CENGAGE

INTRODUCCIÓN A LA PROBABILIDAD Y ESTADÍSTICA

PARA INGENIERÍA Y CIENCIAS

JAY L. DEVORE

Capítulos digitales para estudiantes y docentes*

13 Regresión múltiple y no lineal

14 Pruebas de bondad de ajuste y análisis de datos categóricos

15 Procedimientos libres de distribución

16 Métodos de control de calidad

Respuestas impares a ejercicios seleccionados de los capítulos 12 a 16

Para tener acceso a estos capítulos:

1. Ingresa a **latinoamerica.cengage.com** y busca tu libro de texto por título.
2. Dirígete a los materiales de apoyo de estudiante o del profesor según corresponda.
3. Sigue las indicaciones del sitio para descargar los complementos digitales.

* Material disponible sin costo adicional.

INTRODUCCIÓN A LA PROBABILIDAD Y ESTADÍSTICA

PARA INGENIERÍA Y CIENCIAS

INTRODUCCIÓN A LA PROBABILIDAD Y ESTADÍSTICA

PARA INGENIERÍA Y CIENCIAS

JAY L. DEVORE

California Polytechnic State University, San Luis Obispo

Traducción

Javier León Cárdenas
Formación básica ESIQIE/IPN

Jesús Miguel Torres Flores
ENCB/IPN

Revisión técnica

Universidad Nacional Autónoma de México
Á. Leonardo Bañuelos Saucedo

TEcNM/Instituto Tecnológico de Chihuahua
Leticia del Pilar De la Torre González

Universidad Autónoma de Ciudad Juárez, Chihuahua
Víctor Manuel Carrillo Saucedo
Juan Ernesto Chávez Pierce

Universidad Autónoma de Querétaro
Pablo Talamantes Contreras

CENGAGE

Australia • Brasil • Estados Unidos • México • Reino Unido • Singapur

CENGAGE

***Introducción a la probabilidad y estadística para ingeniería y ciencias*, primera edición**
Jay L. Devore

Director Higher Education Latinoamérica:
Renzo Casapía Valencia

Gerente editorial Latinoamérica:
Jesús Mares Chacón

Editoras:
Karen Estrada Arriaga
y Abril Vega Orozco

Coordinador de manufactura:
Rafael Pérez González

Diseño de portada original:
C Miller Design

Adaptación de portada:
María Eugenia Hernández Granados

Composición tipográfica:
Arturo Rocha Hernández

Traducido del libro:
Probability and Statistics for Engineering and the Sciences
Ninth Edition
Jay Devore
© 2016
ISBN: 978-1-305-25180-9

Datos para catalogación bibliográfica:
Devore, Jay L.
Introducción a la probabilidad y estadística para ingeniería y ciencias
Primera edición
ISBN: 978-607-526-794-4

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

Impreso en México
1 2 3 4 5 6 7 22 21 20 19

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

Contenido

8 Pruebas de hipótesis basadas en una sola muestra

9 Inferencias basadas en dos muestras

10 Análisis de varianza

11 Análisis multifactorial de la varianza

Capítulos digitales disponibles en **latinoamerica.cengage.com**

Capítulos digitales disponibles en **latinoamerica.cengage.com**

16 Métodos de control de calidad

Apéndice de tablas

Prefacio

Propósito

El uso de modelos de probabilidad y métodos estadísticos para el análisis 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 ciencias naturales. Aun cuando los ejemplos y ejercicios se diseñaron pensando en los científicos y en 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 pueden, en principio, dudar 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 buscado intensamente 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 con base en el análisis de 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 he percatado de que los problemas reales con un contexto un tanto desconocido resultan más atractivos para los estudiantes que aquellos definitivamente artificiales planteados en un entorno conocido.

Nivel matemático

La presentación del contenido de esta obra es relativamente modesta en función del desarrollo matemático. Solo en el capítulo 4 y en fragmentos de los capítulos 5 y 6, se utiliza el cálculo de manera sustancial. En particular, con excepción de una observación o nota ocasional, el cálculo se presenta en la parte de inferencia del libro, es decir, en la segunda sección del capítulo 6. No se utiliza álgebra matricial. Por tanto, la mayor parte del contenido es accesible para quienes hayan cursado un semestre o dos trimestres de cálculo diferencial e integral.

Contenido

El capítulo 1 inicia con algunos conceptos y terminología básicos (población, muestra, estadística descriptiva e inferencial, estudios enumerativos contra analíticos y así sucesivamente) y continúa con el estudio de métodos descriptivos gráficos y numéricos importantes. En el capítulo 2 se presenta el tema de probabilidad desde una perspectiva un tanto tradicional, seguido por distribuciones de probabilidad de variables aleatorias continuas y discretas en los capítulos 3 y 4, respectivamente. Las distribuciones conjuntas y sus propiedades se analizan en la primera parte del capítulo 5. En la última parte de este capítulo se introduce la estadística y sus distribuciones muestrales, las cuales constituyen

el puente entre la probabilidad y la inferencia. En los siguientes tres capítulos se aborda la estimación puntual, los intervalos estadísticos y la comprobación de hipótesis basados en una muestra única. Los métodos de inferencia que implican dos muestras independientes y datos apareados se presentan en el capítulo 9. El análisis de la varianza es el tema de los capítulos 10 y 11 (unifactorial y multifactorial, respectivamente). Se aborda, por primera vez, el tema de regresión en el capítulo 12 (modelo de regresión lineal simple y correlación) y se retoma, de manera más amplia, en el capítulo 13. En los últimos tres capítulos se presentan métodos de ji-cuadrada, procedimientos sin distribución (no paramétricos) y técnicas de control estadístico de calidad.

