la resistencia después de 28 días cuando la resistencia acelerada = 2500?

b. ¿En cuánto se debe esperar que cambie la resistencia después de 28 días cuando la resistencia acelerada se incrementa en 1 lb/pulg^2?

c. Responda el inciso b) para un incremento de 100 lb/pulg^2.

d. Responda el inciso b) para una reducción de 100 lb/pulg^2.

8. Recurra al ejercicio 7 y suponga que la desviación estándar de la desviación aleatoria ϵ es de 350 lb/pulg^2.

a. ¿Cuál es la probabilidad de que el valor observado de la resistencia después de 28 días exceda 5000 lb/pulg^2 cuando el valor de la resistencia acelerada es de 2000?

b. Repita el inciso a) con 2500 en lugar de 2000.

c. Considere hacer dos observaciones independientes de resistencia después de 28 días, la primera con una resistencia acelerada de 2000 y la segunda con $x = 2500$. ¿Cuál es la probabilidad de que la segunda observación exceda la primera por más de 1000 lb/pulg^2?

d. Sean Y_1 y Y_2 las observaciones de resistencia después de 28 días cuando $x = x_1$ y $x = x_2$, respectivamente. ¿Por cuánto tendría que exceder x_2 a x_1 para que $P(Y_2 > Y_1) = 0.95$?

9. La velocidad de flujo y(m^3/min) en un dispositivo utilizado para medir la calidad del aire depende de la caída de presión x (pulg. de agua) a través del filtro del dispositivo. Suponga que con valores de x entre 5 y 20, las dos variables están relacionadas de acuerdo con el modelo de regresión lineal simple con recta de regresión verdadera $y = -0.12 + 0.095x$.

a. ¿Cuál es el cambio esperado de la velocidad de flujo asociado con un incremento de 1 pulg en la caída de presión? Explique.

b. ¿Qué cambio de la velocidad de flujo puede ser esperado cuando la caída de presión se reduce en 5 pulg?

c. ¿Cuál es la velocidad de flujo esperada con una caída de presión de 10 pulg? ¿Y en una caída de presión de 15 pulg?

d. Suponga $\sigma = 0.025$ y considere una caída de presión de 10 pulg. ¿Cuál es la probabilidad de que el valor observado de la velocidad de flujo exceda 0.835?, ¿y de que la velocidad de flujo observada exceda 0.840?

e. ¿Cuál es la probabilidad de que una observación de la velocidad de flujo, cuando la caída de presión es de 10 pulg, exceda una observación de la velocidad de flujo cuando la caída de presión es de 11 pulg?

10. Suponga que el costo esperado de una corrida de producción está relacionado con el tamaño de la corrida mediante la ecuación $y = 4000 + 10x$. Sea Y una observación sobre el costo de una corrida. Si el *tamaño* de las variables y el *costo* están relacionados de acuerdo con el modelo de regresión lineal simple, ¿podría ser el caso de que $P(Y > 5500$ cuando $x = 100) = 0.05$ y $P(Y > 6500$ cuando $x = 200) = 0.10$? Explique.

11. Suponga que en un cierto proceso químico el tiempo de reacción y (h) está relacionado con la temperatura (°F) en la cámara en la cual la reacción ocurre de acuerdo con el modelo de regresión lineal simple con la ecuación $y = 5.00 - 0.01x$ y $\sigma = 0.075$.

a. ¿Cuál es el cambio esperado del tiempo de reacción con un incremento de 1°F de la temperatura? ¿Con un incremento de 10°F de la temperatura?

b. ¿Cuál es el tiempo de reacción esperado cuando la temperatura es de 200°F? ¿Y cuando la temperatura es de 250°F?

c. Suponga que se realizan cinco observaciones independientemente del tiempo de reacción, cada una para una temperatura de 250°F. ¿Cuál es la probabilidad de que las cinco observaciones resulten entre 2.4 y 2.6 h?

d. ¿Cuál es la probabilidad de que dos tiempos de reacción independientemente observados a temperaturas con 1° de diferencia sean tales que el tiempo a la temperatura más alta exceda el tiempo a la temperatura más baja?

9.2 Estimación de parámetros de modelo

En esta y en las siguientes secciones se supondrá que las variables x y y están relacionadas de acuerdo con el modelo de regresión lineal simple. Un investigador casi nunca conocerá los valores de β_0, β_1 o σ^2. En cambio, estará disponible una muestra de datos compuesta de n pares observados $(x_1, y_1),..., (x_n, y_n)$, con la cual los parámetros de modelo y la recta de regresión verdadera pueden ser estimados. Se supone que estas observaciones se obtuvieron independientemente una de otra. Es decir, y_i es el valor observado de Y_i, donde $Y_i = \beta_{0+}\beta_1 x_1 + \boldsymbol{\epsilon}_i$ y las n desviaciones $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2,..., \boldsymbol{\epsilon}_n$ son variables aleatorias independientes. La independencia de $Y_1, Y_2,..., Y_n$ se desprende de la independencia de las $\boldsymbol{\epsilon}_i$.

De acuerdo con el modelo los puntos observados estarán distribuidos en torno a la recta de regresión verdadera de una manera aleatoria. La figura 9.6 muestra una gráfica típica de pares observados junto con dos candidatos para la recta de regresión estimada. Intuitivamente, la recta $y = a_0 + a_1x$ no es una estimación razonable de la recta verdadera $y = \beta_0 + \beta_1x$ porque si $y = a_0 + a_1x$ fuera la recta verdadera, los puntos observados con toda seguridad habrían quedado más cerca de la misma. La recta $y = b_0 + b_1x$ es una estimación más factible porque los puntos observados están dispersos, en lugar de estar cerca de esta recta.

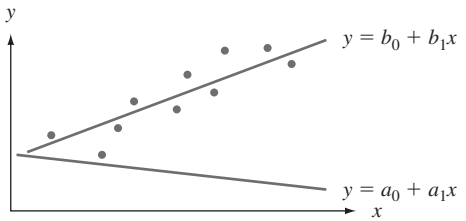


Figura 9.6 Dos estimaciones diferentes de la recta de regresión verdadera

La figura 9.6 y la discusión anterior sugieren que la estimación de $y = \beta_0 + \beta_1x$ deberá ser una recta que en un cierto sentido se ajuste mejor a los puntos de los datos observados. Esto es lo que motiva el principio de mínimos cuadrados el cual puede rastrearse hacia atrás en el tiempo hasta el matemático alemán Gauss (1777–1855). De acuerdo con este principio, una recta proporciona un buen ajuste para los datos si las distancias verticales (desviaciones) de los puntos observados a la recta son pequeñas (véase la figura 9.7). La medida de la bondad del ajuste es la suma de cuadrados de estas desviaciones. La recta de mejor ajuste es entonces la que tiene la suma más pequeña posible de desviaciones al cuadrado.

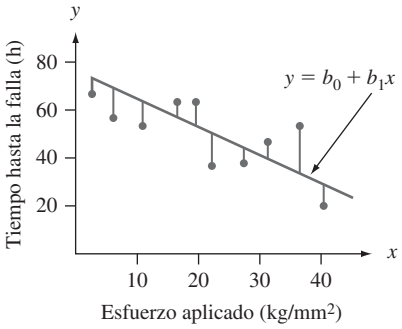


Figura 9.7 Desviaciones de los datos observados respecto a la recta $y = b_0 + b_1x$

Principio de mínimos cuadrados

La desviación vertical del punto (x_i, y_i) respecto a la recta $y = b_0 + b_1x$ es

$$\text{altura del punto} - \text{altura de la recta} = y_i - (b_0 + b_1x_i)$$

La suma de las desviaciones verticales al cuadrado de los puntos $(x_1, y_1),..., (x_n, y_n)$ a la recta es entonces

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1x_i)]^2$$

Las estimaciones puntuales de β_0 y β_1, denotadas por $\hat{\beta}_0$ y $\hat{\beta}_1$, llamadas **estimaciones de mínimos cuadrados**, son aquellos valores que reducen al mínimo a $f(b_0, b_1)$. Es decir, $\hat{\beta}_0$ y $\hat{\beta}_1$ son tales que $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ con cualesquier b_0 y b_1. La **recta de regresión estimada** o **recta de mínimos cuadrados** es entonces la recta cuya ecuación es $y = \hat{\beta}_0 + \hat{\beta}_1x$.

Los valores minimizados de b_0 y b_1 se encuentran tomando las derivadas parciales de $f(b_0, b_1)$ respecto tanto a b_0 como b_1, igualándolas a cero [análogamente a $f'(b) = 0$ en cálculo de una variable] y resolviendo las ecuaciones

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1x_i)(-x_i) = 0$$

Al cancelar el factor -2 y reordenar se obtiene el siguiente sistema de ecuaciones, llamado **ecuaciones normales:**

$$nb_0 + \left(\sum x_i\right)b_1 = \sum y_i$$

$$\left(\sum x_i\right)b_0 + \left(\sum x_i^2\right)b_1 = \sum x_iy_i$$

Estas ecuaciones son lineales en las dos incógnitas b_0 y b_1. Siempre que no todas las x_i sean idénticas, las estimaciones de mínimos cuadrados son la única solución de este sistema.

PROPOSICIÓN

La estimación de mínimos cuadrados del coeficiente de la pendiente β_1 de la recta de regresión verdadera es

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad \textbf{(9.2)}$$

Al calcular las fórmulas para el numerador y el denominador de $\hat{\beta}_1$ son

$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n \qquad S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2/n$$

La estimación de mínimos cuadrados de la intersección β_0 de la recta de regresión verdadera es

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \textbf{(9.3)}$$

Las fórmulas para el cálculo de S_{xy} y S_{xx} requieren sólo los estadísticos resumidos Σx_i, Σy_i, Σx_i^2 y $\Sigma x_i y_i$ (más adelante se requerirá Σy_i^2). Al calcular $\hat{\beta}_0$ se utilizan dígitos adicionales en $\hat{\beta}_1$ porque, si $\bar{x}$ es grande en magnitud, el redondeo afectará la respuesta final. En la práctica es preferible usar un paquete de software estadístico en lugar del cálculo manual y los gráficos trazados a mano. Una vez más, asegúrese de que la gráfica de dispersión muestre un patrón lineal con una variación relativamente homogénea antes de ajustar el modelo de regresión lineal simple.

EJEMPLO 9.4 El número de cetano es una propiedad fundamental en la especificación de la calidad de ignición del combustible que se utiliza en un motor diesel. La determinación de este número para un combustible biodiesel es cara y lleva mucho tiempo. El artículo **"Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (*J. of Automobile Engr.*, 2009: 565–583)** incluye los siguientes datos en x = índice de yodo (g) y y = número de cetano para una muestra de 14 biocombustibles. El índice de yodo es la cantidad de yodo necesario para saturar una muestra de 100 g de aceite. Los autores del artículo ajustan el modelo de regresión lineal simple a estos datos, así que vamos a seguir su ejemplo.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Las cantidades resumidas necesarias para el cálculo manual se puede obtener mediante la colocación de los valores de x en una columna y los valores de y en otra y luego crear columnas para x^2, xy y y^2 (por el momento estos últimos valores no son necesarios, pero se utilizarán en breve). Al calcular las sumas por columna, tenemos $\Sigma x_i = 1307.5$, $\Sigma y_i = 779.2$, $\Sigma x_i^2 = 128\,913.93$, $\Sigma x_i y_i = 71\,347.30$, $\Sigma y_i^2 = 43\,745.22$, de donde

$$S_{xx} = 128\,913.93 - (1307.5)^2/14 = 6802.7693$$
$$S_{xy} = 71\,347.30 - (1307.5)(779.2)/14 = -1424.41429$$

La pendiente estimada de la recta de regresión real (es decir, la pendiente de la recta de mínimos cuadrados) es

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -0.20938742$$

Estimamos que el cambio esperado en el promedio real del número de cetano asociado con un incremento de 1 g en el índice de yodo es −0.209, es decir, una disminución de 0.209. Puesto que $\bar{x} = 93.392857$ y $\bar{y} = 55.657143$, la intersección estimada de la recta de regresión real (es decir, la intersección de la recta de mínimos cuadrados) es

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 55.657143 - (-0.20938742)(93.392857) = 75.212432$$

La ecuación de la recta de regresión estimada (recta de mínimos cuadrados) es $y = 75.212 - 0.209x$, exactamente la descrita en el artículo citado. La figura 9.8 muestra un diagrama de dispersión de los datos con la recta de mínimos cuadrados superpuesta. Esta recta ofrece un resumen muy bueno de la relación entre las dos variables.

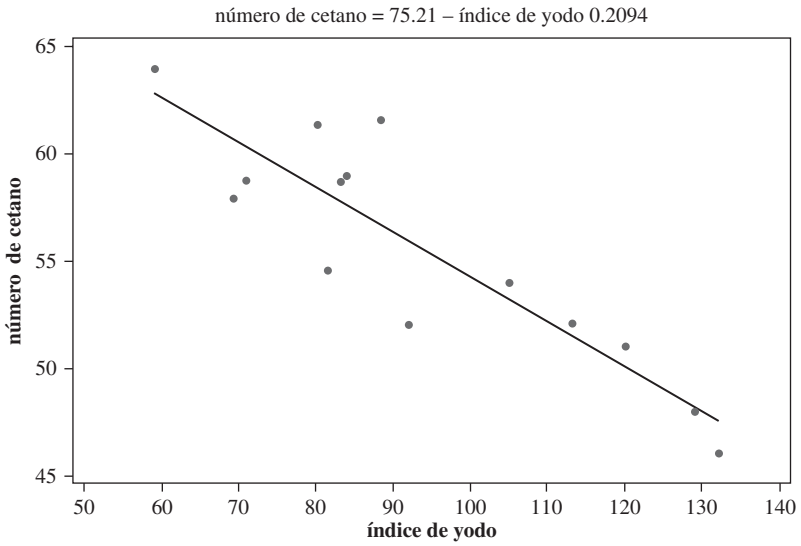


Figura 9.8 Diagrama de dispersión de los datos con la recta de mínimos cuadrados sobrepuesta con Minitab para el ejemplo 9.4

La recta de regresión estimada puede ser utilizada de manera inmediata para dos propósitos diferentes. Con un valor fijo de x, x^*, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ (la altura de la recta sobre x^*) da 1) una estimación puntual del valor esperado de Y cuando $x = x^*$; o 2) una predicción puntual del valor Y que resultará de una nueva observación realizada con $x = x^*$.

EJEMPLO 9.5 Remítase al escenario del valor del índice de yodo para el número de cetano descrito en el ejemplo anterior. La ecuación de regresión estimada fue $y = 75.212 - 0.2094x$. Una estimación puntual del verdadero número de cetano promedio de todos los biocombustibles, cuyo índice de yodo es 100 es

$$\hat{\mu}_{Y \cdot 100} = \hat{\beta}_0 + \hat{\beta}_1(100) = 75.212 - 0.2094(100) = 54.27$$

Si se selecciona una muestra de biocombustible cuyo índice de yodo es 100, también 54.27 es un punto de predicción para el número de cetano resultante.

La recta de mínimos cuadrados no deberá utilizarse para predecir un valor de x mucho más allá del rango de los datos, de tal suerte que $x = 40$ o $x = 150$ en el ejemplo 9.4. El **peligro de extrapolación** es que la relación ajustada (una recta en este caso) puede no ser válida para tales valores de x.

Estimación de σ^2 y σ

El parámetro σ^2 determina la cantidad de variabilidad, inherente en el modelo de regresión. Un valor grande de σ^2 conducirá a (x_i, y_i) observados que están bastante dispersos en torno a la recta de regresión verdadera; mientras σ^2 sea pequeña los puntos observados tenderán a quedar cerca de la recta verdadera (véase la figura 9.9). Se utilizará una estimación de σ^2 en las fórmulas de los intervalos de confianza (IC) y los procedimientos de prueba de hipótesis que se presentan en las dos secciones siguientes. Debido a que la ecuación de la recta verdadera es desconocida, la estimación se basa en el grado al cual las observaciones muestrales se desvían de la recta *estimada*. Muchas desviaciones grandes (residuos) sugieren un valor grande de σ^2, mientras que las desviaciones de pequeña magnitud sugieren que σ^2 es pequeña.

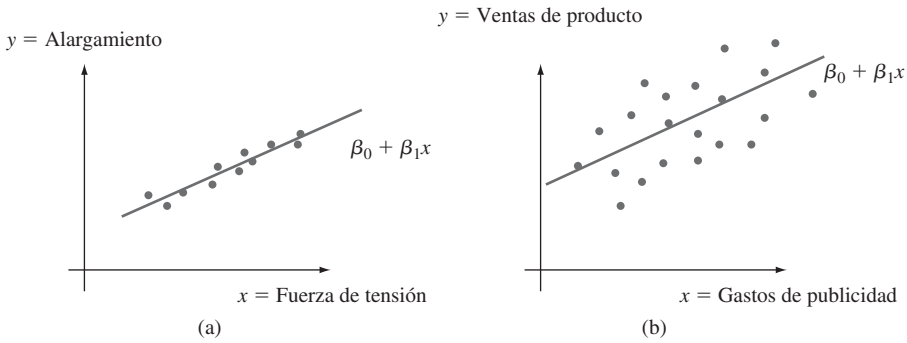


Figura 9.9 Muestra típica para σ^2: (a) pequeña; (b) grande

DEFINICIÓN

Los **valores ajustados** (o **pronosticados**) $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$,se obtienen sustituyendo sucesivamente $x_1, \ldots, x_n$ en la ecuación de la recta de regresión estimada: $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$. Los **residuos** son las diferencias $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$ entre los valores observados y los valores ajustados y.

Es decir, el valor pronosticado $\hat{y}_i$ es el valor de y pronosticado o esperado cuando se utiliza la recta de regresión estimada con $x = x_i$; $\hat{y}_i$ es la altura de la recta de regresión estimada por encima del valor x_i con el cual se realizó la i-ésima observación. El residuo $y_i - \hat{y}_i$ entra es la desviación vertical entre el punto (x_i, y_i) y la recta de mínimos cuadrados es un número positivo si el punto está sobre la recta y negativo si está debajo de esta. Si todos los residuos son pequeños, entonces mucha de la variabilidad en los valores y observados al parecer se debe a la relación lineal entre x y y, mientras que muchos residuos grandes sugieren un poco de variabilidad inherente en y respecto a la cantidad debida a la relación lineal. Suponiendo que la recta en la figura 9.7 es la recta de mínimos cuadrados, los residuos están identificados por segmentos de recta verticales que parten de los puntos observados a la recta. Cuando se obtiene la recta de regresión estimada mediante el principio de mínimos cuadrados, la suma de los residuos en teoría debe ser cero. En la práctica, la suma puede desviarse un poco de cero debido al redondeo.

EJEMPLO 9.6 La alta densidad de población de Japón ha provocado un sinnúmero de problemas de consumo de recursos. Una dificultad especialmente seria tiene que ver con la eliminación de los desechos. El artículo **"Innovative Sludge Handling Through Pelletization Thickening" (*Water Research*, 1999: 3245–3252)** reporta el desarrollo de una nueva máquina de compresión para procesar lodos de albañal. Una parte importante de la investi-

gación implicó relacionar el contenido de humedad de gránulos comprimidos (y, en %) con la velocidad de filtración de la máquina (x, en kg-DS/m/h). Los siguientes datos se tomaron de una gráfica que muestra el artículo:

x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Las cantidades resumidas pertinentes (*estadísticos resumidos*) son $\Sigma x_i = 2817.9$, $\Sigma y_i = 1574.8$, $\Sigma x_i^2 = 415\,949.85$, $\Sigma x_i y_i = 222\,657.88$ y $\Sigma y_i^2 = 124\,039.58$, de donde $\bar{x} = 140.895$, $\bar{y} = 78.74$, $S_{xx} = 18\,921.8295$ y $S_{xy} = 776.434$. Por tanto,

$$\hat{\beta}_1 = \frac{776.434}{18\,921.8295} = 0.04103377 \approx 0.041$$

$$\hat{\beta}_0 = 78.74 - (0.04103377)(140.895) = 72.958547 \approx 72.96$$

por lo que la ecuación de la recta de mínimos cuadrados es $y = 72.96 + 0.041x$. Para precisión numérica, los valores ajustados se calcularon con $\hat{y}_i = 72.958547 + 0.04103377x_i$:

$$\hat{y}_1 = 72.958547 + 0.04103377(125.3) \approx 78.100,\ y_1 - \hat{y}_1 \approx -0.200, \text{ etc.}$$

Nueve de los 20 residuos son negativos, por lo que los nueve puntos correspondientes en un diagrama de dispersión de los datos se encuentran por debajo de la recta de regresión estimada. Todos los valores previstos (ajustes) y residuos aparecen en la siguiente tabla.

Obs	Filtrado	Contenido de humedad	Ajuste	Residuo
1	125.3	77.9	78.100	−0.200
2	98.2	76.8	76.988	−0.188
3	201.4	81.5	81.223	0.277
4	147.3	79.8	79.003	0.797
5	145.9	78.2	78.945	−0.745
6	124.7	78.3	78.075	0.225
7	112.2	77.5	77.563	−0.063
8	120.2	77.0	77.891	−0.891
9	161.2	80.1	79.573	0.527
10	178.9	80.2	80.299	−0.099
11	159.5	79.9	79.503	0.397
12	145.8	79.0	78.941	0.059
13	75.1	76.7	76.040	0.660
14	151.4	78.2	79.171	−0.971
15	144.2	79.5	78.876	0.624
16	125.0	78.1	78.088	0.012
17	198.8	81.5	81.116	0.384
18	132.5	77.0	78.396	−1.396
19	159.6	79.0	79.508	−0.508
20	110.7	78.6	77.501	1.099

Casi de la misma forma en que las desviaciones de la media en en el caso de la muestra se combinaron para obtener la estimación $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$, la estimación de σ^2 en un análisis de regresión se basa en elevar al cuadrado y sumar los residuos. Para esta varianza estimada se continuará utilizando el símbolo s^2, así que no hay que confundirla con la s^2 previa.

DEFINICIÓN

La **suma de cuadrados debido al error** (o su equivalente, la suma de cuadrados residuales) denotada por SSE, es

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

y la estimación de σ^2 es

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

El divisor $n - 2$ en s^2 es el número de grados de libertad (gl) asociado con SSE y la estimación s^2. Esto es porque para obtener s^2 primero se deben estimar los dos parámetros β_0 y β_1, lo que hace que se pierdan 2 grados de libertad (tal y como μ tuvo que estimarse en los problemas de la muestra, con el resultado de una varianza estimada basada en $n - 1$ grados de libertad). Al sustituir cada y_i en la fórmula para s^2 mediante la variable aleatoria Y_i, obtenemos el estimador S^2. Se puede demostrar que S^2 es un estimador insesgado para σ^2 (aunque el estimador S no sea insesgado para σ). Aquí la interpretación de s es similar a lo que se sugirió antes para la desviación estándar muestral: a grandes rasgos, es el tamaño de una desviación típica vertical dentro de la muestra desde la recta de regresión estimada.

