

CAPÍTULO 1



Datos y estadísticas

CONTENIDO

LA ESTADÍSTICA EN LA PRÁCTICA: *BUSINESSWEEK*

1.1 APLICACIONES EN LOS NEGOCIOS Y EN LA ECONOMÍA
Contaduría
Finanzas
Marketing
Producción
Economía

1.2 DATOS
Elementos, variables y observaciones
Escala de medición

Datos cualitativos y cuantitativos
Datos de sección transversal y de series de tiempo

1.3 FUENTES DE DATOS
Fuentes existentes
Estudios estadísticos
Errores en la adquisición de datos

1.4 ESTADÍSTICA DESCRIPTIVA

1.5 INFERENCIA ESTADÍSTICA

1.6 LAS COMPUTADORAS Y EL ANÁLISIS ESTADÍSTICO



LA ESTADÍSTICA *en* LA PRÁCTICA

*BUSINESSWEEK** NUEVA YORK, NUEVA YORK

Con una circulación mundial de más de 1 millón de ejemplares, *BusinessWeek* es la revista más leída en el mundo. Más de 200 reporteros y editores especializados en 26 oficinas alrededor del mundo producen diversos artículos de interés para la comunidad interesada en los negocios y la economía. Junto a los artículos principales y los tópicos de actualidad, la revista presenta diversas secciones regulares sobre negocios internacionales, análisis económicos, procesamiento de la información y ciencia y tecnología. La información en las secciones regulares ayuda a los lectores a mantenerse al día de los avances y novedades y a evaluar el impacto de éstos en los negocios y en las condiciones económicas.

La mayor parte de los números de *BusinessWeek* contienen un artículo de fondo sobre algún tema de interés actual. Por ejemplo, el número del 6 de diciembre de 2004 contenía un reportaje especial sobre los precios de los artículos hechos en China; el número del 3 de enero de 2005 proporcionaba información acerca de dónde invertir en 2005 y el número del 4 de abril de 2005 proporcionaba una panorámica de *BusinessWeek 50*, un grupo diverso de empresas de alto desempeño. Además, la revista semanal *BusinessWeek Investor* proporciona artículos sobre el estado de la economía, que comprenden índices de producción, precios de las acciones de fondos mutualistas y tasas de interés.

BusinessWeek también usa métodos e información estadísticos en la administración de su propio negocio. Por ejemplo, una encuesta anual hecha a sus suscriptores le permitió tener datos demográficos sobre sus hábitos de lectura, compras probables, estilo de vida, etc. Los directivos de *BusinessWeek* usan resúmenes estadísticos obtenidos a partir de las encuestas para dar un mejor servicio a sus sus-

*Los autores agradecen a Charlene Trentham, Director de investigación de *BusinessWeek* por proporcionar este artículo para La estadística en la práctica.



BusinessWeek usa datos y resúmenes estadísticos en muchos de sus artículos. © Terri Millar/E-Visual Communications, Inc.

criptores y anunciantes. Mediante una encuesta reciente entre los suscriptores estadounidenses se supo que 90% de los suscriptores de *BusinessWeek* tienen una computadora personal en casa y que 64% de ellos realizan en el trabajo compras por computadora. Estas estadísticas indican a los directivos de *BusinessWeek* que los avances en computación serán de interés para sus suscriptores. Los resultados de la encuesta también le son proporcionados a sus anunciantes potenciales. Los elevados porcentajes de personas que tienen una computadora en casa y que realizan compras por computadora en el trabajo podría ser un incentivo para que los fabricantes de computadoras se anunciaran en *BusinessWeek*.

Este capítulo muestra los tipos de datos con que se cuenta en un análisis estadístico y describe cómo se obtienen los datos. Presenta la estadística descriptiva y la inferencia estadística como medios para convertir los datos en información estadística que tienen un significado y que es fácil de interpretar.

Con frecuencia aparece en los periódicos y revistas el siguiente tipo de información:

- La asociación de agentes inmobiliarios informó que la mediana del precio de venta de una casa en Estados Unidos es de \$215 000 (*The Wall Street Journal*, 16 de enero de 2006).
- Durante el Super Bowl de 2006 el costo promedio de un spot publicitario de 30 segundos en televisión fue de \$2.5 millones (*USA Today*, 27 de enero de 2007).

- En una encuesta de Jupiter Media se encontró que 31% de los hombres adultos ven más de 10 horas de televisión a la semana. Entre las mujeres sólo 26% (*The Wall Street Journal*, 26 de enero de 2004).
- General Motors, uno de los líderes automotrices en descuentos en efectivo da, en promedio, \$4300 de incentivo en efectivo por vehículo (*USA Today*, 27 de enero de 2006).
- Más de 40% de los directivos de Marriott Internacional ascienden por escalafón (*Fortune*, 20 de enero de 2003).
- Los Yankees de Nueva York tienen la nómina más alta dentro de la liga mayor de béisbol. En el año 2005 la nómina del equipo fue de \$208 306 817, siendo la mediana por jugador de \$5 833 334 (*USA Today*, febrero 2006).
- El promedio industrial Dow Jones cerró en 11 577 (*Barron's*, 6 de mayo de 2006).

A los datos numéricos de las frases anteriores se les llama estadísticas. En este sentido el término *estadística* se refiere a datos numéricos, tales como promedios, medianas, porcentajes y números índices que ayudan a entender una gran variedad de negocios y situaciones económicas. Sin embargo, como se verá, el campo de la estadística es mucho más que datos numéricos. En un sentido amplio, la **estadística** se define como el arte y la ciencia de reunir datos, analizarlos, presentarlos e interpretarlos. Especialmente en los negocios y en la economía, la información obtenida al reunir datos, analizarlos, presentarlos e interpretarlos proporciona a directivos, administradores y personas que deben tomar decisiones una mejor comprensión del negocio o entorno económico, permitiéndoles así tomar mejores decisiones con base en mejor información. En este libro se hace hincapié en el uso de la estadística para la toma de decisiones en los negocios y en la economía.

El capítulo 1 empieza con algunos ejemplos de aplicaciones de la estadística en los negocios y en la economía. En la sección 1.2 se define el término *datos* y se introduce el concepto de conjunto de datos. En esta sección se introducen también términos clave como *variables* y *observaciones*, se muestra la diferencia entre datos cualitativos y cuantitativos y se ilustra el uso de datos transversales y de serie de tiempo. En la sección 1.3 se enseña a obtener datos de fuentes ya existentes o mediante encuestas y estudios experimentales diseñados para obtener datos nuevos. Se resalta también el papel tan importante que tiene ahora Internet en la obtención de datos. En las secciones 1.4 y 1.5 se describe el uso de los datos en la estadística descriptiva y para hacer inferencias estadísticas.

1.1

Aplicaciones en los negocios y en la economía

En el entorno mundial actual de los negocios y de la economía, todo mundo tiene acceso a enormes cantidades de información estadística. Los directivos y los encargados de tomar decisiones que tienen éxito entienden la información y saben usarla de manera eficiente. En esta sección se proporcionan ejemplos que ilustran algunos de los usos de la estadística en los negocios y en la economía.

Contaduría

Las empresas de contadores públicos al realizar auditorías para sus clientes emplean procedimientos de muestreo estadístico. Por ejemplo, suponga que una empresa de contadores desea determinar si las cantidades en cuentas por cobrar que aparecen en la hoja de balance del cliente representan la verdadera cantidad en cuentas por cobrar. Por lo general, el gran número de cuentas por cobrar hace que su revisión tome demasiado tiempo y sea muy costosa. Lo que se hace en estos casos es que el personal encargado de la auditoría selecciona un subconjunto de las cuentas al que se le llama muestra. Después de revisar la exactitud de las cuentas tomadas en la muestra (muestreadas) los auditores concluyen si la cantidad en cuentas por cobrar que aparece en la hoja de balance del cliente es aceptable.

Finanzas

Los analistas financieros emplean una diversidad de información estadística como guía para sus recomendaciones de inversión. En el caso de acciones, el analista revisa diferentes datos financieros como la relación precio/ganancia y el rendimiento de los dividendos. Al comparar la información sobre una determinada acción con la información sobre el promedio en el mercado de acciones, el analista empieza a obtener conclusiones para saber si una determinada acción está sobre o subvaluada. Por ejemplo, *Barron's* (12 de septiembre de 2005) informa que la relación promedio precio/ganancia de 30 acciones del promedio industrial Dow Jones fue 16.5. La relación precio/ganancia de JPMorgan es 11.8. En este caso la información estadística sobre las relaciones precio/ganancia indican un menor precio en comparación con la ganancia para JPMorgan que el promedio en las acciones Dow Jones. Por tanto el analista financiero concluye que JPMorgan está subvaluada. Ésta y otras informaciones acerca de JPMorgan ayudarán al analista a comprar, vender o a recomendar mantener las acciones.

Marketing

Escáneres electrónicos en las cajas de los comercios minoristas recogen datos para diversas aplicaciones en la investigación de mercado. Por ejemplo, proveedores de datos como ACNielsen e Information Research Inc. compran estos datos a las tiendas de abarrotes, los procesan y luego venden los resúmenes estadísticos a los fabricantes; quienes gastan cientos de miles de dólares por producto para obtener este tipo de datos. Los fabricantes también compran datos y resúmenes estadísticos sobre actividades promocionales como precios o *displays* promocionales. Los administradores de marca revisan estas estadísticas y las propias de las actividades promocionales para analizar la relación entre una actividad promocional y las ventas. Estos análisis suelen resultar útiles para establecer futuras estrategias de marketing para diversos productos.

Producción

La importancia que se le da actualmente a la calidad hace del control de calidad una aplicación importante de la estadística a la producción. Para vigilar el resultado de los procesos de producción se usan diversas gráficas de control estadístico de calidad. En particular, para vigilar los resultados promedio se emplea una gráfica x -barra. Suponga, por ejemplo, que una máquina llena botellas con 12 onzas de algún refresco. Periódicamente un empleado del área de producción toma una muestra de botellas y mide el contenido promedio de refresco. Este promedio o valor x -barra se marca como un punto en una gráfica x -barra. Si este punto queda arriba del límite de control superior de la gráfica, hay un exceso en el llenado, y si queda debajo del límite de control inferior de la gráfica hay falta de llenado. Se dice que el proceso está “bajo control” y puede continuar, siempre que los valores x -barra se encuentren entre los límites de control inferior y superior. Con una interpretación adecuada, una gráfica de x -barra ayuda a determinar si es necesario hacer algún ajuste o corrección a un proceso de producción.

Economía

Los economistas suelen hacer pronósticos acerca del futuro de la economía o sobre algunos aspectos de la misma. Usan una variedad de información estadística para hacer sus pronósticos. Por ejemplo, para pronosticar las tasas de inflación, emplean información estadística sobre indicadores como el índice de precios al consumidor, la tasa de desempleo y la utilización de la capacidad de producción. Estos indicadores estadísticos se utilizan en modelos computarizados de pronósticos que predicen las tasas de inflación.

Aplicaciones de la estadística como las descritas en esta sección integran este libro. Dichos ejemplos proporcionan una visión general de la diversidad de las aplicaciones estadísticas. Como complemento de estos ejemplos, profesionales en los campos de los negocios y de la economía proporcionan los artículos de *La estadística en la práctica* que se encuentran al principio de cada capítulo, en los que se presenta el material que se estudiará en el capítulo. Las aplicaciones en *La estadística en la práctica* muestran su importancia en diversas situaciones de los negocios y la economía.

1.2 Datos

Datos son hechos/informaciones y cifras que se recogen, analizan y resumen para su presentación e interpretación. A todos los datos reunidos para un determinado estudio se les llama **conjunto de datos** para el estudio. La tabla 1.1 muestra un conjunto de datos que contiene información sobre 25 empresas que forman parte del S&P 500. El S&P 500 consta de 500 empresas elegidas por Standard & Poor's. Estas empresas representan 76% de la capitalización de mercado de todas las acciones de Estados Unidos. Las acciones de S&P 500 son estrechamente observadas por los inversionistas y por los analistas de Wall Street.

TABLA 1.1 CONJUNTO DE DATOS DE 25 EMPRESAS S&P 500

Empresa	Bolsa de valores	Denominación abreviada Ticker	Posición en <i>BusinessWeek</i>	Precio por acción (\$)	Ganancia por acción (\$)
Abbott Laboratories	N	ABT	90	46	2.02
Altria Group	N	MO	148	66	4.57
Apollo Group	NQ	APOL	174	74	0.90
Bank of New York	N	BK	305	30	1.85
Bristol-Myers Squibb	N	BMJ	346	26	1.21
Cincinnati Financial	NQ	CINF	161	45	2.73
Comcast	NQ	CMCSA	296	32	0.43
Deere	N	DE	36	71	5.77
eBay	NQ	EBAY	19	43	0.57
Federated Dept. Stores	N	FD	353	56	3.86
Hasbro	N	HAS	373	21	0.96
IBM	N	IBM	216	93	4.94
International Paper	N	IP	370	37	0.98
Knight-Ridder	N	KRI	397	66	4.13
Manor Care	N	HCR	285	34	1.90
Medtronic	N	MDT	53	52	1.79
National Semiconductor	N	NSM	155	20	1.03
Novellus Systems	NQ	NVLS	386	30	1.06
Pitney Bowes	N	PBI	339	46	2.05
Pulte Homes	N	PHM	12	78	7.67
SBC Communications	N	SBC	371	24	1.52
St. Paul Travelers	N	STA	264	38	1.53
Teradyne	N	TER	412	15	0.84
UnitedHealth Group	N	UNH	5	91	3.94
Wells Fargo	N	WFC	159	59	4.09

Fuente: Business Week (4 de abril de 2005).

Elementos, variables y observaciones

Elementos son las entidades de las que se obtienen los datos. En el conjunto de datos de la tabla 1.1, cada acción de una empresa es un elemento; los nombres de los elementos aparecen en la primera columna. Como se tienen 25 acciones, el conjunto de datos contiene 25 elementos.

Una **variable** es una característica de los elementos que es de interés. El conjunto de datos de la tabla 1.1 contiene las cinco variables siguientes:

- *Bolsa de valores (mercado bursátil)*: Dónde se comercializa (cotiza) la acción: N (Bolsa de Nueva York) y NQ (Mercado Nacional Nasdaq).
- *Ticker (denominación abreviada)*: Abreviación usada para identificar la acción en la lista de la bolsa
- *Posición en BusinessWeek*: Número del 1 al 500 que indica la fortaleza de la empresa.
- *Precio por acción (\$)*: El precio de cierre (28 de febrero de 2005).
- *Ganancia por acción (\$)*: Las ganancias por acción en los últimos 12 meses.

Los valores encontrados para cada variable en cada uno de los elementos constituyen los datos. Al conjunto de mediciones obtenidas para un determinado elemento se le llama **observación**. Volviendo a la tabla 1.1, el conjunto de mediciones para la primera observación (Abbott Laboratories) es N, ABT, 90, 46 y 2.02. El conjunto de mediciones para la segunda observación (Altria Group) es N, MO, 148, 66 y 4.57, etc. Un conjunto de datos que tiene 25 elementos contiene 25 observaciones.

Escalas de medición

La recolección de datos requiere alguna de las escalas de medición siguientes: nominal, ordinal, de intervalo o de razón. La escala de medición determina la cantidad de información contenida en el dato e indica la manera más apropiada de resumir y de analizar estadísticamente los datos.

Cuando el dato de una variable es una etiqueta o un nombre que identifica un atributo de un elemento, se considera que la escala de medición es una **escala nominal**. Por ejemplo, en relación con la tabla 1.1 la escala de medición para la variable bolsa de valores (mercado bursátil) es nominal porque N y NQ son etiquetas que se usan para indicar dónde cotiza la acción de la empresa. Cuando la escala de medición es nominal, se usa un código o una etiqueta no numérica. Por ejemplo, para facilitar la recolección de los datos y para guardarlos en una base de datos en una computadora puede emplearse un código numérico en el que 1 denote la Bolsa de Nueva York y 2 el Mercado Nacional Nasdaq. En este caso los números 1 y 2 son las etiquetas empleadas para identificar dónde cotizan las acciones. La escala de medición es nominal aun cuando los datos aparezcan como valores numéricos.

Una escala de medición para una variable es **ordinal** si los datos muestran las propiedades de los datos nominales y además tiene sentido el orden o jerarquía de los datos. Por ejemplo, una empresa automovilística (Eastside Automotive) envía a sus clientes cuestionarios para obtener información sobre su servicio de reparación. Cada cliente evalúa el servicio de reparación como excelente, bueno o malo. Como los datos obtenidos son las etiquetas excelente, bueno o malo, tienen las propiedades de los datos nominales, pero además pueden ser ordenados o jerarquizados en relación con la calidad del servicio. Un dato excelente indica el mejor servicio, seguido por bueno y, por último, malo. Por lo que la escala de medición es ordinal. Observe que los datos ordinales también son registrados mediante un código numérico. Por ejemplo, en la tabla 1.1 la posición de los datos en *BusinessWeek* es un dato ordinal. Da una jerarquía del 1 al 500 de acuerdo con la evaluación de *BusinessWeek* sobre la fortaleza de la empresa.

Una escala de medición para una variable es una **escala de intervalo** si los datos tienen las características de los datos ordinales y el intervalo entre valores se expresa en términos de una unidad de medición fija. Los datos de intervalo siempre son numéricos. Las calificaciones en una prueba de aptitudes escolares son un ejemplo de datos de intervalo. Por ejemplo, las ca-

lificaciones obtenidas por tres alumnos en la prueba de matemáticas con 620, 550 y 470, pueden ser ordenadas en orden de mejor a peor. Además las diferencias entre las calificaciones tienen significado. Por ejemplo, el estudiante 1 obtuvo $620 - 550 = 70$ puntos más que el estudiante 2 mientras que el estudiante 2 obtuvo $550 - 470 = 80$ puntos más que el estudiante tres.

Una variable tiene una **escala de razón** si los datos tienen todas las propiedades de los datos de intervalo y la proporción entre dos valores tiene significado. Variables como distancia, altura, peso y tiempo usan la escala de razón en la medición. Esta escala requiere que se tenga el valor cero para indicar que en este punto no existe la variable. Por ejemplo, considere el costo de un automóvil. El valor cero para el costo indica que el automóvil no cuesta, que es gratis. Además, si se compara el costo de un automóvil de \$30 000, con el costo de otro automóvil, \$15 000, la propiedad de razón muestra que $\$30\,000/\$15\,000 = 2$: el primer automóvil cuesta el doble del costo del segundo.

Datos cualitativos y cuantitativos

A los datos cualitativos se les suele llamar datos categóricos.

Los datos también son clasificados en cualitativos y cuantitativos. Los **datos cualitativos** comprenden etiquetas o nombres que se usan para identificar un atributo de cada elemento. Los datos cualitativos emplean la escala nominal o la ordinal y pueden ser numéricos o no. Los **datos cuantitativos** requieren valores numéricos que indiquen cuánto o cuántos. Los datos cuantitativos se obtienen usando las escalas de medición de intervalo o de razón.

El método estadístico adecuado para resumir los datos depende de si los datos son cualitativos o cuantitativos.

Una **variable cualitativa** es una variable con datos cualitativos. El análisis estadístico adecuado para una determinada variable depende de si la variable es cualitativa o cuantitativa. Si la variable es cualitativa, el análisis estadístico es bastante limitado. Tales datos se resumen contando el número de observaciones o calculando la proporción de observaciones en cada categoría cualitativa. Sin embargo, aun cuando para los datos cualitativos se use un código numérico, las operaciones aritméticas de adición, sustracción, multiplicación o división no tienen sentido. En la sección 2.1 se ven las formas de resumir datos cualitativos.

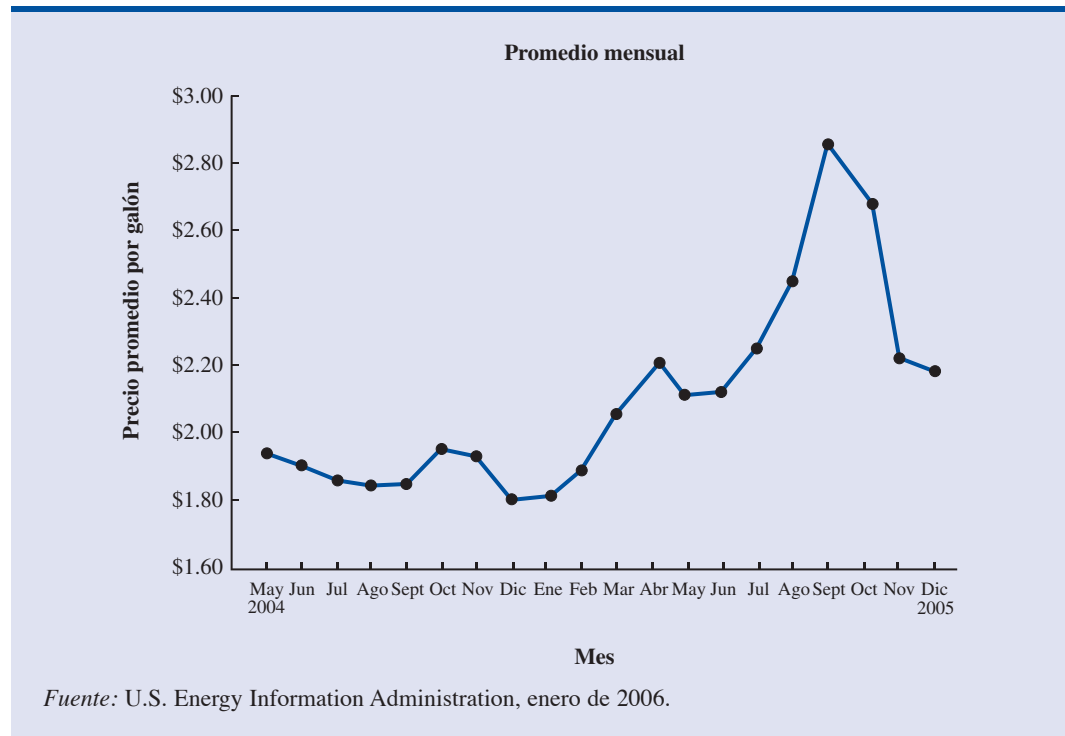
Por otro lado, las operaciones aritméticas sí tienen sentido en las variables cuantitativas. Por ejemplo, cuando se tienen variables cuantitativas, los datos se pueden sumar y luego dividir entre el número de observaciones para calcular el valor promedio. Este promedio suele ser útil y fácil de interpretar. En general hay más alternativas para el análisis estadístico cuando se tienen datos cuantitativos. La sección 2.2 y el capítulo 3 proporcionan condiciones para resumir datos cuantitativos.

Datos de sección transversal y de series de tiempo

Para los propósitos del análisis estadístico la distinción entre datos transversales y datos de series de tiempo es importante. **Datos de sección transversal** son los obtenidos en el mismo o aproximadamente el mismo momento (punto en el tiempo). Los datos de la tabla 1.1 son datos transversales porque describen las cinco variables de las 25 empresas del 25 S&P en un mismo momento. Los **datos de series de tiempo** son datos obtenidos a lo largo de varios periodos. Por ejemplo, la figura 1.1 presenta una gráfica de los precios promedio por galón de gasolina normal en las ciudades de Estados Unidos. En la gráfica se observa que los precios son bastantes estables entre \$1.80 y \$2.00 desde mayo de 2004 hasta febrero de 2005. Después el precio de la gasolina se vuelve volátil. Se eleva en forma notable culminando en un agudo pico en septiembre de 2005.

En las publicaciones sobre negocios y economía se encuentran con frecuencia gráficas de series de tiempo. Estas gráficas ayudan a los analistas a entender lo que ocurrió en el pasado, a identificar cualquier tendencia en el transcurso del tiempo y a proyectar niveles futuros para la series de tiempo. Las gráficas de datos de series de tiempo toman formas diversas como se muestra en la figura 1.2. Con un poco de estudio, estas gráficas suelen ser fáciles de entender y de interpretar.

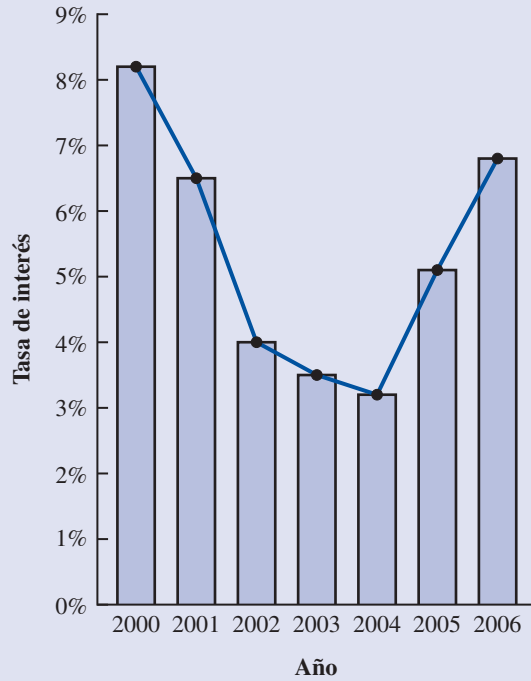
FIGURA 1.1 PRECIO PROMEDIO POR GALÓN DE GASOLINA NORMAL EN LAS CIUDADES DE ESTADOS UNIDOS



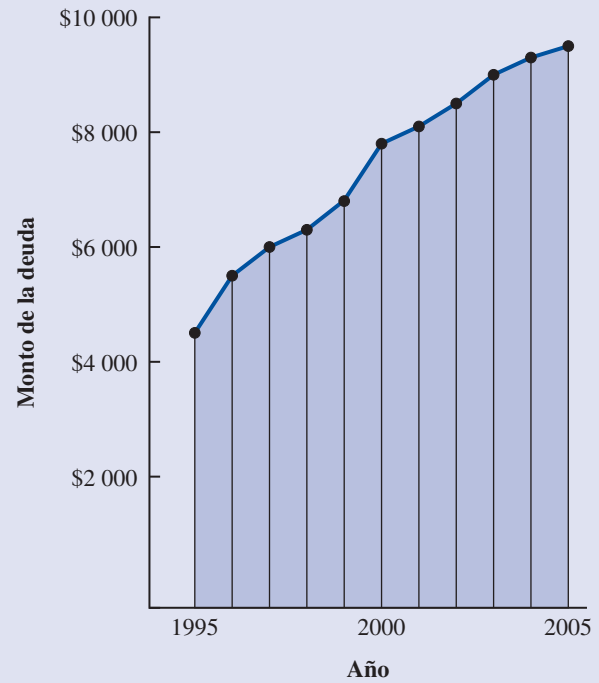
Por ejemplo, la gráfica (A) de la figura 1.2, muestra las tasas de interés en Stafford Loans para los estudiantes entre el año 2000 y el 2006. Después del año 2000 las tasas de interés disminuyen y llegan al nivel más bajo, 3.2%, en el año 2004. Pero, después de este año se observa un marcado aumento en estas tasas de interés, y llegan a 6.8% en el año 2006. El Departamento de Educación de Estados Unidos estima que más de 50% de los estudiantes terminan sus estudios con una deuda; esta creciente tasa de interés es una gran carga financiera para muchos estudiantes recién egresados.

En la gráfica (B) se observa un inquietante aumento en el adeudo promedio por hogar en tarjetas de crédito durante un periodo de 10 años, de 1995 a 2005. Advierta cómo en la serie de tiempo se nota un aumento anual casi constante en el adeudo promedio por hogar en tarjetas de crédito que va de \$4500 en 1995 a \$9500 en 2005. En 2005 un adeudo promedio de 10 000 no parece lejano. La mayor parte de las empresas de tarjetas de crédito ofrecen tasas de interés iniciales relativamente bajas. Sin embargo, después de este periodo inicial, tasas de interés anuales del 18%, 20% y más son frecuentes. Estas tasas dificultan a los hogares pagar los adeudos de las tarjetas de crédito.

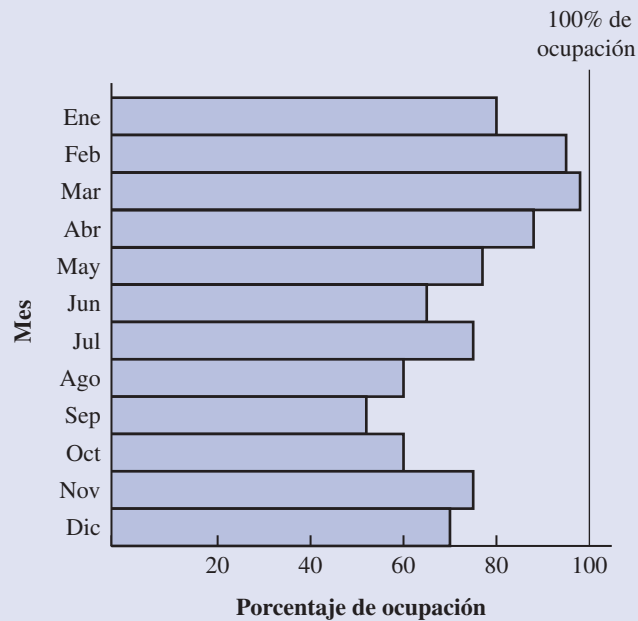
En la gráfica (C) se observan las tasas de ocupación en los hoteles de Florida del sur durante un año. Observe que la forma de esta gráfica es diferente a (A) y (B); en esta gráfica el tiempo en meses se encuentra en el eje vertical y no en el horizontal. Las tasas de ocupación más altas, 95% y 98%, se encuentran en los meses de febrero y marzo que es cuando el clima en Florida del sur es atractivo para los turistas. En efecto, de enero a abril es la estación de mayor ocupación en los hoteles de Florida del sur. Por otro lado, las tasas de ocupación más bajas se observan de agosto a octubre, siendo la menor ocupación en septiembre. Las temperaturas demasiado elevadas y la estación de huracanes son las principales razones de la caída de la ocupación en este periodo.

FIGURA 1.2 DIVERSAS GRÁFICAS DE DATOS DE SERIES DE TIEMPO

(A) Tasas de interés en los Stafford Loans para estudiantes



(B) Adeudo promedio en tarjetas de crédito por hogar



(C) Tasas de ocupación en hoteles de Florida del sur

Las series de tiempo y los pronósticos con series de tiempo se verán en el capítulo 16 cuando se estudien los métodos de pronóstico. Fuera del capítulo 16, los métodos estadísticos que se presentan en este libro son para datos de sección transversal y no para series de tiempo

NOTAS Y COMENTARIOS

1.

Una observación es el conjunto de mediciones obtenidas para cada elemento de un conjunto de datos. Por tanto, el número de observaciones es siempre igual al número de elementos. El número de mediciones de cada elemento es igual al número de variables. Entonces, el número total de datos se determina multiplicando el número de observaciones por el número de variables.
2.

Los datos cuantitativos son discretos o continuos. Datos cuantitativos que miden cuántos (por ejemplo, el número de llamadas recibidas en 5 minutos) son discretos. Datos cuantitativos que miden cuánto (por ejemplo, peso o tiempo) son continuos porque entre los posibles valores de los datos no hay separación.

1.3

Fuentes de datos

Los datos se obtienen de fuentes ya existentes o por medio de encuestas y estudios experimentales realizados con objeto de recolectar nuevos datos.

Fuentes existentes

En algunos casos los datos que se necesitan para una determinada aplicación ya existen. Las empresas cuentan con diversas bases de datos sobre sus empleados, clientes y operaciones de negocios. Datos sobre los salarios de los empleados, sus edades y los años de experiencia suelen obtenerse de los registros internos del personal. Otros registros internos contienen datos sobre ventas, gastos de publicidad, costos de distribución, inventario y cantidades de producción. La mayor parte de las empresas cuentan también con datos detallados de sus clientes. En la tabla 1.2 se muestran algunos de los datos obtenibles de los registros internos de las empresas.

De las organizaciones que se especializan en la recolección y almacenamiento de datos se obtienen cantidades importantes de datos económicos y de negocios. Las empresas disponen de estas fuentes externas de datos si los compran o mediante acuerdos de arrendamiento con opción de compra. Tres empresas que proporcionan amplios servicios de bases de datos a clientes son Dun & Bradstreet, Bloomberg y Dow Jones & Company. ACNielsen e Information Resources, Inc. han hecho un exitoso negocio recolectando y procesando datos que venden a publicistas y a fabricantes de productos.

TABLA 1.2 EJEMPLOS DE DATOS DISPONIBLES DE LOS REGISTROS DE EMPRESAS INTERNACIONALES

Fuente	Algunos de los datos disponibles
Registros sobre los empleados	Nombre, dirección, número de seguridad social, salario, días de vacaciones, días de enfermedad y bonos
Registros de producción	Parte o número de producto, cantidad producida, costo de mano de obra y costo de materiales
Registros de inventario	Parte o número de producto, cantidad de unidades disponibles, nivel de reaprovisionamiento, cantidad económica a ordenar y programa de descuento
Registros de ventas	Número del producto, volumen de ventas, volumen de ventas por región y volumen de ventas por tipo de cliente
Registros de créditos	Nombre del cliente, dirección, número de teléfono, crédito límite y cuentas por cobrar
Perfil de clientes	Edad, género, nivel de ingresos, número de miembros en la familia, dirección y preferencias

También se obtienen datos de diversas asociaciones industriales y de organizaciones de interés especial. La asociación Travel Industry Association of America cuenta con información relacionada con los viajes como número de turistas y gastos en viajes por estado. Estos datos interesan a empresas e individuos de la industria turística. El Graduate Management Admission Council cuenta con datos sobre calificaciones en exámenes, características de los estudiantes y programas de educación para administradores/directivos. La mayor parte de los datos de estas fuentes están a disposición de los usuarios calificados a un costo moderado.

La importancia de Internet como fuente de datos y de información estadística sigue creciendo. Casi todas las empresas cuentan con una página Web que proporciona información general acerca de la empresa así como datos sobre ventas, cantidad de empleados, cantidad de productos, precios de los productos y especificaciones de los productos. Además, muchas empresas se especializan ahora en proporcionar información a través de Internet. Con lo que uno puede tener acceso a cotizaciones de acciones, precios de comidas en restaurantes, datos de salarios y a una variedad casi infinita de información.

Las dependencias de los gobiernos son otra fuente importante de datos. Por ejemplo, el Departamento del Trabajo de Estados Unidos cuenta con una cantidad considerable de datos sobre tasas de empleo, tasas de salarios, magnitud de la fuerza laboral y pertenencia a sindicatos. En la tabla 1.3 se presentan algunas de las dependencias de gobierno junto con los datos que proporcionan. La mayor parte de las dependencias de los gobiernos que recolectan y procesan datos también los ponen a disposición a través de una página en la Web. Por ejemplo, la Oficina de Censos de Estados Unidos tiene una abundancia de datos en el sitio www.census.gov. En la figura 1.3 se muestra la página Web de la Oficina de Censos de Estados Unidos.

Estudios estadísticos

Algunas veces, los datos necesarios para una aplicación particular no se pueden obtener de las fuentes existentes. En tales casos los datos suelen conseguirse realizando un estudio estadístico. Dichos estudios se clasifican como *experimentales* u *observacionales*.

En los estudios experimentales se identifica primero la variable de interés. Después se ubica otra u otras variables que son controladas para lograr datos de cómo ésta influye sobre la variable de interés. Por ejemplo, a una empresa farmacéutica le interesa realizar un experimento para saber la forma en que un medicamento afecta la presión sanguínea. La variable que interesa en el estudio es la presión sanguínea. Otra variable es la dosis del nuevo medicamento que se espera tenga un efecto causal sobre la presión sanguínea. Para obtener estos datos acerca del nuevo medicamento, los investigadores eligen una muestra de individuos. La dosis del medicamento se controla dando diferentes dosis a distintos grupos de individuos. Antes y después se mide la pre-

El mayor estudio estadístico experimental jamás realizado se cree que es el experimento del Servicio de Salud Pública para la vacuna Salk contra la polio. Se eligieron casi 2 millones de niños de 1o., 2o. y 3er. grados en Estados Unidos.

TABLA 1.3 EJEMPLO DE LOS DATOS DISPONIBLES DE ALGUNAS DEPENDENCIAS GUBERNAMENTALES

Dependencia gubernamental	Algunos de los datos disponibles
Oficina de Censos www.census.gov	Datos poblacionales, número de hogares e ingresos de los hogares
Junta de la Reserva Federal www.federalreserve.gov	Datos sobre dinero en circulación, créditos a plazos, tasas de cambio y tasas de interés
Oficina de Administración y Presupuesto www.whitehouse.gov/omb	Datos sobre ingresos, gastos y deudas del gobierno federal
Departamento de Comercio www.doc.gov	Datos sobre las actividades comerciales, valor de los embarques por industria, nivel de ganancia por industria e industrias en crecimiento y en decremento
Oficina de Estadística Laboral www.bls.gov	Gasto de los consumidores, salarios por hora, tasa de desempleo y estadísticas internacionales

FIGURA 1.3 PÁGINA DE INICIO DEL SITIO WEB DE LA OFICINA DE CENSOS DE ESTADOS UNIDOS

Los estudios sobre fumadores y no fumadores son estudios observacionales porque los investigadores no determinan o controlan quién fuma y quién no.

sión sanguínea en cada grupo. El análisis estadístico de los datos experimentales ayuda a determinar el efecto del nuevo medicamento sobre la presión sanguínea.

En los estudios estadísticos no experimentales y observacionales, no se controlan las variables de interés. El tipo más usual de estudio observacional es quizá una encuesta. Por ejemplo, en una encuesta mediante entrevistas personales, primero se identifican las preguntas de la investigación. Después se presenta un cuestionario a los individuos de la muestra. Algunos restaurantes emplean estudios observacionales para obtener datos acerca de la opinión de sus clientes respecto a la calidad de los alimentos, del servicio, de la atmósfera, etc. En la figura 1.4 se presenta un cuestionario empleado por el restaurante Lobster Pot de Florida. Observe que en el cuestionario se pide a los clientes evaluar cinco variables: calidad de los alimentos, amabilidad en el servicio, prontitud en el servicio, limpieza y gestión. Las categorías para las respuestas de excelente, bueno, satisfactorio e insatisfactorio proporcionan datos ordinales que permiten a los directivos de Lobster Pot evaluar la calidad de operación del restaurante.

Los directivos que deseen emplear datos y análisis estadístico como ayuda en la toma de decisiones deben estar conscientes del tiempo y costo que requiere la obtención de los datos. Cuando es necesario obtener los datos en poco tiempo, es deseable el uso de fuentes de datos ya existentes. Si no es posible obtener con facilidad datos importantes de fuentes ya existentes, debe tomarse en cuenta el tiempo y el costo necesarios para obtener los datos. En todos los casos, las personas encargadas de tomar las decisiones deben considerar la contribución del análisis estadístico en el proceso de la toma de decisiones. El costo de la adquisición de datos y del subsiguiente análisis no deben exceder a los ahorros generados por el uso de esta información para tomar una decisión mejor.

Errores en la adquisición de datos

Los directivos siempre deben estar conscientes de la posibilidad de errores en los datos de los estudios estadísticos. Usar datos erróneos es peor que no usar ningún dato. Un error en la adquisición de datos se tiene siempre que el valor del dato obtenido no es igual al verdadero valor o al valor real que se hubiera obtenido con un procedimiento correcto. Estos errores ocurren de va-

FIGURA 1.4 CUESTIONARIO PARA CONOCER LA OPINIÓN DE LOS CLIENTES EMPLEADO EN EL RESTAURANTE THE LOBSTER POT DE REDINGTON SHORES, FLORIDA

TheLOBSTERPot

RESTAURANT

Nos alegramos de su visita al restaurante Lobster Pot y queremos estar seguros de que volverá. De manera que si tiene unos minutos le agradeceríamos mucho que nos llenara esta tarjeta. Sus comentarios y sugerencias son extremadamente importantes para nosotros. Gracias.

Nombre de la persona que lo atendió _____

	Excelente	Bueno	Satisfactorio	Insatisfactorio
Calidad de los alimentos	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Amabilidad en el servicio	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Prontitud en el servicio	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Limpieza	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Gestión	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Comentarios _____

¿Qué lo motivó a visitarnos? _____

Favor de depositarlo en el buzón de sugerencias que se encuentra a la entrada.

rias maneras. Por ejemplo, un entrevistador puede cometer un error de escritura, como una transposición al escribir la edad de una persona y en lugar de 24 años escribir 42 años, o en una entrevista, el entrevistado puede malinterpretar una pregunta y dar una respuesta incorrecta.

Los analistas de datos con experiencia tienen sumo cuidado tanto al recolectar los datos como al registrarlos para garantizar que no se cometan errores. Para comprobar la consistencia interna de los datos se emplean procedimientos especiales. Tales procedimientos indican al analista, por ejemplo, que debe revisar la consistencia de los datos cuando un entrevistado aparece con 22 años de edad pero informa tener 20 años de experiencia en el trabajo. El analista de datos también debe revisar datos que tengan valores inusualmente grande o pequeños, llamados observaciones atípicas, que son candidatos a posibles errores en los datos. En el capítulo 3 se muestran algunos de los métodos estadísticos útiles para identificar observaciones atípicas.

Los errores suelen presentarse durante la adquisición de datos. Emplear a ciegas cualquier dato que se tenga o valerse de datos que fueron adquiridos con poco cuidado da como resultado información desorientadora y malas decisiones. Así, tomar medidas para adquirir datos precisos ayuda a garantizar información confiable y valiosa para la toma de decisiones.

1.4

Estadística descriptiva

La mayor parte de la información estadística en periódicos, revistas, informes de empresas y otras publicaciones consta de datos que se resumen y presentan en una forma fácil de leer y de entender. A estos resúmenes de datos, que pueden ser tabulares, gráficos o numéricos se les conoce como **estadística descriptiva**.

TABLA 1.4 FRECUENCIAS Y FRECUENCIAS PORCENTUALES DE LA VARIABLE BOLSA DE VALORES

Bolsa de valores	Frecuencia	Frecuencia porcentual
Bolsa de Nueva York	20	80
Mercado Nacional Nasdaq	5	20
Totales	25	100

Vuelva al conjunto de datos de la tabla 1.1 que presenta 25 de las empresas de S&P 500. Los métodos de la estadística descriptiva pueden emplearse para resumir la información en este conjunto de datos. Por ejemplo, en la tabla 1.4 se presenta un resumen tabular de los datos de la variable bolsa de valores. Un resumen gráfico de los mismos datos, al que se le llama gráfica de barras aparece en la figura 1.5. Estos tipos de resúmenes, tabular y gráfico, permiten que los datos sean más fáciles de interpretar. Al revisar la tabla 1.4 y la figura 1.5 es fácil entender que la mayor parte de las acciones del conjunto de datos cotizan en la bolsa de Nueva York. Si emplea porcentajes: 80% cotizan en la bolsa de Nueva York y 20% en el Nasdaq.

En la figura 1.6 se presenta un resumen gráfico, llamado histograma, de los datos de la variable cuantitativa precio por acción. El histograma facilita ver que los precios por acción van de \$0 a \$100, con una mayor concentración entre \$20 y \$60.

Además de las presentaciones tabular y gráfica para resumir datos se emplea también la estadística descriptiva numérica. El estadístico descriptivo más común para resumir datos es el promedio o media. Mediante los datos de la variable ganancia por acción de las acciones S&P de la tabla 1.1, el promedio se calcula sumando las ganancias por acción de las 25 acciones y dividiendo

FIGURA 1.5 GRÁFICA DE BARRAS DE LA VARIABLE BOLSA DE VALORES

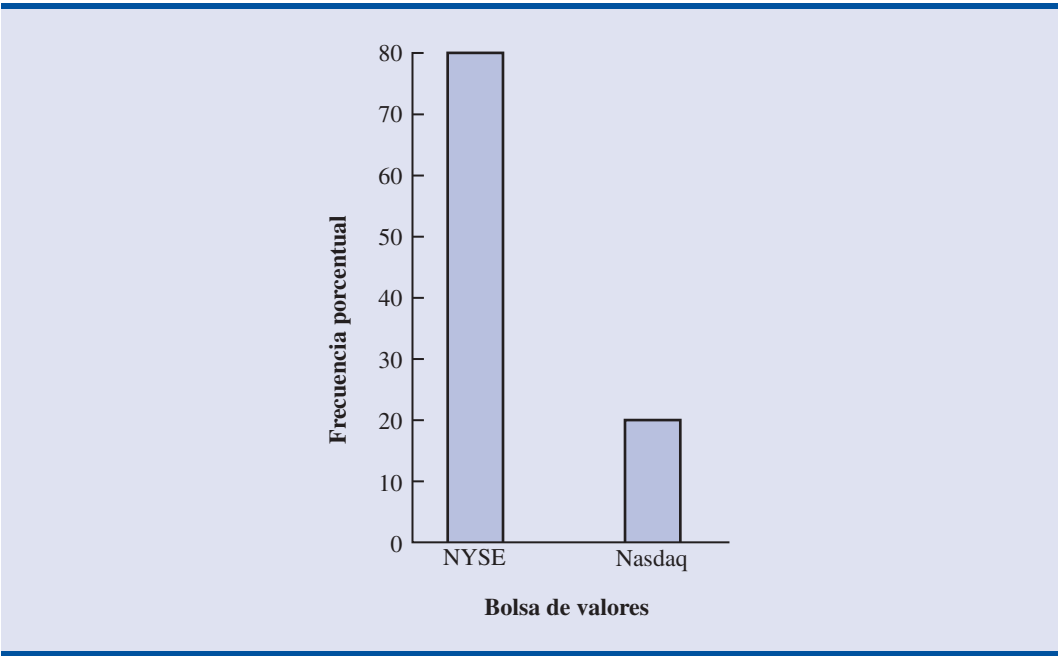
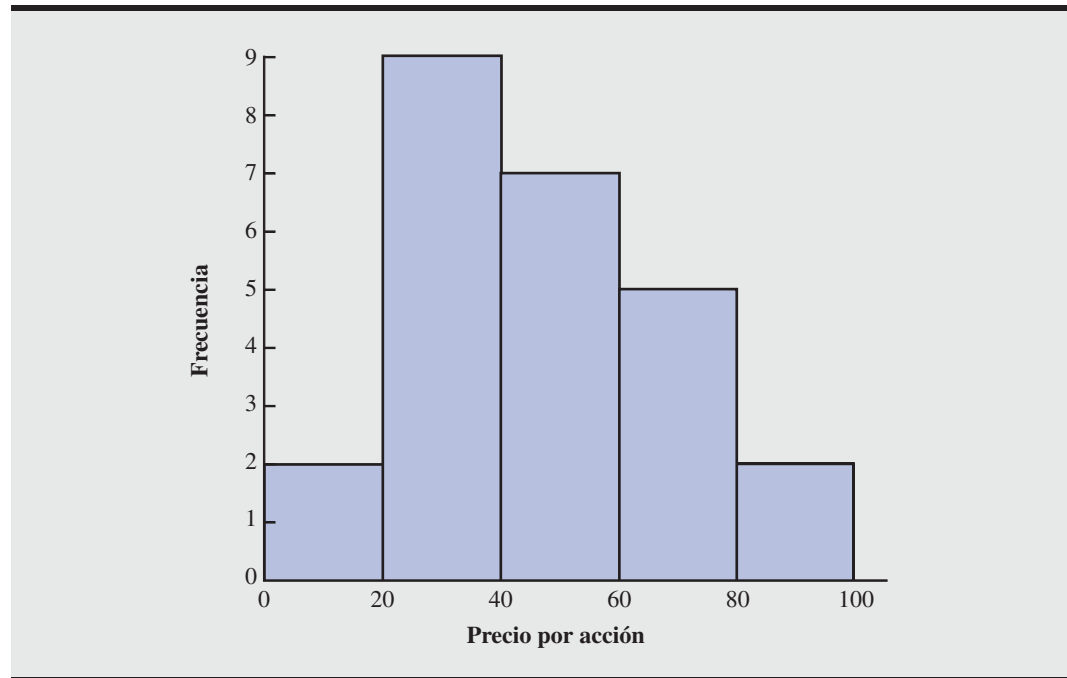


FIGURA 1.6 HISTOGRAMA DE LOS PRECIOS POR ACCIÓN DE 25 ACCIONES S&P

do entre 25. Al hacer esto se obtiene como ganancia promedio por acción \$2.49. Este promedio da una tendencia central, o posición central, de los datos de la variable.

En numerosos campos sigue creciendo el interés por los métodos estadísticos que son aplicables para elaborar y presentar estadísticas descriptivas. En los capítulos 2 y 3 se dedica la atención a los métodos tabulares, gráficos y numéricos de la estadística descriptiva.

1.5

Inferencia estadística

En muchas situaciones se requiere información acerca de grupos grandes de elementos (individuos, empresas, votantes, hogares, productos, clientes, etc.). Pero, debido al tiempo, costo y a otras consideraciones, sólo es posible recolectar los datos de una pequeña parte de este grupo. Al grupo grande de elementos en un determinado estudio se le llama **población** y al grupo pequeño **muestra**. En términos formales se emplean las definiciones siguientes.

POBLACIÓN

La población es el conjunto de todos los elementos de interés en un estudio determinado.

MUESTRA

La muestra es un subconjunto de la población.

El gobierno de Estados Unidos realiza un censo cada 10 años. Las empresas de investigación de mercado realizan estudios muestrales cada día.

Al proceso de realizar un estudio para recolectar datos de toda una población se le llama **censo**. Al proceso de efectuar un estudio para recolectar datos de una muestra se le llama **encuesta muestral**. Una de las principales contribuciones de la estadística es emplear datos de una muestra para hacer estimaciones y probar hipótesis acerca de las características de una población mediante un proceso al que se le conoce como **inferencia estadística**.

Como un ejemplo de inferencia estadística, considere un estudio realizado por Norris Electronics. Norris fabrica focos de alta intensidad que se emplean en diversos productos electrónicos. Con objeto de incrementar la vida útil de estos focos, el grupo de diseño del producto elaboró un filamento nuevo. En este caso, la población está definida por todos los focos que se produzcan con el filamento nuevo. Para evaluar las ventajas del filamento, se fabricaron 200 focos. Los datos recolectados de esta muestra dan el número de horas que duró cada foco hasta que se quemara el filamento. Véase la tabla 1.5.

Suponga que Norris desea usar estos datos muestrales para hacer una inferencia acerca del número de horas promedio de vida útil de todos los focos que se producen con el filamento nuevo. Al sumar los 200 valores de la tabla 1.5 y dividir la suma entre 200 se obtiene el promedio del tiempo de vida de los focos: 76 horas. Este resultado muestral sirve para estimar que el tiempo de vida promedio de los focos de la población es 76 horas. En la figura 1.7 se proporciona un resumen gráfico del proceso de inferencia estadística empleado por Norris Electronics.

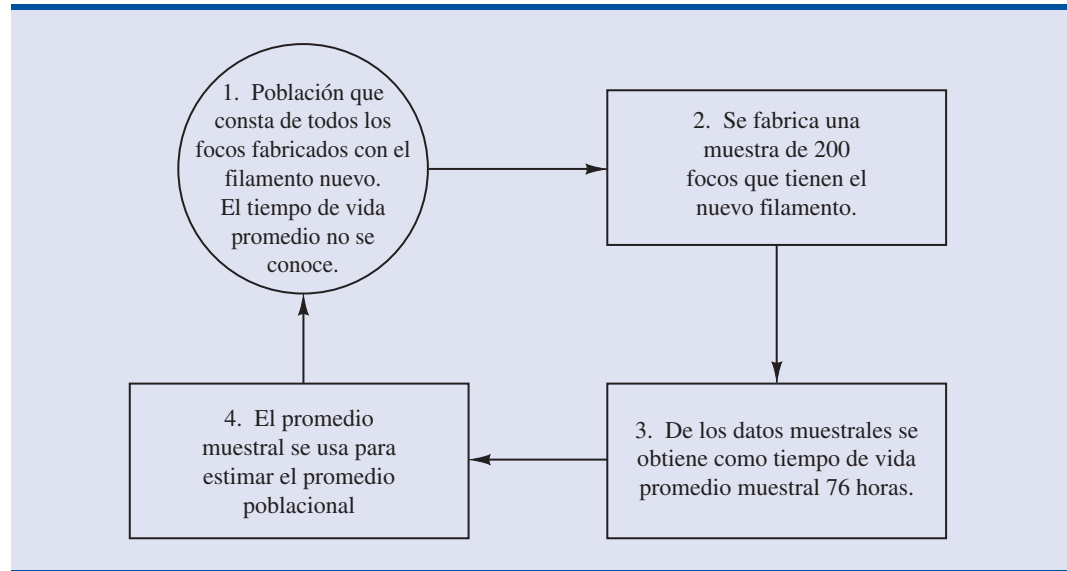
Siempre que un estadístico usa una muestra para estimar una característica poblacional que interesa, suele proporcionar información acerca de la calidad o precisión de la estimación. En el ejemplo de Norris, el estadístico puede informar que la estimación puntual del tiempo de vida promedio de la población de los nuevos focos es 76 horas con un margen de error de ± 4 horas. Entonces, el intervalo de estimación del tiempo de vida promedio de los focos fabricados con el nuevo filamento es de 72 a 80 horas. El estadístico también puede informar qué tan confiado está de que el intervalo de 72 a 80 horas contenga el promedio poblacional.

TABLA 1.5 HORAS DE DURACIÓN DE UNA MUESTRA DE 200 FOCOS DE NORRIS

107	73	68	97	76	79	94	59	98	57
54	65	71	70	84	88	62	61	79	98
66	62	79	86	68	74	61	82	65	98
62	116	65	88	64	79	78	79	77	86
74	85	73	80	68	78	89	72	58	69
92	78	88	77	103	88	63	68	88	81
75	90	62	89	71	71	74	70	74	70
65	81	75	62	94	71	85	84	83	63
81	62	79	83	93	61	65	62	92	65
83	70	70	81	77	72	84	67	59	58
78	66	66	94	77	63	66	75	68	76
90	78	71	101	78	43	59	67	61	71
96	75	64	76	72	77	74	65	82	86
66	86	96	89	81	71	85	99	59	92
68	72	77	60	87	84	75	77	51	45
85	67	87	80	84	93	69	76	89	75
83	68	72	67	92	89	82	96	77	102
74	91	76	83	66	68	61	73	72	76
73	77	79	94	63	59	62	71	81	65
73	63	63	89	82	64	85	92	64	73



FIGURA 1.7 PROCESO DE INFERENCIA ESTADÍSTICA EMPLEADO EN EL EJEMPLO DE NORRIS ELECTRONICS



1.6

Las computadoras y el análisis estadístico

Como en el análisis estadístico suelen emplearse grandes cantidades de datos, los analistas usan software para realizar estos trabajos. Por ejemplo, calcular el tiempo de vida promedio de los 200 focos del ejemplo de Norris Electronics (véase tabla 1.5) resultaría muy tedioso si no se contara con una computadora. Para facilitar el uso de una computadora, los conjuntos de datos de este libro se proporcionan en el disco compacto que viene con el libro. Un logotipo al margen izquierdo del texto identifica a estos conjuntos de datos. Los archivos de datos se encuentran en formatos para Minitab y para Excel. Además, en los apéndices de los capítulos aparecen las instrucciones para llevar a cabo los procedimientos estadísticos usando Minitab y Excel.

Resumen

La estadística es el arte y la ciencia de recolectar, analizar, presentar e interpretar datos. Casi todos los estudiantes de áreas relacionadas con los negocios o la economía necesitan tomar un curso de estadística. Este libro empezó describiendo las aplicaciones típicas de la estadística a los negocios y a la economía.

Los datos consisten en hechos/informaciones y cifras que se recolectan y analizan. Las cuatro escalas de medición que se usan para obtener datos sobre una determinada variable son nominal, ordinal, de intervalo y de razón. La escala de medición para una variable es nominal cuando los datos son etiquetas o nombres que se usan para identificar un atributo de un elemento. La escala es ordinal si los datos presentan las propiedades de los datos nominales y tiene sentido hablar del orden o jerarquía de los datos. La escala es de intervalo si los datos presentan las propiedades de los datos ordinales y los intervalos entre valores se expresan en términos de una unidad fija de medición. Por último, la escala de medición es de razón si los datos presentan las propiedades de los datos de intervalo y tiene sentido hablar de la razón entre dos valores.

Para los propósitos del análisis estadístico, los datos son clasificables en cuantitativos y cualitativos. Los datos cualitativos emplean etiquetas o nombres para identificar un atributo en cada elemento. Los datos cualitativos emplean las escalas de medición nominal u ordinal y pueden ser no numéricos o numéricos. Los datos cuantitativos son valores numéricos que indican cuánto o cuántos. Los datos cuantitativos emplean las escalas de medición de intervalo o de razón. Las operaciones aritméticas usuales sólo tienen sentido si los datos son cuantitativos. Por tanto, los cálculos estadísticos usados para datos cuantitativos no siempre son apropiados para datos cualitativos.

En las secciones 1.4 y 1.5 se introdujeron los temas de estadística descriptiva e inferencia estadística. Estadística descriptiva son los métodos tabulares, gráficos o numéricos que se usan para resumir datos. El proceso de la inferencia estadística emplea los datos obtenidos de una muestra para hacer estimaciones o probar hipótesis acerca de las características de la población. En la última sección del capítulo se indicó que las computadoras facilitan el análisis estadístico. Los conjuntos de datos grandes en los archivos de Minitab o de Excel se encuentran en el disco compacto que va con el libro.

Glosario

Estadística El arte y la ciencia de recolectar, analizar, presentar e interpretar datos.

Datos Los hechos y las cifras que se recolectan, analizan y resumen para su presentación e interpretación.

Conjunto de datos Todos los datos recolectados en un estudio determinado.

Elementos Entidades sobre las que se recolectan los datos.

Variable Una característica que interesa de un elemento.

Observación El conjunto de mediciones obtenidas de un elemento determinado.

Escala nominal Escala de medición de una variable cuando los datos son etiquetas o nombres que se emplean para identificar un atributo de un elemento. Los datos nominales pueden ser no numéricos o numéricos.

Escala ordinal Escala de medición de una variable cuando los datos presentan las propiedades de los datos nominales y el orden o jerarquía de los datos tiene sentido. Los datos ordinales pueden ser no numéricos o numéricos.

Escala de intervalo Escala de medición de una variable cuando los datos presentan las propiedades de los datos ordinales y los intervalos entre valores se expresan en términos de una unidad o medida fija. Los datos de intervalo siempre son numéricos.

Escala de razón Escala de medición de una variable cuando los datos presentan todas las propiedades de los datos de intervalo y la razón entre dos valores tiene sentido. Los datos de razón siempre son numéricos.

Datos cualitativos Etiquetas o nombres utilizados para identificar un atributo de cada elemento. Los datos cualitativos usan las escalas de medición nominal y ordinal y pueden ser no numéricos o numéricos.

Datos cuantitativos Valores numéricos que indican cuánto o cuántos de algo. Los datos cuantitativos se obtienen mediante la escala de intervalo o de razón.

Variable cualitativa Una variable con datos cualitativos.

Variable cuantitativa Una variable con datos cuantitativos.

Datos de sección transversal Datos recolectados en el mismo o aproximadamente en el mismo momento.

Datos de series de tiempo Datos recolectados a lo largo de varios periodos de tiempo.

Estadística descriptiva Resúmenes tabulares, gráficos o numéricos de datos.

Población Conjunto de todos los elementos que interesan en un estudio determinado.

Muestra Un subconjunto de la población.

Censo Un estudio para recolectar los datos de toda la población.

Encuesta muestral Un estudio para recolectar los datos de una muestra.

Inferencia estadística El proceso de emplear los datos obtenidos de una muestra para hacer estimaciones o probar hipótesis acerca de las características de la población.

Autoexamen

Autoexamen

1. Describa la diferencia entre estadística como dato numérico y estadística como disciplina o campo de estudio.
2. La revista *Condé Nast Traveler* realiza una encuesta anual entre sus suscriptores con objeto de determinar los mejores alojamientos del mundo. En la tabla 1.6 se presenta una muestra de nueve hoteles europeos (*Condé Nast Traveler*, enero de 2000). Los precios de una habitación doble estándar van de \$(precio más bajo) a \$\$\$\$ (precio más alto). La calificación general corresponde a la evaluación de habitaciones, servicio, restaurante, ubicación/atmósfera y áreas públicas; cuanto más alta sea la calificación general, mayor es el nivel de satisfacción.
 - a. ¿Cuántos elementos hay en este conjunto de datos?
 - b. ¿Cuántas variables hay en este conjunto de datos?
 - c. ¿Cuáles variables son cualitativas y cuáles cuantitativas?
 - d. ¿Qué tipo de escala de medición se usa para cada variable?
3. Vaya a la tabla 1.6.
 - a. ¿Cuál es el número promedio de habitaciones en los nueve hoteles?
 - b. Calcule la calificación general promedio.
 - c. ¿Qué porcentaje de los hoteles se encuentra en Inglaterra?
 - d. ¿En qué porcentaje de los hoteles el precio de la habitación es de \$\$?
4. Los equipos de sonido todo en uno, llamados minicomponentes, cuentan con sintonizador AM/FM, casetera doble, cargador para un disco compacto con bocinas separadas. En la tabla 1.7 se muestran los precios de menudeo, calidad de sonido, capacidad para discos compactos, sensibilidad y selectividad de la sintonización y cantidad de caseteras en los artículos de una muestra de 10 minicomponentes (*Consumer Report Buying Guide 2002*).
 - a. ¿Cuántos elementos contiene este conjunto de datos?
 - b. ¿Cuál es la población?
 - c. Calcule el precio promedio en la muestra.
 - d. Con los resultados del inciso c, estime el precio promedio para la población.
5. Considere el conjunto de datos de la muestra de los 10 minicomponentes que se muestra en la tabla 1.7.
 - a. ¿Cuántas variables hay en este conjunto de datos?
 - b. De estas variables, ¿cuáles son cualitativas y cuáles son cuantitativas?
 - c. ¿Cuál es la capacidad promedio de CD en la muestra?
 - d. ¿Qué porcentaje de los minicomponentes tienen una sintonización de FM buena o excelente?
 - e. ¿Qué porcentaje de los minicomponentes tienen dos caseteras?

TABLA 1.6 CALIFICACIONES PARA NUEVE LUGARES DONDE ALOJARSE EN EUROPA

Nombre del lugar	País	Precio de la habitación	Número de habitaciones	Calificación general
Graveteye Manor	Inglaterra	\$\$	18	83.6
Villa d'Este	Italia	\$\$\$\$	166	86.3
Hotel Prem	Alemania	\$	54	77.8
Hotel d'Europe	Francia	\$\$	47	76.8
Palace Luzern	Suiza	\$\$	326	80.9
Royal Crescent Hotel	Inglaterra	\$\$\$	45	73.7
Hotel Sacher	Austria	\$\$\$	120	85.5
Duc de Bourgogne	Bélgica	\$	10	76.9
Villa Gallici	Francia	\$\$	22	90.6

Fuente: *Condé Nast Traveler*, enero de 2000.

TABLA 1.7 UNA MUESTRA DE 10 MINICOMPONENTES

Marca y modelo	Precio (\$)	Calidad de sonido	Capacidad para CD	Sintonización FM	Caseteras
Aiwa NSX-AJ800	250	Buena	3	Regular	2
JVC FS-SD1000	500	Buena	1	Muy buena	0
JVC MX-G50	200	Muy buena	3	Excelente	2
Panasonic SC-PM11	170	Regular	5	Muy buena	1
RCA RS 1283	170	Buena	3	Mala	0
Sharp CD-BA2600	150	Buena	3	Buena	2
Sony CHC-CL1	300	Muy buena	3	Muy buena	1
Sony MHC-NX1	500	Buena	5	Excelente	2
Yamaha GX-505	400	Muy buena	3	Excelente	1
Yamaha MCR-E100	500	Muy buena	1	Excelente	0



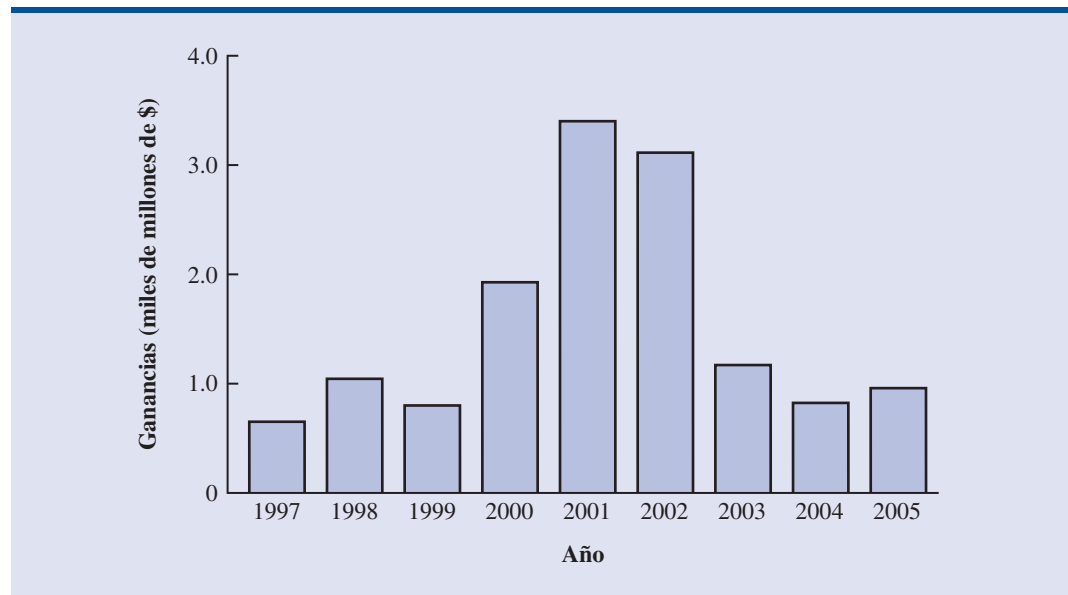
6. La Columbia House vende discos compactos a los miembros de su club de venta por correo. En una encuesta sobre música se les pidió a los nuevos miembros del club que llenaran un cuestionario con 11 preguntas. Algunas de las preguntas eran:
 - a. ¿Cuántos discos compactos has comprado en los últimos 12 meses?
 - b. ¿Eres miembro de algún club de venta de libros por correo (Sí o No)?
 - c. ¿Cuál es tu edad?
 - d. Incluyéndote a ti, de cuántas personas (adultos y niños) consta tu familia.
 - e. ¿Qué tipo de música te interesa comprar? Se presentaban quince categorías entre las que se encontraban rock pesado, rock ligero, música contemporánea para adultos, rap y rancheras. Responde si los datos que se obtienen con cada pregunta son cualitativos o cuantitativos.
7. El hotel Ritz Carlton emplea un cuestionario de opinión del cliente para obtener datos sobre la calidad de sus servicios de restaurante y entretenimiento (The Ritz-Carlton Hotel, Naples, Florida, febrero de 2006). Se les pidió a los clientes que evaluaran seis puntos: recibimiento, servicio, alimentos, menú, atención y atmósfera. Los datos registrados para cada factor fueron 1 para Pasadero, 2 Regular, 3 Bueno y 4 Excelente.
 - a. Las respuestas de los clientes proporcionan datos para seis variables. ¿Son estas variables cualitativas o cuantitativas?
 - b. ¿Qué escala de medición se usa?
8. La empresa Gallup realizó una encuesta telefónica empleando una muestra aleatoria nacional compuesta de 1005 adultos de 18 años o más. En la encuesta se les preguntó a los participantes “Cómo considera que es su salud física en este momento” (www.gallup.com, 7 de febrero de 2002). Las respuestas podían ser Excelente, Buena, Regular o Ninguna opinión.
 - a. ¿Cuál es el tamaño de la muestra de esta investigación?
 - b. ¿Son estos datos cualitativos o cuantitativos?
 - c. ¿Sería conveniente usar promedios o porcentajes para resumir los datos de estas preguntas?
 - d. De las personas que respondieron, 29% dijo que su salud era excelente. ¿Cuántos fueron los individuos que dieron esta respuesta?
9. El Departamento de Comercio informa haber recibido las siguientes solicitudes para concursar por el Malcolm Baldrige National Quality Award: 23 de empresas fabricantes grandes, 18 de empresas grandes de servicios y 30 de negocios pequeños.
 - a. ¿Es el tipo de empresa una variable cualitativa o cuantitativa?
 - b. ¿Qué porcentaje de las solicitudes venían de negocios pequeños?
10. En una encuesta de *The Wall Street Journal* (13 de octubre de 2003) se les hacen a los suscriptores 46 preguntas acerca de sus características e intereses. De cada una de las preguntas si-

guientes, ¿cuál proporciona datos cualitativos o cuantitativos e indica la escala de medición apropiada?

- ¿Cuál es su edad?
 - ¿Es usted hombre o mujer?
 - ¿Cuándo empezó a leer el *WSJ*? Preparatoria, universidad al comienzo de la carrera, a la mitad de la carrera, al final de la carrera o ya retirado.
 - ¿Cuánto tiempo hace que tiene su trabajo o cargo actual?
 - ¿Qué tipo de automóvil piensa comprarse la próxima vez que compre uno? Ocho categorías para las respuestas, entre las que se encontraban sedán, automóvil deportivo, miniván, etcétera.
- Diga de cada una de las variables siguientes si es cualitativa o cuantitativa e indique la escala de medición a la que pertenece.
 - Ventas anuales.
 - Tamaño de los refrescos (pequeño, mediano, grande).
 - Clasificación como empleado (GS 1 a GS 18).
 - Ganancia por acción.
 - Modo de pago (al contado, cheque, tarjeta de crédito).
 - La Oficina de Visitantes a Hawai recolecta datos de los visitantes. Entre las 16 preguntas hechas a los pasajeros de un vuelo de llegada en junio de 2003 estaban las siguientes.
 - Este viaje a Hawai es mi 1o., 2o., 3o., 4o. etc.
 - La principal razón de este viaje es: (10 categorías para escoger entre las que se encontraban vacaciones, luna de miel, una convención).
 - Dónde voy a alojarme: (11 categorías entre las que se encontraban hotel, departamento, parientes, acampar).
 - Total de días en Hawai
 - ¿Cuál es la población que se estudia?
 - ¿El uso de un cuestionario es una buena manera de tener información de los pasajeros en los vuelos de llegada?
 - Diga de cada una de las cuatro preguntas si los datos que suministra son cualitativos o cuantitativos.
 - En la figura 1.8 se presenta una gráfica de barras que resume las ganancias de Volkswagen de los años 1997 a 2005 (*BusinessWeek*, 26 de diciembre de 2005).



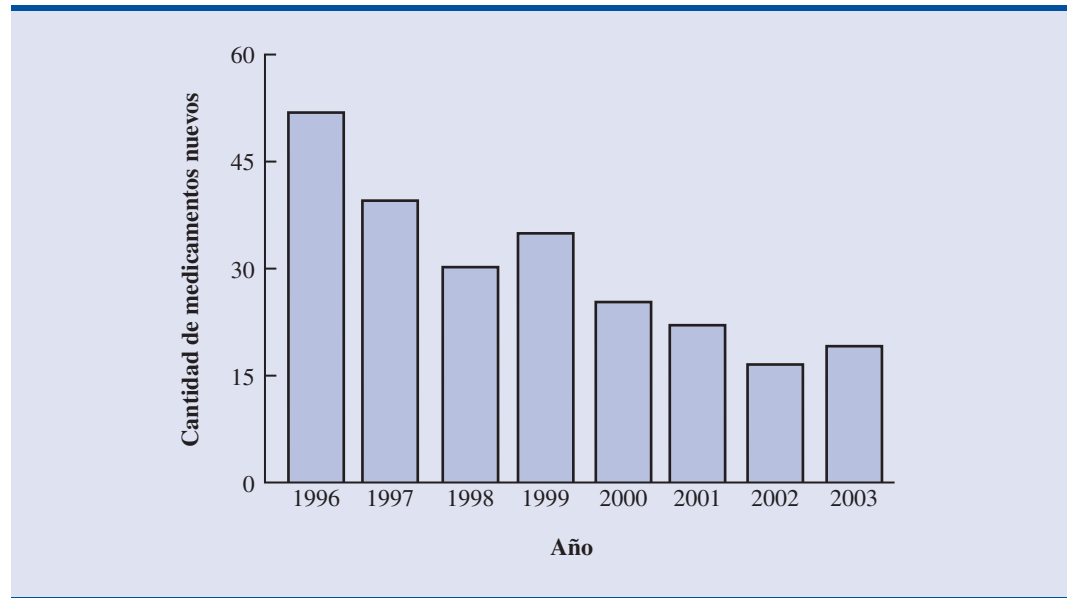
FIGURA 1.8 GANANCIAS DE VOLKSWAGEN



- a. ¿Estos son datos cualitativos o cuantitativos?
 - b. ¿Son datos de series de tiempo o datos de sección transversal?
 - c. ¿Cuál es la variable de interés?
 - d. Comente la tendencia en las ganancias de Volkswagen a lo largo del tiempo. El artículo de *BusinessWeek* (26 de diciembre de 2005) estimó las ganancias en 2006 en \$600 millones o \$0.6 mil millones. ¿Indica la figura si esta estimación parece ser razonable?
 - e. Un artículo similar que apareció en *BusinessWeek* el 23 de julio de 2001 sólo contaba con los datos de 1997 a 2000 junto con elevadas ganancias proyectadas para 2001. ¿Cómo era la perspectiva de las ganancias de Volkswagen en julio de 2001? En 2001, ¿parecía promotor invertir en Volkswagen? Explique.
 - f. ¿Qué advertencia sugiere esta gráfica acerca de la proyección de datos como los de las ganancias de Volkswagen hacia el futuro?
14. CSM Worldwide pronostica la producción mundial de todos los fabricantes de automóviles. Los datos siguientes de CSM muestran el pronóstico de la producción mundial para General Motors, Ford, DaimlerChrysler y Toyota para los años 2004 a 2007 (*USA Today*, 21 de diciembre de 2005). Estos datos están dados en millones de vehículos.

Fabricante	2004	2005	2006	2007
General Motors	8.9	9.0	8.9	8.8
Ford	7.8	7.7	7.8	7.9
DaimlerChrysler	4.1	4.2	4.3	4.6
Toyota	7.8	8.3	9.1	9.6

- a. Haga una gráfica de series de tiempo para los años 2004 a 2007 en la que se observe la cantidad de vehículos fabricados por cada empresa. Muestre las series de tiempo de los cuatro fabricantes en la misma gráfica.
 - b. General Motors ha sido sin discusión el principal fabricante de automóviles desde 1931. En esta gráfica de series de tiempo, ¿cuál es el mayor fabricante de automóviles? Explique.
 - c. Haga una gráfica que muestre los vehículos producidos por los fabricantes de automóviles usando los datos de 2007. ¿Está basada en datos de series de tiempo o en datos de sección transversal?
15. La Food and Drug Administration (FDA) da información sobre la cantidad de medicamentos aprobados en un periodo de ocho años (*The Wall Street Journal*, 12 de enero de 2004). En la figura 1.9 se presenta una gráfica de barras que resume el número de medicamentos nuevos aprobados cada año.
- a. ¿Estos datos son cualitativos o cuantitativos?
 - b. ¿Son datos de series de tiempo o son datos de sección transversal?
 - c. ¿Cuántos medicamentos fueron aprobados en 2003?
 - d. ¿En qué año se aprobaron menos medicamentos? ¿Cuántos fueron?
 - e. Presente un comentario sobre la tendencia en el número de medicamentos nuevos aprobados por la FDA en este periodo de ocho años.
16. El departamento de marketing de su empresa elabora un refresco dietético que dice captará una gran parte del mercado de adultos jóvenes.
- a. ¿Qué datos desearía ver antes de invertir una cantidad importante para introducir el nuevo producto en el mercado?
 - b. ¿Cómo esperaría que se obtuvieran los datos mencionados en el inciso a?
17. El directivo de una empresa grande recomienda un aumento de \$10 000 para evitar que un empleado se cambie a otra empresa. ¿Qué fuentes de datos internas y externas pueden usarse para decidir si es apropiado ese incremento de salario?

FIGURA 1.9 NÚMERO DE MEDICAMENTOS NUEVOS APROBADOS POR LA FDA

18. En una encuesta a 430 viajeros de negocios se encontró que 155 de ellos empleaban los servicios de un agente de viajes para la preparación de sus viajes (*USA Today*, 20 de noviembre de 2003).
 - a. Elabore una estadística descriptiva que sirva para estimar el porcentaje de viajeros de negocios que emplean un agente de viajes para preparar su viaje.
 - b. Con la encuesta se encontró que la manera más frecuente en que los viajeros de negocios hacen los preparativos de su viaje es mediante un sitio en línea. Si 4% de los viajeros de negocios encuestados hacen los preparativos de su viaje de esta manera, ¿cuántos de los 430 encuestados emplearon un sitio en línea?
 - c. Estos datos sobre cómo se hacen los preparativos, ¿son cualitativos o cuantitativos?
19. En un estudio sobre los suscriptores de *BusinessWeek* de Estados Unidos se recogen datos de una muestra de 2861 suscriptores. Cincuenta y nueve por ciento de los encuestados señalaron tener un ingreso de \$75 000 o más y 50% indicaron poseer una tarjeta de crédito de American Express.
 - a. ¿Cuál es la población de interés en este estudio?
 - b. ¿Es el ingreso anual un dato cualitativo o cuantitativo?
 - c. ¿Es la posesión de una tarjeta de crédito de American Express una variable cualitativa o cuantitativa?
 - d. ¿Hacer este estudio requiere datos de series de tiempo o de sección transversal?
 - e. Describa cualquier inferencia estadística posible para *BusinessWeek* con base en esta encuesta.
20. En una encuesta a 131 directores de inversión en Barron's se encontró lo siguiente (Barron's 28 de octubre de 2002):
 - De los dirigentes 43% se clasificaron como optimistas o muy optimistas sobre el mercado de acciones.
 - El rendimiento promedio esperado en los 12 meses siguientes en títulos de capital fue 11.2%.
 - La atención a la salud fue elegida por 21% como el sector con más probabilidad de ir a la cabeza del mercado en los próximos 12 meses.
 - Cuando se les preguntó cuánto tiempo se necesitaría para que las acciones de tecnología y telecomunicación recobraran un crecimiento sostenible, la respuesta promedio de los directivos fue 2.5 años.