Ayuda para el aprendizaje de los estudiantes

Aunque el nivel matemático del libro representará poca dificultad para la mayoría de los estudiantes de ciencias e ingeniería, 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 mejoren su comprensión y apreciación he proporcionado numerosos ejercicios de dificultad variable; algunos implican la aplicación rutinaria del material incluido en el texto y otros requieren que el lector aplique los conceptos analizados en situaciones un tanto nuevas. Existen muchos ejercicios más que la mayoría de los profesores desearía asignar durante cualquier curso, pero recomiendo que se les pida a los estudiantes que resuelvan un número sustancial de ejercicios; 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. Las respuestas a la mayoría de los ejercicios impares aparecen en la sección de respuestas al final del texto.

Para más información acerca de los recursos adicionales de esta obra, consulte a su representante local de Cengage.

Nuevo en esta edición

- El mayor cambio en esta edición es que se eliminó la sección de pruebas de hipótesis por rechazo de región. Las conclusiones de los análisis por pruebas de hipótesis se basan ahora únicamente en valores *P*. La sección 8.1 fue reescrita y ahora incluye las hipótesis y los procedimientos de prueba basados en valores *P*. Por tanto, se requirió una revisión sustancial de las secciones del capítulo 8 y de las secciones y subsecciones basadas en pruebas de hipótesis de los capítulos 9-15.
- Se incluyen muchos ejemplos nuevos y ejercicios, casi todos basados en datos reales o problemas reales. Algunos de estos escenarios son menos técnicos o de alcance más amplio que los presentados en las ediciones anteriores; por ejemplo, investigar el efecto nocebo (experimentación de los efectos secundarios de un fármaco provocada por predisposición) la comparación de los contenidos de sodio en cereales producidos por tres fabricantes distintos, la predicción de la altura a partir de una medición simple de una característica anatómica, el modelaje de la relación entre la edad de una madre adolescente y el peso de su bebé al nacer, la medición de la abstinencia a corto plazo de un fumador sobre la percepción correcta del tiempo transcurrido y la exploración del impacto del fraseo en una prueba literaria cuantitativa.
- Ejemplos y ejercicios adicionales acerca del material de probabilidad (capítulos 2-5) basados en información de fuentes publicadas.
- En la medida de lo posible, se ha mejorado la presentación de los temas con el objetivo de ayudar a los estudiantes a adquirir una comprensión intuitiva de los diferentes conceptos.

Reconocimientos

A lo largo de los años, he recibido apoyo y alimentación invaluables de mis colegas en Cal Poly a quienes les estoy agradecido por ello. También agradezco a los muchos usuarios de ediciones previas que me sugirieron mejoras (y en ocasiones identificaron erratas). 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 expreso 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. En esta edición agradezco de manera 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 valoro la labor estelar de todos los representantes de ventas de Cengage Learning que han promovido mis libros en 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 inspirarme a través de sus propios logros.

Jay L. Devore

Generalidades y estadística descriptiva

1

"Cursé estadística en la escuela de negocios y fue una experiencia transformadora. La formación analítica le brinda un conjunto de habilidades que le permiten distinguirse de la mayoría de las personas en el mercado laboral."

—Laszlo Bock, vicepresidente senior de People Operations (a cargo de todas las contrataciones) en Google.

The New York Times, entrevista con el columnista Thomas Friedman, 20 de abril de 2014.

"No soy muy dado a lamentar, pero esto me desconcertó un tiempo. Creo que debería haber estudiado mucho más estadística en la universidad."

—Max Levchin, cofundador de Paypal, fundador de Slide Founder.

Cita de la semana tomada del sitio web de la American Statistical Association, 23 de noviembre de 2010.

"Sigo diciendo que el trabajo sexy en los próximos 10 años será la estadística, y no estoy bromeando."

—Hal Varian, economista en jefe de Google.

The New York Times, 6 de agosto de 2009.

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.

Al determinar la forma "más verde" de viajar surgió una interesante muestra de la variación. El artículo **"Carbon Conundrum" (*Consumer Reports*, 2008: 9)** identifica organizaciones que ayudan a los consumidores a calcular la producción de carbono. Se registraron los siguientes resultados en el despegue de un vuelo de Nueva York a Los Ángeles:

Cálculo de carbono	CO_2 (lb)
Terra Pass	1924
Conservation International	3000
Cool It	3049
World Resources Institute/Safe Climate	3163
National Wildlife Federation	3465
Sustainable Travel International	3577
Native Energy	3960
Environmental Defense	4000
Carbonfund.org	4820
The Climate Trust/CarbonCounter.org	5860
Bonneville Environmental Foundation	6732

Claramente hay un importante desacuerdo entre estos cálculos respecto a la cantidad exacta de carbono emitido, caracterizado en el articulo como la diferencia que hay entre "una bailarina y Pie Grande". El artículo proporciona una dirección web donde los lectores podrán aprender más acerca del funcionamiento de estas calculadoras.