EJEMPLO 9.7 Los residuos de los datos de contenido de filtración humedad-velocidad fueron calculados con anterioridad. La suma de cuadrados debido al error correspondiente es

$$\text{SSE} = (-0.200)^2 + (-0.188)^2 + \cdots + (1.099)^2 = 7.968$$

La estimación de σ^2 es entonces $\hat{\sigma}^2 = s^2 = 7.968/(20-2) = 0.4427$ y la desviación estándar estimada es $\hat{\sigma} = s = \sqrt{0.4427} = 0.665$. En términos generales, 0.665 es la magnitud de una desviación típica respecto a la recta de regresión estimada; algunos puntos están cerca de la recta así como otros están alejados. ■

El cálculo de SSE con la fórmula de definición implica mucha aritmética tediosa, ya que primero se deben calcular los valores pronosticados y los residuos. La siguiente fórmula de cálculo no requiere estas cantidades

$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i = S_{yy} - \hat{\beta}_1 S_{xy}$$

Esta expresión de en medio se obtiene al sustituir $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ por $\Sigma(y_i - \hat{y}_i)^2$, elevando al cuadrado el sumando, realizando la suma y continuarla hasta obtener los tres términos resultantes y simplificar. La fórmula de cálculo es especialmente sensible a los efectos de redondeo en $\hat{\beta}_0$ y $\hat{\beta}_1$, así que conservar tantos dígitos como sea posible en cálculos intermedios protegerá contra errores de redondeo.

EJEMPLO 9.8 El artículo **"Promising Quantitative Nondestructive Evaluation Techniques for Composite Materials" (*Materials Evaluation*, 1985: 561–565)** reporta sobre el estudio para investigar cómo la propagación de una onda de esfuerzo ultrasónica a través de una sustancia depende de las propiedades de esta última. Los siguientes datos sobre la resistencia a la fractura (x, como porcentaje de resistencia a la tensión más reciente) y la atenuación (y, en neper/cm, la disminución de la amplitud de la onda de esfuerzo) en compuestos de poliéster reforzados con fibra de vidrio se tomaron de una gráfica que aparece en el artículo. El patrón lineal sustancial que se incluye en la gráfica de dispersión sugiere el modelo de regresión lineal simple.

x	12	30	36	40	45	57	62	67	71	78	93	94	100	105
y	3.3	3.2	3.4	3.0	2.8	2.9	2.7	2.6	2.5	2.6	2.2	2.0	2.3	2.1

Las cantidades necesarias resumidas son $n = 14$, $\Sigma x_i = 890$, $\Sigma x_i^2 = 67\,182$, $\Sigma y_i = 37.6$, $\Sigma y_i^2 = 103.54$ y $\Sigma x_i y_i = 2234.30$, de donde $S_{xx} = 10\,603.4285714$, $S_{xy} = -155.98571429$, $\hat{\beta}_1 = -0.0147109$ y $\hat{\beta}_0 = 3.6209072$. Por tanto,

$$\begin{aligned} \text{SSE} &= 103.54 - (3.6209072)(37.6) - (-0.0147109)(2234.30) \\ &= 0.2624532 \end{aligned}$$

Se obtiene el mismo resultado de

$$\text{SSE} = S_{yy} - \hat{\beta}_1 S_{xy} = 103.54 - (37.6)^2/14 - (-0.0147109)(-155.98571429)$$

Así $s^2 = 0.2624532/12 = 0.0218711$ y $s = 0.1479$. Cuando $\hat{\beta}_0$ y $\hat{\beta}_1$ se redondean a tres cifras decimales en la fórmula de cálculo para SSE, el resultado es

$$\text{SSE} = 103.54 - (3.621)(37.6) - (-0.015)(2234.30) = 0.905$$

la cual es más de tres veces el valor correcto. ■

Coeficiente de determinación

La figura 9.10 muestra tres diferentes gráficas de dispersión de datos bivariantes. En las tres gráficas las alturas de los diferentes puntos varían sustancialmente, lo que indica que existe mucha variabilidad en los valores y observados. Todos los puntos en la primera gráfica quedan exactamente en una línea recta. En este caso toda la variación (100%) de y puede ser atribuida al hecho de que x y y están linealmente relacionadas en combinación con la variación de x. Los puntos en la figura 9.10(b) no están exactamente en una recta, pero su variabilidad se compara a la variabilidad total de y, y las desviaciones respecto a la recta de mínimos cuadrados son pequeñas. Es razonable concluir en este caso que gran parte de la variación de y observada puede atribuirse a la relación lineal aproximada entre las variables postuladas por el modelo de regresión lineal simple. Cuando la gráfica de dispersión es como la de la figura 9.10(c) existe una variación sustancial en torno a la recta de mínimos cuadrados respecto a la variación total de y, así que el modelo de regresión lineal simple no explica la variación de y al relacionar y con x.

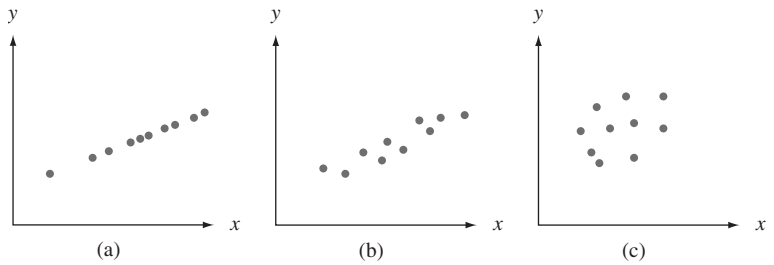


Figura 9.10 Utilización del modelo para explicar la variación de y: (a) datos para los que se explica toda la variación; (b) datos para los cuales se explica la mayor parte de la variación; (c) datos para los cuales se explica poca variación

La suma de cuadrados SSE debido al error puede ser interpretada como una medida de cuánta variación de y permanece sin ser explicada por el modelo; es decir, cuánta no puede ser atribuida a una relación lineal. En la figura 9.10(a), SSE = 0 y no existe ninguna variación no explicada, mientras que esta es pequeña con los datos de la figura 9.10(b) y mucho más grande en la figura 9.10(c). La **suma de cuadrados total** da una medida cuantitativa de la cantidad de variación total en los valores y observados

$$\text{SST} = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 / n$$

La suma de cuadrados total es la suma de las desviaciones al cuadrado respecto a la media muestral de los valores y observados. Por consiguiente, se resta el mismo número $\bar{y}$ de cada y_i presente en SST, mientras que SSE implica restar cada valor diferente pronosticado $\hat{y}_i$ de la y_i correspondiente observada. Así como SSE es la suma de desviaciones al cuadrado respecto a la recta de mínimos cuadrados $y = \hat{\beta}_0 + \hat{\beta}_1 x$, SST es la suma de desviaciones al cuadrado respecto a la recta horizontal a la altura $\bar{y}$ (en tal caso las desviaciones verticales son $y_i - \bar{y}$), como se ilustra en la figura 9.11. Además, puesto que la suma de desviaciones al cuadrado respecto a la recta de mínimos cuadrados es más pequeña que la suma de desviaciones al cuadrado respecto a *cualquier* otra recta, SSE < SST a menos que la propia recta horizontal sea la recta de mínimos cuadrados. El cociente SSE/SST es la proporción de variación total que no puede explicarse mediante el modelo de regresión lineal simple y 1 − SSE/SST (un número entre 0 y 1) es la proporción de variación de la y observada explicada por el modelo.

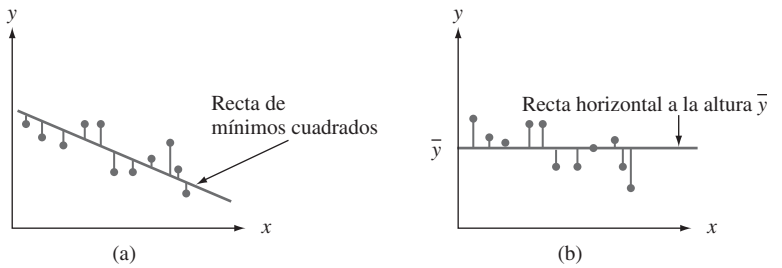


Figura 9.11 Sumas de cuadrados ilustradas: (a) SSE = suma de desviaciones cuadráticas en torno a la recta de mínimos cuadrados; (b) SST = suma de desviaciones al cuadrado en torno a la recta horizontal

DEFINICIÓN

El **coeficiente de determinación,** denotado con r^2, está dado por

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

Se interpreta como la proporción de variación y observada que puede explicarse mediante el modelo de regresión lineal simple (atribuida a una relación lineal aproximada entre y y x).

Mientras más alto es el valor de r^2, más exitoso es el modelo de regresión lineal simple al explicar la variación de y. Cuando se realiza un análisis de regresión mediante un programa estadístico, r^2 o $100r^2$ (el porcentaje de variación explicado por el modelo) son una parte prominente de los resultados. Si r^2 es pequeño, un analista normalmente deseará buscar un modelo alternativo (ya sea un modelo no lineal o uno de regresión múltiple que implique más de una sola variable independiente) que explique con más eficacia la variación de y.

EJEMPLO 9.9 La gráfica de dispersión de los datos para el valor del índice de yodo del número de cetano de la figura 9.8 ciertamente pretende un valor r^2 razonablemente alto. Con

$$\hat{\beta}_0 = 75.212432 \qquad \hat{\beta}_1 = -0.20938742 \qquad \sum y_i = 779.2$$
$$\sum x_i y_i = 71\,347.30 \qquad \sum y_i^2 = 43\,745.22$$

se tiene

$$\text{SST} = 43\,745.22 - (779.2)^2/14 = 377.174$$
$$\text{SSE} = 43745.22 - (75.212432)(779.2) - (2.20938742)(71\,347.30) = 78.920$$

El coeficiente de determinación es entonces

$$r^2 = 1 - \text{SSE/SST} = 1 - (78.920)/(377.174) = 0.791$$

Esto es, 79.1% de la variación observada en el número de cetano es atribuible a (puede explicarse mediante) la regresión lineal simple que relaciona el número de cetano y el valor del índice de yodo (los valores de r^2 son incluso más altos que esto en muchos contextos científicos, pero los científicos sociales generalmente estarían extasiados con ¡un valor grande cerca de este!).

La figura 9.12 muestra resultados parciales generados por Minitab con los datos del número de cetano en el índice de yodo. El programa también proporciona los valores y residuos pronosticados y demás información, si se le solicita. Los formatos utilizados por otros programas difieren un poco de los de Minitab, pero el contenido de la información es muy similar. La suma de cuadrados de regresión se estudiará más adelante. Las demás cantidades en la figura 9.12 que aún no han sido estudiadas serán abordadas en la sección 9.3 [excepto R-Sq(adj), que se contempla en el capítulo 13 cuando se estudian los modelos de regresión múltiple].

```
The regression equation is
número de cetano = 75.2 − 0.209 iod val

                          β̂0  β̂1
Predictor          Coef        SE Coef          T          P
Constant         75.212        2.984       25.21      0.000
índiceyodo    −0.20939     0.03109      −6.73      0.000
                                   100r²
s = 2.56450  R-sq = 79.1%   R-sq(adj) = 77.3%

Analysis of Variance              SSE

SOURCE         DF         SS          MS          F          P
Regression      1     298.25      298.25      45.35      0.000
Error          12      78.92        6.58
Total          13     377.17                    SST
```

Figura 9.12 Resultados obtenidos con Minitab para la regresión de los ejemplos 9.4 y 9.9 ■

El coeficiente de determinación se escribe en una forma un poco diferente al introducir una tercera suma de cuadrados; la **suma de cuadrados debida a la regresión**, SSR, dada por $\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 = \text{SST} - \text{SSE}$. La suma de cuadrados debido a la regresión se interpreta como la cantidad de variación total que *se* explica mediante el modelo. En tal caso se tiene

$$r^2 = 1 - \text{SSE/SST} = (\text{SST} - \text{SSE})/\text{SST} = \text{SSR/SST}$$

la relación de la variación explicada a la variación total. La tabla ANOVA que aparece en la figura 9.12 muestra que SSR = 298.25 de donde $r^2 = 298.25/377.17 = 0.791$ como antes.

Terminología y alcance del análisis de regresión

El término *análisis de regresión* fue introducido por primera vez por Francis Galton a finales del siglo XIX, en conexión con su trabajo sobre la relación entre la estatura del padre *x*

y la estatura del hijo y. Después de recopilar varios pares (x_i, y_i) Galton utilizó el principio de mínimos cuadrados para obtener la ecuación de la recta de regresión estimada con el objetivo de utilizarla para predecir la estatura del hijo a partir de la estatura del padre. Al utilizar la recta obtenida, Galton encontró que si la estatura del padre estaba por encima del promedio, era de esperarse que la del hijo también estuviera por encima del promedio, *pero no tanto como la del padre.* Asimismo, el hijo de un padre cuya estatura es más baja que la del promedio también tendría una estatura más baja que el promedio, pero no tanto como la del padre. Por tanto, la estatura pronosticada del hijo era "retrocedida" hacia la media; porque *regresión* significa un regreso o vuelta. Galton adoptó la terminología *recta de regresión*. Este fenómeno de retroceder hacia la media ha sido observado en muchas otras situaciones (p. ej., promedios de bateo de un año al otro en el béisbol) y se llama **efecto de regresión**.

El análisis hasta ahora ha supuesto que la variable independiente está bajo el control del investigador, así que sólo la variable dependiente Y es aleatoria. Este no fue, sin embargo, el caso con el experimento de Galton: las estaturas de los padres no fueron preseleccionadas, sino que en su lugar tanto X como Y fueron aleatorias. Se pueden aplicar métodos y conclusiones de análisis de regresión cuando los valores de la variable independiente se fijan de antemano y cuando son aleatorios, pero como las deducciones e interpretaciones son más directas en el primer caso, se continuará trabajando explícitamente con él. Para más comentarios, véase el excelente libro de John Neter y colaboradores, citado en la bibliografía del capítulo.

EJERCICIOS Sección 9.2 (12–29)

12. Consulte los datos del ejercicio 4, en el que y = concentración del amonio (mg/L) y x = transpiración (ml/h). Las cantidades resumidas incluyen $n = 13$, $\Sigma x_i = 303.7$, $\Sigma y_i = 52.8$, $S_{xx} = 1585.230769$, $S_{xy} = -341.959231$ y $S_{yy} = 77.270769$.

- **a.** Obtenga la ecuación de la recta de regresión estimada y úsela para calcular una predicción del punto de concentración de amonio de una observación futura cuando la concentración de amonio es 25 ml/h.
- **b.** ¿Qué sucede si la recta de regresión estimada se utiliza para calcular una estimación del punto de concentración medio verdadero cuando la transpiración es de 45 ml/h? ¿Por qué no tiene sentido calcular la estimación de este punto?
- **c.** Calcule e interprete s.
- **d.** ¿Piensa que el modelo de regresión lineal simple es efectivo para explicar la variación observada en la concentración? Explique.

13. Los siguientes datos adjuntos sobre x = densidad de corriente (mA/cm^2) y y = tasa de deposición (μm/min) aparecieron en el artículo **"Plating of 60/40 Tin/Lead Solder for Head Termination Metallurgy" (*Plating and Surface Finishing*, enero de 1997: 38–40)**. ¿Está de acuerdo con la afirmación del autor del artículo en el sentido de que "se obtuvo una re-lación lineal a partir de la tasa de deposición de estaño-plomo como función de la densidad de corriente"? Explique su razonamiento.

x	20	40	60	80
y	0.24	1.20	1.71	2.22

14. Remítase a los datos de relación de temperatura del tanque-eficiencia dada en el ejercicio 1.

- **a.** Determine la ecuación de la recta de regresión estimada.
- **b.** Calcule una estimación puntual de la relación eficiencia promedio verdadera cuando la temperatura del tanque es de 182.
- **c.** Calcule los valores de los residuos con la recta de mínimos cuadrados de las cuatro observaciones con las cuales la temperatura es de 182. ¿Por qué no todas tienen el mismo signo?
- **d.** ¿Qué proporción de la variación observada en la relación de eficiencia puede ser atribuida a la relación de regresión lineal simple entre las dos variables?

15. Se determinaron los valores del módulo de elasticidad (MOE, la relación de esfuerzo, es decir, fuerza por unidad de área a deformación por unidad de longitud, en GPa) y la resistencia a la flexión (una medida de la capacidad para resistir la falla en la flexión en MPa) con una muestra de vigas de concreto de un cierto tipo, y se obtuvieron los siguientes datos (tomados de una gráfica que aparece en el artículo **"Effects of Aggregates**

and Microfillers on the Flexural Properties of Concrete", *Magazine of Concrete Research*, 1997: 81–98):

MOE	29.8	33.2	33.7	35.3	35.5	36.1	36.2
Resistencia	5.9	7.2	7.3	6.3	8.1	6.8	7.0
MOE	36.3	37.5	37.7	38.7	38.8	39.6	41.0
Resistencia	7.6	6.8	6.5	7.0	6.3	7.9	9.0
MOE	42.8	42.8	43.5	45.6	46.0	46.9	48.0
Resistencia	8.2	8.7	7.8	9.7	7.4	7.7	9.7
MOE	49.3	51.7	62.6	69.8	79.5	80.0	
Resistencia	7.8	7.7	11.6	11.3	11.8	10.7	

a. Construya una gráfica de tallo y hojas de los valores de MOE y comente sobre cualquier característica interesante.

b. ¿Será que el valor de resistencia es completa y únicamente determinado por el valor del MOE? Explique.

c. Use los siguientes resultados generados por Minitab para obtener la ecuación de la recta de mínimos cuadrados para predecir la resistencia a partir del módulo de elasticidad y luego para predecir la resistencia de una viga cuyo módulo de elasticidad es de 40. ¿Se sentiría cómodo si utilizara la recta de mínimos cuadrados para predecir la resistencia cuando el módulo de elasticidad es de 100? Explique

```
Predictor      Coef     Stdev  t-ratio      P
Constant     3.2925    0.6008     5.48  0.000
mod elas    0.10748   0.01280     8.40  0.000

s = 0.8657  R-sq = 73.8%  R-sq(adj) = 72.8%

Analysis of Variance

SOURCE       DF       SS      MS      F      P
Regression    1   52.870  52.870  70.55  0.000
Error        25   18.736   0.749
Total        26   71.605
```

d. ¿Cuáles son los valores de SSE, SST y el coeficiente de determinación? ¿Sugieren estos valores que el modelo de regresión lineal simple describe de forma efectiva la relación entre las dos variables? Explique.

16. El artículo **"Characterization of Highway Runoff in Austin, Texas, Area" (*J. of Envir. Engr.*, 1998: 131–137)** incluye una gráfica de dispersión junto con una recta de mínimos cuadrados, de x = volumen de precipitación pluvial (m^3) y y = volumen de escurrimiento (m^3) en un lugar particular. Los siguientes valores fueron tomados de la gráfica.

x	5	12	14	17	23	30	40	47
y	4	10	13	15	15	25	27	46
x	55	67	72	81	96	112	127	
y	38	46	53	70	82	99	100	

a. ¿Será que una gráfica de dispersión de los datos apoya el uso del modelo de regresión lineal simple?

b. Calcule las estimaciones puntuales de la pendiente y la intercepción de la recta de regresión de población.

c. Calcule una estimación puntual del volumen de escurrimiento promedio verdadero cuando el volumen de precipitación pluvial es de 50.

d. Calcule una estimación puntual de la desviación estándar σ.

e. ¿Qué proporción de la variación observada del volumen de escurrimiento puede atribuirse a la relación de regresión lineal simple entre el escurrimiento y la precipitación pluvial?

17. El agregado fino de concreto, hecho a partir de un agregado secundario clasificado de manera uniforme y una pasta de cemento-agua, es beneficioso en las zonas propensas a las lluvias excesivas debido a sus excelentes propiedades de drenaje. El artículo **"Pavement Thickness Design for No-Fines Concrete Parking Lots" (*J. of Trans. Engr.*, 1995: 476–484)** empleó un análisis de mínimos cuadrados en el estudio de cómo y = porosidad (%) se relaciona con x = peso unitario (por pie cúbico) en muestras de concreto. Considere los siguientes datos representativos:

x	99.0	101.1	102.7	103.0	105.4	107.0	108.7	110.8
y	28.8	27.9	27.0	25.2	22.8	21.5	20.9	19.6
x	112.1	112.4	113.6	113.8	115.1	115.4	120.0	
y	17.1	18.9	16.0	16.7	13.0	13.6	10.8	

Un resumen de las cantidades pertinentes es $\Sigma x_i = 1640.1$, $\Sigma y_i = 299.8$, $\Sigma x_i^2 = 179\,849.73$, $\Sigma x_i y_i = 32\,308.59$, $\Sigma y_i^2 = 6430.06$.

a. Obtenga la ecuación de la recta de regresión estimada. A continuación, cree un gráfico de dispersión de los datos y el gráfico de la recta estimada. ¿Será que el modelo de la relación puede explicar una gran parte de la variación observada en y?

b. Interprete la pendiente de la recta de mínimos cuadrados.

c. ¿Qué sucede si la estimación lineal se utiliza para predecir la porosidad cuando el peso unitario es de 135? ¿Por qué esto no es una buena idea?

d. Calcule los residuos correspondientes a las dos primeras observaciones.

e. Calcule e interprete una estimación puntual de σ.

f. ¿Qué proporción de la variación observada en la porosidad se puede atribuir a la relación lineal aproximada entre el peso unitario y la porosidad?

18. Durante la última década el polvo de caucho se ha utilizado en cemento asfáltico para mejorar el rendimiento. El artículo **"Experimental Study of Recycled Rubber-Filled High-Strength Concrete" (*Magazine of Concrete Res.*, 2009: 549–556)** incluye una regresión de y = *esfuerzo axial* (MPa) en x = *esfuerzo cúbico* (Mpa) basada en los siguientes datos de muestra:

x	112.3	97.0	92.7	86.0	102.0	99.2	95.8	103.5	89.0	86.7
y	75.0	71.0	57.7	48.7	74.3	73.3	68.0	59.3	57.8	48.5

a. Obtenga la ecuación de la recta de mínimos cuadrados e interprete su pendiente.

b. Calcule e interprete el coeficiente de determinación.

c. Calcule e interprete una estimación de la desviación estándar σ del error en el modelo de regresión lineal simple.

19. Los siguientes datos son representativos de los reportados en el artículo **"An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data" (*J. of Engr. for Power*, julio de 1973: 165–170)**, con x = tasa de liberación debido a área de quemador (MBtu/h-pie^2) y y = tasa de emisión de NO_x (ppm):

x	100	125	125	150	150	200	200
y	150	140	180	210	190	320	280
x	250	250	300	300	350	400	400
y	400	430	440	390	600	610	670

a. Suponiendo que el modelo de regresión lineal simple es válido obtenga la estimación de mínimos cuadrados de la recta de regresión verdadera.

b. ¿Cuál es la estimación de la tasa de emisión de NO_x esperada cuando la tasa de liberación debido al área del quemador es igual a 225?

c. Estime la cantidad en la cual espera que cambie la tasa de emisiones de NO_x cuando la tasa de liberación debida al área del quemador disminuye en 50.

d. ¿Utilizaría la recta de regresión estimada para predecir la tasa de emisión con una tasa de liberación de 500? ¿Por qué sí, o por qué no?