- a. Cite dos estadísticas descriptivas.
 - b. Haga una inferencia sobre la población de todos los directivos de inversiones respecto al rendimiento promedio esperado en los títulos de capital durante los siguientes 12 meses.
 - c. Haga una inferencia acerca de la cantidad de tiempo que se necesitará para que las acciones de tecnología y telecomunicación recobren un crecimiento sostenible.
21. En una investigación médica que duró siete años se encontró que las mujeres cuyas madres habían tomado el medicamento DES durante el embarazo, tenían el doble de posibilidades de presentar anomalías en los tejidos que pudieran conducir a un cáncer, que aquellas cuyas madres no habían tomado este medicamento.
- a. En este estudio se compararon dos poblaciones. ¿Cuáles son?
 - b. ¿Es posible pensar que los datos se obtuvieron mediante una encuesta o mediante un experimento?
 - c. De la población de las mujeres cuyas madres habían tomado el medicamento DES durante el embarazo, se encontró que en una muestra de 3980 mujeres 63 presentaban anomalías en tejidos que podrían conducir a un cáncer. Dé un estadístico descriptivo útil para estimar el número de mujeres, de cada 1000, de esta población que pueden presentar anomalías en los tejidos.
 - d. De la población de mujeres cuyas madres no tomaron el medicamento DES durante el embarazo, ¿cuál es el número estimado de mujeres, de cada 1000, que pueden presentar anomalías en los tejidos?
 - e. Estudios médicos a menudo utilizan muestras grandes (en este caso, 3980). ¿Por qué?
22. En otoño de 2003, Arnold Schwarzenegger disputó al gobernador Gray Davis la gobernatura de California. En una encuesta realizada entre los votantes registrados se encontró que Arnold Schwarzenegger iba a la cabeza con un porcentaje estimado de 54% (*Newsweek*, 8 de septiembre de 2003).
- a. ¿Cuál fue la población en este estudio?
 - b. ¿Cuál fue la muestra en este estudio?
 - c. ¿Por qué se empleó una muestra en esta situación? Explique.
23. Nielsen Media Research realiza cada semana un sondeo entre los televidentes de Estados Unidos y publica datos tanto de índice de audiencia como de participación en el mercado. El índice de audiencia de Nielsen es el porcentaje de hogares que tienen televisión y que están viendo un programa, mientras que la participación de Nielsen es el porcentaje de hogares que están viendo un programa, entre los hogares que tiene la televisión en uso. Por ejemplo, los resultados de Nielsen Media Research para la Serie Mundial de Béisbol de 2003 entre los Yankees de Nueva York y los Marlins de Florida dieron un índice de audiencia de 12.8% y una participación de 22% (Associated Press, 27 de octubre de 2003). Por tanto, 12.8% de los hogares que tenían televisión estaban viendo la Serie Mundial y 22% de los hogares que estaban viendo la televisión, estaban viendo la Serie Mundial. A partir de los datos de índices de audiencia y de participación, Nielsen publica un ranking semanal de los programas de televisión así como un ranking semanal de las cuatro principales cadenas de televisión en Estados Unidos: ABC, CBS, NBC y Fox.
- a. ¿Qué trata de medir Nielsen Media Research?
 - b. ¿Cuál es la población?
 - c. ¿Por qué se usaría una muestra en esta situación?
 - d. ¿Qué tipo de decisiones o de acciones están basadas en los rankings de Nielsen?
24. En una muestra con cinco calificaciones de los estudiantes en un determinado examen los datos fueron: 72, 65, 82, 90, 76. ¿Cuáles de las afirmaciones siguientes son correctas y cuáles deben cuestionarse como una generalización excesiva?
- a. La calificación promedio de este examen en la muestra de las calificaciones de cinco estudiantes es 77.
 - b. La calificación promedio de todos los estudiantes en este examen es 77.
 - c. Una estimación para la calificación promedio de todos los estudiantes que hicieron el examen es 77.
 - d. Más de la mitad de los estudiantes que hicieron el examen tendrán calificaciones entre 70 y 85.
 - e. Si se incluyen en la muestra otros cinco estudiantes, sus calificaciones estarán entre 65 y 90.

TABLA 1.8 CONJUNTO DE DATOS DE 25 ACCIONES SHADOW

Empresa	Bolsa de valores	Denominación abreviada Symbol	Capacidad de mercado (millones de \$)	Relación precio/ganancia	Margen de ganancia bruta (%)
DeWolfe Companies	AMEX	DWL	36.4	8.4	36.7
North Coast Energy	OTC	NCEB	52.5	6.2	59.3
Hansen Natural Corp.	OTC	HANS	41.1	14.6	44.8
MarineMax, Inc.	NYSE	HZO	111.5	7.2	23.8
Nanometrics Incorporated	OTC	NANO	228.6	38.0	53.3
TeamStaff, Inc.	OTC	TSTF	92.1	33.5	4.1
Environmental Tectonics	AMEX	ETC	51.1	35.8	35.9
Measurement Specialties	AMEX	MSS	101.8	26.8	37.6
SEMCO Energy, Inc.	NYSE	SEN	193.4	18.7	23.6
Party City Corporation	OTC	PCTY	97.2	15.9	36.4
Embrex, Inc.	OTC	EMBX	136.5	18.9	59.5
Tech/Ops Sevcon, Inc.	AMEX	TO	23.2	20.7	35.7
ARCADIS NV	OTC	ARCAF	173.4	8.8	9.6
Qiao Xing Universal Tele.	OTC	XING	64.3	22.1	30.8
Energy West Incorporated	OTC	EWST	29.1	9.7	16.3
Barnwell Industries, Inc.	AMEX	BRN	27.3	7.4	73.4
Innodata Corporation	OTC	INOD	66.1	11.0	29.6
Medical Action Industries	OTC	MDCI	137.1	26.9	30.6
Instrumentarium Corp.	OTC	INMRY	240.9	3.6	52.1
Petroleum Development	OTC	PETD	95.9	6.1	19.4
Drexler Technology Corp.	OTC	DRXR	233.6	45.6	53.6
Gerber Childrenswear Inc.	NYSE	GCW	126.9	7.9	25.8
Gaiam, Inc.	OTC	GAIA	295.5	68.2	60.7
Artesian Resources Corp.	OTC	ARTNA	62.8	20.5	45.5
York Water Company	OTC	YORW	92.2	22.9	74.2



25. En la tabla 1.8 aparece un conjunto de datos con información sobre 25 de las acciones shadow vigiladas por la American Association of Individual Investors (aaii.com, febrero de 2002). Acciones shadow son acciones comunes de empresas pequeñas que no son estrechamente vigiladas por los analistas de Wall Street. Este conjunto de datos se encuentra también en el disco compacto que se incluye en este libro, en el archivo Shadow02.
- ¿Cuántas variables hay en este conjunto de datos?
 - ¿Qué variables son cualitativas y cuáles son cuantitativas?
 - Par la variable bolsa de valores muestre la frecuencia y la frecuencia porcentual de AMEX, NYSE y OTC. Construya una gráfica de barras como la de la figura 1.5.
 - Muestre la distribución de frecuencias del margen de ganancia bruta empleando cinco intervalos: 0–14.9, 15–29.9, 30–44.9, 45–59.9 y 60–74.9. Construya un histograma como el de la figura 1.6.
 - ¿Cuál es la proporción precio/ganancia promedio?



CAPÍTULO 2

Estadística descriptiva: presentaciones tabulares y gráficas

CONTENIDO

LA ESTADÍSTICA EN LA
PRÁCTICA: LA EMPRESA
COLGATE-PALMOLIVE

**2.1 RESUMEN DE DATOS
CUALITATIVOS**
Distribución de frecuencia
relativa y de frecuencia
porcentual
Gráficas de barra y gráficas
de pastel

**2.2 RESUMEN DE DATOS
CUANTITATIVOS**
Distribución de frecuencia
Distribuciones de frecuencia
relativa y de frecuencia
porcentual

Gráficas de puntos
Histograma
Distribuciones acumuladas
Ojiva

**2.3 ANÁLISIS EXPLORATORIO
DE DATOS: EL DIAGRAMA
DE TALLO Y HOJAS**

**2.4 TABULACIONES CRUZADAS
Y DIAGRAMAS DE
DISPERSIÓN**
Tabulación cruzada
Paradoja de Simpson
Diagrama de dispersión y línea
de tendencia

LA ESTADÍSTICA en LA PRÁCTICA

LA EMPRESA COLGATE-PALMOLIVE* NUEVA YORK, NUEVA YORK

La empresa Colgate-Palmolive empezó en la Ciudad de Nueva York en 1806 como una pequeña tienda de jabones y velas. Hoy, Colgate-Palmolive emplea más de 4000 personas que trabajan en 200 países y territorios del mundo. Aunque es más conocida por sus marcas Colgate, Palmolive, Ajax y Fab, la empresa comercializa los productos Mennen, Hill's Science Diet y Hill's Prescription Diet.

La empresa Colgate-Palmolive aplica la estadística en su programa de aseguramiento de la calidad en los detergentes caseros para la ropa. Le interesa la satisfacción del cliente con la cantidad de detergente en los paquetes. Todos los paquetes de cierto tamaño se llenan con la misma cantidad de detergente en peso, aunque el volumen del detergente varía de acuerdo con la densidad del polvo detergente. Por ejemplo, si la densidad del detergente es alta, se necesita una cantidad menor de detergente para tener el peso señalado en el paquete. El resultado es que cuando el cliente abre el paquete le parece que no ha sido bien llenado.

Para controlar el problema del peso del polvo de detergente, se han establecido límites en el nivel aceptable de la densidad del polvo. Con periodicidad se toman muestras estadísticas y se mide la densidad de la muestra de polvo. Los resúmenes de los datos se les proporcionan a los operarios para que de ser necesario lleven a cabo acciones correctivas, de manera que la densidad se mantenga dentro de las especificaciones de calidad establecidas.

En la tabla y figura adjuntas se presentan una distribución de frecuencia y un histograma obtenidos con 150 muestras tomadas en una semana. Densidades mayores a 0.40 son inaceptablemente altas. De acuerdo con la distribución de frecuencia y al histograma la operación satisface los lineamientos de calidad ya que todas las densidades son menores o iguales a 0.40. A la vista de estos resúmenes estadísticos los directivos estarán satisfechos con la calidad del proceso de producción de detergente.

En este capítulo se estudiarán métodos tabulares y gráficos de la estadística descriptiva como distribuciones de frecuencia, gráficas de barras, histogramas, diagramas de tallo y hoja, tabulaciones cruzadas y otros. El objeto de



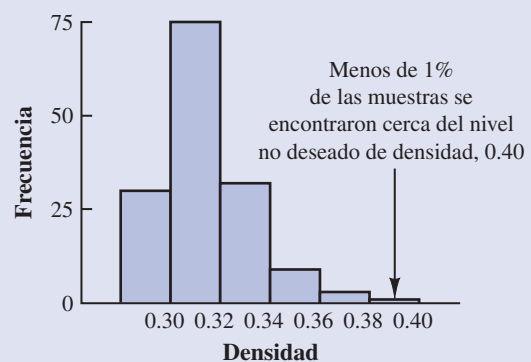
Los resúmenes estadísticos ayudan a mantener la calidad de estos productos de Colgate-Palmolive
© Joe Higgins/South Western.

estos métodos es resumir los datos de manera que sean entendibles e interpretables con facilidad.

Distribución de frecuencia de los datos de densidad

Densidad	Frecuencia
0.29–0.30	30
0.31–0.32	75
0.33–0.34	32
0.35–0.36	9
0.37–0.38	3
0.39–0.40	1
Total	150

Histograma de los datos de densidad



*Los autores agradecen a William R. Fawle, director de aseguramiento de la calidad de la empresa Colgate-Palmolive por proporcionarles este artículo para *La estadística en la práctica*.

Como se indicó en el capítulo 1, los datos se clasifican en cualitativos o cuantitativos. Los **datos cualitativos** emplean etiquetas o nombres para determinar categorías de elementos iguales. Los **datos cuantitativos** son números que indican cuánto o cuántos.

En este capítulo se presentan los métodos tabulares y gráficos empleados para datos cualitativos y cuantitativos. Los resúmenes gráficos o tabulares de datos se encuentran en reportes anuales, en artículos en los periódicos y en estudios de investigación. Todo mundo se encuentra con este tipo de presentaciones. Por tanto, es útil saber cómo se hacen y se interpretan. Se empezará con los métodos tabulares y gráficos para resumir datos que se refieren a una sola variable. En la última sección se introducen los métodos para resumir datos cuando lo que interesa es la relación entre dos variables.

Los paquetes modernos de software para estadística proporcionan muchas posibilidades para resumir datos y elaborar presentaciones gráficas. Minitab y Excel son dos paquetes muy empleados. En los apéndices de este capítulo se muestran algunas de sus posibilidades.

2.1

Resumen de datos cualitativos

Distribución de frecuencia

Conviene iniciar el estudio acerca del uso de los métodos tabulares y gráficos para resumir datos cualitativos con la definición de **distribución de frecuencia**.

DISTRIBUCIÓN DE FRECUENCIA

Una distribución de frecuencia es un resumen tabular de datos que muestra el número (frecuencia) de elementos en cada una de las diferentes clases disyuntas (que no se superponen).

Con el ejemplo siguiente se muestra la elaboración e interpretación de una distribución de frecuencia de datos cualitativos. Cinco refrescos muy conocidos son Coca cola clásica (Coke Classic), Coca cola de dieta (Diet Coke), Dr. Pepper, Pepsi y Sprite. Suponga que los datos de la tabla 2.1 muestran los refrescos que fueron comprados en una muestra de 50 ventas de refresco.

TABLA 2.1 DATOS DE UNA MUESTRA DE 50 VENTAS DE REFRESCO

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	



TABLA 2.2

DISTRIBUCIÓN DE
FRECUENCIA DE
LAS VENTAS DE
REFRESCO

Refresco	Frecuencia
Coke Classic	19
Diet Coke	8
Dr. Pepper	5
Pepsi	13
Sprite	5
Total	50

Para elaborar una distribución de frecuencia con estos datos, se cuenta el número de veces que aparece cada refresco en la tabla 2.1. La Coca cola clásica (Coke Classic) aparece 19 veces, la Coca cola de dieta (Diet Coke) 8 veces, Dr. Pepper 5 veces, Pepsi 13 veces y Sprite 5 veces. Esto queda resumido en la distribución de frecuencia de la tabla 2.2.

Esta distribución de frecuencia proporciona un resumen de cómo se distribuyeron las 50 ventas entre los cinco refrescos. El resumen aporta más claridad que los datos originales de la tabla 2.1. Al observar esta distribución de frecuencia, es claro que Coca cola clásica es el refresco que más se vende, Pepsi el segundo, Coca cola de dieta el tercero y Sprite y Dr. Pepper están empatados en el cuarto lugar. La distribución de frecuencia resume la información sobre la popularidad de los cinco refrescos.

Distribuciones de frecuencia relativa y de frecuencia porcentual

En una distribución de frecuencia se aprecia el número (frecuencia) de los elementos de cada una de las diversas clases disjuntas. Sin embargo, con frecuencia lo que interesa es la proporción o porcentaje de elementos en cada clase. La *frecuencia relativa* de una clase es igual a la parte o proporción de los elementos que pertenecen a cada clase. En un conjunto de datos, en el que hay n observaciones, la frecuencia relativa de cada clase se determina como sigue:

FRECUENCIA RELATIVA

$$\text{Frecuencia relativa de una clase} = \frac{\text{Frecuencia de la clase}}{n} \quad (2.1)$$

La *frecuencia porcentual* de una clase es la frecuencia relativa multiplicada por 100.

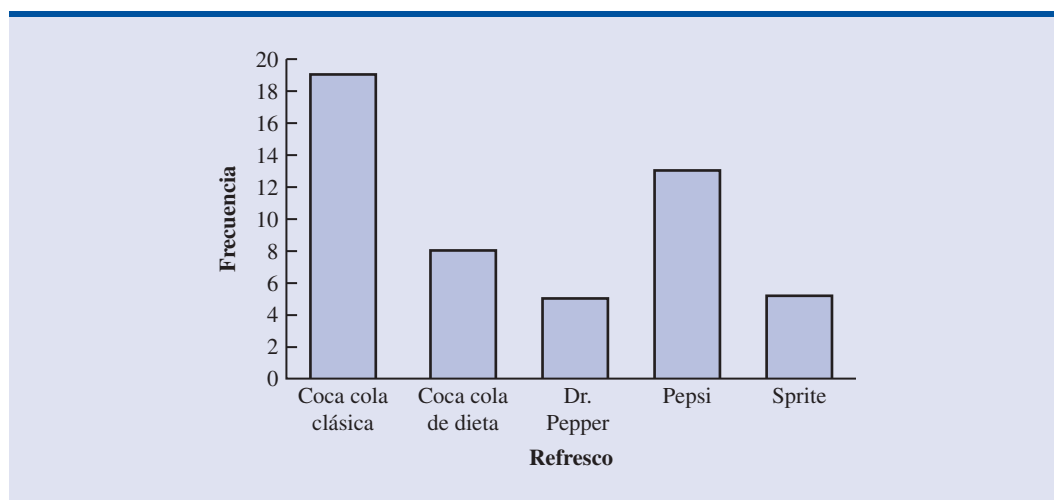
Una **distribución de frecuencia relativa** da un resumen tabular de datos en el que se muestra la frecuencia relativa de cada clase. Una **distribución de frecuencia porcentual** da la frecuencia porcentual de los datos de cada clase. En la tabla 2.3 se presenta una distribución de frecuencia relativa y una distribución de frecuencia porcentual de los datos de los refrescos. En esta tabla se observa que la frecuencia relativa de la Coca cola clásica es $19/50 = 0.38$, la de la Coca cola de dieta es $8/50 = 0.16$, etc. En la distribución de frecuencia porcentual, se muestra que 38% de las ventas fueron de Coca cola clásica, 16% de Coca cola de dieta, etc. También resulta que $38\% + 26\% + 16\% = 80\%$ de las ventas fueron de los tres refrescos que más se venden.

Gráficas de barra y gráficas de pastel

Una **gráfica de barras** o un diagrama de barras, es una gráfica para representar los datos cualitativos de una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual. En uno de los ejes de la gráfica (por lo general en el horizontal), se especifican las etiquetas empleadas para las clases (categorías). Para el otro eje de la gráfica (el vertical) se usa una escala para

TABLA 2.3 DISTRIBUCIONES DE FRECUENCIA RELATIVA Y FRECUENCIA PORCENTUAL DE LAS VENTAS DE REFRESCOS

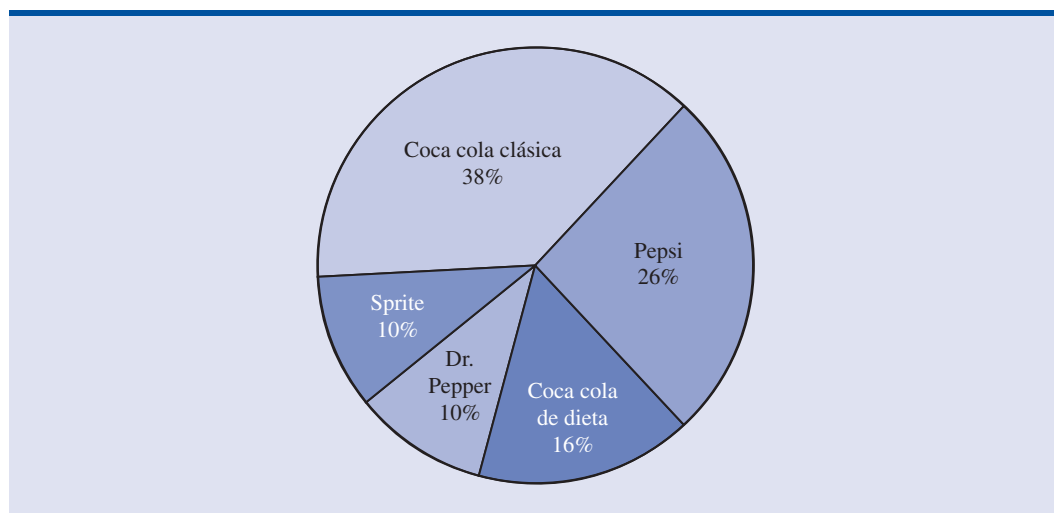
Refresco	Frecuencia relativa	Frecuencia porcentual
Coke Classic	0.38	38
Diet Coke	0.16	16
Dr. Pepper	0.10	10
Pepsi	0.26	26
Sprite	0.10	10
Total	1.00	100

FIGURA 2.1 GRÁFICA DE BARRAS PARA LAS VENTAS DE REFRESCOS

En el control de calidad, las gráficas de barras se usan para identificar las principales causas de problemas. Las graficas se acomodan en orden de alturas descendentes de izquierda a derecha colocando primero la causa de frecuencia más común en primer lugar. A esta gráfica de barras se le llama diagrama de Pareto en honor a su inventor Wilfredo Pareto, un economista italiano.

frecuencia, frecuencia relativa o frecuencia porcentual. Después, empleando un ancho de barra fijo, se dibuja sobre cada etiqueta de las clases una barra que se extiende hasta la frecuencia, frecuencia relativa o frecuencia porcentual de la clase. Cuando se tienen datos cualitativos, las barras deben estar separadas para hacer énfasis en que cada clase está separada. En la figura 2.1 se muestra una gráfica de barras correspondiente a la distribución de frecuencia de las 50 ventas de refrescos. Advierta cómo en esta representación gráfica se observa que Coca cola clásica, Pepsi y Coca cola de dieta son los refrescos preferidos.

La **gráfica de pastel** proporciona otra gráfica para presentar distribuciones de frecuencia relativa y de frecuencia porcentual de datos cualitativos. Para elaborar una gráfica de pastel, primero se dibuja un círculo que representa todos los datos. Después se usa la frecuencia relativa para subdividir el círculo en sectores, o partes, que corresponden a la frecuencia relativa de cada clase. Por ejemplo, como un círculo tiene 360 grados y Coca cola clásica presenta una frecuencia relativa de 0.38, el sector de la gráfica de pastel correspondiente a Coca cola clásica resultará de $0.38(360) = 136.8$ grados. El sector del pastel para Coca cola de dieta constará de

FIGURA 2.2 GRÁFICA DE PASTEL PARA LAS VENTAS DE REFRESCOS

$0.16(360) = 57.6$ grados. Mediante cálculos semejantes para las demás clases se obtiene la gráfica de pastel de la figura 2.2. Los números que aparecen en cada sector pueden ser frecuencia, frecuencia relativa o frecuencia porcentual.

NOTAS Y COMENTARIOS

1. A menudo el número de clases en una distribución de frecuencia es el mismo que el número de categorías encontradas en los datos, como en los datos de las ventas de refresco en esta sección. Los datos comprenden cinco refrescos y para cada uno se definió una clase en la distribución de frecuencia. Si los datos incluyeran todos los refrescos se requerirían muchas categorías, la mayor parte de las cuales sólo tendrían muy pocas ventas. La mayoría de los profesionistas de la estadística aconsejan que las clases con frecuencia pequeña, se agrupen en una sola clase a la que se le llama “otros”. Cualquier clase con 5% o menos se trata de esta manera.
2. La suma de las frecuencias en una distribución de frecuencia es siempre igual al número de observaciones. La suma de las frecuencias relativas en una distribución de frecuencia relativa es siempre igual a 1.00, y la suma de los porcentajes en una distribución de frecuencia porcentual es siempre igual a 100.

Ejercicios

Métodos

1. Como respuesta a una pregunta hay tres alternativas: A, B y C. En una muestra de 120 respuestas, 60 fueron A, 24 B y 36 C. Dé las distribuciones de frecuencia y de frecuencia relativa.
2. Se da una distribución de frecuencia relativa.

Clase	Frecuencia relativa
A	0.22
B	0.18
C	0.40
D	

- a. ¿Cuál es la frecuencia relativa de la clase D?
 - b. El tamaño de la muestra es 200. ¿Cuál es la frecuencia de la clase D?
 - c. Muestre la distribución de frecuencia.
 - d. Dé la distribución de frecuencia porcentual.
3. Un cuestionario proporciona como respuestas 58 Sí, 42 No y 20 ninguna opinión.
 - a. En la construcción de una gráfica de pastel, ¿cuántos grados le corresponderán del pastel a la respuesta Sí?
 - b. ¿Cuántos grados le corresponderán del pastel a la respuesta No?
 - c. Construya una gráfica de pastel.
 - d. Construya una gráfica de barras.

Autoexamen

Aplicaciones

4. Los cuatro programas con horario estelar de televisión son *CSI*, *ER*, *Everybody Loves Raymond* y *Friends* (Nielsen Media Research, 11 de enero de 2004). A continuación se presentan los datos sobre las preferencias de los 50 televidentes de una muestra.

CSI	Friends	CSI	CSI	CSI
CSI	CSI	Raymond	ER	ER
Friends	CSI	ER	Friends	CSI
ER	ER	Friends	CSI	Raymond
CSI	Friends	CSI	CSI	Friends
ER	ER	ER	Friends	Raymond
CSI	Friends	Friends	CSI	Raymond
Friends	Friends	Raymond	Friends	CSI
Raymond	Friends	ER	Friends	CSI
CSI	ER	CSI	Friends	ER

- ¿Estos datos son cualitativos o cuantitativos?
 - Proporcione las distribuciones de frecuencia y de frecuencia relativa.
 - Construya una gráfica de barras y una gráfica de pastel.
 - De acuerdo con la muestra, ¿qué programa de televisión tiene la mayor audiencia? ¿Cuál es el segundo?
5. Los cinco apellidos más comunes en Estados Unidos, en orden alfabético son, Brown, Davis, Johnson, Jones, Smith y Williams (*The World Almanac, 2006*). Suponga que en una muestra de 50 personas con uno de estos apellidos se obtienen los datos siguientes.



Brown	Williams	Williams	Williams	Brown
Smith	Jones	Smith	Johnson	Smith
Davis	Smith	Brown	Williams	Johnson
Johnson	Smith	Smith	Johnson	Brown
Williams	Davis	Johnson	Williams	Johnson
Williams	Johnson	Jones	Smith	Brown
Johnson	Smith	Smith	Brown	Jones
Jones	Jones	Smith	Smith	Davis
Davis	Jones	Williams	Davis	Smith
Jones	Johnson	Brown	Johnson	Davis

Resuma estos datos construyendo:

- Distribuciones de frecuencia relativa y porcentual.
 - Una gráfica de barras.
 - Una gráfica de pastel.
 - De acuerdo con estos datos, ¿cuáles son los tres apellidos más comunes?
6. El índice de audiencia de televisión de Nielsen Media Research mide el porcentaje de personas que tienen televisión y que están viendo un determinado programa. El programa de televisión con el mayor índice de audiencia en la historia de la televisión (en Estados Unidos) fue *M*A*S*H Last Episode Special* transmitido el 28 de febrero de 1983. El índice de audiencia de 60.2 indicó que 60.2% de todas las personas que tenían televisión estaban viendo este programa. Nielsen Media Research publicó la lista de los 50 programas de televisión con los mayores índices de audiencia en la historia de la televisión (*The New York Times Almanac, 2006*). Los datos siguientes presentan las cadenas de televisión que produjeron estos 50 programas con mayor índice de audiencia.



ABC	ABC	ABC	NBC	CBS
ABC	CBS	ABC	ABC	NBC
NBC	NBC	CBS	ABC	NBC
CBS	ABC	CBS	NBC	ABC
CBS	NBC	NBC	CBS	NBC
CBS	CBS	CBS	NBC	NBC
FOX	CBS	CBS	ABC	NBC
ABC	ABC	CBS	NBC	NBC
NBC	CBS	NBC	CBS	CBS
ABC	CBS	ABC	NBC	ABC

- Con estos datos construya una distribución de frecuencia, una de frecuencia porcentual y una gráfica de barras.

Autoexamen

- b. ¿Cuál o cuáles cadenas de televisión han presentado los programas de mayor índice de audiencia? Compare los desempeños de ABC, CBS y NBC.
7. Un restaurante de Florida emplea cuestionarios en los que pide a sus clientes que evalúen el servicio, la calidad de los alimentos, los cocteles, los precios y la atmósfera del restaurante. Cada uno de estos puntos se evalúa con una escala de óptimo (O), muy bueno (V), bueno (G), regular (A) y malo (P). Emplee la estadística descriptiva para resumir los datos siguientes respecto a la calidad de los alimentos. ¿Qué piensa acerca de la evaluación de la calidad de los alimentos de este restaurante?

G	O	V	G	A	O	V	O	V	G	O	V	A
V	O	P	V	O	G	A	O	O	O	G	O	V
V	A	G	O	V	P	V	O	O	G	O	O	V
O	G	A	O	V	O	O	G	V	A	G		

8. A continuación se muestran datos de 55 miembros de un equipo de béisbol. Cada observación indica la posición principal que juegan los miembros del equipo: *pitcher* (P), *catcher* (H), primera base (1), segunda base (2), tercera base (3), shortstop (S), left field (L), center field (C) y right field (R).

L	P	C	H	2	P	R	1	S	S	1	L	P	R	P
P	P	P	R	C	S	L	R	P	C	C	P	P	R	P
2	3	P	H	L	P	1	C	P	P	P	S	1	L	R
R	1	2	H	S	3	H	2	L	P					

- a. Para resumir estos datos use una distribución de frecuencia y otra de frecuencia relativa.
- b. ¿Cuál es la posición que ocupan más miembros del equipo?
- c. ¿Cuál es la posición que ocupan menos miembros del equipo?
- d. ¿Qué posición de campo (L, R, C) es la que juegan más miembros del equipo?
- e. Compare las posiciones L, R, y C con las posiciones 1, 2, 3 y S.
9. Cerca del 60% de las empresas pequeñas y medianas son empresas familiares. En un estudio de TEC International se preguntaba al gerente general (CEO, por sus siglas en inglés) cómo había llegado a ese cargo (*The Wall Street Journal*, 16 de diciembre de 2003). Las respuestas fueron que el CEO heredó el negocio, que el CEO formó la empresa o que el CEO estaba contratado por con la empresa. En una muestra de 26 CEOs de empresas familiares, los datos obtenidos acerca de cómo el CEO había llegado a ese puesto fueron los siguientes:

Formó	Formó	Formó	Heredó
Heredó	Formó	Heredó	Formó
Heredó	Formó	Formó	Formó
Formó	Contrató	Contrató	Contrató
Heredó	Heredó	Heredó	Formó
Formó	Formó	Formó	Contrató
Formó	Heredó		

- a. Dé una distribución de frecuencias.
- b. Dé una distribución de frecuencias porcentuales.
- c. Presente una gráfica de barras.
- d. ¿Qué porcentaje de los CEOs de empresas familiares llegaron a ese puesto por heredar la empresa? ¿Cuál es la razón principal por la que una persona llega al puesto de CEO en una empresa familiar?
10. Netflix, Inc., de San José California, renta, por correo, más de 50 000 títulos de DVD. Los clientes ordenan en línea los DVDs que deseen ver. Antes de ordenar un DVD, el cliente puede ver una descripción del mismo y, si así lo desea, un resumen de las evaluaciones del mismo. Netflix emplea un sistema de evaluación de cinco estrellas que tienen el significado siguiente:

1 estrella	Me disgustó
2 estrellas	No me disgustó
3 estrellas	Me gustó
4 estrellas	Me gustó mucho
5 estrellas	Me fascinó

Dieciocho críticos, entre los que se encontraban Roger Ebert de *Chicago Sun Times* y Ty Burr de *Boston Globe*, proporcionaron evaluaciones en Hispanoamérica de la película *Batman inicia* (Netflix.com, 1 de marzo de 2006). Las evaluaciones fueron las siguientes:

4, 2, 5, 2, 4, 3, 3, 4, 4, 3, 4, 4, 2, 4, 4, 5, 4

- Diga por qué son cualitativos estos datos.
- Dé una distribución de frecuencias y una distribución de frecuencia relativa.
- Dé una gráfica de barras.
- Haga un comentario sobre las evaluaciones que dieron los críticos a esta película.

2.2

Resumen de datos cuantitativos

Distribución de frecuencia

TABLA 2.4

AUDITORÍA ANUAL
(DÍAS DE DURACIÓN)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Como se definió en la sección 2.1, una distribución de frecuencia es un resumen de datos tabular que presenta el número de elementos (frecuencia) en cada una de las clases disyuntas. Esta definición es válida tanto para datos cualitativos como cuantitativos. Sin embargo, cuando se trata de datos cuantitativos se debe tener más cuidado al definir las clases disyuntas que se van a usar en la distribución de frecuencia.

Considere, por ejemplo, los datos cuantitativos de la tabla 2.4. En esta tabla se presenta la duración en días de una muestra de auditorías de fin de año de 20 clientes de una empresa pequeña de contadores públicos. Los tres pasos necesarios para definir las clases de una distribución de frecuencia con datos cuantitativos son

- Determinar el número de clases disyuntas.
- Determinar el ancho de cada clase
- Determinar los límites de clase.

Se mostrarán estos pasos elaborando una distribución de frecuencia con los datos de la tabla 2.4.



Número de clases Las clases se forman especificando los intervalos que se usarán para agrupar los datos. Se recomienda emplear entre 5 y 20 clases. Cuando los datos son pocos, cinco o seis clases bastan para resumirlos. Si son muchos, se suele requerir más clases. La idea es tener las clases suficientes para que se muestre la variación en los datos, pero no deben ser demasiadas si algunas de ellas contienen sólo unos cuantos datos. Como el número de datos en la tabla 2.4 es relativamente pequeña ($n = 20$), se decide elaborar una distribución de frecuencia con cinco clases.

Ancho de clase El segundo paso al construir una distribución de frecuencia para datos cuantitativos es elegir el ancho de las clases. Como regla general es recomendable que el ancho sea el mismo para todas las clases. Así, el ancho y el número de clases no son decisiones independientes. Entre mayor sea el número de clases menor es el ancho de las clases y viceversa. Para determinar el ancho de clase apropiada se empieza por identificar el mayor y el menor de los valores de los datos. Después, usando el número de clases deseado, se emplea la expresión siguiente para determinar el ancho aproximada de clase.

$$\text{Ancho aproximada de clase} = \frac{\text{Valor mayor en los datos} - \text{Valor menor en los datos}}{\text{Número de clase}} \quad (2.2)$$

El ancho aproximado de clase que se obtiene con la ecuación (2.2) se redondea a un valor más adecuado de acuerdo con las preferencias de la persona que elabora la distribución de frecuencia. Por ejemplo, si el ancho de clase aproximado es 9.28, se redondea a 10 porque 10 es un ancho de clase más adecuado para la presentación de la distribución de la frecuencia.

En los datos sobre las duraciones de las auditorías de fin de año el valor mayor en los datos es 33 y el valor menor es 12. Como se ha decidido resumir los datos en cinco clases, empleando

Hacer las clases de una misma amplitud reduce la posibilidad de que los usuarios hagan interpretaciones inapropiadas.

No hay una distribución de frecuencia que sea la mejor para un conjunto de datos. Distintas personas elaboran diferentes, pero igual de aceptables, distribuciones de frecuencia para un conjunto de datos dado. El objetivo es hacer notar el agrupamiento y la variación natural de los datos.

TABLA 2.5

DISTRIBUCIÓN DE FRECUENCIA DE LAS AUDITORÍAS

Duración de las auditorías (días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

la ecuación (2.2) el ancho aproximado de clase que se obtiene es $(33 - 12)/5 = 4.2$. Por tanto, al redondear, en la distribución de frecuencia se usa como ancho de clase cinco días.

En la práctica el número de clases y su ancho adecuado se determinan por prueba y error. Una vez que se elige una determinado número de clases, se emplea la ecuación 2.2 para determinar el ancho aproximado de clase. El proceso se repite con distintos números de clases. El analista determina la combinación de número y ancho de clases que le proporciona la mejor distribución de frecuencia para resumir los datos.

En el caso de los datos de la tabla 2.4, una vez que se ha decidido emplear cinco clases, cada una con ancho de cinco días, el paso siguiente es especificar los límites de cada clase.

Límites de clase Los límites de clase deben elegirse de manera que cada dato pertenezca a una y sólo una de las clases. El *límite de clase inferior* indica el menor valor de los datos a que pertenece esa clase. El *límite de clase superior* indica el mayor valor de los datos a que pertenece esa clase. Al elaborar distribuciones de frecuencia para datos cualitativos, no es necesario especificar límites de clase porque cada dato corresponde de manera natural a una de las clases disjuntas. Pero con datos cuantitativos, como la duración de las auditorías de la tabla 2.4, los límites de clase son necesarios para determinar dónde colocar cada dato.

Mediante los datos de la duración de las auditorías de la tabla 2.4, se elige 10 días como límite inferior y 14 como límite superior de la primera clase. En la tabla 2.5, esta clase se denota como 10–14. El valor menor, 12 (de la tabla), pertenece a la clase 10–14. Después se elige 15 días como límite inferior y 19 como límite superior de la clase siguiente. Así, se continúan definiendo los límites inferior y superior de las clases hasta tener las cinco clases: 10–14, 15–19, 20–24, 25–29 y 30–34. El valor mayor en los datos, 33, pertenece a la clase 30–34. Las diferencias entre los límites inferiores de clase de clases adyacentes es el ancho de clase. Con los dos primeros límites inferiores de clase, 10 y 15, se ve que el ancho de clase es $15 - 10 = 5$.

Una vez determinados números, ancho y límites de las clases, la distribución de frecuencia se obtiene contando el número de datos que corresponden a cada clase. Por ejemplo, en la tabla 2.4 se observa que hay cuatro valores, 12, 14, 14 y 13, que pertenecen a la clase 10–14. Por tanto, la frecuencia de la clase 10–14 es 4. Al continuar con este proceso de conteo para las clases 15–19, 20–24, 25–29 y 30–34 se obtiene la distribución de frecuencia que se muestra en la tabla 2.5. En esta distribución de frecuencia se observa lo siguiente:

1. Las duraciones de las auditorías que se presentan con más frecuencia son de la clase 15–19 días. Ocho de las 20 auditorías caen en esta clase.
2. Sólo una auditoría requirió 30 o más días.

También se obtienen otras conclusiones, dependiendo de los intereses de quien observa la distribución de frecuencia. La utilidad de una distribución de frecuencia es que proporciona claridad acerca de los datos, la cual no es fácil de obtener con la forma desorganizada de éstos.

Punto medio de clase En algunas aplicaciones se desea conocer el punto medio de las clases de una distribución de frecuencia de datos cuantitativos. El **punto medio de clase** es el valor que queda a la mitad entre el límite inferior y el límite superior de la clase. En el caso de las duraciones de las auditorías, los cinco puntos medios de clase son 12, 17, 22, 27 y 32.

Distribuciones de frecuencia relativa y de frecuencia porcentual

Las distribuciones de frecuencia relativa y de frecuencia porcentual para datos cuantitativos se definen de la misma forma que para datos cualitativos. Primero debe recordar que la frecuencia relativa es el cociente, respecto al total de observaciones, de las observaciones que pertenecen a una clase. Si el número de observaciones es n ,

$$\text{Frecuencia relativa de la clase} = \frac{\text{Frecuencia de la clase}}{n}$$

La frecuencia porcentual de una clase es la frecuencia relativa multiplicada por 100.

Con base en la frecuencia de las clases de la tabla 2.5 y dado que $n = 20$, en la tabla 2.6 se muestran las distribuciones de frecuencia relativa y de frecuencia porcentual de los datos de las

TABLA 2.6 DISTRIBUCIONES DE FRECUENCIA RELATIVA Y DE FRECUENCIA PORCENTUAL CON LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de las auditorías (días)	Frecuencia relativa	Frecuencia porcentual
10–14	0.20	20
15–19	0.40	40
20–24	0.25	25
25–29	0.10	10
30–34	0.05	5
Total	1.00	100

duraciones de las auditorías. Observe que 0.40 de las auditorías, o 40%, necesitaron entre 15 y 19 días. Sólo 0.05%, o 5%, requirió 30 o más días. De nuevo, hay más interpretaciones o ideas que se obtienen de la tabla 2.6.

Gráficas de puntos

Uno de los más sencillos resúmenes gráficos de datos son las **gráficas de puntos**. En el eje horizontal se presenta el intervalo de los datos. Cada dato se representa por un punto colocado sobre este eje. La figura 2.3 es la gráfica de puntos de los datos de la tabla 2.4. Los tres puntos que se encuentran sobre el 18 del eje horizontal indican que hubo tres auditorías de 18 días. Las gráficas de puntos muestran los detalles de los datos y son útiles para comparar la distribución de los datos de dos o más variables.

Histograma

Una presentación gráfica usual para datos cuantitativos es el **histograma**. Esta gráfica se hace con datos previamente resumidos mediante una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual. Un histograma se construye colocando la variable de interés en el eje horizontal y la frecuencia, la frecuencia relativa o la frecuencia porcentual en el eje vertical. La frecuencia, frecuencia relativa o frecuencia porcentual de cada clase se indica dibujando un rectángulo cuya base está determinada por los límites de clase sobre el eje horizontal y cuya altura es la frecuencia, la frecuencia relativa o la frecuencia porcentual correspondiente.

La figura 2.4 es un histograma de las duraciones de las auditorías. Observe que la clase con mayor frecuencia se indica mediante el rectángulo que se encuentra sobre la clase 15–19 días. La altura del rectángulo muestra que la frecuencia de esta clase es 8. Un histograma de las distribuciones de frecuencia relativa o porcentual de estos datos se ve exactamente igual que el histograma de la figura 2.4, excepto que en el eje vertical se colocan los valores de frecuencia relativa o porcentual.

Como se muestra en la figura 2.4, los rectángulos adyacentes de un histograma se tocan uno a otro. A diferencia de las gráficas de barras, en un histograma no hay una separación natural en-

FIGURA 2.3 GRÁFICA DE PUNTOS PARA LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS

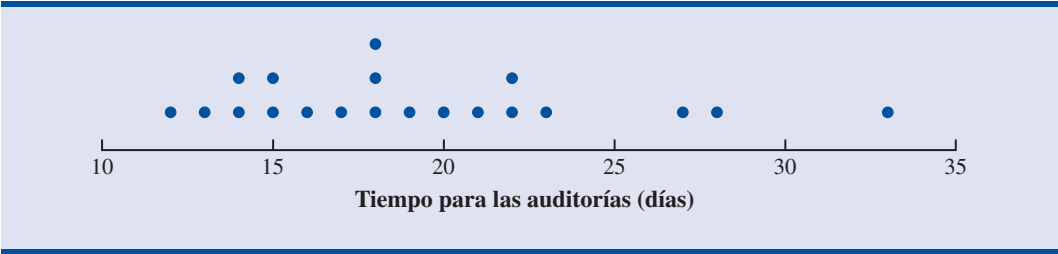
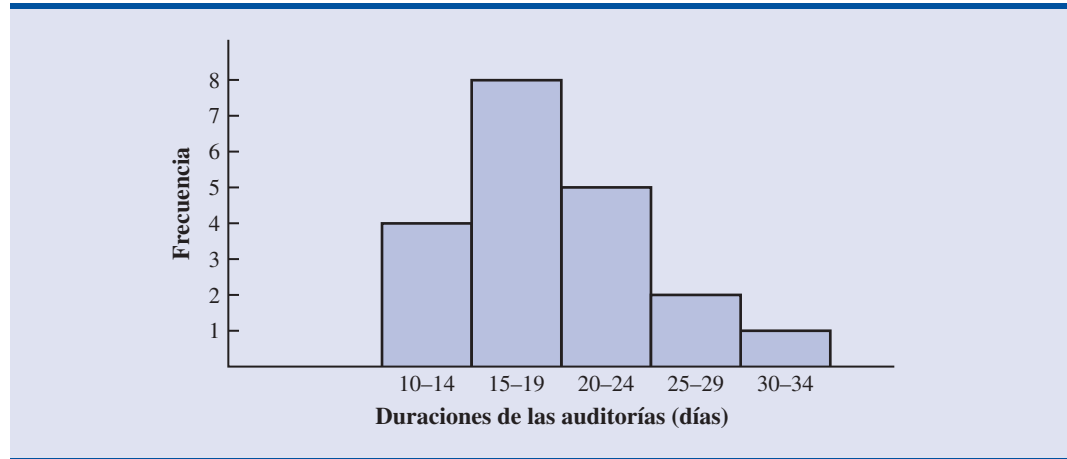


FIGURA 2.4 HISTOGRAMA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS

tre los rectángulos de clases adyacentes. Este formato es el usual para histogramas. Como las clases de las duraciones de las auditorías son 10–14, 15–19, 20–24, 25–29 y 30–34 parecería que se necesitara una unidad de espacio entre las clases, de 14 a 15, de 19 a 20, de 24 a 25 y de 29 a 30. Cuando se construye un histograma se eliminan estos espacios. Eliminar los espacios entre las clases del histograma de las duraciones de las auditorías sirve para indicar que todos los valores entre el límite inferior de la primera clase y el superior de la última son posibles.

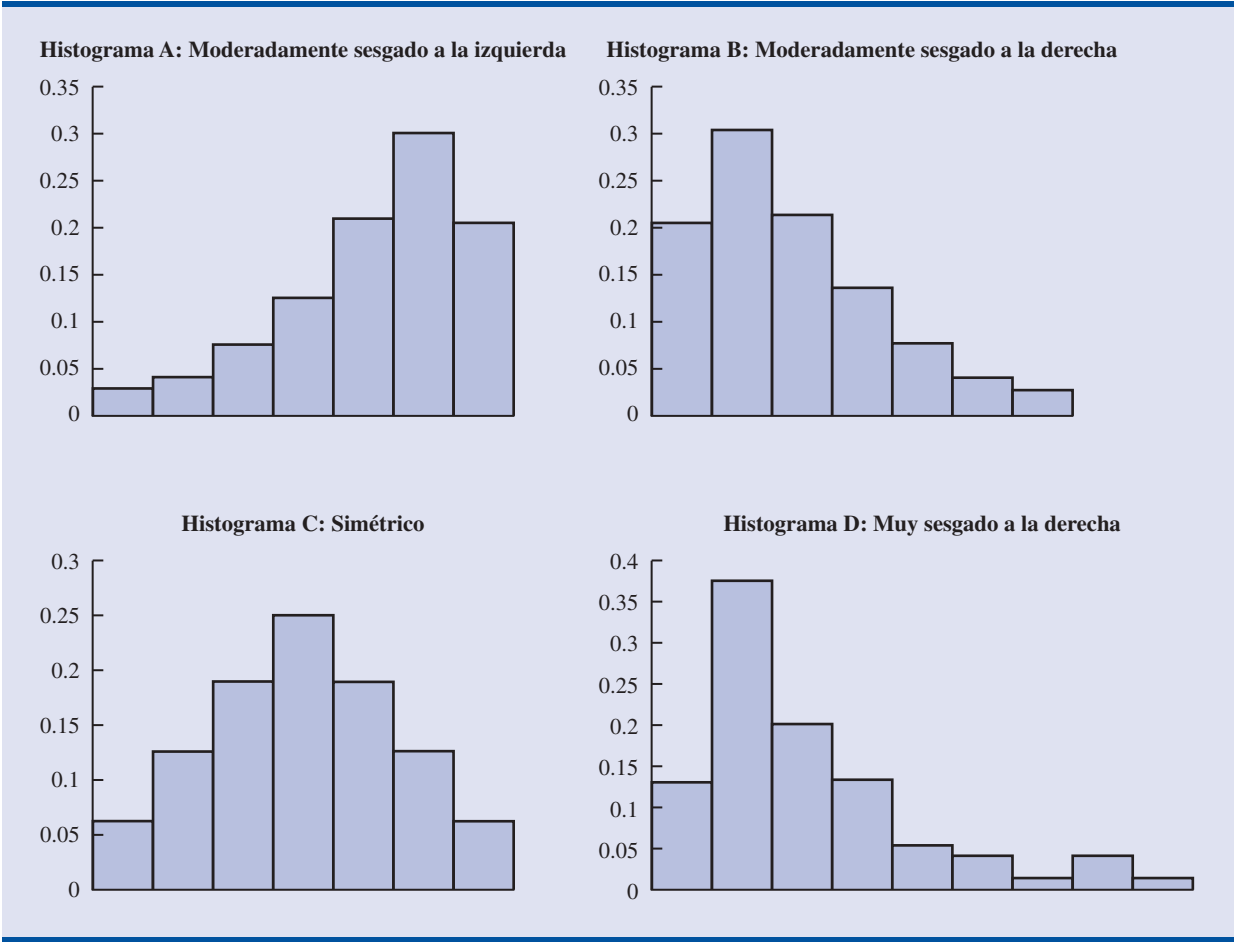
Uno de los usos más importantes de un histograma es proveer información acerca de la forma de la distribución. En la figura 2.5 se muestran cuatro histogramas contruidos a partir de distribuciones de frecuencia relativa. En el histograma A se muestra un conjunto de datos moderadamente sesgado a la izquierda. Se dice que un histograma es sesgado a la izquierda si su cola se extiende más hacia la izquierda. Dichos histogramas son típicos para calificaciones: no hay calificaciones mayores a 100%, la mayor parte están arriba de 70% y sólo hay unas cuantas bajas. En el histograma B se muestra un conjunto de datos moderadamente sesgado a la derecha. Un histograma está sesgado a la derecha si su cola se extiende más hacia la derecha. Ejemplos de este tipo de histogramas son los datos de los precios de las casas; unas cuantas casas caras crean el sesgo a la derecha.

En C se observa un histograma simétrico. En éste la cola izquierda es la imagen de la cola derecha. Los histogramas de datos para aplicaciones nunca son perfectamente simétricos, pero en muchas aplicaciones suelen ser más o menos simétricos. En D se observa un histograma muy sesgado a la derecha. Éste se elaboró con datos sobre la cantidad de compras a lo largo de un día en una tienda de ropa para mujeres. Los datos de aplicaciones de negocios o economía suelen conducir a histogramas sesgados a la derecha. Por ejemplo datos de los precios de las casas, de los salarios, de las cantidades de las compras, etc., suelen dar histogramas sesgados a la derecha.

Distribuciones acumuladas

Una variación de las distribuciones de frecuencia que proporcionan otro resumen tabular de datos cuantitativos es la **distribución de frecuencia acumulada**. La distribución de frecuencia acumulada usa la cantidad, las amplitudes y los límites de las clases de la distribución de frecuencia. Sin embargo, en lugar de mostrar la frecuencia de cada clase, la distribución de frecuencia acumulada muestra la cantidad de datos que tienen un valor *menor o igual* al límite superior de cada clase. Las primeras dos columnas de la tabla 2.7 corresponden a la distribución de frecuencia acumulada de los datos de las duraciones de las auditorías.

FIGURA 2.5 HISTOGRAMAS CON DISTINTOS TIPOS DE SESGO



Para entender cómo se determina la frecuencia acumulada, considere la clase que dice “menor o igual que 24”. La frecuencia acumulada en esta clase es simplemente la suma de la frecuencia de todas las clases en que los valores de los datos son menores o iguales que 24. En la distribución de frecuencia de la tabla 2.5 la suma de las frecuencias para las clases 10–14, 15–29 y 20–24 indica que los datos cuyos valores son menores o iguales que 24 son $4 + 8 + 5 = 17$. Por lo tanto, en esta clase la frecuencia acumulada es 17. Además, en la distribución de frecuen-

TABLA 2.7 DISTRIBUCIONES DE FRECUENCIA ACUMULADA, FRECUENCIA RELATIVA ACUMULADA Y FRECUENCIA PORCENTUAL ACUMULADA

Duración de la auditoría en días	Frecuencia acumulada	Frecuencia relativa acumulada	Frecuencia porcentual acumulada
Menor o igual que 14	4	0.20	20
Menor o igual que 19	12	0.60	60
Menor o igual que 24	17	0.85	85
Menor o igual que 29	19	0.95	95
Menor o igual que 34	20	1.00	100

cias acumuladas de la tabla 2.7 se observa que cuatro auditorías duraron 14 días o menos y que 19 auditorías duraron 29 días o menos.

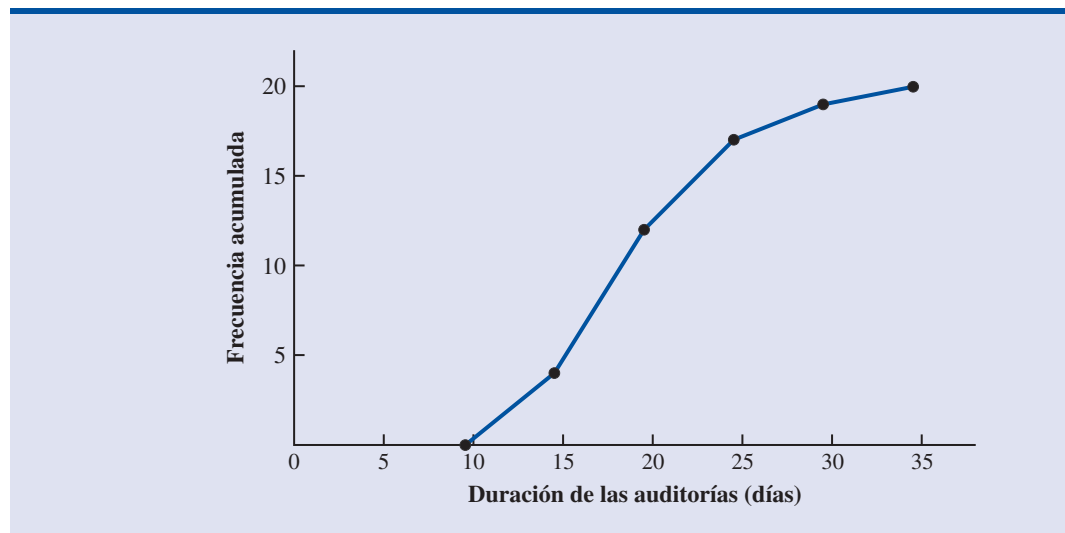
Por último, se tiene que la **distribución de frecuencias relativas acumuladas** indica la proporción de todos los datos que tienen valores menores o iguales al límite superior de cada clase, y la **distribución de frecuencias porcentuales acumuladas** indica el porcentaje de todos los datos que tienen valores menores o iguales al límite superior de cada clase. La distribución de frecuencias relativas acumuladas se calcula ya sea sumando las frecuencias relativas que aparecen en la distribución de frecuencias relativas o dividiendo la frecuencia acumulada entre la cantidad total de datos. Empleando el último método, las frecuencias relativas acumuladas que aparecen en la columna 3 de la tabla 2.7 se obtienen dividiendo las frecuencias acumuladas de la columna 2 entre la cantidad total de datos ($n = 20$). Las frecuencias porcentuales acumuladas se obtienen multiplicando las frecuencias relativas por 100. Estas distribuciones de frecuencias acumuladas relativas y porcentuales indican que 0.85 o el 85% de las auditorías se realizaron en 24 días o menos, 0.95 o 95% de las auditorías se realizaron en 29 días o menos, etcétera.

Ojiva

La gráfica de una distribución acumulada, llamada **ojiva**, es una gráfica que muestra los valores de los datos en el eje horizontal y las frecuencias acumuladas, las frecuencias relativas acumuladas o las frecuencias porcentuales acumuladas en el eje vertical. En la figura 2.6 se muestra una ojiva correspondiente a las frecuencias acumuladas de las duraciones de las auditorías.

La ojiva se construye al graficar cada uno de los puntos correspondientes a la frecuencia acumulada de las clases. Como las clases de las duraciones de las auditorías son 10–14, 15–19, 20–24, etc., hay huecos de una unidad entre 14 y 15, 19 y 20, etc. Estos huecos se eliminan al graficar puntos a la mitad entre los dos límites de clase. Así, para la clase 10–14 se usa 14.5, para la clase 15–19 se usa 19.5 y así en lo sucesivo. En la ojiva de la figura 2.6 la clase “menor o igual que 14” cuya frecuencia acumulada es 4 se grafica mediante el punto que se localiza a 14.5 unidades sobre el eje horizontal y a 4 unidades sobre el vertical. La clase “menor o igual que 19” cuya frecuencia acumulada es 12 se representa por un punto que se encuentra a 19.5 unidades sobre el eje horizontal y 12 unidades sobre el vertical. Observe que en el extremo izquierdo de la ojiva se ha graficado un punto más. Este punto inicia la ojiva mostrando que en los datos no hay valores que se encuentren abajo de la clase 10–14. Este punto se encuentra a 9.5 unidades sobre el eje horizontal y a 0 unidades sobre el vertical. Para terminar los puntos graficados se conectan mediante líneas rectas.

FIGURA 2.6 OJIVA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS



NOTAS Y COMENTARIOS

1. Una gráfica de barras y un histograma son en esencia lo mismo; ambas son representaciones gráficas de una distribución de frecuencia. Un histograma es sólo una gráfica de barras sin separación entre las barras. Para algunos datos cuantitativos discretos, también se puede tener separación entre las barras. Considere por ejemplo, el número de materias en que está inscrito un estudiante universitario. Los datos sólo tienen valores enteros. No hay valores intermedios como 1.5, 2.73, etc. Sin embargo cuando se tienen datos cuantitativos continuos, como en las auditorías, no es apropiado tener separación entre las barras.
2. Los valores adecuados para los límites de clase cuando se tienen datos cuantitativos depende del nivel de precisión de los datos. Por ejemplo, en el caso de los datos de la tabla 2.4, sobre la duración de las auditorías, los límites usados fueron números enteros. Si los datos hubieran estado redondeados a la décima de día más cercana (es decir, 12.3, 14.4, etc.), entonces los límites se hubieran dado con décimas de día. La primera clase, por ejemplo, hubiera sido de 10.0 a 14.9. Si los datos se hubieran registrado hasta la centésima de día más cercana (es decir, 12.34, 14.45, etc.), los límites se hubieran dado con centésimas de días. Por ejemplo la primera clase hubiera sido de 10.00–14.99.
3. Una clase *abierta* sólo necesita el límite inferior de la clase o el límite superior de la clase. Por ejemplo, suponga que en los datos de la tabla 2.4 sobre las duraciones de las auditorías dos de éstas hubieran durado 58 y 65 días. En lugar de haber seguido con clases de amplitud 5 de 35–39, de 40–44, de 45 a 49, etc., podría haber simplificado la distribución de frecuencia mediante una clase abierta de “35 o más”. La frecuencia de esta clase habría sido 2. La mayor parte de las clases abiertas aparecen en el extremo superior de la distribución. Algunas veces se encuentran clases abiertas en el extremo inferior y rara vez están en ambos extremos.
4. En una distribución de frecuencia acumulada, la última frecuencia siempre es igual al número total de observaciones. En una distribución de frecuencia relativa acumulada la última frecuencia siempre es igual a 1.00 y en una distribución de frecuencia porcentual acumulada la última frecuencia es siempre 100.

Ejercicios

Métodos

11. Considere los datos siguientes.

14	21	23	21	16
19	22	25	16	16
24	24	25	19	16
19	18	19	21	12
16	17	18	23	25
20	23	16	20	19
24	26	15	22	24
20	22	24	22	20

- a. Elabore una distribución de frecuencia usando las clases 12–14, 15–17, 18–20, 21–23 y 24–26.
- b. Elabore una distribución de frecuencia relativa y una de frecuencia porcentual usando las clases del inciso a.

12. Considere la distribución de frecuencia siguiente.

Clases	Frecuencia
10–19	10
20–29	14
30–39	17
40–49	7
50–59	2

Construya una distribución de frecuencia acumulada y otra de frecuencia relativa acumulada.



13. Con los datos del ejercicio 12 elabore un histograma y una ojiva.
14. Considere los datos siguientes.

8.9	10.2	11.5	7.8	10.0	12.2	13.5	14.1	10.0	12.2
6.8	9.5	11.5	11.2	14.9	7.5	10.0	6.0	15.8	11.5

- Construya un diagrama de punto.
- Elabore una distribución de frecuencia.
- Construya una distribución de frecuencia porcentual.

Aplicaciones

Autoexamen

15. El personal de un consultorio analiza los tiempos de espera de los pacientes que requieren servicio de emergencia. Los datos siguientes son los tiempos de espera en minutos recolectados a lo largo de un mes.

2 5 10 12 4 4 5 17 11 8 9 8 12 21 6 8 7 13 18 3

Con las clases 0–4, 5–9, etcétera.

- Muestre la distribución de la frecuencia.
 - Expresé la distribución de la frecuencia relativa.
 - Muestre la distribución de frecuencia acumulada.
 - Presente la distribución de frecuencia relativa acumulada.
 - ¿Cuál es la proporción de los pacientes que requieren servicio de emergencia y esperan 9 minutos o menos?
16. Considere las dos distribuciones de frecuencias siguientes. La primera distribución de frecuencia proporciona el ingreso anual bruto ajustado de Estados Unidos (Internal Revenue Service, marzo 2003). La segunda distribución de frecuencia muestra las calificaciones de exámenes de un grupo de estudiantes universitarios en un curso de estadística.

Ingreso (en miles de \$)	Frecuencia (en millones)	Calificaciones de examen	Frecuencia
0–24	60	20–29	2
25–49	33	30–39	5
50–74	20	40–49	6
75–99	6	50–59	13
100–124	4	60–69	32
125–149	2	70–79	78
150–174	1	80–89	43
175–199	1	90–99	21
Total	127	Total	200

- Con los datos del ingreso anual elabore un histograma. ¿Qué evidencia de sesgo observa? ¿Es razonable este sesgo? Explique.
 - Con los datos de las calificaciones elabore un histograma. ¿Qué evidencia de sesgo observa? Explique.
 - Con los datos del ejercicio 11 elabore un histograma. ¿Qué evidencia de sesgo observa? ¿Cuál es la forma general de la distribución?
17. ¿Cuál es el precio típico de las acciones de las 30 empresas del promedio industrial Dow Jones? Los datos siguientes son los precios de las acciones, al dólar más cercano, en enero de 2006 (*The Wall Street Journal*, 16 de enero de 2006).



Empresa	\$/Acción	Empresa	\$/Acción
AIG	70	Home Depot	42
Alcoa	29	Honeywell	37
Altria Group	76	IBM	83
American Express	53	Intel	26
AT&T	25	Johnson & Johnson	62
Boeing	69	JPMorgan Chase	40
Caterpillar	62	McDonald's	35
Citigroup	49	Merck	33
Coca-Cola	41	Microsoft	27
Disney	26	3M	78
DuPont	40	Pfizer	25
ExxonMobil	61	Procter & Gamble	59
General Electric	35	United Technologies	56
General Motors	20	Verizon	32
Hewlett-Packard	32	Wal-Mart	45

- Con estos datos elabore una distribución de frecuencia.
 - Con estos datos elabore un histograma. Interprete el histograma, presente un análisis de la forma general del histograma, el precio medio de cada intervalo de acciones, el precio más frecuente por intervalo de acciones, los precios más alto y más bajo por acción.
 - ¿Cuáles son las acciones que tienen el precio más alto y el más bajo?
 - Use *The Wall Street Journal* para encontrar los precios actuales por acción de estas empresas. Elabore un histograma con estos datos y discuta los cambios en comparación con enero de 2006.
18. NRF/BIG proporciona los resultados de una investigación sobre las cantidades que gastan en vacaciones los consumidores (*USA Today*, 20 de diciembre de 2005). Los datos siguientes son las cantidades gastadas en vacaciones por los 25 consumidores de una muestra.



1200	850	740	590	340
450	890	260	610	350
1780	180	850	2050	770
800	1090	510	520	220
1450	280	1120	200	350

- ¿Cuál es la menor cantidad gastada en vacaciones? ¿Cuál la mayor?
 - Use \$250 como amplitud de clase para elaborar con estos datos una distribución de frecuencia y una distribución de frecuencia porcentual.
 - Elabore un histograma y comente la forma de la distribución.
 - ¿Qué observaciones le permiten hacer las cantidades gastadas en vacaciones?
19. El correo no deseado afecta la productividad de los oficinistas. Se hizo una investigación con oficinistas para determinar la cantidad de tiempo por día que pierden en estos correos no deseados. Los datos siguientes corresponden a los tiempos en minutos perdidos por día observados en una muestra.

2	4	8	4
8	1	2	32
12	1	5	7
5	5	3	4
24	19	4	14

Resuma estos datos construyendo:

- Una distribución de frecuencia (con las clases 1–5, 6–10, 11–15, 16–20, etc.)
- Una distribución de frecuencia relativa
- Una distribución de frecuencia acumulada.

- d. Una distribución de frecuencia relativa acumulada.
 - e. Una ojiva.
 - f. ¿Qué porcentaje de los oficinistas pierde 5 minutos o menos en revisar el correo no deseado?
¿Qué porcentaje pierde más de 10 minutos por día en esto?
20. A continuación se presentan las 20 mejores giras de concierto y el precio promedio del costo de sus entradas en Estados Unidos. Esta lista se basa en datos proporcionados por los promotores y administradores de los locales a la publicación *Pollstar* (*Associated Press*, 21 de noviembre de 2003).



Gira de conciertos	Precio de la entrada	Gira de conciertos	Precio de la entrada
Bruce Springsteen	\$72.40	Toby Keith	\$37.76
Dave Matthews Band	44.11	James Taylor	44.93
Aerosmith/KISS	69.52	Alabama	40.83
Shania Twain	61.80	Harper/Johnson	33.70
Fleetwood Mac	78.34	50 Cent	38.89
Radiohead	39.50	Steely Dan	36.38
Cher	64.47	Red Hot Chili Peppers	56.82
Counting Crows	36.48	R.E.M.	46.16
Timberlake/Aguilera	74.43	American Idols Live	39.11
Mana	46.48	Mariah Carey	56.08

Resuma los datos construyendo:

- a. Una distribución de frecuencia y una distribución de frecuencia porcentual.
 - b. Un histograma.
 - c. ¿Qué concierto tiene el precio promedio más alto? ¿Qué concierto tiene el precio promedio menos caro?
 - d. Haga un comentario sobre qué indican los datos acerca de los precios promedio de las mejores giras de concierto.
21. *Nielsen Home Technology Report* informa sobre la tecnología en el hogar y su uso. Los datos siguientes son las horas de uso de computadora por semana en una muestra de 50 personas.



4.1	1.5	10.4	5.9	3.4	5.7	1.6	6.1	3.0	3.7
3.1	4.8	2.0	14.8	5.4	4.2	3.9	4.1	11.1	3.5
4.1	4.1	8.8	5.6	4.3	3.3	7.1	10.3	6.2	7.6
10.8	2.8	9.5	12.9	12.1	0.7	4.0	9.2	4.4	5.7
7.2	6.1	5.7	5.9	4.7	3.9	3.7	3.1	6.1	3.1

Resuma estos datos construyendo:

- a. Una distribución de frecuencia (como ancho de clase use tres horas).
- b. Una distribución de frecuencia relativa.
- c. Un histograma.
- d. Una ojiva.
- e. Haga un comentario sobre lo que indican los datos respecto al uso de la computadora en el hogar.

2.3

Análisis exploratorio de datos: el diagrama de tallo y hojas

Las técnicas del **análisis exploratorio de datos** emplean aritmética sencilla y gráficas fáciles de dibujar útiles para resumir datos. La técnica conocida como **diagrama de tallo y hojas** muestra en forma simultánea el orden jerárquico y la forma de un conjunto de datos.

TABLA 2.8 NÚMERO DE PREGUNTAS CONTESTADAS CORRECTAMENTE EN UN EXAMEN DE APTITUDES

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

Para ilustrar el uso de los diagramas de tallo y hojas, considere la tabla 2.8. Estos datos son el resultado de un examen de aptitudes con 150 preguntas presentado por 50 personas que aspiraban a un puesto en una empresa. Los datos indican el número de respuestas correctas por examen.

Para elaborar un diagrama de tallo y hoja inicie acomodando los primeros dígitos de cada uno de los datos a la izquierda de una línea vertical. A la derecha de la línea vertical se anota el último dígito de cada dato. Con base en el primer renglón de la tabla 2.8 (112, 72, 69, 97 y 107), los primeros cinco datos al elaborar el diagrama de tallo y hojas serían los siguientes:

6	9
7	2
8	
9	7
10	7
11	2
12	
13	
14	

Por ejemplo, para el dato 112, se observa que los primeros dígitos, 11, se encuentran a la izquierda de la línea y el último dato, 2, a la derecha. De manera similar, el primer dígito, 7, del dato 72 se encuentra a la izquierda de la línea y el 2 a la derecha. Si continúa colocando el último dígito de cada dato en el renglón correspondiente a sus primeros dígitos obtiene:

6	9	8							
7	2	3	6	3	6	5			
8	6	2	3	1	1	0	4	5	
9	7	2	2	6	2	1	5	8	8
10	7	4	8	0	2	6	6	0	6
11	2	8	5	9	3	5	9		
12	6	8	7	4					
13	2	4							
14	1								

Una vez organizados los datos de esta manera, ordenar los datos de cada renglón de menor a mayor es sencillo. Entonces obtiene el diagrama de tallo y hojas que se muestra aquí.

[illegible]

Los números a la izquierda de la línea vertical (6, 7, 8, 9, 10, 11, 12, 13 y 14) forman el *tallo*, y cada dígito a la derecha de la línea vertical es una *hoja*. Por ejemplo, considere el primer renglón que tiene como tallo el 6 y como hojas 8 y 9.

6 | 8 9

Este renglón indica que hay dos datos que tienen como primer dígito el seis. Las hojas indican que estos datos son 68 y 69. De manera similar, el segundo renglón

7 | 2 3 3 5 6 6

indica que hay seis datos que tienen como primer dígito el 7. Las hojas indican que estos datos son 72, 73, 73, 75, 76 y 76.

Para atender a la forma del diagrama de tallo y hojas, se usan rectángulos que contienen las hojas de cada tallo; con esto se obtiene lo siguiente.

6	8	9																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			</
---	---	---	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	----

Al rotar la página sobre su costado en contra de las manecillas del reloj se obtiene una imagen de los datos que es parecida a un histograma y en el que las clases son 60–69, 70–79, 80–89, etcétera.

Aunque el diagrama de tallo y hojas parece proporcionar la misma información que un histograma, tiene dos ventajas fundamentales.

1. El diagrama de tallo y hojas es más fácil de construir a mano.
2. En cada intervalo de clase proporciona más información que un histograma debido a que el tallo y la hoja proporcionan el dato.

Así como para una distribución de frecuencia o para un histograma no hay un determinado número de clases, tampoco para el diagrama de tallo y hojas hay un número determinado de renglones a tallos. Si piensa que el diagrama de tallo y hojas original condensa demasiado los datos, es fácil expandirlo empleando dos o más tallos por cada primer dígito. Por ejemplo, para usar

En un diagrama expandido de tallo y hojas, siempre que un tallo aparece dos veces, al primero le corresponden las hojas 0–4 y al segundo las hojas 5–9.

dos tallos por cada primer dígito se ponen todos los datos que terminen en 0, 1, 2, 3 o 4 en un renglón y todos los datos que terminen en 5, 6, 7, 8 o 9 en otro. Este método se ilustra en el siguiente diagrama expandido de tallo y hojas.

6	8	9
7	2	3 3
7	5	6 6
8	0	1 1 2 3 4
8	5	6
9	1	2 2 2 4
9	5	5 6 7 8 8
10	0	0 2 4
10	6	6 6 7 8
11	2	3
11	5	5 8 9 9
12	4	
12	6	7 8
13	2	4
13		
14	1	

Observe que las hojas de los datos 72, 73 y 73 pertenecen al intervalo 0–4 y aparecen con el primer tallo que tiene el valor 7. Las hojas de los valores 75, 76 y 76 pertenecen al intervalo 5–9 y aparecen con el segundo tallo que tiene el valor 7. Este diagrama expandido de tallo y hojas es semejante a una distribución con los intervalos 65–69, 70–74, 75–79, etcétera.

El ejemplo anterior muestra un diagrama de tallo y hojas con datos de hasta tres dígitos. Estos diagramas también se elaboran con datos de más de tres dígitos. Por ejemplo, considere los datos siguientes sobre el número de hamburguesas vendidas en un restaurante de comida rápida en cada una de 15 semanas.

1565	1852	1644	1766	1888	1912	2044	1812
1790	1679	2008	1852	1967	1954	1733	

A continuación se presenta un diagrama de tallo y hojas de estos datos.

Unidad de hoja = 10

15	6
16	4 7
17	3 6 9
18	1 5 5 8
19	1 5 6
20	0 4

En un diagrama de tallo y hojas se usa un solo dígito para definir cada hoja. La unidad de hoja indica por qué número debe multiplicar los números del tallo y la hoja para aproximar el dato original. Las unidades de hoja son 100, 10, 1, 0.1 etcétera.

Observe que para definir cada hoja se emplea un solo dígito y que para construir el diagrama sólo se usaron los primeros tres dígitos de cada dato. En la parte superior del diagrama se ha especificado que la Unidad de hoja = 10. Para ilustrar cómo se interpretan los datos de este diagrama considere el primer tallo 15 y su hoja correspondiente 6. Al unir estos números obtiene 156. Para lograr una aproximación al dato original es necesario multiplicar este número por 10, el valor de la *unidad de hoja*. Por tanto, $156 \times 10 = 1560$ es una aproximación al dato original empleado para construir el diagrama de tallo y hoja. Aunque a partir de este diagrama no es posible reconstruir los datos exactos, la convención de usar un solo dígito para cada hoja, permite construir diagramas de tallo y hojas con datos que tengan un gran número de dígitos. En diagramas de tallo y hojas en los que no se especifica la unidad de hoja, se supone que la unidad es 1.

Ejercicios

Métodos

22. Con los datos siguientes construya un diagrama de tallo y hojas.

70	72	75	64	58	83	80	82
76	75	68	65	57	78	85	72

23. Con los datos siguientes construya un diagrama de tallo y hojas.

11.3	9.6	10.4	7.5	8.3	10.5	10.0
9.3	8.1	7.7	7.5	8.4	6.3	8.8

24. Con los datos siguientes construya un diagrama de tallo y hojas. Use 10 como unidad de hoja.

1161	1206	1478	1300	1604	1725	1361	1422
1221	1378	1623	1426	1557	1730	1706	1689

Aplicaciones

25. Un psicólogo elabora una nueva prueba de inteligencia para adultos. Aplica la prueba a 20 individuos y obtiene los datos siguientes.

114	99	131	124	117	102	106	127	119	115
98	104	144	151	132	106	125	122	118	118

Construya un diagrama de tallo y hojas.

26. La asociación estadounidense de inversionistas individuales realiza una investigación anual sobre intermediarios de descuento. Las siguientes son las comisiones en una muestra de 24 intermediarios (*AII Journal*, enero de 2003). Estas son dos tipos de operaciones con asistencia de 100 acciones a \$50 cada una y una operación en línea de 500 acciones a \$50 cada una.

Corredor	Operación con asistencia de 100 acciones \$50/acción	Operación en línea de 500 acciones a \$50 /acción	Corredor	Operación con asistencia de 100 acciones \$50/acción	Operación en línea de 500 acciones a \$50/acción
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

archivo
en
CD
Broker

- a. Redondee los precios al dólar más cercano y elabore un diagrama de tallo y hojas de las 100 acciones a \$50 por acción. Haga un comentario sobre la información que obtuvo acerca de estos precios.
- b. Redondee los precios al dólar más cercano y elabore un diagrama de tallo y hojas de las 500 acciones a \$50 por acción. Haga un comentario sobre estos precios.
27. La mayor parte de los centros turísticos importantes de esquí de Estados Unidos ofrecen programas familiares con clases de esquí para niños. Por lo general proporcionan 4 a 6 horas de clase con un instructor certificado. A continuación se presentan las cuotas diarias en 15 centros turísticos. (*The Wall Street Journal*, 20 de enero de 2006).

Centro turístico	Ubicación	Cuota diaria	Centro turístico	Ubicación	Cuota diaria
Beaver Creek	Colorado	\$ 137	Okemo	Vermont	\$ 86
Deer Valley	Utah	115	Park City	Utah	145
Diamond Peak	California	95	Butternut	Massachusetts	75
Heavenly	California	145	Steamboat	Colorado	98
Hunter	New York	79	Stowe	Vermont	104
Mammoth	California	111	Sugar Bowl	California	100
Mount Sunapee	New Hampshire	96	Whistler-Blackcomb	British Columbia	104
Mount Bachelor	Oregon	83			

- Con estos datos elabore un diagrama de tallo y hojas.
 - Interprete el diagrama de tallo y hojas en términos de lo que expresa de las cuotas diarias de estos programas.
28. Para un maratón (13.1 millas) en Florida en 2004 hubo 1228 registrados (*Naples Daily News*, 17 de enero de 2004). Para esta competencia hubo seis grupos de edades. Los datos siguientes son las edades encontradas en una muestra de 40 participantes.



49	33	40	37	56
44	46	57	55	32
50	52	43	64	40
46	24	30	37	43
31	43	50	36	61
27	44	35	31	43
52	43	66	31	50
72	26	59	21	47

- Realice un diagrama expandido de tallo y hojas.
- ¿En qué grupo de edad hubo más participantes?
- ¿Qué edad se presenta con más frecuencia?
- En un artículo del *Naples Daily News* se hace énfasis sobre la cantidad de corredores de veintitantos años. ¿Qué porcentaje de los corredores pertenecían al grupo de veintitantos años? ¿Cuál supone qué era el tema del artículo?

2.4

Tabulaciones cruzadas y diagramas de dispersión

Las tabulaciones cruzadas y los diagramas de dispersión son empleados para presentar un resumen de datos, de tal manera que revele la relación entre las dos variables.

Este capítulo, hasta ahora, se ha concentrado en los métodos tabulares y gráficos empleados para resumir datos de una *sola variable*. Con frecuencia, los directivos o quienes deben tomar decisiones requieren métodos tabulares o gráficos que les ayuden a entender la *relación entre dos variables*. La tabulación cruzada y los diagramas de dispersión son dos métodos de este tipo.

Tabulación cruzada

Una **tabulación cruzada** es un resumen tabular de los datos de dos variables. El uso de la tabulación cruzada se ilustrará con los datos de la aplicación siguiente, que se basan en datos de *Zagat's Restaurant Review*. Se recolectaron los datos correspondientes a la calidad y precios de 300 restaurantes en el área de Los Ángeles. La tabla 2.9 muestra los datos de los 10 primeros restaurantes. Se presentan los datos de calidad y precio característicos de estos restaurantes. La calidad es una variable cualitativa que tiene como categorías bueno, muy bueno y excelente. El precio es una variable cuantitativa que va desde \$10 hasta \$49.

En la tabla 2.10 se muestra una tabulación cruzada con los datos de esta aplicación. El encabezado de la primera columna y el primer renglón definen las clases para las dos variables. Los encabezados de los renglones en el margen izquierdo (buena, muy buena y excelente) corresponden a las tres categorías de calidad. Los encabezados de las columnas (\$10–19, \$20–29, \$30–39 y

**TABLA 2.9** EVALUACIÓN DE LA CALIDAD Y PRECIOS DE 300 RESTAURANTES DE LOS ÁNGELES

Restaurante	Calidad	Precio
1	Bueno	18
2	Muy bueno	22
3	Bueno	28
4	Excelente	38
5	Muy bueno	33
6	Bueno	28
7	Muy bueno	19
8	Muy bueno	11
9	Muy bueno	23
10	Bueno	13
.	.	.
.	.	.
.	.	.