¿Cómo se pueden utilizar técnicas estadísticas para reunir información y sacar conclusiones? Suponga, por ejemplo, que un ingeniero de materiales ha inventado un recubrimiento para retardar la corrosión en tuberías de metal en circunstancias específicas. Si este recubrimiento se aplica a diferentes segmentos de la tubería, la variación de las condiciones ambientales y de los segmentos mismos producirá más corrosión en algunos segmentos que en otros. Para decidir si la cantidad *promedio* de corrosión excede un límite superior especificado de alguna clase, o cuánta corrosión ocurrirá en una sola pieza de tubería se puede utilizar un análisis estadístico con los datos del experimento.

Por otra parte, suponga que el ingeniero inventó el recubrimiento con la creencia de que será superior al recubrimiento que se utiliza actualmente. Se podría realizar un experimento comparativo para investigar esta cuestión aplicando a algunos segmentos de la tubería el recubrimiento actual y el nuevo a otros segmentos. Esto debe realizarse con cuidado o se obtendrá una conclusión errónea. Por ejemplo, tal vez la cantidad promedio de corrosión sea idéntica con ambos recubrimientos. Sin embargo, el recubrimiento nuevo puede ser aplicado a segmentos que tengan una resistencia superior a la corrosión y en condiciones ambientales menos severas en comparación con los segmentos y condiciones del recubrimiento actual. El investigador probablemente observaría una diferencia entre los dos recubrimientos atribuible no a los recubrimientos mismos, sino a extrañas variaciones. La estadística no sólo ofrece métodos para analizar resultados de experimentos una vez que se

han realizado sino también sugerencias sobre cómo pueden llevarse a cabo los experimentos de manera eficiente para mitigar los efectos de la variación y tener 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. De manera que 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 realimentar la calidad de los programas de estudio de ingeniería.

Por lo general, sólo interesan algunas 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 en la que 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 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 Estados Unidos. El sitio web **charitynavigator.com** proporciona información de aproximadamente 6 000 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*

```
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
```

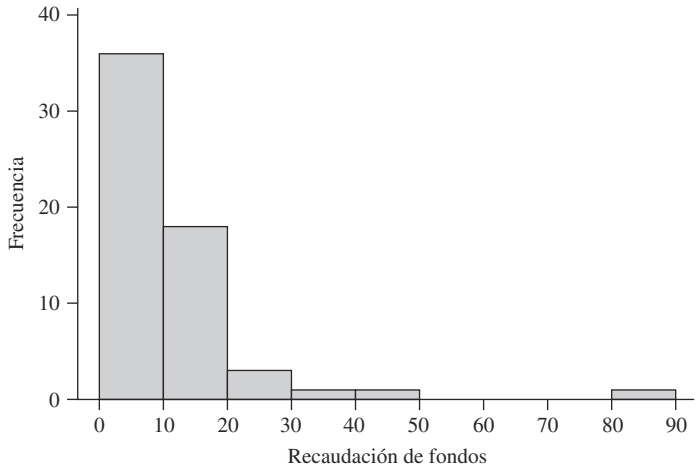


Figura 1.1 Gráfica de tallos y hojas (truncada a diez dígitos) de Minitab e histograma para los datos del porcentaje de recaudación de fondos para caridad

de tallos 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 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 con respecto a la 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 este concreto había sido investigada previamente, pero no se sabía mucho de 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 6 al 8 y luego se utilizan escenarios más complicados en los capítulos 9 al 16. El resto de este capítulo presenta métodos de estadística descriptiva que se utilizan mucho en el desarrollo de la inferencia.

Los capítulos 2 al 5 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.

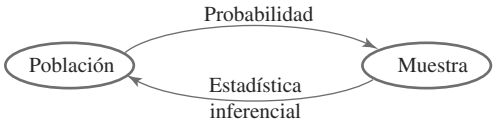


Figura 1.2 Relación entre probabilidad y estadística inferencial

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.

Ámbito de la estadística moderna

Actualmente la metodología estadística la emplean los investigadores de prácticamente todas las disciplinas, incluyendo áreas como

- biología molecular (en el análisis de datos de microarreglos)
- ecología (en la descripción cuantitativa del modo en que se distribuyen espacialmente los individuos de varias poblaciones de plantas y animales)
- ingeniería de materiales (en el estudio de las propiedades de distintos procesos para retardar la corrosión)
- marketing (en el desarrollo de estudios de mercado y estrategias para la comercialización de nuevos productos)
- salud pública (en la identificación de las fuentes de enfermedades y sus formas de tratamiento)
- ingeniería civil (en la evaluación del efecto del esfuerzo en los elementos estructurales y los impactos del flujo de vehículos en las comunidades)

A medida que se avance en la lectura de este libro se encontrará un amplio espectro de escenarios diferentes en los ejemplos y ejercicios que ilustran la aplicación de técnicas de probabilidad y estadística. Muchos de estos escenarios involucran datos u otros materiales extraídos de artículos sobre ingeniería y de revistas de ciencia. Los métodos aquí presentados convierten herramientas establecidas y confiables en el arsenal de todo aquel que trabaja con datos. Mientras tanto, los estadísticos continúan desarrollando nuevos modelos para describir aleatoriedad, incertidumbre y una metodología nueva para el análisis de datos. Como evidencia de los continuos esfuerzos creativos en la comunidad estadística, existen títulos y cápsulas con descripciones de artículos de publicación reciente en revistas de estadística (*Journal of the American Statistical Association* y los *Annals of Applied Statistics* cuyas siglas son *JASA* y *AAS*, respectivamente, dos de las revistas más importantes en esta disciplina):