20. El comportamiento de las barras de unión de refuerzo es determinante e importante en la fuerza y la estabilidad. El artículo **"Experimental Study on the Bond Behavior of Reinforcing Bars Embedded in Concrete Subjected to Lateral Pressure" (*J. of Materials in Civil Engr.*, 2012: 125–133)** reporta los resultados de un experimento en el que se aplicaron diferentes niveles de presión lateral a 21 especímenes cúbicos de concreto, cada uno con una barra redonda de acero liso de 16 mm incrustada, y se determinó la capacidad de unión correspondiente. Debido a las diferentes fuerzas sobre un cubo de hormigón (f_{cu}, en MPa), la presión lateral aplicada fue equivalente a una proporción fija de la f_{cu} de la muestra (0, $0.1f_{cu}$,..., $0.6f_{cu}$). También, puesto que la fuerza de unión puede ser influenciada por la f_{cu} del espécimen, la capacidad de enlace se expresa como el cociente de la fuerza de unión (MPa) para $\sqrt{f_{cu}}$.

Presión	0	0	0	0.1	0.1	0.1	0.2
Cociente	0.123	0.100	0.101	0.172	0.133	0.107	0.217
Presión	0.2	0.2	0.3	0.3	0.3	0.4	0.4
Cociente	0.172	0.151	0.263	0.227	0.252	0.310	0.365
Presión	0.4	0.5	0.5	0.5	0.6	0.6	0.6
Cociente	0.239	0.365	0.319	0.312	0.394	0.386	0.320

a. ¿Será que una gráfica de dispersión de datos apoya el uso del modelo de regresión lineal simple?

b. Utilice los siguientes resultados de Minitab para dar estimaciones puntuales de la pendiente y de la intercepción de la recta de regresión de la población.

c. Calcule una estimación puntual de la capacidad de unión promedio real cuando la presión lateral es $0.45f_{cu}$.

d. ¿Qué es una estimación puntual de la desviación estándar de error σ, y ¿cómo la interpreta?

e. ¿Cuál es el valor de la variación total y qué proporción de este se explica mediante el modelo de la relación?

```
The regression equation is
Cociente = 0.101 + 0.461 Presión

Predictor       Coef   SE Coef       T       P
Constant     0.10121   0.01308    7.74   0.000
Pressure     0.46071   0.03627   12.70   0.000

S=0.0332397  R-Sq=89.5%  R-Sq(adj)=88.9%

Analysis of Variance
Source          DF      SS      MS      F     P
Regression       1 0.17830 0.17830 161.37 0.000
Residual Error  19 0.02099 0.00110
Total           20 0.19929
```

21. El ángulo de recuperación de pliegues y la resistencia a la tensión son las dos características más importantes para evaluar el desempeño de la tela de algodón entrelazada. Un incremento en el ángulo de entrelazado, determinado por la absorbencia de una banda de éster carboxilo, mejora la resistencia a los pliegues de la tela (a expensas de reducir la resistencia mecánica). Los siguientes datos sobre x = absorbencia y y = resistencia al ángulo de los pliegues fueron tomados de una gráfica que se muestra en el artículo **"Predicting the Performance of Durable Press Finished Cotton Fabric with Infrared Spectroscopy" (*Textile Res. J.*, 1999: 145–151)**.

x	0.115	0.126	0.183	0.246	0.282	0.344	0.355	0.452	0.491	0.554	0.651
y	334	342	355	363	365	372	381	392	400	412	420

He aquí los resultados obtenidos con Minitab:

```
Predictor      Coef    SE Coef        T       P
Constant    321.878      2.483   129.64   0.000
absorb      156.711      6.464    24.24   0.000

S = 3.60498  R-Sq = 98.5%  R-Sq(adj) = 98.3%

Source          DF      SS      MS      F      P
Regression       1  7639.0  7639.0 587.81  0.000
Residual Error   9   117.0    13.0
Total           10  7756.0
```

a. ¿Será apropiado el modelo de regresión lineal simple? Explique.

b. ¿Qué ángulo de resistencia a las arrugas pronosticaría para un espécimen de tela cuya absorbencia es de 0.300?

c. ¿Cuál sería la estimación del ángulo de resistencia a los pliegues esperado cuando la absorbencia es de 0.300?

22. El cemento de fosfato de calcio cada vez está ganando más atención para su uso en aplicaciones de reparación ósea. El artículo **"Short-Fibre Reinforcement of Calcium Phosphate Bone Cement" (*J. of Engr. in Med.*, 2007: 203–211)** informa sobre un estudio en el que se utilizan fibras de polipropileno en un intento de mejorar el comportamiento de la fractura. Los siguientes datos de x = peso de la fibra (%) y y = resistencia a la compresión (MPa) fueron proporcionado por los autores del artículo.

x	0.00	0.00	0.00	0.00	0.00	1.25	1.25	1.25	1.25
y	9.94	11.67	11.00	13.44	9.20	9.92	9.79	10.99	11.32

x	2.50	2.50	2.50	2.50	2.50	5.00	5.00	5.00	5.00
y	12.29	8.69	9.91	10.45	10.25	7.89	7.61	8.07	9.04

x	7.50	7.50	7.50	7.50	10.00	10.00	10.00	10.00
y	6.63	6.43	7.03	7.63	7.35	6.94	7.02	7.67

a. Ajuste el modelo de regresión lineal simple a estos datos. Después, determine la proporción de la variación observada en la resistencia que se puede atribuir a la relación entre el modelo de resistencia y el peso de la fibra. Por último, obtenga una estimación puntual de la desviación estándar de ϵ, la desviación aleatoria de la ecuación de modelo.

b. Los valores de resistencia promedio de los seis niveles diferentes de peso de la fibra son 11.05, 10.51, 10.32, 8.15, 6.93 y 7.24, respectivamente. El citado documento incluye una figura en la que se compara la resistencia promedio contra el peso promedio de la fibra. Obtenga la ecuación de esta recta de regresión y calcule el coeficiente de determinación correspondiente. Explique la diferencia entre el valor de r^2 para esta regresión y el valor de r^2 obtenido en a).

23. **a.** Obtenga la SSE con los datos del ejercicio 19 a partir de la fórmula definitoria [SSE $= \Sigma(y_i - \overline{y}_i)^2$] y compare con el valor determinado mediante la fórmula de cálculo.

b. Calcule el valor de la suma de cuadrados total. ¿Será que el modelo de regresión lineal simple explica con eficacia la variación en la tasa de emisión? Justifique su aseveración.

24. Las especies de diatomeas invasivas *Didymosphenia geminata* tienen potencial para causar sustanciales daños ecológicos y económicos en los ríos. En el artículo **"Substrate Characteristics Affect Colonization by the Bloom-Forming Diatom Didymosphenia geminata (*Aquatic Ecology*, 2010: 33–40)** se describe una investigación del comportamiento de la colonia. Un aspecto de particular interés fue si y = densidad de la colonia estaba relacionada con x = área superficial de la roca. El artículo contiene un diagrama de dispersión y el resumen de un análisis de regresión. A continuación se presentan los datos representativos:

x	50	71	55	50	33	58	79	26
y	152	1929	48	22	2	5	35	7

x	69	44	37	70	20	45	49
y	269	38	171	13	43	185	25

a. Ajuste el modelo de regresión lineal simple a estos datos, prediga la densidad de la colonia cuando el área superficial = 70 y cuando el área superficial = 71 y calcule los residuos correspondientes. ¿Cómo se comparan entre sí?

b. Calcule e interprete el coeficiente de determinación.

c. La segunda observación tiene un valor y muy extremo y (en el conjunto de datos completo que consta de 72 observaciones, había dos de estos). Esta observación puede haber tenido un impacto sustancial en el ajuste del modelo y en las conclusiones posteriores. Elimine y vuelva a calcular la ecuación de la recta de regresión estimada. ¿Será que difiere sustancialmente de la ecuación antes de la eliminación? ¿Cuál es el impacto en r^2 y en s?

25. Compruebe que b_1 y b_0 de las expresiones (9.2) y (9.3) satisfacen las ecuaciones normales.

26. Demuestre que el "promedio de puntos" $(\overline{x}, \overline{y})$ queda en la recta de regresión estimada.

27. Suponga que un investigador cuenta con los datos sobre la cantidad de espacio del anaquel x dedicado a la exhibición de un producto particular y sobre los ingresos por ventas y del mismo producto. Quizá el investigador desee adaptar un modelo para el cual la recta de regresión verdadera pase a través de (0, 0). El modelo apropiado es $Y = \beta_1 x + \epsilon$. Suponga que $(x_1, y_1),\ldots, (x_n, y_n)$ son pares observados generados con este modelo y deduzca el estimador de mínimos cuadrados de β_1. [*Sugerencia*: Escriba la suma de desviaciones al cuadrado como una función de b_1, un valor de prueba y use el cálculo para determinar el valor minimizante de b_1.]

28. **a.** Considere los datos del ejercicio 20. Suponga que en lugar de la recta de mínimos cuadrados que pasa por los puntos $(x_1, y_1),\ldots, (x_n, y_n)$, se desea que la recta de mínimos cuadrados pase por $(x_1 - \overline{x}, y_1),\ldots, (x_n - \overline{x}, y_n)$. Construya una gráfica de dispersión con los puntos (x_i, y_i) y luego con los puntos $(x_i - \overline{x}, y_i)$. Use las gráficas para explicar intuitivamente cómo están relacionadas entre sí las dos rectas de mínimos cuadrados.

b. Suponga que en lugar del modelo $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i (i = 1,\ldots, n)$, se desea ajustar un modelo de la forma $Y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \epsilon_i$ $(i = 1,\ldots, n)$. ¿Cuáles son los estimadores de mínimos cuadrados de β_0^* y β_1^* y cómo están relacionados con $\hat{\beta}_0$ y $\hat{\beta}_1$?

29. Considere los siguientes tres conjuntos de datos en los cuales las variables de interés son x = distancia de la casa al trabajo y y = tiempo de recorrido de la distancia de la casa al trabajo. Basado en una gráfica de dispersión y los valores de s y r^2, ¿en qué situación la regresión lineal simple sería más (al menos) efectiva y por qué?

Conjunto de datos	**1**		**2**		**3**	
	x	***y***	***x***	***y***	***x***	***y***
	15	42	5	16	5	8
	16	35	10	32	10	16
	17	45	15	44	15	22
	18	42	20	45	20	23
	19	49	25	63	25	31
	20	46	50	115	50	60
S_{xx}	17.50		1270.8333		1270.8333	
S_{xy}	29.50		2722.5		1431.6667	
$\hat{\beta}_1$	1.685714		2.142295		1.126557	
$\hat{\beta}_0$	13.666672		7.868852		3.196729	
SST	114.83		5897.5		1627.33	
SSE	65.10		65.10		14.48	

9.3 Inferencias sobre el parámetro de la pendiente β_1

Virtualmente en todo el trabajo inferencial realizado hasta ahora, el concepto de variabilidad de muestreo ha sido persistente. En particular, las propiedades de las distribuciones de muestreo de varios estadísticos han sido la base para desarrollar fórmulas de intervalo de confianza y métodos de prueba de hipótesis. La idea clave en este caso es que el valor de cualquier cantidad calculada a partir de datos muestrales, el valor de cualquier estadístico, variará de una muestra a otra.

EJEMPLO 9.10 Reconsidere los datos sobre x = tasa de liberación debido al área del quemador y y = tasa de emisiones de NO_x del ejercicio 9.19 en la sección previa. Existen 14 observaciones realizadas con los valores x 100, 125, 125, 150, 150, 200, 200, 250, 250, 300, 300, 350, 400 y 400, respectivamente. Suponga que la pendiente y la intercepción de la recta de regresión verdadera son $\beta_1 = 1.70$ y $\beta_0 = -50$, con $\sigma = 35$ (consistente con los valores $\hat{\beta}_1 = 1.7114$, $\hat{\beta}_0 = -45.55$, $s = 36.75$). Se procedió a generar una muestra de desviaciones aleatorias $\tilde{\boldsymbol{\epsilon}}_1, \ldots, \tilde{\boldsymbol{\epsilon}}_{14}$ respecto a una distribución normal con media 0 y desviación estándar 35, y luego se sumó $\tilde{\boldsymbol{\epsilon}}_i$ a $\beta_0 + \beta_1 x_i$ para obtener 14 valores y correspondientes. Se realizaron entonces los cálculos de regresión para obtener la pendiente, la intercepción y la desviación estándar estimados. Este proceso se repitió un total de 20 veces y los valores resultantes se proporcionan en la tabla 9.1.

Tabla 9.1 Resultados de simulación del ejemplo 9.10

$\hat{\beta}_1$	$\hat{\beta}_0$	s	$\hat{\beta}_1$	$\hat{\beta}_0$	s
1. 1.7559	−60.62	43.23	11. 1.7843	−67.36	41.80
2. 1.6400	−49.40	30.69	12. 1.5822	−28.64	32.46
3. 1.4699	−4.80	36.26	13. 1.8194	−83.99	40.80
4. 1.6944	−41.95	22.89	14. 1.6469	−32.03	28.11
5. 1.4497	5.80	36.84	15. 1.7712	−52.66	33.04
6. 1.7309	−70.01	39.56	16. 1.7004	−58.06	43.44
7. 1.8890	−95.01	42.37	17. 1.6103	−27.89	25.60
8. 1.6471	−40.30	43.71	18. 1.6396	−24.89	40.78
9. 1.7216	−42.68	23.68	19. 1.7857	−77.31	32.38
10. 1.7058	−63.31	31.58	20. 1.6342	−17.00	30.93

Claramente existe variación en los valores de la pendiente y la intercepción estimadas, así como también en la desviación estándar estimada. La ecuación de la recta de mínimos cuadrados varía, por tanto, de una muestra a la siguiente. La figura 9.13 en la página 511 muestra una gráfica de puntos de las pendientes estimadas, así como también gráficas de la recta de regresión verdadera y las 20 rectas de regresión muestrales. ■

La pendiente β_1 de la recta de regresión de población es el cambio promedio verdadero en la variable dependiente y asociada con un incremento de 1 unidad en la variable independiente x. La pendiente de la recta de mínimos cuadrados, $\hat{\beta}_1$ da una estimación puntual de β_1. Del mismo modo que un intervalo de confianza para μ y los procedimientos para probar hipótesis respecto a μ se basaron en propiedades de la distribución de muestreo de $\overline{X}$, las inferencias adicionales sobre β_1 están basadas en considerar a $\hat{\beta}_1$ como un estadístico e investigar su distribución de muestreo.

Se supone que los valores de las x_i se eligen antes de realizar el experimento, así que sólo las Y_i son aleatorias. Los estimadores (estadísticos y, por tanto, las variables aleatorias) de β_0 y β_1 se obtienen al reemplazar y_i por Y_i en (9.2) y (9.3):

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \overline{Y})}{\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \frac{\sum Y_i - \hat{\beta}_1 \sum x_i}{n}$$

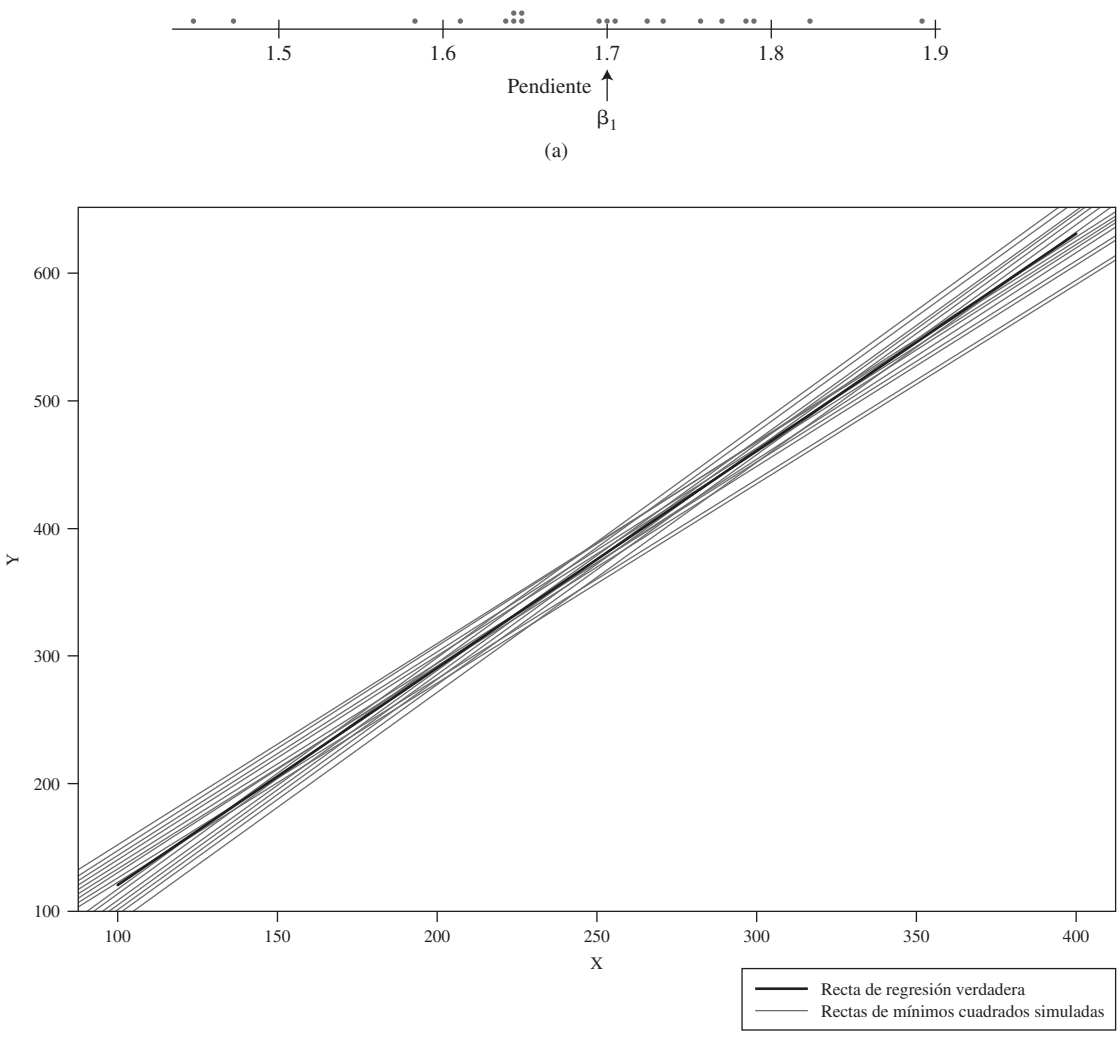


Figura 9.13 Resultados de simulación del ejemplo 9.10; (a) Gráfica de puntos de pendientes estimadas; (b) gráficas de la recta de regresión verdadera y de las 20 rectas de mínimos cuadrados (obtenidas con S-Plus)

Asimismo, el estimador de σ^2 se obtiene al reemplazar cada y_i en la fórmula para s^2 por la variable aleatoria Y_i:

$$\hat{\sigma}^2 = S^2 = \frac{\sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum x_i Y_i}{n - 2}$$

El denominador de $\hat{\beta}_1$, $S_{xx} = \Sigma(x_i - \bar{x})^2$, depende sólo de las x_i y no de las Y_i, así que es una constante. Entonces, puesto que $\Sigma(x_i - \bar{x})\bar{Y} = \bar{Y}\ \Sigma(x_i - \bar{x}) = Y \cdot 0 = 0$, el estimador de la pendiente se escribe como

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) Y_i}{S_{xx}} = \sum c_i Y_i \qquad \text{donde } c_i = (x_i - \bar{x})/S_{xx}$$

Es decir, $\hat{\beta}_1$ es una función lineal de las variables aleatorias independientes $Y_1, Y_2, \ldots, Y_n$, cada una de las cuales está normalmente distribuida. Invocar las propiedades de una función lineal de variables aleatorias conduce a los siguientes resultados.

PROPOSICIÓN

1. La media de $\hat{\beta}_1$ es $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, así que $\hat{\beta}_1$ es un estimador insesgado de β_1 (la distribución de $\hat{\beta}_1$ siempre está centralizada en el valor de β_1).
2. La varianza y la desviación estándar de $\hat{\beta}_1$ son

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} \tag{9.4}$$

   donde $S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2/n$. Reemplazando σ por su estimación s da una estimación para $\sigma_{\hat{\beta}_1}$ (la desviación estándar estimada, es decir, el error estándar estimado de $\hat{\beta}_1$):

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$$

   (Esta estimación también puede ser denotada por $\hat{\sigma}_{\hat{\beta}_1}$.)
3. El estimador $\hat{\beta}_1$ tiene una distribución normal (debido a que es una función lineal de variables aleatorias estandarizadas independientes).

De acuerdo con (9.4), la varianza de $\hat{\beta}_1$ es igual a la varianza σ^2 del término de error aleatorio o, de forma equivalente, de cualquier Y_i, dividida entre $\Sigma(x_i - \bar{x})^2$. Puesto que mide la dispersión de las x_i en torno a $\bar{x}$, se concluye que si se realizan observaciones a valores x_i que están bastante dispersos se obtiene un estimador más preciso del parámetro de la pendiente (la varianza más pequeña de $\hat{\beta}_1$), mientras que los valores de x_i muy cercanos entre sí implican un estimador altamente variable. Desde luego, si las x_i están demasiado dispersas, un modelo lineal quizá no sea apropiado a lo largo del rango de observación.

Muchos procedimientos inferenciales previamente analizados se basaron en estandarizar un estimador restando primero su media y luego dividiéndolo entre su desviación estándar estimada. En particular, los procedimientos de prueba y un intervalo de confianza para la media μ de una población normal utilizaron el hecho de que la variable estandarizada $(\overline{X} - \mu)/(S/\sqrt{n})$, es decir $(\overline{X} - \mu)/S_{\overline{X}}$, tenía una distribución t con $n - 1$ grados de libertad. Un resultado similar en este caso abre la puerta a más inferencias sobre β_1.

TEOREMA

La suposición del modelo de regresión lineal simple implica que la variable estandarizada

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

tiene una distribución t con $n - 2$ grados de libertad.

Un intervalo de confianza para β_1

Tal como en la deducción de intervalos de confianza previos, se inicia con un enunciado de probabilidad

$$P\left(-t_{\alpha/2,\,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

La manipulación de las desigualdades entre paréntesis para aislar β_1 y la sustitución de las estimaciones en lugar de los estimadores da la fórmula del intervalo de confianza.

Un **intervalo de confianza** de $100(1-\alpha)\%$ **para la pendiente** β_1 de la recta de regresión verdadera es

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

Este intervalo tiene la misma forma general de muchos de los intervalos previos. Está centrado en la estimación puntual del parámetro y la cantidad que se extiende a cada lado depende del nivel de confianza deseado (a través del valor crítico t) y de la cantidad de variabilidad del estimador $\hat{\beta}_1$ (a través de $s_{\hat{\beta}_1}$, el cual tenderá a ser más pequeño cuando existe poca variabilidad en la distribución de $\hat{\beta}_1$ y grande en caso contrario).