\$40–49) corresponden a las cuatro clases de la variable precio. Para cada restaurante de la muestra se tiene el nivel de calidad y el precio. Por tanto, a cada restaurante de la muestra le corresponde una celda en un renglón y en una columna de la tabla. Por ejemplo, si el restaurante 5 tiene muy buena calidad y su precio es \$33, a este restaurante le corresponde el renglón 2 y la columna 3 de la tabla 2.10. Así que para elaborar una tabulación cruzada, simplemente se cuenta el número de restaurantes que pertenecen a cada una de las celdas de la tabla de tabulación cruzada.

La tabla 2.10 muestra que la mayor parte de los restaurantes de la muestra (64) tienen muy buena calidad y su precio está en el intervalo \$20–29. También se ve que sólo dos restaurantes tienen una calidad excelente y un precio en el intervalo \$10–19. Así es posible hacer interpretaciones semejantes con el resto de las frecuencias. Observe además que en el margen derecho y en el renglón inferior de la tabulación cruzada aparecen las distribuciones de frecuencia de la calidad y de los precios, por separado. En la distribución de frecuencia de la calidad, en el margen derecho, se observa que hay 84 restaurantes buenos, 150 muy buenos y 66 restaurantes excelentes. De manera semejante, en el renglón inferior se tiene la distribución de frecuencia de la variable precios.

Al dividir los totales del margen derecho de la tabulación cruzada entre el total de esa columna se obtienen distribuciones de frecuencia relativa y frecuencia porcentual de la variable calidad.

Calidad	Frecuencia relativa	Frecuencia porcentual
Bueno	0.28	28
Muy bueno	0.50	50
Excelente	0.22	22
Total	1.00	100

TABLA 2.10 TABULACIÓN CRUZADA DE CALIDAD Y PRECIO DE 300 RESTAURANTES DE LOS ÁNGELES

Calidad	Precio				Total
	\$10–19	\$20–29	\$30–39	\$40–49	
Buena	42	40	2	0	84
Muy buena	34	64	46	6	150
Excelente	2	14	28	22	66
Total	78	118	76	28	300

En esta distribución de frecuencia porcentual se observa que 28% de los restaurantes son calificados como buenos, 50% como muy buenos y 22% excelentes.

Si divide los totales del renglón inferior de la tabulación cruzada entre el total de ese renglón obtiene distribuciones de frecuencia relativa y de frecuencia porcentual de los precios.

Precio	Frecuencia relativa	Frecuencia porcentual
\$10–19	0.26	26
\$20–29	0.39	39
\$30–39	0.25	25
\$40–49	0.09	9
Total	1.00	100

Observe que la suma de los valores en cada columna no tiene correspondencia exacta con el total de la columna debido a que los valores que se suman han sido redondeados. En esta distribución de frecuencia porcentual 26% de los precios se encuentran en la clase de los precios más bajos, 39% se encuentran en la clase siguiente, etcétera.

Las distribuciones de frecuencia y de frecuencia relativa obtenidas de los márgenes de las tabulaciones cruzadas proporcionan información de cada una de las variables por separado, pero no dan ninguna luz acerca de la relación entre las variables. El principal valor de una tabulación cruzada es que permite ver la relación entre las variables. Una observación de la tabulación cruzada de la tabla 2.10 es que los precios más altos están relacionados con la mejor calidad de los restaurantes y los precios bajos están relacionados con menor calidad.

Si se convierten las cantidades de una tabulación cruzada en porcentajes de columna o de renglón, se obtiene más claridad sobre la relación entre las variables. En la tabla 2.11 se presentan los porcentajes de renglón, que son el resultado de dividir cada frecuencia de la tabla 2.10 entre el total del renglón correspondiente. Entonces, cada renglón de la tabla 2.11 es una distribución de frecuencia porcentual de los precios en esa categoría de calidad. Entre los restaurantes de menor calidad (buenos), el mayor porcentaje corresponde a los menos caros (50% tiene precios en el intervalo \$10–19 y 47.6% en el intervalo \$20–29). De los restaurantes de mayor calidad (excelentes), los porcentajes mayores corresponden a los más caros (42.4% tiene precios de \$30–39 y 33.4% de \$40–49). Así que un precio más elevado está relacionado con una mejor calidad de los restaurantes.

La tabulación cruzada se utiliza mucho para examinar la relación entre dos variables. En la práctica, los informes finales de muchos estudios estadísticos contienen una gran cantidad de tabulaciones cruzadas. En este estudio sobre los restaurantes de Los Ángeles, en la tabulación cruzada se emplea una variable cualitativa (las calidades) y una cuantitativa (los precios). También se elaboran tabulaciones cruzadas con dos variables cualitativas o cuantitativas. Cuando se usan variables cuantitativas, primero es necesario crear las clases para los valores de las variables. Por ejemplo, en el caso de los restaurantes se agruparon los precios en cuatro categorías (\$10–19, \$20–29, \$30–39 y \$40–49).

TABLA 2.11 PORCENTAJES DE RENGLÓN DE CADA CATEGORÍA DE CALIDAD

Calidad	Precio				Total
	\$10–19	\$20–29	\$30–39	\$40–49	
Buena	50.0	47.6	2.4	0.0	100
Muy buena	22.7	42.7	30.6	4.0	100
Excelente	3.0	21.2	42.4	33.4	100

Paradoja de Simpson

Es posible combinar o agregar los datos de dos o más tabulaciones cruzadas para obtener una tabulación cruzada resumida que muestre la relación entre dos variables. En tales casos hay que tener mucho cuidado al sacar conclusiones acerca de la relación entre las dos variables de la tabulación cruzada agregada. En algunos casos las conclusiones obtenidas de la tabulación cruzada agregada se invierten por completo al observar los datos no agregados, situación conocida como **paradoja de Simpson**. Para ilustrar la paradoja de Simpson, se proporciona un ejemplo en el que se analizan las sentencias de dos jueces en dos tipos de tribunales.

Los jueces Ron Luckett y Dennis Kendall, presidieron los tres últimos años dos tipos de tribunales, de primera instancia y municipal. Algunas de las sentencias por ellos dictadas fueron apeladas. En la mayor parte de los casos los tribunales de apelación ratificaron las sentencias, pero en algunos casos fueron revocadas. Para cada juez se elabora una tabulación cruzada con las variables: sentencia (ratificada o revocada) y tipo de tribunal (de primera instancia y municipal). Suponga que después se combinan las dos tabulaciones cruzadas agregando los datos de los dos tipos de tribunales. La tabulación cruzada agregada que se obtiene tiene dos variables: sentencia (ratificada o revocada) y juez (Luckett o Kendall). En esta tabulación cruzada para cada uno de los jueces se da la cantidad de sentencias que fueron ratificadas y la cantidad de sentencias que fueron revocadas. En la tabla siguiente se presentan estos resultados junto a los porcentajes de columna entre paréntesis al lado de cada valor.

Sentencia	Juez		Total
	Luckett	Kendall	
Ratificada	129 (86%)	110 (88%)	239
Revocada	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

Al analizar la columna de porcentajes resulta que 14% de las sentencias del juez Luckett fueron revocadas, pero del juez Kendall sólo 12% de las sentencias lo fueron. Por tanto, el juez Kendall tuvo un mejor desempeño, ya que de sus sentencias se ratificó un porcentaje mayor. Sin embargo, de esta conclusión surge un problema.

En la tabla siguiente se muestran los casos atendidos por cada uno de los jueces en los dos tribunales; aquí también se dan los porcentajes entre paréntesis al lado de los valores.

Juez Luckett				Juez Kendall			
Sentencia	Tribunal de primera instancia	Tribunal municipal	Total	Sentencia	Tribunal de primera instancia	Tribunal municipal	Total
Ratificada	29 (91%)	100 (85%)	129	Ratificada	90 (90%)	20 (80%)	110
Revocada	3 (9%)	18 (15%)	21	Revocada	10 (10%)	5 (20%)	15
Total (%)	32 (100%)	118 (100%)	150	Total (%)	100 (100%)	25 (100%)	125

Respecto de los porcentajes de Luckett, en el tribunal de primera instancia 91% de sus sentencias fueron ratificadas y en el tribunal municipal 85% lo fueron. En cuanto a los porcentajes de Kendall, 90% de sus sentencias del tribunal de primera instancia y 80% del tribunal municipal fueron ratificadas. Al comparar los porcentajes de columna de los dos jueces, es obvio que el juez Luckett tuvo un mejor desempeño en ambos tribunales que el Juez Kendall. Esto contradice las conclusiones obtenidas al agregar los datos de los dos tribunales en la primera tabulación cruzada. Se pensó que el juez Kendall tenía un mejor desempeño. Este ejemplo ilustra la paradoja de Simpson.

La primera tabulación cruzada se obtuvo agregando los datos de los dos tribunales de dos tabulaciones cruzadas. Observe que los dos jueces tuvieron porcentajes mayores de sentencias revocadas en las sentencias del tribunal municipal que en las del tribunal de primera instancia. Como el juez Luckett tuvo un porcentaje mayor de casos del tribunal municipal, los datos agregados favorecieron al juez Kendall. Sin embargo, si presta atención a las tabulaciones cruzadas de cada uno de los jueces, es claro que el juez Luckett tuvo un mejor desempeño. Por tanto, en la primera tabulación cruzada el *tipo de tribunal* es una variable oculta que no debe ser ignorada al evaluar el desempeño de estos dos jueces.

Debido a la paradoja de Simpson, es necesario tener mucho cuidado al sacar conclusiones cuando se usan datos agregados. Antes de cualquier conclusión acerca de la relación entre dos variables, en una tabulación cruzada en la que se usan datos agregados, es preciso investigar si no existen variables ocultas que afecten los resultados.

Diagrama de dispersión y línea de tendencia

Un **diagrama de dispersión** es una representación gráfica de la relación entre dos variables cuantitativas y una **línea de tendencia** es una línea que da una aproximación de la relación. Como ejemplo, considere la relación publicidad/ventas en una tienda de equipos de sonido. Durante los últimos tres meses, en 10 ocasiones la tienda apareció en comerciales de televisión, en el fin de semana, para promover sus ventas. Los directivos quieren investigar si hay relación entre el número de comerciales emitidos el fin de semana y las ventas en la semana siguiente. En la tabla 2.12 se presentan datos muestrales de las 10 semanas dando las ventas en cientos de dólares.

En la figura 2.7 aparece el diagrama de dispersión y la línea de tendencia* de los datos de la tabla 2.12. El número de comerciales (*x*) aparece en el eje horizontal y las ventas (*y*) en el eje vertical. En la semana 1, *x* = 2 y *y* = 50. En el diagrama de dispersión se grafica un punto con estas coordenadas. Para las otras nueve semanas se grafican puntos similares. Observe que en dos semanas sólo hubo un comercial, en otras dos semanas hubo dos comerciales, etcétera.

De nuevo, respecto de la figura 2.7, se observa una relación positiva entre el número de comerciales y las ventas. Más ventas corresponden a más comerciales. La relación no es perfecta ya que los puntos no trazan una línea recta. Sin embargo, el patrón que siguen los puntos y la línea de tendencia indican que la relación es positiva.

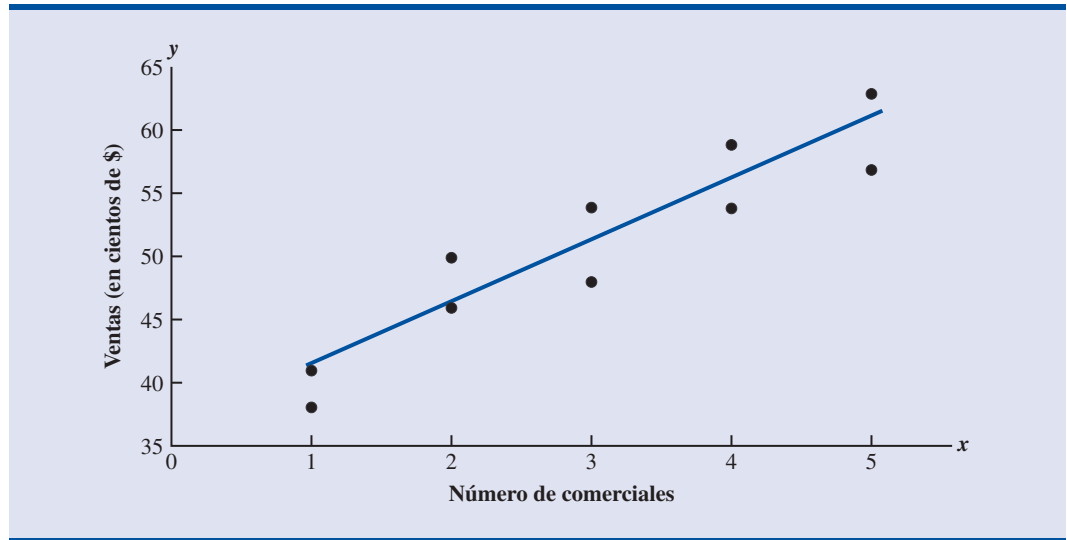
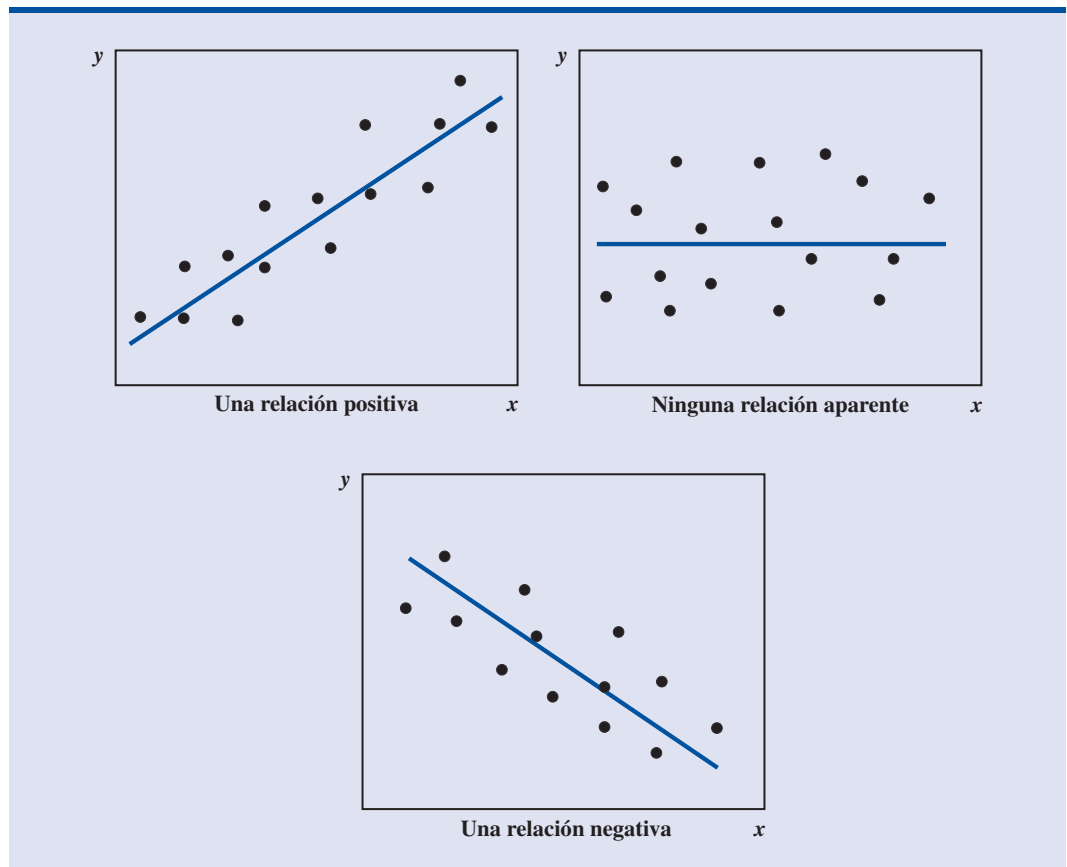
En la figura 2.8 se muestran los patrones de los diagramas de dispersión y el tipo de relación que sugieren. La gráfica arriba a la izquierda representa una relación positiva parecida a la del

TABLA 2.12 DATOS MUESTRALES DE UNA TIENDA DE EQUIPOS DE SONIDO

Semana	Número de comerciales <i>x</i>	Ventas (en cientos de dólares) <i>y</i>
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



*La ecuación de la línea de tendencia es $y = 36.15 + 4.95x$. La pendiente de la línea de tendencia es 4.95 y la intersección con el eje *y* (el punto en que la recta interseca el eje *y*) es 36.15. La interpretación de la pendiente *y* de la intersección con el eje *y* de una línea de tendencia lineal lo verá con detalle en el capítulo 12, cuando estudie la regresión lineal simple.

FIGURA 2.7 DIAGRAMA DE DISPERSIÓN Y LÍNEA DE TENDENCIA DE LA TIENDA DE EQUIPOS DE SONIDO**FIGURA 2.8** TIPOS DE RELACIÓN QUE APARECEN EN LOS DIAGRAMAS DE DISPERSIÓN

ejemplo de la cantidad de comerciales y las ventas. En la gráfica de arriba a la derecha no aparece ninguna relación entre las dos variables. La gráfica inferior representa una relación negativa en la que y tiende a disminuir a medida que x aumenta.

Ejercicios

Métodos

29. Los siguientes son datos de 30 observaciones en las que intervienen dos variables, x y y . Las categorías para x son A, B, y C; para y son 1 y 2.



Observación	x	y	Observación	x	y
1	A	1	16	B	2
2	B	1	17	C	1
3	B	1	18	B	1
4	C	2	19	C	1
5	B	1	20	B	1
6	C	2	21	C	2
7	B	1	22	B	1
8	C	2	23	C	2
9	A	1	24	A	1
10	B	1	25	B	1
11	A	1	26	C	2
12	B	1	27	C	2
13	C	2	28	A	1
14	C	2	29	B	1
15	C	2	30	B	2

- Con estos datos elabore una tabulación cruzada en la que x sea la variable para los renglones y y para las columnas.
- Calcule los porcentajes de los renglones.
- Calcule los porcentajes de las columnas.
- ¿Cuál es la relación, si hay alguna, entre las variables x y y ?

30. Las siguientes 20 observaciones corresponden a 20 variables cuantitativas, x y y .



Observación	x	y	Observación	x	y
1	-22	22	11	-37	48
2	-33	49	12	34	-29
3	2	8	13	9	-18
4	29	-16	14	-33	31
5	-13	10	15	20	-16
6	21	-28	16	-3	14
7	-13	27	17	-15	18
8	-23	35	18	12	17
9	14	-5	19	-20	-11
10	3	-3	20	-7	-22

- Elabore un diagrama de dispersión para la relación entre x y y .
- ¿Cuál es la relación, si hay alguna, entre x y y ?

Aplicaciones

31. En la siguiente tabulación cruzada se muestra el ingreso familiar de acuerdo con el nivel de estudios del jefe de familia, (*Statistical Abstract of the United States, 2002*).

Nivel de estudios	Ingreso por familia (en miles de dólares)					Total
	Menos de 25	25.0–49.9	50.0–74.9	75.0–99.9	100 o más	
No terminó secundaria	9 285	4 093	1 589	541	354	15 862
Terminó secundaria	10 150	9 821	6 050	2 737	2 028	30 786
Parte de bachillerato	6 011	8 221	5 813	3 215	3 120	26 380
Título universitario	2 138	3 985	3 952	2 698	4 748	17 521
Posgrado	813	1 497	1 815	1 589	3 765	9 479
Total	28 397	27 617	19 219	10 780	14 015	100 028

- Calcule los porcentajes por renglón e identifique las distribuciones de frecuencia porcentual del ingreso en los hogares en que el jefe de familia terminó secundaria y en los hogares en que el jefe de familia tiene un título universitario.
 - ¿Qué porcentaje de las familias en que el jefe de familia terminó secundaria gana \$75 000 o más? ¿Qué porcentaje de las familias en que el jefe de familia tienen un título universitario gana 75 000 o más?
 - Con los ingresos de los hogares en que el jefe de familia terminó secundaria elabore un histograma de la frecuencia porcentual, y otro con los ingresos de las familias en que el jefe de familia tiene un grado universitario. ¿Se observa alguna relación clara entre el ingreso familiar y el nivel de educación?
32. Consulte la tabulación cruzada del ingreso familiar de acuerdo con el nivel de estudios del ejercicio 31.
- Calcule los porcentajes e identifique las distribuciones de frecuencia porcentual. ¿Qué porcentaje de jefes de familia no terminó la secundaria?
 - ¿Qué porcentaje de los hogares que perciben \$100 000 o más tienen como jefe de familia a una persona con un posgrado? ¿Qué porcentaje de los hogares que tienen como jefe de familia a una persona con un posgrado perciben más de \$100 000? ¿Por qué son diferentes estos dos porcentajes?
 - Compare las distribuciones de frecuencia porcentual de aquellos hogares que perciben “Menos que 25”, “100 o más” y del “Total”. Haga un comentario sobre la relación entre ingreso familiar y nivel de estudios del jefe de familia.
33. Hace poco los administradores de un campo de golf recibieron algunas quejas acerca de las condiciones de los *greens*. Varios jugadores se quejaron de que estaban demasiado rápidos. En lugar de reaccionar a los comentarios de unos cuantos, la asociación de golf realizó un sondeo con 100 jugadoras y 100 jugadores. Los resultados del sondeo se presentan a continuación.

Jugadores			Jugadoras		
Hándicap	Condición de los greens		Hándicap	Condición de los greens	
	Demasiado rápido	Bien		Demasiado rápido	Bien
Menos de 15	10	40	Menos de 15	1	9
15 o más	25	25	15 o más	39	51

- Combine estas dos tabulaciones cruzadas utilizando como encabezados de renglón Jugadores y Jugadoras y como encabezados de columnas Demasiado rápido y Bien. ¿En qué grupo se encuentra el mayor porcentaje de los que dicen que los *greens* están demasiado rápidos?

- b. Vuelva a las tabulaciones cruzadas iniciales. De los jugadores con bajo hándicap (mejores jugadores), ¿en qué grupo (jugadoras o jugadores) se encuentra un porcentaje mayor de quienes dicen que los *greens* están demasiado rápidos?
 - c. Regrese a las tabulaciones cruzadas iniciales. De los jugadores con alto hándicap, ¿en qué grupo (jugadoras o jugadores) se encuentra un porcentaje mayor para quienes los *greens* están demasiado rápidos?
 - d. ¿Qué conclusiones obtiene acerca de mujeres y hombres respecto a la velocidad de los *greens*? ¿Las conclusiones que obtuvo en el inciso a son consistentes con los incisos b y c? Explique cualquier inconsistencia aparente.
34. En la tabla 2.13 se presentan datos financieros de 36 empresas de una muestra cuyas acciones cotizan en la bolsa de valores de Nueva York (*Investor's Business Daily*, 7 de abril de 2000). Los datos de la columna Ventas/margen/ROE son evaluaciones financieras compuestas que se basan en la tasa de crecimiento de las ventas de una empresa, su margen de ganancia y su rendimiento de los activos (ROE *return on capital employed*). La calificación EPS es una medida del crecimiento por acción.

TABLA 2.13 DATOS FINANCIEROS DE 36 EMPRESAS QUE CONFORMAN UNA MUESTRA

Empresa	EPS	Fuerza relativa del precio	Fuerza relativa del grupo de industrias	Ventas/margen/ ROE
Advo	81	74	B	A
Alaska Air Group	58	17	C	B
Alliant Tech	84	22	B	B
Atmos Energy	21	9	C	E
Bank of Am.	87	38	C	A
Bowater PLC	14	46	C	D
Callaway Golf	46	62	B	E
Central Parking	76	18	B	C
Dean Foods	84	7	B	C
Dole Food	70	54	E	C
Elec. Data Sys.	72	69	A	B
Fed. Dept. Store	79	21	D	B
Gateway	82	68	A	A
Goodyear	21	9	E	D
Hanson PLC	57	32	B	B
ICN Pharm.	76	56	A	D
Jefferson Plt.	80	38	D	C
Kroger	84	24	D	A
Mattel	18	20	E	D
McDermott	6	6	A	C
Monaco	97	21	D	A
Murphy Oil	80	62	B	B
Nordstrom	58	57	B	C
NYMAGIC	17	45	D	D
Office Depot	58	40	B	B
Payless Shoes	76	59	B	B
Praxair	62	32	C	B
Reebok	31	72	C	E
Safeway	91	61	D	A
Teco Energy	49	48	D	B
Texaco	80	31	D	C
US West	60	65	B	A
United Rental	98	12	C	A
Wachovia	69	36	E	B
Winnebago	83	49	D	A
York International	28	14	D	B

Fuente: *Investor's Business Daily*, 7 de abril de 2000.

- a. Elabore una tabulación cruzada con los datos Ventas/margen/ROE (renglones) y EPS (columnas). Para el EPS emplee las clases 0–19, 20–39, 40–59, 60–79 y 80–99.
 - b. Calcule los porcentajes de las columnas y haga un comentario sobre la relación entre las variables.
35. Regrese a la tabla 2.13.
- a. Elabore una tabulación cruzada con los datos Ventas/margen/ROE y Fuerza relativa del grupo de industrias.
 - b. Elabore una distribución de frecuencia de los datos Ventas/margen/ROE.
 - c. Elabore una distribución de frecuencia de los datos Fuerza relativa del grupo de industrias.
 - d. ¿Le ayudó la tabulación cruzada en la elaboración de las distribuciones de frecuencia de los incisos b y c?
36. De nuevo, a la tabla 2.13.
- a. Elabore un diagrama de dispersión con los datos EPS y Fuerza relativa del precio.
 - b. Haga un comentario sobre la relación entre las variables. (El significado del EPS se describe en el ejercicio 34. La Fuerza relativa del precio es una medida de la variación en el precio de una acción en los últimos 12 meses. Valores altos indican gran variación.)
37. La National Football League de Estados Unidos evalúa a los candidatos posición por posición con una escala que va de 5 a 9. La evaluación se interpreta como sigue: 8–9 debe empezar el primer año; 7.0–7.9 debe empezar; 6.0–6.9 será un apoyo para el equipo, y 5.0–5.9 puede pertenecer al club y contribuir. En la tabla 2.14 se presentan posición, peso, tiempo (segundos en correr 40 yardas), y evaluación de 40 candidatos (*USA Today*, 14 de abril de 2000).
- a. Con los datos posición (renglones) y tiempo (columnas) elabore una tabulación cruzada. Para el tiempo emplee las clases 4.00–4.49, 4.50–4.99, 5.00–5.49 y 5.50–5.99.
 - b. Haga un comentario acerca de la relación entre posición y tiempo, con base en la tabulación cruzada que elaboró en el inciso a.
 - c. Con los datos tiempo y calificación obtenida en la evaluación elabore un diagrama de dispersión, coloque la calificación obtenida en la evaluación en el eje vertical.
 - d. Haga un comentario sobre la relación entre tiempo y calificación obtenida en la evaluación.

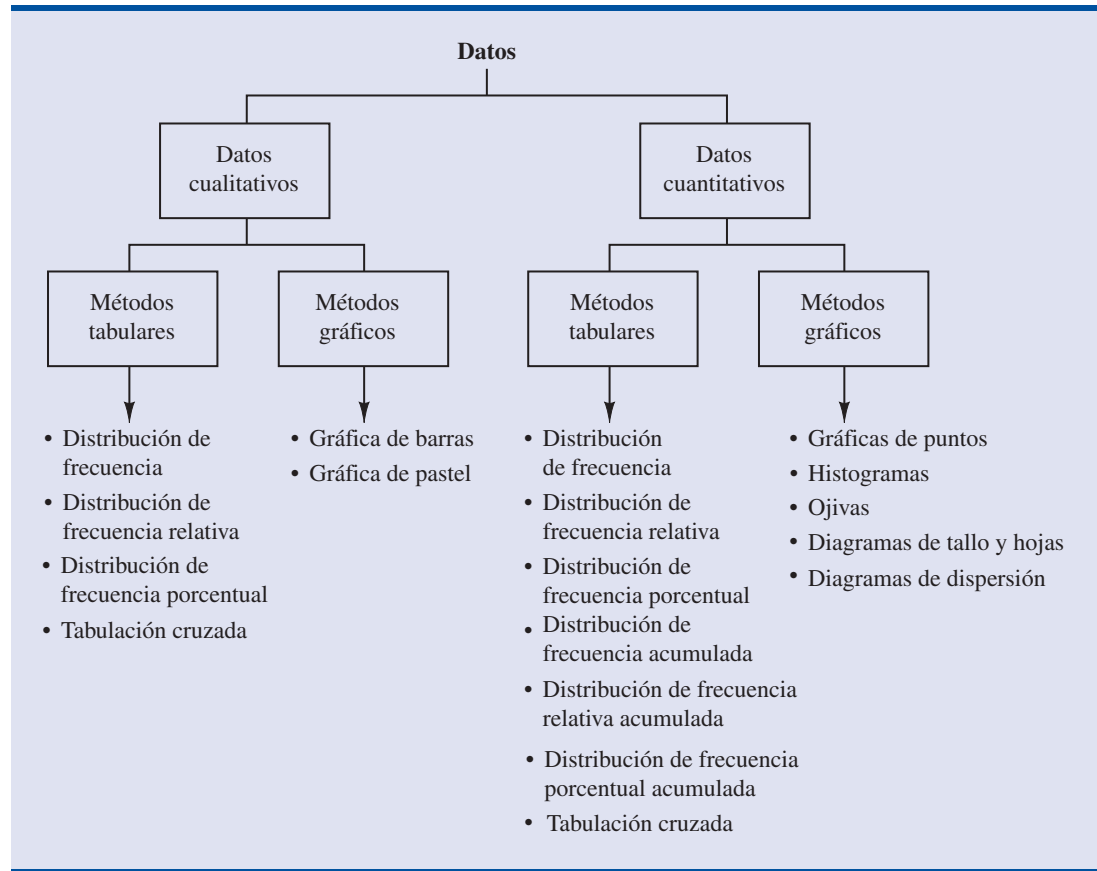
Resumen

Un conjunto de datos, aunque sea de tamaño modesto, es difícil de interpretar con los datos tal y como se han recolectado. Los métodos tabulares y los métodos gráficos permiten organizar y resumir los datos para que muestren algún patrón y sean factibles de interpretación. Para resumir datos cualitativos se presentaron las distribuciones de frecuencia, de frecuencia relativa y las de frecuencia porcentual, las gráficas de barras y las gráficas de pastel. Las distribuciones de frecuencia, de frecuencia relativa, de frecuencia porcentual, los histogramas, las distribuciones de frecuencia acumulada, de frecuencia relativa acumulada, de frecuencia porcentual acumulada y las ojivas se presentaron como métodos para resumir datos cuantitativos. Los diagramas de tallo y hojas son una técnica para el análisis exploratorio de datos que se usa para resumir datos cuantitativos. La tabulación cruzada se presentó como un método para resumir datos para dos variables. Los diagramas de dispersión se presentaron como un método gráfico para mostrar la relación entre dos variables cuantitativas. En la figura 2.9 se resumen los métodos tabulares y gráficos que se presentaron en este capítulo.

Cuando se tienen grandes conjuntos de datos es indispensable usar paquetes de software para la elaboración de resúmenes tabulares o gráficos de los datos. En los dos apéndices de este capítulo se explica el uso de Minitab y de Excel con tal propósito.

TABLA 2.14 DATOS DE 40 CANDIDATOS A LA NATIONAL FOOTBALL LEAGUE DE ESTADOS UNIDOS

Observación	Nombre	Posición	Peso	Tiempo	Evaluación
1	Peter Warrick	Receptor abierto	194	4.53	9
2	Plaxico Burress	Receptor abierto	231	4.52	8.8
3	Sylvester Morris	Receptor abierto	216	4.59	8.3
4	Travis Taylor	Receptor abierto	199	4.36	8.1
5	Laveranues Coles	Receptor abierto	192	4.29	8
6	Dez White	Receptor abierto	218	4.49	7.9
7	Jerry Porter	Receptor abierto	221	4.55	7.4
8	Ron Dugans	Receptor abierto	206	4.47	7.1
9	Todd Pinkston	Receptor abierto	169	4.37	7
10	Dennis Northcutt	Receptor abierto	175	4.43	7
11	Anthony Lucas	Receptor abierto	194	4.51	6.9
12	Darrell Jackson	Receptor abierto	197	4.56	6.6
13	Danny Farmer	Receptor abierto	217	4.6	6.5
14	Sherrod Gideon	Receptor abierto	173	4.57	6.4
15	Trevor Gaylor	Receptor abierto	199	4.57	6.2
16	Cosey Coleman	Guardia	322	5.38	7.4
17	Travis Claridge	Guardia	303	5.18	7
18	Kaulana Noa	Guardia	317	5.34	6.8
19	Leander Jordan	Guardia	330	5.46	6.7
20	Chad Clifton	Guardia	334	5.18	6.3
21	Manula Savea	Guardia	308	5.32	6.1
22	Ryan Johanningmeir	Guardia	310	5.28	6
23	Mark Tauscher	Guardia	318	5.37	6
24	Blaine Saipaia	Guardia	321	5.25	6
25	Richard Mercier	Guardia	295	5.34	5.8
26	Damion McIntosh	Guardia	328	5.31	5.3
27	Jeno James	Guardia	320	5.64	5
28	Al Jackson	Guardia	304	5.2	5
29	Chris Samuels	Tacle ofensivo	325	4.95	8.5
30	Stockar McDougale	Tacle ofensivo	361	5.5	8
31	Chris McIngosh	Tacle ofensivo	315	5.39	7.8
32	Adrian Klemm	Tacle ofensivo	307	4.98	7.6
33	Todd Wade	Tacle ofensivo	326	5.2	7.3
34	Marvel Smith	Tacle ofensivo	320	5.36	7.1
35	Michael Thompson	Tacle ofensivo	287	5.05	6.8
36	Bobby Williams	Tacle ofensivo	332	5.26	6.8
37	Darnell Alford	Tacle ofensivo	334	5.55	6.4
38	Terrance Beadles	Tacle ofensivo	312	5.15	6.3
39	Tutan Reyes	Tacle ofensivo	299	5.35	6.1
40	Greg Robinson-Ran	Tacle ofensivo	333	5.59	6

FIGURA 2.9 MÉTODOS TABULARES Y GRÁFICOS PARA RESUMIR DATOS

Glosario

Datos cualitativos Etiquetas o nombres que se usan para identificar las categorías de elementos semejantes.

Datos cuantitativos Valores numéricos que indican cuánto o cuántos.

Distribución de frecuencia Resumen tabular de datos que muestra el número (frecuencia) de los datos que pertenecen a cada una de varias clases disyuntas.

Distribución de frecuencia relativa Resumen tabular de datos que muestra la proporción o la fracción de datos propios de cada una de varias clases disyuntas.

Distribución de frecuencia porcentual Resumen tabular de datos que muestra el porcentaje de datos que corresponden a cada una de varias clases disyuntas.

Gráfica de barras Gráfica para representar datos cualitativos que hayan sido resumidos en una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual.

Gráfica de pastel Gráfica para representar datos resumidos mediante una distribución de frecuencia relativa y que se basa en la subdivisión de un círculo en sectores que corresponden a la frecuencia relativa de las clases.

Punto medio de clase Valor que se encuentra a la mitad entre el límite de clase inferior y el límite de clase superior.

Gráfica de puntos Gráfica que resume datos mediante la cantidad de puntos sobre los valores de los datos que se encuentran en un eje horizontal.

Histograma Representación gráfica de una distribución de frecuencia, de frecuencia relativa o de frecuencia porcentual que se construye colocando los intervalos de clase sobre un eje horizontal y la frecuencia, la frecuencia relativa o la frecuencia porcentual sobre un eje vertical.

- Distribución de frecuencia acumulada** Síntesis tabular de datos cuantitativos, en la que se muestra el número de datos que son menores o iguales que el límite superior de cada clase.
- Distribución de frecuencia relativa acumulada** Resumen tabular de datos cuantitativos, en el que se muestra la proporción o fracción de datos que son menores o iguales que el límite superior de cada clase.
- Distribución de frecuencia porcentual acumulada** Síntesis tabular de datos cuantitativos, en la que se muestra el porcentaje de datos que son menores o iguales que el límite superior de cada clase.
- Ojiva** Gráfica de una distribución acumulada.
- Análisis exploratorio de datos** Métodos en los que se emplean cálculos aritméticos sencillos y gráficas fáciles de elaborar para resumir datos en forma rápida.
- Diagrama de tallo y hojas** Técnica para el análisis exploratorio de datos que tanto ordena por jerarquía datos cuantitativos como proporciona claridad acerca de la forma de la distribución.
- Tabulación cruzada** Resumen tabular de datos de dos variables. Las clases de una de las variables se representan como renglones; las clases de la otra variable como columnas.
- Paradoja de Simpson** Conclusiones que se obtienen de dos o más tabulaciones cruzadas y que se invierten cuando se agregan los datos en una sola tabulación cruzada.
- Diagrama de dispersión** Representación gráfica de la relación entre dos variables cuantitativas. A una variable se le asigna un eje horizontal y a la otra un eje vertical.
- Línea de tendencia** Línea que da una aproximación de la relación entre dos variables.

Fórmulas clave

Frecuencia relativa

$$\frac{\text{Frecuencia de la clase}}{n}$$

(2.1)

Ancho aproximado de clase

$$\frac{\text{Dato mayor} - \text{Dato menor}}{\text{Número de clases}}$$

(2.2)

Ejercicios complementarios

38. Los cinco automóviles más vendidos en Estados Unidos durante 2003 fueron la camioneta Chevrolet Silverado/C/K, la camioneta Dodge Ram, la camioneta Ford F-Series, el Honda Accord y el Toyota Camry (*Motor Trend*, 2003). En la tabla 2.15 se presenta una muestra de 50 compras de automóviles.

TABLA 2.15 DATOS DE 50 COMPRAS DE AUTOMÓVILES

Silverado	Ram	Accord	Camry	Camry
Silverado	Silverado	Camry	Ram	F-Series
Ram	F-Series	Accord	Ram	Ram
Silverado	F-Series	F-Series	Silverado	Ram
Ram	Ram	Accord	Silverado	Camry
F-Series	Ram	Silverado	Accord	Silverado
Camry	F-Series	F-Series	F-Series	Silverado
F-Series	Silverado	F-Series	F-Series	Ram
Silverado	Silverado	Camry	Camry	F-Series
Silverado	F-Series	F-Series	Accord	Accord





- a. Elabore una distribución de frecuencia y otra de frecuencia porcentual.
b. ¿Cuál es la camioneta y el automóvil de pasajeros más vendidos?
c. Haga una gráfica de pastel.
39. El Higher Education Research Institute de UCLA cuenta con estadísticas sobre las áreas que son más elegidas por los estudiantes de nuevo ingreso. Las cinco más elegidas son arte y humanidades (A), administración de negocios (B), ingeniería (E), política (P) y ciencias sociales (S) (*The New York Times Almanac*, 2006). Otras áreas (O), entre las que se encuentran biología, física, ciencias de la computación y educación se agruparon todas en una sola categoría. Las siguientes fueron las áreas elegidas por 64 estudiantes de recién ingreso de una muestra.

S	P	P	O	B	E	O	E	P	O	O	B	O	O	O	A
O	E	E	B	S	O	B	O	A	O	E	O	E	O	B	P
B	A	S	O	E	A	B	O	S	S	O	O	E	B	O	B
A	E	B	E	A	A	P	O	O	E	O	B	B	O	P	B

- a. Dé una distribución de frecuencia y otra de frecuencia porcentual.
b. Elabore una gráfica de barras.
c. ¿Que porcentaje de los estudiantes de nuevo ingreso elige una de las cinco áreas más elegidas?
d. ¿Cuál es el área más elegida por los estudiantes de nuevo ingreso? ¿Qué porcentaje de los estudiantes de nuevo ingreso elige esta área?
40. A los 100 mejores entrenadores de golf la revista *Golf Magazine* les preguntó, “¿Cuál es el aspecto más relevante que impide a los jugadores de golf desarrollar todo su potencial?” Las respuestas fueron falta de precisión, técnica de golpe inadecuada, actitud mental inadecuada, falta de energía, práctica insuficiente, tiro al hoyo inadecuado, juego corto inadecuado y estrategia de decisión inadecuada. A continuación se presentan los datos obtenidos (*Golf Magazine*, febrero de 2002):



Actitud mental	Actitud mental	Juego corto	Juego corto	Juego corto
Práctica	Precisión	Actitud mental	Precisión	Tiro al hoyo
Energía	Técnica de golpe	Precisión	Juego corto	Tiro al hoyo
Precisión	Actitud mental	Actitud mental	Precisión	Energía
Precisión	Precisión	Juego corto	Energía	Juego corto
Precisión	Tiro al hoyo	Actitud mental	Estrategia de decisión	Precisión
Juego corto	Energía	Actitud mental	Técnica de golpe	Juego corto
Práctica	Práctica	Actitud mental	Energía	Energía
Actitud mental	Juego corto	Actitud mental	Juego corto	Estrategia de decisión
Precisión	Juego corto	Precisión	Actitud mental	Juego corto
Actitud mental	Tiro al hoyo	Actitud mental	Actitud mental	Tiro al hoyo
Práctica	Tiro al hoyo	Práctica	Juego corto	Tiro al hoyo
Energía	Actitud mental	Juego corto	Práctica	Estrategia de decisión
Precisión	Juego corto	Precisión	Práctica	Tiro al hoyo
Precisión	Juego corto	Precisión	Juego corto	Tiro al hoyo
Precisión	Técnica de golpe	Juego corto	Actitud mental	Práctica
Juego corto	Juego corto	Estrategia de decisión	Juego corto	Juego corto
Práctica	Práctica	Juego corto	Práctica	Estrategia de decisión
Actitud mental	Estrategia de decisión	Estrategia de decisión	Energía	Juego corto
Precisión	Práctica	Práctica	Práctica	Precisión

- a. Elabore una distribución de frecuencia y otra de frecuencia porcentual.
b. ¿Cuáles son los aspectos más relevantes que impiden a un jugador de golf desarrollar su potencial?
41. El rendimiento de dividendos son los beneficios anuales que paga una empresa, expresado como porcentaje del precio de una acción ($\text{Dividendo/precio de la acción} \times 100$). En la tabla 2.16 se presenta el rendimiento de dividendos de las empresas del promedio industrial Dow Jones (*The Wall Street Journal*, 3 de marzo de 2006).
- a. Haga una distribución de frecuencia y una distribución de frecuencia porcentual.
b. Haga un histograma.
c. Aporte un comentario sobre la forma de la distribución.

TABLA 2.16 RENDIMIENTO DE DIVIDENDOS DE LAS EMPRESAS DEL PROMEDIO INDUSTRIAL DOW JONES.

Empresa	Rendimiento de dividendos	Empresa	Rendimiento de dividendos
AIG	0.9	Home Depot	1.4
Alcoa	2.0	Honeywell	2.2
Altria Group	4.5	IBM	1.0
American Express	0.9	Intel	2.0
AT&T	4.7	Johnson & Johnson	2.3
Boeing	1.6	JPMorgan Chase	3.3
Caterpillar	1.3	McDonald's	1.9
Citigroup	4.3	Merck	4.3
Coca-Cola	3.0	Microsoft	1.3
Disney	1.0	3M	2.5
DuPont	3.6	Pfizer	3.7
ExxonMobil	2.1	Procter & Gamble	1.9
General Electric	3.0	United Technologies	1.5
General Motors	5.2	Verizon	4.8
Hewlett-Packard	0.9	Wal-Mart Stores	1.3

- d. ¿Qué indican los resúmenes tabular y gráfico acerca de los dividendos de las empresas del promedio industrial Dow Jones?
- e. ¿Qué empresa tiene el más alto rendimiento de dividendos? Si hoy el precio de las acciones de esta empresa es \$20 por acción y usted compra 500 acciones, ¿cuál será el ingreso por dividendos que genere anualmente esta inversión?
42. Cada año en Estados Unidos, aproximadamente 1.5 millones de los estudiantes de educación superior presentan un examen de aptitud escolar (SAT, por sus siglas en inglés). Cerca de 80% de las universidades e instituciones de educación superior emplean las puntuaciones obtenidas por los estudiantes en este examen como criterio de admisión (*College Board*, marzo de 2006). A continuación se presentan las puntuaciones obtenidas en las áreas de matemáticas y expresión verbal por una muestra de estudiantes.



1025	1042	1195	880	945
1102	845	1095	936	790
1097	913	1245	1040	998
998	940	1043	1048	1130
1017	1140	1030	1171	1035

- a. Presente una distribución de frecuencia y un histograma de estas puntuaciones. La primera clase debe empezar en la puntuación 750 y la amplitud de clase deberá ser 100.
- b. Dé un comentario sobre la forma de la distribución.
- c. ¿Qué otras observaciones puede hacer acerca de estas puntuaciones con base en los resúmenes tabulares y gráficos?
43. La Asociación estadounidense de inversionistas independientes informa sobre 94 acciones fantasma. El término *fantasma* se refiere a que son acciones de empresas pequeñas o medianas que no son seguidas de cerca por las principales casas de bolsa. A continuación se presenta, de una muestra de 20 acciones fantasma, información sobre el lugar donde se comercializa la acción —bolsa



Acción	Bolsa de cambio	Ganancia por acción (\$)	Relación Precio/ganancia
Chemi-Trol	OTC	0.39	27.30
Candie's	OTC	0.07	36.20
TST/Impreso	OTC	0.65	12.70

(continúa)

Acción	Bolsa de cambio	Ganancia por acción	Relación precio/ganancia
Unimed Pharm.	OTC	0.12	59.30
Skyline Chili	AMEX	0.34	19.30
Cyanotech	OTC	0.22	29.30
Catalina Light.	NYSE	0.15	33.20
DDL Elect.	NYSE	0.10	10.20
Euphonix	OTC	0.09	49.70
Mesa Labs	OTC	0.37	14.40
RCM Tech.	OTC	0.47	18.60
Anuhco	AMEX	0.70	11.40
Hello Direct	OTC	0.23	21.10
Hilite Industries	OTC	0.61	7.80
Alpha Tech.	OTC	0.11	34.60
Wegener Group	OTC	0.16	24.50
U.S. Home & Garden	OTC	0.24	8.70
Chalone Wine	OTC	0.27	44.40
Eng. Support Sys.	OTC	0.89	16.70
Int. Remote Imaging	AMEX	0.86	4.70

de Nueva York (NYSE), American Stock Exchange (AMEX) o directamente (OTC)— la ganancia por acción y la relación precio/ganancia.

- Con los datos de bolsa de cambio haga una distribución de frecuencia y otra de frecuencia relativa. ¿Cuál tiene más acciones fantasma?
 - Con los datos ganancia por acción y relación precio/ganancia elabore distribuciones de frecuencia y de frecuencia relativa. Para las ganancias por acción emplee las clases 0.00–0.19, 0.20–0.39, etc.; para la relación precio/ganancia use las clases 0.0–9.9, 10.0–19.9, etc. ¿Qué observaciones y comentarios puede hacer acerca de las acciones fantasma?
44. Los datos siguientes de la oficina de los censos de Estados Unidos proporcionan la población en millones de personas por estado (*The World Almanac*, 2006).

Estado	Población	Estado	Población	Estado	Población
Alabama	4.5	Louisiana	4.5	Ohio	11.5
Alaska	0.7	Maine	1.3	Oklahoma	3.5
Arizona	5.7	Maryland	5.6	Oregon	3.6
Arkansas	2.8	Massachusetts	6.4	Pennsylvania	12.4
California	35.9	Michigan	10.1	Rhode Island	1.1
Colorado	4.6	Minnesota	5.1	South Carolina	4.2
Connecticut	3.5	Mississippi	2.9	South Dakota	0.8
Delaware	0.8	Missouri	5.8	Tennessee	5.9
Florida	17.4	Montana	0.9	Texas	22.5
Georgia	8.8	Nebraska	1.7	Utah	2.4
Hawai	1.3	Nevada	2.3	Vermont	0.6
Idaho	1.4	New Hampshire	1.3	Virginia	7.5
Illinois	12.7	New Jersey	8.7	Washington	6.2
Indiana	6.2	New Mexico	1.9	West Virginia	1.8
Iowa	3.0	New York	19.2	Wisconsin	5.5
Kansas	2.7	North Carolina	8.5	Wyoming	0.5
Kentucky	4.1	North Dakota	0.6		



- Elabore una distribución de frecuencia, una de frecuencia porcentual y un histograma. Use como ancho de clase 2.5 millones.
- Explique el sesgo de la distribución.
- ¿Qué observaciones puede hacer acerca de la población en los 50 estados?

45. *Drug Store News* (septiembre de 2002) proporciona datos sobre ventas de medicamentos de las principales farmacias de Estados Unidos. Los datos siguientes son ventas anuales en millones.

Farmacia	Ventas	Farmacia	Ventas
Ahold USA	\$ 1 700	Medicine Shoppe	\$ 1 757
CVS	12 700	Rite-Aid	8 637
Eckerd	7 739	Safeway	2 150
Kmart	1 863	Walgreens	11 660
Kroger	3 400	Wal-Mart	7 250

- Dé un diagrama de tallo y hojas.
 - Indique cuáles son las ventas anuales menores, mayores e intermedias.
 - ¿Cuáles son las dos farmacias mayores?
46. A continuación se presentan las temperaturas diarias más altas y más bajas registradas en 20 ciudades de Estados Unidos (*USA Today*, 3 de marzo 2006).

Ciudad	Alta	Baja	Ciudad	Alta	Baja
Albuquerque	66	39	Los Angeles	60	46
Atlanta	61	35	Miami	84	65
Baltimore	42	26	Minneapolis	30	11
Charlotte	60	29	New Orleans	68	50
Cincinnati	41	21	Oklahoma City	62	40
Dallas	62	47	Phoenix	77	50
Denver	60	31	Portland	54	38
Houston	70	54	St. Louis	45	27
Indianapolis	42	22	San Francisco	55	43
Las Vegas	65	43	Seattle	52	36

- Con las temperaturas altas elabore un diagrama de tallo y hojas.
 - Con las temperaturas bajas elabore un diagrama de tallo y hojas.
 - Compare los dos diagramas y haga comentarios acerca de las diferencias entre las temperaturas más altas y las más bajas.
 - Proporcione una distribución de frecuencia de las temperaturas más altas y de las más bajas.
47. Vuelva al conjunto de datos sobre las temperaturas más altas y las temperaturas más bajas en 20 ciudades presentado en el ejercicio 46.
- Elabore un diagrama de dispersión que muestre la relación entre las dos variables, temperatura más alta y temperatura más baja.
 - Aporte sus comentarios sobre la relación entre las temperaturas más elevadas y las más bajas.
48. Se realizó un estudio sobre satisfacción en el empleo en cuatro ocupaciones. La satisfacción en el empleo se midió mediante un cuestionario de 18 puntos en el que a cada punto había que calificarlo con una escala del 1 al 5; las puntuaciones más altas correspondían a mayor satisfacción en el empleo. La suma de las calificaciones dadas a los 18 puntos proporcionaba una medida de



Ocupación	Satisfacción	Ocupación	Satisfacción	Ocupación	Satisfacción
Abogado	42	Terapeuta físico	78	Analista de sistemas	60
Terapeuta físico	86	Analista de sistemas	44	Terapeuta físico	59
Abogado	42	Analista de sistemas	71	Ebanista	78
Analista de sistemas	55	Abogado	50	Terapeuta físico	60

(continúa)

Ocupación	Satisfacción	Ocupación	Satisfacción	Ocupación	Satisfacción
Abogado	38	Abogado	48	Terapeuta físico	50
Ebanista	79	Ebanista	69	Ebanista	79
Abogado	44	Terapeuta físico	80	Analista de sistemas	62
Analista de sistemas	41	Analista de sistemas	64	Abogado	45
Terapeuta físico	55	Terapeuta físico	55	Ebanista	84
Analista de sistemas	66	Ebanista	64	Terapeuta físico	62
Abogado	53	Ebanista	59	Analista de sistemas	73
Ebanista	65	Ebanista	54	Ebanista	60
Abogado	74	Analista de sistemas	76	Abogado	64
Terapeuta físico	52				

la satisfacción en el empleo de cada uno de los individuos de la muestra. Los datos obtenidos fueron los siguientes.

- Dé una tabulación cruzada para ocupación y satisfacción en el trabajo.
 - En la tabulación cruzada del inciso a calcule los porcentajes de renglones.
 - ¿Qué observaciones puede hacer respecto a la satisfacción en el trabajo en estas ocupaciones?
49. ¿Generan más ingresos las grandes empresas? Los datos siguientes muestran la cantidad de empleados y el ingreso anual de 20 de las empresas de *Fortune* 1000 (*Fortune*, 17 de abril de 2000).



Empresa	Empleados	Ingreso (en millones de \$)	Empresa	Empleados	Ingreso (en millones de \$)
Sprint	77 600	19 930	American Financial	9 400	3 334
Chase Manhattan	74 801	33 710	Fluor	53 561	12 417
Computer Sciences	50 000	7 660	Phillips Petroleum	15 900	13 852
Wells Fargo	89 355	21 795	Cardinal Health	36 000	25 034
Sunbeam	12 200	2 398	Borders Group	23 500	2 999
CBS	29 000	7 510	MCI Worldcom	77 000	37 120
Time Warner	69 722	27 333	Consolidated Edison	14 269	7 491
Steelcase	16 200	2 743	IBP	45 000	14 075
Georgia-Pacific	57 000	17 796	Super Value	50 000	17 421
Toro	1 275	4 673	H&R Block	4 200	1 669

- Haga un diagrama de dispersión para mostrar la relación entre las variables ingreso y empleados.
 - Haga un comentario sobre la relación entre estas variables.
50. En un sondeo realizado entre los edificios comerciales que son clientes de Cincinnati Gas & Electric Company se preguntaba cuál era el principal combustible que empleaban para la calefacción y en qué año se había construido el edificio. A continuación se presenta una parte del diagrama cruzado que se obtuvo con los datos.

Año de construcción	Tipo de combustible				
	Electricidad	Gas natural	Petróleo	Propano	Otros
1973 o antes	40	183	12	5	7
1974–1979	24	26	2	2	0
1980–1986	37	38	1	0	6
1987–1991	48	70	2	0	1

- a. Complete esta tabulación cruzada dando los totales de los renglones y de las columnas.
 - b. Dé las distribuciones de frecuencia de año de construcción y de tipo de combustible empleado.
 - c. Haga una tabulación cruzada en la que se muestren los porcentajes de columnas.
 - d. Elabore una tabulación cruzada en la que se muestren los porcentajes de renglones.
 - e. Comente acerca de la relación entre año de construcción y tipo de combustible empleado.
51. La tabla 2.17 contiene parte de los datos que se encuentran en el archivo titulado Fortune en el disco compacto que viene con el libro. Este archivo proporciona fondos propios, valor de mercado y ganancias de las 50 empresas en una muestra de *Fortune 500*.

TABLA 2.17 DATOS EN UNA MUESTRA DE 50 EMPRESAS DE *FORTUNE 500*

Empresa	Fondos propios (en miles de \$)	Valor de mercado (en miles de \$)	Ganancias (en miles de \$)
AGCO	982.1	372.1	60.6
AMP	2 698.0	12 017.6	2.0
Apple Computer	1 642.0	4 605.0	309.0
Baxter International	2 839.0	21 743.0	315.0
Bergen Brunswick	629.1	2 787.5	3.1
Best Buy	557.7	10 376.5	94.5
Charles Schwab	1 429.0	35 340.6	348.5
.	.	.	.
.	.	.	.
.	.	.	.
Walgreen	2 849.0	30 324.7	511.0
Westvaco	2 246.4	2 225.6	132.0
Whirlpool	2 001.0	3 729.4	325.0
Xerox	5 544.0	35 603.7	395.0



- a. Con las variables fondos propios y ganancia elabore una tabulación cruzada. Para las ganancias emplee las clases 0–200, 200–400, ..., 1000–1200. Para fondos propios emplee las clases 0–1200, 1200–2400, ..., 4800–6000.
 - b. En la tabulación cruzada del inciso a calcule los porcentajes de renglón.
 - c. ¿Observa alguna relación entre ganancia y fondos propios?
52. Vuelva a la tabla 2.17.
- a. Con las variables valor de mercado y ganancia elabore una tabulación cruzada.
 - b. En la tabulación cruzada del inciso a calcule los porcentajes de renglón.
 - c. Haga un comentario sobre la relación entre las variables.
53. Vuelva a la tabla 2.17.
- a. Elabore un diagrama de dispersión que muestre la relación entre las variables ganancia y fondos propios.
 - b. Haga un comentario sobre la relación entre las variables.
54. Vuelva a la tabla 2.17.
- a. Elabore un diagrama de dispersión que muestre la relación entre las variables valor de mercado y fondos propios.
 - b. Haga un comentario sobre la relación entre las variables.

Caso problema 1 Las tiendas Pelican

Las tiendas Pelican, una división de National Clothing, es una cadena de tiendas de ropa para mujer que tiene sucursales por todo Estados Unidos. Hace poco la tienda realizó una promoción en la que envió cupones de descuento a todos los clientes de otras tiendas de National Clothing. Los datos obtenidos en una muestra de 100 pagos con tarjeta de crédito en las tiendas Pelican durante un día de la promoción se presentan en el archivo titulado PelicanStores. En la tabla 2.18 se mues-

TABLA 2.18 DATOS DE 100 COMPRAS CON TARJETA DE CRÉDITO REALIZADAS EN LAS TIENDAS PELICAN

Cliente	Tipo de cliente	Artículos	Ventas netas	Modo de pago	Género	Estado civil	Edad
1	Regular	1	39.50	Discover	Masculino	Casado	32
2	Promocional	1	102.40	Proprietary Card	Femenino	Casada	36
3	Regular	1	22.50	Proprietary Card	Femenino	Casada	32
4	Promocional	5	100.40	Proprietary Card	Femenino	Casada	28
5	Regular	2	54.00	MasterCard	Femenino	Casada	34
.
.
.
96	Regular	1	39.50	MasterCard	Femenino	Casada	44
97	Promocional	9	253.00	Proprietary Card	Femenino	Casada	30
98	Promocional	10	287.59	Proprietary Card	Femenino	Casada	52
99	Promocional	2	47.60	Proprietary Card	Femenino	Casada	30
100	Promocional	1	28.44	Proprietary Card	Femenino	Casada	44

tra parte de este conjunto de datos. El modo de pago Proprietary card se refiere a pagos realizados usando una tarjeta de crédito de National Clothing. A los clientes que hicieron compras usando un cupón de descuento se les denomina aquí promocionales y a quienes hicieron sus compras sin emplear cupón de descuento se les denomina regulares. Como a los clientes de las tiendas Pelican no se les enviaron cupones promocionales, los directivos consideran que las ventas hechas a quienes presentaron un cupón de descuento son ventas que de otro modo no se hubieran hecho. Es claro que Pelican espera que los clientes promocionales continúen comprando con ellos.

La mayor parte de las variables que aparecen en la tabla 2.18 se explican por sí mismas, pero dos de las variables deben ser aclaradas.

Artículos	El número total de artículos comprados
Ventas netas	Cantidad total cargada a la tarjeta de crédito

Los directivos de Pelican desean emplear estos datos muestrales para tener información acerca de sus clientes y para evaluar la promoción utilizando los cupones de descuento.

Informe para los directivos

Emplee los métodos tabulares y gráficos de la estadística descriptiva para ayudar a los directivos de Pelican a elaborar un perfil de sus clientes y a evaluar la promoción. Su informe debe contener, por lo menos, lo siguiente:

1. Distribuciones de frecuencia porcentual de las variables clave.
2. Una gráfica de barras o una gráfica de pastel que muestre el número de clientes correspondiente a cada modo de pago.
3. Una tabulación cruzada con el tipo de cliente (regular o promocional) frente a ventas netas. Haga un comentario sobre las semejanzas o diferencias que observe.
4. Un diagrama de dispersión para investigar la relación entre ventas netas y edad del cliente.

Caso problema 2 Industria cinematográfica

La industria cinematográfica es un negocio muy competido. En más de 50 estudios se producen de 300 a 400 películas por año y el éxito financiero de estas películas varía considerablemente. Las variables usuales para medir el éxito de una película son ventas brutas (en millones de \$) en el fin de semana del estreno, ventas brutas totales (en millones de \$), número de salas en que se presenta la película, semanas en las que la película se encuentra entre las 60 mejores en ventas

TABLA 2.19 DATOS DEL ÉXITO DE 10 PELÍCULAS

Película	Ventas brutas en el estreno (en millones de \$)	Ventas brutas totales (en millones de \$)	Número de salas	Semanas en las 60 mejores
<i>Coach Carter</i>	29.17	67.25	2574	16
<i>Ladies in Lavender</i>	0.15	6.65	119	22
<i>Batman Begins</i>	48.75	205.28	3858	18
<i>Unleashed</i>	10.90	24.47	1962	8
<i>Pretty Persuasion</i>	0.06	0.23	24	4
<i>Fever Pitch</i>	12.40	42.01	3275	14
<i>Harry Potter and the Goblet of Fire</i>	102.69	287.18	3858	13
<i>Monster-in-Law</i>	23.11	82.89	3424	16
<i>White Noise</i>	24.11	55.85	2279	7
<i>Mr. and Mrs. Smith</i>	50.34	186.22	3451	21



brutas. Los datos de una muestra de 100 películas producidas en 2005 se encuentran en el archivo titulado Movies. La tabla 2.19 muestra los datos de las 10 primeras películas que se encuentran en este archivo.

Informe para los directivos

Emplee los métodos tabulares y gráficos de la estadística descriptiva para saber cómo contribuyen estas variables al éxito de una película. Su informe debe contener lo siguiente.

1. Resúmenes tabular y gráfico de las cuatro variables interpretando cada resumen acerca de la industria cinematográfica.
2. Un diagrama de dispersión para investigar la relación entre ventas brutas totales y ventas brutas en el fin de semana del estreno. Analícelo.
3. Un diagrama de dispersión para investigar la relación entre ventas brutas totales y número de salas. Analícelo.
4. Un diagrama de dispersión para investigar la relación entre ventas brutas totales y número de semanas entre las 60 mejores. Analícelo.

Apéndice 2.1 Uso de Minitab para presentaciones gráficas y tabulares

Minitab ofrece amplias posibilidades para la elaboración de resúmenes tabulares y gráficos de datos. Minitab se usa para elaborar diversos resúmenes gráficos y tabulaciones cruzadas. Los métodos gráficos son: gráfica de puntos, histograma, diagrama de tallo y hojas y diagrama de dispersión.

Gráficas de puntos

Para esta demostración emplee los datos de la tabla 2.4 sobre las duraciones de las auditorías. Los datos se encuentran en la columna C1 de la hoja de cálculo de Minitab. Con los pasos siguientes se generará una gráfica de puntos.



- Paso 1.** Seleccionar el menú **Graph** y elegir **Dotplot**
- Paso 2.** Seleccionar **One Y, Simple** y hacer clic en **OK**
- Paso 3.** Cuando aparezca el cuadro de diálogo de Dotplot-One Y, Simple:
Ingresar C1 en el cuadro **Graph Variables**.
Hacer clic en **OK**



Histograma

Empleando los datos de la tabla 2.4 sobre las duraciones de las auditorías se explicará cómo se construye un histograma con las frecuencias sobre el eje vertical. Los datos están en la columna C1 de la hoja de cálculo de Minitab. Con los pasos siguientes se generará un histograma de las duraciones de las auditorías.

- Paso 1.** Seleccionar el menú **Graph**
- Paso 2.** Elegir **Histogram**
- Paso 3.** Seleccionar **Simple** y hacer clic en **OK**
- Paso 4.** Cuando aparezca el cuadro de diálogo Histogram-Simple:
 - Ingresar C1 en el cuadro **Graph Variables**
 - Hacer clic en **OK**
- Paso 5.** Cuando aparezca el histograma:
 - Posicionar el cursor del mouse sobre cualquiera de las barras
 - Dar doble clic
- Paso 6.** Cuando aparezca el cuadro de diálogo Edit Bars:
 - Hacer clic en la pestaña **Binning**
 - Seleccionar **Cutpoint** en Interval Type
 - Seleccionar **Midpoint/Cutpoint positions** en Interval Definition
 - Ingresar 10:35/5 en el cuadro **Midpoint/Cutpoint positions***
 - Hacer clic en **OK**

Observe que Minitab también proporciona la posibilidad de mostrar los puntos medios de los rectángulos del histograma como escala en el eje x . Si se desea esta opción, se modifica el paso 6 seleccionando **Midpoint** en Interval Definition e ingresando 12:32/5 en el cuadro **Midpoint/Cutpoint positions**. Con estos pasos se obtiene el mismo histograma pero con los puntos medios, 12, 17, 22, 27 y 32, marcados en los rectángulos del histograma.

Diagrama de tallo y hojas



Se emplearán los datos de la tabla 2.8 sobre el examen de aptitudes para mostrar la construcción de un diagrama de tallo y hojas. Los datos se encuentran en la columna C1 de la hoja de cálculo de Minitab. Mediante los pasos siguientes se genera el diagrama extendido de tallo y hojas que se muestra en la sección 2.3.

- Paso 1.** Seleccionar el menú **Graph**
- Paso 2.** Elegir **Steam-and-Leaf**
- Paso 3.** Cuando aparezca el cuadro de diálogo Steam-and-Leaf:
 - Ingresar C1 en el cuadro **Graph Variables**
 - Hacer clic en **OK**

Diagrama de dispersión



Para demostrar la elaboración de un diagrama de dispersión se emplearán los datos de la tienda de equipos de sonido que se presentan en la tabla 2.12. Las semanas están numeradas del 1 al 10 en la columna C1, los datos del número de comerciales se encuentran en la columna C2 y los datos de las ventas están en la columna C3 de la hoja de cálculo de Minitab. Con los pasos siguientes se generará el diagrama de dispersión que se muestra en la figura 2.7.

*10:35/5 indica que 10 es el valor inicial del histograma, 35 es el valor final del histograma y 5 es el ancho de clase.

- Paso 1.** Seleccionar el menú **Graph**
- Paso 2.** Elegir **Scatterplot**
- Paso 3.** Elegir **Simple** y dar clic en **OK**
- Paso 4.** Cuando aparezca el cuadro de diálogo Scatterplot-Simple:
 Ingresar C3 bajo **Y variables** y C2 bajo **X variables**.
 Hacer clic en **OK**

Tabulación cruzada



Para demostrar la elaboración de una tabulación cruzada se usan los datos de *Zagat's Restaurant Review*, parte de los cuales se muestran en la tabla 2.9. Los restaurantes se encuentran numerados del 1 al 300 en la columna C1 de la hoja de cálculo de Minitab. Los datos sobre la calidad en la columna C2 y los precios en la columna C3.

Minitab sólo puede elaborar una tabulación cruzada con variables cualitativas y el precio es una variable cuantitativa. De manera que primero necesita codificar los precios especificando la clase a la que pertenece cada precio. Con los pasos siguientes se codificarán los precios haciendo cuatro clases de precios en la columna C4: \$10–19, \$20–29, \$30–39 y \$40–49.

- Paso 1.** Seleccionar el menú **Data**
- Paso 2.** Elegir **Code**
- Paso 3.** Elegir **Numeric to Text**
- Paso 4.** Cuando aparezca el cuadro de diálogo Code-Numeric to Text:
 Ingresar C3 en el cuadro **Code data from columns**
 Ingresar C4 en el cuadro **Into Columns**
 Ingresar 10:19 en el primer cuadro **Original values** y \$10–19 en el cuadro adyacente **New**
 Ingresar 20:29 en el primer cuadro **Original values** y \$20–29 en el cuadro adyacente **New**
 Ingresar 30:39 en el primer cuadro **Original values** y \$30–39 en el cuadro adyacente **New**
 Ingresar 40:49 en el primer cuadro **Original values** y \$40–49 en el cuadro adyacente **New**
 Hacer clic en **OK**

Para cada precio de la columna C3 aparecerá ahora su categoría correspondiente en la columna C4. Ahora puede elaborar una tabulación cruzada para calidad y categoría de los precios usando los datos de las columnas C2 y C4. Con los pasos siguientes se creará una tabulación cruzada que contendrá la misma información que la tabla 2.10.

- Paso 1.** Seleccionar el menú **Stat**
- Paso 2.** Elegir **Tables**
- Paso 3.** Elegir **Cross Tabulation and Chi-Square**
- Paso 4.** Cuando aparezcan los cuadros: Cross Tabulation y Chi-Square:
 Ingresar C2 en el cuadro **For rows** y C4 en el cuadro **For columns**
 Seleccionar **Counts**
 Hacer clic en **OK**

Apéndice 2.2 Uso de Excel para presentaciones gráficas y tabulares

Excel ofrece amplias posibilidades para la elaboración de resúmenes tabulares y gráficos de datos. En este capítulo se muestra cómo usar Excel para elaborar una distribución de frecuencia, gráficas de barras, gráficas de pastel, histogramas, tabulaciones cruzadas y diagramas de dispersión. Se presentan dos de las herramientas más potentes de Excel: el asistente para gráficos y el informe de tabla dinámica

Distribución de frecuencia y gráficas de barras con datos cualitativos

En esta sección se muestra el uso de Excel para la elaboración de una distribución de frecuencia y de una gráfica de barras con datos cualitativos. Ambas cosas se ilustran empleando los datos de la tabla 2.1 sobre ventas de refrescos.

Distribución de frecuencia Se empezará por mostrar el uso de la función COUNTIF para elaborar una distribución de frecuencia con los datos de la tabla 2.1. Consulte la figura 2.10 a medida que se presentan los pasos de esta explicación. La hoja de cálculo con las fórmulas (en la que se ven las funciones y fórmulas empleadas) aparece en segundo plano y la hoja de cálculo con los valores (en la que aparecen los resultados obtenidos con las funciones y fórmulas usadas) aparece en primer plano.

En las celdas A1:A51 se encuentra el título “Ventas de refrescos” y los datos de 50 ventas de refrescos. En las celdas C1:D1 también se ingresaron los títulos “Refresco” y “Frecuencia”. Los nombres de los cinco refrescos se ingresaron en las celdas C2:C6. Ahora se puede usar la función COUNTIF de Excel para contar cuántas veces aparece cada refresco en las celdas A2:A51. Para esto se siguen los pasos:

Paso 1. Seleccionar la celda D2

Paso 2. Ingresar =COUNTIF(\$A\$2:\$A\$51,C2)

Paso 3. Copiar la celda D2 a las celdas D3:D6

FIGURA 2.10 DISTRIBUCIÓN DE FRECUENCIA DE LAS VENTAS DE REFRESCOS CONSTRUIDA EMPLEANDO LA FUNCIÓN COUNTIF DE EXCEL

	A	B	C	D	E
1	Ventas de refrescos		Refresco	Frecuencia	
2	Coke Classic		Coke Classic	=COUNTIF(\$A\$2:\$A\$51,C2)	
3	Diet Coke		Diet Coke	=COUNTIF(\$A\$2:\$A\$51,C3)	
4	Pepsi		Dr. Pepper	=COUNTIF(\$A\$2:\$A\$51,C4)	
5	Diet Coke		Pepsi	=COUNTIF(\$A\$2:\$A\$51,C5)	
6	Coke Classic		Sprite	=COUNTIF(\$A\$2:\$A\$51,C6)	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

	A	B	C	D	E
1	Ventas de refrescos		Refresco	Frecuencia	
2	Coke Classic		Coke Classic	19	
3	Diet Coke		Diet Coke	8	
4	Pepsi		Dr. Pepper	5	
5	Diet Coke		Pepsi	13	
6	Coke Classic		Sprite	5	
7	Coke Classic				
8	Dr. Pepper				
9	Diet Coke				
10	Pepsi				
45	Pepsi				
46	Pepsi				
47	Pepsi				
48	Coke Classic				
49	Dr. Pepper				
50	Pepsi				
51	Sprite				
52					

Nota: Los renglones 11–44 están ocultos.



En la hoja de cálculo con las fórmulas de la figura 2.10 se observan en las celdas las fórmulas ingresadas al seguir estos pasos. En la hoja de cálculo con los valores se observan los valores obtenidos con las fórmulas de cada celda. En esta hoja de cálculo se aprecia la misma distribución de frecuencia de la tabla 2.2



Gráfica de barras Aquí se muestra cómo usar el asistente para gráficos de Excel para elaborar una gráfica de barras con los datos de las ventas de refrescos. En la figura 2.10 obsérvese la distribución de frecuencia que se presenta en la hoja de cálculo con los valores. La gráfica de barras que se va a construir es una extensión de esta hoja de cálculo. En la figura 2.11 se muestra esta misma hoja de cálculo con la gráfica de barras elaborada usando el asistente para gráficos. Los pasos a seguir son:

Paso 1. Seleccionar las celdas C1:D6

Paso 2. Hacer clic en el botón **Asistente para gráficos** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

Paso 3. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 1 de 4: tipo de gráfico:

Elegir **Columnas** de la lista **Tipo de gráfico**

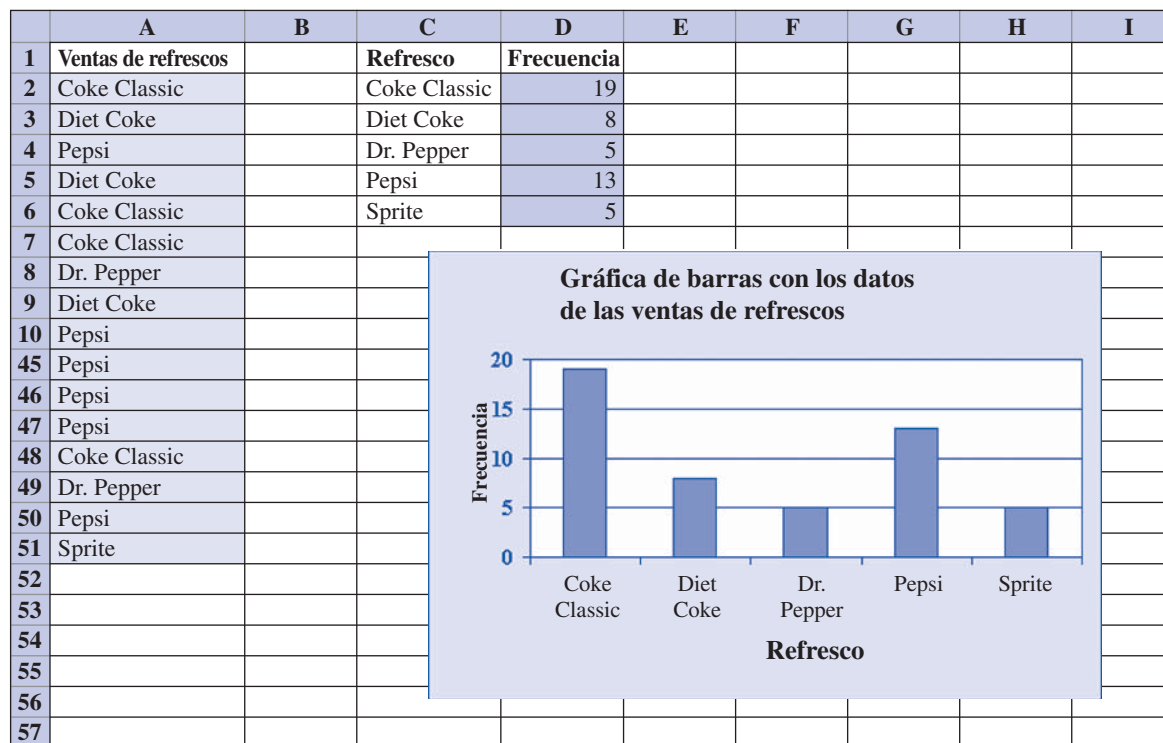
Elegir **Columnas agrupadas** en la visualización **Subtipo de gráfico**

Hacer clic en **Siguiente >**

Paso 4. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 2 de 4: datos de origen:

Hacer clic en **Siguiente >**

FIGURA 2.11 GRÁFICA DE BARRAS CON LOS DATOS DE LAS VENTAS DE REFRESCOS ELABORADA MEDIANTE EL ASISTENTE PARA GRÁFICOS DE EXCEL



Paso 5. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 3 de 4: opciones de gráfico:

- Seleccionar la pestaña **Títulos** y después
 - Digitar Gráfica de barras con los datos de las ventas de refrescos en el cuadro **Título del gráfico**
 - Digitar Refresco en el cuadro **Eje de categorías (X)**
 - Digitar Frecuencia en el cuadro **Eje de valores (Y)**
- Seleccionar la pestaña **Leyenda** y después
 - Quitar la paloma (marca de verificación) que aparece en el cuadro **Mostrar leyenda**
 - Hacer clic en **Siguiente >**

Paso 6. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 4 de 4: ubicación:

- Especificar una ubicación para la nueva gráfica (aquí se utilizó la misma hoja de cálculo que se estaba empleando por lo que se eligió la opción **Como objeto en**)
- Hacer clic en **Finalizar**

En la figura 2.11 se muestra la gráfica de barras que se obtuvo.*

De manera similar, Excel puede elaborar una gráfica de pastel con los datos de las ventas de refrescos. La diferencia principal es que en el paso 3 se elige **Circular** de la lista Tipo de gráfico.

Distribuciones de frecuencia e histogramas para datos cuantitativos

En esta sección se muestra cómo usar Excel para elaborar una distribución de frecuencia y un histograma con datos cuantitativos. Para ilustrar esto se usan los datos de la tabla 2.4 sobre la duración de las auditorías.

Distribución de frecuencia Para elaborar una distribución de frecuencia con datos cuantitativos se puede usar la función FREQUENCY de Excel. Consulte la figura 2.12 a medida que se presentan los pasos a seguir. La hoja de cálculo con las fórmulas aparece en segundo plano y la hoja de cálculo con los valores aparece en primer plano. El título “Duración de la auditoría” se encuentra en la celda A1 y los datos de las 20 auditorías están en las celdas A2:A21. Siguiendo los procedimientos indicados en el texto, introduzca las cinco clases 10–14, 15–19, 20–24, 25–29 y 30–34. El título “Duración de la auditoría” y las cinco clases se ingresan en las celdas C1:C6. El título “Límite superior” y los cinco límites superiores de las clases se ingresan en las celdas D1:D6. Ingrese también el título “Frecuencia” en la celda E1. La función FREQUENCY de Excel se usará para obtener la frecuencia en las celdas E2:E6. Los pasos siguientes describen cómo elaborar una distribución de frecuencia con los datos de las duraciones de las auditorías.