- **"How Many People Do You Know? Efficiently Estimating Personal Network Size" (*JASA*, 2010: 59-70).** ¿A cuántas de las N personas de su colegio conoce? Se podría seleccionar una muestra aleatoria de alumnos de la población y usar una estimación basada en la fracción de personas que conoce de dicha muestra. Lamentablemente esto no es muy efectivo para grandes poblaciones porque la fracción de la población que una persona conoce suele ser muy pequeña. Los autores proponen un "modelo de mezcla latente" con el que afirman haber remediado las deficiencias en las técnicas utilizadas anteriormente. También fue incluido un estudio de simulación de la eficacia del método, basado en grupos formados por nombres ("¿Cuánta gente llamada Michael conoce?"), así como una aplicación del método de datos de la encuesta real. El artículo concluye con algunas recomendaciones prácticas para la construcción de futuros estudios diseñados para estimar el tamaño de la red social.
- **"Active Learning Through Sequential Design, with Applications to the Detection of Money Laundering" (*JASA*, 2009: 969-981).** El lavado de dinero consiste en ocultar el origen de los fondos obtenidos mediante actividades ilegales. El enorme número de transacciones que ocurren a diario en las instituciones financieras dificulta la detección del lavado de capitales. El planteamiento más común ha sido extraer un resumen de diversas cantidades de la historia de las transacciones y llevar a cabo una investigación de mucho tiempo de las actividades sospechosas. El artículo propone un método estadístico más eficiente e ilustra su uso en un caso de estudio.

- **"Robust Internal Benchmarking and False Discovery Rates for Detecting Racial Bias in Police Stops" (*JASA*, 2009: 661-668).** Los alegatos de las acciones de la policía atribuidas, al menos en parte, a los prejuicios raciales se han convertido en un tema polémico en muchas comunidades. En este artículo se propone un nuevo método diseñado para reducir el riesgo de marcar un número sustancial de "falsos positivos" (personas falsamente identificadas como la manifestación de un sesgo). El método se aplicó a los datos de 500 000 peatones detenidos en la ciudad de Nueva York en 2006; de los 3 000 agentes que participan regularmente en la detención de peatones, 15 fueron identificados por haber detenido una fracción mucho mayor de negros e hispanos de lo que podría predecirse en ausencia del sesgo.
- **"Records in Athletics Through Extreme Value Theory" (*JASA*, 2008: 1382-1391).** El documento se centra en el modelado de los extremos relacionados con los récords mundiales en atletismo. Los autores comienzan planteando dos cuestiones: 1) ¿Cuál es el último récord mundial en un evento específico (por ejemplo, el salto de altura para las mujeres)? y 2) ¿Cuán "bueno" es el actual récord mundial y cómo es la calidad de los actuales récords del mundo al comparar los diferentes eventos? Se considera un total de 28 eventos (ocho carreras, tres lanzamientos y tres saltos para hombres y mujeres). Por ejemplo, una conclusión es que el récord masculino de maratón sólo se ha reducido 20 segundos, pero el que registran las mujeres actualmente en el maratón es casi 5 minutos más de lo que en última instancia se puede lograr. La metodología también tiene aplicaciones para problemas como asegurarse de que las pistas de aterrizaje sean lo suficientemente largas o que los diques en Holanda sean lo suficientemente altos.
- **"Self-Exciting Hurdle Models for Terrorist Activity" (*AAS*, 2012: 106-124).** Los autores desarrollaron un modelo de predicción de actividad terrorista teniendo en cuenta el número diario de ataques terroristas en Indonesia desde 1994 hasta 2007. El modelo estima la probabilidad de futuros ataques en función de los tiempos a partir de los últimos ataques. Una característica del modelo considera muchos días sin que ocurra un ataque junto con la presencia de múltiples ataques coordinados en un mismo día. El artículo proporciona una interpretación de diversas características del modelo y evalúa su funcionamiento predictivo.
- **"Prediction of Remaining Life of Power Transformers Based on Left Truncated and Right Censored Lifetime Data" (*AAS*, 2009: 857-879).** Hay aproximadamente 150 000 transformadores de transmisión de energía de alta tensión en Estados Unidos. Fallas inesperadas pueden causar grandes pérdidas económicas, por lo que es importante contar con las predicciones de vida de los transformadores. Los datos pertinentes pueden ser complicados porque la vida útil de algunos transformadores se extiende por varias décadas durante las cuales los registros no son necesariamente completos. En particular, los autores del artículo utilizan datos de una empresa de energía que comenzó a llevar registros detallados en 1980. Sin embargo, algunos transformadores se habían instalado antes del 1 de enero de 1980 y todavía funcionaban después de esa fecha ("truncamiento a la izquierda" de datos), mientras que otras unidades estaban aún en servicio en el momento de la investigación, por lo que su vida completa no está disponible ("truncamiento a la derecha" de datos). El artículo describe los diversos procedimientos para obtener un intervalo de valores posibles (un *intervalo de predicción*) para toda la vida restante y el número acumulado de fallas en un periodo especificado.
- **"The BARISTA: A Model for Bid Arrivals in Online Auctions" (*AAS*, 2007: 412-441).** Las subastas en línea como eBay y uBid a menudo tienen características que las diferencian de las subastas tradicionales. Una diferencia muy importante es que el número de oferentes al inicio de muchas subastas tradicionales es fijo, mientras que en las subastas en línea este número y el número de ofertas resultantes no está predeterminado. El artículo propone un nuevo modelo de BARISTA (Bid ARrivals In STAges) para describir la forma en que llegan las ofertas en línea. El modelo permite mayor intensidad de ofertas al inicio de la subasta y también cuando ésta llega a su fin. Varias propiedades del modelo

son investigadas y validadas con datos de las subastas de eBay.com para asistentes personales Palm M515, juegos de Microsoft Xbox y relojes Cartier.