EJEMPLO 9.11 Cuando se daña la estructura de la madera, puede ser más económico reparar la zona afectada en vez de sustituir toda la estructura. El artículo **"Simplified Model for Strength Assessment of Timber Beams Joined by Bonded Plates" (*J. of Materials in Civil Engr.*, 2013: 980–990)** investigó una estrategia particular para su reparación. Los siguientes datos fueron usados por los autores del artículo como base para ajustar el modelo de regresión lineal simple. La variable dependiente es y = carga de ruptura (N) y la variable independiente es la longitud de anclaje (la longitud adicional del material utilizado para adherir en la unión, en mm).

x	50	50	80	80	110	110	140	140	170	170
y	17 052	14 063	26 264	19 600	21 952	26 362	26 362	26 754	31 654	32 928

Observe que la relación entre la longitud de anclaje y la carga de ruptura claramente no es determinista, ya que hay observaciones con valores x idénticos pero diferentes valores y. La figura 9.14 muestra un diagrama de dispersión de los datos (también se presenta en el artículo citado) donde parece haber una relación lineal positiva bastante sustancial entre las dos variables.

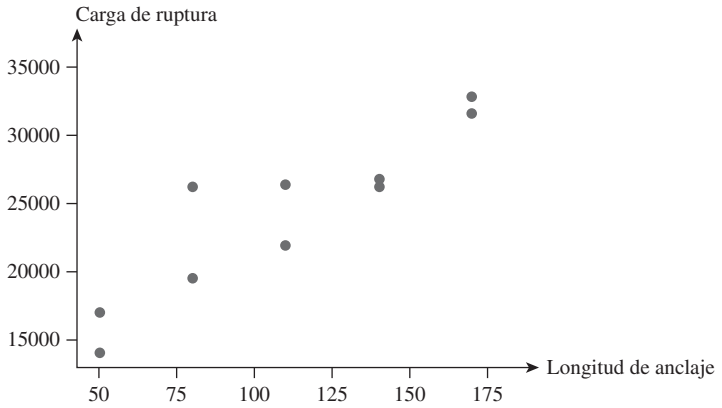


Figura 9.14 Gráfica de dispersión de los datos del ejemplo 9.11

Las cantidades resumidas incluyen $S_{xx} = 18\,000$, $S_{xy} = 2\,225\,579.40$, $S_{yy} = \text{SST} = 331\,839\,568.9$, $\hat{\beta}_1 = 123.6433$, $\hat{\beta}_0 = 10\,698.33$, $\text{SSE} = 56\,661\,439.1$ y $r^2 = 0.829$. Aproximadamente 83% de la variación observada en la carga de ruptura puede atribuirse a la relación del modelo de regresión lineal simple entre carga de ruptura y longitud de anclaje. El grado de libertad de error es $10 - 2 = 8$, de lo cual $s^2 = 56\,661\,439.1/8 = 7\,082\,679.89$ y $s = 2661.33$. El error estándar estimado de $\hat{\beta}_1$ es

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{2661.33}{\sqrt{18\,000}} = 19.836$$

Un nivel de confianza de 95% requiere que $t_{0.025,8} = 2.306$. El intervalo de confianza es

$$123.64 \pm (2.306)(19.836) = 123.64 \pm 45.74 = (77.90, 169.38)$$

Con un alto grado de confianza, se estima que un aumento en la resistencia a la ruptura media verdadera de entre 77.90 N y 169.38 N está asociado con un aumento de 1 mm de longitud de anclaje (al menos para las longitudes de anclaje entre 50 y 170 mm). Este intervalo no es demasiado estrecho, es una consecuencia del pequeño tamaño muestral y la variabilidad sustancial sobre la recta de regresión estimada. Observe que el intervalo incluye sólo valores positivos, por lo que podemos estar absolutamente seguros de que la fuerza aumenta conforme aumenta la longitud de anclaje.

La figura 9.15 presenta los resultados de regresión del paquete SAS. El valor de $s_{\hat{\beta}_1}$ se encuentra debajo de las estimaciones del parámetro como el segundo número en la columna del error estándar. También es un error estándar estimado para $\hat{\beta}_0$, del cual se puede calcular un intervalo de confianza para la intercepción de la recta de regresión de población. Las dos últimas columnas de la tabla de parámetros estimados dan información sobre la comprobación de ciertas hipótesis, nuestro siguiente tema de análisis.

Análisis de varianza					
Fuente	Grados de libertad	Suma de los cuadrados	Media de los cuadrados	Valor P	Pr > F
Modelo	1	275178130	275178130	38.85	0.0003
Error	8	56661439	7082680		
Total corregido	9	331839569			

Raíz MSE	2661.33047	R cuadrada	0.8293
Media dependiente	24299	R Adj. cuadrada	0.8079
Coef. Var	10.95238		

Parámetros estimados					
Variable	Grados de libertad	Parámetro estimado	Error estándar	Valor *t*	Pr >|t|
Intercepción	1	10698	2338.67544	4.57	0.0018
Longitud de anclaje	1	123.64333	19.83639	6.23	0.0003

Figura 9.15 Resultados obtenidos con SAS con los datos del ejemplo 9.11

Procedimientos de prueba de hipótesis

Tal como antes, la hipótesis nula en una prueba respecto a β_1 será un enunciado de igualdad. El valor nulo (valor de β_1 supuesto verdadero por la hipótesis nula) será denotado por β_{10} (léase "beta uno cero", *no* "beta diez"). El estadístico de prueba se obtiene reemplazando β_1 en la variable estandarizada T por el valor nulo β_{10}; es decir, estandarizando el estimador de β_1 conforme a la suposición de que H_0 es verdadera. El estadístico de prueba tiene, por tanto, una distribución t con $n - 2$ grados de libertad cuando H_0 es verdadera, lo que permite determinar un valor P tal como se describe en los capítulos 8 y 9 para las pruebas t.

El par más comúnmente encontrado de hipótesis sobre β_1 es H_0: $\beta_1 = 0$ contra H_a: $\beta_1 \neq 0$. Cuando esta hipótesis nula es verdadera, $\mu_{Y \cdot x} = \beta_0$ independiente de x. Entonces conocer x no da ninguna información sobre el valor de la variable dependiente. Una prueba de estas dos hipótesis se refiere a menudo como la *prueba de la utilidad del modelo* de

regresión lineal simple. A menos que n sea muy pequeño, H_0 se rechazará y la utilidad del modelo se confirma precisamente cuando r^2 es razonablemente grande. El modelo de regresión lineal simple no debe utilizarse para otras inferencias (estimaciones de la media o predicciones de valores futuros) a menos que los resultados en las pruebas del modelo de utilidad rechace H_0 para una α convenientemente pequeña.

Hipótesis nula: H_0: $\beta_1 = \beta_{10}$

Valor del estadístico de prueba: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Hipótesis alternativa	Determinación del valor *P*
H_a: $\beta_1 > \beta_{10}$	Área bajo la curva t_{n-2} a la derecha de t
H_a: $\beta_1 < \beta_{10}$	Área bajo la curva t_{n-2} a la izquierda de t
H_a: $\beta_1 \neq \beta_{10}$	$2 \cdot$ (Área bajo la curva t_{n-2} a la derecha de $\lvert t \rvert$)

La **prueba de utilidad del modelo** es la prueba de H_0: $\beta_1 = 0$ contra H_a: $\beta_1 \neq 0$, en cuyo caso el valor del estadístico de prueba es el **cociente *t***, $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

EJEMPLO 9.12 Los ciclomotores o mopeds son muy populares en Europa debido a su costo y su fácil manejo. Sin embargo, pueden ser peligrosos si se modifican las características de rendimiento. Una de las características comúnmente manipulada es la velocidad máxima. El artículo **"Procedure to Verify the Maximum Speed of Automatic Transmission Mopeds in Periodic Motor Vehicle Inspections" (*J. of Automotive Engr.*, 2008: 1615–1623)** incluyó un análisis de regresión lineal simple de las variables x = velocidad de la pista de prueba (km/h) y y = velocidad de prueba de rodamiento. He aquí los datos leídos de una gráfica en el artículo:

x	42.2	42.6	43.3	43.5	43.7	44.1	44.9	45.3	45.7
y	44	44	44	45	45	46	46	46	47
x	45.7	45.9	46.0	46.2	46.2	46.8	46.8	47.1	47.2
y	48	48	48	47	48	48	49	49	49

Una gráfica de dispersión de los datos muestra un patrón lineal sustancial. El resultado de Minitab en la figura 9.16 da el coeficiente de determinación como $r^2 = 0.923$, que sin duda presagia una útil relación lineal. Vamos a llevar a cabo la prueba de utilidad del modelo en un nivel de significancia $\alpha = 0.01$.

```
The regression equation is
vel rodamiento = −2.22 + 1.08 vel pista

Predictor          Coef       SE Coef          T          P
Constant         −2.224         3.528      −0.63      0.537
vel pista       1.08342       0.07806      13.88      0.000
S = 0.506890    R-Sq = 92.3%    R-Sq(adj) = 91.9%

Analysis of Variance

Source               DF        SS        MS         F        P
Regression            1    49.500    49.500    192.65    0.000
Residual Error       16     4.111     0.257
Total                17    53.611
```

$s_{\hat{\beta}_1}$ $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$ ← Valor *P* para la prueba de la utilidad del modelo

Figura 9.16 Resultados obtenidos con Minitab del ejemplo 9.12

El parámetro de interés es β_1, el cambio esperado en la velocidad de rodadura asociado con un incremento de 1 km/h en la prueba de velocidad. La hipótesis nula H_0: $\beta_1 = 0$ será rechazada en favor de la alternativa H_a: $\beta_1 \neq 0$ si el cociente t, $t = \hat{\beta}_1/s_{\hat{\beta}_1}$ se encuentra demasiado lejos, es decir en la cola de la curva t_{n-2} (lo que resulta en un pequeño valor P). De acuerdo con la figura 9.16, $\hat{\beta}_1 = 1.08342$, $s_{\hat{\beta}_1} = 0.07806$, y

$$t = \frac{1.08342}{0.07806} = 13.88 \quad \text{(también en los resultados)}$$

El valor P es dos veces el área capturada bajo la curva t de 16 grados de libertad a la derecha de 13.88. Minitab da un valor $P = 0.000$. Por ello se puede rechazar la hipótesis nula de no relación lineal útil a cualquier nivel de significancia razonable. Esta confirmación de la utilidad del modelo de regresión simple permite calcular varias estimaciones y predicciones tal como se describe en la sección 9.4. ■

Regresión y ANOVA

La descomposición de la suma total de cuadrados $\Sigma(y_i - \bar{y})^2$ en una parte SSE, la cual mide la variación no explicada y una parte SSR, la cual mide la variación explicada por la relación lineal, hace recordar fuertemente el ANOVA unidireccional. De hecho, la hipótesis nula H_0: $\beta_1 = 0$ puede ser probada contra H_a: $\beta_1 \neq 0$ con una tabla ANOVA (tabla 9.2) y determinando el valor P para la prueba F.

Tabla 9.2 Tabla ANOVA para regresión lineal simple

Origen de la variación	gl	Suma de cuadrados	Media cuadrática	f
Regresión	1	SSR	SSR	$\dfrac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \dfrac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

La prueba F da exactamente el mismo resultado que la prueba t de utilidad de modelo $t^2 = f$ y $t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$. Virtualmente todos los programas de computadora que cuentan con opciones de regresión incluyen dicha tabla ANOVA en los resultados. Por ejemplo, la figura 12.15 muestra los resultados obtenidos con SAS con los datos de mortero del ejemplo 12.11. La tabla ANOVA en la parte superior de los resultados tiene $f = 38.85$ con un valor P de 0.0003 para la prueba de utilidad de modelo. La tabla de estimaciones de parámetro da $t = 6.23$ de nuevo con $P = 0.0003$ y $38.85 \approx (6.23)^2$ (que serían idénticos si se mostrara mayor precisión decimal).

EJERCICIOS **Sección 9.3 (30–43)**

30. Reconsidere la situación descrita en el ejercicio 7, en el cual x = resistencia acelerada del concreto y y = resistencia después de 28 días de curado. Suponga que el modelo de regresión lineal simple es válido con x entre 1000 y 4000 y que $\beta_1 = 1.25$ y $\sigma = 350$. Considere un experimento en el cual $n = 7$ y los valores x a los cuales se realizan las observaciones son $x_1 = 1000$, $x_2 = 1500$, $x_3 = 2000$, $x_4 = 2500$, $x_5 = 3000$, $x_6 = 3500$ y $x_7 = 4000$.

a. Calcule $\sigma_{\hat{\beta}_1}$, la desviación estándar de $\hat{\beta}_1$.

b. ¿Cuál es la probabilidad de que la pendiente estimada con base en las observaciones esté entre 1.00 y 1.50?

c. Suponga que también es posible hacer una sola observación con cada uno de los $n = 11$ valores $x_1 = 2000$, $x_2 = 2100$,..., $x_{11} = 3000$. Si un objetivo importante es estimar β_1 con tanta precisión como sea posible, ¿se preferiría el experimento con $n = 11$ a uno con $n = 7$?

31. Durante las operaciones de perforación de pozos de petróleo, los componentes del ensamble de perforación pueden sufrir rompimiento por esfuerzo a partir de sulfuros. El artículo **"Composition Optimization of High-Strength Steels for Sulfide Cracking Resistance Improvement" (*Corrosion Science*, 2009: 2878–2884)** informa sobre un estudio en el que se analizó la composición de un acero de grado estándar. Los siguientes datos sobre el umbral de esfuerzo y = umbral de tensión (SMYS%) y x = límite elástico (MPa) se obtuvieron de una gráfica en el artículo (que también incluye la ecuación de la recta de mínimos cuadrados).

x	635	644	711	708	836	820	810	870	856	923	878	937	948
y	100	93	88	84	77	75	74	63	57	55	47	43	38

$$\sum x_i = 10\,576,\ \sum y_i = 894,\ \sum x_i^2 = 8\,741\,264,$$
$$\sum y_i^2 = 66\,224,\ \sum x_i y_i = 703\,192$$

a. ¿Qué proporción de la variación observada en el esfuerzo puede ser atribuida a la relación lineal aproximada entre las dos variables?

b. Calcule la desviación estándar estimada $s_{\hat{\beta}_1}$.

c. Calcule un intervalo de confianza usando el nivel de confianza de 95% de la variación esperada del esfuerzo asociado con un aumento de 1 MPa en la resistencia. ¿Será que este promedio real de cambio ha sido estimado con precisión?

32. El ejercicio 16 de la sección 9.2 aporta datos sobre x = volumen de precipitación pluvial y y = volumen de escurrimiento (ambos en m^3). Use los resultados adjuntos obtenidos con Minitab para decidir si existe una relación lineal útil entre la precipitación pluvial y el escurrimiento, y luego calcule un intervalo de confianza para el cambio promedio verdadero del volumen de escurrimiento asociado con 1 m^3 de incremento del volumen de precipitación pluvial.

```
The regression equation is
escurrimiento  = −1.13 + 0.827 precipitación pluvial

Predictor            Coef      Stdev    t-ratio        P
Constant           −1.128      2.368     −0.48    0.642
precipitación     0.82697    0.03652      22.64    0.000
pluvial

s = 5.240          R-sq = 97.5%     R-sq(adj) = 97.3%
```

33. El ejercicio 15 de la sección 9.2 incluyó resultados generados por Minitab del módulo de elasticidad con una regresión de resistencia a la flexión de vigas de concreto.

a. Úselos para calcular un intervalo de confianza con un nivel de confianza de 95% para la pendiente β_1 de la recta de regresión de población e interprete el intervalo resultante.

b. Suponga que anteriormente se creía que cuando el módulo de elasticidad se incrementa en 1 GPa, el cambio promedio verdadero asociado de la resistencia a la flexión era cuando mucho de 0.1 MPa. ¿Contradicen los datos esta creencia? Formule y pruebe las hipótesis pertinentes.

34. Las tecnologías electromagnéticas ofrecen eficaces técnicas de detección no destructivas para determinar las características del pavimento. La propagación de las ondas electromagnéticas a través del material depende de sus propiedades dieléctricas. Los siguientes datos, amablemente proporcionados por los autores del artículo **"Dielectric Modeling of Asphalt Mixtures and Relationship with Density" (*J. of Transp. Engr.*, 2011: 104–111)**, se utilizaron para relacionar y = constante dieléctrica a x = aire vacío (%) para 18 muestras que tienen un contenido de asfalto de 5%:

y	4.55	4.49	4.50	4.47	4.47	4.45	4.40	4.34	4.43
x	4.35	4.79	5.57	5.20	5.07	5.79	5.36	6.40	5.66
y	4.43	4.42	4.40	4.33	4.44	4.40	4.26	4.32	4.34
x	5.90	6.49	5.70	6.49	6.37	6.51	7.88	6.74	7.08

El siguiente resultado de R es de una regresión lineal simple de y sobre x:

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  4.858691    0.059768   81.283     <2e-16
Airevacío   −0.074676    0.009923   −7.526   1.21e-06

Residual standard error: 0.03551 on 16 DF Multiple
R-squared: 0.7797, Adjusted R-squared: 0.766 F-statistic:
56.63 on 1 and 16 DF, p-value: 1.214e-06