Paso 1. Seleccionar las celdas E2:E6

Paso 2. Digitar, pero no ingresar, la fórmula siguiente:

=FREQUENCY(A2:A21,D2:D6)

Paso 3. Pulsar las teclas CTRL+SHIFT(mayúsculas)+ENTER con lo que la fórmula matricial será ingresada en cada una de las celdas E2:E6

El resultado se muestra en la figura 2.12. Los valores que aparecen en las celdas E2:E6 son las frecuencias de las clases correspondientes. Regrese a la función FREQUENCY, vea que el intervalo de las celdas para los límites superiores de clase (D2:D6) sirve de argumento a la función. Estos límites superiores de clase a los que Excel llama *bins*, le dicen a Excel qué frecuencia poner en las celdas del intervalo de salida (E2:E6). Por ejemplo, la frecuencia de la clase que tiene el límite superior, o *bin*, 14 será colocada en la primera celda (E2), la frecuencia de la clase que tiene el límite superior, o *bin*, 19 será colocada en la segunda celda (E3), y así sucesivamente.



Para ingresar una fórmula matricial es necesario mantener oprimidas las teclas Ctrl y Shift(mayúsculas) mientras se pulsa la tecla Enter.

*La gráfica de barras de la figura 2.11 no es del mismo tamaño que la obtenida con Excel después de seleccionar **Finalizar**. Modificar el tamaño de una gráfica de Excel no es difícil. Primero se selecciona la gráfica, en los bordes de la gráfica aparecerán unos cuadritos negros llamados manillas de tamaño. Hacer clic sobre las manillas de tamaño y arrastrarlas para darle a la figura el tamaño deseado.

FIGURA 2.12 DISTRIBUCIÓN DE FRECUENCIA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS CON LA FUNCIÓN FREQUENCY DE EXCEL

	A	B	C	D	E
1	D. auditoría		D. auditoría	Límite superior	Frecuencia
2	12		10-14	14	=FREQUENCY(A2:A21,D2:D6)
3	15		15-19	19	=FREQUENCY(A2:A21,D2:D6)
4	20		20-24	24	=FREQUENCY(A2:A21,D2:D6)
5	22		25-29	29	=FREQUENCY(A2:A21,D2:D6)
6	14		30-34	34	=FREQUENCY(A2:A21,D2:D6)
7	14				
8	15				
9	27				
10	21				
11	18				
12	19				
13	18				
14	22				
15	33				
16	16				
17	18				
18	17				
19	23				
20	28				
21	13				

	A	B	C	D	E
1	D. auditoría		D. auditoría	Límite superior	Frecuencia
2	12		10-14	14	4
3	15		15-19	19	8
4	20		20-24	24	5
5	22		25-29	29	2
6	14		30-34	34	1
7	14				
8	15				
9	27				
10	21				
11	18				
12	19				
13	18				
14	22				
15	33				
16	16				
17	18				
18	17				
19	23				
20	28				
21	13				

Histograma Para usar el ayudante para gráficos de Excel para construir un histograma con las duraciones de las auditorías parta de la distribución de frecuencia de la figura 2.12. En la figura 2.13 se presenta la hoja de trabajo con la distribución de frecuencia y el histograma. Los pasos siguientes indican cómo emplear el asistente para gráficos al elaborar un histograma con los datos de las duraciones de las auditorías.

Paso 1. Seleccionar las celdas E2:E6

Paso 2. Hacer clic en el botón **Asistente para gráficos** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

Paso 3. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 1 de 4: tipo de gráfico

Elegir **Columnas** en la lista **Tipo de gráfico**

Elegir **Columnas agrupadas** en la visualización **Subtipo de gráfico**

Hacer clic en **Siguiente >**

Paso 4. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 2 de 4: datos de origen:

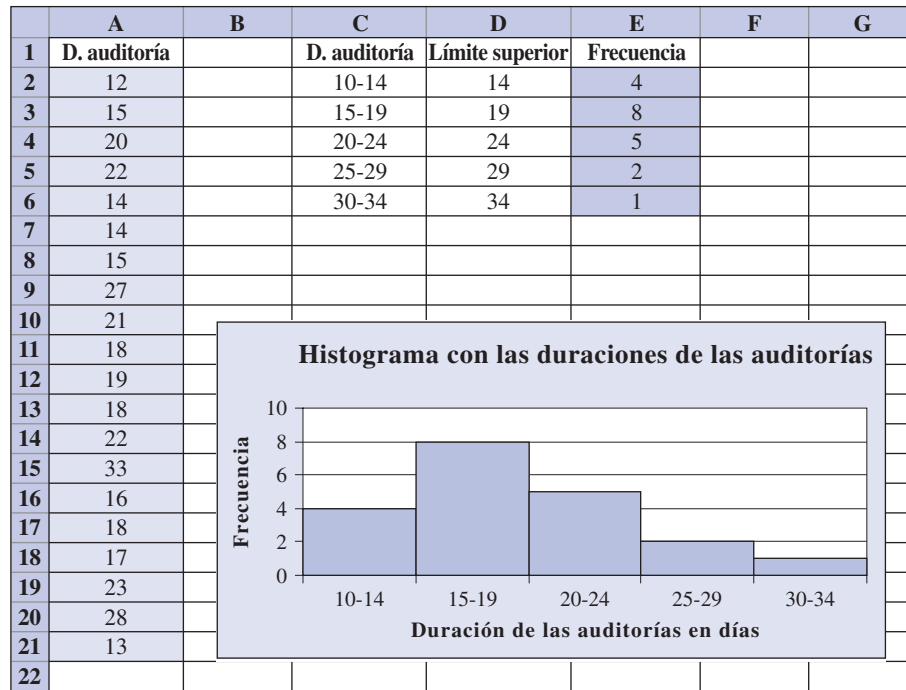
Seleccionar la pestaña **Serie** y después

Hacer clic en el cuadro **Rótulos del eje de categorías (X)**

Seleccionar las celdas C2:C6

Hacer clic en **Siguiente >**

FIGURA 2.13 HISTOGRAMA CON LAS DURACIONES DE LAS AUDITORÍAS



Paso 5. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 3 de 4: opciones de gráfico:

Seleccionar la pestaña **Títulos** y después

Digitar Histograma de las duraciones de las auditorías en el cuadro

Título del gráfico

Digitar Duración de las auditorías en días en el cuadro **Eje de categorías (X):**

Digitar Frecuencia en el cuadro **Eje de valores (Y):**

Seleccionar la pestaña **Leyenda** y después

Quitar la paloma (marca de verificación) que aparece en el cuadro

Mostrar leyenda

Hacer clic en **Siguiente >**

Paso 6. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 4 de 4: ubicación:

Especificar una ubicación para la nueva gráfica (aquí se utilizó la misma hoja de cálculo que se estaba empleando por lo que se eligió la opción

Como objeto en)

Hacer clic en **Finalizar**

Ahora en la hoja de cálculo aparecerá una gráfica de columnas elaborada por Excel. Pero entre las columnas habrá espacios. Como en un histograma no hay espacios entre las columnas, es necesario modificar esta gráfica para eliminar los espacios entre las columnas. Los pasos siguientes describen cómo hacerlo.

Paso 1. Dar doble clic en cualquiera de las columnas de la gráfica.

Paso 2. Cuando aparezca el cuadro de diálogo Formato de punto de datos:

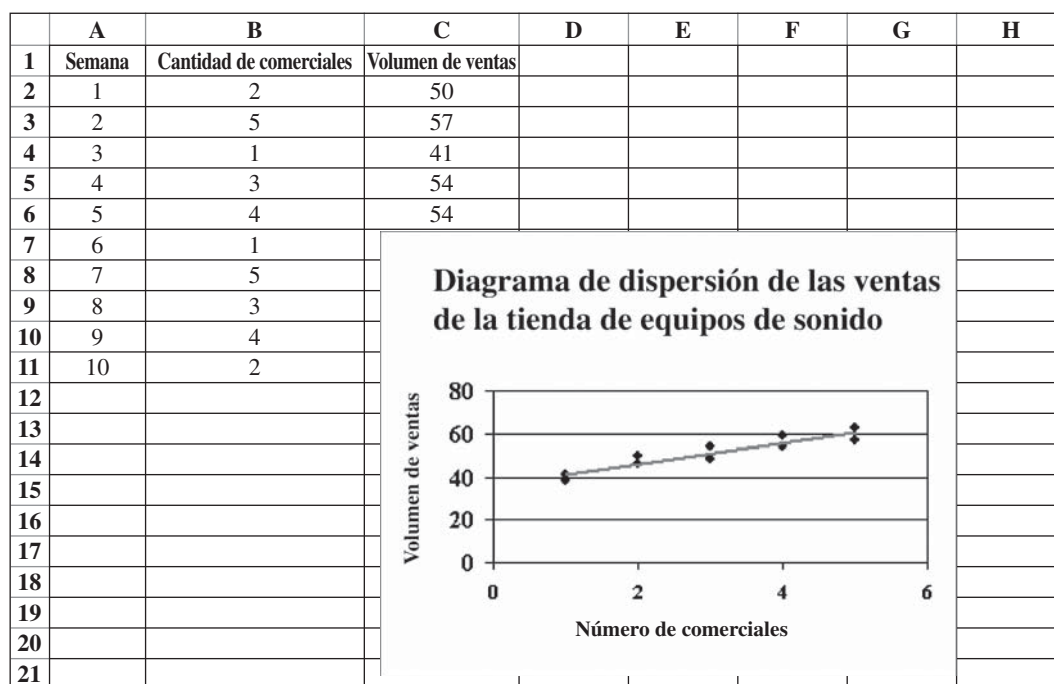
Seleccionar la pestaña **Opciones**

Ingresar 0 en el cuadro **Ancho del rango**

Hacer clic en **Aceptar**

El histograma se verá como el que aparece en la figura 2.13.

Por último, un aspecto interesante de la hoja de cálculo de la figura 2.13 es que Excel ha relacionado los datos que aparecen en las celdas A2:A21 con las frecuencias que aparecen en las celdas E2:E6 y con el histograma. Si se modifica alguno de los datos de las celdas A2:A21 se

FIGURA 2.14 DIAGRAMA DE DISPERSIÓN DE LAS VENTAS DE LA TIENDA DE EQUIPOS DE SONIDO

modificarán automáticamente las frecuencias de las celdas E2:E6 y también el histograma y aparecerán una distribución de frecuencias y un histograma modificados. Se aconseja probar cómo se realiza esta modificación automática modificando uno o dos de los datos.

Diagrama de dispersión

Se usarán los datos de la tienda de equipo de sonido que aparecen en la tabla 2.12 para mostrar cómo se usa el asistente para gráficos de Excel al elaborar un diagrama de dispersión. Consulte la figura 2.14 a medida que se describen los pasos para elaborar esta gráfica. La hoja de cálculo con los valores aparece en segundo plano y el diagrama de dispersión elaborado por el asistente para gráficos en primer plano. Los pasos a seguir son los siguientes.

Paso 1. Seleccionar la celda B1:C11

Paso 2. Hacer clic en el botón **Asistente para gráficos** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

Paso 3. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 1 de 4: tipo de gráfico:

Elegir **XY (Dispersión)** en la lista **Tipo de gráfico**

Elegir **Dispersión** en la visualización **Subtipo de gráfico**

Hacer clic en **Siguiente >**

Paso 4. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 2 de 4: datos de origen:

Hacer clic en **Siguiente >**

Paso 5. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 3 de 4: opciones de gráfico:

Seleccionar la pestaña **Títulos**

Digitar Diagrama de dispersión de las ventas de la tienda de equipos de sonido en el cuadro **Título del gráfico**

Digitar Número de comerciales en el cuadro **Eje de categorías (X)**:

Digitar Volumen de ventas en el cuadro **Eje de valores (Y)**:

Seleccionar la pestaña **Leyenda**

Quitar la paloma (marca de verificación) que aparece en el cuadro

Mostrar leyenda

Hacer clic en **Siguiente >**

Paso 6. Cuando aparezca el cuadro de diálogo Asistente para gráficos – paso 4 de 4: ubicación:

Especificar una ubicación para la nueva gráfica (aquí se utilizó la misma hoja de cálculo que se estaba empleando por lo que se eligió la opción

Como objeto en)

Hacer clic en **Finalizar**

En el diagrama de dispersión puede trazar una línea de tendencia de la manera siguiente.

Paso 1. Colocar el cursor del mouse sobre cualquiera de los puntos del diagrama de dispersión y dar clic con el botón derecho del mouse. Aparecerá una lista de opciones

Paso 2. Elegir **Agregar línea de tendencia**

Paso 3. Cuando aparezca el cuadro agregar línea de tendencia:

Seleccionar la pestaña **Tipo**

Elegir **Lineal** en la visualización **Tipo de tendencia o regresión**

Hacer clic en **Aceptar**

En la hoja de cálculo de la figura 2.14 se observa el diagrama de dispersión con la línea de tendencia.

Informe en tabla dinámica

El informe en tabla dinámica de Excel proporciona una valiosa herramienta para la manipulación de un conjunto de datos en que se tiene más de una variable. Se ilustrará su uso mostrando cómo elaborar una tabulación cruzada.

Tabulación cruzada Se ilustra la elaboración de una tabulación cruzada empleando los datos de los restaurantes que aparecen en la figura 2.15. Los títulos se han ingresado en el renglón 1 y los datos de los 300 restaurantes se han ingresado en las celdas A2:C301

FIGURA 2.15 HOJA DE CÁLCULO DE EXCEL CON LOS DATOS DE LOS RESTAURANTES



Nota: los renglones 12–291 están ocultos.

	A	B	C	D
1	Restaurante	Calidad	Precio (\$)	
2	1	Bueno	18	
3	2	Muy bueno	22	
4	3	Bueno	28	
5	4	Excelente	38	
6	5	Muy bueno	33	
7	6	Bueno	28	
8	7	Muy bueno	19	
9	8	Muy bueno	11	
10	9	Muy bueno	23	
11	10	Bueno	13	
292	291	Muy bueno	23	
293	292	Muy bueno	24	
294	293	Excelente	45	
295	294	Bueno	14	
296	295	Bueno	18	
297	296	Bueno	17	
298	297	Bueno	16	
299	298	Bueno	15	
300	299	Muy bueno	38	
301	300	Muy bueno	31	
302				

- Paso 1.** Seleccionar el menú **Datos**
- Paso 2.** Elegir **Informe de tabla y datos dinámicos**
- Paso 3.** Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 1 de 3:
Elegir **Lista o base de datos de Microsoft Excel**
Elegir **Tabla dinámica**
Hacer clic en **Siguiente**
- Paso 4.** Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 2 de 3:
Ingresar A1:C301 en el cuadro **Rango**
Hacer clic en **Siguiente**
- Paso 5.** Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 3 de 3:
Seleccionar **Hoja de cálculo nueva**
Seleccionar **Diseño**
- Paso 6.** Cuando aparezca el diagrama Asistente para tablas y gráficos dinámicos – diseño (véase figura 2.16):
Arrastre el botón de campo **Calidad (Quality)** a la sección **FILA (ROW)** del diagrama
Arrastre el botón de campo **Precio (Meal Price)** a la sección **COLUMNA (COLUMN)** del diagrama
Arrastre el botón de campo **Restaurante (Restaurant)** a la sección **DATOS (DATA)** del diagrama
Dar doble clic en el botón de campo **Suma de Restaurante** en la sección DATOS
Cuando aparezca el cuadro de diálogo Campo de la tabla dinámica:
Elegir **Cuenta** bajo **Resumir por**
Hacer clic en **Aceptar** (la figura 2.17 muestra el diseño completo del diagrama)
Hacer clic en **Aceptar**
- Paso 7.** Cuando aparezca el cuadro de diálogo Asistente para tablas y gráficos dinámicos – paso 3 de 3:
Hacer clic en **Finalizar**

En la figura 2.18 se muestra parte del resultado generado por Excel. Observe que las columnas D a AK se han ocultado para que se puedan mostrar los resultados en una figura de tamaño razo-

FIGURA 2.16 ASISTENTE PARA TABLAS Y GRÁFICOS DINÁMICOS: DISEÑO

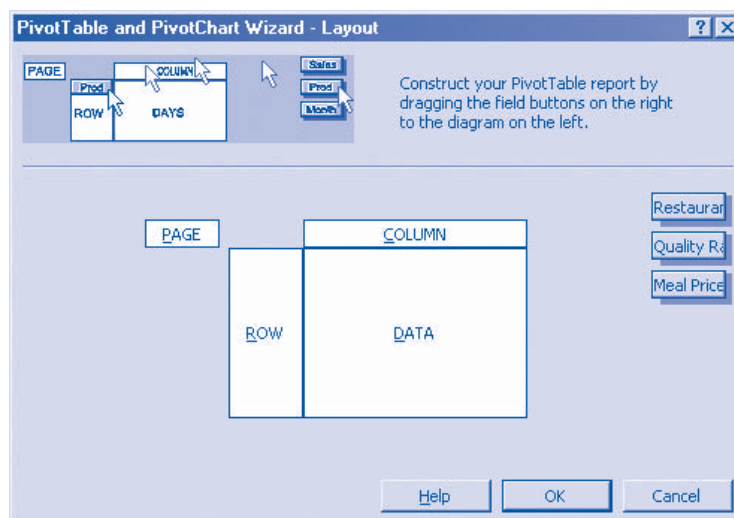


FIGURA 2.17 ASISTENTE PARA TABLAS Y GRÁFICOS DINÁMICOS: DISEÑO

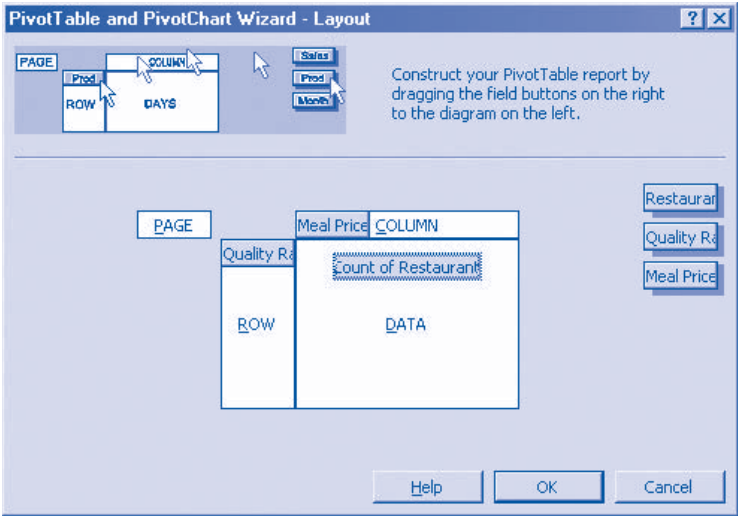


FIGURA 2.18 RESULTADO INICIAL DEL INFORME DE TABLA DINÁMICA (LAS COLUMNAS D:AK ESTÁN OCULTAS)

	A	B	C	AL	AM	AN	AO
1							
2							
3	Suma de restaurantes	Precio (\$) ▼					
4	Calidad ▼	10	11	47	48	Gran total	
5	Excelente			2	2	66	
6	Bueno	6	4			84	
7	Muy bueno	1	4		1	150	
8	Gran total	7	8	2	3	300	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

FIGURA 2.19 INFORME DE TABLA DINÁMICA FINAL CON LOS DATOS DE LOS RESTAURANTES

	A	B	C	D	E	F	G
1							
2							
3	Suma de restaurantes	Precio (\$) ▼					
4	Calidad ▼	10-19	20-29	30-39	40-49	Gran total	
5	Bueno	42	40	2		84	
6	Muy bueno	34	64	46	6	150	
7	Excelente	2	14	28	22	66	
8	Gran total	78	118	76	28	300	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							

nable. Los títulos de los renglones (Excelente, Bueno y Muy bueno) y los totales de los renglones (66, 84, 150 y 300) de la figura 2.18 son los mismos que en la tabla 2.10, sólo que en distinto orden. Para colocarlos en el orden Bueno, Muy bueno, Excelente hay que seguir los siguientes pasos:

Paso 1. Hacer clic con el botón derecho sobre la celda A5

Paso 2. Elegir **Ordenar**

Paso 3. Elegir **Mover al final**

En la figura 2.18 hay una columna para cada precio. Por ejemplo, en la columna B se encuentran los restaurantes cuyo precio es \$10, en la columna C los restaurantes cuyo precio es \$11, etc. Para que el informe en tabla dinámica se vea como en la tabla 2.10, se deben agrupar las columnas en cuatro categorías de precios: \$10–19, \$20–29, \$30–39 y \$40–49. Los pasos necesarios para agrupar las columnas de la hoja de cálculo que aparece en la figura 2.18 son:

Paso 1. Hacer clic con el botón derecho en Precio(\$) en la celda B3 de la Tabla dinámica

Paso 2. Elegir **Agrupar y mostrar detalles**

Elegir **Agrupar**

Paso 3. Cuando aparezca el cuadro de diálogo **Agrupar**

Ingresar 10 en el cuadro **Comenzar en**

Ingresar 49 en el cuadro **Terminar en**

Ingresar 10 en el cuadro **Por**

Hacer clic en **Aceptar**

La tabla dinámica que se obtiene se presenta en la figura 2.19. Es la tabla dinámica final. Observe que esta tabla proporciona la misma información que la tabla cruzada de la tabla 2.10.

CAPÍTULO 3



Estadística descriptiva: medidas numéricas

CONTENIDO

LA ESTADÍSTICA EN LA PRÁCTICA:

SMALL FRY DESIGN

3.1 MEDIDAS DE LOCALIZACIÓN

Media
Mediana
Moda
Percentiles
Cuartiles

3.2 MEDIDAS DE VARIABILIDAD

Rango
Rango intercuartílico
Varianza
Desviación estándar
Coeficiente de variación

3.3 MEDIDAS DE LA FORMA DE LA DISTRIBUCIÓN, DE LA POSICIÓN RELATIVA Y LA DETECCIÓN DE OBSERVACIONES ATÍPICAS

Forma de la distribución
Puntos z
Teorema de Chebyshev

Regla empírica

Detección de observaciones atípicas

3.4 ANÁLISIS EXPLORATORIO DE DATOS

Resumen de cinco números
Diagrama de caja

3.5 MEDIDAS DE ASOCIACIÓN ENTRE DOS VARIABLES

Covarianza
Interpretación de la covarianza
Coeficiente de correlación
Interpretación del coeficiente de correlación

3.6 LA MEDIA PONDERADA Y EL EMPLEO DE DATOS AGRUPADOS

Media ponderada
Datos agrupados

LA ESTADÍSTICA *en* LA PRÁCTICA

SMALL FRY DESIGN*

SANTA ANA, CALIFORNIA

Fundada en 1997, Small Fry Design es una empresa de juguetes y accesorios que diseña e importa productos para niños pequeños. La línea de productos de la empresa incluye muñecos de peluche, móviles, juguetes musicales, sonajeros y mantas de seguridad y ofrece diseños de juguetes de alta calidad para bebés, con énfasis especial en los colores, texturas y sonidos. Los productos son diseñados en Estados Unidos y manufacturados en China.

Small Fry Design emplea representantes independientes para la venta de sus productos a tiendas de mobiliario para niños, tiendas de accesorios y ropa para niños, tiendas de regalos, tiendas exclusivas de departamentos e importantes empresas de ventas por catálogo. En la actualidad los productos de Small Fry Design se distribuyen en más de 1000 negocios en todo Estados Unidos.

La administración del flujo de efectivo es una de las actividades más relevantes del funcionamiento cotidiano de esta empresa. Garantizar suficiente ingreso de efectivo para cumplir con la deuda corriente y la deuda a corto plazo es la diferencia entre el éxito y el fracaso de la empresa. Un factor importante de la administración del flujo de efectivo es el análisis y control de las cuentas por cobrar. Al medir el tiempo promedio y el valor en dólares que tienen las facturas pendientes, los administradores pronostican la disponibilidad de dinero y vigilan la situación de las cuentas por cobrar. La empresa se ha planteado los objetivos siguientes: el tiempo promedio de una factura pendiente no debe ser más de 45 días y el valor en dólares de las facturas que tengan más de 60 días no debe ser superior a 5% del valor en dólares de todas las cuentas por cobrar.

En un resumen reciente sobre el estado de las cuentas por cobrar se presentaron los siguientes estadísticos descriptivos sobre el tiempo que tenían las facturas pendientes.

Media	40 días
Mediana	35 días
Moda	31 días

*Los autores agradecen a John A. McCarthy, presidente de Small Fry Design por proporcionar este artículo para *La estadística en la práctica*.



Móvil “El rey de la selva” de Small Fry Design.
© Foto cortesía de Small Fry Design, Inc.

La interpretación de dichos estadísticos indica que el tiempo promedio de una factura pendiente es 40 días. La mediana revela que la mitad de las facturas se quedan pendientes 35 días o más. La moda, 31 días, muestra que el tiempo que con más frecuencia permanece pendiente una factura es 31 días. Este resumen estadístico indica también que sólo 3% del valor en dólares de todas las cuentas por cobrar tienen más de 60 días. De acuerdo con esta información estadística, la administración está satisfecha de que las cuentas por cobrar y el flujo de efectivo entrante estén bajo control.

En este capítulo aprenderá a calcular e interpretar algunas de las medidas estadísticas empleadas por Small Fry Design. Además de la media, la mediana y la moda usted estudiará otros estadísticos descriptivos como el rango, la varianza, la desviación estándar, los percentiles y la correlación. Estas medidas numéricas ayudan a la comprensión e interpretación de datos.

En el capítulo 2 estudió las presentaciones tabular y gráfica para resumir datos. En este capítulo se le presentan varias medidas numéricas que proporcionan otras opciones para resumir datos.

Empezará con medidas numéricas para conjuntos de datos que constan de una sola variable. Si el conjunto de datos consta de más de una variable, empleará estas mismas medidas numéricas para cada una de las variables por separado. Sin embargo, en el caso de dos variables, estudiará también medidas de la relación entre dos variables.

Se presentan medidas numéricas de localización, dispersión, forma, y asociación. Si estas medidas las calcula con los datos de una muestra, se llaman **estadísticos muestrales**. Si estas medidas las calcula con los datos de una población se llaman **parámetros poblacionales**. En inferencia estadística, al estadístico muestral se le conoce como el **estimador puntual** del correspondiente parámetro poblacional. El proceso de estimación puntual será estudiado con más detalle en el capítulo 7.

En los dos apéndices del capítulo se le muestra cómo usar Minitab y Excel para calcular muchas de las medidas descritas en este capítulo.

3.1

Medidas de localización

Media

La medida de localización más importante es la **media**, o valor promedio, de una variable. La media proporciona una medida de localización central de los datos. Si los datos son datos de una muestra, la media se denota \bar{x} ; si los datos son datos de una población, la media se denota con la letra griega μ .

En las fórmulas estadísticas se acostumbra denotar el valor de la primera observación de la variable x con x_1 , el valor de la segunda observación de la variable x con x_2 y así con lo siguiente. En general, el valor de la i -ésima observación de la variable x se denota x_i . La fórmula para la media muestral cuando se tiene una muestra de n observaciones es la siguiente.

La media muestral \bar{x} es un estadístico muestral.

MEDIA MUESTRAL

$$\bar{x} = \frac{\sum x_i}{n} \quad (3.1)$$

En la fórmula anterior el numerador es la suma de los valores de las n observaciones. Es decir,

$$\sum x_i = x_1 + x_2 + \cdots + x_n$$

La letra griega Σ es el símbolo de sumatoria (suma)

Para ilustrar el cálculo de la media muestral, considere los siguientes datos que representan el tamaño de cinco grupos de una universidad.

$$46 \quad 54 \quad 42 \quad 46 \quad 32$$

Se emplea la notación x_1, x_2, x_3, x_4, x_5 para representar el número de estudiantes en cada uno de los cinco grupos.

$$x_1 = 46 \quad x_2 = 54 \quad x_3 = 42 \quad x_4 = 46 \quad x_5 = 32$$

Por tanto, para calcular la media muestral, escriba

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{46 + 54 + 42 + 46 + 32}{5} = 44$$

La media muestral del tamaño de estos grupos es 44 alumnos.

Otra ilustración del cálculo de la media muestral aparece en la situación siguiente. Suponga que la bolsa de trabajo de una universidad envía cuestionarios a los recién egresados de la carrera de administración solicitándoles información sobre sus sueldos mensuales iniciales. En la ta-

TABLA 3.1 SUELDOS MENSUALES INICIALES EN UNA MUESTRA DE 12 RECIÉN EGRESADOS DE LA CARRERA DE ADMINISTRACIÓN

Egresado	Sueldo mensual inicial (\$)	Egresado	Sueldo mensual inicial (\$)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

En la tabla 3.1 se presentan estos datos. El sueldo mensual inicial medio de los 12 recién egresados se calcula como sigue.

$$\begin{aligned}
 \bar{x} &= \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12} \\
 &= \frac{3450 + 3550 + \cdots + 3480}{12} \\
 &= \frac{42\,480}{12} = 3540
 \end{aligned}$$

En la ecuación (3.1) se muestra cómo se calcula la media en una muestra de n observaciones. Para calcular la media de una población use la misma fórmula, pero con una notación diferente para indicar que trabaja con toda la población. El número de observaciones en una población se denota N y el símbolo para la media poblacional es μ .

La media muestral \bar{x} es un estimador puntual de la media poblacional μ .

MEDIA POBLACIONAL

$$\mu = \frac{\sum x_i}{N} \quad (3.2)$$

Mediana

La **mediana** es otra medida de localización central. Es el valor de enmedio en los datos ordenados de menor a mayor (en forma ascendente). Cuando tiene un número impar de observaciones, la mediana es el valor de enmedio. Cuando la cantidad de observaciones es par, no hay un número enmedio. En este caso, se sigue una convención y la mediana es definida como el promedio de las dos observaciones de enmedio. Por conveniencia, la definición de mediana se plantea así:

MEDIANA

Ordenar los datos de menor a mayor (en forma ascendente).

- Si el número de observaciones es impar, la mediana es el valor de enmedio.
- Si el número de observaciones es par, la mediana es el promedio de las dos observaciones de enmedio.

Apliquemos esta definición para calcular la mediana del número de alumnos en un grupo a partir de la muestra de los cinco grupos de universidad. Los datos en orden ascendente son

32 42 46 46 54

Como $n = 5$ es impar, la mediana es el valor de enmedio. De manera que la mediana del tamaño de los grupos es 46. Aun cuando en este conjunto de datos hay dos observaciones cuyo valor es 46, al poner las observaciones en orden ascendente se toman en consideración todas las observaciones.

Suponga que también desea calcular la mediana del salario inicial de los 12 recién egresados de la carrera de administración de la tabla 3.1. Primero ordena los datos de menor a mayor

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925
 Los dos valores
 de en medio

Como $n = 12$ es par, se localizan los dos valores de enmedio: 3490 y 3520. La mediana es el promedio de estos dos valores.

$$\text{Mediana} = \frac{3490 + 3520}{2} = 3505$$

La mediana es la medida de localización más empleada cuando se trata de ingresos anuales y valores de propiedades, debido a que la media puede inflarse por unos cuantos ingresos o valores de propiedades muy altos. En tales casos, la mediana es la medida de localización central preferida.

Aunque la media es la medida de localización central más empleada, en algunas situaciones se prefiere la mediana. A la media la influyen datos en extremo pequeños o considerablemente grandes. Por ejemplo, suponga que uno de los recién graduados de la tabla 3.1 tuviera un salario inicial de \$10 000 mensuales (quizá su familia sea la dueña de la empresa). Si reemplaza el mayor sueldo inicial mensual de la tabla 3.1, \$3925, por \$10 000 y vuelve a calcular la media, la media muestral cambia de \$3540 a \$4046. Sin embargo, la mediana, \$3505, permanece igual ya que \$3490 y \$3520 siguen siendo los dos valores de en medio. Si hay algunos sueldos demasiado altos, la mediana proporciona una medida de tendencia central mejor que la media. Al generalizar lo anterior, es posible decir que cuando los datos contengan valores extremos, es preferible usar a la mediana como medida de localización central.

Moda

La tercera medida de localización es la **moda**. La moda se define como sigue.

MODA

La moda es el valor que se presenta con mayor frecuencia.

Para ilustrar cómo identificar a la moda, considere la muestra del tamaño de los cinco grupos de la universidad. El único valor que se presenta más de una vez es el 46. La frecuencia con que se presenta este valor es 2, por lo que es el valor con mayor frecuencia, entonces es la moda. Para ver otro ejemplo, considere la muestra de los sueldos iniciales de los recién egresados de la carrera de administración. El único salario mensual inicial que se presenta más de una vez es \$3480. Como este valor tiene la frecuencia mayor, es la moda.

Hay situaciones en que la frecuencia mayor se presenta con dos o más valores distintos. Cuando esto ocurre hay más de una moda. Si los datos contienen más de una moda se dice que los datos son *bimodales*. Si contienen más de dos modas, son *multimodales*. En los casos multimodales casi nunca se da la moda, porque dar tres o más modas no resulta de mucha ayuda para describir la localización de los datos.

Percentiles

Un **percentil** aporta información acerca de la dispersión de los datos en el intervalo que va del menor al mayor valor de los datos. En los conjuntos de datos que no tienen muchos valores repetidos, el percentil p divide a los datos en dos partes. Cerca de p por ciento de las observaciones tienen valores menores que el percentil p y aproximadamente $(100 - p)$ por ciento de las observaciones tienen valores mayores que el percentil p . El percentil p se define como sigue:

PERCENTIL

El percentil p es un valor tal que por lo menos p por ciento de las observaciones son menores o iguales que este valor y por lo menos $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor.

Las puntuaciones en los exámenes de admisión de escuelas y universidades se suelen dar en términos de percentiles. Por ejemplo, suponga que un estudiante obtiene 54 puntos en la parte verbal del examen de admisión. Esto no dice mucho acerca de este estudiante en relación con los demás estudiantes que realizaron el examen. Sin embargo, si esta puntuación corresponde al percentil 70, entonces 70% de los estudiantes obtuvieron una puntuación menor a la de dicho estudiante y 30% de los estudiantes obtuvieron una puntuación mayor.

Para calcular el percentil p se emplea el procedimiento siguiente.

CÁLCULO DEL PERCENTIL p

Paso 1. Ordenar los datos de menor a mayor (colocar los datos en orden ascendente).

Paso 2. Calcular el índice i

$$i = \left(\frac{p}{100} \right) n$$

donde p es el percentil deseado y n es el número de observaciones.

Paso 3. (a) Si i no es un número entero, debe redondearlo. El primer entero mayor que i denota la posición del percentil p .

(b) Si i es un número entero, el percentil p es el promedio de los valores en las posiciones i e $i + 1$.

Seguir estos pasos facilita el cálculo de los percentiles.

Para ilustrar el empleo de este procedimiento, determine el percentil 85 en los sueldos mensuales iniciales de la tabla 3.1.

Paso 1. Ordenar los datos de menor a mayor

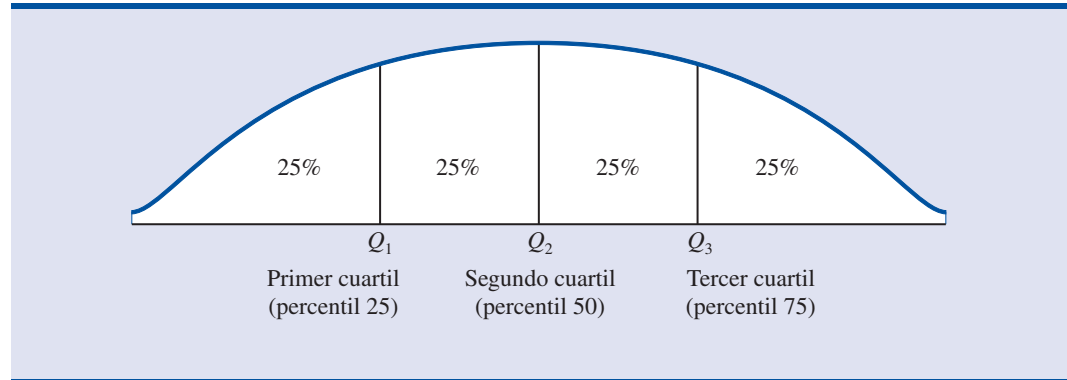
3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Paso 2.

$$i = \left(\frac{p}{100} \right) n = \left(\frac{85}{100} \right) 12 = 10.2$$

Paso 3. Como i no es un número entero, se debe *redondear*. La posición del percentil 85 es el primer entero mayor que 10.2, es la posición 11.

Observe ahora los datos, entonces el percentil 85 es el dato en la posición 11, o sea 3730.

FIGURA 3.1 LOCALIZACIÓN DE LOS CUARTILES

Para ampliar la formación en el uso de este procedimiento, calculará el percentil 50 en los sueldos mensuales iniciales. Al aplicar el paso 2 obtiene.

$$i = \left(\frac{50}{100} \right) 12 = 6$$

Como i es un número entero, de acuerdo con el paso 3 b) el percentil 50 es el promedio de los valores de los datos que se encuentran en las posiciones seis y siete; de manera que el percentil 50 es $(3490 + 3520)/2 = 3505$. Observe que el *percentil 50 coincide con la mediana*.

Cuartiles

Los cuartiles sólo son percentiles determinados; así que los pasos para calcular los percentiles también se emplean para calcular los cuartiles.

Con frecuencia es conveniente dividir los datos en cuatro partes; así, cada parte contiene una cuarta parte o 25% de las observaciones. En la figura 3.1 se muestra una distribución de datos dividida en cuatro partes. A los puntos de división se les conoce como **cuartiles** y están definidos como sigue:

Q_1 = primer cuartil, o percentil 25

Q_2 = segundo cuartil, o percentil 50

Q_3 = tercer cuartil, o percentil 75

Una vez más se ordenan los sueldos iniciales de menor a mayor. Q_2 , el segundo cuartil (la mediana), ya se tiene identificado, es 3505.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Para calcular los cuartiles Q_1 y Q_3 use la regla para hallar el percentil 25 y el percentil 75. A continuación se presentan estos cálculos.

Para hallar Q_1 ,

$$i = \left(\frac{p}{100} \right) n = \left(\frac{25}{100} \right) 12 = 3$$

Como i es un entero, el paso 3 b) indica que el primer cuartil, o el percentil 25, es el promedio del tercer y cuarto valores de los datos; esto es, $Q_1 = (3450 + 3480)/2 = 3465$.

Para hallar Q_3 ,

$$i = \left(\frac{p}{100} \right) n = \left(\frac{75}{100} \right) 12 = 9$$

Como i es un entero, el paso 3 b) indica que el tercer cuartil, o el percentil 75, es el promedio del noveno y décimo valores de los datos; esto es, $Q_3 = (3550 + 3650)/2 = 3600$.

Los cuartiles dividen los datos de los sueldos iniciales en cuatro partes y cada parte contiene 25% de las observaciones.

3310	3355	3450		3480	3480	3490		3520	3540	3550		3650	3730	3925
$Q_1 = 3465$			$Q_2 = 3505$					$Q_3 = 3600$						
(Mediana)														

Los cuartiles han sido definidos como el percentil 25, el percentil 50 y el percentil 75. Por lo que los cuartiles se calculan de la misma manera que los percentiles. Sin embargo, algunas veces se siguen otras convenciones para calcular los cuartiles, por ello los valores que se dan para los cuartiles varían ligeramente, dependiendo de la convención que se siga. De cualquier manera, el objetivo de calcular los cuartiles siempre es dividir los datos en cuatro partes iguales.

NOTAS Y COMENTARIOS

Cuando el conjunto de datos contiene valores extremos, es preferible usar la mediana que la media como unidad de localización central. Otra medida que suele ser usada cuando hay valores extremos es la *media recortada*. La media recortada se obtiene eliminando del conjunto de datos un determinado porcentaje de los valores menores y mayores y calculando después la media de los valores restantes. Por ejemplo, la media recortada a 5% se ob-

tiene eliminando el 5% menor y el 5% mayor de los valores y calculando después la media de los valores restantes. Con la muestra de los 12 sueldos iniciales, $0.05(12) = 0.6$. Redondear este valor a 1, indica que en la media recortada a 5% se elimina el valor (1) menor y el valor (1) mayor. La media recortada a 5% usando las 10 observaciones restantes es 3524.50.

Ejercicios

Método

- Los valores de los datos en una muestra son 10, 20, 12, 17 y 16. Calcule la media y la mediana.
- Los datos en una muestra son 10, 20, 21, 17, 16 y 25. Calcule la media y la mediana.
- Los valores en una muestra son 27, 25, 20, 15, 30, 34, 28 y 25. Calcule los percentiles 20, 25, 65 y 75.
- Una muestra tiene los valores 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 y 53. Calcule la media, la mediana y la moda.

Aplicaciones

- El Dow Jones Travel Index informa sobre lo que pagan por noche en un hotel en las principales ciudades de Estados Unidos los viajeros de negocios (*The Wall Street Journal*, 16 de enero de 2004). Los precios promedio por noche en 20 ciudades son los siguientes:

Atlanta	\$163	Minneapolis	\$125
Boston	177	New Orleans	167
Chicago	166	New York	245
Cleveland	126	Orlando	146
Dallas	123	Phoenix	139
Denver	120	Pittsburgh	134
Detroit	144	San Francisco	167
Houston	173	Seattle	162
Los Angeles	160	St. Louis	145
Miami	192	Washington, D.C.	207

Autoexamen

archivo
en
Hotels

CD

- ¿Cuál es la media en el precio de estas habitaciones?
 - ¿Cuál es la mediana en el precio de estas habitaciones?
 - ¿Cuál es la moda?
 - ¿Cuál es el primer cuartil?
 - ¿Cuál es el tercer cuartil?
6. Una asociación recaba información sobre sueldos anuales iniciales de los recién egresados de universidades de acuerdo con su especialidad. El salario anual inicial de los administradores de empresas es \$39 580 (*CNNMoney.com*, 15 de febrero de 2006). A continuación se presentan muestras de los sueldos anuales iniciales de especialistas en marketing y en contaduría (los datos están en miles):



Egresados de marketing

34.2 45.0 39.5 28.4 37.7 35.8 30.6 35.2 34.2 42.4

Egresados de contaduría

33.5 57.1 49.7 40.2 44.2 45.2 47.8 38.0
 53.9 41.1 41.7 40.8 55.5 43.5 49.1 49.9

- Para cada uno de los grupos de sueldos iniciales calcule moda, mediana y media.
 - Para cada uno de los grupos de sueldos iniciales calcule el primer y el tercer cuartil.
 - Los egresados de contaduría suelen tener mejores salarios iniciales. ¿Qué indican los datos muestrales acerca de la diferencia entre los sueldos anuales iniciales de egresados de marketing y de contaduría?
7. La Asociación Estadounidense de Inversionistas Individuales realiza una investigación anual sobre los corredores de bolsa (*AAII Journal*, enero de 2003). En la tabla 3.2 se muestran las comisiones que cobran los corredores de bolsa con descuento por dos tipos de transacciones: transacción con ayuda del corredor de 100 acciones a \$50 por acción y transacción en línea de 500 acciones a \$50 por acción.
- Calcule la media, mediana y moda de las comisiones que se cobran por una transacción con ayuda del corredor de 100 acciones a \$50 por acción.
 - Calcule la media, mediana y moda de las comisiones que se cobran por una transacción en línea de 500 acciones a \$50 por acción.
 - ¿Qué cuesta más, una transacción con ayuda del corredor de 100 acciones a \$50 por acción o una transacción en línea de 500 acciones a \$50 por acción?
 - ¿Está relacionado el costo de la transacción con el monto de la transacción?

TABLA 3.2 COMISIONES QUE COBRAN LOS CORREDORES DE BOLSA

Corredor	Con ayuda del corredor de 100 acciones \$50/acción	En línea 500 acciones a \$50/acción	Corredor	Con ayuda del corredor de 100 acciones \$50/acción	En línea 500 acciones a \$50/acción
Accutrade	30.00	29.95	Merrill Lynch Direct	50.00	29.95
Ameritrade	24.99	10.99	Muriel Siebert	45.00	14.95
Banc of America	54.00	24.95	NetVest	24.00	14.00
Brown & Co.	17.00	5.00	Recom Securities	35.00	12.95
Charles Schwab	55.00	29.95	Scottrade	17.00	7.00
CyberTrader	12.95	9.95	Sloan Securities	39.95	19.95
E*TRADE Securities	49.95	14.95	Strong Investments	55.00	24.95
First Discount	35.00	19.75	TD Waterhouse	45.00	17.95
Freedom Investments	25.00	15.00	T. Rowe Price	50.00	19.95
Harrisdirect	40.00	20.00	Vanguard	48.00	20.00
Investors National	39.00	62.50	Wall Street Discount	29.95	19.95
MB Trading	9.95	10.55	York Securities	40.00	36.00

Fuente: *AAII Journal*, enero de 2003.



Autoexamen

8. Millones de estadounidenses trabajan para sus empresas desde sus hogares. A continuación se presenta una muestra de datos que dan las edades de estas personas que trabajan desde sus hogares.

18	54	20	46	25	48	53	27	26	37
40	36	42	25	27	33	28	40	45	25

- Calcule la media y la moda.
 - La edad mediana de la población de todos los adultos es de 36 años (*The World Almanac*, 2006). Use la edad mediana de los datos anteriores para decir si las personas que trabajan desde sus hogares tienden a ser más jóvenes o más viejos que la población de todos los adultos.
 - Calcule el primer y el tercer cuartil.
 - Calcule e interprete el percentil 32.
9. J. D. Powers and Associates hicieron una investigación sobre el número de minutos por mes que los usuarios de teléfonos celulares usan sus teléfonos (Associated Press, junio de 2002). A continuación se muestran los minutos por mes hallados en una muestra de 15 usuarios de teléfonos celulares

615	135	395
430	830	1180
690	250	420
265	245	210
180	380	105

- ¿Cuál es la media de los minutos de uso por mes?
 - ¿Cuál es la mediana de los minutos de uso por mes?
 - ¿Cuál es el percentil 85?
 - J. D. Powers and Associates informa que los planes promedio para usuarios de celulares permiten hasta 750 minutos de uso por mes. ¿Qué indican los datos acerca de la utilización que hacen los usuarios de teléfonos celulares de sus planes mensuales?
10. En una investigación hecha por la Asociación Estadounidense de Hospitales se encontró que la mayor parte de las salas de emergencias de los hospitales estaban operando a toda su capacidad (Associated Press, 9 de abril de 2002). En esta investigación se reunieron datos de los tiempos de espera en las salas de emergencias de hospitales donde éstas operaban a toda su capacidad y de hospitales en que operan de manera equilibrada y rara vez manejan toda su capacidad.

Tiempos de espera para las SE en hospitales a toda capacidad		Tiempos de espera para las SE en hospitales en equilibrio	
87	59	60	39
80	110	54	32
47	83	18	56
73	79	29	26
50	50	45	37
93	66	34	38
72	115		

- Calcule la media y la mediana de estos tiempos de espera en los hospitales a toda capacidad.
- Calcule la media y la mediana de estos tiempos de espera en los hospitales en equilibrio.
- Con base en estos resultados, ¿qué observa acerca de los tiempos de espera para las salas de emergencia? ¿Preocuparán a la Asociación Estadounidense de Hospitales los resultados estadísticos encontrados aquí?

11. En una prueba sobre consumo de gasolina se examinaron a 13 automóviles en un recorrido de 100 millas, tanto en ciudad como en carretera. Se obtuvieron los datos siguientes de rendimiento en millas por galón.

Ciudad: 16.2 16.7 15.9 14.4 13.2 15.3 16.8 16.0 16.1 15.3 15.2 15.3 16.2
Carretera: 19.4 20.6 18.3 18.6 19.2 17.4 17.2 18.6 19.0 21.1 19.4 18.5 18.7

Use la media, la mediana y la moda para indicar cuál es la diferencia en el consumo entre ciudad y carretera.

12. La empresa Walt Disney compró en 7.4 mil millones de dólares Pixar Animation Studios Inc. (CNNMoney.com 24 de enero de 2006). A continuación se presentan las películas animadas producidas por cada una de estas empresas (Disney y Pixar). Las ganancias están en millones de dólares. Calcule las ganancias totales, la media, la mediana y los cuartiles para comparar el éxito de las películas producidas por ambas empresas. ¿Sugieren dichos estadísticos por lo menos una razón por la que Disney haya podido estar interesada en comprar Pixar? Analice.



Películas de Disney	Ganancias (millones de \$)	Películas de Pixar	Ganancias (millones de \$)
<i>Pocahontas</i>	346	<i>Toy Story</i>	362
<i>Hunchback of Notre Dame</i>	325	<i>A Bug's Life</i>	363
<i>Hercules</i>	253	<i>Toy Story 2</i>	485
<i>Mulan</i>	304	<i>Monsters, Inc.</i>	525
<i>Tarzan</i>	448	<i>Finding Nemo</i>	865
<i>Dinosaur</i>	354	<i>The Incredibles</i>	631
<i>The Emperor's New Groove</i>	169		
<i>Lilo & Stitch</i>	273		
<i>Treasure Planet</i>	110		
<i>The Jungle Book 2</i>	136		
<i>Brother Bear</i>	250		
<i>Home on the Range</i>	104		
<i>Chicken Little</i>	249		

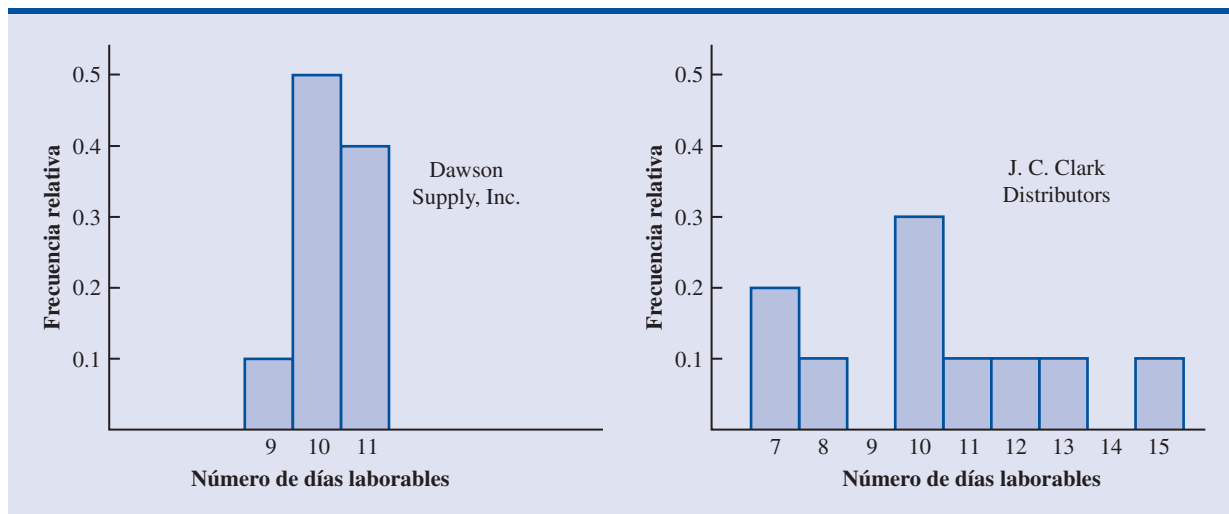
3.2

Medidas de variabilidad

La variabilidad en los tiempos de entrega produce incertidumbre en la planeación de la producción. Los métodos que se presentan en esta sección ayudan a medir y entender la variabilidad.

Además de las medidas de localización, suele ser útil considerar las medidas de variabilidad o de dispersión. Suponga que usted es el encargado de compras de una empresa grande y que con regularidad envía órdenes de compra a dos proveedores. Después de algunos meses de operación, se percata de que el número promedio de días que ambos proveedores requieren para surtir una orden es 10 días. En la figura 3.2 se presentan los histogramas que muestran el número de días que cada uno de los proveedores necesita para surtir una orden. Aunque en ambos casos este número promedio de días es 10 días, ¿muestran los dos proveedores el mismo grado de confiabilidad en términos de tiempos para surtir los productos? Observe la dispersión, o variabilidad, de estos tiempos en ambos histogramas. ¿Qué proveedor preferiría usted?

Para la mayoría de las empresas es importante recibir a tiempo los materiales que necesitan para sus procesos. En el caso de J. C. Clark Distributors sus tiempos de entrega, de siete u ocho días, parecen muy aceptables; sin embargo, sus pocos tiempos de entrega de 13 a 15 días resul-

FIGURA 3.2 DATOS HISTÓRICOS QUE MUESTRAN EL NÚMERO DE DÍAS REQUERIDOS PARA COMPLETAR UNA ORDER

tan desastrosos en términos de mantener ocupada a la fuerza de trabajo y de cumplir con el plan de producción. Este ejemplo ilustra una situación en que la variabilidad en los tiempos de entrega puede ser la consideración más importante en la elección de un proveedor. Para la mayor parte de los encargados de compras, la poca variabilidad que muestra en los tiempos de entrega de Dawson Supply, Inc. hará de esta empresa el proveedor preferido.

Ahora mostramos el estudio de algunas de las medidas de variabilidad más usadas.

Rango

La medida de variabilidad más sencilla es el **rango**.

RANGO

$$\text{Rango} = \text{Valor mayor} - \text{Valor menor}$$

De regreso a los datos de la tabla 3.1 sobre sueldos iniciales de los recién egresados de la carrera de administración, el mayor sueldo inicial es 3925 y el menor 3310. El rango es $3925 - 3310 = 615$.

Aunque el rango es la medida de variabilidad más fácil de calcular, rara vez se usa como única medida. La razón es que el rango se basa sólo en dos observaciones y, por tanto, los valores extremos tienen una gran influencia sobre él. Suponga que uno de los recién egresados haya tenido \$10 000 como sueldo inicial, entonces el rango será $10\,000 - 3310 = 6690$ en lugar de 615. Un valor así no sería muy descriptivo de la variabilidad de los datos ya que 11 de los 12 sueldos iniciales se encuentran entre 3310 y 3730.

Rango intercuartílico

Una medida que no es afectada por los valores extremos es el **rango intercuartílico (RIC)**. Esta medida de variabilidad es la diferencia entre el tercer cuartil Q_3 y el primer cuartil Q_1 . En otras palabras, el rango intercuartílico es el rango en que se encuentra el 50% central de los datos.

RANGO INTERCUARTÍLICO

$$\text{IQR} = Q_3 - Q_1 \quad (3.3)$$

En los datos de los sueldos mensuales iniciales, los cuartiles son $Q_3 = 3600$ y $Q_1 = 3465$. Por lo tanto el rango intercuartílico es $3600 - 3465 = 135$.

Varianza

La **varianza** es una medida de variabilidad que utiliza todos los datos. La varianza está basada en la diferencia entre el valor de cada observación (x_i) y la media. A la diferencia entre cada valor x_i y la media (\bar{x} cuando se trata de una muestra, μ cuando se trata de una población) se le llama *desviación respecto de la media*. Si se trata de una muestra, una desviación respecto de la media se escribe $(x_i - \bar{x})$, y si se trata de una población se escribe $(x_i - \mu)$. Para calcular la varianza, estas desviaciones respecto de la media *se elevan al cuadrado*.

Si los datos son de una población, el promedio de estas desviaciones elevadas al cuadrado es la *varianza poblacional*. La varianza poblacional se denota con la letra griega σ^2 . En una población en la que hay N observaciones y la media poblacional es μ , la varianza poblacional se define como sigue.

VARIANZA POBLACIONAL

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad (3.4)$$

En la mayor parte de las aplicaciones de la estadística, los datos a analizar provienen de una muestra. Cuando se calcula la varianza muestral, lo que interesa es estimar la varianza poblacional σ^2 . Aunque una explicación detallada está más allá del alcance de este libro, es posible demostrar que si la suma de los cuadrados de las desviaciones respecto de la media se divide entre $n - 1$, en lugar de entre n , la varianza muestral que se obtiene constituye un estimador no sesgado de la varianza poblacional. Por esta razón, la *varianza muestral*, que se denota por s^2 , se define como sigue.

La varianza muestral s^2 es el estimador de la varianza poblacional σ^2 .

VARIANZA MUESTRAL

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \quad (3.5)$$

Para ilustrar el cálculo de la varianza muestral, se emplean los datos de los tamaños de cinco grupos de una universidad, presentados en la sección 3.1. En la tabla 3.3 aparece un resumen de los datos con el cálculo de las desviaciones respecto de la media y de los cuadrados de las desviaciones respecto de la media. La suma de los cuadrados de las desviaciones respecto de la media es $\sum (x_i - \bar{x})^2 = 256$. Por tanto, siendo $n - 1 = 4$, la varianza muestral es

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

Antes de continuar, hay que hacer notar que las unidades correspondientes a la varianza muestral suelen causar confusión. Como los valores que se suman para calcular la varianza, $(x_i - \bar{x})^2$, están elevados al cuadrado, las unidades correspondientes a la varianza muestral tam-

TABLA 3.3 CÁLCULO DE LAS DESVIACIONES Y DE LOS CUADRADOS DE LAS DESVIACIONES RESPECTO DE LA MEDIA EMPLEANDO LOS DATOS DE LOS TAMAÑOS DE CINCO GRUPOS DE ESTADOUNIDENSES

Número de estudiantes en un grupo (x_i)	Número promedio de alumnos en un grupo (\bar{x})	Desviación respecto a la media ($x_i - \bar{x}$)	Cuadrado de la desviación respecto de la media ($(x_i - \bar{x})^2$)
46	44	2	4
54	44	10	100
42	44	-2	4
46	44	2	4
32	44	-12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

La varianza sirve para comparar la variabilidad de dos o más variables.

bién están *elevadas al cuadrado*. Por ejemplo, la varianza muestral en los datos de la cantidad de alumnos en los grupos es $s^2 = 64$ (estudiantes)². Las unidades al cuadrado de la varianza dificultan la comprensión e interpretación intuitiva de los valores numéricos de la varianza. Aquí lo recomendable es entender la varianza como una medida útil para comparar la variabilidad de dos o más variables. Al comparar variables, la que tiene la varianza mayor, muestra más variabilidad. Otra interpretación del valor de la varianza suele ser innecesaria.

Para tener otra ilustración del cálculo de la varianza muestral, considere los sueldos iniciales de 12 recién egresados de la carrera de administración, presentados en la tabla 3.1. En la sección 3.1 se vio que la media muestral de los sueldos mensuales iniciales era 3540. En la tabla 3.4 se muestra el cálculo de la varianza muestral ($s^2 = 27\,440.91$).

TABLA 3.4 CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS DE LOS SUELDOS INICIALES

Sueldo mensual (x_i)	Media muestral (\bar{x})	Desviación respecto de la media ($x_i - \bar{x}$)	Cuadrado de la desviación respecto de la media ($(x_i - \bar{x})^2$)
3450	3540	-90	8 100
3550	3540	10	100
3650	3540	110	12 100
3480	3540	-60	3 600
3355	3540	-185	34 225
3310	3540	-230	52 900
3490	3540	-50	2 500
3730	3540	190	36 100
3540	3540	0	0
3925	3540	385	148 225
3520	3540	-20	400
3480	3540	-60	3 600
		0	301 850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Empleando la ecuación (3.5),

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301\,850}{11} = 27\,440.91$$

En las tablas 3.3 y 3.4 se presenta la suma, tanto de las desviaciones respecto de la media como de los cuadrados de las desviaciones respecto de la media. En todo conjunto de datos, la suma de las desviaciones respecto de la media será *siempre igual a cero*. Observe que en las tablas 3.3 y 3.4 $\sum(x_i - \bar{x}) = 0$. Las desviaciones positivas y las desviaciones negativas se anulan mutuamente haciendo que la suma de las desviaciones respecto a la media sea igual a cero.

Desviación estándar

La **desviación estándar** se define como la raíz cuadrada positiva de la varianza. Continuando con la notación adoptada para la varianza muestral y para la varianza poblacional, se emplea s para denotar la desviación estándar muestral y σ para denotar la desviación estándar poblacional. La desviación estándar se obtiene de la varianza como sigue.

La desviación estándar muestral s es el estimador de la desviación estándar poblacional σ .

DESVIACIÓN ESTÁNDAR

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \quad (3.6)$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \quad (3.7)$$

Recuerde que la varianza muestral para los tamaños de cinco grupos de una universidad es $s^2 = 64$. Por tanto, la desviación estándar muestral es $s = \sqrt{64} = 8$. En los datos de los sueldos iniciales, la desviación estándar es $s = \sqrt{27\,440.91} = 165.65$.

La desviación estándar es más fácil de interpretar que la varianza debido a que la desviación estándar se mide en las mismas unidades que los datos.

¿Qué se gana con convertir la varianza en la correspondiente desviación estándar? Recuerde que en la varianza las unidades están elevadas al cuadrado. Por ejemplo, la varianza muestral de los datos de los sueldos iniciales de los egresados de administración es $s^2 = 27,440.91$ (dólares)². Como la desviación estándar es la raíz cuadrada de la varianza, las unidades de la varianza, dólares al cuadrado, se convierten en dólares en la desviación estándar. Por tanto, la desviación estándar de los sueldos iniciales es \$165.65. En otras palabras, la desviación estándar se mide en las mismas unidades que los datos originales. Por esta razón es más fácil comparar la desviación estándar con la media y con otros estadísticos que se miden en las mismas unidades que los datos originales.

Coefficiente de variación

El coeficiente de variación es una medida relativa de la variabilidad; mide la desviación estándar en relación con la media.

En algunas ocasiones se requiere un estadístico descriptivo que indique cuán grande es la desviación estándar en relación con la media. Esta medida es el **coeficiente de variación** y se representa como porcentaje.

COEFICIENTE DE VARIACIÓN

$$\left(\frac{\text{Desviación estándar}}{\text{Media}} \times 100 \right) \% \quad (3.8)$$

En los datos de los tamaños de los cinco grupos de estudiantes, se encontró una media muestral de 44 y una desviación estándar muestral de 8. El coeficiente de variación es $[(8/44) \times 100]\% = 18.2\%$. Expresado en palabras, el coeficiente de variación indica que la desviación estándar muestral es 18.2% del valor de la media muestral. En los datos de los sueldos iniciales, la media muestral encontrada es 3540 y la desviación estándar muestral es 165.65, el coeficiente de variación, $[(165.65/3540) \times 100]\% = 4.7\%$, indica que la desviación estándar muestral es sólo 4.7% del valor de la media muestral. En general, el coeficiente de variación es un estadístico útil para comparar la variabilidad de variables que tienen desviaciones estándar distintas y medias distintas.

NOTAS Y COMENTARIOS

1. Los paquetes de software para estadística y las hojas de cálculo sirven para buscar los estadísticos descriptivos presentados en este capítulo. Una vez que los datos se han ingresado en una hoja de cálculo, basta emplear unos cuantos comandos sencillos para obtener los estadísticos deseados. En los apéndices 3.1 y 3.2 se muestra cómo usar Minitab y Excel para lograrlo.
2. La desviación estándar suele usarse como medida del riesgo relacionado con una inversión en acciones o en fondos de acciones (*BussinesWeek*, 7 de enero de 2000). Proporciona una medida de cómo fluctúa la rentabilidad mensual respecto de la rentabilidad promedio a largo plazo.
3. Redondear los valores de la media muestral \bar{x} y de los cuadrados de las desviaciones $(x_i - \bar{x})^2$

puede introducir errores cuando se emplea una calculadora para el cálculo de la varianza y de la desviación estándar. Para reducir los errores de redondeo se recomienda conservar por lo menos seis dígitos significativos en los cálculos intermedios. La varianza o la desviación estándar obtenidos se redondean entonces a menos dígitos significativos.

4. Otra fórmula alterna para el cálculo de la varianza muestral es

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

donde $\sum x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$.

Ejercicios

Métodos

13. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule el rango y el rango intercuartílico.
14. Considere una muestra que tiene como valores 10, 20, 12, 17 y 16. Calcule la varianza y la desviación estándar.
15. Considere una muestra con valores 27, 25, 0, 15, 30, 34, 28 y 25. Calcule el rango, el rango intercuartílico, la varianza y la desviación estándar.

Aplicaciones

16. Las puntuaciones obtenidas por un jugador de boliche en seis juegos fueron 182, 168, 184, 190, 170 y 174. Use estos datos como una muestra y calcule los estadísticos descriptivos siguientes
 - a. Rango
 - b. Varianza
 - c. Desviación estándar
 - d. Coeficiente de variación
17. *A home theater in a box* es la manera más sencilla y económica de tener sonido envolvente en un centro de entretenimiento en casa. A continuación se presenta una muestra de precios (*Consumer Report Buying Guide* 2004). Los precios corresponden a modelos con y sin reproductor de DVD.

Modelos con reproductor de DVD	Precio	Modelos sin reproductor de DVD	Precio
Sony HT-1800DP	\$450	Pioneer HTP-230	\$300
Pioneer HTD-330DV	300	Sony HT-DDW750	300
Sony HT-C800DP	400	Kenwood HTB-306	360
Panasonic SC-HT900	500	RCA RT-2600	290
Panasonic SC-MTI	400	Kenwood HTB-206	300

- a. Calcule el precio medio de los modelos con reproductor de DVD y el precio medio de los modelos sin reproductor de DVD. ¿Cuánto es lo que se paga de más por tener un reproductor de DVD en casa?
- b. Calcule el rango, la varianza y la desviación estándar de las dos muestras. ¿Qué le dice esta información acerca de los precios de los modelos con y sin reproductor de DVD?

Autoexamen

Autoexamen

18. Las tarifas de renta de automóviles por día en siete ciudades del este de Estados Unidos son las siguientes (*The Wall Street Journal* 16 de enero de 2004).

Ciudad	Tarifa por día
Boston	\$43
Atlanta	35
Miami	34
New York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

- Calcule la media, la varianza y la desviación estándar de estas tarifas.
 - En una muestra similar de siete ciudades del oeste la media muestral de las tarifas fue de \$38 por día. La varianza y la desviación estándar fueron 12.3 y 3.5 cada una. Analice la diferencia entre las tarifas de las ciudades del este y del oeste.
19. *Los Angeles Times* informa con regularidad sobre el índice de la calidad del aire en varias regiones del sur de California. En una muestra de los índices de calidad del aire en Pomona se tienen los datos siguientes: 28, 42, 58, 48, 45, 55, 60, 49 y 50.
- Calcule el rango y el rango intercuartílico.
 - Calcule la varianza muestral y la desviación estándar muestral.
 - En una muestra de índices de calidad del aire en Anaheim, la media muestral es 48.5, la varianza muestral es 136 y la desviación estándar muestral es 11.66. Con base en estos estadísticos descriptivos compare la calidad del aire en Pomona y en Anaheim.
20. A continuación se presentan los datos que se usaron para elaborar los histogramas sobre el número de días necesarios para surtir una orden (véase la figura 3.2).

Días de entrega de Dawson Supply, Inc.: 11 10 9 10 11 11 10 11 10 10
Días de entrega de Clark Distributors: 8 10 13 7 10 11 10 7 15 12

Use el rango y la desviación estándar para sustentar la observación hecha antes de que Dawson Supply proporcione los tiempos de entrega más consistentes.

21. ¿Cómo están los costos de abarrotes en el país? A partir de una canasta alimenticia de 10 artículos entre los que se encuentran carne, leche, pan, huevos, café, papas, cereal y jugo de naranja, la revista *Where to Retire* calculó el costo de la canasta alimenticia en seis ciudades y en seis zonas con personas jubiladas en todo el país (*Where to Retire* noviembre/diciembre de 2003). Los datos encontrados, al dólar más cercano, se presentan a continuación.

Ciudad	Costo	Zona de jubilados	Costo
Buffalo, NY	\$33	Biloxi-Gulfport, MS	\$29
Des Moines, IA	27	Asheville, NC	32
Hartford, CT	32	Flagstaff, AZ	32
Los Angeles, CA	38	Hilton Head, SC	34
Miami, FL	36	Fort Myers, FL	34
Pittsburgh, PA	32	Santa Fe, NM	31

- Calcule la media, varianza y desviación estándar de las ciudades y de las zonas de jubilados.
- ¿Qué observaciones puede hacer con base en estas dos muestras?



22. La Asociación Estadounidense de Inversionistas Individuales realiza cada año una investigación sobre los corredores de bolsa con descuento (*AAII Journal*, enero de 2003). En la tabla 3.2 se muestran las comisiones que cobran 24 corredores de bolsa con descuento por dos tipos de transacciones: transacción con ayuda del corredor de 100 acciones a \$50 la acción y transacción en línea de 500 acciones a \$50 la acción.
- Calcule el rango y el rango intercuartílico en cada tipo de transacción.
 - Calcule la varianza y la desviación estándar en cada tipo de transacción.
 - Calcule el coeficiente de variación en cada tipo de transacción.
 - Compare la variabilidad en el costo que hay en los dos tipos de transacciones
24. Las puntuaciones de un jugador de golf en el 2005 y 2006 son las siguientes:

2005	74	78	79	77	75	73	75	77
2006	71	70	75	77	85	80	71	79

- Use la media y la desviación estándar para evaluar a este jugador de golf en estos dos años.
 - ¿Cuál es la principal diferencia en su desempeño en estos dos años? ¿Se puede ver algún progreso en sus puntuaciones del 2006?, ¿cuál?
24. Los siguientes son los tiempos que hicieron los velocistas de los equipos de pista y campo de una universidad en un cuarto de milla y en una milla (los tiempos están en minutos).

<i>Tiempos en un cuarto de milla:</i>	0.92	0.98	1.04	0.90	0.99
<i>Tiempos en una milla:</i>	4.52	4.35	4.60	4.70	4.50

Después de ver estos datos, el entrenador comentó que en un cuarto de milla los tiempos eran más homogéneos. Use la desviación estándar y el coeficiente de variación para resumir la variabilidad en los datos. El uso del coeficiente de variación, ¿indica que la aseveración del entrenador es correcta?

3.3

Medidas de la forma de la distribución, de la posición relativa y de la detección de observaciones atípicas

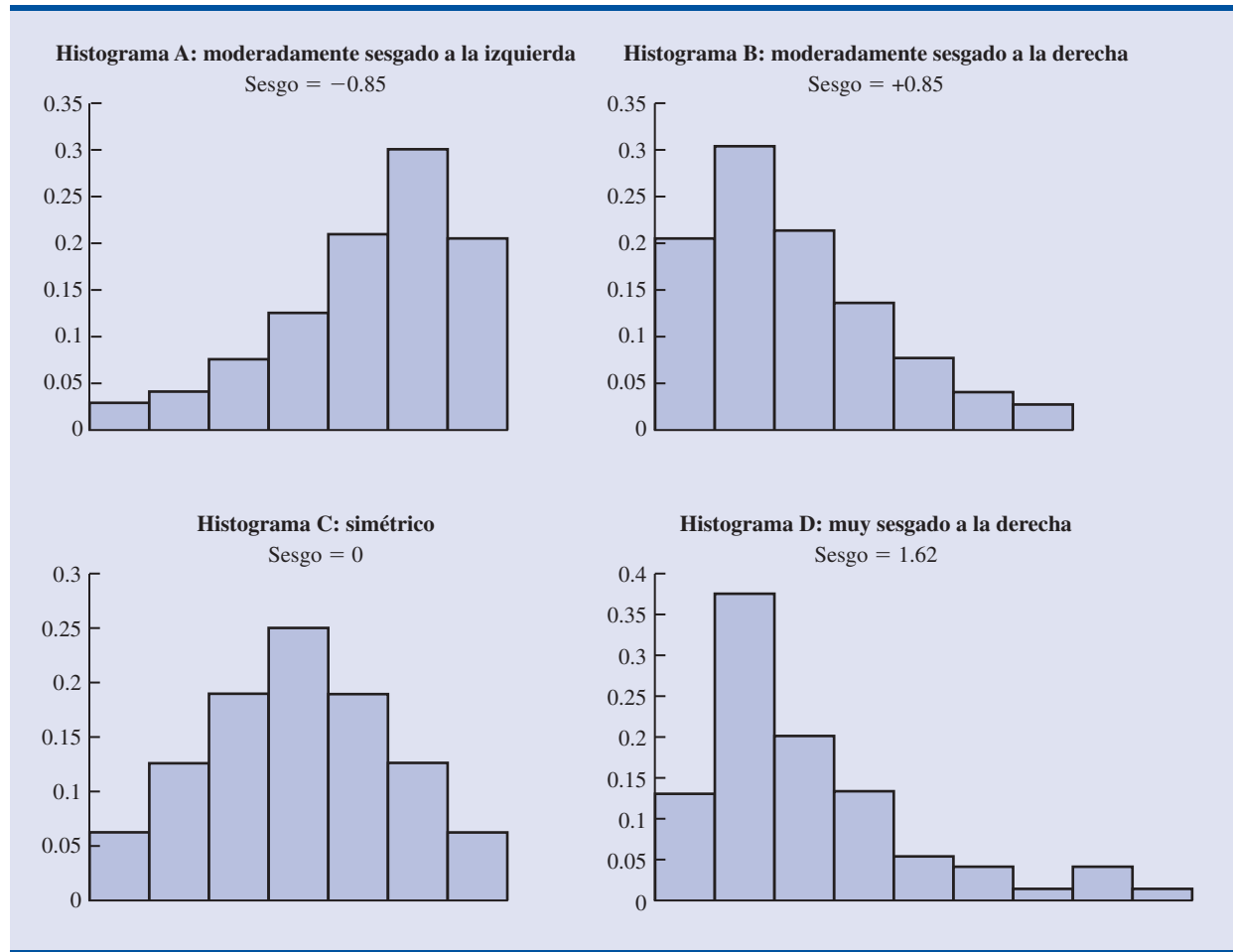
Se han descrito ya varias medidas de localización y de variabilidad de los datos. Además de estas medidas se necesita una medida de la forma de la distribución. En el capítulo 2 se vio que un histograma es una representación gráfica que muestra la forma de una distribución. Una medida numérica importante de la forma de una distribución es el **sesgo**.

Forma de la distribución

En la figura 3.3 se muestran cuatro histogramas elaborados a partir de distribuciones de frecuencias relativas. Los histogramas A y B son moderadamente sesgados. El histograma A es sesgado a la izquierda, su sesgo es -0.85 . El histograma B es sesgado a la derecha, su sesgo es $+0.85$. El histograma C es simétrico; su sesgo es cero. El histograma D es muy sesgado a la derecha; su sesgo es 1.62 . La fórmula que se usa para calcular el sesgo es un poco complicada.* Sin embargo, es fácil de calcular empleando el software para estadística (véase los apéndices 3.1 y 3.2). En

*La fórmula para calcular el sesgo de datos muestrales es:

$$\text{Sesgo} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

FIGURA 3.3 HISTOGRAMAS QUE MUESTRAN EL SESGO DE CUATRO DISTRIBUCIONES

los datos sesgados a la izquierda, el sesgo es negativo; en datos sesgados a la derecha, el sesgo es positivo. Si los datos son simétricos, el sesgo es cero.

En una distribución simétrica, la media y la mediana son iguales. Si los datos están sesgados a la derecha, la media será mayor que la mediana; si los datos están sesgados a la izquierda, la media será menor que la mediana. Los datos que se emplearon para elaborar el histograma D son los datos de las compras realizadas en una tienda de ropa para dama. El monto medio de las compras es \$77.60 y el monto mediano de las compras es \$59.70. Los pocos montos altos de compras tienden a incrementar la media, mientras que a la mediana no le afectan estos montos elevados de compras. Cuando los datos están ligeramente sesgados, se prefiere la mediana como medida de localización.

Puntos z

Además de las medidas de localización, variabilidad y forma, interesa conocer también la ubicación relativa de los valores de un conjunto de datos. Las medidas de localización relativa ayudan a determinar qué tan lejos de la media se encuentra un determinado valor.

A partir de la media y la desviación estándar, se puede determinar la localización relativa de cualquier observación. Suponga que tiene una muestra de n observaciones, en que los valores se

denotan x_1, x_2, \dots, x_n . Suponga además que ya determinó la media muestral, que es \bar{x} y la desviación estándar muestral, que es s . Para cada valor x_i existe otro valor llamado **punto z** . La ecuación (3.9) permite calcular el punto z correspondiente a cada x_i .

PUNTO z

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.9}$$

donde

z_i = punto z para x_i
 \bar{x} = media muestral
 s = desviación estándar muestral

Al punto z también se le suele llamar *valor estandarizado*. El punto z_i puede ser interpretado como el *número de desviaciones estándar a las que x_i se encuentra de la media \bar{x}* . Por ejemplo si $z_1 = 1.2$, esto indica que x_1 es 1.2 desviaciones estándar mayor que la media muestral. De manera similar, $z_2 = -0.5$ indica que x_2 es 0.5 o 1/2 desviación estándar menor que la media muestral. Puntos z mayores a cero corresponden a observaciones cuyo valor es mayor a la media, y puntos z menores que cero corresponden a observaciones cuyo valor es menor a la media. Si el punto z es cero, el valor de la observación correspondiente es igual a la media.

El punto z de cualquier observación se interpreta como una medida relativa de la localización de la observación en el conjunto de datos. Por tanto, observaciones de dos conjuntos de datos distintos que tengan el mismo punto z tienen la misma localización relativa; es decir, se encuentran al mismo número de desviaciones estándar de la media.

En la tabla 3.5 se calculan los puntos z correspondientes a los tamaños de los grupos de estudiantes. Recuerde que ya calculó la media muestral, $\bar{x} = 44$, y la desviación estándar muestral, $s = 8$. El punto z de la quinta observación, que es -1.50 , indica que esta observación está más alejada de la media; esta observación está 1.50 desviaciones estándar más abajo de la media.

Teorema de Chebyshev

El **teorema de Chebyshev** permite decir qué proporción de los valores que se tienen en los datos debe estar dentro de un determinado número de desviaciones estándar de la media.

TABLA 3.5 PUNTOS z CORRESPONDIENTES A LOS DATOS DE LOS TAMAÑOS DE LOS GRUPOS DE ESTUDIANTES

Número de estudiantes en un grupo (x_i)	Desviación respecto de la media ($x_i - \bar{x}$)	Puntos z $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	$2/8 = 0.25$
54	10	$10/8 = 1.25$
42	-2	$-2/8 = -0.25$
46	2	$2/8 = 0.25$
32	-12	$-12/8 = -1.50$

TEOREMA DE CHEBYSHEV

Por lo menos $(1 - 1/z^2)$ de los valores que se tienen en los datos deben encontrarse dentro de z desviaciones estándar de la media, donde z es cualquier valor mayor que 1.