- **"Statistical Challenges in the Analysis of Cosmic Microwave Background Radiation" (*AAS*, 2009: 61-95).** El fondo cósmico de microondas (CMB, por sus siglas en inglés) es una fuente importante de información sobre la historia temprana del universo. Su nivel de radiación es uniforme y para medir las fluctuaciones se han desarrollado instrumentos extremadamente delicados. Los autores proporcionan una revisión de las cuestiones estadísticas del CMB con el análisis de datos, también proporcionan muchos ejemplos de aplicación de procedimientos estadísticos a los datos obtenidos de una reciente misión del satélite de la NASA, la *Wilkinson Microwave Anisotropy Probe*.

La información estadística aparece con mayor frecuencia en los medios populares y en ocasiones el centro de atención son los estadísticos. Por ejemplo, el **23 de noviembre de 2009**, el ***New York Times*** reportó en el artículo "Behind Cancer Guidelines, Quest for Data" que la nueva ciencia para la investigación del cáncer y los métodos más sofisticados para el análisis de datos realizados por los servicios preventivos de Estados Unidos impulsaron un grupo de trabajo para reexaminar las directrices respecto a la frecuencia con que las mujeres de mediana edad en adelante deben someterse a una mamografía. El panel formó seis grupos independientes para hacer modelos estadísticos. El resultado fue un nuevo conjunto de conclusiones, incluida la afirmación de que las mamografías cada dos años son tan beneficiosas para las pacientes como las mamografías anuales, pero otorga la mitad del riesgo de sufrir daños. Se cita a Donald Berry, un prominente bioestadístico, quien dice estar gratamente sorprendido de que el grupo de trabajo tomara en serio la nueva investigación para formular sus recomendaciones. El informe del grupo de trabajo ha generado mucha controversia entre organizaciones del cáncer, políticos y las propias mujeres.

Esperamos que usted se convenza cada vez más de la importancia y la pertinencia de la disciplina de la estadística, y que profundice en el libro y en el tema. Así también motivarlo lo suficiente para que continúe con su aprendizaje de la estadística más allá de este curso.

Estudios enumerativos contra analíticos

W. E. Deming, influyente estadístico estadounidense y un fuerte propulsor de la revolución de calidad de Japón durante las décadas de 1950 y 1960, introdujo la distinción entre *estudios enumerativos* y *estudios analíticos*. En los primeros el interés se enfoca en un conjunto finito, identificable y no cambiante de individuos u objetos que conforman una población. Un *marco de muestreo*, es decir, una lista de los individuos u objetos que tienen que ser muestreados, se halla disponible para el investigador o puede ser construido. Por ejemplo, el marco se podría componer de todas las firmas incluidas en una petición para calificar cierta iniciativa respecto a las boletas de una próxima votación electoral; por lo general se elige una muestra para indagar si el número de firmas *válidas* sobrepasa un valor especificado. En otro ejemplo, el marco puede contener números de serie de todos los hornos fabricados por una compañía particular durante cierto tiempo; se puede seleccionar una muestra para inferir algo sobre la duración promedio de estas unidades. El uso de los métodos inferenciales que se presentan en este libro es razonablemente no controversial en tales escenarios (aun cuando los estadísticos continúan debatiendo sobre cuáles métodos en particular deben utilizarse).

Un estudio analítico se define como aquel que no es de naturaleza enumerativa. Este tipo de estudios a menudo se realizan con el objetivo de mejorar un producto al actuar sobre un proceso de cierta clase (p. ej., recalibrar el equipo o ajustar el nivel de alguna sustancia como puede ser la cantidad de un catalizador). A menudo se obtienen datos sólo sobre un proceso existente, uno que puede diferir en aspectos importantes del proceso futuro. No existe, por tanto, un marco de muestreo que incluya los individuos o los objetos de interés. Por ejemplo, una muestra de cinco turbinas con un nuevo diseño puede ser fabricada y

probada para investigar su eficiencia. Estas cinco turbinas podrían ser consideradas como una muestra de la población conceptual de todos los prototipos que podrían ser fabricados experimentalmente en condiciones similares, pero *no* necesariamente representativas de la población de las unidades fabricadas una vez que la producción futura esté en proceso. Los métodos para utilizar la información sobre las muestras para sacar conclusiones sobre las unidades de producción futuras pueden ser problemáticos. Se debe acudir con alguien que tenga los conocimientos necesarios en el área de diseño e ingeniería de turbinas (o de cualquier otra área pertinente) para que juzgue si tal extrapolación es sensible. Una buena exposición de estos temas se encuentra en el artículo **"Assumptions for Statistical Inference", de Gerald Hahn y William Meeker (*The American Statistician*, 1993: 1-11)**.