Analysis of Variance Table

Response: Dieléctrico
           Df    Sum Sq   Mean Sq  F value      Pr(>F)
Airevacío   1  0.071422  0.071422   56.635  1.214e-06
Residuals  16   0.20178  0.001261
```

a. Obtenga la ecuación de la recta de mínimos cuadrados e interprete su pendiente.

b. ¿Qué proporción de la variación observada en la constante dieléctrica se puede atribuir a la relación lineal aproximada entre la constante dieléctrica y el aire vacío?

c. ¿Habrá una relación lineal útil entre la constante dieléctrica y el aire vacío? Establezca y pruebe las hipótesis apropiadas.

d. Suponga que anteriormente se creía que cuando el aire vacío aumenta en 1 por ciento, el cambio promedio verdadero asociado con la constante dieléctrica sería al menos de -0.05. ¿Los datos de la muestra contradicen esta creencia? Realice una prueba de hipótesis apropiadas usando un nivel de significancia de 0.01.

35. ¿Cómo afecta la aceleración lateral, en las fuerzas laterales experimentadas en las curvas, que en gran medida están bajo el control del conductor, respecto a la sensación de náusea perciben los pasajeros de un autobús? El artículo **"Motion Sickness in Public Road Transport: The Effect of Driver, Route, and Vehicle" (*Ergonomics*, 1999: 1646-1664)** reporta datos sobre x = sensación de mareo provocado por el movimiento (calculado de acuerdo con una norma británica para evaluar movimientos similares en el mar) y y = sensación de náusea reportada (%). Las cantidades pertinentes son

$$n = 17,\ \sum x_i = 222.1,\ \sum y_i = 193,\ \sum x_i^2 = 3056.69,$$
$$\sum x_i y_i = 2759.6,\ \sum y_i^2 = 2975$$

Los valores de las sensaciones en la muestra oscilaron entre 6.0 y 17.6.

a. Suponiendo que el modelo de regresión lineal simple es válido para relacionar estas dos variables (esto es apoyado por los datos sin procesar), calcule e interprete un estimador del parámetro de pendiente que dé información sobre la precisión y la confiabilidad de la estimación.

b. ¿Habrá una relación lineal útil entre estas dos variables? Pruebe las hipótesis adecuadas usando $\alpha = 0.01$.

c. ¿Sería sensato utilizar el modelo de regresión lineal simple como base para predecir el % de náusea cuando la sensación = 5.0? Explique su razonamiento.

d. Cuando se utilizó Minitab para ajustar el modelo de regresión lineal simple a los datos sin procesar, la observación (6.0, 2.50) se señaló como un posible impacto sustancial en el ajuste. Elimine esta observación de la muestra y recalcule la estimación del inciso a). Con base en esto, ¿ejercerá la observación una influencia indebida?

36. Cuando se utilizan fluidos para remover metales durante las operaciones de maquinado para enfriar y lubricar la herramienta y la pieza de trabajo suele producirse bruma (gotas transportadas por el aire o aerosoles). La generación de bruma es una preocupación para la OSHA, la cual recientemente ha reducido sustancialmente la norma en los lugares de trabajo. El artículo **"Variables Affecting Mist Generation from Metal Removal Fluids" (*Lubrication Engr*., 2002: 10–17)** aporta los siguientes datos sobre x = velocidad de flujo de un aceite soluble al 5% (cm/s) y y = la cantidad de gotas de bruma con diámetro menor que 10 μm (mg/m^3):

x	89	177	189	354	362	442	965
y	0.40	0.60	0.48	0.66	0.61	0.69	0.99

a. Los investigadores realizaron un análisis de regresión lineal simple para relacionar las dos variables. ¿Apoya la gráfica de dispersión esta estrategia?

b. ¿Qué proporción de la variación observada de la bruma puede ser atribuida a la relación de regresión lineal simple entre velocidad y bruma?

c. A los investigadores les interesaba particularmente el impacto en la bruma de la velocidad creciente de 100 a 1000 (un factor de 10 correspondiente a la diferencia entre los valores x más pequeños y más grandes presentes en la muestra). Cuando x se incrementa de esta manera, ¿existe evidencia sustancial de que el incremento promedio verdadero de y es menor que 0.6?

d. Estime el cambio promedio verdadero de la bruma asociado con un incremento de 1 cm/s en la velocidad y hágalo de modo que dé información sobre precisión y confiabilidad.

37. La obtención de imágenes mediante resonancia magnética (MRI, por sus siglas en inglés) está bien establecida como una herramienta para medir velocidades de la sangre y en flujos de volúmenes. El artículo **"Correlation Analysis of Stenotic Aortic Valve Flow Patterns Using Phase Contrast MRI"**, citado en el ejercicio 1.67, propone utilizar esta metodología para determinar el área valvular en pacientes con estenosis aórtica. Los siguientes datos sobre velocidad pico (m/s), obtenidos en exámenes de 23 pacientes en dos planos diferentes, se tomaron de una gráfica que aparece en el artículo citado.

Nivel-	0.60	0.82	0.85	0.89	0.95	1.01	1.01	1.05
Nivel--	0.50	0.68	0.76	0.64	0.68	0.86	0.79	1.03
Nivel-	1.08	1.11	1.18	1.17	1.22	1.29	1.28	1.32
Nivel--	0.75	0.90	0.79	0.86	0.99	0.80	1.10	1.15
Nivel-	1.37	1.53	1.55	1.85	1.93	1.93	2.14	
Nivel--	1.04	1.16	1.28	1.39	1.57	1.39	1.32	

a. ¿Habrá alguna diferencia entre la velocidad promedio verdadera en los dos diferentes planos? Realice una prueba de hipótesis apropiada (como lo hicieron los autores del artículo).

b. Los autores del artículo también regresaron el nivel de velocidad-- contra nivel de velocidad-. La intersección y la pendiente estimadas resultantes son 0.14701 y 0.65393 con errores estándar estimados correspondientes de 0.07877 y 0.05947, coeficiente de determinación de 0.852 y s = 0.110673. El artículo señala que esta regresión mostraba evidencias de una fuerte relación lineal pero una pendiente de regresión muy por debajo de 1. ¿Está usted de acuerdo?

38. Remítase a los datos sobre x = tasa de liberación y y = tasa de emisión de NO_x dados en el ejercicio 19.

a. ¿Especifica el modelo de regresión lineal simple una relación útil entre las dos tasas? Use el procedimiento de prueba apropiado para obtener información sobre el valor P y luego saque una conclusión a nivel de significancia 0.01.

b. Calcule un intervalo de confianza de 95% para el cambio esperado en la tasa de emisiones asociado con un incremento de 10 MBtu/h-pie^2 en la tasa de liberación.

39. Realice la prueba de utilidad de modelo mediante el método ANOVA con los datos de contenido de humedad-tasa de filtración del ejemplo 9.6. Verifique que se obtenga un resultado equivalente al de la prueba t.

40. Use las reglas del valor esperado para demostrar que es un estimador insesgado de β_0 (suponiendo que $\hat{\beta}_1$ es insesgado para β_1).

41. **a.** Verifique que $E(\hat{\beta}_1) = \beta_1$ usando las reglas de valor esperado del capítulo 5.

b. Use las reglas de varianza para verificar la expresión para $V(\hat{\beta}_1)$ dada en esta sección.

42. Si cada x_i se multiplica por una constante positiva c y cada y_i se multiplica por otra constante positiva d, verifique que el estadístico t para probar que H_0: $\beta_1 = 0$ contra H_a: $\beta_1 \neq 0$ no cambia de valor (el valor de $\hat{\beta}_1$ sí cambiará, lo cual demuestra que la magnitud de $\hat{\beta}_1$ no es indicativo por sí mismo de la utilidad de modelo).

43. La probabilidad de un error tipo II con la prueba t para H_0: $\beta_1 = \beta_{10}$ se calcula del mismo modo que para las pruebas t del

capítulo 8. Si el valor alternativo de β_1 es denotado por β_1', el valor de

$$d = \frac{|\beta_{10} - \beta_1'|}{\sigma\sqrt{\dfrac{n-1}{S_{xx}}}}$$

se calcula primero, luego se ingresa al conjunto apropiado de curvas de la tabla A.17 del apéndice por el eje horizontal con el valor de *d,* y β se lee en la curva de $n - 2$ grados de libertad. Un artículo que apareció en el ***Journal of Public Health Engineering*** reporta los resultados de un análisis de regresión basado en $n = 15$ observaciones en las cuales x = temperatura de aplicación de filtro (°C) y y = % de eficiencia de eliminación de BOD. Las cantidades calculadas incluyen $\Sigma x_i = 402$, $\Sigma x_i^2 = 11\,098$, $s = 3.725$ y $\hat{\beta}_1 = 1.7035$. Considere probar a un nivel de 0.01 H_0: $\beta_1 = 1$, la que manifiesta que el incremento esperado en el % de eliminación de BOD es 1 cuando la temperatura de aplicación del filtro se incrementa 1°C, contra la alternativa H_a: $\beta_1 > 1$. Determine P(error tipo II) cuando $\beta_1' = 2$, $\sigma = 4$.

9.4 Inferencias sobre $\mu_{Y\cdot x^*}$ y predicción de valores Y futuros

Sea x^* un valor específico de la variable independiente x. Una vez que las $\hat{\beta}_0$ y $\hat{\beta}_1$ estimadas han sido calculadas, $\hat{\beta}_1 + \hat{\beta}_1 x^*$ puede ser considerada una estimación puntual de $\mu_{Y\cdot x^*}$ (el valor esperado o el valor promedio real de Y cuando $x = x^*$) o un predicción del valor Y que resultará de una sola observación realizada cuando $x = x^*$. La estimación puntual o predicción por sí misma no da información de con cuánta precisión ha sido estimada $\mu_{Y\cdot x^*}$ o pronosticada Y. Esto se remedia desarrollando un intervalo de confianza para $\mu_{Y\cdot x^*}$ y un intervalo de predicción (IP) para un solo valor de Y.

Antes de obtener datos muestrales, tanto $\hat{\beta}_0$ como $\hat{\beta}_1$ están sujetas a variabilidad de muestreo; es decir, ambas son estadísticos cuyos valores variarán de muestra en muestra. Suponga, por ejemplo, que $\beta_0 = 50$ y $\beta_1 = 2$. Entonces una primera muestra de pares (x, y) podría dar $\hat{\beta}_0 = 52.35$, $\hat{\beta}_1 = 1.895$; una segunda muestra podría dar $\hat{\beta}_0 = 46.52$, $\hat{\beta}_1 = 2.056$; etcétera. De aquí se desprende que $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ misma cambia de valor de muestra en muestra, así que es un estadístico. Si la intersección y la pendiente de la recta de la población son los valores antes mencionados 50 y 2, respectivamente, y $x^* = 10$, entonces este estadístico está tratando de estimar el valor $50 + 2(10) = 70$. La estimación con una primera muestra podría ser $52.35 + 1.895(10) = 71.30$, con una segunda muestra podría ser $46.52 + 2.056(10) = 67.08$, etcétera.

Esta variación en el valor de $\hat{\beta}_0 + \hat{\beta}_1 x^*$ se puede visualizar regresando a la figura 9.13 en la página 511. Considere el valor $x^* = 300$. Las alturas de las 20 rectas de regresión estimada por encima de este valor son un poco diferentes entre sí. Lo mismo puede decirse de las alturas de las rectas por encima del valor $x^* = 350$. De hecho, al parecer hay más variación en el valor de $\hat{\beta}_0 + \hat{\beta}_1(350)$ que en el de $\hat{\beta}_0 + \hat{\beta}_1(300)$. Más adelante veremos que esto es porque 350 está más lejos de $\bar{x} = 235.71$ (el "centro de los datos") que 300.

Los métodos para hacer inferencias acerca de β_1 se basan en las propiedades de la distribución muestral del estadístico $\hat{\beta}_1$. De la misma manera, las inferencias sobre la media Y del valor $\hat{\beta}_0 + \hat{\beta}_1 x^*$ se basan en las propiedades de la distribución muestral del estadístico $\hat{\beta}_0 + \hat{\beta}_1 x^*$. La sustitución de las expresiones para $\hat{\beta}_0$ y $\hat{\beta}_1$ en $\hat{\beta}_0 + \hat{\beta}_1 x^*$ seguida de alguna manipulación algebraica lleva a la representación de $\hat{\beta}_0 + \hat{\beta}_1 x^*$ como una función lineal de las Y_i:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\right] Y_i = \sum_{i=1}^{n} d_i Y_i$$

Los coeficientes $d_1, d_2, \ldots, d_n$ en esta función lineal implican a las x_i y a las x^*, las cuales son fijas. La aplicación de las reglas de la probabilidad conjunta a esta función lineal aporta las siguientes propiedades.

PROPOSICIÓN

Sea $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, donde x^* es algún valor fijo de x. Entonces

1. La media de $\hat{Y}$ es

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

Así pues $\hat{\beta}_0 + \hat{\beta}_1 x^*$ es un estimador insesgado para $\beta_0 + \beta_1 x^*$, (es decir, para $\mu_{Y\cdot x^*}$).

2. La varianza de $\hat{Y}$ es

$$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum x_i^2 - \left(\sum x_i\right)^2/n}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

y la desviación estándar $\sigma_{\hat{Y}}$ es la raíz cuadrada de esta expresión. La desviación estándar estimada de $\hat{\beta}_0 + \hat{\beta}_1 x^*$, denotada por $s_{\hat{Y}}$ o $s_{\hat{\beta}_0} + s_{\hat{\beta}_1 x^*}$ se obtiene al reemplazar σ por su estimación s:

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y}$ tiene una distribución normal.

La varianza de $\hat{\beta}_0 + \hat{\beta}_1 x^*$ es más pequeña cuando $x^* = \bar{x}$ y se incrementa a medida que x^* se aleja de $\bar{x}$ en una u otra dirección. Por consiguiente, la estimación de $\mu_{Y\cdot x^*}$ es más precisa cuando x^* está cerca del centro de las x_i que cuando está lejos de los valores x a los cuales se les realizaron las observaciones. Esto implicará que el intervalo de confianza así como el intervalo de predicción sean más angostos con una x^* cerca de $\bar{x}$ que con una x^* lejos de $\bar{x}$. La mayoría de los programas de computadora dan tanto $\hat{\beta}_0 + \hat{\beta}_1 x^*$ como $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$ con cualquier x^* especificado.

Inferencias sobre $\mu_{Y\cdot x^*}$

Así como los procedimientos inferenciales para β_1 se basaron en la variable t, obtenida estandarizando β_1, aquí una variable t obtenida estandarizando $\hat{\beta}_0 + \hat{\beta}_1 x^*$ conduce a un intervalo de confianza y procedimientos de prueba.

TEOREMA

La variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}} \qquad (9.5)$$

tiene una distribución t con $n - 2$ grados de libertad.

Un enunciado de probabilidad implica que esta variable estandarizada ahora puede ser manipulada para producir un intervalo de confianza para $\mu_{Y\cdot x^*}$.

Un **intervalo de confianza** de $100(1 - \alpha)\%$ **para** $\mu_{Y\cdot x^*}$, el valor esperado de Y cuando $x = x^*$, es

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2,\,n-2} \cdot s_{\hat{Y}} \qquad (9.6)$$

Este intervalo de confianza está centrado en la estimación puntual de $\mu_{Y\cdot x^*}$ y se extiende a cada lado en una cantidad que depende del nivel de confianza y del grado de variabilidad del estimador en el cual está basada la estimación puntual.

EJEMPLO 9.13 La corrosión de varillas de refuerzo de acero es el problema de durabilidad más importante de las estructuras de concreto reforzadas. La carbonatación del concreto ocurre a consecuencia de una reacción química que reduce el pH lo suficiente para iniciar la corrosión de las varillas de refuerzo. A continuación se dan datos representativos sobre x = profundidad de carbonatación (mm) y y = resistencia (MPa) para una muestra de especímenes testigo tomados de un edificio particular (obtenidos de una gráfica que se ilustra en el artículo **"The Carbonation of Concrete Structures in the Tropical Environment of Singapore", *Magazine of Concrete Res*., 1996: 293–300**).

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4
x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8

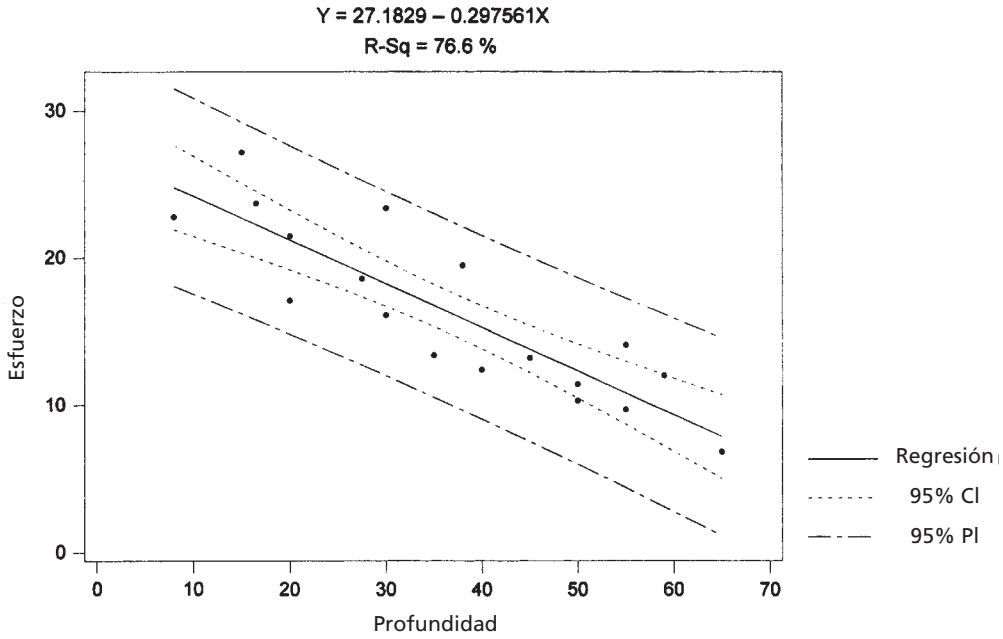


Figura 9.17 Diagrama de dispersión generado por Minitab con intervalos de confianza e intervalos de predicción con los datos del ejemplo 9.13

Una gráfica de dispersión de los datos (véase la figura 9.17) apoya fuertemente el uso del modelo de regresión lineal simple. Las siguientes son cantidades pertinentes:

$$\sum x_i = 659.0 \qquad \sum x_i^2 = 28\,967.50 \qquad \bar{x} = 36.6111 \qquad S_{xx} = 4840.7778$$
$$\sum y_i = 293.2 \qquad \sum x_i y_i = 9293.95 \qquad \sum y_i^2 = 5335.76$$
$$\hat{\beta}_1 = -0.297561 \qquad \hat{\beta}_0 = 27.182936 \qquad \text{SSE} = 131.2402$$
$$r^2 = 0.766 \qquad s = 2.8640$$

Calcule ahora un intervalo de confianza, utilizando un nivel de confianza de 95% para la resistencia media de todos los especímenes testigo que tienen una profundidad de carbonatación de 45 mm; es decir, un intervalo de confianza para $\beta_0 + \beta_1(45)$. El intervalo está centrado en

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(45) = 27.18 - 0.2976(45) = 13.79$$

La desviación estándar estimada del estadístico $\hat{Y}$ es

$$s_{\hat{Y}} = 2.8640\sqrt{\frac{1}{18} + \frac{(45 - 36.6111)^2}{4840.7778}} = 0.7582$$

El valor crítico t de 16 grados de libertad para un nivel de confianza de 95% es 2.120, con el cual se determina que el intervalo deseado es

$$13.79 \pm (2.120)(0.7582) = 13.79 \pm 1.61 = (12.18, 15.40)$$

La angostura de este intervalo sugiere que se tiene información razonablemente precisa sobre la media que se está estimando. Recuerde que si recalcula este intervalo muestra tras muestra, a la larga aproximadamente 95% de los intervalos calculados incluiría $\beta_0 + \beta_1(45)$. Sólo se puede esperar que esta media quede en el intervalo único que se calculó.

La figura 9.18 muestra resultados Minitab obtenidos por una solicitud de ajustar el modelo de regresión lineal simple y calcular intervalos de confianza para la media de resistencia a profundidades de 45 mm y 35 mm. Los intervalos aparecen en la parte inferior de los resultados; observe que el segundo intervalo es más angosto que el primero, porque 35 está mucho más cerca de $\bar{x}$ que 45. La figura 9.17 muestra 1) curvas correspondientes a los límites de confianza con cada valor x diferente y 2) límites de predicción que se discutirán en breve. Observe cómo las curvas se alejan cada vez más a medida que x se aleja de $\bar{x}$.

```
The regression equation is   resistencia = 27.2 − 0.298 profundidad

Predictor           Coef         Stdev      t-ratio       P
 Constant          27.183         1.651       16.46   0.000
   profundidad   −0.29756       0.04116      −7.23   0.000
s = 2.864        R-sq = 76.6%      R-sq(adj) = 75.1%

Analysis of Variance

SOURCE             DF         SS         MS          F         P
Regression          1     428.62     428.62      52.25     0.000
     Error         16     131.24       8.20
     Total         17     559.86

   Fit         Stdev.Fit           95.0% C.I.           95.0% P.I.
13.793             0.758     (12.185, 15.401)      (7.510, 20.075)