De acuerdo con este teorema para $z = 2, 3$ y 4 desviaciones estándar se tiene

- Por lo menos 0.75, o 75%, de los valores de los datos deben estar dentro de $z = 2$ desviaciones estándar de la media.
- Al menos 0.89, o 89%, de los valores deben estar dentro de $z = 3$ desviaciones estándar de la media.
- Por lo menos 0.94, o 94%, de los valores deben estar dentro de $z = 4$ desviaciones estándar de la media.

Para dar un ejemplo del uso del teorema de Chebyshev, suponga que en las calificaciones obtenidas por 100 estudiantes en un examen de estadística para la administración, la media es 70 y la desviación estándar es 5. ¿Cuántos estudiantes obtuvieron puntuaciones entre 60 y 80?, ¿y cuántos tuvieron puntuaciones entre 58 y 82?

En el caso de las puntuaciones entre 60 y 80 observe que 60 está dos desviaciones estándar debajo de la media y que 80 está dos desviaciones estándar sobre la media. Mediante el teorema de Chebyshev encuentre que por lo menos 0.75, o por lo menos 75%, de las observaciones deben tener valores dentro de dos desviaciones estándar de la media. Así que por lo menos 75% de los estudiantes deben haber tenido puntuaciones entre 60 y 80.

En el caso de las puntuaciones entre 58 y 82, se encuentra que $(58 - 70)/5 = -2.4$, por lo que 58 se encuentra 2.4 desviaciones estándar debajo de la media, y que $(82 - 70)/5 = +2.4$, entonces 82 se encuentra 2.4 desviaciones estándar sobre la media. Al aplicar el teorema de Chebyshev con $z = 2.4$, se tiene

$$\left(1 - \frac{1}{z^2}\right) = \left(1 - \frac{1}{(2.4)^2}\right) = 0.826$$

Por lo menos 82.6% de los estudiantes deben tener puntuaciones entre 58 y 82.

Regla empírica

Una de las ventajas del teorema de Chebyshev es que se aplica a cualquier conjunto de datos, sin importar la forma de la distribución de los datos. En efecto se usa para cualquiera de las distribuciones de la figura 3.3. Sin embargo, en muchas aplicaciones prácticas los datos muestran una distribución simétrica con forma de montaña o de campana como en la figura 3.4. Cuando se cree que los datos tienen aproximadamente esta distribución, se puede emplear la **regla empírica** para determinar el porcentaje de los valores de los datos que deben encontrarse dentro de un determinado número de desviaciones estándar de la media.

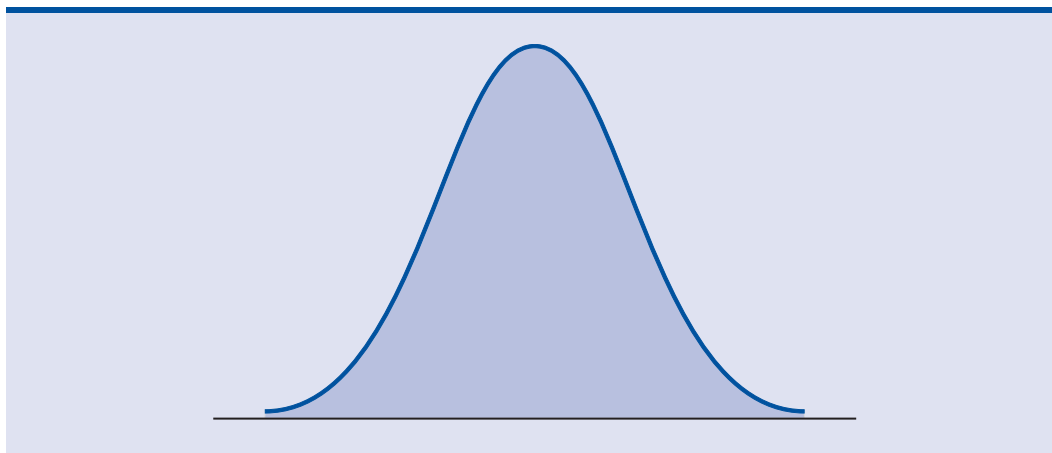
REGLA EMPÍRICA

Cuando los datos tienen una distribución en forma de campana:

- Cerca de 68% de los valores de los datos se encontrarán a no más de una desviación estándar desde la media.
- Aproximadamente 95% de los valores de los datos se encontrarán a no más de dos desviaciones estándar desde la media.
- Casi todos los valores de los datos estarán a no más de tres desviaciones estándar de la media.

En el teorema de Chebyshev se requiere que $z > 1$, pero z no tiene que ser entero.

La regla empírica está basada en la distribución de probabilidad normal, la cual se estudiará en el capítulo 6. La distribución normal se emplea mucho en todo el libro

FIGURA 3.4 DISTRIBUCIÓN EN FORMA DE MONTAÑA O DE CAMPANA

Por ejemplo, los envases con detergente líquido se llenan en forma automática en una línea de producción. Los pesos de llenado suelen tener una distribución en forma de campana. Si el peso medio de llenado es de 16 onzas y la desviación estándar de 0.25 onzas, la regla empírica es aplicada para sacar las conclusiones siguientes:

- Aproximadamente 68% de los envases llenados pesarán entre 15.75 y 16.25 onzas (estarán a no más de una desviación estándar de la media).
- Cerca de 95% de los envases llenados pesarán entre 15.50 y 16.50 onzas (estarán a no más de dos desviaciones estándar de la media).
- Casi todos los envases llenados pesarán entre 15.25 y 16.75 onzas (estarán a no más de tres desviaciones estándar de la media).

Detección de observaciones atípicas

Algunas veces un conjunto de datos tiene una o más observaciones cuyos valores son mucho más grandes o mucho más pequeños que la mayoría de los datos. A estos valores extremos se les llama **observaciones atípicas**. Las personas que se dedican a la estadística y con experiencia en ella toman medidas para identificar estas observaciones atípicas y después las revisan con cuidado. Una observación extraña quizá sea el valor de un dato que se anotó de modo incorrecto. Si es así puede corregirse antes de continuar con el análisis. Una observación atípica tal vez provenga, también, de una observación que se incluyó indebidamente en el conjunto de datos; si es así se puede eliminar. Por último, una observación atípica quizá es un dato con un valor inusual, anotado correctamente y que sí pertenece al conjunto de datos. En tal caso debe conservarse.

Para identificar las observaciones atípicas se emplean los valores estandarizados (puntos z). Recuerde que la regla empírica permite concluir que en los datos con una distribución en forma de campana, casi todos los valores se encuentran a no más de tres desviaciones estándar de la media. Por tanto, si usa los puntos z para identificar las observaciones atípicas, es recomendable considerar cualquier dato cuyo punto z sea menor que -3 o mayor que $+3$ como una observación atípica. Debe examinar la exactitud de tales valores y si en realidad pertenecen al conjunto de datos.

De regreso a los puntos z correspondientes a los datos de los tamaños de grupos de estudiantes de la tabla 3.5, la puntuación -1.50 indica que el tamaño del quinto grupo es el que se encuentra más alejado de la media. Sin embargo, este valor estandarizado queda completamente dentro de los límites de -3 y $+3$. Por tanto, los puntos z no indican que haya observaciones atípicas en estos datos.

Es conveniente determinar si hay observaciones atípicas antes de tomar decisiones con base en el análisis de los datos. Al escribir los datos o al ingresarlos en la computadora suelen cometerse errores. Las observaciones atípicas no necesariamente deben ser eliminadas, pero sí debe verificarse su exactitud y que sean adecuadas.

NOTAS Y COMENTARIOS

1. El teorema de Chebyshev es aplicable a cualquier conjunto de datos y se usa para determi-

nar el número mínimo de los valores de los datos que estarán a no más de un determinado nú-

mero de desviaciones estándar de la media. Si se sabe que los datos tienen forma de campana se puede decir más. Por ejemplo, la regla empírica permite decir que *cerca de 95%* de los valores de los datos estarán a no más de dos desviaciones estándar de la media. El teorema de Chebyshev sólo permite concluir que por lo menos *75%* de los valores de los datos estarán en ese intervalo.

2. Antes de analizar un conjunto de datos, los estadísticos suelen hacer diversas verificaciones para confirmar la validez de los datos. En estudios grandes no es poco común que se cometan errores al anotar los datos o al ingresarlos en la computadora. Identificar las observaciones atípicas es una herramienta usada para verificar la validez de los datos.

Ejercicios

Métodos

25. Considere una muestra cuyos datos tienen los valores 10, 20, 12, 17 y 16. Calcule el punto z de cada una de estas cinco observaciones.
26. Piense en una muestra en que la media es 500 y la desviación estándar es 100. ¿Cuáles son los puntos z de los datos siguientes: 520, 650, 500, 450 y 280?
27. Considere una muestra en que la media es 30 y la desviación estándar es 5. Utilice el teorema de Chebyshev para determinar el porcentaje de los datos que se encuentra dentro de cada uno de los rangos siguientes.
 - a. 20 a 40
 - b. 15 a 45
 - c. 22 a 38
 - d. 18 a 42
 - e. 12 a 48
28. Suponga datos que tienen una distribución en forma de campana cuya media es 30 y desviación estándar 5. Utilice la regla empírica para determinar el porcentaje de los datos que se encuentra dentro de cada uno de los rangos siguientes.
 - a. 20 a 40
 - b. 15 a 45
 - c. 25 a 35

Aplicaciones

29. En una encuesta nacional se encontró que los adultos duermen en promedio 6.9 horas por noche. Suponga que la desviación estándar es 1.2 horas.
 - a. Emplee el teorema de Chebyshev para hallar el porcentaje de individuos que duermen entre 4.5 y 9.3 horas.
 - b. Mediante el teorema de Chebyshev encuentre el porcentaje de individuos que duermen entre 3.9 y 9.9 horas.
 - c. Suponga que el número de horas de sueño tiene una distribución en forma de campana. Use la regla empírica para calcular el porcentaje de individuos que duermen entre 4.5 y 9.3 horas por día. Compare este resultado con el valor que obtuvo en el inciso a empleando este resultado.
30. La Administración de Información de Energía informó que el precio medio del galón de gasolina fue \$2.30 (*Energy Information Administration*, 27 de febrero de 2006). Admita que la desviación estándar haya sido \$0.10 y que el precio del galón de gasolina tenga una distribución en forma de campana.
 - a. ¿Qué porcentaje de la gasolina se vendió entre \$2.20 y \$2.40 por galón?
 - b. ¿Qué porcentaje de la gasolina se vendió entre \$2.20 y \$2.50 por galón?
 - c. ¿Qué porcentaje de la gasolina se vendió a más de \$2.50 por galón?
31. El promedio de los puntos obtenidos en una sección de un examen a nivel nacional fue 507. Si la desviación estándar es aproximadamente 100, conteste las preguntas siguientes usando una distribución en forma de campana y la regla empírica.

Autoexamen

Autoexamen

- a. ¿Qué porcentaje de los estudiantes obtuvo una puntuación superior a 607?
 - b. ¿Qué porcentaje de los estudiantes obtuvo una puntuación superior a 707?
 - c. ¿Qué porcentaje de los estudiantes obtuvo una puntuación entre 407 y 507?
 - d. ¿Qué porcentaje de los estudiantes obtuvo una puntuación entre 307 y 607?
32. En California los altos costos del mercado inmobiliario han obligado a las familias que no pueden darse el lujo de comprar casas grandes, a construir cobertizos como extensión alternativa de sus viviendas. Estos cobertizos suelen aprovecharse como oficinas, estudios de arte, áreas recreativas, etc. El precio medio de un cobertizo es de \$3100 (*Newsweek*, 29 de septiembre de 2003). Asuma que la desviación estándar es de \$1200.
- a. ¿Cuál es el punto z de un cobertizo cuyo precio es de \$2300?
 - b. ¿Cuál es el punto z de un cobertizo cuyo precio es de \$4900?
 - c. Interprete los valores z de los incisos a y b. Diga si alguno de ellos debe ser considerado como una observación atípica.
 - d. El artículo de *Newsweek* describe una combinación oficina-cobertizo cuyo precio fue de \$13 000. ¿Puede considerar este precio como una observación atípica? Explique.
33. La empresa de luz y fuerza de Florida tiene fama de que después de las tormentas repara muy rápidamente sus líneas. Sin embargo en la época de huracanes del 2004 y 2005, la realidad fue otra, su rapidez para reparar sus líneas no fue suficientemente buena (*The Wall Street Journal*, 16 de enero de 2006). Los siguientes datos son de los días que fueron necesarios para restablecer el servicio después de los huracanes del 2004 y 2005.

Huracán	Días para restablecer el servicio
Charley	13
Frances	12
Jeanne	8
Dennis	3
Katrina	8
Rita	2
Wilma	18

Con base en esta muestra de siete, calcule los estadísticos descriptivos siguientes

- a. Media, mediana y moda.
 - b. Rango y desviación estándar.
 - c. ¿En el caso del huracán Wilma considera el tiempo requerido para restablecer el servicio como una observación atípica?
 - d. Estos siete huracanes ocasionaron 10 millones de interrupciones del servicio a los clientes. ¿Indican dichas estadísticas que la empresa debe mejorar su servicio de reparación en emergencias? Discuta.
34. A continuación se presentan los puntos que obtuvieron los equipos en una muestra de 10 juegos universitarios de la NCAA (*USA Today*, 26 de febrero de 2004).

Equipo ganador	Puntos	Equipo perdedor	Puntos	Margen de ganancia
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6

Equipo ganador	Puntos	Equipo perdedor	Puntos	Margen de ganancia
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

- Calcule la media y la desviación estándar de los puntos obtenidos por los equipos ganadores.
 - Suponga que los puntos obtenidos por los equipos ganadores de la NCAA tienen una distribución en forma de campana. Mediante la media y la desviación estándar halladas en el inciso a, estime cuál es el porcentaje de todos los juegos de la NCAA en que el equipo ganador obtuvo 84 puntos o más. Calcule el porcentaje en todos los juegos de la NCAA en que el equipo ganador obtuvo más de 90 puntos.
 - Aproxime la media y la desviación estándar del margen de ganancia. ¿Hay en estos datos alguna observación atípica? Explique.
35. *Consumer Review* publica en Internet estudios y evaluaciones de diversos productos. La siguiente es una lista de 20 sistemas de sonido con sus evaluaciones (www.audioreview.com). La escala de evaluación es de 1 a 5, siendo 5 lo mejor.



Sistema de sonido	Evaluación	Sistema de sonido	Evaluación
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan Aeries	4.26
Mission 73li	4.33	Omni Audio SA 12.3	2.32
PSB 400i	4.50	Polk Audio RT12	4.50
Snell Acoustics D IV	4.64	Sunfire True Subwoofer	4.17
Thiel CS1.5	4.20	Yamaha NS-A636	2.17

- Calcule la media y la mediana.
- Aproxime el primer y el tercer cuartil.
- Estime la desviación estándar.
- El sesgo de estos datos es -1.67 . Comente la forma de esta distribución.
- Calcule los puntos z correspondientes a Allison One y a Ommi Audio
- ¿Hay en estos datos alguna observación atípica? Explique.

3.4

Análisis exploratorio de datos

En el capítulo 2 se introdujeron el diagrama de tallo y hojas como una técnica para el análisis exploratorio de datos. Recuerde que el análisis exploratorio de datos permite usar operaciones aritméticas sencillas y representaciones gráficas fáciles de dibujar para resumir datos. En esta sección, para continuar con el análisis exploratorio de datos, se considerarán los resúmenes de cinco números y los diagramas de caja.

Resumen de cinco números

En el **resumen de cinco números** se usan los cinco números siguientes para resumir los datos.

- El valor menor.
- El primer cuartil (Q_1).
- La mediana (Q_2).

4. El tercer cuartil (Q_3).
5. El valor mayor.

La manera más fácil de elaborar un resumen de cinco números es, primero, colocar los datos en orden ascendente. Hecho esto, es fácil identificar el valor menor, los tres cuartiles y el valor mayor. A continuación se presentan los salarios iniciales de los 12 recién egresados de la carrera de administración, que se presentaron en la tabla 3.1, ordenados de menor a mayor.

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925
		$Q_1 = 3465$			$Q_2 = 3505$ (Mediana)			$Q_3 = 3600$			

La media, que es 3505 y los cuartiles $Q_1 = 3465$ y $Q_3 = 3600$ se calcularon ya en la sección 3.1. Si revisa los datos encontrará que el valor menor es 3310 y el valor mayor es 3925. Así, el resumen de cinco números correspondiente a los datos de los salarios iniciales es 3310, 3465, 3505, 3600, 3925. Entre cada dos números adyacentes del resumen de cinco números se encuentran aproximadamente 25% de los datos.

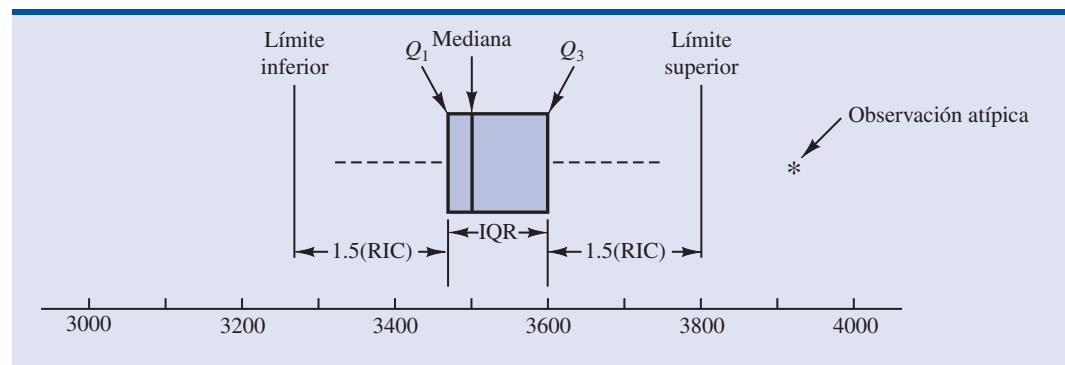
Diagrama de caja

Un **diagrama de caja** es un resumen gráfico de los datos con base en el resumen de cinco números. La clave para la elaboración de un diagrama de caja es el cálculo de la mediana y de los cuartiles Q_1 y Q_3 . También se necesita el rango intercuartílico, $RIC = Q_3 - Q_1$. En la figura 3.5 se presenta el diagrama de caja de los datos de los salarios mensuales iniciales. Los pasos para elaborar un diagrama de caja son los siguientes.

1. Se dibuja una caja cuyos extremos se localicen en el primer y tercer cuartiles. En los datos de los salarios iniciales $Q_1 = 3465$ y $Q_3 = 3600$. Esta caja contiene 50% de los datos centrales.
2. En el punto donde se localiza la mediana (3505 en los datos de los salarios) se traza una línea vertical.
3. Usando el rango intercuartílico, $RIC = Q_3 - Q_1$, se localizan los *límites*. En un diagrama de caja los límites se encuentran 1.5(RIC) abajo del Q_1 y 1.5(RIC) arriba del Q_3 . En el caso de los salarios, $RIC = Q_3 - Q_1 = 3600 - 3465 = 135$. Por tanto, los límites son $3465 - 1.5(135) = 3262.5$ y $3600 + 1.5(135) = 3802.5$. Los datos que quedan fuera de estos límites se consideran *observaciones atípicas*.
4. A las líneas punteadas que se observan en la figura 3.5 se les llama *bigotes*. Los bigotes van desde los extremos de la caja hasta los valores menor y mayor de los límites calculados en el paso 3. Por tanto, los bigotes terminan en los salarios cuyos valores son 3310 y 3730.
5. Por último mediante un asterisco se indica la localización de las observaciones atípicas. En la figura 3.5 se observa que hay una observación atípica, 3925.

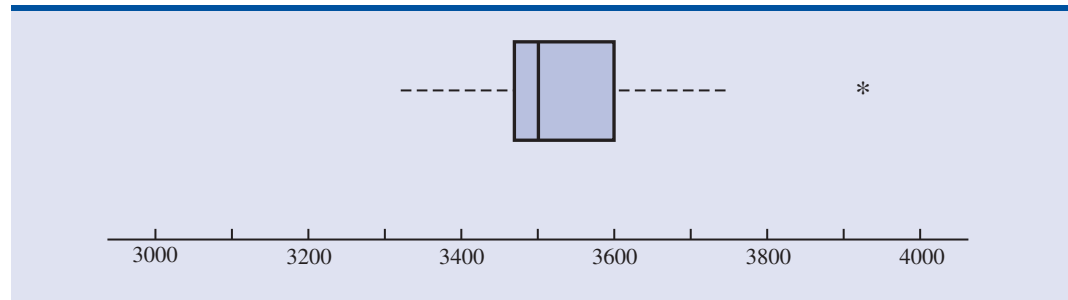
Los diagramas de caja proporcionan otra manera de identificar observaciones atípicas. Pero no necesariamente se identifican los mismos valores que los correspondientes a un punto z menor que -3 o mayor que $+3$. Puede emplear cualquiera de estos procedimientos, o los dos.

FIGURA 3.5 DIAGRAMA DE CAJA DE LOS SALARIOS INICIALES, EN EL QUE SE MUESTRAN LAS LÍNEAS QUE INDICAN LOS LÍMITES INFERIOR Y SUPERIOR



En la figura 3.5 se incluyeron las líneas que indican la localización de los límites superior e inferior. Estas líneas se dibujaron para mostrar cómo se calculan los límites y dónde se localizan en los datos de los salarios iniciales. Los límites, aunque siempre se calculan, por lo general no se dibujan en el diagrama de caja. En la figura 3.6 se muestra la apariencia usual del diagrama de caja de los datos de los salarios iniciales.

FIGURA 3.6 DIAGRAMA DE CAJA DE LOS DATOS DE LOS SALARIOS INICIALES



NOTAS Y COMENTARIOS

1. Una ventaja de los procedimientos del análisis exploratorio de datos es que son fáciles de usar; son necesarios pocos cálculos. Simplemente se ordenan los datos de menor a mayor y se identifican los cinco números del resumen de cinco números. Después se construye el diagrama de caja. No es necesario calcular la media ni la desviación estándar de los datos.
2. En el apéndice 3.1 se muestra cómo elaborar el diagrama de caja de los datos de los salarios iniciales empleando Minitab. El diagrama de caja que se obtiene es similar al de la figura 3.6, pero puesto de lado.

Ejercicios

Métodos

36. Considere una muestra cuyos valores son 27, 25, 20, 15, 30, 34, 28 y 25. Dé el resumen de cinco números de estos datos
37. Muestre diagrama de caja para los datos del ejercicio 36.
38. Elabore el resumen de cinco números y el diagrama de caja de los datos: 5, 15, 18, 10, 8, 12, 16, 10, 6.
39. En un conjunto de datos, el primer cuartil es 42 y el tercer cuartil es 50. Calcule los límites inferior y superior del diagrama de caja correspondiente. El dato con el valor 65, ¿debe considerarse como una observación atípica?

Aplicaciones

40. Ebby Halliday Realtors suministra publicidad sobre propiedades exclusivas ubicadas en Estados Unidos. A continuación se dan los precios de 22 propiedades (*The Wall Street Journal*, 16 de enero de 2004). Los precios se dan en miles

1500	700	2995
895	619	880
719	725	3100
619	739	1699
625	799	1120
4450	2495	1250
2200	1395	912
1280		

Autoexamen

archivo
en
Property CD

Autoexamen

- Muestre el resumen de cinco números.
 - Calcule los límites inferior y superior.
 - La propiedad de mayor precio, \$4 450 000, domina el lago White Rock en Dallas, Texas. ¿Esta propiedad se puede considerar como un valor atípico? Explique.
 - La segunda propiedad más cara que aparece en la lista es de \$3 100 000, ¿debe considerarse como valor atípico? Explique.
 - Dibuje el diagrama de caja.
41. A continuación se presentan las ventas, en millones de dólares, de 21 empresas farmacéuticas.

8 408	1 374	1872	8879	2459	11 413
608	14 138	6452	1850	2818	1 356
10 498	7 478	4019	4341	739	2 127
3 653	5 794	8305			

- Proporcione el resumen de cinco números.
 - Calcule los límites superior e inferior.
 - ¿Hay alguna observación atípica en estos datos?
 - Las ventas de Johnson & Johnson son las mayores de la lista, \$14 138 millones. Suponga que se comete un error al registrar los datos (un error de transposición) y en lugar del valor dado se registra \$41 138 millones. ¿Podría detectar este problema con el método de detección de observaciones atípicas del inciso c, de manera que se pudiera corregir este dato?
 - Dibuje el diagrama de caja.
42. Las nóminas en la liga mayor de béisbol siguen aumentando. Las nóminas de los equipos, en millones, son las siguientes (*USA Today* Online Database, marzo de 2006).



Equipo	Nómina	Equipo	Nómina
Arizona	\$ 62	Milwaukee	\$ 40
Atlanta	86	Minnesota	56
Baltimore	74	NY Mets	101
Boston	124	NY Yankees	208
Chi Cubs	87	Oakland	55
Chi White Sox	75	Philadelphia	96
Cincinnati	62	Pittsburgh	38
Cleveland	42	San Diego	63
Colorado	48	San Francisco	90
Detroit	69	Seattle	88
Florida	60	St. Louis	92
Houston	77	Tampa Bay	30
Kansas City	37	Texas	56
LA Angels	98	Toronto	46
LA Dodgers	83	Washington	49

- ¿Cuál es la mediana de la nómina?
 - Proporcione el resumen de cinco números.
 - ¿Es una observación atípica la nómina de \$208 millones de los Yankees de Nueva York? Explique.
 - Dibuje un diagrama de caja.
43. El presidente de la Bolsa de Nueva York, Richard Grasso, y su junta directiva se vieron cuestionados por el gran paquete de compensaciones pagado a Grasso. El salario más bonos de Grasso, \$8.5 millones, superó el de todos los altos ejecutivos de las principales empresas de servicios financieros. Los datos siguientes muestran los salarios anuales más bonos pagados a los altos eje-

cutivos de 14 empresas de servicios financieros (*The Wall Street Journal*, 17 de septiembre de 2003). Los datos se dan en millones.

Empresa	Salario/bono	Empresa	Salario/bono
Aetna	\$3.5	Fannie Mae	\$4.3
AIG	6.0	Federal Home Loan	0.8
Allstate	4.1	Fleet Boston	1.0
American Express	3.8	Freddie Mac	1.2
Chubb	2.1	Mellon Financial	2.0
Cigna	1.0	Merrill Lynch	7.7
Citigroup	1.0	Wells Fargo	8.0

- a. ¿Cuál es la mediana del salario más bono pagado a los altos ejecutivos de las 14 empresas de servicios financieros?
 - b. Obtenga el resumen de cinco números.
 - c. ¿Se debe considerar el salario más bonos de Grasso, \$8.5 millones, como una observación atípica en el grupo de altos ejecutivos? Explique.
 - d. Presente el diagrama de caja.
44. En la tabla 3.6 se presentan 46 fondos mutualistas y sus rendimientos porcentuales anuales. (*Smart Money*, febrero de 2004.)
- a. ¿Cuáles son los rendimientos porcentuales promedio y la mediana de estos fondos mutualistas?
 - b. ¿Cuáles son el primer y tercer cuartil?
 - c. Obtenga el resumen de cinco números.
 - d. ¿Hay alguna observación atípica en estos datos? Presente el diagrama de caja.



TABLA 3.6 RENDIMIENTOS PORCENTUALES ANUALES EN FONDOS MUTUALISTAS

Fondo mutualista	Rendimiento (%)	Fondo mutualista	Rendimiento (%)
Alger Capital Appreciation	23.5	Nations Small Company	21.4
Alger LargeCap Growth	22.8	Nations SmallCap Index	24.5
Alger MidCap Growth	38.3	Nations Strategic Growth	10.4
Alger SmallCap	41.3	Nations Value Inv	10.8
AllianceBernstein Technology	40.6	One Group Diversified Equity	10.0
Federated American Leaders	15.6	One Group Diversified Int'l	10.9
Federated Capital Appreciation	12.4	One Group Diversified Mid Cap	15.1
Federated Equity-Income	11.5	One Group Equity Income	6.6
Federated Kaufmann	33.3	One Group Int'l Equity Index	13.2
Federated Max-Cap Index	16.0	One Group Large Cap Growth	13.6
Federated Stock	16.9	One Group Large Cap Value	12.8
Janus Adviser Int'l Growth	10.3	One Group Mid Cap Growth	18.7
Janus Adviser Worldwide	3.4	One Group Mid Cap Value	11.4
Janus Enterprise	24.2	One Group Small Cap Growth	23.6
Janus High-Yield	12.1	PBHG Growth	27.3
Janus Mercury	20.6	Putnam Europe Equity	20.4
Janus Overseas	11.9	Putnam Int'l Capital Opportunity	36.6
Janus Worldwide	4.1	Putnam International Equity	21.5
Nations Convertible Securities	13.6	Putnam Int'l New Opportunity	26.3
Nations Int'l Equity	10.7	Strong Advisor Mid Cap Growth	23.7
Nations LargeCap Enhd. Core	13.2	Strong Growth 20	11.7
Nations LargeCap Index	13.5	Strong Growth Inv	23.2
Nation MidCap Index	19.5	Strong Large Cap Growth	14.5

3.5

Medidas de la asociación entre dos variables

Hasta ahora se han examinado métodos numéricos que resumen datos en *una sola variable*. Con frecuencia los administradores o quienes toman decisiones necesitan conocer la *relación entre dos variables*. En esta sección se presentan la covarianza y la correlación como medidas descriptivas de la relación entre dos variables.

Se empieza retomando la aplicación concerniente a la tienda de equipos de sonido que se presentó en la sección 2.4. El administrador de la tienda desea determinar la relación entre el número de comerciales televisados en un fin de semana y las ventas de la tienda durante la semana siguiente. En la tabla 3.7 se presentan datos muestrales de las ventas expresadas en cientos de dólares. En esta tabla se presentan 10 observaciones ($n = 10$), una por cada semana. El diagrama de dispersión en la figura 3.7 muestra una relación positiva, en que las mayores ventas (y) están asociadas con mayor número de comerciales (x). En efecto, el diagrama de dispersión sugiere que podría emplearse una línea recta como aproximación a esta relación. En la argumentación siguiente se introduce la **covarianza** como una medida descriptiva de la asociación entre dos variables.

Covarianza

En una muestra de tamaño n con observaciones (x_1, y_1) , (x_2, y_2) , etc., la covarianza muestral se define como sigue:

COVARIANZA MUESTRAL

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

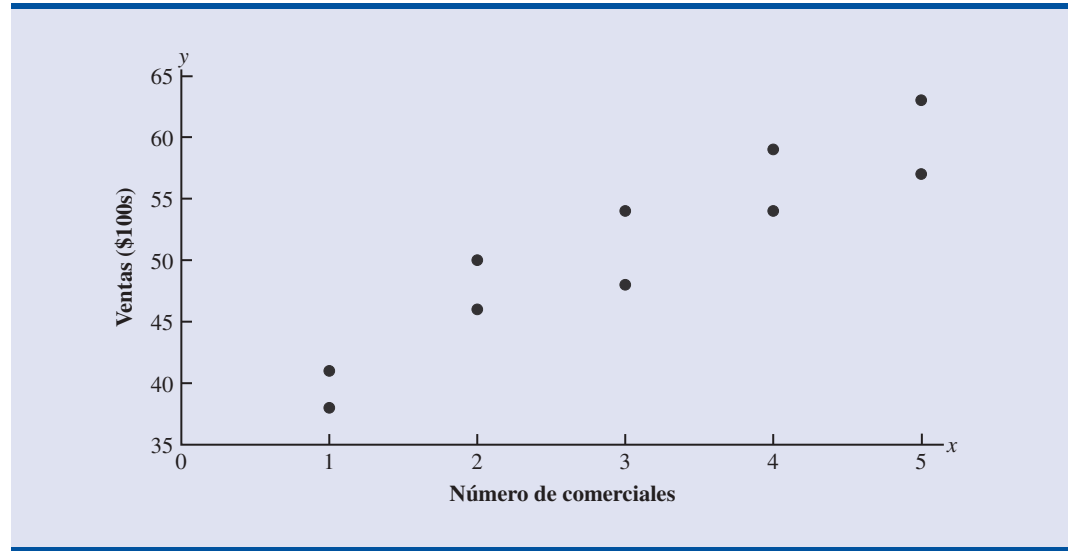
(3.10)

Esta fórmula aparea cada x_i con una y_i . Después se suman los productos obtenidos al multiplicar la desviación de cada x_i de su media muestral \bar{x} por la desviación de la y_i correspondiente de su media muestral \bar{y} ; esta suma se divide entre $n - 1$.

TABLA 3.7 DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

Semana	Número de comerciales x	Volumen de ventas (\$100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



FIGURA 3.7 DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

Para medir, en el problema de la tienda de equipo de sonido, la fuerza de la relación lineal entre el número de comerciales x y el volumen de ventas y , se usa la ecuación (3.10) para calcular la covarianza muestral. En la tabla 3.8 se muestra el cálculo de $\sum(x_i - \bar{x})(y_i - \bar{y})$. Observe que $\bar{x} = 30/10 = 3$ y $\bar{y} = 510/10 = 51$. Empleando la ecuación (3.10) se encuentra que la covarianza muestral es

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

TABLA 3.8 CÁLCULO DE LA COVARIANZA MUESTRAL

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	-1	-1	1
	5	57	2	6	12
	1	41	-2	-10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	-2	-13	26
	5	63	2	12	24
	3	48	0	-3	0
	4	59	1	8	8
	2	46	-1	-5	5
Totales	30	510	0	0	99

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

La fórmula para calcular la covarianza de una población de tamaño N es semejante a la ecuación (3.10), pero la notación usada es diferente para indicar que se está trabajando con toda la población.

COVARIANZA POBLACIONAL

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} \quad (3.11)$$

En la ecuación (3.11) μ_x se usa para denotar la media poblacional de la variable x y μ_y para denotar la media poblacional de la variable y . La covarianza σ_{xy} está definida para una población de tamaño N .

Interpretación de la covarianza

Para ayudar a la interpretación de la covarianza muestral, considere la figura 3.8; presenta el mismo diagrama de dispersión de la figura 3.7 pero con una línea vertical punteada en $\bar{x} = 3$ y una línea horizontal punteada en $\bar{y} = 51$. Estas líneas dividen a la gráfica en cuatro cuadrantes. Los puntos del cuadrante I corresponden a x_i mayor que \bar{x} y y_i mayor que \bar{y} , los puntos del cuadrante II corresponden a x_i menor que \bar{x} y y_i mayor que \bar{y} , etc. Por tanto, los valores de $(x_i - \bar{x})(y_i - \bar{y})$ serán positivos para los puntos del cuadrante I, negativos para los puntos del cuadrante II, positivos para los puntos del cuadrante III y negativos para los puntos del cuadrante IV.

Si el valor de s_{xy} es positivo, los puntos que más influyen sobre s_{xy} deberán encontrarse en los cuadrantes I y III. Por tanto, s_{xy} positivo indica que hay una asociación lineal positiva entre x y y ; es decir, que a medida que el valor de x aumenta, el valor de y aumenta. Si s_{xy} es negativo, los puntos que más influyen sobre s_{xy} deberán encontrarse en los cuadrantes II y IV. Entonces, s_{xy} negativo indica que hay una asociación lineal negativa entre x y y ; esto es, conforme el valor de x aumenta, el valor de y disminuye. Por último, si los puntos tienen distribución uniforme en los cuatro cuadrantes, s_{xy} tendrá un valor cercano a cero, lo que indicará que no hay asociación lineal entre x y y . En la figura 3.9 se muestran los valores de s_{xy} esperables en tres tipos de diagramas de dispersión.

La covarianza es una medida de la asociación lineal entre dos variables.

FIGURA 3.8 DIAGRAMA DE DISPERSIÓN DIVIDIDO PARA LA TIENDA DE EQUIPOS DE SONIDO

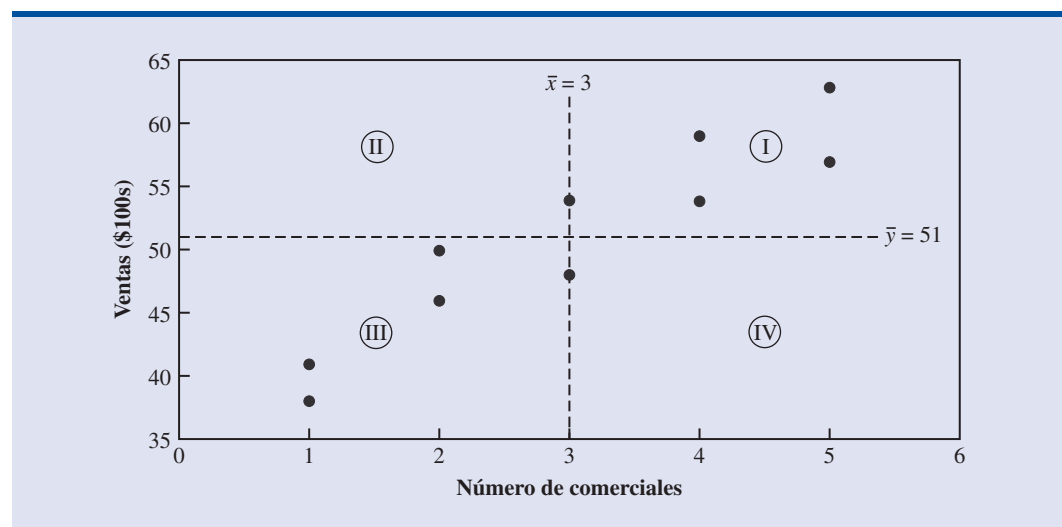
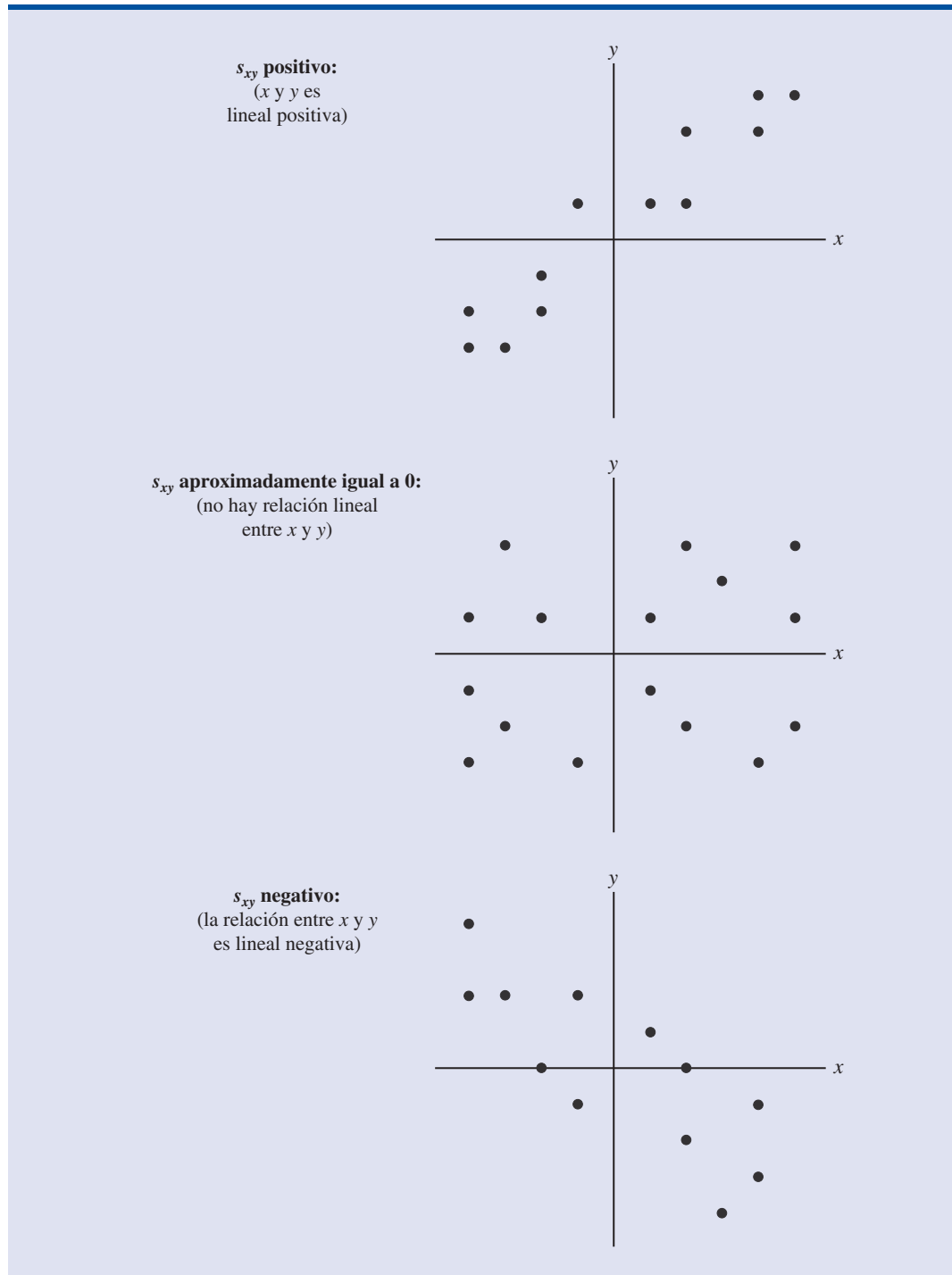


FIGURA 3.9 INTERPRETACIÓN DE LA COVARIANZA MUESTRAL

Si observa otra vez la figura 3.8, encontrará que el diagrama de dispersión de la tienda de equipos de sonido tiene un patrón similar a la gráfica superior de la figura 3.9. Como es de esperarse, el valor de la covarianza muestral indica que hay una relación lineal positiva en la que $s_{xy} = 11$.

Por la argumentación anterior parece que un valor positivo grande de la varianza indica una relación lineal positiva fuerte y que un valor negativo grande indica una relación lineal negativa fuerte. Sin embargo, un problema en el uso de la covarianza, como medida de la fuerza de la relación lineal, es que el valor de la covarianza depende de las unidades de medición empleadas para x y y . Suponga, por ejemplo, que se desea medir la relación entre la estatura x y el peso y de las personas. Es claro que la fuerza de la relación deberá ser la misma, ya sea que la altura se mida en pies o en pulgadas. Sin embargo, cuando la estatura se mide en pulgadas, los valores de $(x_i - \bar{x})$ son mayores que cuando se mide en pies. En efecto, cuando la estatura se mide en pulgadas, el valor del numerador $\sum(x_i - \bar{x})(y_i - \bar{y})$ de la ecuación (3.10) es mayor —entonces la covarianza es mayor— siendo que en realidad la relación no varía. Una medida de la relación entre dos variables, a la cual no le afectan las unidades de medición empleadas para x y y , es el **coeficiente de correlación**.

Coeficiente de correlación

Para datos muestrales el coeficiente de correlación del producto–momento de Pearson está definido como sigue.

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO–MOMENTO DE PEARSON: DATOS MUESTRALES

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad (3.12)$$

donde

r_{xy} = coeficiente de correlación muestral

s_{xy} = covarianza muestral

s_x = desviación estándar muestral de x

s_y = desviación estándar muestral de y

En la ecuación (3.12) se observa que el coeficiente de correlación del producto–momento de Pearson para datos muestrales (llamado *coeficiente de correlación muestral*) se calcula dividiendo la covarianza muestral entre el producto de la desviación estándar muestral de x por la desviación estándar muestral de y .

A continuación se calcula el coeficiente de correlación de los datos de la tienda de equipos para sonido. A partir de la tabla 3.8, se calcula la desviación estándar muestral de las dos variables.

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{20}{9}} = 1.49$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{566}{9}} = 7.93$$

Ahora, como $s_{xy} = 11$, el coeficiente de correlación muestral es igual a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = +0.93$$

La fórmula para calcular el coeficiente de correlación de una población que se denota con la letra griega ρ_{xy} (ro) es la siguiente.

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO-MOMENTO DE PEARSON:
DATOS POBLACIONALES

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (3.13)$$

donde

ρ_{xy} = coeficiente de correlación poblacional

σ_{xy} = covarianza poblacional

σ_x = desviación estándar poblacional de x

σ_y = desviación estándar poblacional de y

El coeficiente de correlación muestral r_{xy} proporciona un estimador del coeficiente de correlación poblacional ρ_{xy} .

El coeficiente de correlación muestral r_{xy} proporciona un estimador del coeficiente de correlación poblacional ρ_{xy} .

Interpretación del coeficiente de correlación

Primero se considerará un ejemplo sencillo que ilustra el concepto de una relación lineal positiva perfecta. En el diagrama de dispersión en la figura 3.10 se representa la relación entre x y y con base en los datos muestrales siguientes.

x_i	y_i
5	10
10	30
15	50

FIGURA 3.10 DIAGRAMA DE DISPERSIÓN QUE REPRESENTA UNA RELACIÓN LINEAL POSITIVA PERFECTA

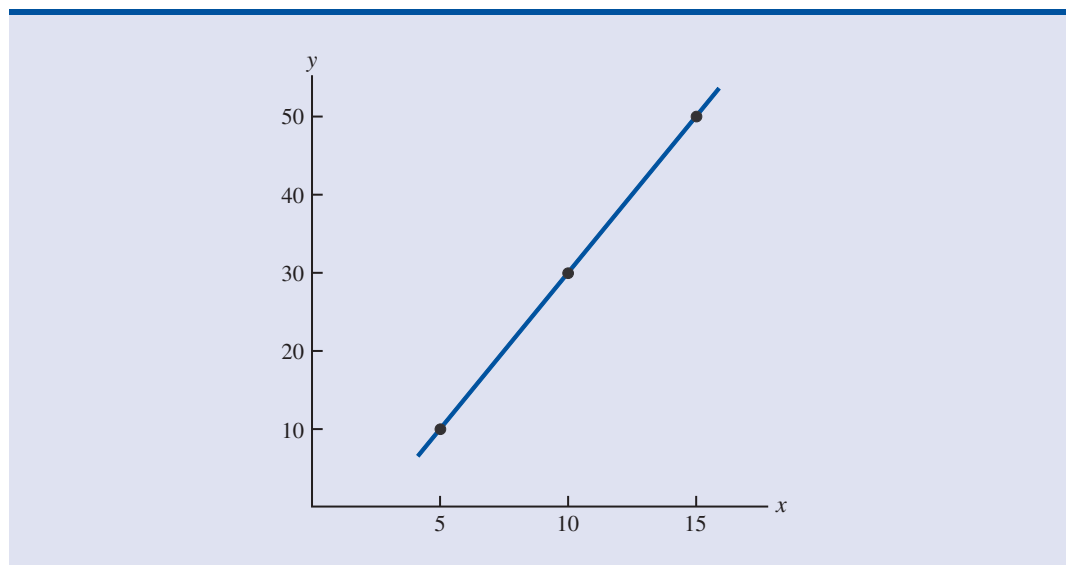


TABLA 3.9 CÁLCULOS PARA OBTENER EL COEFICIENTE DE CORRELACIÓN MUESTRAL

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	-5	25	-20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totales	30	90	0	50	0	800	200
	$\bar{x} = 10 \quad \bar{y} = 30$						

La línea recta trazada a través de los tres puntos expresa una relación lineal perfecta entre x y y . Para emplear la ecuación (3.12) en el cálculo de la correlación muestral, es necesario calcular primero s_{xy} , s_x y s_y . En la tabla 3.9 se muestran parte de los cálculos. Con los resultados de la tabla 3.9 se tiene

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

El coeficiente de correlación va desde -1 hasta $+1$. Los valores cercanos a -1 o a $+1$ corresponden a una relación lineal fuerte. Entre más cercano a cero sea el valor de la correlación, más débil es la relación lineal.

De manera que el valor del coeficiente de correlación muestral es 1.

En general, puede demostrar que si todos los valores del conjunto de datos caen en una línea recta con pendiente positiva, el coeficiente de correlación será $+1$; es decir, un coeficiente de correlación de $+1$ corresponde a una relación lineal positiva perfecta entre x y y . Por otra parte, si los puntos del conjunto de datos caen sobre una línea recta con pendiente negativa, el coeficiente de correlación muestral será -1 ; un coeficiente de correlación de -1 corresponde a una relación lineal negativa perfecta entre x y y .

Suponga ahora que un conjunto de datos muestra una relación lineal positiva entre x y y , pero que la relación no es perfecta. El valor de r_{xy} será menor a 1, indicando que no todos los puntos del diagrama de dispersión se encuentran en una línea recta. Entre más se desvíen los puntos de una relación lineal positiva perfecta, más pequeño será r_{xy} . Si r_{xy} es igual a cero, entonces no hay relación lineal entre x y y ; si r_{xy} tiene un valor cercano a cero, la relación lineal es débil.

Recuerde que en el caso de los datos de la tienda de equipo de sonido $r_{xy} = +0.93$. Entonces se concluye que existe una relación lineal fuerte entre el número de comerciales y las ventas. Más en específico, un aumento en el número de comerciales se asocia con un incremento en las ventas.

Para terminar, es preciso destacar que la correlación proporciona una medida de la asociación lineal y no necesariamente de la causalidad. Que la correlación entre dos variables sea alta no significa que los cambios en una de las variables ocasionen modificaciones en la otra. Por ejemplo, quizá encuentre que las evaluaciones de la calidad y los precios de los restaurantes tengan una correlación positiva. Sin embargo, aumentar los precios de un restaurante no hará que las evaluaciones mejoren.

Ejercicios

Métodos

Autoexamen

45. Las siguientes son cinco observaciones de dos variables

x_i	4	6	11	3	16
y_i	50	50	40	60	30

- Elabore un diagrama de dispersión con x en el eje horizontal.
 - ¿Qué indica el diagrama de dispersión elaborado en el inciso a respecto a la relación entre las dos variables?
 - Calcule e interprete la covarianza muestral.
 - Calcule e interprete el coeficiente de correlación muestral.
46. Las siguientes son cinco observaciones de dos variables.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

- Elabore un diagrama de dispersión con estas variables.
- ¿Qué indica este diagrama de dispersión respecto de la relación entre x y y ?
- Calcule e interprete la covarianza muestral.
- Calcule e interprete el coeficiente de correlación muestral.

Aplicaciones

47. Nielsen Media Research proporciona dos medidas de la audiencia que tienen los programas de televisión: un *rating* de los programas, porcentaje de hogares que tienen televisión y están viendo determinado programa, y un *share* de los programas de televisión, porcentaje de hogares que tienen la televisión encendida y están viendo un determinado programa. Los datos siguientes muestran los datos de *rating* y *share* de Nielsen para la final de la liga mayor de básquetbol en un periodo de nueve años. (Associated Press, 27 de octubre de 2003).

Rating	19	17	17	14	16	12	15	12	13
Share	32	28	29	24	26	20	24	20	22

- Elabore un diagrama de dispersión con los *ratings* en el eje horizontal.
 - ¿Cuál es la relación entre *rating* y *share*? Explique.
 - Calcule e interprete la covarianza muestral.
 - Calcule el coeficiente de correlación muestral. ¿Qué dice este valor acerca de la relación entre *rating* y *share*?
48. En un estudio del departamento de transporte sobre la velocidad y el rendimiento de la gasolina en automóviles de tamaño mediano se obtuvieron los datos siguientes.

Velocidad	30	50	40	55	30	25	60	25	50	55
Rendimiento	28	25	25	23	30	32	21	35	26	25

Calcule e interprete el coeficiente de correlación muestral.

49. *PC World* proporciona evaluaciones de 15 *notebook* PCs (*PC World*, febrero de 2000). La puntuación de funcionamiento mide cuán rápido corre una PC un conjunto de aplicaciones usadas en administración, en comparación con una máquina de línea base. Por ejemplo una PC cuya puntuación de funcionamiento es 200 es dos veces más rápida que una máquina de línea base. Para proporcionar una evaluación general de cada *notebook* probada en el estudio se empleó una escala de 100 puntos. Una puntuación general alrededor de 90 es excepcional, mientras que una de 70 es buena. En la tabla 3.10 se muestran las puntuaciones de funcionamiento y las puntuaciones generales de 15 *notebooks*.

TABLA 3.10 PUNTUACIONES DE FUNCIONAMIENTO Y PUNTUACIONES GENERALES DE 15 *NOTEBOOK* PC

<i>Notebook</i>	Puntuación de funcionamiento	Puntuación general
AMS Tech Roadster 15CTA380	115	67
Compaq Armada M700	191	78
Compaq Prosignia Notebook 150	153	79
Dell Inspiron 3700 C466GT	194	80
Dell Inspiron 7500 R500VT	236	84
Dell Latitude Cpi A366XT	184	76
Enpower ENP-313 Pro	184	77
Gateway Solo 9300LS	216	92
HP Pavilion Notebook PC	185	83
IBM ThinkPad I Series 1480	183	78
Micro Express NP7400	189	77
Micron TransPort NX PII-400	202	78
NEC Versa SX	192	78
Sceptre Soundx 5200	141	73
Sony VAIO PCG-F340	187	77

- Calcule el coeficiente de correlación muestral.
 - ¿Qué indica el coeficiente de correlación muestral acerca de la relación entre la puntuación de funcionamiento y la puntuación general?
50. El Promedio Industrial Dow Jones (DJIA, por sus siglas en inglés) y el Standard & Poor's 500 Index (S&P 500) se usan para medir el mercado bursátil. El DJIA se basa en el precio de las acciones de 30 empresas grandes; el S&P 500 se basa en los precios de las acciones de 500 empresas. Si ambas miden el mercado bursátil, ¿cuál es la relación entre ellas? En los datos siguientes se muestra el aumento porcentual diario o la disminución porcentual diaria del DJIA y del S&P 500 en una muestra de nueve días durante tres meses (*The Wall Street Journal*, 15 de enero a 10 de marzo de 2006).



DJIA	0.20	0.82	-0.99	0.04	-0.24	1.01	0.30	0.55	-0.25
S&P 500	0.24	0.19	-0.91	0.08	-0.33	0.87	0.36	0.83	-0.16

- Muestre el diagrama de dispersión.
 - Calcule el coeficiente de correlación muestral de estos datos.
 - Discuta la asociación entre DJIA y S&P 500. ¿Es necesario consultar ambos para tener una idea general sobre el mercado bursátil diario?
51. Las temperaturas más altas y más bajas en 12 ciudades de Estados Unidos son las siguientes. (Weather Channel, 25 de enero de 2004.)



Ciudad	Alta	Baja	Ciudad	Alta	Baja
Albany	9	-8	Los Angeles	62	47
Boise	32	26	New Orleans	71	55
Cleveland	21	19	Portland	43	36
Denver	37	10	Providence	18	8
Des Moines	24	16	Raleigh	28	24
Detroit	20	17	Tulsa	55	38

- ¿Cuál es la media muestral de las temperaturas diarias más elevadas?
- ¿Cuál es la media muestral de las temperaturas diarias más bajas?
- ¿Cuál es la correlación entre temperaturas más elevadas y temperaturas más bajas?

3.6

La media ponderada y el empleo de datos agrupados

En la sección 3.1 se presentó la media como una de las medidas más importantes de localización central. La fórmula para la media de una muestra en la que hay n observaciones se escribe como sigue.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (3.14)$$

En esta fórmula, a cada x_i se le da la misma importancia o el mismo peso. Aunque esto es lo más común, en algunas situaciones la media se calcula dando a cada observación un peso que refleja su importancia. A una media calculada de esta manera se le llama **media ponderada**.

Media ponderada

La media ponderada se calcula:

MEDIA PONDERADA

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad (3.15)$$

donde

x_i = valor de la observación i

w_i = peso para la observación i

Si los datos provienen de una muestra, la ecuación (3.15) proporciona la media ponderada muestral. Si son de una población, μ se sustituye por \bar{x} en la ecuación (3.15) y se obtiene la media ponderada poblacional.

Como ejemplo de la necesidad de la media ponderada muestral, considere la muestra siguiente de cinco compras de materia prima realizadas en los últimos tres meses.

Compra	Costo por libra (\$)	Número de libras
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

Observe que el costo por libra varía desde \$2.80 hasta \$3.40 y la cantidad comprada varía desde 500 hasta 2 750 libras. Suponga que el administrador quiere información sobre el costo medio por libra de la materia prima. Como las cantidades compradas varían, es necesario emplear la fórmula para la media ponderada. Los valores de los datos de los cinco costos por libra son $x_1 = 3.00$, $x_2 = 3.40$, $x_3 = 2.80$, $x_4 = 2.90$, y $x_5 = 3.25$. El costo medio ponderado por libra se ob-

tiene ponderando cada costo con su cantidad correspondiente. Por ejemplo, los pesos (de ponderación) son $w_1 = 1200$, $w_2 = 500$, $w_3 = 2750$, $w_4 = 1000$ y $w_5 = 800$. De acuerdo con la ecuación (3.15) la media ponderada se calcula:

$$\begin{aligned}\bar{x} &= \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800} \\ &= \frac{18\,500}{6250} = 2.96\end{aligned}$$

Así, los cálculos de la media ponderada indican que el costo medio por libra de materia prima es \$2.96. Observe que si hubiera usado la ecuación (3.14) en lugar de la fórmula para la media ponderada, hubiera obtenido resultados engañosos. En ese caso la media de los valores de los cinco costos por libra sería $(3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \3.07 , valor que exagera el costo medio real por libra comprada.

La selección de las ponderaciones para el cálculo de una determinada media ponderada dependen de la aplicación. Un ejemplo muy conocido por los estudiantes es el promedio de las calificaciones (en Estados Unidos). En este caso los valores de los datos son 4 que corresponde a A, 3 que corresponde a B, 2 que corresponde a C, 1 que corresponde a D y 0 que corresponde a F. Los pesos son los créditos por hora de cada materia. El ejercicio 54 al final de esta sección es un ejemplo del cálculo de esta media ponderada. En otros cálculos de la media ponderada se emplean como pesos cantidades como libras, dólares o volumen. En cualquier caso, si la importancia de las observaciones varía, el analista debe elegir los pesos que mejor reflejen la relevancia de cada observación en la determinación de la media.

El cálculo de las calificaciones es un buen ejemplo del uso de la media ponderada.

Datos agrupados

En la mayor parte de los casos, las medidas de localización y variabilidad se calculan mediante los valores individuales de los datos. Sin embargo, otras veces sólo se tienen datos agrupados o datos en una distribución de frecuencias. En la argumentación siguiente se muestra cómo usar la fórmula de la media ponderada para obtener aproximaciones a la media, la varianza y la desviación estándar de **datos agrupados**.

En la sección 2.2 se presentó una distribución de las duraciones en días en una muestra de auditorías de fin de año de una empresa pequeña de contadores públicos. La distribución de frecuencias de las duraciones de las auditorías que se obtuvo de una muestra de 20 clientes se presenta de nuevo en la tabla 3.11. Con base en esta distribución de frecuencias, ¿cuál es la media muestral de la duración de las auditorías?

Para calcular la media usando datos agrupados, considere el punto medio de cada clase como representativo de los elementos de esa clase. Si M_i denota el punto medio de la clase i y f_i denota la frecuencia de la clase i . Entonces la fórmula para la media ponderada (3.15) se usa con los valores de los datos denotados por M_i y los pesos dados por las frecuencias f_i . En este caso, el denominador de la ecuación (3.15) es la suma de las frecuencias, que es el tamaño de la muestra n .

TABLA 3.11 DISTRIBUCIÓN DE FRECUENCIAS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (en días)	Frecuencia
10–14	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

Es decir, $\sum f_i = n$. De manera que la ecuación para la media muestral de datos agrupados es la siguiente:

MEDIA MUESTRAL DE DATOS AGRUPADOS

$$\bar{x} = \frac{\sum f_i M_i}{n} \quad (3.16)$$

donde

M_i = punto medio de la clase i

f_i = frecuencia de la clase i

n = tamaño de la muestra

Como el punto medio de clase, M_i , se encuentra a la mitad entre los límites de clase, en tabla 3.11 el punto medio de la primera clase, 10–14, es $(10 + 14)/2 = 12$. En la tabla 3.12 se presentan los cinco puntos medios de clase y los cálculos de la media ponderada de los datos de la duración de las auditorías. Como puede ver, la media muestral de la duración de las auditorías es 19 días.

Para calcular la varianza de datos agrupados se emplea una versión ligeramente modificada de la fórmula para la varianza dada en la ecuación (3.5). En la ecuación (3.5) los cuadrados de las desviaciones de los datos respecto a la media muestral se escribieron como $(x_i - \bar{x})^2$. Pero cuando se tienen datos agrupados no se conocen los valores. En este caso, se considera el punto medio de clase, M_i , como representativo de los valores x_i de la clase correspondiente. Por tanto, los cuadrados de las desviaciones respecto a la media $(x_i - \bar{x})^2$ son sustituidos por $(M_i - \bar{x})^2$. Entonces, igual que en el cálculo de la media muestral de datos agrupados, pondere cada valor por la frecuencia de la clase, f_i . La suma de los cuadrados de las desviaciones respecto a la media de todos los datos se aproxima mediante $\sum f_i (M_i - \bar{x})^2$. En el denominador aparece el término $n - 1$ en lugar de n , con objeto de hacer que la varianza muestral sea un estimador de la varianza poblacional. Por consiguiente, la fórmula usada para obtener la varianza muestral de datos agrupados es:

VARIANZA MUESTRAL PARA DATOS AGRUPADOS

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n - 1} \quad (3.17)$$

TABLA 3.12 CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS AGRUPADOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (días)	Punto medio de clase (M_i)	Frecuencia (f_i)	$f_i M_i$
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		<u>20</u>	<u>380</u>

Media muestral $\bar{x} = \frac{\sum f_i M_i}{n} = \frac{380}{20} = 19$ días

TABLA 3.13 CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS AGRUPADOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (días)	Punto medio de clase (M_i)	Frecuencia (f_i)	Desviación ($M_i - \bar{x}$)	Cuadrado de la desviación ($(M_i - \bar{x})^2$)	$f_i(M_i - \bar{x})^2$
10–14	12	4	–7	49	196
15–19	17	8	–2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
					$\Sigma f_i(M_i - \bar{x})^2$

Varianza muestral $s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$

En la tabla 3.13 se presenta el cálculo de la varianza muestral de las duraciones de las auditorías a partir de los datos agrupados de la tabla 3.11, ahí la varianza muestral es 30.

La desviación estándar de datos agrupados es simplemente la raíz cuadrada de la varianza de los datos agrupados. La desviación estándar muestral de los datos de las duraciones de las auditorías es $s = \sqrt{30} = 5.48$.

Antes de terminar esta sección sobre el cálculo de medidas de localización y de dispersión de datos agrupados, debe observar que las fórmulas (3.16) y (3.17) son para muestras. El cálculo de las medidas poblacionales es semejante. A continuación se presentan las fórmulas para la media y la varianza poblacional de datos agrupados.

MEDIA POBLACIONAL DE DATOS AGRUPADOS

$$\mu = \frac{\Sigma f_i M_i}{N}$$

(3.18)

VARIANZA POBLACIONAL DE DATOS AGRUPADOS

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N}$$

(3.19)

NOTAS Y COMENTARIOS

Al calcular los estadísticos descriptivos de datos agrupados, se usan los puntos medios de clase para aproximar los valores de los datos de cada clase. Por tanto, los estadísticos descriptivos de datos agrupados aproximan los estadísticos descriptivos

que se obtendrían si se usaran los datos originales. En consecuencia, es recomendable calcular los estadísticos descriptivos con los datos originales y no con los datos agrupados, siempre que sea posible.

Ejercicios

Métodos

52. Considere los datos siguientes con sus pesos correspondientes

x_i	Peso (w_i)
3.2	6
2.0	3
2.5	2
5.0	8

- Calcule la media ponderada.
- Calcule la media muestral de los cuatro valores de los datos sin los pesos. Observe la diferencia que hay entre los resultados obtenidos con los dos métodos.

53. Considere los datos muestrales de la distribución de frecuencia siguiente.

Clase	Punto medio	Frecuencia
3–7	5	4
8–12	10	7
13–17	15	9
18–22	20	5

- Calcule la media muestral.
- Calcule la varianza muestral y la desviación estándar muestral.

Aplicaciones

54. El promedio de calificaciones de los estudiantes de ciertas escuelas universitarias es el cálculo de una media ponderada. A las calificaciones se les dan los valores siguientes: A (4), B (3), C (2), D (1) y F (0). Después de un semestre de 60 horas de créditos, un estudiante obtuvo las calificaciones siguientes: A en 9 horas de crédito, B en 15 horas, C en 33 horas y D en 3 horas.
- Calcule el promedio de calificaciones de este estudiante.
 - En esta universidad los estudiantes deben tener un promedio de 2.5 para poder seguir sus estudios. ¿Dicho estudiante podrá seguir sus estudios?
55. *Bloomberg Personal Finance* (julio/agosto de 2001) incluye las empresas siguientes en el portafolio de las inversiones que recomienda. A continuación se presentan las cantidades en dólares que asignan a cada acción en un portafolio con valor de \$25 000.

Empresa	Portafolio (\$)	Tasa de crecimiento estimado (%)	Rendimiento de dividendos (%)
Citigroup	3000	15	1.21
General Electric	5500	14	1.48
Kimberly-Clark	4200	12	1.72
Oracle	3000	25	0.00
Pharmacia	3000	20	0.96
SBC Communications	3800	12	2.48
WorldCom	2500	35	0.00

Autoexamen

Autoexamen

- a. Use como pesos las cantidades en dólares del portafolio, ¿cuál es la tasa de crecimiento medio ponderado del portafolio?
 - b. ¿Cuál es el rendimiento medio ponderado de los dividendos en este portafolio?
56. En una investigación realizada entre los suscriptores de la revista *Fortune* se hizo la pregunta siguiente: “De los últimos números ¿cuántos ha leído?” Suponga que en la distribución de frecuencia siguiente se resumen las 500 respuestas.

Números leídos	Frecuencia
0	15
1	10
2	40
3	85
4	350
Total	500

- a. ¿Cuál es la cantidad media de los últimos números que han leído los suscriptores?
 - b. ¿Cuál es la desviación estándar en la cantidad de los últimos números que han leído los suscriptores?
57. La distribución de frecuencias siguiente muestra los precios de las 30 acciones del Promedio Industrial Dow Jones (*The Wall Street Journal*, 16 de enero de 2006).

Precio por acción	Frecuencia
\$20–29	7
\$30–39	6
\$40–49	6
\$50–59	3
\$60–69	4
\$70–79	3
\$80–89	1

Calcule el precio medio por acción y la desviación estándar de los precios por acción en el Promedio Industrial Dow Jones.

Resumen

En este capítulo se presentaron varios estadísticos descriptivos que sirven para resumir la localización, variabilidad y forma de la distribución de un conjunto de datos. A diferencia de los procedimientos gráficos y tabulares presentados en el capítulo 2, las medidas presentadas resumen los datos con valores numéricos. Cuando dichos valores numéricos se obtienen de una muestra, son llamados estadísticos muestrales, cuando se obtienen de una población, son parámetros poblacionales. A continuación se presenta la notación que se acostumbra emplear para estadísticos muestrales y para parámetros poblacionales.

En inferencia estadística a los estadísticos muestrales se les conoce como estimadores puntuales de los parámetros poblacionales.

	Estadístico muestral	Parámetro poblacional
Media	\bar{x}	μ
Varianza	s^2	σ^2
Desviación estándar	s	σ
Covarianza	s_{xy}	σ_{xy}
Correlación	r_{xy}	ρ_{xy}

Como medidas de localización central se definió la media, la mediana y la moda. Después se usó el concepto de percentiles para describir otras localizaciones en el conjunto de datos. A continuación se presentaron el rango, el rango intercuartílico, la varianza, la desviación estándar y el coeficiente de variación como medidas de variabilidad o de dispersión. La primera medida presentada para la forma de la distribución de los datos fue el sesgo; aquí, valores negativos corresponden a distribuciones de datos sesgadas a la izquierda, y valores positivos corresponden a distribuciones de datos sesgadas a la derecha. Después se describió cómo usar la media y la desviación estándar junto con el teorema de Chebyshev y la regla empírica para obtener más información acerca de la distribución de los datos y para identificar observaciones atípicas.

En la sección 3.4 se mostró cómo elaborar un resumen de cinco números y un diagrama de caja para obtener simultáneamente información sobre la localización, variabilidad y forma de una distribución. En la sección 3.5 se presentaron la covarianza y el coeficiente de correlación como medidas de la asociación entre dos variables. En la última sección se vio cómo calcular la media ponderada y cómo calcular media, varianza y desviación estándar de datos agrupados.

Los estadísticos descriptivos, aquí estudiados, pueden calcularse mediante paquetes de software para estadística y hojas de cálculo. En el apéndice 3.1 se muestra cómo obtener la mayor parte de estos estadísticos descriptivos usando Minitab. En el apéndice 3.2 se muestra el uso de Excel para los mismos propósitos.

Glosario

Estadístico muestral Valor numérico usado como una medida que resume una muestra (por ejemplo, la media muestral \bar{x} , la varianza muestral, s^2 y la desviación estándar muestral, s).

Parámetro poblacional Valor numérico que resume una población (por ejemplo, la media poblacional μ , la varianza poblacional, σ^2 y la desviación estándar poblacional, σ).

Estimador puntual Un estadístico muestral como \bar{x} , s^2 y s cuando se usa para estimar el parámetro poblacional correspondiente.

Media Medida de localización central que se calcula sumando los valores de los datos y dividiendo entre el número de observaciones.

Mediana Medida de localización central proporcionada por el valor central de los datos cuando éstos se han ordenado de menor a mayor.

Moda Medida de localización central, definida como el valor que se presenta con mayor frecuencia.

Percentil Un valor tal que por lo menos p por ciento de las observaciones son menores o iguales que este valor y por lo menos $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor. El percentil 50 es la mediana.

Cuartiles Los percentiles 25, 50 y 75, llamados cada uno primer cuartil, segundo cuartil (mediana) y tercer cuartil. Los cuartiles sirven para dividir al conjunto de datos en cuatro partes; cada una contiene aproximadamente 25% de los datos.

Rango Una medida de la variabilidad, que se define como el valor mayor menos el menor.

Rango intercuartílico (RIC) Una medida de la variabilidad, que se define como la diferencia entre el tercer y primer cuartil.

Varianza Una medida de la variabilidad que se basa en los cuadrados de las desviaciones de los datos respecto a la media.

Desviación estándar Una medida de variabilidad obtenida de la raíz cuadrada de la varianza.

Coeficiente de variación Medida de variabilidad relativa que se obtiene al dividir la desviación estándar entre la media y multiplicando el resultado por 100.

Sesgo Medida de la forma de la distribución de los datos. Datos sesgados a la izquierda tienen un sesgo negativo; una distribución de datos simétrica tiene sesgo cero, y datos sesgados a la derecha tienen sesgo positivo.

Punto z Valor que se calcula dividiendo la desviación respecto a la media $(x_i - \bar{x})$ entre la desviación estándar s . A los puntos z también se les conoce como valores estandarizados y denotan el número de desviaciones estándar que x_i se aleja de la media.

Teorema de Chebyshev Un teorema útil para obtener la proporción de valores en los datos que se encuentran a no más de un número determinado de desviaciones estándar de la media.

Regla empírica Regla empleada para calcular el porcentaje de los valores en los datos que se encuentran a no más de una, dos o tres desviaciones estándar de la media, cuando los datos muestran una distribución en forma de campana.

Observación atípica Datos que tienen un valor inusualmente grande o pequeño.

Resumen de cinco números Técnica para el análisis exploratorio de datos, usa cinco números para resumir los datos: el valor menor, el primer cuartil, la mediana, el tercer cuartil, y el valor mayor.

Diagrama de caja Resumen gráfico de los datos que se basa en el resumen de cinco números.

Covarianza Medida de la relación lineal entre dos variables. Si la covarianza es positiva, indica una relación positiva, y si es negativa, una relación negativa.

Coefficiente de correlación Medida de la relación lineal entre dos variables, que puede tener valores desde -1 hasta $+1$. Los valores cercanos a $+1$ indican una fuerte relación lineal positiva; valores cercanos a -1 muestran una fuerte relación lineal negativa, y valores cercanos a cero una ausencia de relación lineal.

Media ponderada Media que se obtiene asignando a cada uno de los valores un peso que refleja su importancia.

Datos agrupados Datos que se dan en intervalos de clase, como cuando se resumen para una distribución de frecuencias. No se tienen los valores de los datos originales.

Fórmulas clave

Media muestral

$$\bar{x} = \frac{\sum x_i}{n} \quad (3.1)$$

Media poblacional

$$\mu = \frac{\sum x_i}{N} \quad (3.2)$$

Rango intercuartílico

$$\text{RIC} = Q_3 - Q_1 \quad (3.3)$$

Varianza poblacional

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \quad (3.4)$$

Varianza muestral

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} \quad (3.5)$$

Desviación estándar

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \quad (3.6)$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \quad (3.7)$$

Coefficiente de variación

$$\left(\frac{\text{Desviación estándar}}{\text{Media}} \times 100 \right) \% \quad (3.8)$$

Punto z

$$z_i = \frac{x_i - \bar{x}}{s} \quad (3.9)$$

Covarianza muestral

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (3.10)$$

Covarianza poblacional

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N} \quad (3.11)$$

Coefficiente de correlación del producto–momento de Pearson: datos muestrales

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \quad (3.12)$$

Coefficiente de correlación del producto–momento de Pearson: datos poblacionales

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (3.13)$$

Media ponderada

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad (3.15)$$

Media muestral de datos agrupados

$$\bar{x} = \frac{\sum f_i M_i}{n} \quad (3.16)$$

Varianza muestral de datos agrupados

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n - 1} \quad (3.17)$$

Media poblacional de datos agrupados

$$\mu = \frac{\sum f_i M_i}{N} \quad (3.18)$$

Varianza poblacional de datos agrupados

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N} \quad (3.19)$$

Ejercicios complementarios

58. De acuerdo con 2003 Annual Consumer Spending Survey, el cargo promedio mensual a una tarjeta de crédito Bank of America Visa fue de \$1838 (*U.S. Airways Attaché Magazine*, diciembre de 2003). En una muestra de cargos mensuales a tarjetas de crédito los datos obtenidos son los siguientes.



236	1710	1351	825	7450
316	4135	1333	1584	387
991	3396	170	1428	1688

- Calcule la media y la mediana.
 - Calcule el primero y tercer cuartil.
 - Calcule el rango y el rango intercuartílico.
 - Calcule la varianza y la desviación estándar.
 - El sesgo en este conjunto de datos es 2.12. Comente la forma de la distribución. ¿Esta es la forma que esperaría? ¿Por qué sí o por qué no?
 - ¿Hay observaciones atípicas en estos datos?
59. La oficina de censos de Estados Unidos proporciona estadísticas sobre las familias en ese país, informaciones como edad al contraer el primer matrimonio, estado civil actual y tamaño de la casa (www.census.gov, 20 de marzo de 2006). Los datos siguientes son edades al contraer el primer matrimonio en una muestra de hombres y en una muestra de mujeres.



Hombres	26	23	28	25	27	30	26	35	28
	21	24	27	29	30	27	32	27	25
Mujeres	20	28	23	30	24	29	26	25	
	22	22	25	23	27	26	19		

- Determine la mediana en la edad de hombres y mujeres al contraer el primer matrimonio.
 - Calcule el primer y tercer cuartil tanto en los hombres como en las mujeres.
 - Hace 30 años la mediana en la edad al contraer el primer matrimonio era 25 años entre los hombres y 22 años entre las mujeres. ¿Qué indica esta información acerca de la edad a la que deciden contraer matrimonio los jóvenes de hoy en día?
60. El rendimiento de los dividendos son los beneficios anuales que paga una empresa por acción dividido entre el precio corriente en el mercado expresado como porcentaje. En una muestra de 10 empresas, los dividendos son los siguientes (*The Wall Street Journal*, 16 de enero de 2004).

Empresa	Porcentaje de rendimiento	Empresa	Porcentaje de rendimiento
Altria Group	5.0	General Motors	3.7
American Express	0.8	JPMorgan Chase	3.5
Caterpillar	1.8	McDonald's	1.6
Eastman Kodak	1.9	United Technology	1.5
ExxonMobil	2.5	Wal-Mart Stores	0.7

- ¿Cuáles son la media y mediana de los rendimientos de dividendos?
- ¿Cuál es la varianza y la desviación estándar?
- ¿Qué empresa proporciona el mayor rendimiento de dividendos?
- ¿Cuál es el punto z correspondiente a McDonalds? Interprete este punto z .
- ¿Cuál es el punto z de General Motors? Interprete este punto z .
- De acuerdo con los puntos z , ¿Hay algún dato atípico en la muestra?

61. El departamento de educación de Estados Unidos informa que cerca de 50% de los estudiantes universitarios toma un préstamo estudiantil como ayuda para cubrir sus gastos (Natural Center for Educational Studies, enero de 2006). Se tomó una muestra de los estudiantes que terminaron sus carreras teniendo una deuda sobre el préstamo estudiantil. Los datos muestran el monto en dólares de estas deudas:

10.1 14.8 5.0 10.2 12.4 12.2 2.0 11.5 17.8 4.0

- Entre los estudiantes que toman un préstamo estudiantil, ¿cuál es la mediana en la deuda que tienen una vez terminados sus estudios?
 - ¿Cuál es la varianza y cuál la desviación estándar?
62. Los propietarios de negocios pequeños suelen contratar a empresas con servicio de nómina para que se encarguen del pago de sus empleados. Las razones son que encuentran regulaciones complicadas para el pago de impuestos y que las multas por errores en los impuestos de los empleados son elevadas. De acuerdo con el Internal Revenue Service, 26% de las declaraciones de impuestos de los empleados contienen errores que ocasionan multas a los dueños. (*The Wall Street Journal*, 30 de enero de 2006). La siguiente es una muestra de 20 multas a propietarios de negocios pequeños.

820 270 450 1010 890 700 1350 350 300 1200
390 730 2040 230 640 350 420 270 370 620

- ¿Cuál es la media en multas?
 - ¿Cuál es la desviación estándar?
 - ¿Es una observación atípica la multa más alta, \$2040?
 - ¿Cuáles son algunas de las ventajas que tienen los propietarios de los negocios pequeños al contratar una empresa de servicio de pago de nómina para que se ocupen del pago a sus empleados, incluyendo la declaración de impuestos de los empleados?
63. El transporte público y el automóvil son los dos medios que usa un empleado para ir a su trabajo cada día. Se presenta una muestra del tiempo requerido con cada medio. Los tiempos se dan en minutos.