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 en 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, descrito en el capítulo 5.

Los ingenieros y científicos a menudo reúnen datos realizando alguna clase de experimento. Esto implica decidir cómo asignar varios tratamientos diferentes (como fertilizantes o recubrimientos anticorrosivos) a las diferentes 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 (como rendimiento de un proceso de producción).

EJEMPLO 1.4 Un artículo en el ***New York Times*** **(27 de enero de 1987)** informa 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 de manera subsecuente 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 arma 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.

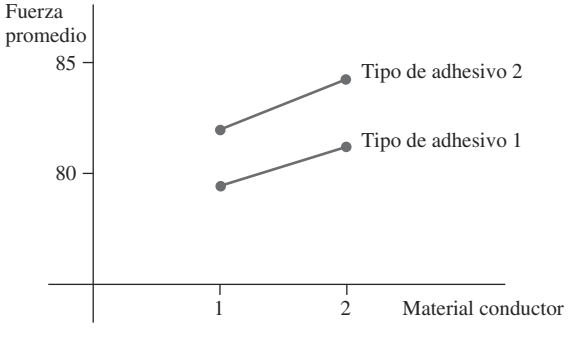


Figura 1.3 Fuerzas adhesivas promedio en el ejemplo 1.5

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. En el capítulo 11 se verá cómo la cuidadosa selección de una fracción de estas posibilidades usualmente permitirá obtener la información deseada. ■

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 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 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 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 registro, 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 $x_1, x_2, x_3, \ldots, x_n$. 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 $x_1, x_2, \ldots, x_n$ 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	55555555666666666777777888888888999999999999
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 .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, quizá 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 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 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 Estados Unidos debido a que no se gradúan suficientes estudiantes de la universidad. Estados Unidos solía 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.

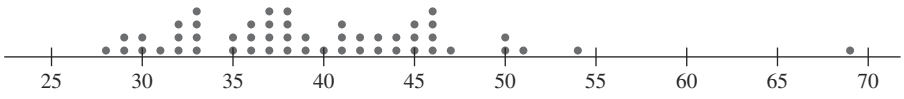


Figura 1.6 Gráfica de puntos para los datos del ejemplo 1.8 ■

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

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

$$\text{frecuencia relativa del valor } x = 3: \quad \frac{70}{200} = .35$$

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 .35 y .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	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.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".

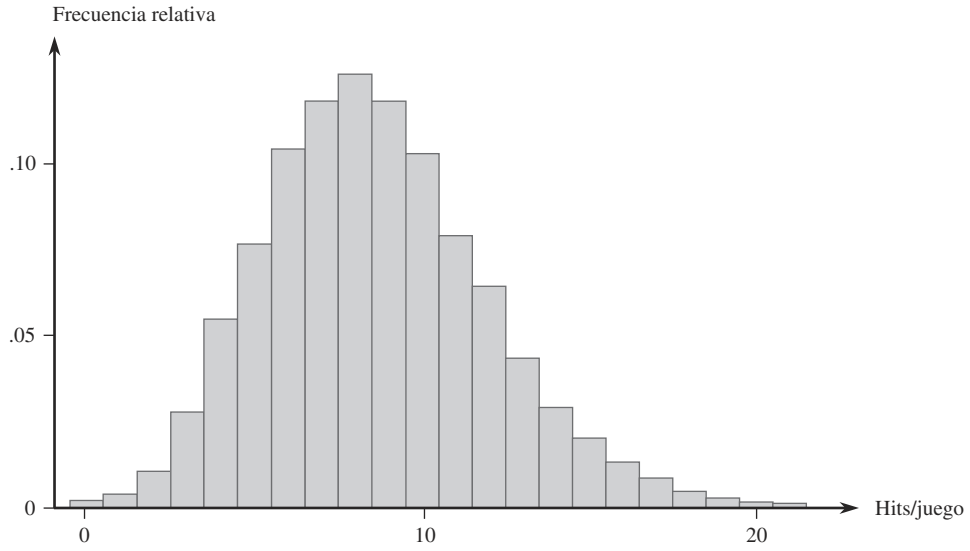


Figura 1.7 Histograma del número de hits por juego de nueve entradas

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

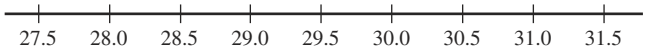
$$\begin{aligned}\text{proporción de juegos de dos hits a lo sumo} &= \text{frecuencia relativa para } x = 0 + \text{frecuencia relativa para } x = 1 + \text{frecuencia relativa para } x = 2 \\ &= .0010 + .0037 + .0108 = .0155\end{aligned}$$

De manera similar,

$$\text{proporción de juegos con entre 5 y 10 hits (inclusive)} = .0752 + .1026 + \cdots + .1015 = .6361$$

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	.011	.011	.122	.233	.278	.189	.100	.044	.011