   Fit         Stdev.Fit           95.0% C.I.           95.0% P.I.
16.768             0.678     (15.330, 18.207)     (10.527, 23.009)
```

Figura 9.18 Resultados de regresión obtenidos con Minitab con los datos del ejemplo 9.13 ■

En algunas situaciones se desea un intervalo de confianza no sólo para un solo valor x, sino para dos o más valores x. Suponga que un investigador desea un intervalo de confianza tanto para $\mu_{Y\cdot\nu}$ como para $\mu_{Y\cdot w}$, donde ν y w son dos valores diferentes de la variable independiente. Es tentador calcular el intervalo de confianza (9.6) primero con $x = \nu$ y luego con $x = w$. Suponga que se utiliza $\alpha = 0.05$ en cada cálculo para obtener dos intervalos de 95%. Luego, si las variables implicadas al calcular los dos intervalos fueran independientes una de otra, el nivel de confianza conjunto sería $(0.95)\cdot(0.95) \approx 0.90$.

Sin embargo, los intervalos no son independientes porque se utilizan las mismas $\hat{\beta}_0$, $\hat{\beta}_1$, y S en cada uno. Por consiguiente, no se puede aseverar que el nivel de confianza conjunto para los dos intervalos sea exactamente de 90%. Se puede demostrar, no obstante, que si el intervalo de confianza de $100(1 - \alpha)\%$ (9.6) se calcula tanto con $x = \nu$ como con $x = w$ para obtener intervalos de confianza conjuntos para $\mu_{Y\cdot\nu}$ y $\mu_{Y\cdot w}$, entonces *el nivel de confianza conjunto en el par de intervalos resultante es al menos de* $100(1 - 2\alpha)\%$. En particular, si se utiliza $\alpha = 0.05$ se obtiene un nivel de confianza conjunto de *al menos* 90%, en tanto que si se utiliza $\alpha = 0.01$ se obtiene una confianza de al menos 98%. Así, en el ejemplo 9.13, un intervalo de confianza de 95% para $\mu_{Y\cdot 45}$ fue (12.185, 15.401) y un intervalo de confianza de 95% para $\mu_{Y\cdot 35}$ fue (15.330, 18.207). El nivel de confianza simultáneo o conjunto para las dos proposiciones $12.185 < \mu_{Y\cdot 45} < 15.401$ y $15.330 < \mu_{Y\cdot 35} < 18.207$ es al menos de 90%.

La validez de estos intervalos de confianza conjuntos o simultáneos se fundamenta en un resultado de probabilidad llamado **desigualdad de Bonferroni**, así que los intervalos de confianza conjuntos se conocen como **intervalos de Bonferroni**. El método es fácil de generalizar para que dé intervalos conjuntos para k diferentes $\mu_{Y \cdot x}$. *Utilizando el intervalo* (12.6) *por separado primero con* $x = x_1^*$, *luego con* $x = x_2^*$,..., *y finalmente con* $x = x_k^*$ *se obtiene un conjunto de k intervalos de confianza para los cuales está garantizado que el nivel de confianza simultáneo o conjunto es al menos de* $100(1 - k\alpha)\%$.

Las pruebas de hipótesis respecto a $\beta_0 + \beta_1 x^*$ están basadas en el estadístico de prueba T obtenido al reemplazar $\beta_0 + \beta_1 x^*$ en el numerador de la expresión (9.5) por el valor nulo de μ_0. Por ejemplo, H_0: $\beta_0 + \beta_1(45) = 15$ en el ejemplo 9.13 expresa que cuando la profundidad de carbonatación es de 45, la resistencia esperada (es decir, el promedio verdadero) es de 15. El valor del estadístico de prueba es entonces $t = [\hat{\beta}_0 + \hat{\beta}_1(45) - 15]/s_{\hat{\beta}_0 + \hat{\beta}_1(45)}$, y la prueba es de cola superior, inferior o de dos colas de acuerdo con la desigualdad en H_a.

Intervalo de predicción para un valor futuro de Y

A menos que se calcule un intervalo estimado para $\mu_{Y \cdot x^*}$ un investigador quizá desee obtener un intervalo de valores factibles para el valor de Y asociado con alguna observación futura cuando la variable independiente tiene el valor x^*. Por ejemplo, el tamaño del vocabulario y está relacionado con la edad x de un niño. El intervalo de confianza (9.6) con $x^* = 6$ podría proporcionar un estimado del tamaño de vocabulario promedio verdadero de todos los niños de 6 años. Alternativamente, se podría desear un intervalo de valores factibles para el tamaño del vocabulario de un niño particular de 6 años.

Un intervalo de confianza se refiere a un parámetro, o característica de población, cuyo valor es fijo pero desconocido. En contraste, un valor futuro de Y no es un parámetro sino una variable aleatoria; por esta razón se hace referencia a un intervalo de valores factibles para un valor Y futuro como **intervalo de predicción** en lugar de intervalo de confianza. El error de estimación es $\beta_0 + \beta_1 x^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, una diferencia entre una cantidad fija (pero desconocida) y una variable aleatoria. El error de predicción es $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, una diferencia entre dos variables aleatorias. Existe, por tanto, más incertidumbre en la predicción que en la estimación, así que un intervalo de predicción será más ancho que un intervalo de confianza. Debido a que el valor futuro Y es independiente de las Y_i observadas,

$$\begin{aligned} V[Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)] &= \text{varianza del error de predicción} \\ &= V(Y) + V(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

Además, puesto que $E(Y) = \beta_0 + \beta_1 x^*$ y $E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \beta_0 + \beta_1 x^*$, el valor esperado del error de predicción es $E(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)) = 0$. Se puede demostrar entonces que la variable estandarizada

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

tiene una distribución t con $n - 2$ grados de libertad. Al sustituir esta T en la proposición de probabilidad $P(-t_{\alpha/2,\, n-2} < T < t_{\alpha/2,\, n-1}) = 1 - \alpha$ y manipulándola para aislar Y entre las dos desigualdades se obtiene el siguiente intervalo

Un **intervalo de predicción de** $100(1-\alpha)\%$ **para una observación** Y **futura que se va a realizar cuando** $x = x^*$ es

$$\begin{aligned}&\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} \\ &= \hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{Y}}}\end{aligned} \tag{9.7}$$

La interpretación del nivel de predicción de $100(1-\alpha)\%$ es idéntica al de los niveles de confianza previos; si se utiliza la fórmula (9.7) repetidamente, a la larga los intervalos resultantes en realidad contendrán los valores y observados $100(1-\alpha)\%$ del tiempo. Observe que el 1 debajo de la raíz cuadrada inicial hace que el intervalo de predicción (9.7) sea más ancho que el intervalo de confianza (9.6), aun cuando ambos intervalos estén centrados en $\hat{\beta}_0 + \hat{\beta}_1 x^*$. Además, a medida que $n \to \infty$, el ancho del intervalo de confianza tiende a cero, en tanto que el ancho del intervalo de predicción no (porque incluso con el perfecto conocimiento de β_0 y β_1, seguirá habiendo incertidumbre en la predicción).

EJEMPLO 12.14 Regrese a los datos de profundidad de carbonatación-resistencia del ejemplo 9.13 y calcule un intervalo de predicción de 95% para un valor de resistencia que resultaría de seleccionar un solo espécimen testigo cuya profundidad de carbonatación es de 45 mm. Cantidades pertinentes del ejemplo son

$$\hat{y} = 13.79 \qquad s_{\hat{Y}} = 0.7582 \qquad s = 2.8640$$

Para un nivel de predicción de 95% basado en $n - 2 = 16$ grados de libertad, el valor crítico es 2.120, exactamente el que se utilizó previamente para un nivel de confianza de 95%. El intervalo de predicción es entonces

$$\begin{aligned}13.79 \pm (2.120)\sqrt{(2.8640)^2 + (0.7582)^2} &= 13.79 \pm (2.120)(2.963) \\ &= 13.79 \pm 6.28 = (7.51, 20.07)\end{aligned}$$

Valores factibles para una sola observación de resistencia cuando la profundidad es de 45 mm son (al nivel de predicción de 95%) entre 7.51 y 20.07 MPa. El intervalo de confianza de 95% para una resistencia media, cuando la profundidad es de 45, fue (12.18, 15.40). El intervalo de predicción es mucho más ancho que esto debido a los $(2.8640)^2$ extra debajo de la raíz cuadrada. La figura 9.18, los resultados Minitab del ejemplo 9.13, muestra este intervalo, así como también el intervalo de confianza. ■

La técnica Bonferroni puede emplearse como en el caso del intervalo de confianza. Si se calcula un intervalo de predicción de $100(1-\alpha)\%$ para cada uno de los k valores diferentes de x, el nivel de predicción simultánea o conjunta para los k intervalos es al menos de $100(1-k\alpha)\%$.

EJERCICIOS **Sección 9.4 (44–56)**

44. El ajuste del modelo de regresión lineal simple a las $n = 27$ observaciones de x = módulo de elasticidad y y = resistencia a la flexión dados en el ejercicio 15 de la sección 12.2 dio por resultado en $\hat{y} = 7.592$, $s_{\hat{y}} = 0.179$ cuando $x = 40$ y $\hat{y} = 9.741$, $s_{\hat{y}} = 0.253$ para $x = 60$.

a. Explique por qué $s_{\hat{y}}$ es más grande cuando $x = 60$ que cuando $x = 40$.

b. Calcule un intervalo de confianza con un nivel de confianza de 95% para la resistencia promedio verdadera de todas las vigas cuyo módulo de elasticidad es 40.

c. Calcule un intervalo de predicción con un nivel de predicción 95% para la resistencia de una sola viga cuyo módulo de elasticidad es 40.

d. Si se calcula un intervalo de confianza de 95% para la resistencia promedio verdadera cuando el módulo de elasticidad es 60, ¿cuál será el nivel de confianza simultáneo tanto para este intervalo como para el intervalo calculado en el inciso b)?

45. Reconsidere los datos de contenido de humedad-tasa de filtración introducidos en el ejemplo 9.6 (véase también el ejemplo 9.7).

a. Calcule un intervalo de confianza de 90% para $\beta_0 + 125\beta_1$, el contenido de humedad promedio verdadero cuando la tasa de filtración es 125.

b. Pronostique el valor del contenido de humedad con un solo experimento en el cual la tasa de filtración es 125 utilizando un nivel de predicción de 90%. ¿Cómo se compara este intervalo al del inciso a)? ¿Por qué es este el caso?

c. ¿Cómo se compararán los intervalos de los incisos a) y b) con un intervalo de confianza y un intervalo de predicción cuando la tasa de filtración es 115? Responda sin calcular en realidad estos nuevos intervalos.

d. Interprete las hipótesis H_0: $\beta_0 + 125\beta_1 = 80$ y H_a: $\beta_0 + 125\beta_1 < 80$ y luego realice una prueba a un nivel de significancia de 0.01

46. La astringencia es la calidad de un vino que hace que la boca del bebedor lo perciba un poco áspero, seco y astringente. El documento **"Analysis of Tannins in Red Wine Using Multiple Methods: Correlation with Perceived Astringency" (*Amer. J. of Enol. and Vitic*., 2006: 481–485)** informó sobre una investigación para evaluar la relación entre la percepción de la astringencia y la concentración de taninos mediante diversos métodos analíticos. He aquí los datos proporcionados por los autores en x = concentración de taninos por la precipitación de proteínas y y = la astringencia percibida determinada por un panel de catadores.

x	0.718	0.808	0.924	1.000	0.667	0.529	0.514	0.559
y	0.428	0.480	0.493	0.978	0.318	0.298	−0.224	0.198
x	0.766	0.470	0.726	0.762	0.666	0.562	0.378	0.779
y	0.326	−0.336	0.765	0.190	0.066	−0.221	−0.898	0.836
x	0.674	0.858	0.406	0.927	0.311	0.319	0.518	0.687
y	0.126	0.305	−0.577	0.779	−0.707	−0.610	−0.648	−0.145
x	0.907	0.638	0.234	0.781	0.326	0.433	0.319	0.238
y	1.007	−0.090	−1.132	0.538	−1.098	−0.581	−0.862	−0.551

Las cantidades importantes se resumen como sigue:

$$\sum x_i = 19.404,\ \sum y_i = -0.549,\ \sum x_i^2 = 13.248032,$$
$$\sum y_i^2 = 11.835795,\ \sum x_i y_i = 3.497811$$
$$S_{xx} = 13.248032 - (19.404)^2/32 = 1.48193150,$$
$$S_{yy} = 11.82637622$$
$$S_{xy} = 3.497811 - (19.404)(-0.549)/32 = 3.83071088$$

a. Ajuste el modelo de regresión lineal simple a estos datos. Después, determine la proporción de la variación observada en la astringencia que se puede atribuir a la relación entre el modelo de astringencia y concentración de taninos.

b. Calcule e interprete un intervalo de confianza para la pendiente de la recta de regresión real.

c. Estime un promedio real de astringencia cuando la concentración de taninos es 0.6 y hágalo de una manera que transmita información acerca de la fiabilidad y la precisión.

d. Estime la astringencia del vino de una muestra única cuya concentración de taninos es 0.6 y hágalo de una manera que transmita información acerca de la fiabilidad y la precisión.

e. ¿Le parece que el promedio real de astringencia de una concentración de taninos de 0.7 es algo diferente de 0? Establezca y pruebe las hipótesis adecuadas.

47. El modelo de regresión lineal simple se ajusta muy bien a los datos de precipitación pluvial y volumen de escurrimiento dados en el ejercicio 16 de la sección 9.2. La ecuación de la recta de mínimos cuadrados es $y = -1.128 + 0.82697\,x$, $r^2 = 9.75$ y $s = 5.24$.

a. Use el hecho de que $s_{\hat{Y}} = 1.44$ cuando el volumen de la precipitación pluvial es de 40 m^3 para predecir el escurrimiento en una forma que transmita información sobre confiabilidad y precisión. ¿Sugiere el intervalo resultante que se dispone de información precisa sobre el valor de escurrimiento para esta futura observación? Explique su razonamiento.

b. Calcule un intervalo de predicción para escurrimiento cuando la precipitación pluvial es de 50, utilizando el mismo nivel de predicción del inciso a). ¿Qué se puede decir sobre el nivel de predicción simultáneo para los dos intervalos que calculó?

48. Una alcantarilla en un colector pluvial es la superficie de contacto entre el escurrimiento superficial y el conductor de desagüe. El dispositivo de la alcantarilla es un dispositivo que mejora las propiedades supresoras de contaminantes. El artículo **"An Evaluation of the Urban Stormwater Pollutant Demoral Efficiency of Catch Basin Inserts" (*Water Envir. Res*., 2005: 500–510)** reporta pruebas de varios dispositivos en condiciones controladas en las que el flujo de entrada es muy parecido al que se puede esperar en el campo. Considere los siguientes datos, tomados de una gráfica que aparece en el artículo, para un tipo particular de dispositivo sobre x cantidad filtrada (miles de litros) y y = % total de sólidos suspendidos eliminados.

x	23	45	68	91	114	136	159	182	205	228
y	53.3	26.9	54.8	33.8	29.9	8.2	17.2	12.2	3.2	11.1

Las cantidades resumidas son

$$\sum x_i = 1251,\ \sum x_i^2 = 199\,365,\ \sum y_i = 250.6,$$
$$\sum y_i^2 = 9249.36,\ \sum x_i y_i = 21\,904.4$$

a. ¿Avala la gráfica de dispersión la selección del modelo de regresión lineal simple? Explique.

b. Obtenga la ecuación de la recta de mínimos cuadrados.

c. ¿Qué proporción de la variación observada en el % de eliminación puede ser atribuida a la relación de modelo?

d. ¿Será que el modelo de regresión lineal simple especifica una relación útil? Realice una prueba de hipótesis apropiada con un nivel de significancia de 0.05.

e. ¿Será que hay una fuerte evidencia para concluir que al menos existe 2% de reducción promedio verdadera en la eliminación de sólidos suspendidos, asociada a un incremento

de 10 000 litros de la cantidad filtrada? Pruebe las hipótesis apropiadas con $\alpha = 0.05$.

f. Calcule e interprete un intervalo de confianza de 95% para el promedio verdadero de % eliminado cuando la cantidad filtrada es de 100 000 litros. ¿Cómo se compara este intervalo de ancho con el intervalo de confianza cuando la cantidad filtrada es de 200 000 litros?

g. Calcule e interprete un intervalo de predicción de 95% para el % eliminado cuando la cantidad filtrada es de 100 000 litros. ¿Cómo se compara el ancho de este intervalo con el del intervalo de confianza calculado en (f) y con el del intervalo de predicción cuando la cantidad filtrada es de 200 000 litros?

49. Le informan que un intervalo de confianza de 95% para el contenido de plomo esperado, cuando el flujo de tráfico es de 15 basado en una muestra de $n = 10$ observaciones, es (462.1, 597.7). Calcule un intervalo de confianza de 99% para el contenido de plomo esperado cuando el flujo de tráfico es de 15.

50. Se han utilizado aleaciones de silicio-germanio en ciertos tipos de celdas solares. El artículo **"Silicon-Germanium Films Deposited by Low-Frequency Plasma-Enhanced Chemical Vapor Deposition" (*J. of Material Res*., 2006: 88–104)** reporta sobre un estudio de varias propiedades estructurales y eléctricas. Considere los siguiente datos sobre x = concentración de Ge en fase sólida (desde 0 hasta 1) y y = posición de nivel Fermi (eV).

x	0	0.42	0.23	0.33	0.62	0.60	0.45	0.87	0.90	0.79	1	1	1
y	0.62	0.53	0.61	0.59	0.50	0.55	0.59	0.31	0.43	0.46	0.23	0.22	0.19

Un diagrama de dispersión muestra una relación lineal sustancial. He aquí una salida Minitab de un ajuste de mínimos cuadrados. [*Nota*: Existen varias inconsistencias entre los datos dados en el artículo, la gráfica que ahí se ilustra y la información resumida sobre un análisis de regresión.]

```
The regression equation is
Posición de nivel Fermi =
0.7217 − 0.4327 concentración de Ge

S = 0.0737573   R-Sq = 80.2%   R-Sq(adj) = 78.4%

Analysis of Variance

Source       DF        SS        MS       F      P
Regression    1  0.241728  0.241728   44.43  0.000
Error        11  0.059842  0.005440
Total        12  0.301569
```

a. Obtenga una estimación de intervalo del cambio esperado en la posición del nivel Fermi asociado con un incremento de 0.1 en la concentración de Ge, e interprete su estimación.

b. Obtenga una estimación de intervalo para la posición media del nivel Fermi cuando la concentración es de 0.50, e interprete su estimación.

c. Obtenga un intervalo de valores factibles para la posición que resulta de una sola observación que ha de realizarse cuando la concentración es de 0.50, interprete su intervalo y compare con el intervalo de b).

d. Obtenga intervalos de confianza simultáneos para la posición esperada cuando la concentración es de 0.3, 0.5 y 0.7; el nivel de confianza conjunto deberá ser al menos de 97%.

51. Remítase al ejemplo 12.12 en el cual x = velocidad en la pista de pruebas y y = velocidad de rodamiento de prueba.

a. Minitab dio $s_{\hat{\beta}_0+\hat{\beta}_1(45)} = 0.120$ y $s_{\hat{\beta}_0+\hat{\beta}_1(47)} = 0.186$. ¿Por qué la primera desviación estándar estimada es más pequeña que la segunda?

b. Use los resultados obtenidos con Minitab del ejemplo para calcular un intervalo de confianza de 95% para la velocidad de rodamiento de prueba esperada cuando la velocidad de prueba es = 45.

c. Use los resultados obtenidos con Minitab para calcular un intervalo de predicción de 95% para un solo valor de la velocidad de rodamiento cuando la velocidad de prueba es = 47.

52. El grabado con plasma es esencial en la transferencia de patrones de líneas finas en procesos de semiconductores de corriente. El artículo **"Ion Beam-Assisted Etching of Aluminum with Chlorine" (*J. of the Electrochem. Soc*., 1985: 2010–2012)** da los siguientes datos (tomados de una gráfica) sobre flujo de cloro (x, en SCCM) a través de una tobera utilizada en el mecanismo de grabado y en la velocidad de grabado (y, en 100 A/min).

x	1.5	1.5	2.0	2.5	2.5	3.0	3.5	3.5	4.0
y	23.0	24.5	25.0	30.0	33.5	40.0	40.5	47.0	49.0

Las cantidades estadísticas resumidas son $\Sigma x_i = 24.0$, $\Sigma y_i = 312.5$, $\Sigma x_i^2 = 70.50$, $\Sigma x_i y_i = 902.25$, $\Sigma y_i^2 = 11\ 626.75$, $\hat{\beta}_0 = 6.448718$, $\hat{\beta}_1 = 10.602564$.

a. ¿Será que el modelo de regresión lineal simple especifica una relación útil entre el flujo de cloro y la velocidad de grabado?

b. Estime el cambio promedio verdadero en la velocidad de grabado asociado con un incremento de 1 SCCM en la velocidad de flujo utilizando un intervalo de confianza de 95%, e interprete el intervalo.

c. Calcule a un intervalo de confianza de 95% para $\mu_{Y\cdot 3.0}$ la velocidad de grabado promedio verdadera cuando el flujo = 3.0. ¿Ha sido estimado con precisión este promedio?

d. Calcule un intervalo de predicción de 95% para una sola observación de velocidad de grabado que se realizará cuando el flujo = 3.0. ¿Es probable que la predicción sea precisa?

e. Cuando el flujo = 2.5 sea más ancho o más angosto que los intervalos correspondientes de los incisos c) y d), ¿serán los intervalos de confianza y predicción de 95%? Responda sin calcular en realidad los intervalos.

f. ¿Recomendaría calcular un intervalo de predicción de 95% para un flujo de 6.0? Explique.

53. Considere los siguientes cuatro intervalos basados en los datos del ejercicio 9.17 (sección 9.2):

a. Un intervalo de confianza de 95% para la porosidad media cuando el peso unitario es de 110.

b. Un intervalo de predicción de 95% para la porosidad cuando el peso unitario es de 110.

c. Un intervalo de confianza de 95% para la porosidad media cuando el peso unitario es de 115.

d. Un intervalo de predicción de 95% para la porosidad cuando el peso unitario es de 115.

Sin calcular ninguno de estos intervalos, ¿qué se puede decir sobre sus anchos uno respecto al otro?

54. La estatura de un paciente es útil para una variedad de propósitos médicos, tales como la estimación de volumen en una

unidad de cuidados intensivos que requiere ventilación artificial. Sin embargo, puede ser difícil hacer una medición exacta si el paciente está confuso, inconsciente o sedado. Y medir la estatura cuando un individuo está acostado no es sencillo. En cambio, las mediciones de longitud del cúbito son generalmente rápidas y fáciles de obtener, incluso en pacientes confinados a una silla o una cama. Los siguientes datos sobre x = longitud del cúbito (cm) y y = estatura (cm) para los varones mayores de 65 se obtuvieron de un gráfico en el artículo **"Ulna Length to Predict Height in English and Portuguese Patient Populations" (*European J. of Clinical Nutr.*, 2012: 209–215)**.

x	22.5	22.8	22.8	23.3	23.3	24.4	25.0
y	158	155	156	160	161	162	164

x	25.0	25.0	25.0	26.0	26.0	26.8	28.2
y	166	167	170	166	173	178	174

Las cantidades resumidas incluyen $\Sigma x_i = 346.1$, $\Sigma y_i = 2310$, $S_{xx} = 36.463571$, $S_{xy} = 137.60$, $S_{yy} = 626.00$.

a. Obtenga la ecuación de la recta de regresión estimada e interprete su pendiente.

b. Calcule e interprete el coeficiente de determinación.

c. Realice una prueba de la utilidad del modelo.

d. Calcule los intervalos de predicción para las estaturas de dos individuos cuya longitud del cúbito es 23 y 25, respectivamente; utilice un nivel de predicción de 95% para cada intervalo.

e. Con base en las predicciones de d), ¿está de acuerdo con el siguiente señalamiento en el citado artículo?: "la estatura puede predecirse con precisión a partir de la longitud del cúbito".

55. Verifique que $V(\hat{\beta}_0 + \hat{\beta}_1 x)$ se obtuvo, de hecho, mediante la expresión que aparece en el texto. [*Sugerencia*: $V(\Sigma d_i Y_i) = \Sigma d_i^2 \cdot V(Y_i)$.]

56. El artículo **"Bone Density and Insertion Torque as Predictors of Anterior Cruciate Ligament Graft Fixation Strength" (*The Amer. J. of Sports Med.*, 2004: 1421–1429)** proporciona los siguientes datos sobre el torque de inserción máximo (N · m) y carga de deformación (N), donde ésta mide la resistencia del injerto, correspondientes a 15 especímenes diferentes.

Torque	1.8	2.2	1.9	1.3	2.1	2.2	1.6	2.1
Carga	491	477	598	361	605	671	466	431

Torque	1.2	1.8	2.6	2.5	2.5	1.7	1.6
Carga	384	422	554	577	642	348	446

a. ¿Es factible que la carga de cedencia esté normalmente distribuida?

b. Estime la carga de deformación promedio verdadera calculando un intervalo de confianza con un nivel de confianza de 95% e interprételo.

c. Los siguientes son resultados obtenidos con Minitab para la regresión de la carga de deformación generada por el torque. ¿Será que el modelo de regresión lineal simple especifica una relación útil entre las variables?

```
Predictor       Coef    SE Coef       T       P
Constant      152.44      91.17    1.67   0.118
Torque        178.23      45.97    3.88   0.002