Transporte público: 28 29 32 37 33 25 29 32 41 34
Automóvil: 29 31 33 32 34 30 31 32 35 33

- Calcule la media muestral en el tiempo que se necesita con cada transporte.
 - Calcule la desviación estándar para cada transporte.
 - De acuerdo con los resultados en los incisos a y b ¿cuál será el medio de transporte preferido? Explique.
 - Para cada medio de transporte elabore un diagrama de caja. ¿Se confirma la conclusión que dio en el inciso c mediante una comparación de los diagramas de caja?
64. La National Association of Realtors informa sobre la mediana en el precio de una casa en Estados Unidos y sobre el aumento de esta mediana en los últimos cinco años. Use la muestra de precios de casas para responder a las preguntas siguientes.

995.9 48.8 175.0 263.5 298.0 218.9 209.0
628.3 111.0 212.9 92.6 2325.0 958.0 212.5

- ¿Cuál es la mediana muestral de los precios de las casas?
 - En enero del 2001 la National Association of Realtors informó que la mediana en el precio de una casa en Estados Unidos era \$139 300. ¿Cuál ha sido el incremento porcentual de la mediana en el precio de una casa en cinco años?
 - ¿Cuáles son el primer y tercer cuartiles de los datos muestrales?
 - Dé el resumen de cinco números para los precios de las casas.
 - ¿Existe alguna observación atípica en los datos?
 - ¿En la muestra cuál es la media en el precio de una casa? ¿Por qué prefiere la National Association of Realtors usar en sus informes la mediana en el precio de las casas?
65. Los datos siguientes son los gastos en publicidad (en millones de dólares) y los envíos en millones de barriles (bbls.) de las 10 principales marcas de cerveza.





Marca	Gastos en publicidad (millones de dólares)	Despachos en bbls (millones)
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Lite	5.3	4.3
Milwaukee's Best	1.7	4.3

- a. ¿Cuál es la covarianza muestral? ¿Indica que hay una relación positiva o negativa?
- b. ¿Cuál es el coeficiente de correlación?
66. Road & Track proporciona la muestra siguiente de desgaste en llantas y la capacidad de carga máxima de llantas de automóviles.

Desgaste en llantas	Capacidad de carga máxima
75	853
82	1047
85	1135
87	1201
88	1235
91	1356
92	1389
93	1433
105	2039

- a. Con estos datos elabore un diagrama de dispersión en el que el desgaste ocupe el eje x .
- b. Calcule el coeficiente de correlación muestral. ¿Qué indica el coeficiente de correlación muestral acerca de la relación entre el desgaste y la capacidad de carga máxima?
67. Los datos siguientes presentan el seguimiento de la rentabilidad primaria por acción durante 52 semanas y los valores contables reportados por 10 empresas (*The Wall Street Journal*, 13 de marzo de 2000).

Empresa	Valor contable	Rentabilidad
Am Elec	25.21	2.69
Columbia En	23.20	3.01
Con Ed	25.19	3.13
Duke Energy	20.17	2.25
Edison Int'l	13.55	1.79
Enron Cp.	7.44	1.27
Peco	13.61	3.15
Pub Sv Ent	21.86	3.29
Southn Co.	8.77	1.86
Unicom	23.22	2.74

- a. Elabore un diagrama de dispersión, que los valores contables ocupen el eje x .
 - b. Calcule el coeficiente de correlación muestral. ¿Qué indica este coeficiente acerca de la relación entre la rentabilidad por acción y el valor contable?
68. Una técnica de pronóstico conocida como promedios móviles emplea el promedio o la media de los n periodos más recientes para pronosticar el valor siguiente en los datos de una serie de tiempo. En un promedio móvil de tres periodos, se usan los datos de los tres periodos más recientes para calcular el pronóstico. Considere un producto que en los primeros tres meses de este año tuvo la demanda siguiente: enero (800 unidades), febrero (750 unidades) y marzo (900 unidades).
- a. ¿Cuál es pronóstico para abril empleando un promedio móvil de tres meses?
 - b. A una variación de esta técnica se le conoce como promedios móviles ponderados. La ponderación permite que al calcular el pronóstico se le dé más importancia a los datos recientes de la serie de tiempo. Por ejemplo, en un promedio móvil de tres meses a los datos que tienen un mes de antigüedad se les da 3 como peso, 2 a los que tienen dos meses de antigüedad y 1 a los que tienen un mes. Con tales datos, calcule el pronóstico para abril usando promedios móviles de tres meses.
69. A continuación se presentan los días de plazo de vencimiento en una muestra de cinco fondos de mercado de dinero. Aparecen también las cantidades, en dólares, invertidas en los fondos. Emplee la media ponderada para determinar el número medio de días en los plazos de vencimiento de los dólares invertidos en estos cinco fondos de mercado de dinero.

Días de plazo de vencimiento	Valor en dólares
20	20
12	30
7	10
5	15
6	10

70. Un sistema de radar de la policía vigila los automóviles en una carretera que permite una velocidad máxima de 55 millas por hora. La siguiente es una distribución de frecuencias de las velocidades.

Velocidad (millas por hora)	Frecuencia
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10
	<hr/>
Total	475

- a. ¿Cuál es la velocidad media de los automóviles en esta carretera?
- b. Calcule la varianza y la desviación estándar.

Caso problema 1 Las tiendas Pelican

Las tiendas Pelican, una división de National Clothing, es una cadena de tiendas de ropa para mujer con sucursales por todo Estados Unidos. En fechas recientes la cadena realizó una promoción en la que envió cupones de descuento a clientes de otras tiendas de National Clothing. Los datos obtenidos en una muestra de 100 pagos con tarjeta de crédito en las tiendas Pelican, durante un día de la promoción, aparecen en el archivo titulado PelicanStores. En la tabla 3.14 se muestra parte de este conjunto de datos. El modo de pago Proprietary card se refiere a pagos realizados con tarjeta de crédito de National Clothing. A los clientes que hicieron compras con un cupón de descuento se les denomina aquí promocionales y a quienes hicieron sus compras sin emplear cupón de descuento se les denomina regulares. Como a los clientes de las tiendas Pelican no se les enviaron cupones promocionales, los directivos consideran que las ventas hechas a las personas que presentaron un cupón de descuento son ventas que de otro modo no se hubieran realizado. Es obvio que Pelican espera que los clientes promocionales continúen comprando en sus tiendas.

La mayor parte de las variables que aparecen en la tabla 3.14 se explican por sí mismas, pero dos de ellas deben ser aclaradas.

Artículos Número de artículos comprados
Ventas netas Cantidad cargada a la tarjeta de crédito

Los directivos de Pelican desean emplear estos datos muestrales para tener información acerca de sus clientes y evaluar la promoción de los cupones de descuento.

Informe para los directivos

Use los métodos de la estadística descriptiva presentados en este capítulo para resumir los datos y comente sus hallazgos. Su informe debe contener, por lo menos, lo siguiente:

- 1. Estadísticos descriptivos sobre las ventas netas y sobre las ventas a los distintos tipos de clientes.
- 2. Estadísticos descriptivos respecto de la relación entre edad y ventas netas.

TABLA 3.14 MUESTRA DE 100 COMPRAS CON TARJETA DE CRÉDITO REALIZADAS EN LAS TIENDAS PELICAN

Cliente	Tipo de cliente	Ar- tículos	Ventas netas	Modo de pago	Género	Estado civil	Edad
1	Regular	1	39.50	Discover	Masculino	Casado	32
2	Promocional	1	102.40	Proprietary Card	Femenino	Casada	36
3	Regular	1	22.50	Proprietary Card	Femenino	Casada	32
4	Promocional	5	100.40	Proprietary Card	Femenino	Casada	28
5	Regular	2	54.00	MasterCard	Femenino	Casada	34
6	Regular	1	44.50	MasterCard	Femenino	Casada	44
7	Promocional	2	78.00	Proprietary Card	Femenino	Casada	30
8	Regular	1	22.50	Visa	Femenino	Casada	40
9	Promocional	2	56.52	Proprietary Card	Femenino	Casada	46
10	Regular	1	44.50	Proprietary Card	Femenino	Casada	36
.
.
.
96	Regular	1	39.50	MasterCard	Femenino	Casada	44
97	Promocional	9	253.00	Proprietary Card	Femenino	Casada	30
98	Promocional	10	287.59	Proprietary Card	Femenino	Casada	52
99	Promocional	2	47.60	Proprietary Card	Femenino	Casada	30
100	Promocional	1	28.44	Proprietary Card	Femenino	Casada	44



Caso problema 2 Industria cinematográfica

La industria cinematográfica es un negocio muy competido. En más de 50 estudios se producen 300 a 400 películas por año y el éxito financiero de estas películas varía en forma considerable. Las variables usuales para medir el éxito de una película son ventas brutas (en millones de dólares) en el fin de semana del estreno, ventas brutas totales (en millones de dólares), número de salas donde se presenta la película, semanas en las que la película se encuentra entre las 60 mejores en ventas brutas. Los datos de una muestra de 100 películas producidas en 2005 se encuentran en el archivo titulado *Movies*. La tabla 3.15 muestra los datos de las 10 primeras películas que se encuentran en este archivo.

Informe para los directivos

Use los métodos numéricos de la estadística descriptiva presentados en este capítulo para averiguar cómo contribuyen estas variables al éxito de una película. Su informe debe contener lo siguiente.

1. Estadísticos descriptivos para cada una de las cuatro variables con un análisis sobre la información que la estadística descriptiva proporciona acerca de la industria del cine.
2. ¿Hay alguna película que deba ser considerada como una observación atípica de alto desempeño?
3. Los estadísticos descriptivos muestran la relación entre ventas brutas y cada una de las otras variables. Argumente.

TABLA 3.15 DATOS DEL ÉXITO DE 10 PELÍCULAS

Película	Ventas brutas en el estreno (en millones de dólares)	Ventas brutas totales (en millones de dólares)	Número de salas	Semanas en las 60 mejores
<i>Coach Carter</i>	29.17	67.25	2574	16
<i>Ladies in Lavender</i>	0.15	6.65	119	22
<i>Batman Begins</i>	48.75	205.28	3858	18
<i>Unleashed</i>	10.90	24.47	1962	8
<i>Pretty Persuasion</i>	0.06	0.23	24	4
<i>Fever Pitch</i>	12.40	42.01	3275	14
<i>Harry Potter and the Goblet of Fire</i>	102.69	287.18	3858	13
<i>Monster-in-Law</i>	23.11	82.89	3424	16
<i>White Noise</i>	24.11	55.85	2279	7
<i>Mr. and Mrs. Smith</i>	50.34	186.22	3451	21



Caso problema 3 Las escuelas de negocios de Asia-Pacífico

En la actualidad se ha vuelto mundial el interés por tener un grado superior en estudios de negocios. En una investigación se encontró que en Asia cada vez más personas eligen una maestría en administración de negocios como camino hacia el éxito corporativo. De esta manera, en las escuelas de Asia-Pacífico, el número de solicitudes a cursos de maestría en administración de negocios sigue aumentando.

En esa región miles de personas suspenden sus carreras y pasan dos años en estudios para obtener una formación teórica en negocios. Los cursos en estas escuelas son bastante pesados y comprenden economía, banca, marketing, ciencias de la conducta, relaciones laborales, toma de decisiones, pensamiento estratégico, derecho internacional en negocios y otras áreas. En los datos que se presentan en la tabla 3.16 aparecen algunas de las características de las principales escuelas de negocios de Asia-Pacífico.



TABLA 3.16 DATOS DE 25 ESCUELAS DE NEGOCIOS EN ASIA-PACÍFICO

Escuela de negocios	Estudiantes de tiempo completo	Estudiantes por facultad	Colegia- tura para estudiantes		Edad	% de extranjeros	GMAT	Examen de inglés	Experiencia laboral	Salario inicial (\$) (\$)
			locales (\$)	de fuera (\$)						
Melbourne Business School	200	5	24 420	29 600	28	47	Sí	No	Sí	71 400
University of New South Wales (Sydney)	228	4	19 993	32 582	29	28	Sí	No	Sí	65 200
Indian Institute of Management (Ahmedabad)	392	5	4 300	4 300	22	0	No	No	No	7 100
Chinese University of Hong Kong	90	5	11 140	11 140	29	10	Sí	No	No	31 000
International University of Japan (Niiigata)	126	4	33 060	33 060	28	60	Sí	Sí	No	87 000
Asian Institute of Management (Manila)	389	5	7 562	9 000	25	50	Sí	No	Sí	22 800
Indian Institute of Management (Bangalore)	380	5	3 935	16 000	23	1	Sí	No	No	7 500
National University of Singapore	147	6	6 146	7 170	29	51	Sí	Sí	Sí	43 300
Indian Institute of Management (Calcutta)	463	8	2 880	16 000	23	0	No	No	No	7 400
Australian National University (Canberra)	42	2	20 300	20 300	30	80	Sí	Sí	Sí	46 600
Nanyang Technological University (Singapore)	50	5	8 500	8 500	32	20	Sí	No	Sí	49 300
University of Queensland (Brisbane)	138	17	16 000	22 800	32	26	No	No	Sí	49 600
Hong Kong University of Science and Technology	60	2	11 513	11 513	26	37	Sí	No	Sí	34 000
Macquarie Graduate School of Management (Sydney)	12	8	17 172	19 778	34	27	No	No	Sí	60 100
Chulalongkorn University (Bangkok)	200	7	17 355	17 355	25	6	Sí	No	Sí	17 600
Monash Mt. Eliza Business School (Melbourne)	350	13	16 200	22 500	30	30	Sí	Sí	Sí	52 500
Asian Institute of Management (Bangkok)	300	10	18 200	18 200	29	90	No	Sí	Sí	25 000
University of Adelaide	20	19	16 426	23 100	30	10	No	No	Sí	66 000
Massey University (Palmerston North, New Zealand)	30	15	13 106	21 625	37	35	No	Sí	Sí	41 400
Royal Melbourne Institute of Technology Business Graduate School	30	7	13 880	17 765	32	30	No	Sí	Sí	48 900
Jamnalal Bajaj Institute of Management Studies (Bombay)	240	9	1 000	1 000	24	0	No	No	Sí	7 000
Curtin Institute of Technology (Perth)	98	15	9 475	19 097	29	43	Sí	No	Sí	55 000
Lahore University of Management Sciences	70	14	11 250	26 300	23	2.5	No	No	No	7 500
Universiti Sains Malaysia (Penang)	30	5	2 260	2 260	32	15	No	Sí	Sí	16 000
De La Salle University (Manila)	44	17	3 300	3 600	28	3.5	Sí	No	Sí	13 100

Informe para los directivos

Use los métodos de la estadística descriptiva para resumir los datos de la tabla 3.16. Argumente sobre sus hallazgos.

1. Para cada variable presente un resumen del conjunto de datos. Haga comentarios e interpretaciones con base en máximos y mínimos, así como en las medias y proporciones adecuadas. ¿Qué conclusiones nuevas proporcionan estos estadísticos descriptivos respecto de las escuelas de negocios de Asia-Pacífico?
2. Resuma los datos para hacer las comparaciones siguientes:
 - a. Diferencias entre las colegiaturas para alumnos locales y de fuera.
 - b. Diferencias entre los salarios promedio iniciales para egresados de escuelas que exigen experiencia laboral y de escuelas que no la exigen.
 - c. Discrepancias entre los salarios promedio iniciales de egresados de escuelas que exigen una prueba de inglés y de escuelas que no la exigen.
3. ¿Parece haber relación entre los salarios iniciales y las colegiaturas?
4. Presente cualquier gráfica y resumen numérico que pueda servir para comunicar a otras personas la información presentada en la tabla 3.16.

Apéndice 3.1 Estadística descriptiva usando Minitab

En este apéndice se describe cómo usar Minitab para obtener estadísticos descriptivos. En la tabla 3.1 aparecen los sueldos iniciales de 12 recién egresados de la carrera de administración. En el panel A de la figura 3.11 están los estadísticos descriptivos obtenidos para resumir los datos usando Minitab. A continuación se dan las definiciones de los títulos que se observan en el panel A.

N	número de valores en los datos
N*	número de datos faltantes
Mean	media
SE Mean	error estándar de la media
StDev	desviación estándar
Minimum	valor mínimo (menor) en los datos
Q1	primer cuartil
Median	mediana
Q3	tercer cuartil
Maximum	valor máximo (mayor) en los datos

El título SE mean se refiere al *error estándar de la media*. Este valor se obtiene dividiendo la desviación estándar entre la raíz cuadrada de N . La interpretación y uso de esta medición se verá en el capítulo 7, cuando se introduzca el tema del muestreo y de la distribución muestral.

Aunque en los resultados de Minitab no aparecen el rango, el rango intercuartílico, la varianza y el coeficiente de variación, estas medidas son fáciles de calcular a partir de los resultados que aparecen en la figura 3.11; se calculan como sigue.

$$\text{Rango} = \text{Máximo} - \text{Mínimo}$$

$$\text{RIC} = Q_3 - Q_1$$

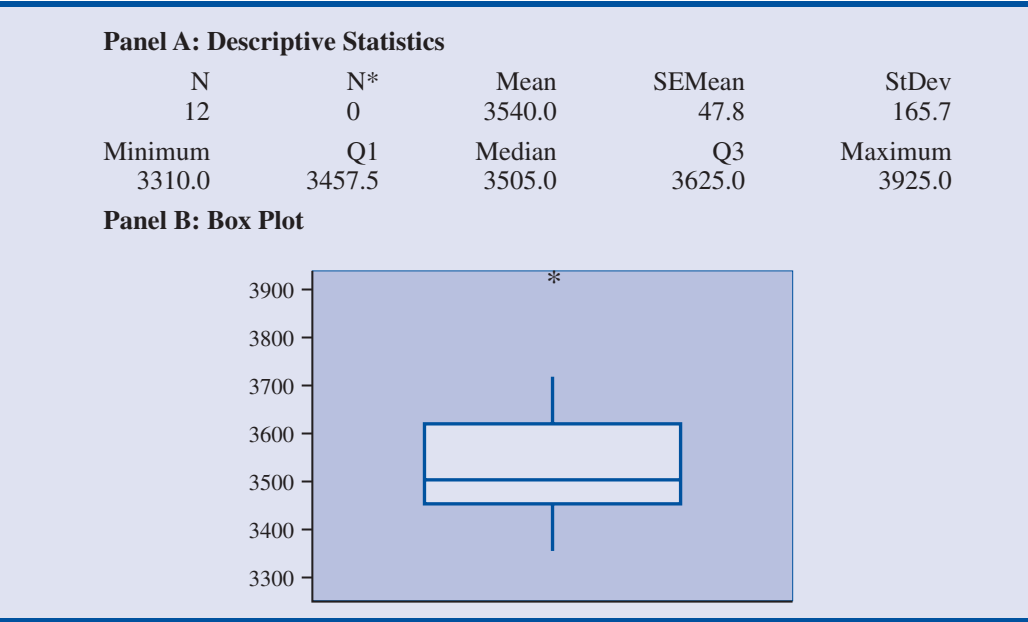
$$\text{Varianza} = (\text{StDev})^2$$

$$\text{Coeficiente de variación} = (\text{StDev}/\text{Media}) \times 100$$

Por último, observe que los cuartiles que da Minitab, $Q_1 = 3457.5$ y $Q_3 = 3625$, son ligeramente diferentes a los calculados en la sección 3.1. Esto se debe al empleo de convenciones* di-

*Cuando se tienen n observaciones ordenadas de menor a mayor (en orden ascendente), para localizar los cuartiles Q_1 y Q_3 Minitab usa las posiciones dadas por $(n + 1)/4$ y $3(n + 1)/4$, respectivamente. Si se obtiene un número fraccionario, Minitab interpola entre los valores de los datos adyacentes ordenados para determinar el cuartil correspondiente.

FIGURA 3.11 ESTADÍSTICOS DESCRIPTIVOS Y DIAGRAMA DE CAJA PROPORCIONADOS POR MINITAB



ferentes para identificar los cuartiles. De manera que los valores Q_1 y Q_3 obtenidos con una convención quizá no sean idénticos a los valores Q_1 y Q_3 obtenidos con otra. Sin embargo, estas diferencias tienden a ser despreciables y los resultados no afectan al hacer las interpretaciones relacionadas con los cuartiles.

Ahora verá cómo se generan los estadísticos que aparecen en la figura 3.11. Los datos de los sueldos iniciales se encuentran en la columna C2 de la hoja de cálculo de Minitab. Para generar los estadísticos descriptivos realice los pasos siguientes:



- Paso 1.** Seleccionar el menú **Stat**
- Paso 2.** Elegir **Basic Statistics**
- Paso 3.** Elegir **Display Descriptive Statistics**
- Paso 4.** Cuando aparece el cuadro de diálogo Display Descriptive Statistics:
 - Ingresar C2 en el cuadro **Variables**
 - Dar clic en **OK**

El panel B de la figura 3.11 es un diagrama de caja obtenido con Minitab y contiene entre el primer y tercer cuartil 50% de los datos. La línea dentro de la caja corresponde a la mediana. El asterisco indica que hay una observación atípica en 3925.

Con los pasos siguientes se genera el diagrama de caja que aparece en la figura 3.11.

- Paso 1.** Seleccionar el menú **Graph**
- Paso 2.** Elegir **Boxplot**
- Paso 3.** Elegir **Simple** y hacer clic en **OK**
- Paso 4.** Cuando aparezca el cuadro de diálogo Boxplot-One Y, Simple:
 - Ingresar C2 en el cuadro **Graph variables**
 - Hacer clic en **OK**

La medida del sesgo tampoco aparece como parte de los resultados estándar de estadística descriptiva que proporciona Minitab. Sin embargo, puede incluirse mediante los pasos siguientes.

FIGURA 3.12 COVARIANZA Y CORRELACIÓN OBTENIDAS USANDO MINITAB CON LOS DATOS DEL NÚMERO DE COMERCIALES Y VENTAS

Covariances: No. of Commercials, Sales Volume		
	No. of Comme	Sales Volume
No. of Comme	2.22222	
Sales Volume	11.00000	62.88889

Correlations: No. of Commercials, Sales Volume		
Pearson correlation of No. of Commercials and Sales Volume = 0.930		
P-Value = 0.000		

Paso 1. Seleccionar el menú **Stat**

Paso 2. Elegir **Basic Statistics**

Paso 3. Elegir **Display Descriptive Statistics**

Paso 4. Cuando aparezca el cuadro de diálogo Display Descriptive Statistics:

Clic en **Statistics**

Elegir **Skewness**

Clic en **OK**

Clic en **OK**

La medida del sesgo, 1.09, aparecerá en su hoja de cálculo.

La figura 3.12 muestra los resultados que da Minitab para la covarianza y la correlación con los datos de la tienda de equipos de sonido presentados en la tabla 3.7. En la parte de la figura que corresponde a la covarianza, *No. of Comme* denota el número de semanas que se televisaron los comerciales y *Sales Volume* las ventas durante la semana siguiente. El valor que aparece en la columna *No. of Comme* y en el renglón *Sales Volume*, 11, es la covarianza muestral que se calculó en la sección 3.5. El valor de la columna *No. of Comme* y en el renglón *No. of Comme*, 2.22222, es la varianza muestral del número de comerciales, y el valor que se encuentra en la columna *Sales Volume* y en el renglón *Sales Volume*, 62.88889, es la varianza muestral de las ventas. El coeficiente de correlación muestral, 0.930, aparece en los resultados, en la parte correspondiente a la correlación. Nota: la interpretación del valor- $p = 0.000$ se verá en el capítulo 9.

Ahora se describe cómo obtener la información que se muestra en la figura 3.12. En la columna C2 de la hoja de cálculo de Minitab ingrese los datos del número de comerciales y en la columna C3 los datos de las ventas. Los pasos necesarios para obtener los resultados que se muestran en los tres primeros renglones de la figura 3.12 son los siguientes.

Paso 1. Seleccionar el menú **Stat**

Paso 2. Elegir **Basic Statistics**

Paso 3. Elegir **Covariance**

Paso 4. Cuando aparezca el cuadro de diálogo Covariance:

Ingresar C2 C3 en el cuadro **Variable**

Clic en **OK**

Para obtener el resultado correspondiente a la correlación, que se observa en la tabla 3.12, sólo hay que hacer una modificación a estos pasos para la covarianza. En el paso 3 seleccione la opción **Correlation**.

Apéndice 3.2 Estadísticos descriptivos usando Excel

Emplee Excel para generar los estadísticos descriptivos vistos en este capítulo. Ahora aprenderá a usar Excel para generar diversas medidas de localización y de variabilidad para una variable, así como la covarianza y el coeficiente de correlación para medir la asociación entre dos variables.



FIGURA 3.13 USO DE LAS FUNCIONES DE EXCEL PARA CALCULAR LA MEDIA, MEDIANA, MODA, VARIANZA Y DESVIACIÓN ESTÁNDAR

	A	B	C	D	E	F
1	Graduate	Starting Salary		Mean	=AVERAGE(B2:B13)	
2	1	3450		Median	=MEDIAN(B2:B13)	
3	2	3550		Mode	=MODE(B2:B13)	
4	3	3650		Variance	=VAR(B2:B13)	
5	4	3480		Standard Deviation	=STDEV(B2:B13)	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

	A	B	C	D	E	F
1	Graduate	Starting Salary		Mean	3540	
2	1	3450		Median	3505	
3	2	3550		Mode	3480	
4	3	3650		Variance	27440.91	
5	4	3480		Standard Deviation	165.65	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

Uso de las funciones de Excel



Excel tiene funciones para calcular media, mediana, moda, varianza muestral y desviación estándar muestral. Con los datos de los sueldos iniciales de la tabla 3.1 ilustrará el uso de las funciones de Excel para calcular la media, mediana, moda, varianza muestral y desviación estándar muestral. Al ir siguiendo los pasos necesarios, consulte la figura 3.13. Ingrese los datos en la columna B.

Para calcular la media emplee la función AVERAGE (PROMEDIO) de Excel ingresando la fórmula siguiente en la celda E1:

=AVERAGE(B2:B13)

De manera similar ingrese en las celdas E2:E5 las fórmulas =MEDIANA(B2:B13), =MODA(B2:B13), =VAR(B2:B13) y =DESVEST(B2:B13) para calcular, respectivamente, la mediana, moda, varianza y desviación estándar. La hoja de cálculo que aparece en primer plano muestra que los valores calculados usando las funciones de Excel son iguales a los ya calculados en este capítulo.

Excel tiene también funciones para calcular la covarianza y el coeficiente de correlación. Al usar estas funciones debe tener cuidado, dado que la función covarianza trata a los datos como población y la función correlación como muestra. Por tanto, los resultados obtenidos con la función covarianza de Excel deben ajustarse para obtener la covarianza muestral. Se le muestra cómo usar estas funciones de Excel para el cálculo de la covarianza muestral y del coeficiente de correlación muestral empleando los datos de la tienda que vende equipos de sonido y que se presentaron en la figura 3.14.



FIGURA 3.14 USO DE LAS FUNCIONES DE EXCEL PARA CALCULAR LA COVARIANZA Y LA CORRELACIÓN

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	=COVAR(B2:B11,C2:C11)	
2	1	2	50		Sample Correlation	=CORREL(B2:B11,C2:C11)	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	9.90	
2	1	2	50		Sample Correlation	0.93	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

La función covarianza de Excel, COVAR, se emplea para calcular la covarianza poblacional ingresando la fórmula siguiente en la celda F1

$$=COVAR(B2:B11,C2:C11)$$

De manera similar ingrese la fórmula: CORREL(B2:B11,C2:C11) para calcular el coeficiente de correlación muestral. En la hoja de cálculo que aparece en primer plano aparecen los valores obtenidos usando estas funciones de Excel. Observe que el valor del coeficiente de correlación muestral (0.93) es el mismo que obtuvo empleando la ecuación (3.12). Sin embargo, el resultado obtenido, 9.9, mediante la función COVAR de Excel, lo obtuvo tratando los datos como población. Por tanto, es necesario ajustar este resultado de Excel para obtener la covarianza muestral. Este ajuste es bastante sencillo. En primer lugar hay que observar que en la fórmula para la covarianza poblacional, ecuación (3.11), requiere dividir entre el número total de observaciones en el conjunto de datos. En cambio, en la fórmula para la covarianza muestral, ecuación (3.10), requiere dividir entre el número total de observaciones menos 1. Entonces, para usar este resultado de Excel, 9.9, para calcular la covarianza muestral, simplemente multiplique 9.9 por $n/(n - 1)$. Como $n = 10$, se tiene

$$s_{xy} = \left(\frac{10}{9}\right)9.9 = 11$$

De esta manera la covarianza muestral de los datos de la tienda de equipos para sonido es 11.

Uso de las herramientas de Excel para estadísticos descriptivos

Como se mostró, Excel tiene funciones estadísticas que permiten calcular los estadísticos descriptivos de un conjunto de datos. Estas funciones sirven para calcular dichos estadísticos de uno en uno (por ejemplo, la media, la varianza, etc.). Excel cuenta también con diversas herramientas para el análisis de datos. Una de estas herramientas llamada Estadística descriptiva, permite calcular varios estadísticos descriptivos de una sola vez. A continuación se le muestra cómo usar

FIGURA 3.15 USO DE LAS HERRAMIENTAS DE EXCEL PARA ESTADÍSTICOS DESCRIPTIVOS

	A	B	C	D	E	F
1	Graduate	Starting Salary		Starting Salary		
2	1	3450				
3	2	3550		Mean	3540	
4	3	3650		Standard Error	47.82	
5	4	3480		Median	3505	
6	5	3355		Mode	3480	
7	6	3310		Standard Deviation	165.65	
8	7	3490		Sample Variance	27440.91	
9	8	3730		Kurtosis	1.7189	
10	9	3540		Skewness	1.0911	
11	10	3925		Range	615	
12	11	3520		Minimum	3310	
13	12	3480		Maximum	3925	
14				Sum	42480	
15				Count	12	
16						



esta herramienta para calcular los estadísticos descriptivos del conjunto de datos referidos a los sueldos iniciales presentados en la tabla 3.1. Consulte la figura 3.15 a medida que se le describen los pasos necesarios.

- Paso 1.** Seleccionar el menú **Herramientas**
- Paso 2.** Elegir **Análisis de datos**
- Paso 3.** Cuando aparezca el cuadro de diálogo Análisis de datos:
 - Elegir **Estadística descriptiva**
 - Clic en **OK**
- Paso 4.** Cuando aparezca el cuadro de diálogo Estadística descriptiva:
 - Ingresar B1:B13 en el cuadro **Rango de entrada**
 - Seleccionar **Agrupados por Columnas**
 - Seleccionar **Rótulos en la primera fila**
 - Seleccionar **Rango de salida**
 - Ingresar D1 en la caja para el rango de salida (para identificar la esquina superior izquierda de la hoja de cálculo en la que aparecerá la estadística descriptiva)
 - Seleccionar **Resumen de estadísticas**
 - Clic en **OK.**

Las celdas D1:D15 de la figura 3.15 muestran la estadística descriptiva obtenida con Excel. Las entradas en negritas son los estadísticos descriptivos que se estudiaron en este capítulo. Los estadísticos descriptivos que no están en negritas se estudiarán en capítulos subsiguientes o en textos más avanzados.

CAPÍTULO 4



Introducción a la probabilidad

CONTENIDO

LA ESTADÍSTICA
EN LA PRÁCTICA:
LA EMPRESA
ROHM AND HASS

- 4.1** EXPERIMENTOS, REGLAS
DE CONTEO Y ASIGNACIÓN
DE PROBABILIDADES
Reglas de conteo, combinaciones
y permutaciones
Asignación de probabilidades
Probabilidades para el proyecto
KP&L

- 4.2** EVENTOS Y SUS
PROBABILIDADES
- 4.3** ALGUNAS RELACIONES
BÁSICAS DE PROBABILIDAD
Complemento de un evento
Ley de la adición
- 4.4** PROBABILIDAD
CONDICIONAL
Eventos independientes
Ley de la multiplicación
- 4.5** TEOREMA DE BAYES
Método tabular



LA ESTADÍSTICA *en* LA PRÁCTICA

LA EMPRESA ROHM AND HASS*

Filadelfia, Pensilvania

Rohm and Hass es el principal productor de materiales especiales, entre los que se encuentran materiales electrónicos, polímeros para pinturas y artículos para el cuidado personal. Los productos de esta empresa permiten la creación de bienes de consumo de vanguardia en mercados como el farmacéutico, el de alimentos, el de suministros para la construcción, equipos de comunicación y productos para el hogar. La fuerza de trabajo de la empresa es de más de 17 000 personas y sus ventas anuales son de \$8 mil millones. Una red de más de 100 puntos de fabricación, investigación técnica y servicio al cliente proporciona los productos y servicios de Rohm and Hass en 27 países.

En el área de productos químicos especiales, la empresa ofrece diversos productos químicos destinados a satisfacer las especificaciones únicas de sus clientes. Para un cliente determinado, la empresa produce un catalizador caro que el cliente emplea en sus procesos químicos. Algunos, pero no todos los lotes que produce la empresa satisfacen las especificaciones del producto. El contrato estipula que el cliente debe probar cada lote después de recibirlo y determinar si el catalizador podrá realizar la función esperada. Los lotes que no pasen la prueba del cliente serán regresados. Con el tiempo, la experiencia ha mostrado que el cliente acepta 60% de los lotes y regresa 40%. Ni el cliente ni la empresa estaban satisfechos con este servicio.

La empresa examinó la posibilidad de, antes de enviar el lote, replicar la prueba que hacía el cliente. Sin embargo, los elevados costos del equipo especial que se necesitaba para la prueba hicieron que esta posibilidad no fuera factible. Los químicos de la empresa encargados del problema propusieron una prueba diferente de costo bajo que se podía practicar antes de enviar el lote al cliente. La empresa creyó que la nueva prueba podría indicar si el catalizador pasaría la compleja prueba que practicaba el cliente.



Una nueva prueba antes de enviar el lote al cliente mejora el servicio al cliente. © Keith Word/Stone.

La pregunta era: ¿cuál es la probabilidad de que el catalizador pase la prueba del cliente dado que pasó la nueva prueba antes de enviar el lote?

La empresa produjo una muestra del catalizador y la sometió a la nueva prueba. Entonces sólo los lotes de catalizador que pasaban la prueba se enviaban al cliente. Mediante el análisis de probabilidad de los datos se supo que si el catalizador pasaba la nueva prueba antes de ser enviado al cliente, la probabilidad de que el catalizador pasara la prueba del cliente era 0.909. O que si el catalizador pasaba la prueba de la empresa, la probabilidad de que no pasara la prueba del cliente y fuera rechazado era 0.091. El análisis de probabilidad aportó evidencias para poner en uso el procedimiento de la prueba antes de enviar el lote. Esta nueva prueba tuvo una mejora inmediata en el servicio al cliente y redujo tanto los costos como los gastos de envío y el manejo de los lotes regresados.

A la probabilidad de que un lote sea aceptado por el cliente, dado que pasó la nueva prueba, se le llama probabilidad condicional. En este capítulo aprenderá cómo calcular la probabilidad condicional y otras probabilidades útiles en la toma de decisiones.

*Los autores agradecen a Michael Haskell, de la subsidiaria Morton International de Rohm and Hass por haberles proporcionado este artículo para *La estadística en la práctica*.

Los administradores sustentan sus decisiones en un análisis de incertidumbres como las siguientes:

1. ¿Qué posibilidades hay de que disminuyan las ventas si aumentamos los precios?
2. ¿Qué posibilidad hay de que un método nuevo de ensamblado aumente la productividad?
3. ¿Cuáles son las posibilidades de que el producto se tenga listo a tiempo?
4. ¿Qué oportunidad existe de que una nueva invención sea rentable?

Algunos de los primeros trabajos sobre probabilidad se dieron en una serie de cartas entre Pierre de Fermat y Blaise Pascal durante el año de 1650.

La **probabilidad** es una medida numérica de la posibilidad de que ocurra un evento. Por tanto, las probabilidades son una medida del grado de incertidumbre asociado con cada uno de los eventos previamente enunciados. Si cuenta con las probabilidades, tiene la capacidad de determinar la posibilidad de ocurrencia que tiene cada evento.

Los valores de probabilidad se encuentran en una escala de 0 a 1. Los valores cercanos a 0 indican que las posibilidades de que ocurra un evento son muy pocas. Los cercanos a 1 indican que es casi seguro que ocurra un evento. Otras probabilidades entre cero y uno representan distintos grados de posibilidad de que ocurra un evento. Por ejemplo, si considera el evento “que llueva mañana”, se entiende que si el pronóstico del tiempo dice “la probabilidad de que llueva es cercana a cero”, implica que casi no hay posibilidades de que llueva. En cambio, si informan que la probabilidad de que llueva es 0.90, sabe que es muy posible que llueva. La probabilidad de 0.50 indica que es igual de posible que llueva como que no llueva. En la figura 4.1 se presenta la probabilidad como una medida numérica de la posibilidad de que ocurra un evento.

4.1

Experimentos, reglas de conteo y asignación de probabilidades

En el contexto de la probabilidad, un **experimento** es definido como un proceso que genera resultados definidos. Y en cada una de las repeticiones del experimento, habrá uno y sólo uno de los posibles resultados experimentales. A continuación se dan varios ejemplos de experimentos con sus correspondientes resultados.

Experimento	Resultado experimental
Lanzar una moneda	Cara, cruz
Tomar una pieza para inspeccionarla	Con defecto, sin defecto
Realizar una llamada de ventas	Hay compra, no hay compra
Lanzar un dado	1, 2, 3, 4, 5, 6
Jugar un partido de fútbol	Ganar, perder, empatar

Al especificar todos los resultados experimentales posibles, está definiendo el **espacio muestral** de un experimento.

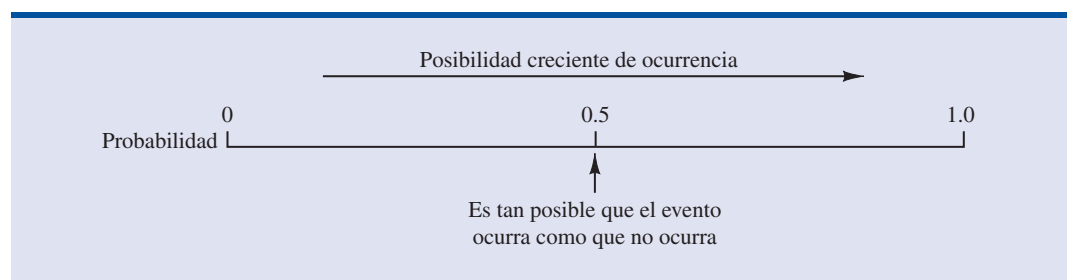
ESPACIO MUESTRAL

El espacio muestral de un experimento es el conjunto de todos los resultados experimentales.

A los resultados experimentales también se les llama puntos muestrales.

A un resultado experimental también se le llama **punto muestral** para identificarlo como un elemento del espacio muestral.

FIGURA 4.1 PROBABILIDAD COMO MEDIDA NUMÉRICA DE LA POSIBILIDAD DE QUE UN EVENTO OCURRA



Considere el primer experimento presentado en la tabla anterior, lanzar una moneda. La cara de la moneda que caiga hacia arriba —cara o cruz— determina el resultado experimental (puntos muestrales). Si denota con S el espacio muestral, puede emplear la notación siguiente para describir el espacio muestral.

$$S = \{\text{Cara, cruz}\}$$

En el segundo experimento de la tabla —tomar una pieza para revisarla— puede describir el espacio muestral como sigue:

$$S = \{\text{Defectuosa, no defectuosa}\}$$

Los dos experimentos descritos tienen dos resultados experimentales (puntos muestrales). Pero, observe ahora el cuarto experimento enumerado en la tabla, lanzar un dado. Los resultados experimentales, definidos por el número de puntos del dado en la cara que cae hacia arriba, son los seis puntos del espacio muestral de este experimento.

$$S = \{1, 2, 3, 4, 5, 6\}$$

Reglas de conteo, combinaciones y permutaciones

Al asignar probabilidades es necesario saber identificar y contar los resultados experimentales. A continuación tres reglas de conteo que son muy utilizadas.

Experimentos de pasos múltiples La primera regla de conteo sirve para experimentos de pasos múltiples. Considere un experimento que consiste en lanzar dos monedas. Defina los resultados experimentales en términos de las caras y cruces que se observan en las dos monedas. ¿Cuántos resultados experimentales tiene este experimento? El experimento de lanzar dos monedas es un experimento de dos pasos: el paso 1 es lanzar la primera moneda y el paso 2 es lanzar la segunda moneda. Si se emplea H para denotar cara y T para denotar cruz, (H, H) será el resultado experimental en el que se tiene cara en la primera moneda y cara en la segunda moneda. Si continúa con esta notación, el espacio muestral (S) en este experimento del lanzamiento de monedas será el siguiente:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

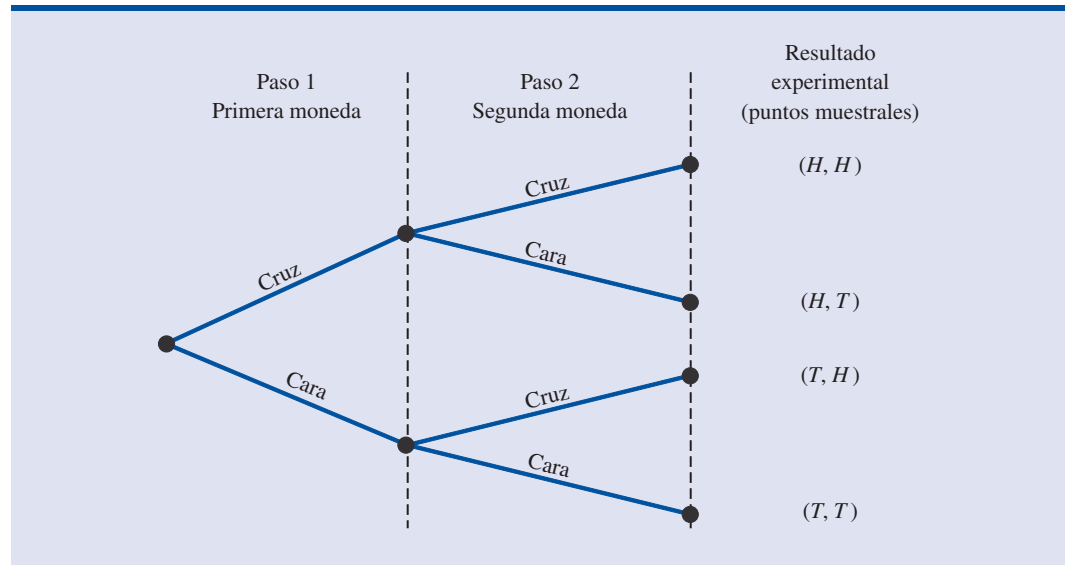
Por tanto, hay cuatro resultados experimentales. En este caso es fácil enumerar todos los resultados experimentales.

La regla de conteo para experimentos de pasos múltiples permite determinar el número de resultados experimentales sin tener que enumerarlos.

REGLA DE CONTEO PARA EXPERIMENTOS DE PASOS MÚLTIPLES

Un experimento se describe como una sucesión de k pasos en los que hay n_1 resultados posibles en el primer paso, n_2 resultados posibles en el segundo paso y así en lo sucesivo, entonces el número total de resultados experimentales es $(n_1)(n_2) \dots (n_k)$.

Si considera el experimento del lanzamiento de dos monedas como la sucesión de lanzar primero una moneda ($n_1 = 2$) y después lanzar la otra ($n_2 = 2$), siguiendo la regla de conteo $(2)(2) = 4$, entonces hay cuatro resultados distintos. Como ya se mostró, estos resultados son $S = \{(H, H), (H, T), (T, H), (T, T)\}$. El número de resultados experimentales de seis monedas es $(2)(2)(2)(2)(2)(2) = 64$.

FIGURA 4.2 DIAGRAMA DE ÁRBOL PARA EL LANZAMIENTO DE DOS MONEDAS

Sin el diagrama de árbol podría pensarse que sólo se pueden tener tres resultados experimentales en dos lanzamientos de una moneda: 0 caras, 1 cara y 2 caras.

Un **diagrama de árbol** es una representación gráfica que permite visualizar un experimento de pasos múltiples. En la figura 4.2 aparece un diagrama de árbol para el experimento del lanzamiento de dos monedas. La secuencia de los pasos en el diagrama va de izquierda a derecha. El paso 1 corresponde al lanzamiento de la primera moneda, el paso 2 al de la segunda moneda. En cada paso, los dos resultados posibles son cruz o cara. Observe que a cada uno de los resultados posibles en el paso 1 pertenecen dos ramas por los dos posibles resultados en el paso 2. Cada uno de los puntos en el extremo derecho del árbol representa un resultado experimental. Cada trayectoria a través del árbol, desde el nodo más a la izquierda hasta uno de los nodos en el extremo derecho del árbol, muestra una secuencia única de resultados.

Ahora una aplicación de la regla de conteo para experimentos de pasos múltiples en el análisis de un proyecto de expansión de la empresa Kentucky Power & Light (KP&L). Kentucky Power & Light ha empezado un proyecto que tiene como objetivo incrementar la capacidad de generación de una de sus plantas en el norte de Kentucky. El proyecto fue dividido en dos etapas o pasos sucesivos: etapa 1 (diseño) y etapa 2 (construcción). A pesar de que cada etapa se planeará y controlará con todo el cuidado posible, a los administrativos no les es posible pronosticar el tiempo exacto requerido en cada una de las etapas del proyecto. En un análisis de proyectos de construcción similares encuentran que la posible duración de la etapa de diseño es de 2, 3, o 4 meses y que la duración de la construcción es de 6, 7 u 8 meses. Además, debido a la necesidad urgente de más energía eléctrica, los administrativos han establecido como meta 10 meses para la terminación de todo el proyecto.

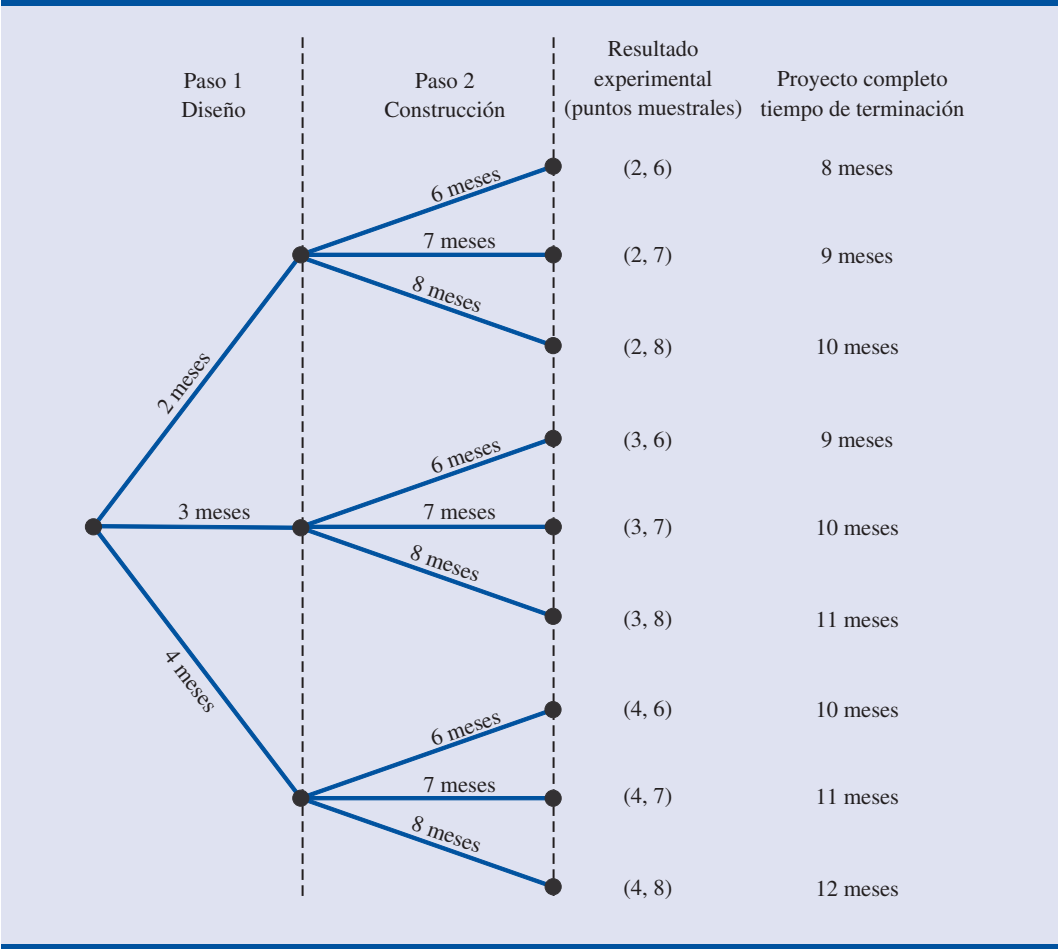
Como hay tres posibles periodos para la etapa del diseño (paso 1) y tres para la etapa de la construcción (paso 2) cabe aplicar la regla de conteo para experimentos de pasos múltiples, entonces el total de resultados posibles es $(3)(3) = 9$. Para describir los resultados experimentales emplean una notación de dos números; por ejemplo, (2, 6) significa que la etapa del diseño durará 2 meses y la etapa de la construcción 6. Esto da como resultado una duración de $2 + 6 = 8$ meses para todo el proyecto. En la tabla 4.1 aparecen los nueve resultados experimentales que hay para el problema de KP&L. El diagrama de árbol de la figura 4.3 muestra como se presentan los nueve resultados (puntos muestrales).

La regla de conteo y el diagrama de árbol ayudan al administrador del proyecto a identificar los resultados experimentales y a determinar la posible duración del proyecto. De acuerdo con la

TABLA 4.1 RESULTADOS EXPERIMENTALES (PUNTOS MUESTRALES) PARA EL PROYECTO KP&L

Duración (meses)			
Etapa 1 Diseño	Etapa 2 Construcción	Notación para los resultados experimentales	Proyecto completo: duración (meses)
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

FIGURA 4.3 DIAGRAMA DE ÁRBOL PARA EL PROYECTO KP&L



información de la figura 4.3, la duración del proyecto es de 8 a 12 meses, y seis de los nueve resultados experimentales tienen la duración deseada de 10 meses o menos. Aun cuando identificar los resultados experimentales ayuda, es necesario considerar cómo asignar los valores de probabilidad a los resultados experimentales antes de evaluar la probabilidad de que el proyecto dure los 10 meses deseados.

Combinaciones Otra regla de conteo útil le permite contar el número de resultados experimentales cuando el experimento consiste en seleccionar n objetos de un conjunto (usualmente mayor) de N objetos. Ésta es la regla de conteo para combinaciones.

REGLA DE CONTEO PARA COMBINACIONES

El número de combinaciones de N objetos tomados de n en n es

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \quad (4.1)$$

donde

$$N! = N(N-1)(N-2) \cdots (2)(1)$$

$$n! = n(n-1)(n-2) \cdots (2)(1)$$

y por definición, $0! = 1$

Cuando se hace un muestreo de una población finita de tamaño N , la regla de conteo para combinaciones sirve para hallar el número de muestras de tamaño n que pueden seleccionarse.

La notación $!$ significa *factorial*; por ejemplo, 5 factorial es $5! = (5)(4)(3)(2)(1) = 120$.

Como ejemplo del uso de la regla de conteo para combinaciones, considere un procedimiento de control de calidad en el que un inspector selecciona al azar dos de cinco piezas para probar que no tengan defectos. En un conjunto de cinco partes, ¿cuántas combinaciones de dos partes pueden seleccionarse? De acuerdo con la regla de conteo de la ecuación (4.1) es claro que con $N = 5$ y $n = 2$ se tiene

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{(5)(4)(3)(2)(1)}{(2)(1)(3)(2)(1)} = \frac{120}{12} = 10$$

De manera que hay 10 resultados posibles en este experimento de la selección aleatoria de dos partes de un conjunto de cinco. Si etiqueta dichas partes como A, B, C, D y E, las 10 combinaciones o resultados experimentales serán AB, AC, AD, AE, BC, BD, BE, CD, CE y DE.

Para ver otro ejemplo, considere la lotería de Florida en la que se seleccionan seis números de un conjunto de 53 números para determinar al ganador de la semana. Para establecer las distintas variables en la selección de seis enteros de un conjunto de 53, se usa la regla de conteo para combinaciones.

$$\binom{53}{6} = \frac{53!}{6!(53-6)!} = \frac{53!}{6!47!} = \frac{(53)(52)(51)(50)(49)(48)}{(6)(5)(4)(3)(2)(1)} = 22\,957\,480$$

La regla de conteo para combinaciones muestra que la probabilidad de ganar en esta lotería es muy pequeña.

La regla de conteo para combinaciones arroja casi 23 millones de resultados experimentales en esta lotería. Si una persona compra un billete de lotería, tiene una en 22 957 480 posibilidades de ganar la lotería.

Permutaciones La tercera regla de conteo que suele ser útil, es para permutaciones. Dicha regla permite calcular el número de resultados experimentales cuando se seleccionan n objetos de

un conjunto de N objetos y el orden de selección es relevante. Los mismos n objetos seleccionados en orden diferente se consideran un resultado experimental diferente.

REGLA DE CONTEO PARA PERMUTACIONES

El número de permutaciones de N objetos tomados de n en n está dado por

$$P_n^N = n! \binom{N}{n} = \frac{N!}{(N-n)!} \quad (4.2)$$

La regla de conteo para permutaciones tiene relación estrecha con la de combinaciones; sin embargo, con el mismo número de objetos, el número de permutaciones que se obtiene en un experimento es mayor que el número de combinaciones, ya que cada selección de n objetos se ordena de $n!$ maneras diferentes.

Para ver un ejemplo, reconsidere el proceso de control de calidad en el que un inspector selecciona dos de cinco piezas para probar que no tienen defectos. ¿Cuántas permutaciones puede seleccionar? La ecuación (4.2) indica que si $N = 5$ y $n = 2$, se tiene

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)} = \frac{120}{6} = 20$$

De manera que el experimento de seleccionar aleatoriamente dos piezas de un conjunto de cinco piezas, teniendo en cuenta el orden en que se seleccionen, tiene 20 resultados. Si las piezas se etiquetan A, B, C, D y E, las 20 permutaciones son AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE y ED.

Asignación de probabilidades

Ahora verá cómo asignar probabilidades a los resultados experimentales. Los tres métodos comúnmente usados son el método clásico, el método de la frecuencia relativa y el método subjetivo. Sin importar el método que se use, es necesario satisfacer los **requerimientos básicos para la asignación de probabilidades**.

REQUERIMIENTOS BÁSICOS PARA LA ASIGNACIÓN DE PROBABILIDADES

1. La probabilidad asignada a cada resultado experimental debe estar entre 0 y 1, inclusive. Si denota con E_i el i -ésimo resultado experimental y con $P(E_i)$ su probabilidad, entonces exprese este requerimiento como

$$0 \leq P(E_i) \leq 1 \text{ para toda } i \quad (4.3)$$

2. La suma de las probabilidades de los resultados experimentales debe ser igual a 1.0. Para resultados experimentales n escriba este requerimiento como

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \quad (4.4)$$

El **método clásico** de asignación de probabilidades es apropiado cuando todos los resultados experimentales tienen la misma posibilidad. Si existen n resultados experimentales, la probabilidad asignada a cada resultado experimental es $1/n$. Cuando emplee este método, satisfará en automático los dos requerimientos básicos de la asignación de probabilidades.

Por ejemplo, considere el experimento del lanzamiento de una moneda, los dos resultados experimentales —cruz o cara— tienen la misma posibilidad. Como uno de los dos resultados igualmente posibles es cara, la probabilidad de que caiga cara es $1/2$ o 0.50 . Asimismo, la probabilidad de que caiga cruz también es $1/2$ o 0.50 .

Otro ejemplo, considere el experimento de lanzar un dado. Es razonable pensar que los seis resultados que pueden presentarse son igualmente posibles y, por tanto, la probabilidad asignada a cada resultado es $1/6$. Si $P(1)$ denota la probabilidad de que la cara del dado que caiga hacia arriba sea la que tiene un punto, entonces $P(1) = 1/6$. De manera similar $P(2) = 1/6$, $P(3) = 1/6$, $P(4) = 1/6$, $P(5) = 1/6$ y $P(6) = 1/6$. Observe que dichas probabilidades satisfacen los dos requerimientos básicos de las ecuaciones (4.3) y (4.4), porque cada una es mayor o igual que cero y juntas suman 1.0 .

El **método de frecuencia relativa** para la asignación de probabilidades es el más conveniente cuando existen datos para estimar la proporción de veces que se presentarán los resultados si el experimento se repite muchas veces. Considere, por ejemplo un estudio sobre los tiempos de espera en el departamento de rayos x de un hospital pequeño. Durante 20 días sucesivos un empleado registra el número de personas que están esperando el servicio a las 9:00 a.m.; los resultados son los siguientes.

Número de personas que esperan	Número de días: resultados de ocurrencia
0	2
1	5
2	6
3	4
4	3
	<hr/>
	Total 20

En estos datos aparece que 2 de los 20 días, había cero pacientes esperando el servicio, 5 días había un paciente en espera y así sucesivamente. Con el método de la frecuencia relativa, la probabilidad que se le asignará al resultado experimental cero pacientes esperan el servicio, será $2/20 = 0.10$; al resultado experimental un paciente espera el servicio, $5/20 = 0.25$; $6/20 = 0.30$ a dos pacientes esperan el servicio; $4/20 = 0.20$ a tres pacientes esperan el servicio y $3/20 = 0.15$ a cuatro pacientes esperan el servicio. Como sucede con el método clásico, al usar el método de frecuencia relativa se satisfacen en automático los dos requerimientos básicos correspondientes a las ecuaciones (4.3) y (4.4).

El **método subjetivo** de asignación de probabilidades es el más indicado cuando no es factible suponer que todos los resultados de un experimento sean igualmente posibles y, además, cuenta con pocos datos relevantes. El método subjetivo de asignación de probabilidades a los resultados de un experimento, usa toda la información disponible, por ejemplo, la propia experiencia o la intuición. Después de considerar dicha información se asigna un valor de probabilidad que expresa el *grado de confianza* (en una escala de 0 a 1) que tiene acerca de que un resultado experimental ocurra. Como la probabilidad subjetiva expresa el grado de confianza que tiene un individuo, es personal. Cuando se usa el método de probabilidad subjetiva, es de esperarse que personas distintas asignen probabilidades diferentes a los mismos resultados de un experimento.

En el método subjetivo hay que tener cuidado de que se satisfagan los dos requerimientos básicos expresados en las ecuaciones (4.3) y (4.4). Sea cual sea el grado de confianza que tenga la persona, el valor de probabilidad asignado a cada resultado experimental debe estar entre 0 y 1, inclusive, y la suma de las probabilidades de todos los resultados experimentales debe ser 1.0 .

Considere el caso en el que Tom y Judy Elsbernd hacen una oferta para la compra de una casa. Hay dos resultados posibles:

E_1 = su oferta será aceptada

E_2 = su oferta no será aceptada

El teorema de Bayes (véase sección 4.5) proporciona un medio para combinar la probabilidad a priori determinada subjetivamente con probabilidades obtenidas por otros medios para obtener probabilidades a posteriori o revisadas.

Judy cree que la probabilidad de que su oferta sea aceptada es 0.8; por tanto, Judy establece que $P(E_1) = 0.8$ y $P(E_2) = 0.2$; Tom, por su parte, cree que la probabilidad de que su oferta sea aceptada es 0.6; por tanto, Tom establecerá $P(E_1) = 0.6$ y $P(E_2) = 0.4$. Observe que la estimación de probabilidad de E_1 que hace Tom refleja bastante pesimismo de que su oferta sea aceptada.

Tanto Judy como Tom asignaron probabilidades que satisfacen los dos requerimientos básicos. El hecho de que sus probabilidades sean diferentes subraya la naturaleza personal del método subjetivo.

Incluso en situaciones de negocios en que es posible emplear el método clásico o el de las probabilidades relativas, los administradores suelen proporcionar estimaciones subjetivas de una probabilidad. En tales casos, la mejor estimación de una probabilidad suele obtenerse combinando las estimaciones del método clásico o del método de las frecuencias relativas con las estimaciones subjetivas de una probabilidad.

Probabilidades para el proyecto KP&L

Para continuar con el análisis del proyecto KP&L hay que hallar las probabilidades de los nueve resultados experimentales enumerados en la tabla 4.1. De acuerdo con la experiencia, los administrativos concluyen que los resultados experimentales no son todos igualmente posibles. Por tanto, no emplean el método clásico de asignación de probabilidades. Entonces deciden hacer un estudio sobre la duración de los proyectos similares realizados por KP&L en los últimos tres años. En la tabla 4.2 se resume el resultado de este estudio considerando 40 proyectos similares.

Después de analizar los resultados de este estudio, los administrativos deciden emplear el método de frecuencia relativa para asignar las probabilidades. Los administrativos podrían haber aportado probabilidades subjetivas, pero se dieron cuenta de que el proyecto actual era muy similar a los 40 proyectos anteriores. Así, consideraron que el método de frecuencia relativa sería el mejor.

Si emplea la tabla 4.2 para calcular las probabilidades, observará que el resultado (2, 6) —duración de la etapa 1, 2 meses, y duración de la etapa 2, 6 meses— se encuentra seis veces en los 40 proyectos. Con el método de las frecuencias relativas, la probabilidad signada a este resultado es $6/40 = 0.15$. También el resultado (2, 7) se encuentra seis veces en los 40 proyectos $6/40 = 0.15$. Continuando de esta manera, se obtienen, para los puntos muestrales del proyecto de KP&L, las asignaciones de probabilidad que se muestran en la tabla 4.3. Observe que $P(2, 6)$ representa la probabilidad del punto muestral (2, 6), $P(2, 7)$ representa la probabilidad del punto muestral (2, 7) y así sucesivamente.

TABLA 4.2 DURACIÓN DE 40 PROYECTOS DE KP&L

Duración (meses)		Punto muestral	Número de proyectos que tuvieron esta duración
Etapla 1 Diseño	Etapla 2 Construcción		
2	6	(2, 6)	6
2	7	(2, 7)	6
2	8	(2, 8)	2
3	6	(3, 6)	4
3	7	(3, 7)	8
3	8	(3, 8)	2
4	6	(4, 6)	2
4	7	(4, 7)	4
4	8	(4, 8)	6
Total			40

TABLA 4.3 ASIGNACIÓN DE PROBABILIDADES PARA EL PROYECTO KP&L, EMPLEANDO EL MÉTODO DE LAS FRECUENCIAS RELATIVAS

Punto muestral	Tiempo de terminación del proyecto	Probabilidad del punto muestral
(2, 6)	8 meses	$P(2, 6) = 6/40 = 0.15$
(2, 7)	9 meses	$P(2, 7) = 6/40 = 0.15$
(2, 8)	10 meses	$P(2, 8) = 2/40 = 0.05$
(3, 6)	9 meses	$P(3, 6) = 4/40 = 0.10$
(3, 7)	10 meses	$P(3, 7) = 8/40 = 0.20$
(3, 8)	11 meses	$P(3, 8) = 2/40 = 0.05$
(4, 6)	10 meses	$P(4, 6) = 2/40 = 0.05$
(4, 7)	11 meses	$P(4, 7) = 4/40 = 0.10$
(4, 8)	12 meses	$P(4, 8) = 6/40 = 0.15$
	Total	1.00

NOTAS Y COMENTARIOS

1. En estadística la noción de experimento difiere un poco del concepto de experimento de las ciencias físicas. En las ciencias físicas, los investigadores suelen realizar los experimentos en laboratorios o en ambientes controlados, con objeto de investigar causas y efectos. En los experimentos estadísticos, la probabilidad determina los resultados. Aun cuando un experimento se repita con exactitud, el resultado puede ser completamente diferente. Debido a esta influencia que tiene la probabilidad sobre los resultados, a los experimentos en estadística también se les conoce como *experimentos aleatorios*.
2. Cuando de una población de tamaño N se extrae una muestra aleatoria sin reemplazarla, se emplea la regla de conteo para combinaciones para calcular la cantidad de muestras de tamaño n que pueden seleccionarse.

Ejercicios

Métodos

1. Un experimento consta de tres pasos; para el primer paso hay tres resultados posibles, para el segundo hay dos resultados posibles y para el tercer paso hay cuatro resultados posibles. ¿Cuántos resultados distintos hay para el experimento completo?
2. ¿De cuántas maneras es posible seleccionar tres objetos de un conjunto de seis objetos? Use las letras A, B, C, D, E y F para identificar a los objetos y enumere todas las combinaciones diferentes de tres objetos.
3. ¿Cuántas permutaciones de tres objetos se pueden seleccionar de un grupo de seis objetos? Use las letras A, B, C, D, E y F para identificar a los objetos y enumere cada una de las permutaciones factibles para los objetos B, D y F.
4. Considere el experimento de lanzar una moneda tres veces.
 - a. Elabore un diagrama de árbol de este experimento.
 - b. Enumere los resultados del experimento.
 - c. ¿Cuál es la probabilidad que le corresponde a cada uno de los resultados?
5. Suponga que un experimento tiene cinco resultados igualmente posibles: E_1, E_2, E_3, E_4 y E_5 . Asigne probabilidades a los resultados y muestre que satisfacen los requerimientos expresados por las ecuaciones (4.3) y (4.4). ¿Qué método empleó?
6. Un experimento que tiene tres resultados es repetido 50 veces y se ve que E_1 aparece 20 veces, E_2 13 veces y E_3 17 veces. Asigne probabilidades a los resultados. ¿Qué método empleó?

Autoexamen

Autoexamen

7. La persona que toma las decisiones asigna las probabilidades siguientes a los cuatro resultados de un experimento: $P(E_1) = 0.10$, $P(E_2) = 0.15$, $P(E_3) = 0.40$ y $P(E_4) = 0.20$. ¿Son válidas estas asignaciones de probabilidades? Argumente.

Aplicaciones

8. En una ciudad las solicitudes de cambio de uso de suelo pasan por un proceso de dos pasos: una revisión por la comisión de planeación y la decisión final tomada por el consejo de la ciudad. En el paso 1 la comisión de planeación revisa la solicitud de cambio de uso de suelo y hace una recomendación positiva o negativa respecto al cambio. En el paso 2 el consejo de la ciudad revisa la recomendación hecha por la comisión de planeación y vota para aprobar o desaprobar el cambio de suelo. Suponga que una empresa dedicada a la construcción de complejos departamentales presenta una solicitud de cambio de uso de suelo. Considere el proceso de la solicitud como un experimento. ¿Cuántos puntos muestrales tiene este experimento? Enumérellos. Construya el diagrama de árbol del experimento.
9. El muestreo aleatorio simple usa una muestra de tamaño n tomada de una población de tamaño N para obtener datos para hacer inferencias acerca de las características de la población. Suponga que, de una población de 50 cuentas bancarias, desea tomar una muestra de cuatro cuentas con objeto de tener información acerca de la población. ¿Cuántas muestras diferentes de cuatro cuentas pueden obtener?
10. El capital de riesgo es una fuerte ayuda para los fondos disponibles de las empresas. De acuerdo con Venture Economics (*Investor's Business Daily*, 28 de abril de 2000) de 2374 desembolsos en capital de riesgo, 1434 son de empresas en California, 390 de empresas en Massachussets, 217 de empresas en Nueva York y 112 de empresas en Colorado. Veintidós por ciento de las empresas que reciben fondos se encuentran en las etapas iniciales de desarrollo y 55% en la etapa de expansión. Suponga que desea tomar en forma aleatoria una de estas empresas para saber cómo son usados los fondos de capital de riesgo.
- ¿Cuál es la probabilidad de que la empresa que seleccione sea de California?
 - ¿De que la empresa no sea de ninguno de los estados citados?
 - ¿De que la empresa elegida no se encuentre en las etapas iniciales de desarrollo?
 - Si admite que las empresas en las etapas iniciales de desarrollo tuvieran una distribución homogénea en todo el país, ¿cuántas empresas de Massachussets que reciben fondos de capital de riesgo se encuentran en las etapas iniciales de desarrollo?
 - La cantidad total de fondos invertidos es \$32.4 mil millones. Estime la cantidad destinada a Colorado.
11. La National Highway Traffic Safety Administration (NHTSA) realizó una investigación para saber si los conductores de Estados Unidos están usando sus cinturones de seguridad (Associated Press, 25 de agosto de 2003). Los datos muestrales fueron los siguientes.

Autoexamen

Autoexamen

Conductores que emplean el cinturón

Región	Sí	No
Noreste	148	52
Oeste medio	162	54
Sur	296	74
Oeste	252	48
Total	858	228

- ¿Cuál es la probabilidad de que en Estados Unidos un conductor lleve puesto el cinturón?
- Un año antes, la probabilidad en Estados Unidos de que un conductor llevara puesto el cinturón era 0.75. El director de NHTSA, doctor Jeffrey Runge esperaba que en 2003 la probabilidad llegara a 0.78. ¿Estará satisfecho con los resultados del estudio del 2003?

- c. ¿Cuál es la probabilidad de que se use el cinturón en las distintas regiones del país? ¿En qué región se usa más el cinturón?
 - d. En la muestra, ¿qué proporción de los conductores provenía de cada región del país? ¿En qué región se seleccionaron más conductores? ¿Qué región viene en segundo lugar?
 - e. Si admite que en todas las regiones la cantidad de conductores es la misma, ¿ve usted alguna razón para que la probabilidad estimada en el inciso a sea tan alta? Explique.
12. En Estados Unidos hay una lotería que se juega dos veces por semana en 28 estados, en las Islas Vírgenes y en el Distrito de Columbia. Para jugar, debe comprar un billete y seleccionar cinco números del 1 al 55 y un número del 1 al 42. Para determinar al ganador se sacan 5 bolas blancas entre 55 bolas blancas y una bola roja entre 42 bolas rojas. Quien atine a los cinco números de bolas blancas y al número de la bola roja es el ganador. Ocho trabajadores de una empresa tienen el récord del mayor premio, ganaron \$365 millones al atinarle a los números 15-17-43-44-49 de las bolas blancas y al 29 de las bolas rojas. En cada juego hay también otros premios. Por ejemplo, quien atina a los cinco números de las bolas blancas se lleva un premio de \$200 000 (www.powerball.com, 19 de marzo de 2006).
- a. ¿De cuántas maneras se pueden seleccionar los primeros cinco números?
 - b. ¿Cuál es la probabilidad de ganar los \$200 000 atinándole a los cinco números de bolas blancas?
 - c. ¿Cuál es la probabilidad de atinarle a todos los números y ganar el premio mayor?
13. Una empresa que produce pasta de dientes está analizando el diseño de cinco empaques diferentes. Suponiendo que existe la misma posibilidad de que los clientes elijan cualquiera de los empaques, ¿cuál es la probabilidad de selección que se le asignaría a cada diseño de empaque? En un estudio, se pidió a 100 consumidores que escogieran el diseño que más les gustara. Los resultados se muestran en la tabla siguiente. ¿Confirman estos datos la creencia de que existe la misma posibilidad de que los clientes elijan cualquiera de los empaques? Explique

Diseño	Número de veces que fue elegido
1	5
2	15
3	30
4	40
5	10

4.2

Eventos y sus probabilidades

En la introducción de este capítulo el término *evento* fue aplicado tal como se usa en el lenguaje cotidiano. Después, en la sección 4.1 se presentó el concepto de experimento y de los correspondientes resultados experimentales o puntos muestrales. Puntos muestrales y eventos son la base para el estudio de la probabilidad. Por tanto, ahora se le presenta la definición formal de **evento** como se emplea en relación con los puntos muestrales. Con esto se tiene la base para poder dar probabilidades de eventos.

EVENTO

Un evento es una colección de puntos muestrales.

Para dar un ejemplo recuerde el proyecto de KP&L. Considere que al encargado del proyecto le interesa conocer la probabilidad de terminar el proyecto en 10 meses o menos. En la tabla 4.3 aparecen los puntos muestrales (2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6) correspondientes a una duración del proyecto de 10 meses o menos. C denota el evento de que el proyecto dura 10 meses o menos:

$$C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$$

Si cualquiera de estos puntos muestrales es el resultado experimental, entonces ocurre el evento C .

Otros eventos de posible interés para el administrador del proyecto KP&L son los siguientes:

L = El evento de que el proyecto esté acabado en *menos* de 10 meses

M = El evento de que el proyecto esté acabado en *más* de 10 meses

De acuerdo con la tabla 4.3 dichos eventos consisten de los siguientes puntos muestrales

$$L = \{(2, 6), (2, 7), (3, 6)\}$$

$$M = \{(3, 8), (4, 7), (4, 8)\}$$

Para el proyecto KP&L existen otros muchos eventos, pero todos serán una colección de puntos muestrales del experimento.

Dadas las probabilidades de los puntos muestrales que se presentan en la tabla 4.3, para calcular la probabilidad de cualquier evento que interese al administrador del proyecto KP&L, se emplea la definición siguiente.

PROBABILIDAD DE UN EVENTO

La probabilidad de cualquier evento es igual a la suma de las probabilidades de los puntos muestrales que forman el evento.

De acuerdo con esta definición, la probabilidad de un determinado evento se calcula sumando las probabilidades de los puntos muestrales (resultados experimentales) que forman el evento. Ahora es posible calcular la probabilidad de que el proyecto dure 10 meses o menos. Como este evento está dado por $C = \{(2, 6), (2, 7), (2, 8), (3, 6), (3, 7), (4, 6)\}$, la probabilidad del evento C denotada por $P(C)$ está dada por

$$P(C) = P(2, 6) + P(2, 7) + P(2, 8) + P(3, 6) + P(3, 7) + P(4, 6)$$

Al consultar las probabilidades de los puntos muestrales de la tabla 4.3, se tiene

$$P(C) = 0.15 + 0.15 + 0.05 + 0.10 + 0.20 + 0.05 = 0.70$$

Así, como el evento de que el proyecto dure menos de 10 meses está dado por $L = \{(2, 6), (2, 7), (3, 6)\}$, la probabilidad de este evento será

$$\begin{aligned} P(L) &= P(2, 6) + P(2, 7) + P(3, 6) \\ &= 0.15 + 0.15 + 0.10 = 0.40 \end{aligned}$$

Por último, el evento de que el proyecto dure más de 10 meses está dado por $M = \{(3, 8), (4, 7), (4, 8)\}$ y por tanto

$$\begin{aligned} P(M) &= P(3, 8) + P(4, 7) + P(4, 8) \\ &= 0.05 + 0.10 + 0.15 = 0.30 \end{aligned}$$

Con estas probabilidades, ahora puede informarle al administrador del proyecto KP&L las probabilidades siguientes: que el proyecto dure 10 meses o menos es 0.70; que dure menos de 10 meses es 0.40 y que dure más de 10 meses es 0.30. Este procedimiento para calcular las probabilidades de los eventos aplica para cualquier evento que interese al administrador del proyecto KP&L.

Siempre que se puedan identificar todos los puntos muestrales de un experimento y asignar a cada uno su probabilidad, es factible calcular la probabilidad de un evento usando la definición. Sin embargo, en muchos experimentos la gran cantidad de puntos muestrales hace en extremo difícil, si no imposible, la determinación de los puntos muestrales, así como la asignación de sus probabilidades correspondientes. En las secciones restantes de este capítulo se presentan algunas relaciones básicas de probabilidad útiles para calcular la probabilidad de un evento, sin necesidad de conocer las probabilidades de todos los puntos muestrales.

NOTAS Y COMENTARIOS

1. El espacio muestral S es un evento. Puesto que contiene todos los resultados experimentales, su probabilidad es 1; es decir $P(S) = 1$.
2. Cuando se usa el método clásico para asignar probabilidades, se parte de que todos los resultados experimentales son igualmente posibles.

En tales casos la probabilidad de un evento es calculable contando el número de resultados experimentales que hay en el evento y dividiendo el resultado entre el número total de resultados experimentales.

Ejercicios

Métodos

14. Para un experimento hay cuatro resultados que son igualmente posibles: E_1 , E_2 , E_3 y E_4 .
 - a. ¿Cuál es la probabilidad de que ocurra E_2 ?
 - b. ¿De que ocurra cualquiera de dos resultados (por ejemplo, E_1 o E_2)?
 - c. ¿De que ocurran tres de estos resultados (E_1 o E_2 o E_4)?
15. Considere el experimento de seleccionar un naipe de una baraja con 52 naipes. Cada naipe es un punto muestral y su probabilidad es $1/52$.
 - a. Enumere los puntos muestrales del evento si selecciona un as.
 - b. Enumere los puntos muestrales del evento si selecciona un trébol.
 - c. Enumere los puntos muestrales del evento si selecciona una figura (sota, rey o reina).
 - d. Halle la probabilidad correspondiente a cada uno de los eventos de los incisos a, b y c.
16. Considere el experimento que consiste en lanzar un par de dados. Suponga que lo relevante es la suma de los puntos en las dos caras que caen hacia arriba.
 - a. ¿Cuántos puntos muestrales habrá? (*Sugerencia:* Use la regla de conteo para experimentos de pasos múltiples.)
 - b. Enumere los puntos muestrales.
 - c. ¿Cuál es la probabilidad de obtener un 7?
 - d. ¿De obtener un 9 o un número mayor?
 - e. Como en cada lanzamiento son factibles seis valores pares (2, 4, 6, 8, 10, y 12) y sólo cinco impares (3, 5, 7, 9 y 11), se tendrán más veces resultados pares que impares. ¿Está de acuerdo? Explique.
 - f. ¿Qué método usó para calcular las probabilidades pedidas?

Autoexamen

Aplicaciones

17. Consulte las tablas 4.2 y 4.3 que muestran los puntos muestrales del proyecto KP&L y sus probabilidades.
 - a. La etapa del diseño (etapa 1) saldrá del presupuesto si su duración es mayor a 4 meses. Liste los puntos muestrales del evento si la etapa del diseño sale del presupuesto.
 - b. ¿Cuál es la probabilidad de que la etapa del diseño salga del presupuesto?
 - c. La etapa de la construcción (etapa 2) saldrá del presupuesto si su duración es mayor a 8 meses. Enumere los puntos muestrales del evento si la etapa de construcción sale del presupuesto.
 - d. ¿Cuál es la probabilidad de que la etapa de construcción salga del presupuesto?
 - e. ¿Cuál es la probabilidad de que las dos etapas salgan del presupuesto?
18. Suponga que el administrador de un complejo grande de departamentos proporciona la siguiente estimación de probabilidades subjetivas acerca del número de departamentos libres que habrá el mes próximo.