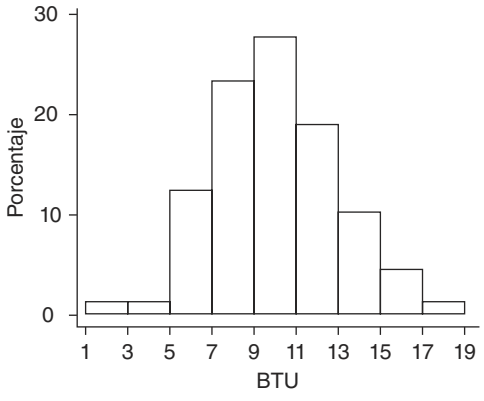


Figura 1.8 Histograma de los datos de consumo de energía del ejemplo 1.10

De acuerdo con el histograma,

$$\begin{pmatrix}\text{proporción de}\\ \text{observaciones}\\ \text{menores que 9}\end{pmatrix} \approx .01 + .01 + .12 + .23 = .37 \text{ (valor exacto} = \frac{34}{90} = .378)$$

La frecuencia relativa para la clase 9–<11 es aproximadamente .27, entonces se estima que aproximadamente la mitad de esta, o .135, queda entre 9 y 10. Por tanto,

$$\begin{pmatrix}\text{proporción de observaciones}\\ \text{menores que 10}\end{pmatrix} \approx .37 + .135 = .505 \text{ (poco más de 50\%)}$$

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

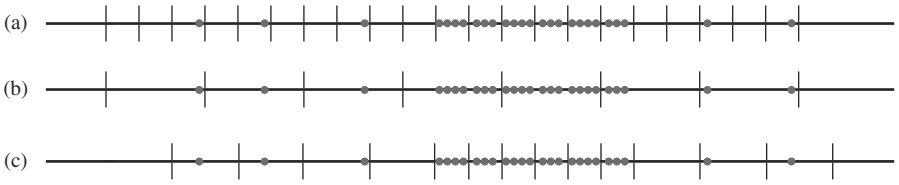


Figura 1.9 Selección de intervalos de clase para datos de "densidad variable": (a) intervalos de ancho igual muy cortos, (b) algunos intervalos de ancho igual, (c) intervalos de ancho desigual

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	.1875	.3125	.1042	.1875	.1667	.0417
Densidad	.094	.156	.052	.047	.021	.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.

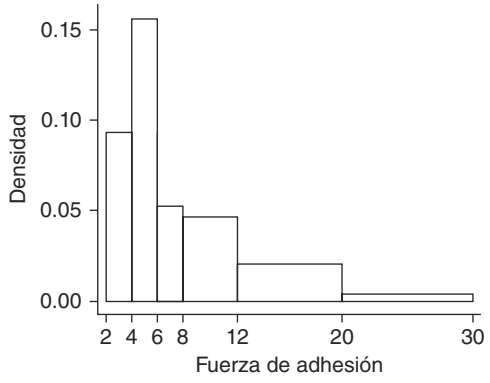


Figura 1.10 Histograma Minitab de densidad para la fuerza de adhesión del ejemplo 1.11 ■

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.).

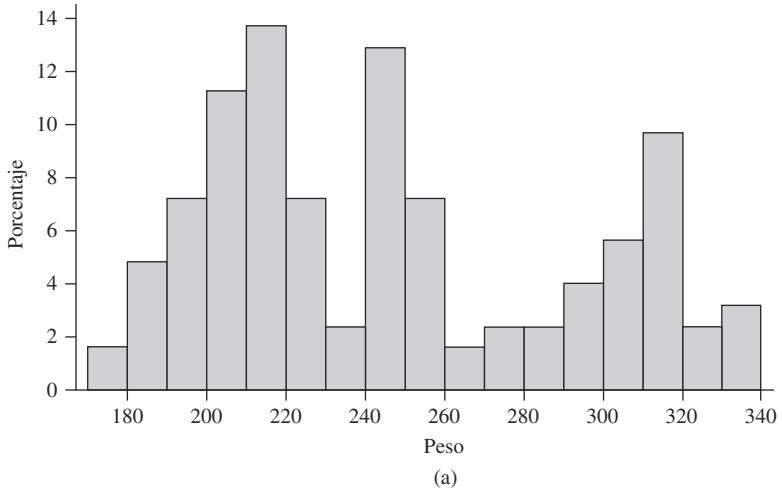


Figura 1.11 Peso de los jugadores de la NFL. (a) histograma y (b) histograma suavizado

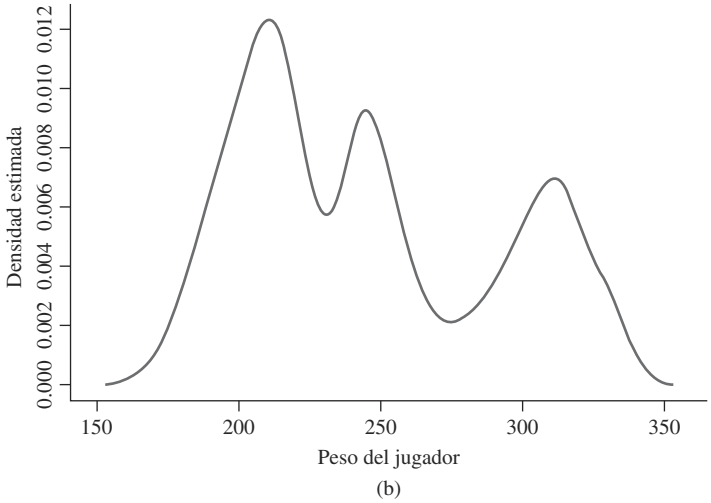


Figura 1.11 (continuación)

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.