S = 73.2141   R-Sq = 53.6%   R-Sq(adj) = 50.0%

Source           DF       SS      MS      F      P
Regression        1    80554   80554  15.03  0.002
Residual Error   13    69684    5360
Total            14   150238
```

d. Los autores del artículo citado aseguran que "por consiguiente, no se puede sino concluir que los métodos basados en el análisis de regresión simple no son clínicamente suficientes para predecir la resistencia de fijación individual". ¿Está usted de acuerdo? [*Sugerencia*: Considere predecir la carga de deformación cuando el torque es de 2.0.]

9.5 Correlación

Hay muchas situaciones en las que al estudiar el comportamiento conjunto de dos variables el objetivo es ver si están relacionadas en lugar de utilizar una para predecir el valor de la otra. En esta sección primero se desarrolla el coeficiente de correlación muestral r como una medida de qué tan fuerte es la relación entre dos variables x y y en una muestra y luego se relaciona r con el coeficiente de correlación ρ.

Coeficiente de correlación muestral *r*

Dados n pares numéricos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, es natural hablar de que x y y tienen una relación positiva si las x grandes se parean con y grandes y las x pequeñas con y pequeñas. Asimismo, si las x grandes se parean con y pequeñas y las x pequeñas con y grandes, se asume una relación negativa entre las variables. Considere la cantidad

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Luego, si la relación es fuertemente positiva, una x_i por encima de la media $\bar{x}$ tenderá a aparearse con una y_i por encima de la media $\bar{y}$, de modo que $(x_i - \bar{x})(y_i - \bar{y}) > 0$ y este producto también será positivo siempre y cuando x_i y y_i estén por debajo de sus medias respectivas. De este modo una relación positiva implica que S_{xy} será positiva. Un argumento análogo demuestra que cuando la relación es negativa, S_{xy} será negativa, puesto que la mayoría de los productos $(x_i - \bar{x})(y_i - \bar{y})$ seguirán siendo negativos. Esto se ilustra en la figura 9.19.

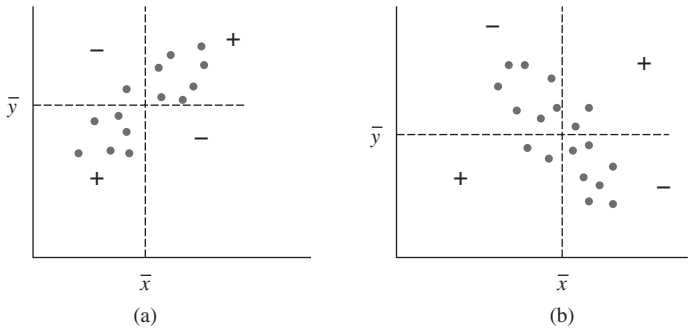


Figura 9.19 (a) Gráfica de dispersión con S_{xy} positiva; (b) gráfica de dispersión con S_{xy} negativa [+ significa $(x_i - \bar{x})(y_i - \bar{y}) > 0$ y − significa $(x_i - \bar{x})(y_i - \bar{y}) < 0$]

Aunque S_{xy} parece ser una medida factible de la fuerza de una relación, aún no se sabe qué tan positiva o negativa pueda ser. Desafortunadamente, S_{xy} tiene un serio defecto. Si se cambian las unidades de medición de x o y, se puede hacer que S_{xy} sea arbitrariamente grande en magnitud o arbitrariamente próxima a cero. Por ejemplo, si $S_{xy} = 25$ cuando x se mide en metros, $S_{xy} = 25\,000$ cuando x se mide en milímetros y 0.025 cuando x está expresada en kilómetros. Una condición razonable para imponer cualquier medida de qué tan fuerte es la relación entre x y y es que la medida calculada no dependa de las unidades particulares utilizadas para medirlas. Esta condición se cumple modificando S_{xy} para obtener el coeficiente de correlación muestral.

DEFINICIÓN

El **coeficiente de correlación muestral** para los n pares es $(x_1, y_1),\ldots, (x_n, y_n)$ es

$$r = \frac{S_{xy}}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{9.8}$$

EJEMPLO 9.15 Una evaluación precisa de la productividad del suelo es crítica para una planificación racional del uso del suelo. Desafortunadamente, tal como argumenta el autor del artículo **"Productivity Ratings Based on Soil Series" (*Prof. Geographer*, 1980: 158–163)**, no es fácil obtener un índice de productividad del suelo aceptable. Una dificultad es que la productividad está determinada en parte por el tipo de cosecha y la relación entre el rendimiento de dos cosechas diferentes plantadas en el mismo suelo puede no ser muy fuerte. Con fines ilustrativos, el artículo presenta los siguientes datos sobre una cosecha de maíz x y una cosecha de cacahuates y (mT/Ha) para ocho tipos diferentes de suelo.

x	2.4	3.4	4.6	3.7	2.2	3.3	4.0	2.1
y	1.33	2.12	1.80	1.65	2.00	1.76	2.11	1.63

Con $\Sigma x_i = 25.7$, $\Sigma y_i = 14.40$, $\Sigma x_i^2 = 88.31$, $\Sigma x_i y_i = 46.856$ y $\Sigma y_i^2 = 26.4324$,

$$S_{xx} = 88.31 - \frac{(25.7)^2}{8} = 5.75 \quad S_{yy} = 26.4324 - \frac{(14.40)^2}{8} = 0.5124$$

$$S_{xy} = 46.856 - \frac{(25.7)(14.40)}{8} = 0.5960$$

de donde

$$r = \frac{0.5960}{\sqrt{5.75}\sqrt{0.5124}} = 0.347$$

■

Propiedades de *r*

Las propiedades más importantes de *r* son las siguientes:

1. El valor de *r* no depende de cuál de las dos variables estudiadas es *x* o cuál es *y*.
2. El valor de *r* es independiente de las unidades en las cuales se midan *x* y *y*.
3. $-1 \leq r \leq 1$
4. $r = 1$ si y sólo si todos los pares (x_i, y_i) quedan en una línea recta con pendiente positiva, y $r = -1$ si y sólo si los pares (x_i, y_i) quedan en una línea recta con pendiente negativa.
5. El cuadrado del coeficiente de correlación muestral da el valor del coeficiente de determinación que resultaría de ajustar el modelo de regresión lineal simple, en símbolos $(r)^2 = r^2$.

La propiedad 1 contrasta con lo que sucede en el análisis de regresión donde virtualmente todas las cantidades de interés (la pendiente estimada, la intercepción *y* estimada, s^2, etc.) dependen de cuál de las dos variables es tratada como la variable dependiente. Sin embargo, la propiedad 5 demuestra que la proporción de variación de la variable dependiente explicada al ajustar el modelo de regresión lineal simple no depende de cuál variable desempeñe este rol.

La propiedad 2 equivale a decir que *r* no cambia si cada x_i es reemplazada por cx_i y si cada y_i es reemplazada por dy_i (un cambio en la escala de medición), así como también si cada x_i es reemplazada por $x_i - a$ y y_i por $y_i - b$ (lo que cambia la ubicación de cero en el eje de medición). Esto implica, por ejemplo, que *r* es el mismo si la temperatura se mide en °F o °C.

La propiedad 3 dice que el valor máximo de *r*, correspondiente al grado más grande posible de relación positiva, es $r = 1$, mientras que la relación más negativa está identificada con $r = -1$. De acuerdo con la propiedad 4, las correlaciones positivas y negativas más grandes se obtienen sólo cuando todos los puntos quedan a lo largo de una línea recta. Cualquier otra configuración de puntos, aun cuando la configuración sugiere una relación determinística entre las variables, dará un valor *r* menor que 1 en magnitud absoluta. Por consiguiente *r mide el grado de relación lineal* entre las variables. Un valor de *r* cercano a cero no es evidencia de la falta de una fuerte relación, sino sólo de la ausencia de una relación lineal, de modo que tal valor de *r* debe ser interpretado con precaución. La figura 9.20 ilustra varias configuraciones de puntos asociadas con diferentes valores de *r*.

Una pregunta que suele plantearse es, "¿cuándo se puede decir que existe una correlación fuerte entre las variables y cuándo una débil?". He aquí una regla de oro informal para caracterizar el valor de *r*:

Débil	**Moderada**	**Fuerte**
$-0.5 \leq r \leq 0.5$	si $-0.8 < r < -0.5$ o $0.5 < r < 0.8$	si $r \geq 0.8$ o $r \leq -0.8$

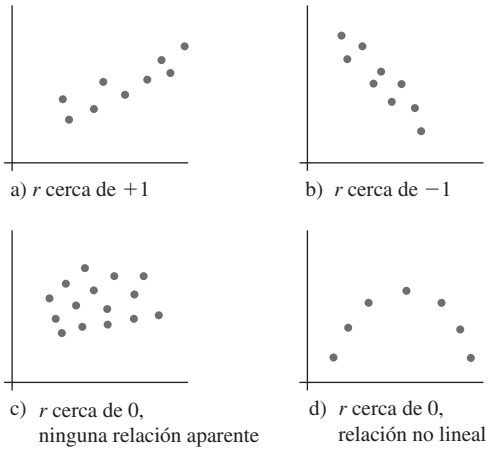


Figura 9.20 Gráficas de dispersión con diferentes valores de r

Quizá le sorprenda que una r tan sustancial como 0.5 o -0.5 aparezca en la categoría débil. La razón es que si $r = 0.5$ o -0.5, entonces $r^2 = 0.25$ en una regresión con cualquiera de las variables en el papel de y. Un modelo de regresión que explica la mayor parte del 25% de la variación observada en realidad no es muy impresionante. En el ejemplo 9.15 la correlación entre la cosecha de maíz y la cosecha de cacahuates se describiría como débil.

Inferencias sobre el coeficiente de correlación de una población

El coeficiente de correlación r mide qué tan fuerte es la relación entre x y y en la muestra observada. Se puede pensar que los pares (x_i, y_i) se sacaron de una población de pares bivariantes, con (X_i, Y_i) teniendo alguna función de masa de probabilidad o función de densidad de probabilidad conjunta. Se define el coeficiente de correlación $\rho(X, Y)$ mediante

$$\rho = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

donde

$$\text{cov}\,(X, Y) = \begin{cases} \displaystyle\sum_x \sum_y (x - \mu_X)(y - \mu_Y) p(x, y) & (X, Y) \text{ discreto} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\, dx\, dy & (X, Y) \text{ continuo} \end{cases}$$

Si se considera que $p(x, y)$ o $f(x, y)$ describen la distribución de pares de valores dentro de toda la población, ρ se transforma en la medida de qué tan fuertemente están relacionadas x y y en la población.

El coeficiente de correlación de la población ρ es un parámetro o característica de la población, exactamente como lo son μ_X, μ_Y, σ_X y σ_Y, así que se puede utilizar el coeficiente de correlación muestral para hacer varias inferencias sobre ρ. En particular, r es una estimación puntual de ρ y el estimador correspondiente es

$$\hat{\rho} = R = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2} \sqrt{\sum (Y_i - \overline{Y})^2}}$$

EJEMPLO 9.16 Los investigadores médicos han observado que las adolescentes tienen mucho más probabilidad de parir bebés con bajo peso que las adultas. Dado que estos bebés tienen mayores tasas de mortalidad numerosas investigaciones se han centrado en la relación entre la edad de la madre y el peso del bebé al nacer. Uno de estos estudios se describe en el artículo **"Body Size and Intelligence in 6-Year-Olds: Are Offspring of Teenage Mothers at Risk? (*Maternal and Child Health J.*, 2009: 847–856)**. La siguiente información en x = edad de la madre (años) y y = peso del bebé al nacer (g) es consistente con las cantidades resumidas dadas en el citado artículo, así como con los datos publicados por el National Center for Health Statistics.

x	15	17	18	15	16	19	17	16	18	19
y	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

Un diagrama de dispersión de los datos muestra un patrón lineal creciente algo sustancial. Las cantidades resumidas correspondientes son $\Sigma x_i = 170$, $\Sigma y_i = 30\,041$, $\Sigma x_i^2 = 3910$, $\Sigma y_i^2 = 91\,785\,351$, $\Sigma x_i y_i = 515\,600$, a partir de lo cual $S_{xx} = 20$, $S_{yy} = 1\,539\,182.90$, $S_{xy} = 4903$. Por tanto,

$$r = \frac{4903}{\sqrt{20}\sqrt{1\,539\,182.90}} = 0.884$$

Con ρ que denota la correlación entre la edad de la madre y peso del bebé en toda la población de adolescentes que parieron, la estimación puntual de ρ es $\hat{\rho} = r = 0.884$. ■

Los intervalos de muestra pequeña y los procedimientos de prueba presentados en los capítulos 7 al 9 están basados en la suposición de normalidad de la población. Para probar las hipótesis sobre ρ es necesario hacer una suposición análoga acerca de la distribución de los pares de valores (x, y) en la población. Ahora se supone que *tanto X como Y* son aleatorios (gran parte del trabajo de regresión se realizó con x fijada por el experimentador) con una distribución bivariada de probabilidad normal tal como se describe en la sección 5.2. Recuerde que en este caso $\rho = 0$ implica que X y Y son independientes de la variable aleatoria.

Suponer que los pares se tomaron de una distribución normal bivariada permite probar hipótesis sobre ρ y construir un intervalo de confianza. No existe una forma completamente satisfactoria de verificar la factibilidad de la suposición de normalidad bivariada. Una verificación parcial implica construir dos gráficas de probabilidad normal separadas, una para las x_i y otra para las y_i, puesto que la normalidad bivariada implica que las distribuciones marginales tanto de X como de Y son normales. Si cualquiera de las gráficas se aparta sustancialmente de un patrón de línea recta, no se deberán utilizar los siguientes procedimientos inferenciales para una n pequeña.

Prueba en cuanto a la ausencia de correlación

Sea R el coeficiente de correlación de la muestra como una variable aleatoria (antes de que se obtengan los datos). Cuando H_0: $\rho = 0$ es verdadera el estadístico de prueba

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

tiene una distribución t con $n - 2$ grados de libertad.

Hipótesis alternativa	Determinación del valor *P*
H_a: $\rho > 0$	Área bajo la curva t_{n-2} a la derecha de t
H_a: $\rho < 0$	Área bajo la curva t_{n-2} a la izquierda de t
H_a: $\rho \neq 0$	2·(Área bajo la curva t_{n-2} a la derecha de $\lvert t \rvert$)

EJEMPLO 9.17 Los efectos neurotóxicos del manganeso son bien conocidos y normalmente son provocados por la prolongada exposición laboral durante largos periodos. En los campos de higiene ocupacional e higiene ambiental no había sido reportada con anterioridad la relación entre la peroxidación de los lípidos, la cual es responsable del deterioro de los alimentos y de los daños en tejidos vivos, y la exposición laboral. El artículo **"Lipid Peroxidation in Workers Exposed to Manganese" (*Scand. J. of Work and Environ. Health*, 1996: 381–386)** proporciona información sobre x = concentración de manganeso en sangre (ppb) y y = concentración ($m$mol/L) de malondialdehído, el cual es el producto estable de la peroxidación de los lípidos, tanto para una muestra de 22 trabajadores expuestos a manganeso como para una muestra de control de 45 individuos. El valor de r para la muestra de control fue de 0.29, por lo que

$$t = \frac{(0.29)\sqrt{45-2}}{\sqrt{1-(0.29)^2}} \approx 2.0$$

El valor P correspondiente para una prueba t de dos colas basada en 43 grados de libertad es aproximadamente de 0.052 (el artículo citado sólo reporta que el valor $P > 0.05$). No se desearía rechazar la aseveración de que $\rho = 0$ al nivel de significancia 0.01 o 0.05. Para la muestra de trabajadores expuestos, $r = 0.83$ y $t \approx 6.7$, existe una clara evidencia de que existe una relación lineal en toda la población de trabajadores expuestos de la cual fue seleccionada la muestra. ■

Puesto que ρ mide el grado al cual existe una relación lineal entre las dos variables en la población, la hipótesis nula H_0: $\rho = 0$ manifiesta que no existe tal relación de población. Se puede utilizar la expresión $t\ \hat{\beta}_1/s_{\hat{\beta}_1}$ para probar una relación lineal entre las dos variables en el contexto del análisis de regresión. Resulta que ambos procedimientos de prueba son completamente equivalentes porque $r\sqrt{n-2}/\sqrt{1-r^2} = \hat{\beta}_1/s_{\hat{\beta}_1}$. Cuando el interés radica sólo en valorar la fuerza de cualquier relación lineal en lugar de ajustarse a un modelo y utilizarlo para estimar o predecir, la fórmula del estadístico de prueba que se acaba de presentar requiere menos cálculos que la relación t.

Otras inferencias sobre r

El procedimiento para probar H_0: $\rho = \rho_0$ cuando $\rho_0 \neq 0$ no es equivalente a cualquier procedimiento de análisis de regresión. El estadístico de prueba se basa en una transformación de R desarrollada por el estadístico famoso R. A. Fisher.

PROPOSICIÓN

Cuando $(X_1, Y_1),\ldots, (X_n, Y_n)$ es la muestra de una distribución normal bivariada, la variable aleatoria

$$V = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \tag{9.9}$$

tiene aproximadamente una distribución normal con media y varianza

$$\mu_V = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \qquad \sigma_V^2 = \frac{1}{n-3}$$

El razonamiento para la transformación es obtener una función de R que tenga una varianza independiente de ρ; este no sería el caso con R misma. Además, no se deberá utilizar la transformación si n es muy pequeña, puesto que la aproximación no será válida.

El estadístico de prueba para demostrar H_0: $\rho = \rho_0$ es

$$Z = \frac{V - \frac{1}{2}\ln[(1 + \rho_0)/(1 - \rho_0)]}{1/\sqrt{n - 3}}$$

Hipótesis alternativa	Determinación del valor *P*
H_a: $\rho > \rho_0$	Área bajo la curva normal estándar a la derecha de z
H_a: $\rho < \rho_0$	Área bajo la curva normal estándar a la izquierda de z
H_a: $\rho \neq \rho_0$	$2 \cdot$ (Área bajo la curva normal estándar a la derecha de $\lvert z \rvert$)

EJEMPLO 9.18 El artículo **"Size Effect in Shear Strength of Large Beams–Behavior and Finite Element Modelling" (*Mag. of Concrete Res.*, 2005: 497–509)** reporta sobre un estudio de las diversas características de gran reforzado de vigas de hormigón profundas y superficiales probadas hasta la falla. Considere los siguientes datos sobre x = resistencia de cubo y y = resistencia de cilindro (ambas en MPa):

x	55.10	44.83	46.32	51.10	49.89	45.20	48.18	46.70	54.31	41.50
y	49.10	31.20	32.80	42.60	42.50	32.70	36.21	40.40	37.42	30.80
x	47.50	52.00	52.25	50.86	51.66	54.77	57.06	57.84	55.22	
y	35.34	44.80	41.75	39.35	44.07	43.40	45.30	39.08	41.89	

Por tanto, $S_{xx} = 367.74$, $S_{yy} = 488.54$ y $S_{xy} = 322.37$ a partir de lo cual $r = 0.761$. ¿Será que este valor proporciona una fuerte evidencia para concluir que las dos medidas de resistencia están al menos moderada y positivamente correlacionadas?

La interpretación previa de correlación positiva moderada fue $0.5 < \rho < 0.8$, así que se desea probar H_0: $\rho = 0.5$ contra H_a: $\rho > 0.5$. El valor calculado de V es entonces

$$v = 0.5 \ln\left(\frac{1 + 0.761}{1 - 0.761}\right) = 0.999 \qquad 0.5 \ln\left(\frac{1 + 0.5}{1 - 0.5}\right) = 0.549$$

Por consiguiente, $z = (0.999 - 0.549)\sqrt{19 - 3} = 0.180$. El valor P para una prueba de cola superior es $1 - \Phi(1.80)\sqrt{19 - 3} = 0.0359$. La hipótesis nula, por consiguiente, puede ser rechazada a un nivel de significancia de 0.05, pero no al nivel de 0.01. El último resultado es algo más sorprendente a la luz de la magnitud de r, pero cuando n es pequeño, puede resultar una r razonablemente grande aun cuando ρ no sea del todo sustancial. A nivel de significancia de 0.01, la evidencia de una correlación moderadamente positiva no es convincente. ■

Para obtener un intervalo de confianza para ρ primero se deduce un intervalo para $\mu_V = \frac{1}{2}\ln[(1 + \rho)/(1 - \rho)]$. Al estandarizar V, escribir una proposición de probabilidad y manipular las desigualdades resultantes se obtiene

$$\left(v - \frac{z_{\alpha/2}}{\sqrt{n - 3}}, v + \frac{z_{\alpha/2}}{\sqrt{n - 3}}\right) \tag{9.10}$$

como un intervalo de $100(1 - \alpha)\%$ para μ_v, donde $v = \frac{1}{2}\ln[(1 + r)/(1 - r)]$. Este intervalo puede entonces ser manipulado para dar el intervalo de confianza deseado.

Un intervalo de confianza de $100(1 - \alpha)\%$ para ρ es

$$\left(\frac{e^{2c_1} - 1}{e^{2c_1} + 1}, \quad \frac{e^{2c_2} - 1}{e^{2c_2} + 1}\right)$$

donde c_1 y c_2 son los puntos extremos izquierdo y derecho, respectivamente, del intervalo (12.11).

EJEMPLO 9.19 Desde la época de Leonardo da Vinci se sabe que x = altura y y = envergadura (medida de dedo a dedo mientras los brazos están extendidos a los lados) están estrechamente relacionadas. Aquí se presentan las mediciones de una muestra aleatoria de estudiantes que están asistiendo a un curso de estadística:

x	63.0	63.0	65.0	64.0	68.0	69.0	71.0	68.0
y	62.0	62.0	64.0	64.5	67.0	69.0	70.0	72.0
x	68.0	72.0	73.0	73.5	70.0	70.0	72.0	74.0
y	70.0	72.0	73.0	75.0	71.0	70.0	76.0	76.5

Un diagrama de dispersión muestra un patrón lineal aproximado, y así se forman diagramas de probabilidad normal de x y y. El coeficiente de correlación de la muestra se calcula como $r = 0.9422$. Su transformación de Fisher es

$$v = 0.5\ln\left(\frac{1 + 0.9422}{1 - 0.9422}\right) = 1.757$$

Un intervalo de confianza de 95% para μ_v es

$$1.757 \pm \frac{1.96}{\sqrt{16 - 3}} = (1.213, 2.301) = (c_1, c_2)$$

El intervalo de confianza para ρ con un nivel de confianza de aproximadamente 95% es, por tanto,

$$\left(\frac{e^{2(1.213)} - 1}{e^{2(1.213)} + 1}, \frac{e^{2(2.301)} - 1}{e^{2(2.301)} + 1}\right) = (0.838, 0.980)$$

Observe que el intervalo incluye sólo valores superiores a 0.8, así que al parecer hay una asociación lineal fuerte entre ambas variables en la población muestreada.

En el capítulo 5 se advirtió que un valor grande del coeficiente de correlación (cercano a 1 o -1) implica sólo asociación y no causalidad. Esto es válido tanto para ρ como para r.

EJERCICIOS **Sección 9.5 (57–67)**

57. El artículo **"Behavioral Effects of Mobile Telephone Use During Simulated Driving" (*Ergonomics*, 1995: 2536–2562)** reporta que para una muestra de 20 sujetos experimentales el coeficiente de correlación muestral con x = edad y y = tiempo desde que el sujeto obtuvo una licencia de manejo (años) fue 0.97. ¿Por qué piensa que el valor de r se aproxima tanto a 1? (Los autores del artículo dan una explicación.)

58. El Turbine Oil Oxidation Test (TOST) y el Rotating Bomb Oxidation Test (RBOT) son dos procedimientos diferentes de evaluar la estabilidad ante la oxidación de aceites para turbina de vapor. El artículo **"Dependence of Oxidation Stability of Steam Turbine Oil on Base Oil Composition" (*J. of the Society of Tribologists and Lubrication Engrs*., octubre de 1997: 19–24)** reporta las siguientes observaciones de 12 especímenes

de aceite. sobre x = tiempo para realizar TOST (h) y y = tiempo para realizar RBOT (min) con

TOST	4200	3600	3750	3675	4050	2770
RBOT	370	340	375	310	350	200
TOST	4870	4500	3450	2700	3750	3300
RBOT	400	375	285	225	345	285

a. Calcule e interprete el valor del coeficiente de correlación muestral (como lo hicieron los autores del artículo).

b. ¿Cómo se vería afectado el valor de r si se hubiera hecho x = tiempo para realizar RBOT y y = tiempo para realizar TOST?

c. ¿Cómo se vería afectado el valor de r si el tiempo para realizar RBOT estuviera expresado en horas?

d. Construya gráficas de probabilidad normal y comente.

e. Realice una prueba de hipótesis para decidir si el tiempo para realizar RBOT y el tiempo para realizar TOST están linealmente relacionados.

59. La firmeza y la fibrosidad de los espárragos son importantes para determinar su calidad. Este fue el enfoque de un estudio reportado en **"Post-Harvest Glyphosphate Application Reduces Toughening, Fiber Content, and Lignification of Stores Asparagus Spears" (*J. of the Amer. Soc. of Hort. Science*, 1988: 569–572)**. El artículo reporta los siguientes datos (tomados de una gráfica) sobre x = fuerza cortante (kg) y y = porcentaje de peso de la fibra en seco.

x	46	48	55	57	60	72	81	85	94
y	2.18	2.10	2.13	2.28	2.34	2.53	2.28	2.62	2.63
x	109	121	132	137	148	149	184	185	187
y	2.50	2.66	2.79	2.80	3.01	2.98	3.34	3.49	3.26

$n = 18$, $\Sigma x_i = 1950$, $\Sigma x_i^2 = 251\,970$,
$\Sigma y_i = 47.92$, $\Sigma y_i^2 = 130.6074$, $\Sigma x_i y_i = 5530.92$

a. Calcule el valor del coeficiente de correlación muestral. Basado en este valor, ¿cómo describiría la naturaleza de la relación entre las dos variables?

b. Si un primer espécimen tiene un valor más grande de fuerza cortante que un segundo espécimen, ¿qué tiende a ser cierto del porcentaje de peso de fibra en seco para los dos especímenes?

c. Si la fuerza cortante se expresa en libras, ¿qué le pasa al valor de r? ¿Por qué?

d. Si el modelo de regresión lineal simple fuera ajustado a estos datos, ¿qué proporción de la variación observada en porcentaje de peso de la fibra en seco podría ser explicada por la relación de modelo?

e. Realice una prueba a un nivel de significancia 0.01 para decidir si existe una asociación lineal positiva entre las dos variables.

60. Las evaluaciones del movimiento de cabeza son importantes porque los individuos, especialmente los discapacitados, pueden ser capaces de operar los dispositivos de ayuda de esta manera. El artículo **"Constancy of Head Turning Recorded in Healthy Young Humans" (*J. of Biomed. Engr.*, 2008: 428–436)** reporta datos sobre los rangos en los ángulos de inclinación máxima de la cabeza en el sentido de las manecillas del reloj para la parte anterior, posterior, derecha e izquierda en 14 sujetos seleccionados al azar. Considere los siguientes datos para el ángulo promedio de inclinación máxima anterior (AMIA), tanto en la dirección de las manecillas del reloj (SH) como en sentido contrario (SA).

Sujeto:	1	2	3	4	5	6	7
SH:	57.9	35.7	54.5	56.8	51.1	70.8	77.3
SA:	44.2	52.1	60.2	52.7	47.2	65.6	71.4
Sujeto:	8	9	10	11	12	13	14
SH:	51.6	54.7	63.6	59.2	59.2	55.8	38.5
SA:	48.8	53.1	66.3	59.8	47.5	64.5	34.5

a. Calcule una estimación puntual del coeficiente de correlación de población entre el SH y el AMIA y entre el AMIA y el SA ($\Sigma SH = 786.7$, $\Sigma SA = 767.9$, $\Sigma SH^2 = 45\,727.31$, $\Sigma SA^2 = 43\,478.07$, $\Sigma SHSA = 44\,187.87$).

b. Suponiendo normalidad bivariada (las gráficas de probabilidad normal del SH y las muestras del SA son razonablemente rectas) lleve a cabo una prueba al nivel de significancia 0.01 para decidir si existe una relación lineal entre las dos variables en la población (al igual que los autores del citado artículo). ¿La conclusión sería la misma si se utiliza un nivel de significancia de 0.001?

61. Los autores del artículo **"Objective Effects of a Six Months' Endurance and Strength Training Program in Outpatients with Congestive Heart Failure" (*Medicine and Science in Sports and Exercise*, 1999: 1102-1107)** presentan un análisis de correlación para investigar la relación entre el nivel de lactato máximo x y la resistencia muscular y. Los siguientes datos fueron tomados de una gráfica en el artículo.

x	400	750	770	800	850	1025	1200
y	3.80	4.00	4.90	5.20	4.00	3.50	6.30
x	1250	1300	1400	1475	1480	1505	2200
y	6.88	7.55	4.95	7.80	4.45	6.60	8.90

$S_{xx} = 36.9839$, $S_{yy} = 2\,628\,930.357$, $S_{xy} = 7377.704$. Un diagrama de dispersión muestra un patrón lineal.

a. Realice una prueba para ver si existe una correlación positiva entre el nivel de lactato máximo y la resistencia muscular en la población de la cual se seleccionaron estos datos.

b. Si se tuviera que realizar un análisis de regresión para predecir la resistencia a consecuencia del nivel de lactato, ¿qué proporción de variación observada en la resistencia podría ser atribuida a la relación lineal aproximada? Responda la pregunta análoga si se utilizara regresión para predecir el nivel de lactato a partir de la resistencia; responda ambas preguntas sin realizar ningún cálculo de regresión.

62. El artículo **"Quantitative Estimation of Clay Mineralogy in Fine-Grained Soils" (*J. of Geotechnical and Geoenvironmental Engr.*, 2011: 997–1008)** informa sobre diferentes propiedades químicas de los suelos naturales y artificiales. Aquí se presentan las observaciones de x = capacidad de intercambio catiónico (CEC, en meq/100 g) y y = área superficial específica (SSA, en m^2/g) de 20 suelos naturales.

x	66	121	134	101	77	89	63	57	117	118
y	175	324	460	288	205	210	295	161	314	265
x	76	125	75	71	133	104	76	96	58	109
y	236	355	240	133	431	306	132	269	158	303

Minitab da la siguiente salida en respuesta a una solicitud de r:

```
correlation of x and y = 0.853
```

Los diagramas de probabilidad normal de x y y son muy rectos.

a. Realice una prueba de hipótesis para ver si hay una asociación lineal positiva en la población de los datos de la muestra que fueron seleccionados.

b. Con $n = 20$, ¿qué tan pequeño tendría que ser el valor de r para que la hipótesis nula de la prueba de a) no sea rechazada en el nivel de significancia 0.01?

c. Calcule un intervalo de confianza para ρ utilizando un nivel de confianza de 95%.

63. En el artículo **"Sensory and Physical Properties of Inherently Flame-Retardant Fabrics" (*Textile Research*, 1984: 61–68)** se investigaron las propiedades físicas de seis muestras de tela resistente a las llamas. Use los siguientes datos y un nivel de significancia de 0.05 para determinar si existe una relación lineal entre la rigidez x (mg-cm) y el espesor y (mm). ¿Es sorprendente el resultado de la prueba a la luz del valor de r?

x	7.98	24.52	12.47	6.92	24.11	35.71
y	0.28	0.65	0.32	0.27	0.81	0.57

64. Los siguientes datos sobre x = índice de transparencia ultravioleta y y = máxima prevalencia de infección se tomaron de una gráfica en el artículo **"Solar Radiation Decreases Parasitism in Daphnia" (*Ecology Letters*, 2012: 47–54)**:

x	1.3	1.4	1.5	2.0	2.2	2.7	2.7	2.7	2.8
y	16	3	32	1	13	0	8	16	2
x	2.9	3.0	3.6	3.8	3.8	4.6	5.1	5.7	
y	1	7	36	25	10	35	58	56	

Las cantidades resumidas incluyen $S_{xx} = 25.5224$, $S_{yy} = 5593.0588$ y $S_{xy} = 264.4882$.

a. Calcule e interprete el valor del coeficiente de correlación de la muestra.

b. Si decide ajustar el modelo de regresión lineal simple a estos datos, ¿qué proporción de la variación observada en la prevalencia máxima podría explicarse mediante la relación del modelo?

c. Si decide hacer la regresión del índice de transparencia UV sobre la máxima prevalencia (es decir, intercambiar los papeles de x y y), ¿qué proporción de la variación observada puede atribuirse a la relación del modelo?

d. Realice una prueba de H_0: $\rho = 0.5$ contra H_a: $\rho > 0.5$ utilizando un nivel de significancia de 0.05. [*Nota*: El artículo citado informó el valor P para probar H_0: $\rho = 0$ contra H_0: $\rho \neq 0$.]

65. La torsión durante la rotación externa de la cadera y la extensión pueden explicar por qué ocurren las lágrimas labral acetabular en atletas profesionales. El artículo **"Hip Rotational Velocities During the Full Golf Swing" (*J. of Sports Science and Med.*, 2009: 296–299)** informa sobre una investigación en la que el pico de la velocidad máxima de rotación interna de la cadera (x) y el pico final de la velocidad de rotación externa de la cadera (y) se han determinado en una muestra de 15 jugadores de golf. Los datos proporcionados por los autores del artículo fueron utilizados para calcular las siguientes cantidades resumidas:

$$\sum(x_i - \bar{x})^2 = 64\,732.83, \sum(y_i - \bar{y})^2 = 130\,566.96,$$
$$\sum(x_i - \bar{x})(y_i - \bar{y}) = 44\,185.87$$

Gráficas separadas de probabilidad normal mostraron patrones lineales muy importantes.

a. Calcule una estimación puntual del coeficiente de correlación de la población.

b. Efectúe una prueba en el nivel de significancia 0.01 para decidir si existe una relación lineal entre las dos velocidades en la población muestreada.

c. ¿La conclusión de b) habría cambiado si se hubiera probado la hipótesis adecuada para decidir si existe una relación lineal positiva entre la población? ¿Qué pasará si se utiliza un nivel de significancia 0.05 en lugar 0.01?

66. Considere una serie de tiempo; es decir, una secuencia de observaciones X_1, X_2,... obtenidas durante el transcurso del tiempo con valores observados x_1, x_2,..., x_n. Suponga que la serie no muestra tendencia hacia arriba ni hacia abajo durante el transcurso del tiempo. Usualmente un investigador deseará saber qué tan fuertemente están relacionados los valores en la serie separados por un número especificado de unidades de tiempo. El *coeficiente de autocorrelación muestral correspondiente a un retardo* r_1 es simplemente el valor del coeficiente de correlación muestral r de los pares (x_1, x_2), (x_2, x_3),..., (x_{n-1}, x_n), es decir, pares de valores separados por una unidad de tiempo. Asimismo, el *coeficiente de autocorrelación muestral correspondiente a dos retardos* r_2 es r para los $n - 2$ pares (x_1, x_3), (x_2, x_4),..., (x_{n-2}, x_n).

a. Calcule los valores de r_1, r_2 y r_3 para los datos de temperatura del ejercicio 82 del capítulo 1 y comente.

b. Análogo al coeficiente de correlación de la población ρ, sean ρ_1, ρ_2,... los coeficientes de autocorrelación teóricos o de largo plazo con los varios retardos. Si todos estos ρ son 0 no existe relación (lineal) con cualquier retraso. En este caso, si n es grande cada R_i tiene aproximadamente una distribución normal con media 0 y desviación estándar $1/\sqrt{n}$ y los diferentes R_i son casi independientes. Por consiguiente H_0:

$\rho_i = 0$ puede probarse mediante una prueba z con el valor estadístico de prueba $z_i = \sqrt{n}r_i$. Si $n = 100$ y $r_1 = 0.16$, $r_2 = -0.09$ y $r_3 = -0.15$, a un nivel de significancia de aproximadamente 0.05, ¿existe alguna evidencia de autocorrelación teórica con los primeros tres retrasos?

c. Si se prueba simultáneamente la hipótesis nula del inciso b) con más de un retraso para más de un retraso, ¿por qué se desearía aumentar el nivel de significancia para cada prueba?

67. Se recopiló una muestra de $n = 500$ pares (x, y) y se realizó una prueba de H_0: $\rho = 0$ contra H_a: $\rho \neq 0$. El valor P resultante se calculó como 0.00032.

a. ¿Qué conclusión sería apropiada al nivel de significancia 0.001?

b. ¿Indicará este pequeño valor P que existe una muy fuerte relación entre x y y (un valor de ρ que difiera considerablemente de 0)? Explique.

c. Suponga ahora que una muestra de $n = 10\,000$ pares (x, y) dio por resultado $r = 0.022$. Pruebe H_0: $\rho = 0$ contra H_a: $\rho \neq 0$ a un nivel 0.05. ¿Será el resultado estadísticamente significativo? Comente sobre el significado práctico de su análisis.

EJERCICIOS SUPLEMENTARIOS (68–75)

68. El avalúo de un almacén puede parecer sencillo en comparación con otros avalúos asignados. El avalúo de un almacén implica comparar una edificación que principalmente es una armadura abierta con otros edificios semejantes. Sin embargo, siguen habiendo varios atributos de un almacén que están factiblemente relacionados con el valor apreciado. El artículo **"Challenges In Appraising 'Simple' Warehouse Properties" (Donald Sonneman, *The Appraisal Journal*, abril de 2001, 174–178)** proporciona los siguientes datos sobre la altura de la armadura (pies), la cual determina qué tan alto pueden ser apilados los productos almacenados y el precio de venta ($) por pie cuadrado.

Altura	12	14	14	15	15	16	18	22	22	24
Precio	35.53	37.82	36.90	40.00	38.00	37.50	41.00	48.50	47.00	47.50
Altura	24	26	26	27	28	30	30	33	36	
Precio	46.20	50.35	49.13	48.07	50.90	54.78	54.32	57.17	57.45	

a. ¿Será el caso de que la altura de la armadura y el precio de venta están "determinísticamente" relacionados, es decir, que el precio de venta está determinado por completo y únicamente por la altura de la armadura? [*Sugerencia*: Examine los datos.]

b. Construya una gráfica de dispersión de los datos. ¿Qué sugieren?

c. Determine la ecuación de la recta de mínimos cuadrados.

d. Proporcione una predicción puntual del precio cuando la altura de la armadura es de 27 pies y calcule el residuo correspondiente.

e. ¿Qué porcentaje de la variación observada del precio de venta puede ser atribuido a la relación lineal aproximada entre la altura de la armadura y el precio?

69. Remítase al ejercicio previo, el cual dio datos sobre alturas de armaduras para una muestra de almacenes y los precios de venta correspondientes.

a. Estime el cambio promedio verdadero del precio de venta asociado con un pie de incremento en la altura de la armadura y hágalo de modo que dé información sobre la precisión de la estimación.

b. Estime el precio de venta promedio real de todos los almacenes cuya armadura tiene una altura de 25 pies y hágalo de modo que dé información sobre la precisión de la estimación.

c. Pronostique el precio de venta de un solo almacén cuya armadura tiene una altura de 25 pies y hágalo de modo que dé información sobre la precisión de la predicción. ¿Cómo se compara esta estimación con la estimación de b)?

d. Sin calcular ningún intervalo, ¿cómo se compararía el ancho de un intervalo de predicción de 95% con el precio de venta cuando la altura de la armadura es de 25 pies con el ancho de un intervalo de 95% cuando la altura es de 30 pies? Explique su razonamiento.

e. Calcule e interprete el coeficiente de correlación muestral.

70. A los científicos forenses usualmente les interesa realizar alguna clase de medición en un cuerpo (vivo o muerto) y luego utilizarla como base para inferir algo sobre la edad del cuerpo. Considere los datos adjuntos sobre edad (años) y % de ácido aspértico D (de aquí en adelante %DAA) de una pieza dental particular **("An Improved Method for Age at Death Determination from the Measurements of D-Aspertic Acid in Dental Collagen", *Archaeometry*, 1990: 61–70).**

Edad	9	10	11	12	13	14	33	39	52	65	69
%DAA	1.13	1.10	1.11	1.10	1.24	1.31	2.25	2.54	2.93	3.40	4.55

Suponga que una pieza dental de otro individuo tiene 2.01%DAA. ¿Podría ser que el individuo tuviera menos de 22 años? Esta pregunta era pertinente en cuanto a si el individuo podía ser sentenciado a cadena perpetua por homicidio o no.

Una estrategia aparentemente razonable es retroceder la edad en %DDA, y luego calcule un intervalo de probabilidad para la edad cuando %DAA = 2.01. No obstante, es más natural en este caso considerar la edad como la variable independiente x y el %DAA como la variable dependiente y, así que el modelo de regresión es $\%\text{DAA} = \beta_0 + \beta_1 x + \epsilon$. Después de estimar los coeficientes de regresión se puede sustituir $y^* = 2.01$ en la ecuación estimada y luego resolverla para una predicción

de edad $\hat{x}$. Este uso "inverso" de la recta de regresión se llama "calibración". Un intervalo de predicción para edad con nivel de predicción aproximadamente de $100(1-\alpha)\%$ es $\hat{x} \pm t_{\alpha/2,n-2} \cdot SE$ donde

$$SE = \frac{s}{\hat{\beta}_1}\left\{1 + \frac{1}{n} + \frac{(\hat{x}-\bar{x})^2}{S_{xx}}\right\}^{1/2}$$

Calcule este intervalo de predicción para $y^* = 2.01$ y luego aborde la pregunta previamente planteada.

71. Los compuestos fenólicos se encuentran en las descargas residuales de procesos de conversión de carbón, refinerías de petróleo, fabricación de herbicidas y fabricación de fibra de vidrio. Estos compuestos son tóxicos, cancerígenos y han contribuido en las últimas décadas a la contaminación ambiental de los ambientes acuáticos. En un estudio que se reporta en **"Photolysis, Biodegradation, and Sorption Behavior of Three Selected Phenolic Compounds on the Surface and Sediment of Rivers" (*J. of Envir. Engr.*, 2011: 1114–1121)** los autores examinaron las características de absorción de tres compuestos fenólicos. Los siguientes datos sobre la y = concentración absorbida (μg/g) y x = concentración de equilibrio (μg/mL) de 2, 4-dinitrofenol (DNP) en un sedimento de río natural particular fueron obtenidos de una gráfica del artículo.

x	0.11	0.13	0.14	0.18	0.29	0.44	0.67	0.78	0.93
y	1.72	2.17	2.33	3.00	5.17	7.61	11.47	12.72	14.78

a. Calcule los puntos estimados de la pendiente y la intersección de la recta de regresión de población.

b. ¿El modelo de regresión lineal simple especifica una relación útil entre y y x?

c. Confirme que $\hat{y} = 3.404$, $S_{\hat{Y}} = 0.107$ cuando $x = 0.2$, y $\hat{y} = 0.6.616$, $S_{\hat{Y}} = 0.088$ cuando $x = 0.4$. Explique por qué $s_{\hat{Y}}$ es mayor cuando $x = 0.2$ que cuando $x = 0.4$.

d. Calcule un intervalo de confianza con un nivel de confianza de 95% para el promedio verdadero de la concentración absorbida de DNP de todas las muestras de sedimento de río que tienen una concentración de equilibrio de 0.4.

e. Calcule un intervalo de predicción con un nivel de predicción de 95% para que la concentración absorbida de DNP en una muestra de sedimento de río sólo tenga una concentración de equilibrio de 0.4.

72. Los resultados SAS dados al final de la página están basados en datos tomados del artículo **"Evidence for and the Rate of Denitrification in the Arabian Sea" (*Deep Sea Research*, 1978: 431–435)**. Las variables estudiadas son x = nivel de salinidad (%) y y = nivel de nitrato (μM/L).

a. ¿Cuál es el tamaño de muestra n? [*Sugerencia*: Busque los grados de libertad para SSE.]

b. Calcule una estimación puntual del nivel de nitrato esperado cuando el nivel de salinidad es de 35.5.

c. ¿Habrá una relación lineal útil entre las dos variables?

d. ¿Cuál es el valor del coeficiente de correlación muestral?

e. ¿Utilizaría el modelo de regresión lineal simple para sacar conclusiones cuando el nivel de salinidad es de 40?

73. La presencia de carburos de aleación duros en aleaciones de hierro blanco al alto cromo produce una excelente resistencia a la abrasión, lo que las hace apropiadas para el manejo de materiales en la industria minera y del procesamiento de materiales. Los datos adjuntos sobre x = contenido de austenita retenido (%) y y = pérdida por desgaste abrasivo (mm^3) en prueba de desgaste de alfileres con granate como el abrasivo fueron tomadas de una gráfica que aparece en el artículo **"Microstructure-Property Relationships in High Chromium White Iron Alloys" (*Intl. Materials Reviews*, 1996: 59-82)**.

x	4.6	17.0	17.4	18.0	18.5	22.4	26.5	30.0	34.0
y	0.66	0.92	1.45	1.03	0.70	0.73	1.20	0.80	0.91
x	38.8	48.2	63.5	65.8	73.9	77.2	79.8	84.0	
y	1.19	1.15	1.12	1.37	1.45	1.50	1.36	1.29	

Resultados obtenidos con SAS para el ejercicio 72

```
Dependent Variable: NIVELNITRATO

                              Analysis of Variance

 Source          DF         Sum of Squares         Mean Square          F Value         Prob > F

Model             1              64.49622             64.49622           63.309           0.0002
Error             6               6.11253              1.01875
C Total           7              70.60875

               Root MSE          1.00933      R-square      0.9134
               Dep Mean         26.91250      Adj R-sq      0.8990
               C.V.              3.75043

                              Parameter Estimates

                          Parameter            Standard          T for HO:
Variable        DF         Estimate               Error       Parameter = 0       Prob > |T|
INTERCEPCION     1       326.976038         37.71380243               8.670           0.0001
SALINIDAD        1        -8.403964          1.05621381              -7.957           0.0002
```

Resultados obtenidos con SAS para el ejercicio 73

```
Dependent Variable: PERDIDADABRASIVA
                                   Analysis of Variance

 Source              DF         Sum of Squares          Mean Square          F Value          Prob > F
  Model               1                0.63690              0.63690           15.444           0.0013
  Error              15                0.61860              0.04124
C Total              16                1.25551

                  Root MSE       0.20308       R-square      0.5073
                  Dep Mean       1.10765       Adj R-sq      0.4744
                  C.V.          18.33410

                                   Parameter Estimates

                            Parameter           Standard           T for H0:
Variable             DF      Estimate              Error       Parameter = 0          Prob > |T|
INTERCEPCION          1      0.787218         0.09525879               8.264              0.0001
CONTENIDOAUSTENITA    1      0.007570         0.00192626               3.930              0.0013
```

Use los datos y los resultados obtenidos con SAS dados al inicio de esta página para responder las siguientes preguntas.

a. ¿Qué proporción de la variación observada de pérdida por desgaste puede atribuirse a la relación de modelo de regresión lineal simple?

b. ¿Cuál es el valor del coeficiente de correlación muestral?

c. Pruebe la utilidad del modelo de regresión lineal simple con $\alpha = 0.01$.

d. Estime la pérdida por desgaste promedio verdadera cuando el contenido es de 50% y hágalo de modo que dé información sobre confiabilidad y precisión.

e. ¿Qué valor de pérdida por desgaste pronosticaría cuando el contenido es de 30% y cuál es valor del residuo correspondiente?

74. Los siguientes datos se tomaron de una gráfica de dispersión del artículo **"Urban Emissions Measured with Aircraft" (*J. of the Air and Waste Mgmt. Assoc.*, 1998: 16–25)**. La variable de respuesta es ΔNO_y la variable explicativa es ΔCO.

ΔCO	50	60	95	108	135
ΔCO_y	2.3	4.5	4.0	3.7	8.2
ΔCO	210	214	315	720	
ΔNO_y	5.4	7.2	13.8	32.1	

a. Adapte un modelo apropiado a los datos y juzgue la utilidad del modelo.

b. Pronostique el valor de ΔNO_y que se obtendría al realizar una observación más cuando ΔCO es 400 y hágalo de modo que dé información sobre precisión y confiabilidad. ¿Será que el ΔNO_y puede ser pronosticado con precisión? Explique.

c. El valor más grande de ΔCO es mucho más grande que los demás valores. ¿Ha tenido esta observación un impacto sustancial en la ecuación ajustada?

75. Se estudió la relación entre la velocidad (pies/s) y la cadencia al correr (número de pasos/s) entre maratonistas. Las cantidades resumidas resultantes fueron $n = 11$, Σ(velocidad) = 205.4, $\Sigma(\text{velocidad})^2 = 3880.08$, Σ(cadencia) = 35.16, $\Sigma(\text{cadencia})^2 = 112.681$ y Σ(velocidad)(cadencia) = 660.130.

a. Calcule la ecuación de la recta de mínimos cuadrados que utilizaría para predecir la cadencia a partir de la velocidad.

b. Calcule la ecuación de la recta de mínimos cuadrados que utilizaría para predecir la velocidad a partir de la cadencia.

c. Calcule el coeficiente de determinación para la regresión de la cadencia basada en la velocidad del inciso a) y para la regresión de la velocidad basada en la cadencia del inciso b). ¿Cómo están relacionadas?

BIBLIOGRAFÍA

Draper, Norman y Harry Smith, *Applied Regression Analysis* (3a. ed.), Wiley, Nueva York, 1999. El libro más completo y autorizado sobre análisis de regresión actualmente en proceso de impresión.

Neter, John, *et al*., *Applied Linear Statistical Models* (5a. ed.), Irwin, Homewood, IL., 2005. Los primeros 14 capítulos constituyen un estudio extremadamente informativo y fácil de leer acerca del análisis de regresión.

Fundamentos de probabilidad y estadística cubre de manera específica el plan de estudios del curso a nivel superior.

La presente obra está integrada por el siguiente contenido:

- Generalidades y estadística descriptiva
- Probabilidad
- Variables aleatorias discretas y distribuciones de probabilidad
- Variables aleatorias continuas y distribuciones de probabilidad
- Estimación puntual
- Intervalos estadísticos basados en una sola muestra
- Pruebas de hipótesis basadas en una sola muestra
- Análisis de la varianza
- Regresión lineal simple y correlación

Estos temas establecen una manera singular de abordar la probabilidad y estadística, ya que introducen al estudiante con amplitud a modelos y métodos mayormente utilizados y de este modo pueden conectarse con la materia mediante ejemplos y ejercicios que combinan experiencias diarias con sus intereses científicos.

Acompáñanos a conocer la ***Probabilidad y estadística*** desde una perspectiva clara y eficaz.

CENGAGE

Visite nuestro sitio en http://latinoamerica.cengage.com

ISBN-13: 978-607-526-663-3
ISBN-10: 607-526-663-1

9 786075 266633