Departamentos libres	Probabilidad
0	0.05
1	0.15
2	0.35
3	0.25
4	0.10
5	0.10

Dé la probabilidad de cada uno de los eventos siguientes.

- a. No haya departamentos libres.
 - b. Haya por lo menos 4 departamentos libres.
 - c. Haya 2 o menos departamentos libres.
19. Una asociación deportiva realiza un sondeo entre las personas mayores a 6 años respecto de su participación en actividades deportivas. (*Statistical Abstract of the United States: 2002*). El total de la población de estas edades fue 248.5 millones, de los cuales 120.9 millones eran hombres y 127.6 millones mujeres. A continuación se presenta el número de participantes en los cinco deportes principales.

Actividad	Participantes (en millones)	
	Hombres	Mujeres
Andar en bicicleta	22.2	21.0
Acampar	25.6	24.3
Caminar	28.7	57.7
Hacer ejercicio con aparatos	20.4	24.4
Nadar	26.4	34.4

- a. Estime la probabilidad de que una mujer, elegida al azar, participe en cada una de estas actividades deportivas.
- b. Estime la probabilidad de que un hombre, elegido en forma aleatoria, participe en cada una de estas actividades deportivas.
- c. Estime la probabilidad de que una persona, elegida en forma aleatoria, haga ejercicio caminando.
- d. Suponga que acaba de ver una persona que pasa caminando para hacer ejercicio. ¿Cuál es la probabilidad de que sea mujer?, ¿de que sea hombre?

20. La revista *Fortune* publica anualmente una lista de las 500 empresas más grandes de Estados Unidos. A continuación se presentan los cinco estados en los que hay más de estas 500 empresas de *Fortune*.

Estado	Número de empresas
Nueva York	54
California	52
Texas	48
Illinois	33
Ohio	30

Suponga que se elige una de las 500 empresas de *Fortune*. ¿Cuál es la probabilidad de cada uno de los eventos siguientes?

- Sea N el evento: la empresa se encuentra en Nueva York. Halle $P(N)$.
 - Sea T el evento: la empresa se encuentra en Texas. Halle $P(T)$.
 - Sea B el evento: la empresa se encuentra en uno de estos cinco estados. Halle $P(B)$.
21. En la tabla siguiente se dan las edades de la población de Estados Unidos (*The World Almanac 2004*). Los datos aparecen en millones de personas.

Edad	Cantidad
19 y menos	80.5
20 a 24	19.0
25 a 34	39.9
35 a 44	45.2
45 a 54	37.7
55 a 64	24.3
65 y más	35.0

Suponga una selección aleatoria de una persona de esta población.

- ¿Cuál es la probabilidad de que la persona tenga entre 20 y 24 años?
- ¿De que la persona tenga entre 20 y 34 años?
- ¿De que tenga 45 años o más?

4.3

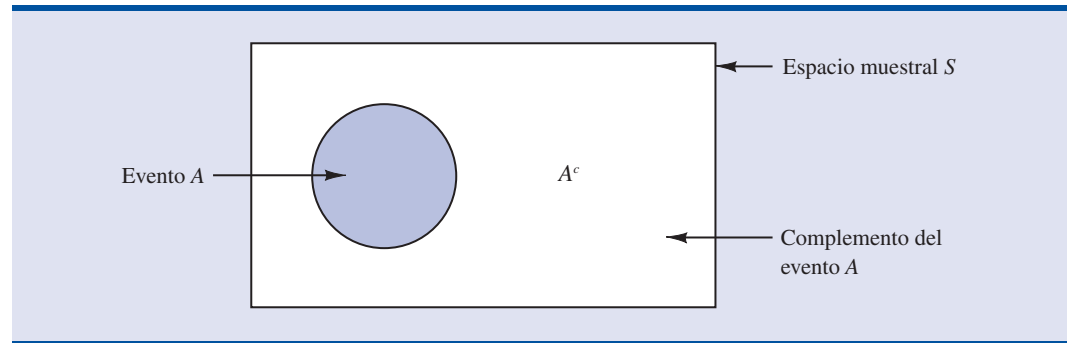
Algunas relaciones básicas de probabilidad

Complemento de un evento

Dado un evento A , el **complemento de A** se define como el evento que consta de todos los puntos muestrales que *no* están en A . El complemento de A se denota A^c . Al diagrama de la figura 4.4 se le llama **diagrama de Venn** e ilustra el concepto del complemento. El área rectangular representa el espacio muestral del experimento y, por tanto, contiene todos los puntos muestrales. El círculo representa el evento A y encierra sólo los puntos muestrales que pertenecen a A . La región del rectángulo que aparece sombreada incluye todos los puntos muestrales que no están en el evento A y es, por definición, el complemento de A .

En cualquier aplicación de la probabilidad ocurre un evento A o su complemento A^c . Por tanto,

$$P(A) + P(A^c) = 1$$

FIGURA 4.4 EL COMPLEMENTO DEL EVENTO A ES EL ÁREA QUE APARECE SOMBREADA

Despejando $P(A)$, obtiene lo siguiente.

CÁLCULO DE UNA PROBABILIDAD USANDO EL COMPLEMENTO

$$P(A) = 1 - P(A^c) \quad (4.5)$$

La ecuación (4.5) indica que la probabilidad de un evento A se puede calcular si se conoce la probabilidad de su complemento, $P(A^c)$.

Por ejemplo, considere el caso de un administrador de ventas que, después de revisar los informes de ventas, encuentra que 80% de los contactos con clientes nuevos no producen ninguna venta. Si A denota el evento hubo venta y A^c el evento no hubo venta, el administrador tiene que $P(A^c) = 0.80$. Mediante la ecuación (4.5) se ve que

$$P(A) = 1 - P(A^c) = 1 - 0.80 = 0.20$$

La conclusión es que la probabilidad de una venta en el contacto con un cliente nuevo es 0.20.

Otro ejemplo, un gerente de compras encuentra que la probabilidad de que el proveedor surta un pedido sin piezas defectuosas es 0.90, empleando el complemento podemos concluir que la probabilidad de que el pedido contenga piezas defectuosas es de $1 - 0.90 = 0.10$.

Ley de la adición

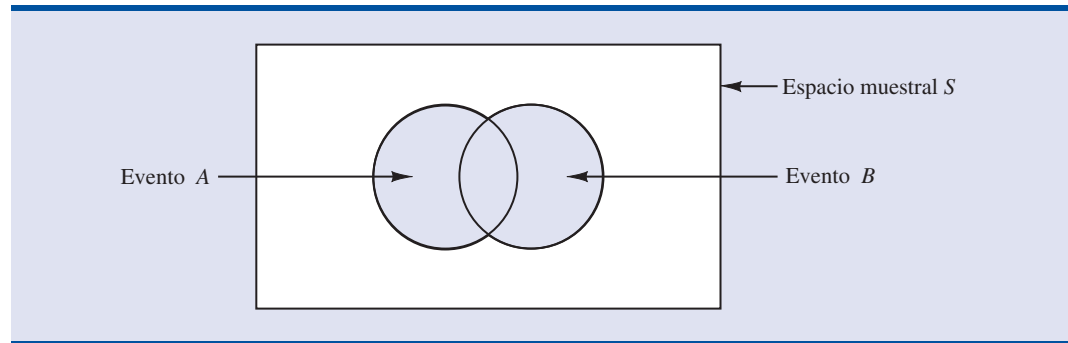
La ley de la adición sirve para determinar la probabilidad de que ocurra por lo menos uno de dos eventos. Es decir, si A y B son eventos, nos interesa hallar la probabilidad de que ocurra el evento A o el B o ambos.

Antes de presentar la ley de la adición es necesario ver dos conceptos relacionados con la combinación de eventos: la *unión* y la *intersección* de eventos. Dados dos eventos, A y B , la **unión de A y B** se define.

UNIÓN DE DOS EVENTOS

La unión de A y B es el evento que contiene todos los puntos muestrales que pertenecen a A o a B o a ambos. La unión se denota $A \cup B$.

El diagrama de Venn de la figura 4.5 representa la unión de los eventos A y B . Observe que en los dos círculos están contenidos todos los puntos muestrales del evento A y todos los puntos

FIGURA 4.5 LA UNIÓN DE LOS EVENTOS A Y B APARECE SOMBREADA

muestrales del evento B . El que los círculos se traslapen indica que algunos puntos muestrales están contenidos tanto en A como en B .

A continuación la definición de la **intersección de A y B** :

INTERSECCIÓN DE DOS EVENTOS

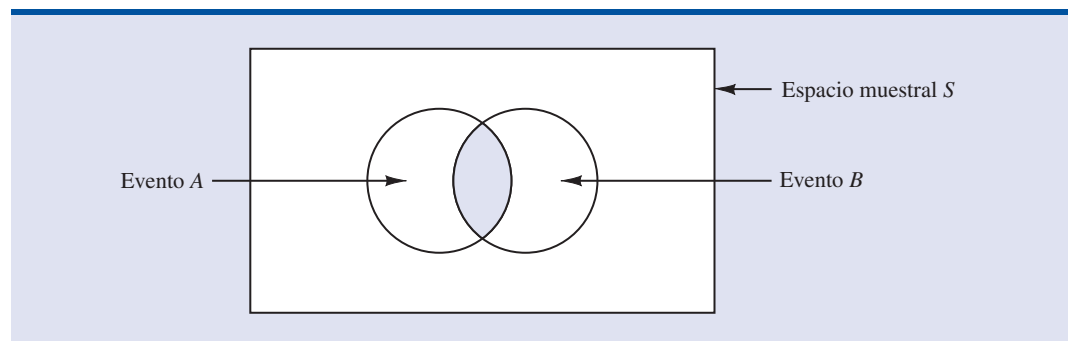
Dados dos eventos A y B , la intersección de A y B es el evento que contiene los puntos muestrales que pertenecen tanto a A como a B .

El diagrama de Venn ilustra la intersección de los eventos A y B mostrados en la figura 4.6. El área donde los círculos se sobreponen es la intersección que contiene una muestra de los puntos que están tanto en A como en B .

Ahora ya puede continuar con la ley de la adición. La **ley de la adición** proporciona una manera de calcular la probabilidad de que ocurra el evento A o el evento B o ambos. En otras palabras, la ley de la adición se emplea para calcular la probabilidad de la unión de los dos eventos. La ley de la adición se expresa.

LEY DE LA ADICIÓN

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.6)$$

FIGURA 4.6 LA INTERSECCIÓN DE LOS EVENTOS A Y B APARECE SOMBREADA

Para que logre un entendimiento intuitivo de la ley de la adición, observe que en la ley de la adición, los dos primeros términos $P(A) + P(B)$, corresponden a los puntos muestrales en $A \cup B$. Pero, como los puntos muestrales que se encuentran en la intersección $A \cap B$ están tanto en A como en B , cuando se calcula $P(A) + P(B)$, los puntos que se encuentran en $A \cap B$ cuentan dos veces. Esto se corrige restando $P(A \cap B)$.

Para ver un ejemplo de la aplicación de la ley de la adición, considere el caso de una pequeña empresa de ensamble en la que hay 50 empleados. Se espera que todos los trabajadores terminen su trabajo a tiempo y que pase la inspección final. A veces, alguno de los empleados no satisface el estándar de desempeño, ya sea porque no termina a tiempo su trabajo o porque no ensambla bien una pieza. Al final del periodo de evaluación del desempeño, el jefe de producción encuentra que 5 de los 50 trabajadores no terminaron su trabajo a tiempo, 6 de los 50 trabajadores ensamblaron mal una pieza y 2 de los 50 trabajadores no terminaron su trabajo a tiempo y armaron mal una pieza.

Sea

L = el evento no se terminó el trabajo a tiempo

D = el evento se armó mal la pieza

La información de las frecuencias relativas lleva a las probabilidades siguientes.

$$P(L) = \frac{5}{50} = 0.10$$

$$P(D) = \frac{6}{50} = 0.12$$

$$P(L \cap D) = \frac{2}{50} = 0.04$$

Después de analizar los datos del desempeño, el jefe de producción decide dar una calificación baja al desempeño de los trabajadores que no terminaron a tiempo su trabajo o que armaron mal alguna pieza; por tanto, el evento de interés es $L \cup D$. ¿Cuál es la probabilidad de que el jefe de producción dé a un trabajador una calificación baja de desempeño?

Observe que esta pregunta sobre probabilidad se refiere a la unión de dos eventos. En concreto, se desea hallar $P(L \cup D)$, usando la ecuación (4.6) se tiene

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

Como conoce las tres probabilidades del lado derecho de esta expresión, se tiene

$$P(L \cup D) = 0.10 + 0.12 - 0.04 = 0.18$$

Estos cálculos indican que la probabilidad de que un empleado elegido al azar obtenga una calificación baja por su desempeño es 0.18

Para ver otro ejemplo de la ley de la adición, considere un estudio reciente efectuado por el director de personal de una empresa importante de software. En el estudio encontró que 30% de los empleados que se van de la empresa antes de dos años, lo hacen por estar insatisfechos con el salario, 20% se van de la empresa por estar descontentos con el trabajo y 12% por estar insatisfechos con las *dos* cosas, el salario y el trabajo. ¿Cuál es la probabilidad de que un empleado

que se vaya de la empresa en menos de dos años lo haga por estar insatisfecho con el salario, con el trabajo o con las dos cosas?

Sea

S = el evento el empleado se va de la empresa por insatisfacción con el salario

W = el evento el empleado se va de la empresa por insatisfacción con el trabajo

Se tiene $P(S) = 0.30$, $P(W) = 0.20$ y $P(S \cap W) = 0.12$. Al aplicar la ecuación (4.6), de la ley de la adición, se tiene

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.30 + 0.20 - 0.12 = 0.38.$$

Así, la probabilidad de que un empleado se vaya de la empresa por el salario o por el trabajo es 0.38.

Antes de concluir el estudio de la ley de la adición se considerará un caso especial que surge cuando los **eventos son mutuamente excluyentes**.

EVENTOS MUTUAMENTE EXCLUYENTES

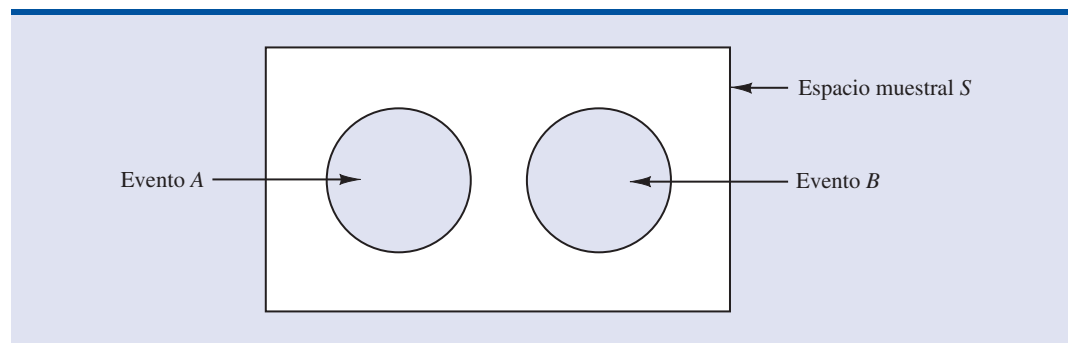
Se dice que dos eventos son mutuamente excluyentes si no tienen puntos muestrales en común.

Los eventos A y B son mutuamente excluyentes si, cuando un evento ocurre, el otro no puede ocurrir. Por tanto, para que A y B sean mutuamente excluyentes, se requiere que su intersección no contenga ningún punto muestral. En la figura 4.7 aparece el diagrama de Venn que representa dos eventos, A y B , mutuamente excluyentes. En este caso $P(A \cap B) = 0$ y la ley de la adición se expresa como sigue:

LEY DE LA ADICIÓN PARA EVENTOS MUTUAMENTE EXCLUYENTES

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

FIGURA 4.7 EVENTOS MUTUAMENTE EXCLUYENTES



Ejercicios

Métodos

22. Suponga que tiene un espacio muestral con cinco resultados experimentales que son igualmente posibles: E_1, E_2, E_3, E_4 y E_5 . Sean

$$A = \{E_1, E_2\}$$

$$B = \{E_3, E_4\}$$

$$C = \{E_2, E_3, E_5\}$$

- Halle $P(A)$, $P(B)$ y $P(C)$.
- Calcule $P(A \cup B)$. ¿ A y B son mutuamente excluyentes?
- Estime A^c , C^c , $P(A^c)$ y $P(C^c)$.
- Halle $A \cup B^c$ y $P(A \cup B^c)$.
- Halle $P(B \cup C)$.

23. Suponga que se tiene el espacio muestral $S = \{E_1, E_2, E_3, E_4, E_5, E_6, E_7\}$, donde E_1, E_2, \dots, E_7 denotan puntos muestrales. La asignación de probabilidades es la siguiente: $P(E_1) = 0.05$, $P(E_2) = 0.20$, $P(E_3) = 0.20$, $P(E_4) = 0.25$, $P(E_5) = 0.15$, $P(E_6) = 0.10$ y $P(E_7) = 0.05$. Sea

$$A = \{E_1, E_4, E_6\}$$

$$B = \{E_2, E_4, E_7\}$$

$$C = \{E_2, E_3, E_5, E_7\}$$

- Halle $P(A)$, $P(B)$ y $P(C)$.
- Encuentre $A \cup B$ y $P(A \cup B)$.
- Halle $A \cap B$ y $P(A \cap B)$.
- ¿Los eventos A y B son mutuamente excluyentes?
- Halle B^c y $P(B^c)$.

Autoexamen

Aplicaciones

24. Las autoridades de Clarkson University realizaron un sondeo entre sus alumnos para conocer su opinión acerca de su universidad. Una pregunta fue si la universidad no satisface sus expectativas, si las satisface o si supera sus expectativas. Encontraron que 4% de los interrogados no dieron una respuesta, 26% respondieron que la universidad no llenaba sus expectativas y 56% indicó que la universidad superaba sus expectativas.
- Si toma un alumno al azar, ¿cuál es la probabilidad de que diga que la universidad supera sus expectativas?
 - Si toma un alumno al azar, ¿cuál es la probabilidad de que diga que la universidad satisface o supera sus expectativas?
25. La Oficina de Censos de Estados Unidos cuenta con datos sobre la cantidad de adultos jóvenes, entre 18 y 24 años, que viven en casa de sus padres.* Sea

M = el evento adulto joven que vive en casa de sus padres

F = el evento adulta joven que vive en casa de sus padres

Si toma al azar un adulto joven y una adulta joven, los datos de dicha oficina permiten concluir que $P(M) = 0.56$ y $P(F) = 0.42$ (*The World Almanac*, 2006). La probabilidad de que ambos vivan en casa de sus padres es 0.24.

- ¿Cuál es la probabilidad de que al menos uno de dos adultos jóvenes seleccionados viva en casa de sus padres?
- ¿Cuál es la probabilidad de que los dos adultos jóvenes seleccionados vivan en casa de sus padres?

*En estos datos se incluye a los adultos jóvenes solteros que viven en los internados de las universidades, porque es de suponer que estos adultos jóvenes vuelven a las casas de sus padres en las vacaciones.

26. Datos sobre las 30 principales acciones y fondos balanceados proporcionan los rendimientos porcentuales anuales y a 5 años para el periodo que termina el 31 de marzo de 2000 (*The Wall Street Journal*, 10 de abril de 2000). Suponga que considera altos un rendimiento anual arriba de 50% y un rendimiento a cinco años arriba de 300%. Nueve de los fondos tienen un rendimiento anual arriba de 50%, siete de los fondos a cinco años lo tienen arriba de 300% y cinco de los fondos tienen tanto un rendimiento anual arriba de 50% como un rendimiento a cinco años arriba de 300%.
- ¿Cuál es la probabilidad de un rendimiento anual alto y cuál es la probabilidad de un rendimiento a cinco años alto?
 - ¿Cuál es la probabilidad de ambos, un rendimiento anual alto y un rendimiento a cinco años alto?
 - ¿Cuál es la probabilidad de que no haya un rendimiento anual alto ni un rendimiento a cinco años alto?
27. En una encuesta en la pretemporada de fútbol americano de la NCAA 2001 se preguntó: “¿Este año habrá un equipo del Big Ten o del Pac-10 en el juego del Rose Bowl?” De los 13 429 interrogados, 2961 dijeron que habría uno del Big Ten, 4494 señalaron que habría uno del Pac-10 y 6823 expresaron que ni el Big Ten ni el Pac-10 tendría un equipo en el Rose Bowl (www.yahoo.com, 30 de agosto de 2001).
- ¿Cuál es la probabilidad de que el interrogado responda que ni el Big Ten ni el Pac-10 tendrán un equipo en el Rose Bowl?
 - ¿De que afirme que el Big Ten o el Pac-10 tendrán un equipo en el campeonato Rose Bowl?
 - Halle la probabilidad de que la respuesta sea que tanto el Big Ten como el Pac-10 tendrán un equipo en el Rose Bowl.
28. En una encuesta aplicada a los suscriptores de una revista se encontró que en los últimos 12 meses 45.8% habían rentado un automóvil por razones de trabajo, 54% por razones personales y 30% por razones de trabajo y personales.
- ¿Cuál es la probabilidad de que un suscriptor haya rentado un automóvil en los últimos 12 meses por razones de trabajo o por razones personales?
 - ¿Cuál es la probabilidad de que un suscriptor no haya rentado un automóvil en los últimos 12 meses ni por razones de trabajo ni por razones personales?
29. En Estados Unidos cada año hay más estudiantes con buenas calificaciones que desean inscribirse a las mejores universidades del país. Como el número de lugares permanece relativamente estable, algunas universidades rechazan solicitudes de admisión anticipadas. La universidad de Pensilvania recibió 2851 solicitudes para admisión anticipada. De éstas admitió a 1033 estudiantes, rechazó definitivamente a 854 estudiantes y dejó a 964 para el plazo de admisión normal. Esta universidad admitió a cerca de 18% de los solicitantes en el plazo normal para hacer un total (número de admisiones anticipadas más número de admisiones normales) de 2375 estudiantes (*USA Today* 24 de enero de 2001). Sean los eventos: E , un estudiante que solicita admisión anticipada es admitido; R rechazado definitivamente y D dejado para el plazo normal de admisión, sea A el evento de que un estudiante es admitido en el plazo normal.
- Use los datos para estimar $P(E)$, $P(R)$ y $P(D)$.
 - ¿Son mutuamente excluyentes los eventos E y D ? Halle $P(E \cap D)$.
 - De los 2375 estudiantes admitidos en esta universidad, ¿cuál es la probabilidad de que un estudiante tomado en forma aleatoria haya tenido una admisión anticipada.
 - Suponga que un estudiante solicita admisión anticipada en esta universidad. ¿Cuál es la probabilidad de que el estudiante tenga una admisión anticipada o en el periodo normal de admisión?

Autoexamen

4.4

Probabilidad condicional

Con frecuencia, en la probabilidad de un evento influye el hecho de que un evento relacionado con él ya haya ocurrido. Suponga que tiene un evento A cuya probabilidad es $P(A)$. Si obtiene información nueva y sabe que un evento relacionado con él, denotado por B , ya ha ocurrido, de-

seará aprovechar esta información y volver a calcular la probabilidad del evento A . A esta nueva probabilidad del evento A se le conoce como **probabilidad condicional** y se expresa $P(A \mid B)$. La notación \mid indica que se está considerando la probabilidad del evento A *dada* la condición de que el evento B ha ocurrido. Por tanto, la notación $P(A \mid B)$ se lee “la probabilidad de A dado B ”.

Como ejemplo de la probabilidad condicional, considere el caso de las promociones de los agentes de policía de una determinada ciudad. La fuerza policiaca consta de 1200 agentes, 960 hombres y 240 mujeres. De éstos, en los últimos dos años, fueron promovidos 340. En la tabla 4.4 se muestra cómo quedaron repartidas estas promociones entre los hombres y mujeres.

Después de analizar el registro de las promociones, un comité feminil protestó, ya que habían sido promovidos 288 agentes hombres, frente a sólo 36 mujeres. Los directivos de la fuerza policiaca argumentaron que el número de mujeres promovidas no se debía a una discriminación, sino a que el número de mujeres que son agentes de policía es una cantidad pequeña. Ahora verá cómo emplear la probabilidad condicional para analizar esta acusación de discriminación.

Sean

- M = el evento que un agente de policía sea hombre
- W = el evento que un agente de policía sea mujer
- A = el evento que un agente de policía sea promovido
- A^c = el evento que un agente de policía no sea promovido

Dividir los valores de los datos de la tabla 4.4 entre el total de agentes de policía, 1200, permite concretar la información que se tiene en las probabilidades siguientes.

$$P(M \cap A) = 288/1200 = 0.24 =$$

probabilidad de que un agente de policía, escogido en forma aleatoria, sea hombre y haya sido promovido

$$P(M \cap A^c) = 672/1200 = 0.56 =$$

probabilidad de que un agente de policía, escogido en forma aleatoria, sea hombre y no haya sido promovido

$$P(W \cap A) = 36/1200 = 0.03 =$$

probabilidad de que un agente de policía, escogido en forma aleatoria, sea mujer y haya sido promovido

$$P(W \cap A^c) = 204/1200 = 0.17 =$$

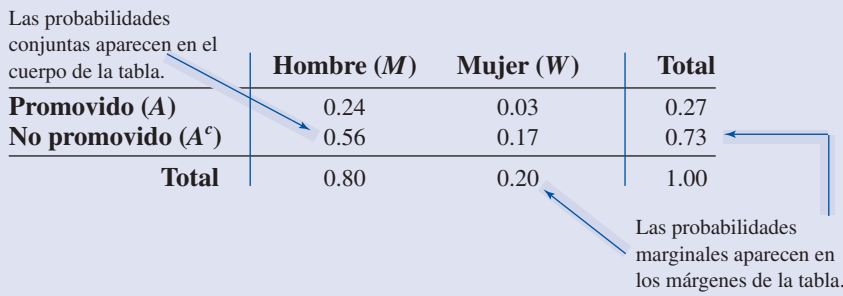
probabilidad de que un agente de policía, escogido en forma aleatoria, sea mujer y no haya sido promovido

Como cada uno de estos valores da la probabilidad de la intersección de dos eventos, se les llama **probabilidades conjuntas**. A la tabla 4.5, que proporciona la información de las probabilidades de promoción de los agentes de policía, se le conoce como *tabla de probabilidades conjuntas*.

Las cantidades que aparecen en los márgenes de una tabla de las probabilidades conjuntas son las probabilidades de cada uno de los eventos por separado. Es decir, $P(M) = 0.80$, $P(W) =$

TABLA 4.4 PROMOCIONES, EN LOS ÚLTIMOS DOS AÑOS, DE LOS AGENTES DE POLICÍA

	Hombre	Mujer	Total
Promovido	288	36	324
No promovido	672	204	876
Total	960	240	1200

TABLA 4.5 TABLA DE PROBABILIDAD CONJUNTA PARA LAS PROMOCIONES


Las probabilidades conjuntas aparecen en el cuerpo de la tabla.

	Hombre (M)	Mujer (W)	Total
Promovido (A)	0.24	0.03	0.27
No promovido (A^c)	0.56	0.17	0.73
Total	0.80	0.20	1.00

Las probabilidades marginales aparecen en los márgenes de la tabla.

0.20, $P(A) = 0.27$, $P(A^c) = 0.73$. A estas probabilidades se les conoce como **probabilidades marginales** por encontrarse en los márgenes de una tabla de probabilidad conjunta.

Observe que las probabilidades marginales se obtienen al sumar las probabilidades conjuntas del renglón o columna correspondiente de la tabla de probabilidades conjuntas. Por ejemplo, la probabilidad marginal de ser promovido es $P(A) = P(M \cap A) + P(W \cap A) = 0.24 + 0.03 = 0.27$. En las probabilidades marginales se observa que 80% de la fuerza policiaca está formada por hombres y 20% por mujeres, que 27% de los agentes de policía fueron promovidos y 73% no fueron promovidos.

Ahora empiece con el análisis de la probabilidad condicional calculando la probabilidad de que un agente de policía sea promovido dado que ese agente sea hombre. Emplee la notación para probabilidad condicional para determinar $P(A | M)$. Para calcular $P(A | M)$ se observa, primero, que esta notación sólo significa que se considera la probabilidad del evento A (promoción) ya que la condición designada como evento M (que el agente de policía sea hombre) está dada. Así que $P(A | M)$ indica que sólo interesan los promovidos dentro de los 960 agentes de policía que son hombres. Como 288 de los 960 agentes de policía que son hombres fueron promovidos, la probabilidad de ser promovido dado que se es un agente hombre es $288/960 = 0.30$. En otras palabras, puesto que un agente de policía es hombre, ese agente tuvo 30% de probabilidades de ser promovido en los dos últimos años.

Resultó fácil aplicar este procedimiento, ya que en la tabla 4.4 se muestra el número de agentes de policía en cada categoría. Ahora es interesante mostrar cómo calcular probabilidades condicionales, como $P(A | M)$, a partir de las probabilidades de eventos relacionados y no a partir de los datos de frecuencias de la tabla 4.4.

Entonces, $P(A | M) = 288/960 = 0.30$. Ahora, tanto el numerador como el denominador de esta fracción se dividen entre 1200, cantidad total de agentes de policía en el estudio.

$$P(A | M) = \frac{288}{960} = \frac{288/1200}{960/1200} = \frac{0.24}{0.80} = 0.30$$

Observe que la probabilidad condicional se obtiene de $0.24/0.80$. Regrese a la tabla de probabilidad conjunta (tabla 4.5) y observe que 0.24 es la probabilidad conjunta de A y M ; es decir, $P(A \cap M) = 0.24$; también que 0.80 es la probabilidad marginal de que un agente de la policía seleccionado aleatoriamente sea hombre. Es decir, $P(M) = 0.80$. Por tanto, la probabilidad condicional $P(A | M)$ se calcula como la razón entre $P(A \cap M)$ y la probabilidad marginal $P(M)$.

$$P(A | M) = \frac{P(A \cap M)}{P(M)} = \frac{0.24}{0.80} = 0.30$$

El hecho de que la probabilidad condicional se pueda calcular como la razón entre una probabilidad conjunta respecto a una probabilidad marginal proporciona la siguiente fórmula para el cálculo de la probabilidad condicional de dos eventos A y B .

PROBABILIDAD CONDICIONAL

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (4.7)$$

o

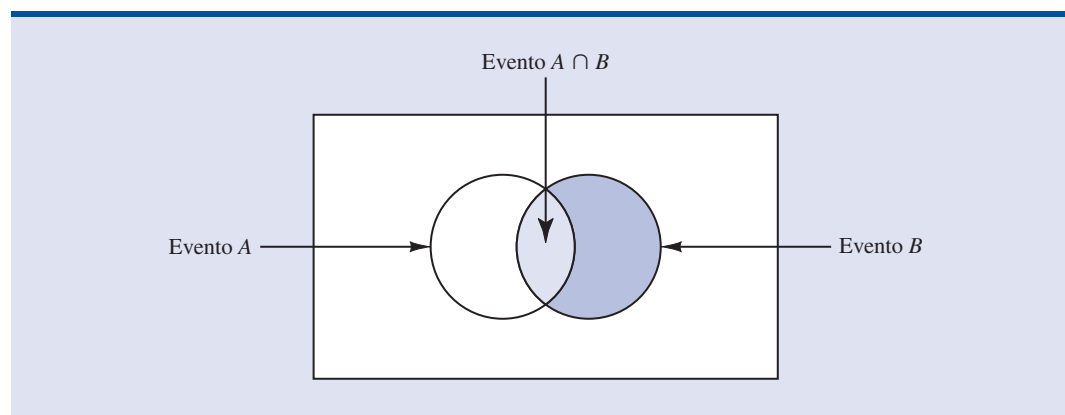
$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad (4.8)$$

El diagrama de Venn de la figura 4.8 ayuda a lograr una comprensión intuitiva de la probabilidad condicional. El círculo de la derecha muestra que el evento B ha ocurrido, la parte del círculo que se superpone con el evento A se denota $(A \cap B)$. Una vez que el evento B ha ocurrido, la única manera de que también sea observable el evento A es que ocurra el evento $(A \cap B)$. De manera que la razón $P(A \cap B)/P(B)$ aporta la probabilidad condicional de que se observe el evento A dado que el evento B ya ha ocurrido.

Ahora, considere de nuevo el asunto de la discriminación contra las mujeres agentes de policía. La probabilidad marginal del renglón 1 de la tabla 4.5 indica que la probabilidad de que un agente de la policía sea promovido (ya sea hombre o mujer) es $P(A) = 0.27$. Sin embargo, la cuestión relevante en el caso de la discriminación tiene que ver con las probabilidades condicionales $P(A | M)$ y $P(A | W)$. Es decir, ¿cuál es la probabilidad de que un agente de la policía sea promovido *dado que* es hombre y cuál es la probabilidad que un agente de la policía sea promovido *dado que* es mujer? Si estas dos probabilidades son iguales, no hay fundamentos para un argumento de discriminación ya que las oportunidades de ser promovidos son las mismas para agentes de la policía hombres o mujeres. Pero, si hay diferencia entre estas dos probabilidades condicionales se confirmará que los hombres y mujeres agentes de policía son considerados de manera distinta cuando se trata de las decisiones para promoverlos.

Ya se determinó que $P(A | M) = 0.30$. Ahora use los valores de probabilidad de la tabla 4.5 y la ecuación (4.7) de probabilidad condicional para calcular la probabilidad de que un agente de

FIGURA 4.8 PROBABILIDAD CONDICIONAL $P(A | B) = P(A \cap B)/P(B)$



la policía sea promovido dado que es mujer; es decir, $P(A | W)$. Use la ecuación (4.7) con W en lugar de B

$$P(A | W) = \frac{P(A \cap W)}{P(W)} = \frac{0.03}{0.20} = 0.15$$

¿Qué conclusión obtiene? La probabilidad de que un agente de policía sea promovido dado que es hombre es 0.30, el doble de 0.15, que es la probabilidad de que un agente de policía sea promovido dado que es mujer. Aunque el uso de la probabilidad condicional no demuestra por sí misma que haya discriminación en este caso, los valores de probabilidad condicional confirman el argumento presentado por las mujeres agentes de policía.

Eventos independientes

En el ejemplo anterior, $P(A) = 0.27$, $P(A | M) = 0.30$ y $P(A | W) = 0.15$. Es claro que a la probabilidad de ser promovido (evento A) le afecta o le influye el que el oficial sea un hombre o una mujer. En concreto, como $P(A | M) \neq P(A)$ los eventos A y M son eventos dependientes. Es decir, a la probabilidad del evento A (ser promovido) la altera o le afecta saber que se da el evento M (que el agente sea hombre). De manera similar, como $P(A | W) \neq P(A)$, los eventos A y W son *eventos dependientes*. Pero, si la probabilidad de un evento A no cambia por la existencia del evento M —es decir, si $P(A | M) = P(A)$ —, entonces los eventos A y M son **eventos independientes**. Esto lleva a la definición de la independencia de dos eventos.

EVENTOS INDEPENDIENTES

Dos eventos A y B son independientes si

$$P(A | B) = P(A) \quad (4.9)$$

o

$$P(B | A) = P(B) \quad (4.10)$$

Si no es así, los eventos son dependientes.

Ley de la multiplicación

Mientras que la ley de la suma de probabilidades sirve para calcular la probabilidad de la unión de dos eventos, la ley de la multiplicación es útil para calcular la probabilidad de la intersección de dos eventos. La ley de la multiplicación se basa en la definición de probabilidad condicional. Al despejar en las ecuaciones (4.7) y (4.8) $P(A \cap B)$, se obtiene la **ley de la multiplicación**.

LEY DE LA MULTIPLICACIÓN

$$P(A \cap B) = P(B)P(A | B) \quad (4.11)$$

o

$$P(A \cap B) = P(A)P(B | A) \quad (4.12)$$

Para ilustrar el uso de la ley de la multiplicación, considere el caso del departamento de circulación de un periódico al que 84% de los hogares de cierta región están suscritos a la edición diaria del periódico. Si D denota el evento un hogar suscrito a la edición diaria, $P(D) = 0.84$. Además, sabe que la probabilidad de que un hogar ya suscrito a la edición diaria se suscriba también a la edición dominical (evento S) es 0.75; esto es, $P(S | D) = 0.75$.

¿Cuál es la probabilidad de que un hogar se suscriba a ambas, a la edición diaria y a la dominical? Emplee la ley de la multiplicación y calcule la probabilidad deseada, $P(S \cap D)$.

$$P(S \cap D) = P(D)P(S | D) = 0.84(0.75) = 0.63$$

Así, sabe que 63% de los hogares se suscriben a ambas ediciones, a la diaria y a la dominical.

Antes de terminar esta sección hay que considerar el caso especial de la ley de la multiplicación cuando los eventos involucrados son independientes. Recuerde que los eventos A y B son independientes si $P(A | B) = P(A)$ o $P(B | A) = P(B)$. Por tanto, con las ecuaciones (4.11) y (4.12) obtiene, para el caso especial de eventos independientes, la siguiente ley de la multiplicación.

LEY DE LA MULTIPLICACIÓN PARA EVENTOS INDEPENDIENTES

$$P(A \cap B) = P(A)P(B)$$

(4.13)

Para calcular la probabilidad de la intersección de dos eventos independientes, simplemente se multiplican las probabilidades correspondientes. Observe que la ley de la multiplicación para eventos independientes proporciona otra manera de determinar si dos eventos son independientes. Es decir, si $P(A \cap B) = P(A)P(B)$, entonces A y B son independientes; si $P(A \cap B) \neq P(A)P(B)$, entonces A y B son dependientes.

Como una aplicación de la ley de la multiplicación para eventos independientes considere el caso del jefe de una gasolinera que por experiencia sabe que 80% de los clientes usan tarjeta de crédito al pagar la gasolina. ¿Cuál es la probabilidad de que los dos siguientes clientes paguen la gasolina con tarjeta de crédito? Sean

A = el evento el primer cliente paga con tarjeta de crédito

B = el evento el segundo cliente paga con tarjeta de crédito

entonces el evento que interesa es $A \cap B$. Si no hay ninguna otra información, será razonable suponer que A y B son eventos independientes. Por tanto,

$$P(A \cap B) = P(A)P(B) = (0.80)(0.80) = 0.64$$

Para concluir esta sección, observe que el interés por la probabilidad condicional surgió porque los eventos suelen estar relacionados. En esos casos, los eventos son dependientes y para calcular la probabilidad de estos eventos se usan las fórmulas para probabilidad condicional de las ecuaciones (4.7) y (4.8). Si dos eventos no están relacionados, son independientes; en este caso a las probabilidades de ninguno de los eventos les afecta el hecho de que el otro evento ocurra o no.

NOTAS Y COMENTARIOS

No hay que confundir la noción de eventos mutuamente excluyentes con la de eventos independientes. Dos eventos cuyas probabilidades no son cero, no pueden ser mutuamente excluyentes e indepen-

dientes. Si uno de los eventos mutuamente excluyentes ocurre, el otro evento no puede ocurrir; por tanto, la probabilidad de que ocurra el otro evento se reduce a cero.

Ejercicios

Métodos

30. Suponga dos eventos, A y B , y que $P(A) = 0.50$, $P(B) = 0.60$ y $P(A \cap B) = 0.40$.
 - a. Halle $P(A | B)$.
 - b. Halle $P(B | A)$.
 - c. ¿ A y B son independientes? ¿Por qué sí o por qué no?

31. Suponga dos eventos, A y B , que son mutuamente excluyentes. Admita, además, que $P(A) = 0.30$ y $P(B) = 0.40$.
- Obtenga $P(A \cap B)$.
 - Calcule $P(A | B)$.
 - Un estudiante de estadística argumenta que los conceptos de eventos mutuamente excluyentes y eventos independientes son en realidad lo mismo y que si los eventos son mutuamente excluyentes deben ser también independientes. ¿Está usted de acuerdo? Use la información sobre las probabilidades para justificar su respuesta.
 - Dados los resultados obtenidos, ¿qué conclusión sacaría usted acerca de los eventos mutuamente excluyentes e independientes?

Aplicaciones

32. Debido al aumento de los costos de los seguros, en Estados Unidos 43 millones de personas no cuentan con un seguro médico (*Time*, 1 de diciembre de 2003). En la tabla siguiente se muestran datos muestrales representativos de la cantidad de personas que cuentan con seguro médico.

		Seguro médico	
		Sí	No
Edad	18 a 34	750	170
	35 o mayor	950	130

- Con estos datos elabore una tabla de probabilidad conjunta y úsela para responder las preguntas restantes.
 - ¿Qué indican las probabilidades marginales acerca de la edad de la población de Estados Unidos?
 - ¿Cuál es la probabilidad de que una persona tomada en forma aleatoria no tenga seguro médico?
 - Si la persona tiene entre 18 y 34 años, ¿cuál es la probabilidad de que no tenga seguro médico?
 - Si la persona tiene 34 años o más ¿cuál es la probabilidad de que no tenga seguro médico?
 - Si la persona no tiene seguro médico, ¿cuál es la probabilidad de que tenga entre 18 y 34 años?
 - ¿Qué indica esta información acerca del seguro médico en Estados Unidos?
33. Una muestra de estudiantes de la maestría en administración de negocios, arrojó la siguiente información sobre la principal razón que tuvieron los estudiantes para elegir la escuela en donde hacen sus estudios.

Autoexamen

		Razones de su elección			
		Calidad de la escuela	Costo de la escuela	Otras	Totales
Tipo de estudiante	Tiempo completo	421	393	76	890
	Medio tiempo	400	593	46	1039
	Totales	821	986	122	1929

- Con estos datos elabore una tabla de probabilidad conjunta.
- Use las probabilidades marginales: calidad de la escuela, costo de la escuela y otras para comentar cuál es la principal razón por la que eligen una escuela.

- c. Si es un estudiante de tiempo completo, ¿cuál es la probabilidad de que la principal razón para su elección de la escuela haya sido la calidad de la escuela?
- d. Si es un estudiante de medio tiempo, ¿cuál es la probabilidad de que la principal razón para su elección de la escuela haya sido la calidad de la escuela?
- e. Si A denota el evento es estudiante de tiempo completo y B denota el evento la calidad de la escuela fue la primera razón para su elección, ¿son independientes los eventos A y B ? Justifique su respuesta.
34. La tabla siguiente muestra las probabilidades de los distintos tipos sanguíneos en la población.

	A	B	AB	O
Rh+	0.34	0.09	0.04	0.38
Rh-	0.06	0.02	0.01	0.06

- a. ¿Cuál es la probabilidad de que una persona tenga sangre tipo O?
- b. ¿De que tenga sangre Rh-?
- c. ¿Cuál es la probabilidad de que una persona sea Rh- dado que la persona tiene sangre tipo O?
- d. ¿Cuál es la probabilidad de que una persona tenga sangre tipo B dado que es Rh+?
- e. ¿Cuál es la probabilidad de que en un matrimonio, los dos sean Rh-?
- f. ¿Cuál es la probabilidad de que en un matrimonio, los dos tengan sangre AB?
35. El Departamento de Estadística Laboral de Estados Unidos reúne datos sobre las ocupaciones de las personas entre 25 y 64 años. La tabla siguiente presenta el número de hombres y mujeres (en millones) en cada una de las categorías ocupacionales.

Ocupación	Hombres	Mujeres
Directivo/Profesional	19 079	19 021
Enseñanza/Ventas/ Administrativo	11 079	19 315
Servicio	4 977	7 947
Producción con precisión	11 682	1 138
Operadores/Obrero	10 576	3 482
Agricultura/Ganadería/Silvicultura/Pesca	1 838	514

- a. Desarrolle una tabla de probabilidad conjunta.
- b. ¿Cuál es la probabilidad de que un trabajador mujer sea directivo o profesional?
- c. ¿Cuál es la probabilidad de que un trabajador hombre esté en producción con precisión?
- d. ¿Es la ocupación independiente del género? Justifique su respuesta con el cálculo de la probabilidad.
36. Reggie Miller de los Indiana Pacers tiene el record de la National Basketball Association de más canastas de 3 puntos anotadas en toda una carrera, acertando en 85% de sus tiros (*USA Today*, 22 de enero de 2004). Suponga que ya casi al final de un juego cometen una falta contra él y le conceden dos tiros.
- a. ¿Cuál es la probabilidad de que acierte en los dos tiros?
- b. ¿De que acierte en por lo menos uno de los dos tiros?
- c. ¿De que no acierte en ninguno de los dos tiros?
- d. Al final de un juego de básquetbol suele ocurrir que cometan faltas contra un jugador del equipo opuesto para detener el reloj del juego. La estrategia usual es cometer una falta contra el peor tirador del otro equipo. Suponga que el centro de los Indiana Pacers acierta 58% de sus tiros. Calcule para él las probabilidades calculadas en los incisos a, b y c y muestre que hacer una falta intencional contra el centro de los Indiana Pacers es mejor que hacerlo contra Reggie Miller.
37. Visa Card de Estados Unidos estudia con qué frecuencia usan sus tarjetas (de débito y de crédito) los consumidores jóvenes, entre 18 y 24 años. Los resultados del estudio proporcionan las probabilidades siguientes.

- La probabilidad de que un consumidor use su tarjeta al hacer una compra es 0.37.
- Dado que un consumidor usa su tarjeta, la probabilidad de que tenga entre 18 y 24 años es 0.19.
- Puesto que un consumidor usa su tarjeta, la probabilidad de que sea mayor de 24 años es 0.81.

Datos de la Oficina de Censos de Estados Unidos indican que 14% de los consumidores tienen entre 18 y 24 años.

- Ya que un consumidor tiene entre 18 y 24 años, ¿cuál es la probabilidad de que use su tarjeta?
 - Dado que un consumidor tiene más de 24 años, ¿cuál es la probabilidad de que use su tarjeta?
 - ¿Qué interpretación se le da a las probabilidades de los incisos a y b?
 - ¿Empresas como Visa, Master Card y Discover deben proporcionar tarjetas a los consumidores entre 18 y 24 años, antes de que tengan una historia crediticia? Si no, explique. Si sí, ¿qué restricciones deben poner las empresas a estos consumidores?
38. En un estudio de Morgan Stanley Consumer Research se muestrearon hombres y mujeres y se les preguntó qué preferían tomar: agua de botella o una bebida deportiva como Gatorade o Propel Fitness (*The Atlanta Journal-Constitution*, 28 de diciembre de 2005). Suponga que en el estudio hayan participado 200 hombres y 200 mujeres y que de todos 280 hayan preferido el agua de botella. En el grupo de los que preferían bebidas deportivas, 80 eran hombres y 40 eran mujeres.

Sea

M = el evento el consumidor es hombre

W = el evento el consumidor es mujer

B = el evento el consumidor prefiere agua de botella

S = el evento el consumidor prefiere una bebida deportiva

- ¿Cuál es la probabilidad de que en este estudio una persona prefiera agua de botella?
- ¿De que en este estudio una persona prefiera una bebida deportiva?
- ¿Cuáles son las probabilidades condicionales $P(M|S)$ y $P(W|S)$?
- ¿Cuáles son las probabilidades conjuntas $P(M \cap S)$ y $P(W \cap S)$?
- Dado que un consumidor es hombre, ¿cuál es la probabilidad de que prefiera una bebida deportiva?
- Ya que un consumidor es mujer, ¿cuál es la probabilidad de que prefiera una bebida deportiva?
- ¿Depende la preferencia por una bebida deportiva de que el consumidor sea hombre o mujer? Explique usando la información sobre las probabilidades.

4.5

Teorema de Bayes

En el estudio de la probabilidad condicional vio que revisar las probabilidades cuando se obtiene más información es parte importante del análisis de probabilidades. Por lo general, se suele iniciar el análisis con una estimación de probabilidad inicial o **probabilidad previa** de los eventos que interesan. Después, de fuentes como una muestra, una información especial o una prueba del producto, se obtiene más información sobre estos eventos. Dada esta nueva información, se modifican o revisan los valores de probabilidad mediante el cálculo de probabilidades revisadas a las que se les conoce como **probabilidades posteriores**. El **teorema de Bayes** es un medio para calcular estas probabilidades. En la figura 4.9 se presentan los pasos de este proceso de revisión de la probabilidad.

FIGURA 4.9 REVISIÓN DE LA PROBABILIDAD USANDO EL TEOREMA DE BAYES

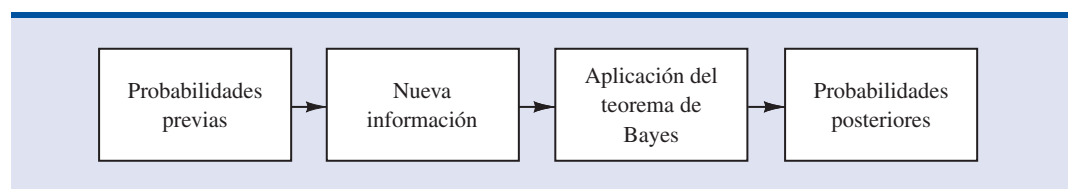


TABLA 4.6 CALIDAD DE DOS PROVEEDORES

	Porcentaje de piezas buenas	Porcentaje de piezas malas
Proveedor 1	98	2
Proveedor 2	95	5

Como aplicación del teorema de Bayes, considere una fábrica que compra piezas de dos proveedores. Sea A_1 el evento la pieza proviene del proveedor 1 y A_2 el evento la pieza proviene del proveedor 2. De las piezas que compra la fábrica, 65% proviene del proveedor 1 y 35% restante proviene del proveedor 2. Por tanto, si toma una pieza aleatoriamente, le asignará las probabilidades previas $P(A_1) = 0.65$ y $P(A_2) = 0.35$.

La calidad de las piezas compradas varía de acuerdo con el proveedor. Por experiencia, sabe que la calidad de los dos proveedores es como muestra la tabla 4.6. Si G denota el evento la pieza está buena y B denota el evento la pieza está mala, la información de la tabla 4.6 proporciona los siguientes valores de probabilidad condicional.

$$\begin{aligned} P(G \mid A_1) &= 0.98 & P(B \mid A_1) &= 0.02 \\ P(G \mid A_2) &= 0.95 & P(B \mid A_2) &= 0.05 \end{aligned}$$

El diagrama de árbol de la figura 4.10 representa el proceso de recibir una pieza, de uno de los dos proveedores, y después determinar si la pieza es buena o mala como experimento de dos pasos. Se observa que existen cuatro resultados experimentales: dos corresponden a que la pieza esté buena y dos corresponden a que la pieza esté mala.

Cada uno de los resultados experimentales es la intersección de dos eventos, de manera que para calcular estas probabilidades puede usar la ley de la multiplicación. Por ejemplo,

$$P(A_1, G) = P(A_1 \cap G) = P(A_1)P(G \mid A_1)$$

FIGURA 4.10 DIAGRAMA DE ÁRBOL PARA EL EJEMPLO DE LOS DOS PROVEEDORES

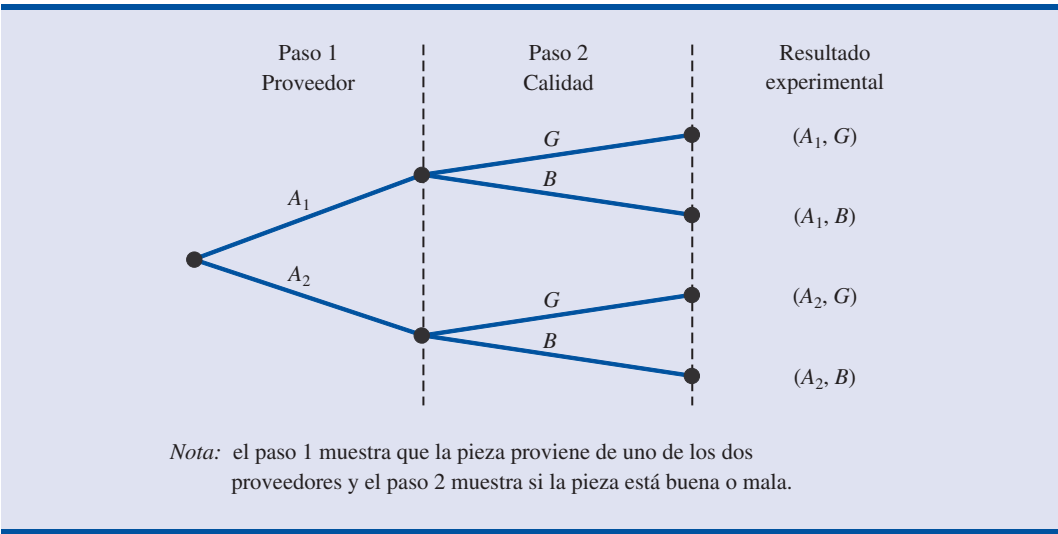
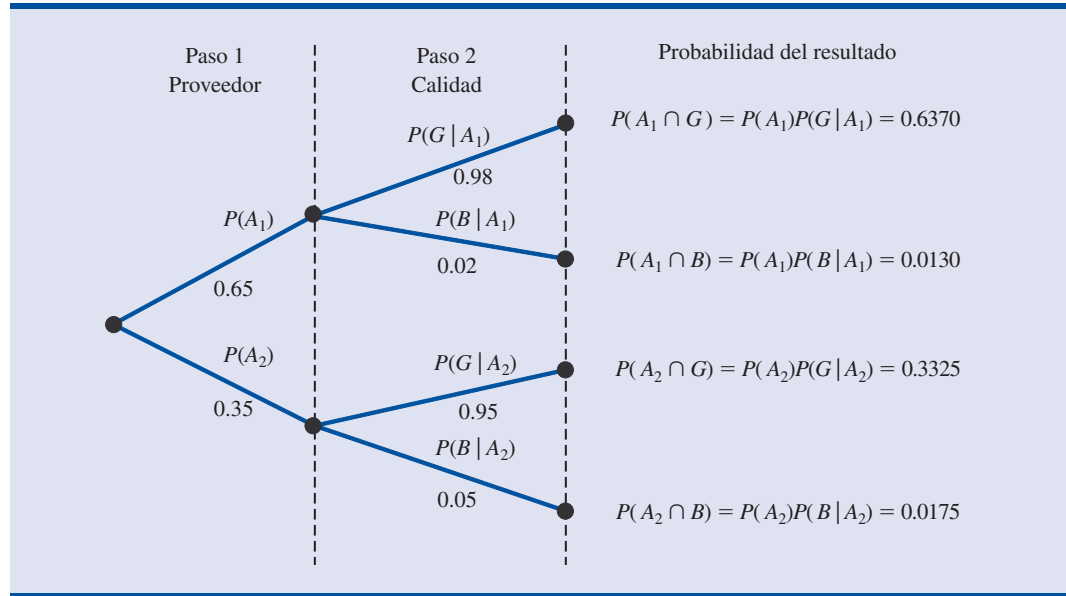


FIGURA 4.11 ÁRBOL DE PROBABILIDAD PARA EL EJEMPLO DE LOS DOS PROVEEDORES

El proceso del cálculo de estas probabilidades conjuntas se representa mediante un árbol de probabilidad (figura 4.11). De izquierda a derecha por el árbol, las probabilidades de cada una de las ramas del paso 1 son probabilidades previas y las probabilidades de cada una de las ramas del paso 2 son probabilidades condicionales. Para hallar la probabilidad de cada uno de los resultados experimentales, simplemente se multiplican las probabilidades de las ramas que llevan a ese resultado. En la figura 4.11 se muestra cada una de estas probabilidades conjuntas junto con las probabilidades en cada rama.

Suponga ahora que las piezas de los dos proveedores se emplean en el proceso de fabricación de esta empresa y que una máquina se descompone al tratar de procesar una pieza mala. Dada la información de que la pieza está mala, ¿cuál es la probabilidad de que sea del proveedor 1 y cuál es la probabilidad de que sea del proveedor 2? Para responder estas preguntas aplique el teorema de Bayes usando la información del árbol de probabilidad (figura 4.11).

Como B es el evento la parte está mala, lo que busca son las probabilidades posteriores $P(A_1 | B)$ y $P(A_2 | B)$. De acuerdo con la ley para la probabilidad condicional

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)} \quad (4.14)$$

Del árbol de probabilidad

$$P(A_1 \cap B) = P(A_1)P(B | A_1) \quad (4.15)$$

Para hallar $P(B)$, se observa que B sólo puede presentarse de dos maneras: $(A_1 \cap B)$ y $(A_2 \cap B)$. Por tanto,

$$\begin{aligned} P(B) &= P(A_1 \cap B) + P(A_2 \cap B) \\ &= P(A_1)P(B | A_1) + P(A_2)P(B | A_2) \end{aligned} \quad (4.16)$$

Sustituyendo las ecuaciones (4.15) y (4.16) en la ecuación (4.14) y expresando de manera similar $P(A_2 | B)$ se obtiene el teorema de Bayes para el caso de dos eventos.

Al reverendo Thomas Bayes, un ministro presbiteriano, se le atribuye la idea inicial que llevó a la versión del teorema de Bayes que se usa en la actualidad.

TEOREMA DE BAYES (CASO DE DOS EVENTOS)

$$P(A_1 | B) = \frac{P(A_1)P(B | A_1)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2)} \quad (4.17)$$

$$P(A_2 | B) = \frac{P(A_2)P(B | A_2)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2)} \quad (4.18)$$

A partir de la ecuación (4.17) y los valores de probabilidad del ejemplo, se tiene

$$\begin{aligned} P(A_1 | B) &= \frac{P(A_1)P(B | A_1)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2)} \\ &= \frac{(0.65)(0.02)}{(0.65)(0.02) + (0.35)(0.05)} = \frac{0.0130}{0.0130 + 0.0175} \\ &= \frac{0.0130}{0.0305} = 0.4262 \end{aligned}$$

Y usando la ecuación (4.18) se encuentra $P(A_2 | B)$.

$$\begin{aligned} P(A_2 | B) &= \frac{(0.35)(0.05)}{(0.65)(0.02) + (0.35)(0.05)} \\ &= \frac{0.0175}{0.0130 + 0.0175} = \frac{0.0175}{0.0305} = 0.5738 \end{aligned}$$

Observe que al principio de este ejemplo, la probabilidad de seleccionar una pieza y que fuera del proveedor 1 era 0.65. Sin embargo, dada la información de que la pieza está mala, la probabilidad de que la pieza provenga del proveedor 1 bajó a 0.4262. En efecto, si la pieza está mala, la posibilidad de que sea del proveedor 2 es mayor que 50-50; es decir, $P(A_2 | B) = 0.5738$.

El teorema de Bayes es aplicable cuando los eventos para los que se quiere calcular la probabilidad revisada son mutuamente excluyentes y su unión es todo el espacio muestral.* En el caso de n eventos mutuamente excluyentes A_1, A_2, \dots, A_n , cuya unión sea todo el espacio muestral, el teorema de Bayes aplica para calcular cualquiera de las probabilidades posteriores $P(A_i | B)$ como se muestra a continuación

TEOREMA DE BAYES

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + \dots + P(A_n)P(B | A_n)} \quad (4.19)$$

*Si la unión de los eventos es todo el espacio muestral, los eventos son *colectivamente exhaustivos*.

Con las probabilidades previas $P(A_1), P(A_2), \dots, P(A_n)$ y las probabilidades condicionales adecuadas $P(B | A_1), P(B | A_2), \dots, P(B | A_n)$, se usa la ecuación (4.19) para calcular la probabilidad posterior de los eventos A_1, A_2, \dots, A_n .

Método tabular

Para realizar los cálculos del teorema de Bayes es útil emplear un método tabular. En la tabla 4.7 se muestra este método aplicado al problema de las piezas de los proveedores. Los cálculos que se muestran ahí se realizan mediante los pasos siguientes.

Paso 1. Se harán las columnas siguientes:

Columna 1: Para los eventos mutuamente excluyentes A_i de los que quiere tener la probabilidad posterior

Columna 2: Para las probabilidades previas $P(A_i)$ de los eventos

Columna 3: Para las probabilidades condicionales $P(B | A_i)$ de la nueva información B dado cada evento

Paso 2. En la columna 4 se calculan las probabilidades conjuntas $P(A_i \cap B)$, de cada evento y la nueva información, empleando la ley de la multiplicación. Estas probabilidades conjuntas se encuentran multiplicando las probabilidades previas de la columna 2 por las correspondientes probabilidades condicionales de la columna 3; es decir, $P(A_i \cap B) = P(A_i)P(B | A_i)$.

Paso 3. Sume las probabilidades de la columna 4. Esta suma es la probabilidad de la nueva información, $P(B)$. Así, en la tabla 4.7 se ve que la probabilidad de que una pieza sea del proveedor 1 y esté mala es 0.0130 y que la probabilidad de que la pieza sea del proveedor 2 y esté mala es 0.0175. Como éstas son las únicas dos maneras de tener una pieza mala, la suma $0.0130 + 0.0175$, que es 0.0305, da la probabilidad de hallar una pieza mala en las piezas recibidas de los dos proveedores.

Paso 4. En la columna 5 se calculan las probabilidades posteriores usando la relación básica de la probabilidad condicional.

$$P(A_i | B) = \frac{P(A_i \cap B)}{P(B)}$$

Observe que las probabilidades conjuntas $P(A_i \cap B)$ están en la columna 4 y que la probabilidad $P(B)$ es la suma de la columna 4.

TABLA 4.7 MÉTODO TABULAR PARA LOS CÁLCULOS DEL TEOREMA DE BAYES APLICADO AL EJEMPLO DE LOS DOS PROVEEDORES

(1) Eventos A_i	(2) Probabilidades previas $P(A_i)$	(3) Probabilidades condicionales $P(B A_i)$	(4) Probabilidades conjuntas $P(A_i \cap B)$	(5) Probabilidades posteriores $P(A_i B)$
A_1	0.65	0.02	0.0130	$0.0130/0.0305 = 0.4262$
A_2	0.35	0.05	0.0175	$0.0175/0.0305 = 0.5738$
	1.00		$P(B) = 0.0305$	1.0000

NOTAS Y COMENTARIOS

1. El teorema de Bayes se usa mucho en la toma de decisiones. Las probabilidades previas suelen ser estimaciones subjetivas dadas por la persona que toma las decisiones. Se obtiene información muestral y se usan las probabilidades posteriores para emplearlas en la toma de decisiones.
2. Un evento y su complemento son mutuamente excluyentes y su unión es todo el espacio muestral. Por tanto, el teorema de Bayes siempre se emplea para calcular la probabilidad posterior de un evento y su complemento.

Ejercicios

Métodos

Autoexamen

39. Las probabilidades previas de los eventos A_1 y A_2 son $P(A_1) = 0.40$ y $P(A_2) = 0.60$. Sabe también que $P(A_1 \cap A_2) = 0$. Suponga que $P(B | A_1) = 0.20$ y $P(B | A_2) = 0.05$.
 - a. ¿ A_1 y A_2 son eventos mutuamente excluyentes? Explique.
 - b. Calcule $P(A_1 \cap B)$ y $P(A_2 \cap B)$.
 - c. Calcule $P(B)$.
 - d. Emplee el teorema de Bayes para calcular $P(A_1 | B)$ y $P(A_2 | B)$.
40. Las probabilidades previas de los eventos A_1, A_2 y A_3 son $P(A_1) = 0.20$, $P(A_2) = 0.50$ y $P(A_3) = 0.30$. Las probabilidades condicionales del evento B dados los eventos A_1, A_2 y A_3 son $P(B | A_1) = 0.50$, $P(B | A_2) = 0.40$ y $P(B | A_3) = 0.30$.
 - a. Calcule $P(B \cap A_1)$, $P(B \cap A_2)$ y $P(B \cap A_3)$.
 - b. Emplee el teorema de Bayes, ecuación (4.19), para calcular la probabilidad posterior $P(A_2 | B)$.
 - c. Use el método tabular para emplear el teorema de Bayes en el cálculo de $P(A_1 | B)$, $P(A_2 | B)$ y $P(A_3 | B)$.

Aplicaciones

41. Una empresa de consultoría presenta una oferta para un gran proyecto de investigación. El director de la firma piensa inicialmente que tiene 50% de posibilidades de obtener el proyecto. Sin embargo, mas tarde, el organismo al que se le hizo la oferta pide más información sobre la oferta. Por experiencia se sabe que en 75% de las ofertas aceptadas y en 40% de las ofertas no aceptadas, este organismo solicita más información.
 - a. ¿Cuál es la probabilidad previa de que la oferta sea aceptada (es decir, antes de la solicitud de más información)?
 - b. ¿Cuál es la probabilidad condicional de que se solicite más información dado que la oferta será finalmente aceptada?
 - c. Calcule la probabilidad posterior de que la oferta sea aceptada dado que se solicitó más información.
42. Un banco local revisa su política de tarjetas de crédito con objeto de retirar algunas de ellas. En el pasado aproximadamente 5% de los tarjetahabientes incumplieron, dejando al banco sin posibilidad de cobrar el saldo pendiente. De manera que el director estableció una probabilidad previa de 0.05 de que un tarjetahabiente no cumpla. El banco encontró también que la probabilidad de que un cliente que es cumplido no haga un pago mensual es 0.20. Por supuesto la probabilidad de no hacer un pago mensual entre los que incumplen es 1.
 - a. Dado que un cliente no hizo el pago de uno o más meses, calcule la probabilidad posterior de que el cliente no cumpla.
 - b. El banco deseará retirar sus tarjetas si la probabilidad de que un cliente no cumpla es mayor que 0.20. ¿Debe retirar el banco una tarjeta si el cliente no hace un pago mensual?

Autoexamen

43. En los automóviles pequeños el rendimiento de la gasolina es mayor, pero no son tan seguros como los coches grandes. Los automóviles pequeños constituyen 18% de los vehículos en circulación, pero en accidentes con automóviles pequeños se registraron 11 898 víctimas mortales en uno de los últimos años (*Reader's Digest*, mayo de 2000). Suponga que la probabilidad de que un automóvil pequeño tenga un accidente es 0.18. La probabilidad de que en un accidente con un automóvil pequeño haya una víctima mortal es 0.128 y la probabilidad de que haya una víctima mortal si el automóvil no es pequeño es 0.05. Usted se entera de un accidente en el que hubo una víctima mortal. ¿Cuál es la probabilidad de que el accidente lo haya tenido un automóvil pequeño?
44. La American Council of Education informa que en Estados Unidos 47% de los estudiantes que ingresan en la universidad terminan sus estudios en un lapso de cinco años (Associated Press, 6 de mayo de 2002). Suponga que en los registros de terminación de estudios encuentra que 50% de los estudiantes que terminan sus estudios en cinco años son mujeres y 45% de quienes no terminan sus estudios en cinco años son mujeres. Los estudiantes que no terminan sus estudios en cinco años son estudiantes que han abandonado sus estudios o que están por terminarlos.
- Sea A_1 = el estudiante termina sus estudios en cinco años
 A_2 = el estudiante no termina sus estudios en cinco años
 W = el estudiante es mujer
 Empleando la información dada, dé las probabilidades siguientes: $P(A_1)$, $P(A_2)$, $P(W|A_1)$ y $P(W|A_2)$.
 - ¿Cuál es la probabilidad de que una estudiante termine sus estudios en cinco años?
 - ¿Cuál es la probabilidad de que un estudiante termine sus estudios en cinco años?
 - Dados los resultados anteriores, ¿cuál es el porcentaje de mujeres y cuál es el porcentaje de hombres que entran en la universidad?
45. En un artículo acerca del crecimiento de las inversiones, la revista *Money* informa que las acciones en medicamentos muestran una poderosa tendencia de largo plazo y ofrecen a los inversionistas potenciales inigualables y duraderas ganancias. La Health Care Financing Administration confirma estas conclusiones con su pronóstico de que para 2010 el consumo de medicamentos llegará a \$366 mil millones, cuando en 2000 era de \$117 mil millones. Muchas de las personas de 65 años o más necesitan medicamentos. Entre estas personas, 82% necesita medicamentos de manera regular, 55% usa tres o más medicamentos de manera regular y 40% necesita cinco o más medicamentos regularmente. En cambio entre las personas menores de 65 años, 49% usa medicamentos de manera regular, 37% necesita tres o más medicamentos de manera regular y 28% usa cinco o más medicamentos regularmente (*Money*, septiembre de 2001). La Oficina de Censos de Estados Unidos informa que de los 281 421 906 habitantes de Estados Unidos, 34 991 753 son personas de 65 años o mayores (U.S. Census Bureau, *Census 2000*).
- Calcule la probabilidad de que en Estados Unidos una persona tenga 65 años o más.
 - Calcule la probabilidad de que una persona necesite medicamentos de manera regular.
 - Calcule la probabilidad de que una persona tenga 65 años o más y necesite cinco o más medicamentos.
 - Dado que una persona usa cinco o más medicamentos, calcule la probabilidad de que tenga 65 años o más.

Resumen

En este capítulo se introdujeron conceptos básicos de probabilidad y se ilustró cómo usar el análisis de probabilidad para obtener información útil para la toma de decisiones. Se describió cómo interpretar la probabilidad como una medida numérica de la posibilidad de que ocurra un evento. Además, se vio que la probabilidad de un evento se puede calcular, ya sea sumando las probabilidades de los resultados experimentales (puntos muestrales) que comprende el evento o usando las relaciones que establecen las leyes de probabilidad de la adición, de la probabilidad condicional y de la multiplicación. En el caso de que se obtenga información adicional, se mostró cómo usar el teorema de Bayes para obtener probabilidades revisadas o posteriores.

Glosario

Probabilidad Medida numérica de la posibilidad de que ocurra un evento.

Experimento Proceso para generar resultados bien definidos.

Espacio muestral Conjunto de todos los resultados experimentales.

Punto muestral Un elemento del espacio muestral. Un punto muestral que representa un resultado experimental.

Diagrama de árbol Representación gráfica que ayuda a visualizar un experimento de pasos múltiples.

Requerimientos básicos en la asignación de probabilidades Dos requerimientos que restringen la manera en que se asignan probabilidades son: 1) Para cada resultado experimental E_i se debe tener $0 \leq P(E_i) \leq 1$; 2) si E_1, E_2, \dots, E_n son todos los resultados experimentales, se debe tener que $P(E_1) + P(E_2) + \dots + P(E_n) = 1.0$.

Método clásico Sirve para la asignación de probabilidades, es apropiado cuando todos los resultados experimentales son igualmente posibles.

Método de las frecuencias relativas Útil para la asignación de probabilidades, es conveniente cuando se tienen datos para estimar la proporción de veces que se presentará un resultado experimental si se repite un gran número de veces.

Método subjetivo Método para la asignación de probabilidades basado en un juicio.

Evento Colección de puntos muestrales

Complemento de A El evento que consta de todos los puntos muestrales que no están en A.

Diagrama de Venn Una representación gráfica para mostrar de manera simbólica el espacio muestral y las operaciones con eventos en la cual el espacio muestral se representa como un rectángulo y los eventos se representan como círculos dentro del espacio muestral.

Unión de A y B Evento que contiene todos los puntos muestrales que pertenecen a A o a B o a ambos. La unión se denota $A \cup B$.

Intersección de A y B Evento que contiene todos los puntos muestrales que pertenecen tanto a A como a B. La intersección se denota $A \cap B$.

Ley de la adición Ley de probabilidad que se usa para calcular la unión de dos eventos. Es $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Si los eventos son mutuamente excluyentes, $P(A \cap B) = 0$; en este caso la ley de la adición se reduce a $P(A \cup B) = P(A) + P(B)$.

Eventos mutuamente excluyentes Eventos que no tienen puntos muestrales en común; es decir, $A \cap B$ es vacío y $P(A \cap B) = 0$.

Probabilidad condicional Probabilidad de un evento dado que otro evento ya ocurrió. La probabilidad condicional de A dado B es $P(A | B) = P(A \cap B)/P(B)$.

Probabilidad conjunta La probabilidad de que dos eventos ocurran al mismo tiempo; es decir, la probabilidad de la intersección de dos eventos.

Probabilidad marginal Los valores en los márgenes de una tabla de probabilidad conjunta que dan las probabilidades de cada evento por separado.

Eventos independientes Son dos eventos, A y B, para los que $P(A | B) = P(A)$ o $P(B | A) = P(B)$; es decir, los eventos no tienen ninguna influencia uno en otro.

Ley de la multiplicación Una ley de probabilidad que se usa para calcular la probabilidad de la intersección de dos eventos. Esto es $P(A \cap B) = P(B)P(A | B)$ o $P(A \cap B) = P(A)P(B | A)$. Para eventos independientes se reduce a $P(A \cap B) = P(A)P(B)$.

Probabilidades previas Estimaciones iniciales de las probabilidades de eventos.

Probabilidades posteriores Probabilidades revisadas de eventos basadas en informaciones adicionales.

Teorema de Bayes Método usado para calcular las probabilidades posteriores.

Fórmulas clave

Regla de conteo para combinaciones

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \quad (4.1)$$

Regla de conteo para permutaciones

$$P_n^N = n! \binom{N}{n} = \frac{N!}{(N-n)!} \quad (4.2)$$

Cálculo de la probabilidad usando el complemento

$$P(A) = 1 - P(A^c) \quad (4.5)$$

Ley de la adición

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.6)$$

Probabilidad condicional

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (4.7)$$

$$P(B | A) = \frac{P(A \cap B)}{P(A)} \quad (4.8)$$

Ley de la multiplicación

$$P(A \cap B) = P(B)P(A | B) \quad (4.11)$$

$$P(A \cap B) = P(A)P(B | A) \quad (4.12)$$

Ley de la multiplicación para eventos independientes

$$P(A \cap B) = P(A)P(B) \quad (4.13)$$

Teorema de Bayes

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2) + \cdots + P(A_n)P(B | A_n)} \quad (4.19)$$

Ejercicios complementarios

46. En un sondeo se les pidió a 1035 adultos su opinión respecto a los negocios (*BusinessWeek*, 11 de septiembre de 2000). Una de las preguntas era: “¿Cómo califica usted a las empresas estadounidenses respecto a la calidad de los productos y competitividad a nivel mundial?” Las respuestas fueron: excelentes, 18%; bastante buenas, 50%; regulares, 26%; malas, 5% y no saben o no contestaron 1%.
 - a. ¿Cuál es la probabilidad de que un interrogado considere a las empresas estadounidenses bastante buenas o excelentes?
 - b. ¿Cuántos de los interrogados consideraron malas a las empresas estadounidenses?
 - c. ¿Cuántos de los interrogados dijo no saber o no contestó?
47. Un administrador financiero realiza dos nuevas inversiones, una en la industria del petróleo y otra en bonos municipales. Después de un año cada una de las inversiones se clasificará como buena o no. Considere como un experimento el resultado que se obtiene con estas dos acciones.
 - a. ¿Cuántos puntos muestrales hay en este experimento?
 - b. Presente un diagrama de árbol y enumere los puntos muestrales.
 - c. Sea O = el evento la inversión en la industria del petróleo es buena y M = el evento la inversión en los fondos municipales es buena. Dé los puntos muestrales de O y de M .
 - d. Enumere los puntos muestrales de la unión de los eventos ($O \cup M$).
 - e. Cuento los puntos muestrales de la intersección de los eventos ($O \cap M$).
 - f. ¿Son mutuamente excluyentes los eventos O y M ? Explique.

48. A principios de 2003, el presidente de Estados Unidos propuso eliminar los impuestos a los dividendos de los accionistas con el argumento de que era un doble impuesto. Las corporaciones pagan impuestos sobre las ganancias que luego son repartidas como dividendos. En un sondeo realizado a 671 estadounidenses, Techno Metrica Market Intelligence halló que 47% estaban a favor de la propuesta, 44% se oponían a ella y 9% no estaban seguros (*Investor's Business Daily*, 13 de enero de 2003). Al analizar las respuestas de acuerdo con la pertenencia a los partidos políticos, se encontró en el sondeo que 29% de los demócratas estaban a favor, 64% de los republicanos estaban a favor y 48% de los independientes estaban a favor.
- ¿Cuántos de los encuestados estuvieron a favor de la eliminación de los impuestos a los dividendos?
 - ¿Cuál es la probabilidad condicional de que una persona esté a favor de la propuesta dado que es demócrata?
 - ¿Es la afiliación partidaria independiente de que una persona esté a favor de la propuesta?
 - Si se supone que las respuestas de las personas estuvieron de acuerdo con sus propios intereses, ¿qué grupo se beneficiará más con la aceptación de la propuesta?
49. En un estudio realizado con 31 000 ingresos a hospitales en el estado de Nueva York se encontró que 4% de los ingresados sufrieron daños a causa del tratamiento. Un séptimo de estos daños condujeron a la muerte y un cuarto se debió a negligencia médica. En uno de cada 7.5 casos de negligencia médica se levantó una demanda y en una de cada dos demandas se tuvo que pagar una indemnización.
- ¿Cuál es la probabilidad de que una persona que ingresa en un hospital sufra un daño a causa del tratamiento debido a negligencia médica?
 - ¿Cuál es la probabilidad de que una persona que ingresa en un hospital muera a causa de daños producidos por el tratamiento?
 - En el caso de daños causado por negligencia médica, ¿cuál es la probabilidad de que la demanda ocasione una indemnización?
50. En una encuesta por teléfono para determinar la opinión de los televidentes respecto a un nuevo programa de televisión se obtuvieron las opiniones siguientes:

Opinión	Frecuencia
Malo	4
Regular	8
Bueno	11
Muy bueno	14
Excelente	13

- ¿Cuál es la probabilidad de que un televidente tomado aleatoriamente opine que el nuevo programa es bueno o le dé un calificativo mejor.
 - ¿Cuál es la probabilidad de que un televidente tomado aleatoriamente opine que el nuevo programa es regular o le dé un calificativo inferior?
51. En la siguiente tabulación cruzada se muestra el ingreso familiar de acuerdo con el nivel de estudios del cabeza de familia (*Statistical Abstract of the United States: 2002*).

Nivel de estudios	Ingreso familiar (en miles de \$)					Total
	Menos de 25	25.0–49.9	50.0–74.9	75.0–99.9	100 o más	
Preparatoria sin terminar	9 285	4 093	1 589	541	354	15 862
Preparatoria terminada	10 150	9 821	6 050	2 737	2 028	30 786
Estudios universitarios sin terminar	6 011	8 221	5 813	3 215	3 120	26 380
Estudios universitarios terminados	2 138	3 985	3 952	2 698	4 748	17 521
Estudios de posgrado	813	1 497	1 815	1 589	3 765	9 479
Total	28 397	27 617	19 219	10 780	14 015	100 028

- a. Elabore una tabla de probabilidad conjunta.
 - b. ¿Cuál es la probabilidad de que el cabeza de familia no haya terminado la preparatoria?
 - c. ¿Cuál es la probabilidad de que el cabeza de familia haya terminado la universidad o tenga estudios de posgrado?
 - d. ¿Cuál es la probabilidad de que si el cabeza de familia terminó la universidad, el ingreso familiar sea \$100 000 o más?
 - e. ¿Cuál es la probabilidad de que el ingreso familiar sea menor a \$25 000?
 - f. ¿Cuál es la probabilidad de que una familia en la que el cabeza de familia terminó la universidad, tenga un ingreso familiar menor a \$25 000?
 - g. ¿El ingreso familiar es independiente del nivel de educación?
52. En un estudio realizado entre los 2010 nuevos estudiantes inscritos a las maestrías de negocios se obtuvieron los datos siguientes.