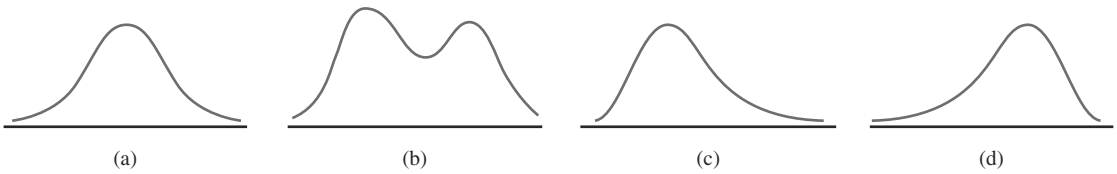


Figura 1.12 Histogramas suavizados: (a) unimodal simétrico, (b) bimodal, (c) positivamente asimétrico y (d) negativamente asimétrico

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 2 501 residentes adultos durante abril de 2006 para indagar lo que pensaban respecto a 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	.191
B	893	.357
C	680	.272
D	178	.071
F	100	.040
No sabe	172	.069
	2501	1.000

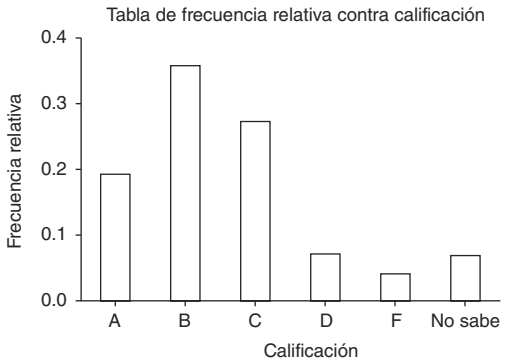


Figura 1.13 Histograma Minitab de los datos de la calificación

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, en particular 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:

.31 .35 .36 .36 .37 .38 .40 .40 .40
.41 .41 .42 .42 .42 .42 .42 .43 .44
.45 .46 .46 .47 .48 .48 .48 .51 .54
.54 .55 .58 .62 .66 .66 .67 .68 .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:

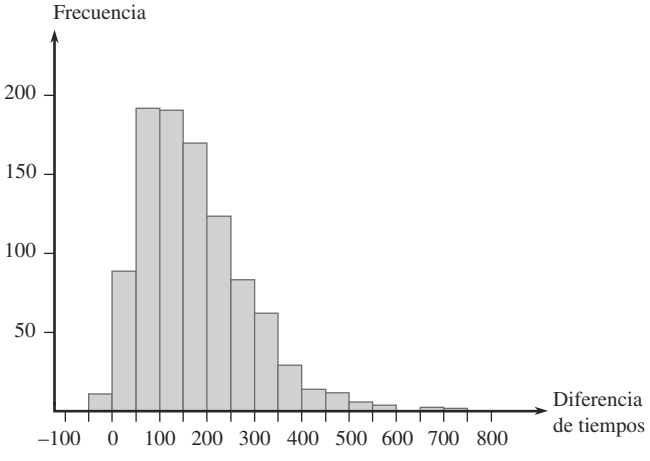
y	1	0	1	0	0	2	0	1	1	1	2	1	0	0	1	1	0	1	1
z	1	8	6	1	1	5	3	0	0	4	4	0	0	1	2	1	4	0	4
y	1	1	0	0	0	1	1	2	0	1	2	2	1	1	0	2	1	1	0
z	0	3	0	1	1	0	1	3	2	4	6	6	0	1	1	8	3	3	5
y	1	5	0	3	0	1	1	0	0										
z	0	5	2	3	1	0	0	0	3										

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:	.177	.166	.175	.136
Clase:	20−<30	30−<40	40−<60	60−<90
Frecuencia relativa:	.194	.078	.044	.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 .18, .50, .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)** informa 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.

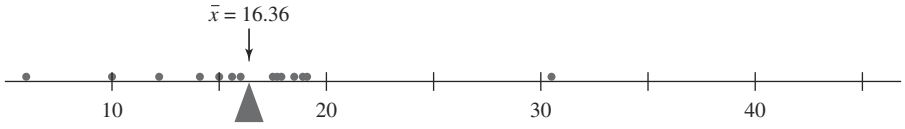


Figura 1.14 Gráfica de puntos de los datos del ejemplo 1.14 ■

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 $\mu =$ (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 $\mu =$ 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 interpretación de una pieza en particular son tan específicas que la duración no depende en absoluto de 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

Introducción a la probabilidad y estadística para ingeniería y ciencias expone, con amplitud, los modelos y métodos para el análisis de datos. Para una mejor comprensión de los temas que se abordan, esta obra presenta varios ejemplos reales, con diferentes grados de dificultad.

Características:

- Recuadros con **definiciones, hipótesis y teoremas** claves de fácil acceso.
- Gran cantidad de **ejemplos y ejercicios basados en datos o problemas reales.**
- **Ejemplos y ejercicios adicionales** acerca del material de probabilidad.
- Presentación de los temas que se enfoca en la **formación de una comprensión intuitiva** de los conceptos presentados.
- **Glosario** de símbolos y abreviaturas.
- **Apéndice** con las tablas más importantes.

Las características de esta obra permiten que el lector se conecte con la probabilidad y estadística ya que combinan sus experiencias cotidianas con sus intereses profesionales.

CENGAGE

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

ISBN-13: 978-607-526-794-4
ISBN-10: 607-526-794-8

9 786075 267944