		Hizo solicitudes en varias universidades	
		Sí	No
Grupos de edades	23 o menos	207	201
	24–26	299	379
	27–30	185	268
	31–35	66	193
	36 o más	51	169

- a. Para un estudiante de maestría tomado en forma aleatoria elabore una tabla de probabilidad conjunta para el experimento que consiste en observar la edad del estudiante y si hizo solicitudes en varias universidades.
 - b. ¿Cuál es la probabilidad de que un estudiante tomado en forma aleatoria tenga 23 años o menos?
 - c. ¿Cuál es la probabilidad de que un estudiante tomado en forma aleatoria tenga más de 26 años?
 - d. ¿Cuál es la probabilidad de que un estudiante tomado en forma aleatoria haya hecho solicitud en varias universidades?
53. Vaya nuevamente a los datos de los nuevos estudiantes inscritos a las maestrías de negocios del ejercicio 52.
- a. Dado que una persona hizo solicitudes en varias universidades, ¿cuál es la probabilidad de que tenga entre 24 y 26 años?
 - b. Ya que una persona tiene 36 años o más, ¿cuál es la probabilidad de que haya hecho solicitudes en varias universidades?
 - c. ¿Cuál es la probabilidad de que una persona entre 24 y 26 años haya hecho solicitudes en varias universidades?
 - d. Suponga que la persona sólo hizo solicitud para una universidad. ¿Cuál es la probabilidad de que la persona tenga 31 años o más?
 - e. ¿La edad y el hacer solicitudes en varias universidades son independientes? Explique.
54. En una encuesta realizada por IBD/TIPPP para obtener información sobre la opinión respecto a las inversiones para el retiro (*Investor's Business Daily*, 5 de mayo de 2000) se les preguntó a los hombres y mujeres interrogados qué tan importante les parecía que era el nivel de riesgo al elegir una inversión para el retiro. Con los datos obtenidos se elaboró la siguiente tabla de probabilidades conjuntas. “Importante” significa que el interrogado respondió que el nivel de riesgo era importante o muy importante.

	Hombre	Mujer	Total
Importante	0.22	0.27	0.49
No importante	0.28	0.23	0.51
Total	0.50	0.50	1.00

- a. ¿Cuál es la probabilidad de que uno de los interrogados diga que es importante?
 - b. ¿Cuál es la probabilidad de que una de las mujeres interrogadas diga que es importante?
 - c. ¿Cuál es la probabilidad de que uno de los hombres interrogados diga que es importante?
 - d. ¿El nivel de riesgo es independiente del género del interrogado?
 - e. ¿La opinión de hombres y mujeres difiere respecto al riesgo?
55. Una empresa grande de productos de consumo transmite por televisión publicidad para uno de sus jabones. De acuerdo con una encuesta realizada, se asignaron probabilidades a los eventos siguientes.

B = una persona compra el producto

S = una persona recuerda haber visto la publicidad

$B \cap S$ = una persona compra el producto y recuerda haber visto la publicidad.

Las probabilidades fueron $P(B) = 0.20$, $P(S) = 0.40$ y $P(B \cap S) = 0.12$.

- a. ¿Cuál es la probabilidad de que una persona compre el producto dado que recuerda haber visto la publicidad? ¿Ver la publicidad aumenta la probabilidad de que el individuo compre el producto? Si usted tuviera que tomar la decisión, ¿recomendaría que continuara la publicidad (suponiendo que los costos sean razonables)?
 - b. Si una persona que no compra el producto de la empresa compra el de la competencia. ¿Cuál sería su estimación de la participación de la empresa en el mercado? ¿Esperaría que continuando con la publicidad aumentara la participación de la empresa en el mercado? ¿Por qué sí o por qué no?
 - c. La empresa probó también otra publicidad y los valores de probabilidad asignados fueron $P(S) = 0.30$, $P(B \cap S) = 0.10$. Dé $P(B | S)$ en el caso de esta otra publicidad. ¿Qué publicidad parece tener mejor efecto en la compra de los clientes?
56. Cooper Realty es una empresa inmobiliaria pequeña que se encuentra en Albany, Nueva York y que se especializa en la venta de casas residenciales. Últimamente quiso saber cuál era la posibilidad de que una de las casas que tiene en venta se vendiera en menos de un determinado número de días. Mediante un análisis de 800 casas vendidas por la empresa en los años anteriores se obtuvieron los datos siguientes.

		Días en venta hasta la compra			Total
		Menos de 30	31–90	Más de 90	
Precio pedido inicialmente	Menos de \$150 000	50	40	10	100
	\$150 000–\$199 999	20	150	80	250
	\$200 000–\$250 000	20	280	100	400
	Más de \$250 000	10	30	10	50
	Total	100	500	200	800

- a. Si A se define como el evento de que la casa esté en venta más de 90 días antes de ser vendida, estime la probabilidad de A .
- b. Si B se define como el evento de que el precio inicial sea menor que \$150 000, estime la probabilidad de B .
- c. ¿Cuál es la probabilidad de $A \cap B$?
- d. Suponga que se acaba de firmar un contrato para vender una casa en un precio inicial menor que \$150 000, ¿cuál es la probabilidad de que a Cooper Realty le tome menos de 90 días venderla?
- e. ¿Los eventos A y B son independientes?

57. Una empresa estudió el número de accidentes ocurridos en su planta de Brownsville, Texas. De acuerdo con información anterior, 6% de los empleados sufrieron accidentes el año pasado. Los directivos creen que un programa especial de seguridad reducirá este año los accidentes a 5%. Se estima además que 15% de los empleados que sufrieron un accidente el año pasado tendrán un accidente este año.
- ¿Qué porcentaje de los empleados sufrirá accidentes en los dos años?
 - ¿Qué porcentaje de los empleados sufrirá por lo menos un accidente en este periodo de dos años?
58. El departamento de recolección de impuestos de Estados Unidos en Dallas, preocupado por las declaraciones de impuestos fraudulentas, cree que la probabilidad de hallar una declaración de impuestos fraudulenta, dado que la declaración contiene deducciones que exceden el estándar, es 0.20. Dado que las deducciones no exceden el estándar, la probabilidad de una declaración fraudulenta disminuye a 0.02. Si 8% de las declaraciones exceden el estándar de deducciones, ¿cuál es la mejor estimación del porcentaje de declaraciones fraudulentas?
59. Una empresa petrolera compra una opción de tierra en Alaska. Los estudios geológicos preliminares asignaron las probabilidades previas siguientes.

$$P(\text{petróleo de alta calidad}) = 0.50$$

$$P(\text{petróleo de calidad media}) = 0.20$$

$$P(\text{que no haya petróleo}) = 0.30$$

- ¿Cuál es la probabilidad de hallar petróleo?
- Después de 200 pies de perforación en el primer pozo, se toma una prueba de suelo. Las probabilidades de hallar el tipo de suelo identificado en la prueba son las siguientes.

$$P(\text{suelo} \mid \text{petróleo de alta calidad}) = 0.20$$

$$P(\text{suelo} \mid \text{petróleo de calidad media}) = 0.80$$

$$P(\text{suelo} \mid \text{que no haya petróleo}) = 0.20$$

¿Cómo debe interpretar la empresa la prueba de suelo? ¿Cuáles son las probabilidades revisadas y cuáles son las nuevas probabilidades de hallar petróleo?

60. Las empresas que hacen negocios por Internet suelen obtener información acerca del visitante de un sitio Web a partir de los sitios visitados previamente. El artículo "Internet Marketing" (*Interfaces*, marzo/abril de 2001) describe cómo los datos sobre el flujo de clics en los sitios Web visitados se usan junto a un modelo de actualización Bayesiano para determinar el género de una persona que visita la Web. ParFore creó un sitio Web para la venta de equipo y ropa para golf. A los directivos de la empresa les gustaría que apareciera una determinada oferta para los visitantes del sexo femenino y otra oferta determinada para los visitantes del sexo masculino. En una muestra de visitas anteriores al sitio Web se sabe que 60% de las personas que visitan el sitio son hombres y 40% mujeres.
- ¿Cuál es la probabilidad previa de que el siguiente visitante del sitio Web sea mujer?
 - Suponga que el actual visitante de ParFore.com visitó previamente el sitio de la Web de Dillard, y que es tres veces más probable que ese sitio sea visitado por mujeres que por hombres. ¿Cuál es la probabilidad revisada de que el visitante actual de ParFore.com sea mujer? ¿Desplegaría la oferta que está dirigida más a hombres o a mujeres?

Caso problema

Los jueces del condado de Hamilton

Los jueces del condado de Hamilton llevan miles de casos cada año. En su inmensa mayoría la sentencia queda dictada. Sin embargo, en algunos casos hay apelaciones y algunas apelaciones revocan la sentencia. Kristen DelGuzzi de *The Cincinnati Enquirer* realizó, durante tres años, un estudio sobre los casos llevados por los jueces del condado de Hamilton. En la tabla 4.8 se muestran los resultados de los 182 908 casos llevados por 38 jueces en tribunales de primera instan-

TABLA 4.8 CASOS DESPACHADOS, APELADOS Y REVOCADOS EN LOS TRIBUNALES DEL CONDADO DE HAMILTON

archivo
en **CD**
Judge

Tribunal de primera instancia			
Juez	Casos despachados	Casos apelados	Casos revocados
Fred Cartolano	3 037	137	12
Thomas Crush	3 372	119	10
Patrick Dinkelacker	1 258	44	8
Timothy Hogan	1 954	60	7
Robert Kraft	3 138	127	7
William Mathews	2 264	91	18
William Morrissey	3 032	121	22
Norbert Nadel	2 959	131	20
Arthur Ney Jr.	3 219	125	14
Richard Niehaus	3 353	137	16
Thomas Nurre	3 000	121	6
John O'Connor	2 969	129	12
Robert Ruehlman	3 205	145	18
J. Howard Sundermann	955	60	10
Ann Marie Tracey	3 141	127	13
Ralph Winkler	3 089	88	6
Total	43 945	1762	199
Tribunal de relaciones domésticas			
Juez	Casos despachados	Casos apelados	Casos revocados
Penelope Cunningham	2 729	7	1
Patrick Dinkelacker	6 001	19	4
Deborah Gaines	8 799	48	9
Ronald Panioto	12 970	32	3
Total	30 499	106	17
Tribunal municipal			
Juez	Casos despachados	Casos apelados	Casos revocados
Mike Allen	6 149	43	4
Nadine Allen	7 812	34	6
Timothy Black	7 954	41	6
David Davis	7 736	43	5
Leslie Isaiah Gaines	5 282	35	13
Karla Grady	5 253	6	0
Deidra Hair	2 532	5	0
Dennis Helmick	7 900	29	5
Timothy Hogan	2 308	13	2
James Patrick Kenney	2 798	6	1
Joseph Luebbers	4 698	25	8
William Mallory	8 277	38	9
Melba Marsh	8 219	34	7
Beth Mattingly	2 971	13	1
Albert Mestemaker	4 975	28	9
Mark Painter	2 239	7	3
Jack Rosen	7 790	41	13
Mark Schweikert	5 403	33	6
David Stockdale	5 371	22	4
John A. West	2 797	4	2
Total	108 464	500	104

cia, tribunales de relaciones domésticas y tribunales municipales. Dos de los jueces (Dinkelacker y Hogan) no prestaron sus servicios en el mismo tribunal durante los tres años completos.

El objetivo del estudio de este periódico fue evaluar el trabajo de los jueces. Las apelaciones suelen ser el resultado de errores cometidos por los jueces, y el periódico deseaba saber qué jueces realizan bien su trabajo y qué jueces cometían demasiados errores. Se le solicita su ayuda para realizar el análisis de datos. Emplee sus conocimientos de probabilidad y de probabilidad condicional para ayudar a la clasificación de los jueces. Podrá analizar también la posibilidad de apelación y de revocación en los casos tratados en los distintos tribunales.

Informe administrativo

Elabore un informe con su clasificación de los jueces. Incluya un análisis de la posibilidad de apelación y de revocación del caso en los tres tribunales. Como mínimo su informe debe contener lo siguiente:

1. La probabilidad de que los casos sean apelados y revocados en los distintos tribunales.
2. La probabilidad, para cada juez, de que un caso sea apelado.
3. La probabilidad, para cada juez, de que un caso sea revocado.
4. La probabilidad, para cada juez, de revocación dada una apelación.
5. Clasifique a los jueces de cada tribunal de mejor a peor. Dé el criterio que usa y proporcione el fundamento que justifique su elección.



CAPÍTULO 5

Distribuciones de probabilidad discreta

CONTENIDO

LA ESTADÍSTICA EN LA PRÁCTICA: CITIBANK

- 5.1** VARIABLES ALEATORIAS
 - Variables aleatorias discretas
 - Variables aleatorias continuas
- 5.2** DISTRIBUCIONES DE PROBABILIDAD DISCRETA
- 5.3** VALOR ESPERADO Y VARIANZAS
 - Valor esperado
 - Varianza
- 5.4** DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL
 - Un experimento binomial
 - El problema de la tienda de ropa Martin Clothing Store

Uso de las tablas de probabilidades binomiales
Valor esperado y varianza en la distribución binomial

- 5.5** DISTRIBUCIÓN DE PROBABILIDAD DE POISSON
 - Un ejemplo con intervalos de tiempo
 - Un ejemplo con intervalos de longitud o de distancia

- 5.6** DISTRIBUCIÓN DE PROBABILIDAD HIPERGEOMÉTRICA

LA ESTADÍSTICA *en* LA PRÁCTICA

CITIBANK*

LONG ISLAND CITY, NUEVA YORK

Citibank, una división de Citigroup, proporciona una amplia gama de servicios financieros, que comprende cuentas de cheques y de ahorro, préstamos e hipotecas, seguros y servicios de inversión, todos dentro del marco de una estrategia única llamada Citibanking. Citibanking significa una identidad de marca consistente en todo el mundo, una oferta coherente de productos y servicios de calidad para el cliente. Citibanking permite al cliente disponer de dinero en cualquier momento, en cualquier parte y de la manera que lo desee. Ya sea que el cliente desee ahorrar para el futuro o solicitar un préstamo para hoy, lo puede hacer en Citibank.

Los cajeros automáticos de Citibank, localizados en los Citicard Banking Center (CBC), permiten al cliente hacer todas sus operaciones bancarias en un solo lugar con un simple toque de su dedo, 24 horas al día y 7 días a la semana. Más de 150 operaciones bancarias diferentes, desde depósitos hasta manejo de inversiones, pueden ser realizadas con facilidad. Los cajeros automáticos Citibanking son mucho más que un simple cajero automático y en la actualidad los clientes realizan en ellos 80% de sus transacciones.

Cada Citibank CBC opera como un sistema de espera en línea al que los clientes llegan en forma aleatoria a solicitar el servicio de uno de los cajeros automáticos. Si todos los cajeros automáticos están ocupados, debe esperar en la fila. Con periodicidad realizan estudios acerca de la capacidad de los CBC para determinar los tiempos de espera para el cliente y establecer si son necesarios más cajeros automáticos.

Los datos recolectados por Citibank muestran que la llegada aleatoria de los clientes sigue una distribución de probabilidad conocida como distribución de Poisson. Mediante la distribución de Poisson, Citibank calcula las pro-



Un vanguardista cajero automático de Citibank.

© Jeff Greenberg/Photo Edit.

babilidades de que llegue un número determinado de clientes a un CBC durante un determinado periodo y decidir cuál es el número de cajeros que necesita. Por ejemplo, sea x la cantidad de clientes que llega en un periodo de un minuto. Suponga que la tasa media de llegadas de clientes a un determinado CBC es dos clientes por minuto, la tabla siguiente da las probabilidades de que llegue un determinado número de clientes por minuto.

x	Probabilidad
0	0.1353
1	0.2707
2	0.2707
3	0.1804
4	0.0902
5 o más	0.0527

Las distribuciones de probabilidad discretas como la empleada por Citibank, son el tema de este capítulo. Además de la distribución de Poisson, verá las distribuciones binomial e hipergeométrica; conocerá también cómo emplear estas distribuciones de probabilidad para obtener información de utilidad.

*Los autores agradecen a Stacey Karter, Citibank, por proporcionarnos este artículo para *La estadística en práctica*.

En este capítulo se continúa con el estudio de la probabilidad introduciendo los conceptos de variable aleatoria y distribuciones de probabilidad. El punto sustancial de este capítulo son las distribuciones de probabilidad discreta de tres distribuciones de probabilidad discreta que serán estudiadas son: la binomial, la de Poisson y la hipergeométrica.

5.1

Variables aleatorias

En el capítulo 4 se definió el concepto de experimento con sus correspondientes resultados experimentales. Una variable aleatoria proporciona un medio para describir los resultados experimen-

Las variables aleatorias deben tomar valores numéricos.

VARIABLE ALEATORIA

Una **variable aleatoria** es una descripción numérica del resultado de un experimento.

tales empleando valores numéricos. Las variables aleatorias deben tomar valores numéricos. En efecto, una variable aleatoria asocia un valor numérico a cada uno de los resultados experimentales. El valor numérico de la variable aleatoria depende del resultado del experimento. Una variable aleatoria puede ser *discreta* o *continua*, depende del tipo de valores numéricos que asuma.

Variables aleatorias discretas

A una variable aleatoria que asuma ya sea un número finito de valores o una sucesión infinita de valores tales como 0, 1, 2, . . . , se le llama **variable aleatoria discreta**. Considere, por ejemplo, el siguiente experimento: un contador presenta el examen para certificarse como contador público. El examen tiene cuatro partes. Defina una variable aleatoria x como x = número de partes del examen aprobadas. Ésta es una variable aleatoria discreta porque puede tomar el número finito de valores 0, 1, 2, 3 o 4.

Para tener otro ejemplo de una variable aleatoria discreta considere el experimento de observar los automóviles que llegan a una caseta de peaje. La variable aleatoria que interesa es x = número de automóviles que llega a la caseta de peaje en un día. Los valores que puede tomar la variable aleatoria son los de la secuencia 0, 1, 2, etc. Así, x es una variable aleatoria discreta que toma uno de los valores de esta sucesión infinita.

Aunque los resultados de muchos experimentos se describen mediante valores numéricos, los de otros no. Por ejemplo, en una encuesta se le puede preguntar a una persona si recuerda el mensaje de un comercial de televisión. Este experimento tiene dos resultados: que la persona no recuerda el mensaje y que la persona recuerda el mensaje. Sin embargo, estos resultados se describen numéricamente definiendo una variable aleatoria x como sigue: sea $x = 0$ si la persona no recuerda el mensaje y sea $x = 1$ si la persona recuerda el mensaje. Los valores numéricos de esta variable son arbitrarios (podría haber usado 5 y 10), pero son aceptables de acuerdo con la definición de una variable aleatoria, es decir, x es una variable aleatoria porque proporciona una descripción numérica de los resultados del experimento.

En la tabla 5.1 aparecen algunos otros ejemplos de variables aleatorias discretas. Observe que en cada ejemplo la variable aleatoria discreta asume un número finito de valores o asume los valores de una secuencia infinita como 0, 1, 2, Este tipo de variables aleatorias discretas se estudia con detalle en este capítulo.

TABLA 5.1 EJEMPLOS DE VARIABLES ALEATORIAS DISCRETAS

Experimento	Variable aleatoria (x)	Valores posibles para la variable aleatoria
Llamar a cinco clientes	Número de clientes que hacen un pedido	0, 1, 2, 3, 4, 5
Inspeccionar un envío de 50 radios	Número de radios que tienen algún defecto	0, 1, 2, . . . , 49, 50
Hacerse cargo de un restaurante durante un día	Número de clientes	0, 1, 2, 3, . . .
Vender un automóvil	Sexo del cliente	0 si es hombre; 1 si es mujer

Variables aleatorias continuas

A una variable que puede tomar cualquier valor numérico dentro de un intervalo o colección de intervalos se le llama **variable aleatoria continua**. Los resultados experimentales basados en escalas de medición tales como tiempo, peso, distancia y temperatura pueden ser descritos por variables aleatorias continuas. Considere, por ejemplo, el experimento de observar las llamadas telefónicas que llegan a la oficina de atención de una importante empresa de seguros. La variable aleatoria que interesa es x = tiempo en minutos entre dos llamadas consecutivas. Esta variable aleatoria puede tomar cualquier valor en el intervalo $x \geq 0$. En efecto, x puede tomar un número infinito de valores, entre los que se encuentran valores como 1.26 minutos, 2.751 minutos, 4.3333 minutos, etc. Otro ejemplo, considere el tramo de 90 millas de una carretera entre Atlanta y Georgia. Para el servicio de ambulancia de emergencia en Atlanta, la variable aleatoria x es x = número de millas hasta el punto en que se localiza el siguiente accidente de tráfico en este tramo de la carretera. En este caso, x es una variable aleatoria continua que toma cualquier valor en el intervalo $0 \leq x \leq 90$. En la tabla 5.2 aparecen otros ejemplos de variables aleatorias continuas. Observe que cada ejemplo describe una variable aleatoria que toma cualquier valor dentro de un intervalo de valores. Las variables aleatorias continuas y sus distribuciones de probabilidad serán tema del capítulo 6.

TABLA 5.2 EJEMPLOS DE VARIABLES ALEATORIAS CONTINUAS

Experimento	Variable aleatoria (x)	Valores posibles para la variable aleatoria
Operar un banco	Tiempo en minutos entre la llegada de los clientes	$x \geq 0$
Llenar una lata de refresco (máx. 12.1 onzas)	Cantidad de onzas	$0 \leq x \leq 12.1$
Construir una biblioteca	Porcentaje del proyecto terminado en seis meses	$0 \leq x \leq 100$
Probar un proceso químico nuevo	Temperatura a la que tiene lugar la reacción deseada (min. 150°F; máx. 212°F)	$150 \leq x \leq 212$

NOTAS Y COMENTARIOS

Un modo de determinar si una variable aleatoria es discreta o continua es imaginar los valores de la variable aleatoria como puntos sobre un segmento de recta. Elegir dos puntos que representen valores de la variable aleatoria. Si todo el segmento de recta entre esos dos puntos representa también valores posibles para la variable aleatoria, entonces la variable aleatoria es continua.

Ejercicios

Métodos

1. Considere el experimento que consiste en lanzar una moneda dos veces.
 - a. Enumere los resultados experimentales.
 - b. Defina una variable aleatoria que represente el número de caras en los dos lanzamientos.
 - c. Dé el valor que la variable aleatoria tomará en cada uno de los resultados experimentales.
 - d. ¿Es una variable aleatoria discreta o continua?

2. Considere el experimento que consiste en un empleado que arma un producto.
 - a. Defina la variable aleatoria que represente el tiempo en minutos requerido para armar el producto.
 - b. ¿Qué valores toma la variable aleatoria?
 - c. ¿Es una variable aleatoria discreta o continua?

Aplicaciones

Autoexamen

3. Tres estudiantes agendan entrevistas para un empleo de verano en el Brookwood Institute. En cada caso el resultado de la entrevista será una oferta de trabajo o ninguna oferta. Los resultados experimentales se definen en términos de los resultados de las tres entrevistas.
 - a. Enumere los resultados experimentales.
 - b. Defina una variable aleatoria que represente el número de ofertas de trabajo. ¿Es una variable aleatoria continua?
 - c. Dé el valor de la variable aleatoria que corresponde a cada uno de los resultados experimentales.
4. Suponga que conoce la tasa hipotecaria de 12 instituciones de préstamo. La variable aleatoria que interesa es el número de las instituciones de préstamo en este grupo que ofrecen una tasa fija a 30 años de 8.5% o menos. ¿Qué valores toma esta variable aleatoria?
5. Para realizar cierto análisis de sangre, los técnicos laboratoristas tienen que llevar a cabo dos procedimientos. En el primero requieren uno o dos pasos y en el segundo requieren uno, dos o tres pasos.
 - a. Enumere los resultados experimentales correspondientes a este análisis de sangre.
 - b. Si la variable aleatoria que interesa es el número de pasos requeridos en todo el análisis (los dos procedimientos), dé los valores que toma la variable aleatoria en cada uno de los resultados experimentales.
6. A continuación se da una serie de experimentos y su variable aleatoria correspondiente. En cada caso determine qué valores toma la variable aleatoria y diga si se trata de una variable aleatoria discreta o continua.

Experimento	Variable aleatoria (x)
a. Hacer un examen con 20 preguntas	Número de preguntas contestadas correctamente
b. Observar los automóviles que llegan a una caseta de peaje en 1 hora	Número de automóviles que llegan a la caseta de peaje
c. Revisar 50 declaraciones de impuestos	Número de declaraciones que tienen algún error
d. Observar trabajar a un empleado	Número de horas no productivas en una jornada de 8 horas
e. Pesar un envío	Número de libras

5.2

Distribuciones de probabilidad discreta

La **distribución de probabilidad** de una variable aleatoria describe cómo se distribuyen las probabilidades entre los valores de la variable aleatoria. En el caso de una variable aleatoria discreta x , la distribución de probabilidad está definida por una **función de probabilidad**, denotada por $f(x)$. La función de probabilidad da la probabilidad de cada valor de la variable aleatoria.

Como ejemplo de una variable aleatoria discreta y de su distribución de probabilidad, considere las ventas de automóviles en DiCarlo Motors en Saratoga, Nueva York. Durante los últimos 300 días de operación, los datos de ventas muestran que hubo 57 días en los que no se vendió ningún automóvil, 117 días en los que se vendió 1 automóvil, 72 días en los que se vendieron 2 automóviles, 42 días en los que se vendieron 3 automóviles, 12 días en los que se vendieron 4 automóviles y 3 días en los que se vendieron 5 automóviles. Suponga que considera el experimento

de seleccionar un día de operación en DiCarlo Motors y se define la variable aleatoria de interés como x = número de automóviles vendidos en un día. De acuerdo con datos del pasado, se sabe que x es una variable aleatoria discreta que puede tomar los valores 0, 1, 2, 3, 4 o 5. En la notación de funciones de probabilidad $f(0)$ da la probabilidad de vender 0 automóviles, $f(1)$ da la probabilidad de vender 1 automóvil, y así en lo sucesivo. Como los datos del pasado indican que en 54 de 300 días se vendieron 0 automóviles, a $f(0)$ se le asigna el valor $54/300 = 0.18$, lo que significa que la probabilidad de que se vendan 0 automóviles en un día es 0.18. De manera similar, como en 117 de los 300 días se vendió un automóvil, a $f(1)$ se le asigna el valor $117/300 = 0.39$, que significa que la probabilidad de que se venda exactamente 1 automóvil en un día es 0.39. Continuando de esta manera con los demás valores de la variable aleatoria, se obtienen los valores de $f(2)$, $f(3)$, $f(4)$ y $f(5)$, valores que se muestran en la tabla 5.3, que es la distribución de probabilidad para el número de automóviles vendidos en un día en DiCarlo Motors.

Una ventaja importante de definir una variable aleatoria y su correspondiente distribución de probabilidad es que una vez que se conoce la distribución de probabilidad, es relativamente fácil determinar la probabilidad de diversos eventos que pueden ser útiles para tomar decisiones. Por ejemplo, empleando la distribución de probabilidad de DiCarlo Motors, tabla 5.3, se observa que el número de automóviles que es más probable vender en un día es 1, ya que es $f(1) = 0.39$. Además se observa que la probabilidad de vender tres o más automóviles en un día es $f(3) + f(4) + f(5) = 0.14 + 0.04 + 0.01 = 0.19$. Estas probabilidades, junto con otras que pueden interesar para tomar decisiones, proporcionan información que sirve de ayuda al encargado de la toma de decisiones para entender la venta de automóviles en DiCarlo Motors.

Al elaborar una función de probabilidad para una variable aleatoria discreta, deben satisfacerse las dos condiciones siguientes.

Estas condiciones son análogas a los dos requerimientos básicos, presentados en el capítulo 4, para asignar probabilidades a los resultados experimentales.

CONDICIONES REQUERIDAS PARA UNA FUNCIÓN DE PROBABILIDAD DISCRETA

$$f(x) \geq 0 \quad (5.1)$$

$$\sum f(x) = 1 \quad (5.2)$$

En la tabla 5.3 se observa que las probabilidades de la variable aleatoria x satisfacen la ecuación (5.1); para todos los valores de x , $f(x)$ es mayor o igual que 0; además, como estas probabilidades suman 1, también se satisface la ecuación (5.2). Por tanto, la función de probabilidad de DiCarlo Motors es una función de probabilidad discreta válida.

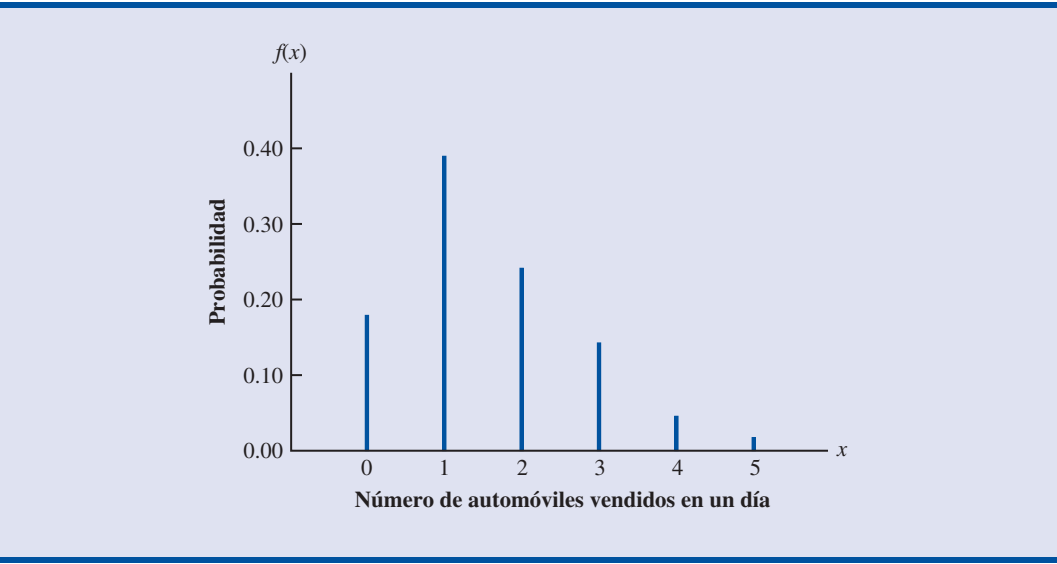
Las distribuciones de probabilidad también se representan gráficamente. En la figura 5.1, en el eje horizontal aparecen los valores de la variable aleatoria x para el caso de DiCarlo Motors y en el eje vertical aparecen las probabilidades correspondientes a estos valores.

Además de tablas y gráficas, para describir las funciones de probabilidad se suele usar una fórmula que da el valor de la función de probabilidad, $f(x)$, para cada valor x . El ejemplo más sencillo

TABLA 5.3 DISTRIBUCIÓN DE PROBABILIDAD PARA EL NÚMERO DE AUTOMÓVILES VENDIDOS EN UN DÍA EN DICARLO MOTORS

x	$f(x)$
0	0.18
1	0.39
2	0.24
3	0.14
4	0.04
5	0.01
Total	1.00

FIGURA 5.1 REPRESENTACIÓN GRÁFICA DE LA DISTRIBUCIÓN DE PROBABILIDAD DEL NÚMERO DE AUTOMÓVILES VENDIDOS EN UN DÍA EN DICARLO MOTORS



de una distribución de probabilidad discreta dada mediante una fórmula es la **distribución de probabilidad uniforme discreta**. Su función de probabilidad está definida por la ecuación (5.3).

FUNCIÓN DE PROBABILIDAD UNIFORME DISCRETA

$$f(x) = 1/n$$

(5.3)

donde

n = número de valores que puede tomar la variable aleatoria.

Por ejemplo, si en el experimento que consiste en lanzar un dado se define una variable aleatoria x como el número de puntos en la cara del dado que cae hacia arriba. En este experimento la variable aleatoria toma $n = 6$ valores; $x = 1, 2, 3, 4, 5, 6$. Por tanto, la función de probabilidad de esta variable aleatoria uniforme discreta es

$$f(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

Los valores de la variable aleatoria con sus probabilidades correspondientes se presentan a continuación.

x	$f(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Otro ejemplo, la variable aleatoria x tiene la siguiente distribución de probabilidad discreta.

x	$f(x)$
1	1/10
2	2/10
3	3/10
4	4/10

Esta distribución de probabilidad se define mediante la fórmula

$$f(x) = \frac{x}{10} \quad \text{para } x = 1, 2, 3 \text{ o } 4$$

Si evalúa $f(x)$ para un valor determinado de la variable aleatoria obtiene la probabilidad correspondiente. Por ejemplo, con la función de probabilidad dada arriba se ve que $f(2) = 2/10$ da la probabilidad de que la variable aleatoria tome el valor 2.

Las funciones de probabilidad discreta más empleadas suelen especificarse mediante fórmulas. Tres casos importantes son las distribuciones binomial, de Poisson e hipergeométrica; estas distribuciones se estudian más adelante en este capítulo

Ejercicios

Métodos

7. A continuación se presenta la distribución de probabilidad de una variable aleatoria x .

x	$f(x)$
20	0.20
25	0.15
30	0.25
35	0.40

- ¿Es válida esta distribución de probabilidad?
- ¿Cuál es la probabilidad de que $x = 30$?
- ¿Cuál es la probabilidad de que x sea menor o igual que 25?
- ¿Cuál es la probabilidad de que x sea mayor que 30?

Aplicaciones

8. Los datos siguientes se obtuvieron contando el número de salas de operaciones de un hospital que fueron usadas en un periodo de 20 días. Tres de estos 20 días sólo se usó una sala de operaciones, cinco de estos 20 días se usaron dos, ocho de estos 20 días se usaron tres salas de operaciones y cuatro de estos 20 días se usaron las cuatro salas de operaciones del hospital.
- Use el método de las frecuencias relativas para elaborar una distribución de probabilidad para el número de salas de operaciones usadas en un día.
 - Elabore una gráfica a partir de la distribución de probabilidad.
 - Muestre que la distribución de probabilidad elaborada satisface las condiciones requeridas para una distribución de probabilidad.

Autoexamen

Autoexamen

9. En Estados Unidos 38% de los niños de cuarto grado no pueden leer un libro adecuado a su edad. La tabla siguiente muestra, de acuerdo con las edades, el número de niños que tienen problemas de lectura. La mayoría de estos niños tienen problemas de lectura que debieron ser detectados y corregidos antes del tercer grado.

Edad	Número de niños
6	37 369
7	87 436
8	160 840
9	239 719
10	286 719
11	306 533
12	310 787
13	302 604
14	289 168

Si desea tomar una muestra de niños que tienen problemas de lectura para que participen en un programa que mejora las habilidades de lectura. Sea x la variable aleatoria que indica la edad de un niño tomado en forma aleatoria.

- a. Con estos datos elabore una distribución de probabilidad para x . Especifique los valores de la variable aleatoria y los correspondientes valores de la función de probabilidad $f(x)$.
 - b. Trace la gráfica de esta distribución de probabilidad.
 - c. Muestre que la distribución de probabilidad satisface las ecuaciones (5.1) y (5.2).
10. En la tabla 5.4 se muestra la distribución de frecuencias porcentuales para las puntuaciones dadas a la satisfacción con el trabajo por una muestra de directivos en sistemas de información de nivel alto y de nivel medio. Las puntuaciones van de 1 (muy insatisfecho) a 5 (muy satisfecho).

TABLA 5.4 DISTRIBUCIÓN DE FRECUENCIA PORCENTUAL DE LAS PUNTUACIONES DADAS POR DIRECTIVOS DE NIVEL ALTO Y DE NIVEL MEDIO A LA SATISFACCIÓN CON EL TRABAJO

Puntuación de la satisfacción con el trabajo	Directivos de alto nivel	Directivos de nivel medio
1	5	4
2	9	10
3	3	12
4	42	46
5	41	28

- a. Elabore una distribución de probabilidad con las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel alto.
 - b. Elabore una distribución de probabilidad con las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel medio.
 - c. ¿Cuál es la probabilidad de que un ejecutivo de nivel alto dé una puntuación de 4 o 5 a su satisfacción con el trabajo?
 - d. ¿Cuál es la probabilidad de que un ejecutivo de nivel medio esté muy satisfecho?
 - e. Haga una comparación entre la satisfacción con el trabajo de los ejecutivos de nivel alto y la que tienen los ejecutivos de nivel medio.
11. Un técnico da servicio a máquinas franquadoras de empresas en el área de Phoenix. El servicio puede durar 1, 2, 3 o 4 horas dependiendo del tipo de falla. Los distintos tipos de fallas se presentan aproximadamente con la misma frecuencia.

- a. Elabore una distribución de probabilidad de las duraciones de los servicios.
 - b. Elabore una gráfica de la distribución de probabilidad.
 - c. Muestre que la distribución de probabilidad que ha elaborado satisface las condiciones requeridas para ser una distribución de probabilidad discreta.
 - d. ¿Cuál es la probabilidad de que un servicio dure tres horas?
 - e. Acaba de llegar una solicitud de servicio y no se sabe cuál es el tipo de falla. Son las 3:00 p.m. y los técnicos de servicio salen a las 5:00 de la tarde. ¿Cuál es la probabilidad de que el técnico de servicio tenga que trabajar horas extras para reparar la máquina hoy?
12. El jefe del departamento de admisión de una universidad calcula subjetivamente una distribución de probabilidad para x , el número de estudiantes que ingresarán en la universidad. A continuación se presenta esta distribución de probabilidad.

x	$f(x)$
1000	0.15
1100	0.20
1200	0.30
1300	0.25
1400	0.10

- a. ¿Es válida esta distribución de probabilidad? Explique.
 - b. ¿Cuál es la probabilidad de que ingresen 1200 o menos estudiantes? Explique.
13. Un psicólogo encuentra que el número de sesiones necesarias para ganarse la confianza de un paciente es 1, 2 o 3. Sea x la variable aleatoria que representa el número de sesiones necesarias para ganarse la confianza de un paciente. Se ha propuesto la función de probabilidad siguiente.

$$f(x) = \frac{x}{6} \quad \text{para } x = 1, 2 \text{ o } 3$$

- a. ¿Es válida esta función de probabilidad? Explique.
 - b. ¿Cuál es la probabilidad de que se necesiten exactamente 2 sesiones para ganarse la confianza del paciente?
 - c. ¿De que se necesiten por lo menos 2 sesiones para ganarse la confianza del paciente?
14. La tabla siguiente es una distribución parcial de probabilidades para las ganancias proyectadas de MRA Company (x ganancias en miles de dólares) durante el primer año de operación (los valores negativos indican pérdida).

x	$f(x)$
-100	0.10
0	0.20
50	0.30
100	0.25
150	0.10
200	

- a. ¿Cuál es el valor adecuado para $f(200)$? ¿Qué interpretación le da a este valor?
- b. ¿Cuál es la probabilidad de que la empresa sea rentable?
- c. ¿Cuál es la probabilidad de que la empresa gane por lo menos \$100 000?

5.3

Valor esperado y varianzas

Valor esperado

El **valor esperado**, o media, de una variable aleatoria es una medida de la localización central de la variable aleatoria. A continuación se da la fórmula para obtener el valor esperado de una variable aleatoria x .

El valor esperado es un promedio ponderado de los valores que toma la variable aleatoria. Los pesos son las probabilidades.

VALOR ESPERADO DE UNA VARIABLE ALEATORIA DISCRETA

$$E(x) = \mu = \sum xf(x)$$

(5.4)

Las dos notaciones $E(x)$ y μ se usan para denotar el valor esperado de una variable aleatoria x . La ecuación (5.4) indica que para calcular el valor esperado de una variable aleatoria discreta se multiplica cada valor de la variable aleatoria por su probabilidad correspondiente $f(x)$ y después se suman estos productos. Usando el ejemplo de la sección 5.2 sobre las ventas de automóviles en DiCarlo Motors, en la tabla 5.5 se muestra cómo se calcula el valor esperado del número de automóviles vendidos en un día. La suma de las entradas en la columna $xf(x)$ indica que el valor esperado es 1.50 automóviles por día. Por tanto, aunque se sabe que en un día las ventas pueden ser de 0, 1, 2, 3, 4 o 5 automóviles, DiCarlo prevé que a la larga se venderán 1.50 automóviles por día. Si en un mes hay 30 días de operación, el valor esperado, 1.50, se emplea para pronosticar que las ventas promedio mensuales serán de $30(1.5) = 45$ automóviles.

El valor esperado no tiene que ser un valor que pueda tomar la variable aleatoria.

Varianza

Aunque el valor esperado proporciona el valor medio de una variable aleatoria, también suele ser necesaria una medida de la variabilidad o dispersión. Así como en el capítulo 3 se usó la **varianza** para resumir la variabilidad de los datos, ahora se usa la **varianza** para resumir la variabilidad en los valores de la variable aleatoria. A continuación se da la fórmula para calcular la

La varianza es un promedio ponderado de los cuadrados de las desviaciones de una variable aleatoria de su media. Los pesos son las probabilidades.

VARIANZA DE UNA VARIABLE ALEATORIA DISCRETA

$$\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 f(x)$$

(5.5)

TABLA 5.5 CÁLCULO DEL VALOR ESPERADO PARA EL NÚMERO DE AUTOS QUE SE VENDEN EN UN DÍA EN DICARLO MOTORS

x	$f(x)$	$xf(x)$
0	0.18	$0(.18) = 0.00$
1	0.39	$1(.39) = 0.39$
2	0.24	$2(.24) = 0.48$
3	0.14	$3(.14) = 0.42$
4	0.04	$4(.04) = 0.16$
5	0.01	$5(.01) = 0.05$
		1.50
$E(x) = \mu = \sum xf(x)$		

TABLA 5.6 CÁLCULO DE LA VARIANZA PARA EL NÚMERO DE AUTOS QUE SE VENDEN EN UN DÍA EN DICARLO MOTORS

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	$0 - 1.50 = -1.50$	2.25	0.18	$2.25(0.18) = 0.4050$
1	$1 - 1.50 = -0.50$	0.25	0.39	$0.25(0.39) = 0.0975$
2	$2 - 1.50 = 0.50$	0.25	0.24	$0.25(0.24) = 0.0600$
3	$3 - 1.50 = 1.50$	2.25	0.14	$2.25(0.14) = 0.3150$
4	$4 - 1.50 = 2.50$	6.25	0.04	$6.25(0.04) = 0.2500$
5	$5 - 1.50 = 3.50$	12.25	0.01	$12.25(0.01) = 0.1225$
				1.2500

$\sigma^2 = \sum (x - \mu)^2 f(x)$

varianza de una variable aleatoria. Como indica la ecuación (5.5), una parte esencial de la fórmula de la varianza es la desviación $x - \mu$, la cual mide qué tan alejado del valor esperado, o media μ , se encuentra un valor determinado de la variable aleatoria. Para calcular la varianza de una variable aleatoria, estas desviaciones se elevan al cuadrado y después se ponderan con el correspondiente valor de la función de probabilidad. A la suma de estas desviaciones al cuadrado, ponderadas, se le conoce como *varianza*. Para denotar la varianza de una variable aleatoria se usan las notaciones $\text{Var}(x)$ y σ^2 .

En la tabla 5.6 aparece en forma resumida el cálculo de la varianza de la distribución de probabilidad del número de automóviles vendidos en un día en DiCarlo Motors. Como ve, la varianza es 1.25. La **desviación estándar**, σ , se define como la raíz cuadrada positiva de la varianza. Por tanto, la desviación estándar del número de automóviles vendidos en un día es

$$\sigma = \sqrt{1.25} = 1.118$$

La desviación estándar se mide en las mismas unidades que la variable aleatoria ($\sigma = 1.1180$ automóviles) y por tanto suele preferirse para describir la variabilidad de una variable aleatoria. La varianza σ^2 se mide en unidades al cuadrado por lo que es más difícil de interpretar.

Ejercicios

Métodos

15. La tabla siguiente muestra la distribución de probabilidad de una variable aleatoria x .

x	$f(x)$
3	0.25
6	0.50
9	0.25

- Calcule $E(x)$, el valor esperado de x .
- Calcule σ^2 , la varianza de x .
- Calcule σ , la desviación estándar de x .

Autoexamen

16. La tabla siguiente muestra la distribución de probabilidad de una variable aleatoria y .

y	$f(y)$
2	0.20
4	0.30
7	0.40
8	0.10

- Calcule $E(y)$.
- Calcule $\text{Var}(y)$ y σ .

Aplicaciones

17. Una ambulancia de voluntarios realiza de 0 a 5 servicios por día. A continuación se presenta la distribución de probabilidad de los servicios por día.

Número de servicios	Probabilidad	Número de servicios	Probabilidad
0	0.10	3	0.20
1	0.15	4	0.15
2	0.30	5	0.10

- ¿Cuál es el valor esperado del número de servicios?
- ¿Cuál es la varianza del número de servicios? ¿Cuál es la desviación estándar?

Autoexamen

18. Los datos siguientes son el número de recámaras en casas rentadas y en casas propias en ciudades centrales de Estados Unidos (www.census.gov, 31 de marzo de 2003).

Recámaras	Número de casas (en miles)	
	Rentadas	Propias
0	547	23
1	5012	541
2	6100	3832
3	2644	8690
4 o más	557	3783

- Defina una variable aleatoria x = número de recámaras en casas rentadas y elabore una distribución de probabilidad para esta variable. ($x = 4$ representará 4 recámaras o más.)
 - Calcule el valor esperado y la varianza del número de recámaras en casas rentadas.
 - Defina una variable aleatoria y = número de recámaras en casas propias y elabore una distribución de probabilidad para esta variable. ($y = 4$ representará 4 recámaras o más.)
 - Calcule el valor esperado y la varianza del número de recámaras en casas propias.
 - ¿Qué observaciones resultan al comparar el número de recámaras en casas rentadas y en casas propias?
19. La National Basketball Association (NBA) lleva diversas estadísticas de cada equipo. Dos se refieren al porcentaje de tiros de campo hechos por un equipo y el porcentaje de tiros de tres puntos hechos por un equipo. En parte de la temporada del 2004, el registro de tiros de los 29 equipos de la NBA indicaba que la probabilidad de anotar dos puntos en un tiro de campo era 0.44, y que la probabilidad de anotar tres puntos en un tiro de tres puntos era 0.34 (www.nba.com, 3 de enero de 2004).

- a. ¿Cuál es el valor esperado para un tiro de dos puntos de estos equipos?
 - b. ¿Cuál es el valor esperado para un tiro de tres puntos de estos equipos?
 - c. Si la probabilidad de hacer un tiro de dos puntos es mayor que la probabilidad de hacer uno de tres puntos, ¿por qué los entrenadores permiten a algunos jugadores hacer un tiro de tres puntos si tienen oportunidad? Use el valor esperado para explicar su respuesta.
20. A continuación se presenta la distribución de probabilidad para los daños pagados por una empresa de seguros para automóviles, en seguros contra choques.

Pago	Probabilidad
0	0.85
500	0.04
1 000	0.04
3 000	0.03
5 000	0.02
8 000	0.01
10 000	0.01

- a. Use el pago esperado para determinar la prima en el seguro de choques que le permitirá a la empresa cubrir los gastos.
 - b. La empresa de seguros cobra una tasa anual de \$520 por la cobertura de choques. ¿Cuál es el valor esperado de un seguro de choques para un asegurado? (*Indicación:* son los pagos esperados de la empresa menos el costo de cobertura.) ¿Por qué compran los asegurados un seguro de choques con este valor esperado?
21. La siguiente distribución de probabilidad sobre puntuaciones dadas a la satisfacción con el trabajo por una muestra de directivos de alto nivel y de nivel medio en sistemas de la información va desde 1 (muy insatisfecho) hasta 5 (muy satisfecho).

Puntuación de la satisfacción con el trabajo	Probabilidad	
	Directivo de nivel alto	Directivo de nivel medio
1	0.05	0.04
2	0.09	0.10
3	0.03	0.12
4	0.42	0.46
5	0.41	0.28

- a. ¿Cuál es el valor esperado en las puntuaciones dadas a la satisfacción con el trabajo por los ejecutivos de nivel alto?
 - b. ¿Cuál es el valor esperado en las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel medio?
 - c. Calcule la varianza de las puntuaciones dadas a la satisfacción con el trabajo por los directivos de nivel medio.
 - d. Calcule la desviación estándar de las puntuaciones dadas a la satisfacción con el trabajo en las dos distribuciones de probabilidad.
 - e. Compare la satisfacción con el trabajo de los directivos de alto nivel con la que tienen los directivos de nivel medio.
22. La demanda de un producto de una empresa varía enormemente de mes a mes. La distribución de probabilidad que se presenta en la tabla siguiente, basada en los datos de los dos últimos años, muestra la demanda mensual de la empresa.

Demanda unitaria	Probabilidad
300	0.20
400	0.30
500	0.35
600	0.15

- a. Si la empresa basa las órdenes mensuales en el valor esperado de la demanda mensual, ¿cuál será la cantidad ordenada mensualmente por la empresa para este producto?
 - b. Suponga que cada unidad demandada genera \$70 de ganancia y que cada unidad ordenada cuesta \$50. ¿Cuánto ganará o perderá la empresa en un mes si coloca una orden con base en su respuesta al inciso a y la demanda real de este artículo es de 300 unidades?
23. El estudio 2002 New York City Housing and Vacancy Survey indicó que había 59 324 viviendas con renta controlada y 236 263 unidades con renta estabilizada construidas en 1947 o después. A continuación se da la distribución de probabilidad para el número de personas que viven en estas unidades (www.census.gov, 12 de enero de 2004).

Número de personas	Renta controlada	Renta estabilizada
1	0.61	0.41
2	0.27	0.30
3	0.07	0.14
4	0.04	0.11
5	0.01	0.03
6	0.00	0.01

- a. ¿Cuál es el valor esperado para el número de personas que viven en cada tipo de unidad?
 - b. ¿Cuál es la varianza para el número de personas que viven en cada tipo de unidad?
 - c. Haga comparaciones entre el número de personas que viven en una unidad de renta controlada y el número de personas que viven en una unidad de renta estabilizada.
24. J. R. Ryland Computer Company está considerando hacer una expansión a la fábrica para empezar a producir una nueva computadora. El presidente de la empresa debe determinar si hacer un proyecto de expansión a mediana gran escala. La demanda del producto nuevo es incierta, la cual, para los fines de planeación puede ser demanda pequeña, mediana o grande. Las probabilidades estimadas para la demanda son 0.20, 0.50 y 0.30, respectivamente. Con x y y representando ganancia anual en miles de dólares, los encargados de planeación en la empresa elaboraron el siguiente pronóstico de ganancias para los proyectos de expansión a mediana y gran escala.

		Ganancia con la expansión a mediana escala		Ganancia con la expansión a gran escala	
		x	$f(x)$	y	$f(y)$
Demanda	Baja	50	0.20	0	0.20
	Mediana	150	0.50	100	0.50
	Alta	200	0.30	300	0.30

- a. Calcule el valor esperado de las ganancias correspondientes a las dos alternativas de expansión. ¿Cuál de las decisiones se prefiere para el objetivo de maximizar la ganancia esperada?
- b. Calcule la varianza de las ganancias correspondientes a las dos alternativas de expansión. ¿Cuál de las decisiones se prefiere para el objetivo de minimizar el riesgo o la incertidumbre?

5.4

Distribución de probabilidad binomial

La distribución de probabilidad binomial es una distribución de probabilidad que tiene muchas aplicaciones. Está relacionada con un experimento de pasos múltiples al que se le llama experimento binomial.

Un experimento binomial

Un **experimento binomial** tiene las cuatro propiedades siguientes.

PROPIEDADES DE UN EXPERIMENTO BINOMIAL

1. El experimento consiste en una serie de n ensayos idénticos.
2. En cada ensayo hay dos resultados posibles. A uno de estos resultados se le llama *éxito* y al otro se le llama *fracaso*.
3. La probabilidad de éxito, que se denota p , no cambia de un ensayo a otro. Por ende, la probabilidad de fracaso, que se denota $1 - p$, tampoco cambia de un ensayo a otro.
4. Los ensayos son independientes.

Jacob Bernoulli (1654-1705), el primero de la familia Bernoulli de matemáticos suizos, publicó un tratado sobre probabilidad que contenía la teoría de las permutaciones y de las combinaciones, así como el teorema del binomio.

Si se presentan las propiedades 2, 3 y 4, se dice que los ensayos son generados por un proceso de Bernoulli. Si, además, se presenta la propiedad 1, se trata de un experimento binomial. En la figura 5.2 se presenta una sucesión de éxitos y fracasos de un experimento binomial con ocho ensayos.

En un experimento binomial lo que interesa es el *número de éxitos en n ensayos*. Si x denota el número de éxitos en n ensayos, es claro que x tomará los valores 0, 1, 2, 3, ..., n . Dado que el número de estos valores es finito, x es una variable aleatoria *discreta*. A la distribución de probabilidad correspondiente a esta variable aleatoria se le llama **distribución de probabilidad binomial**. Por ejemplo, considere el experimento que consiste en lanzar una moneda cinco veces y observar si la cara de la moneda que cae hacia arriba es cara o cruz. Suponga que se desea contar el número de caras que aparecen en los cinco lanzamientos. ¿Presenta este experimento las propiedades de un experimento binomial? ¿Cuál es la variable aleatoria que interesa? Observe que:

1. El experimento consiste en cinco ensayos idénticos; cada ensayo consiste en lanzar una moneda.
2. En cada ensayo hay dos resultados posibles: cara o cruz. Se puede considerar cara como éxito y cruz como fracaso.
3. La probabilidad de éxito y la probabilidad de fracaso son iguales en todos los ensayos, siendo $p = 0.5$ y $1 - p = 0.5$.
4. Los ensayos o lanzamientos son independientes porque al resultado de un ensayo no afecta a lo que pase en los otros ensayos o lanzamientos.

FIGURA 5.2 UNA POSIBLE SUCESIÓN DE ÉXITOS Y FRACASOS EN UN EXPERIMENTO BINOMIAL DE OCHO ENSAYOS

<i>Propiedad 1:</i> El experimento consiste en $n = 8$ ensayos idénticos.	
<i>Propiedad 2:</i> En cada ensayo se obtiene como resultado un éxito o un fracaso.	
Ensayos	→ 1 2 3 4 5 6 7 8
Resultados	→ S F F S S F S S

Por tanto, se satisfacen las propiedades de un experimento binomial. La variable aleatoria que interesa es $x =$ número de caras que aparecen en cinco ensayos. En este caso, x puede tomar los valores 0, 1, 2, 3, 4 o 5.

Otro ejemplo, considere a un vendedor de seguros que visita a 10 familias elegidas en forma aleatoria. El resultado correspondiente de la visita a cada familia se clasifica como éxito si la familia compra un seguro y como fracaso si la familia no compra ningún seguro. Por experiencia, el vendedor sabe que la probabilidad de que una familia tomada aleatoriamente compre un seguro es 0.10. Al revisar las propiedades de un experimento binomial aparece que:

1. El experimento consiste en 10 ensayos idénticos; cada ensayo consiste en visitar a una familia.
2. En cada ensayo hay dos resultados posibles: la familia compra un seguro (éxito) o la familia no compra ningún seguro (fracaso).
3. Las probabilidades de que haya compra y de que no haya compra se supone que son iguales en todas las visitas, siendo $p = 0.10$ y $1 - p = 0.90$.
4. Los ensayos son independientes porque las familias se eligen en forma aleatoria.

Como estos cuatro puntos se satisfacen, este ejemplo es un experimento binomial. La variable aleatoria que interesa es el número de ventas al visitar a las 10 familias. En este caso los valores que puede tomar x son 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 y 10.

La propiedad 3 de un experimento binomial se llama *suposición de estacionaridad* y algunas veces se confunde con la propiedad 4, independencia de los ensayos. Para ver la diferencia entre estas dos propiedades, reconsidere el caso del vendedor que visita a las familias para venderles un seguro. Si a medida que el día avanza, el vendedor se va cansando y va perdiendo entusiasmo, la probabilidad de éxito puede disminuir, por ejemplo, a 0.05 en la décima llamada. En tal caso la propiedad 3 (estacionaridad) no se satisface, y no se tiene un experimento binomial. Incluso si la propiedad 4 se satisface —en cada familia la decisión de comprar o no se hizo de manera independiente— si no se satisface la propiedad 3, no se trata de un experimento binomial.

En las aplicaciones de los experimentos binomiales se emplea una fórmula matemática llamada *función de probabilidad binomial* que sirve para calcular la probabilidad de x éxitos en n ensayos. Empleando los conceptos de probabilidad presentados en el capítulo 4, se mostrará, en el contexto de un ilustrativo problema, cómo se desarrolla la fórmula.

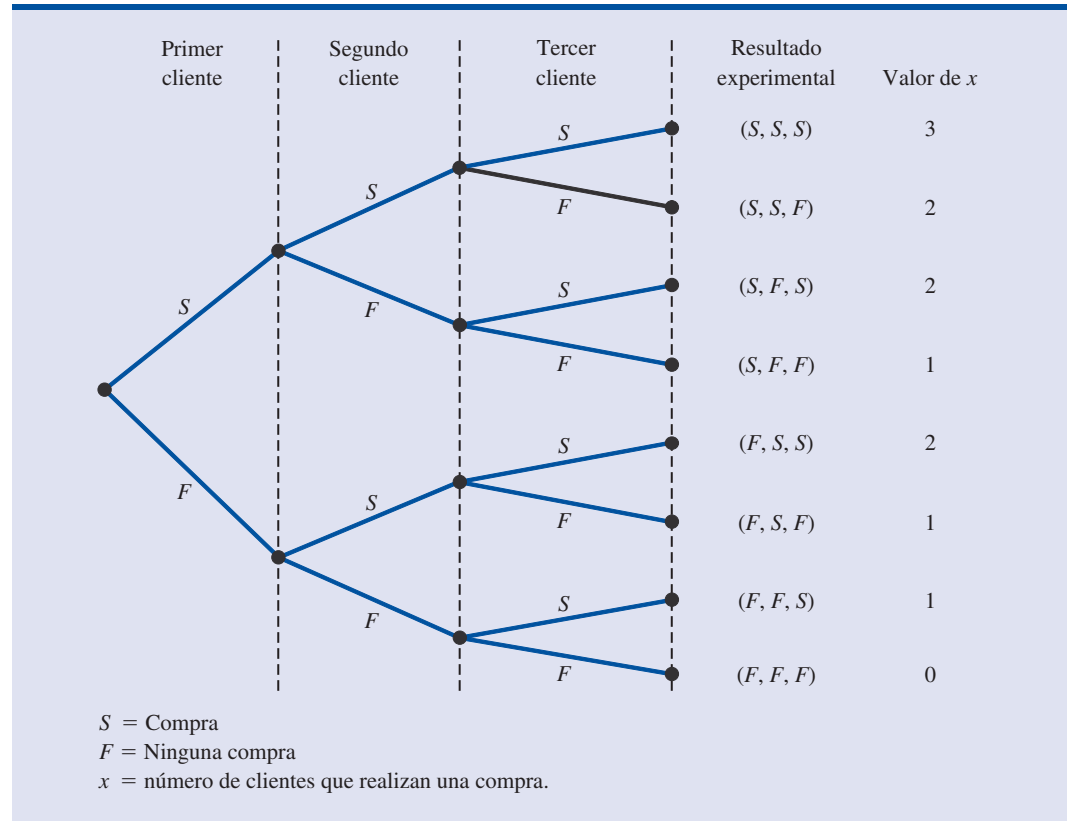
El problema de la tienda de ropa Martin Clothing Store

Considere las decisiones de compra de los próximos tres clientes que lleguen a la tienda de ropa Martin Clothing Store. De acuerdo con la experiencia, el gerente de la tienda estima que la probabilidad de que un cliente realice una compra es 0.30. ¿Cuál es la probabilidad de que dos de los próximos tres clientes realicen una compra?

Un diagrama de árbol (figura 5.3), permite advertir que el experimento de observar a los tres clientes para ver si cada uno de ellos decide realizar una compra tiene ocho posibles resultados. Entonces, si S denota éxito (una compra) y F fracaso (ninguna compra), lo que interesa son los resultados experimentales en los que haya dos éxitos (decisiones de compra) en los tres ensayos. A continuación verifique que el experimento de las tres decisiones de compra es un experimento binomial. Al verificar los cuatro requerimientos de un experimento binomial, se observa que:

1. Es posible describir el experimento como una serie de tres ensayos idénticos, un ensayo por cada uno de los tres clientes que llegan a la tienda.
2. Cada ensayo tiene dos posibles resultados: el cliente hace una compra (éxito) o el cliente no hace ninguna compra (fracaso).
3. La probabilidad de que el cliente haga una compra (0.30) o de que no haga una compra (0.70) se supone que es la misma para todos los clientes.
4. La decisión de comprar de cada cliente es independiente de la decisión de comprar de los otros clientes.

FIGURA 5.3 DIAGRAMA DE ÁRBOL PARA EL PROBLEMA DE LA TIENDA DE ROPA MARTIN CLOTHING STORE



En consecuencia, se satisfacen las propiedades de un experimento binomial.

Con la fórmula siguiente* se calcula el número de resultados experimentales en los que hay exactamente x éxitos en n ensayos.

NÚMERO DE RESULTADOS EXPERIMENTALES EN LOS QUE HAY EXACTAMENTE x ÉXITOS EN n ENSAYOS

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (5.6)$$

donde

$$n! = n(n-1)(n-2) \cdots (2)(1)$$

y por definición,

$$0! = 1$$

Ahora regrese al experimento de las decisiones de compra de tres clientes de la tienda Martin Clothing Store. La ecuación (5.6) sirve para determinar el número de resultados experimentales

* Esta fórmula presentada en el capítulo 4, determina el número de combinaciones de n objetos tomados de x a la vez. En el experimento binomial esta fórmula combinatoria da el número de resultados experimentales (series de n ensayos) en los que hay x éxitos.

en los que hay dos compras; el número de maneras en que son posibles $x = 2$ éxitos en $n = 3$ ensayos. De acuerdo con la ecuación (5.6)

$$\binom{n}{x} = \binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{(3)(2)(1)}{(2)(1)(1)} = \frac{6}{2} = 3$$

La ecuación (5.6) indica que en tres de los resultados experimentales hay dos éxitos. En la figura 5.3 aparecen denotados por (S, S, F) , (S, F, S) y (F, S, S) .

Empleando la ecuación (5.6) para determinar en cuántos resultados experimentales hay tres éxitos (compras) en tres ensayos, se obtiene

$$\binom{n}{x} = \binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3!}{3!0!} = \frac{(3)(2)(1)}{3(2)(1)(1)} = \frac{6}{6} = 1$$

El único resultado experimental con tres éxitos es el identificado por (S, S, S) mostrado en la figura 5.3.

Ya sabe que usando la ecuación (5.6) es posible determinar el número de resultados experimentales en los que hay x éxitos. Sin embargo, si va a determinar la probabilidad de x éxitos en n ensayos, es necesario conocer también la probabilidad correspondiente a cada uno de estos resultados experimentales. Como en un experimento binomial, los ensayos son independientes, para hallar la probabilidad de una determinada sucesión de éxitos y fracasos simplemente se multiplican las probabilidades correspondientes al resultado de cada ensayo.

La probabilidad de que los dos primeros clientes compren y el tercero no compre, denotada por (S, S, F) está dada por

$$pp(1-p)$$

Puesto que la probabilidad de compra en cualquier ensayo es 0.30, la probabilidad de que haya una compra en los dos primeros ensayos y que no haya compra en el tercer ensayo es

$$(0.30)(0.30)(0.70) = (0.30)^2(0.70) = 0.063$$

Hay otros dos resultados experimentales en los que también se obtienen dos éxitos y un fracaso. A continuación se presentan las probabilidades de los tres resultados experimentales en los que hay dos éxitos.

Resultados de los ensayos			Resultado experimental	Probabilidad de este resultado experimental
1er. cliente	2o. cliente	3er. cliente		
Compra	Compra	No hay compra	(S, S, F)	$pp(1-p) = p^2(1-p)$ $= (0.30)^2(0.70) = 0.063$
Compra	Compra	Compra	(S, F, S)	$p(1-p)p = p^2(1-p)$ $= (0.30)^2(0.70) = 0.063$
No hay compra	Compra	Compra	(F, S, S)	$(1-p)pp = p^2(1-p)$ $= (0.30)^2(0.70) = 0.063$

Observe que los tres resultados experimentales en los que hay dos éxitos tienen la misma probabilidad. Esto se cumple en general. En cualquier experimento binomial todas las series de resultados de ensayos en las que hay x éxitos en n ensayos tienen la *misma probabilidad* de ocurrencia. A continuación se presenta la probabilidad de cada una de las series de ensayos en las que hay x éxitos en n ensayos.

Probabilidad de una
determinada serie de $= p^x(1 - p)^{(n-x)}$
resultados de ensayos

En el caso de la tienda de ropa Martin Clothing Store, esta fórmula indica que la probabilidad de cualquier resultado experimental con dos éxitos es $p^2(1 - p)^{(3-2)} = p^2(1 - p)^1 = (0.30)^2(0.70)^1 = 0.63$.

Como la ecuación (5.6) da el número de resultados de un experimento binomial en el que hay x éxitos, y la ecuación (5.7) da la probabilidad de cada serie en la que hay x éxitos, combinando las ecuaciones (5.6) y (5.7) se obtiene la **función de probabilidad binomial** siguiente.

FUNCIÓN DE PROBABILIDAD BINOMIAL

$$f(x) = \binom{n}{x} p^x (1 - p)^{(n-x)} \quad (5.8)$$

donde

$f(x)$ = probabilidad de x éxitos en n ensayos

n = número de ensayos

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

p = probabilidad de un éxito en cualquiera de los ensayos

$1 - p$ = probabilidad de un fracaso en cualquiera de los ensayos

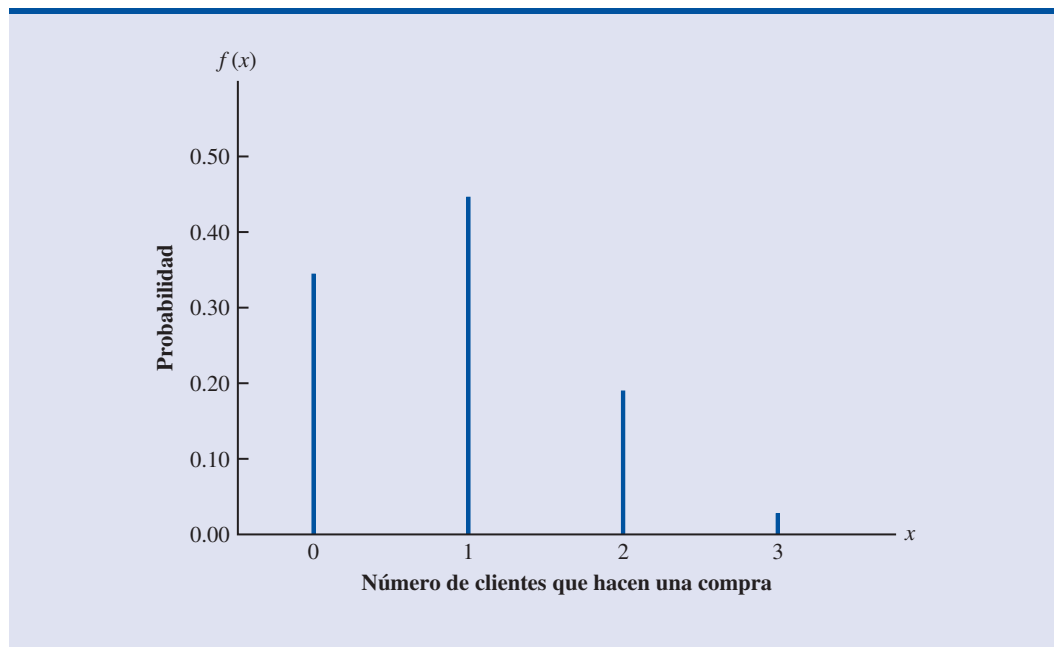
En el ejemplo de la tienda de ropa Martin Clothing Store se calculará ahora la probabilidad de que ningún cliente realice una compra, de que exactamente un cliente realice una compra, de que exactamente dos clientes realicen una compra y de que los tres clientes realicen una compra. Los cálculos se presentan en forma resumida en la tabla 5.7, que da la distribución de probabilidad para el número de clientes que hacen una compra. La figura 5.4 es una gráfica de esta distribución de probabilidad.

La función de probabilidad binomial es aplicable a *cualquier* experimento binomial. Si encuentra que una situación presenta las propiedades de un experimento binomial y conoce los valores de n y p , use la ecuación (5.8) para calcular la probabilidad de x éxitos en n ensayos.

TABLA 5.7 DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL PARA EL NÚMERO DE CLIENTES QUE HACEN UNA COMPRA

x	$f(x)$
0	$\frac{3!}{0!3!} (0.30)^0 (0.70)^3 = 0.343$
1	$\frac{3!}{1!2!} (0.30)^1 (0.70)^2 = 0.441$
2	$\frac{3!}{2!1!} (0.30)^2 (0.70)^1 = 0.189$
3	$\frac{3!}{3!0!} (0.30)^3 (0.70)^0 = \frac{0.027}{1.000}$

FIGURA 5.4 REPRESENTACIÓN GRÁFICA DE LA DISTRIBUCIÓN DE PROBABILIDAD BINOMIAL PARA EL NÚMERO DE CLIENTES QUE HACEN UNA COMPRA



Si considera variaciones del experimento de la tienda de ropa, por ejemplo, que lleguen a la tienda 10 clientes en lugar de tres clientes, también se emplea la función de probabilidad binomial dada por la ecuación (5.8). Suponga que tiene un experimento binomial con $n = 10$, $x = 4$ y $p = 0.30$. La probabilidad de que cuatro de los 10 clientes que entran en la tienda de ropa realicen una compra es

$$f(4) = \frac{10!}{4!6!} (0.30)^4 (0.70)^6 = 0.2001$$

Uso de las tablas de probabilidades binomiales

Existen tablas que dan la probabilidad de x éxitos en n ensayos de un experimento binomial. Estas tablas son fáciles de usar y los resultados se obtienen más rápidamente que con la ecuación (5.8). La tabla 5 del apéndice B es una de estas tablas de probabilidades binomiales. Una parte de esta tabla se presenta en la tabla 5.8. Para usarla es necesario especificar los valores de n , p y x en el experimento binomial de que se trate. En el ejemplo que se presenta en la parte superior de la tabla 5.8 se ve que la probabilidad de $x = 3$ éxitos en un experimento binomial con $n = 10$ y $p = 0.40$ es 0.2150. Use la ecuación (5.8) para verificar que este mismo resultado se obtiene si usa la función de probabilidad binomial directamente.

Ahora se usará la tabla 5.8 para corroborar la probabilidad de 4 éxitos en 10 ensayos en el problema de la tienda de ropa Martin Clothing Store. Observe que el valor de $f(4) = 0.2001$ se lee directamente de la tabla de probabilidades binomiales, eligiendo $n = 10$, $x = 4$ y $p = 0.30$.

Aun cuando las tablas de probabilidades binomiales son relativamente fáciles de utilizar, es imposible contar con tablas que tengan todos los valores de n y p de un experimento binomial. Sin embargo, con las calculadoras de hoy en día, usar la ecuación (5.8) para calcular la probabilidad deseada no es difícil, en especial si el número de ensayos no es grande. En los ejercicios tendrá la oportunidad de usar la ecuación (5.8) para calcular probabilidades binomiales, a menos que el problema pida que use la tabla de probabilidad binomial.

Con las calculadoras modernas estas tablas son casi innecesarias. Es muy fácil evaluar la ecuación (5.8) directamente.

TABLA 5.8 ALGUNOS VALORES DE LA TABLA DE PROBABILIDAD BINOMIAL
EJEMPLO: $n = 10$, $x = 3$, $p = 0.40$; $f(3) = 0.2150$

n	x	0.05	0.10	0.15	0.20	p 0.25	0.30	0.35	0.40	0.45	0.50
9	0	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176
	2	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703
	3	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641
	4	0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461
	5	0.0000	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461
	6	0.0000	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641
	7	0.0000	0.0000	0.0000	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0020
10	0	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098
	2	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439
	3	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172
	4	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051
	5	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461
	6	0.0000	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051
	7	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172
	8	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0005	0.0016	0.0042	0.0098
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010

Los paquetes de software para estadística como Minitab y los paquetes de hojas de cálculo como Excel también están habilitadas para calcular probabilidades binomiales. Considere el ejemplo de la tienda de ropa Martin Clothing Store con $n = 10$ y $p = 0.30$. En la figura 5.5 se muestran las probabilidades binomiales para todos los valores posibles de x , generadas por Minitab. Observe que estos valores son los mismos que se encuentran en la columna $p = 0.30$ de la tabla 5.8. En el apéndice 5.1 se da paso por paso el procedimiento en Minitab para generar el resultado que se muestra en la figura 5.5. En el apéndice 5.2 se describe cómo usar Excel para calcular probabilidades binomiales.

Valor esperado y varianza en la distribución binomial

En la sección 5.3 se dieron las fórmulas para calcular el valor esperado y la varianza de una variable aleatoria discreta. En el caso especial de que la variable aleatoria tenga una distribución binomial para la que se conoce el número de ensayos n y la probabilidad de éxito p , las fórmulas generales para el valor esperado y la varianza se simplifican. El resultado se muestra a continuación.

VALOR ESPERADO Y VARIANZA EN LA DISTRIBUCIÓN BINOMIAL

$$E(x) = \mu = np \quad (5.9)$$

$$\text{Var}(x) = \sigma^2 = np(1 - p) \quad (5.10)$$

FIGURA 5.5 RESULTADOS DE MINITAB QUE MUESTRAN LAS PROBABILIDADES BINOMIALES PARA EL PROBLEMA DE LA TIENDA DE ROPA MARTIN CLOTHING STORE

x	P(X = x)
0.00	0.0282
1.00	0.1211
2.00	0.2335
3.00	0.2668
4.00	0.2001
5.00	0.1029
6.00	0.0368
7.00	0.0090
8.00	0.0014
9.00	0.0001
10.00	0.0000

Para el problema de los tres clientes de la tienda de ropa Martin Clothing Store, use la ecuación (5.9) para calcular el número esperado de clientes que harán una compra.

$$E(x) = np = 3(0.30) = 0.9$$

Suponga que Martin Clothing Store pronostica que el mes próximo 1000 clientes visitarán la tienda. ¿Cuál es el número esperado de clientes que harán una compra? La respuesta es $\mu = np = (1000)(0.30) = 300$. Así, para aumentar el número esperado de compras, Martin debe hacer que más clientes visiten su tienda o de alguna manera aumentar la probabilidad de que una persona que visite la tienda haga una compra.

En el caso de los tres clientes de la tienda de ropa Martin Clothing Store, la varianza y la desviación estándar del número de clientes que harán una compra son

$$\begin{aligned}\sigma^2 &= np(1 - p) = 3(0.3)(0.7) = 0.63 \\ \sigma &= \sqrt{0.63} = 0.79\end{aligned}$$

Para los próximos 1000 clientes que visiten la tienda, la varianza y la desviación estándar del número de clientes que harán una compra son

$$\begin{aligned}\sigma^2 &= np(1 - p) = 1000(0.3)(0.7) = 210 \\ \sigma &= \sqrt{210} = 14.49\end{aligned}$$

NOTAS Y COMENTARIOS

1. En las tablas binomiales del apéndice B los valores de p llegan sólo hasta 0.50. Es posible pensar que estas tablas no son útiles cuando la probabilidad de éxito es mayor a 0.50. Sin embargo, puede usarlas observando que la probabilidad de $n - x$ fracasos es también la probabilidad de x éxitos. Cuando la probabilidad de éxito es mayor que $p = 0.50$, en lugar de la probabilidad de éxito calcule la probabilidad de $n - x$ fracasos. Cuando $p > 0.50$, la probabilidad de fracaso, $1 - p$, será menor que 0.50.
2. En algunas fuentes se presentan tablas binomiales en forma acumulada. Al usar estas tablas para hallar la probabilidad de x éxitos en n ensayos hay que hacer una resta. Por ejemplo, $f(2) = P(x \leq 2) - P(x \leq 1)$. Las tablas que se presentan en este libro dan estas probabilidades. Para calcular probabilidades acumuladas usando las tablas de este libro, sume las probabilidades individuales. Por ejemplo, para calcular $P(x \leq 2)$ usando las tablas del libro, sume $f(0) + f(1) + f(2)$.

Ejercicios

Métodos

Autoexamen

25. Considere un experimento binomial con dos ensayos y $p = 0.4$.
 - a. Dibuje un diagrama de árbol para este experimento (véase figura 5.3).
 - b. Calcule la probabilidad de un éxito, $f(1)$.
 - c. Calcule $f(0)$.
 - d. Calcule $f(2)$.
 - e. Calcule la probabilidad de por lo menos un éxito.
 - f. Calcule el valor esperado, la varianza y la desviación estándar.
26. Considere un experimento binomial con $n = 10$ y $p = 0.10$.
 - a. Calcule $f(0)$.
 - b. Calcule $f(2)$.
 - c. Calcule $P(x \leq 2)$.
 - d. Calcule $P(x \geq 1)$.
 - e. Calcule $E(x)$.
 - f. Calcule $\text{Var}(x)$ y σ .
27. Considere un experimento binomial con $n = 20$ y $p = 0.70$.
 - a. Calcule $f(12)$.
 - b. Calcule $f(16)$.
 - c. Calcule $P(x \geq 16)$.
 - d. Calcule $P(x \leq 15)$.
 - e. Calcule $E(x)$.
 - f. Calcule $\text{Var}(x)$ y σ .

Aplicaciones

28. Una encuesta de Harris Interactive para InterContinental Hotel and Resorts preguntó: “Cuando viaja al extranjero, ¿suele aventurarse usted solo para conocer la cultura o prefiere permanecer con el grupo de su *tour* y apegarse al itinerario?” Se encontró que 23% prefiere permanecer con el grupo de su *tour* (*USA Today*, 21 de enero de 2004).
 - a. ¿Cuál es la probabilidad de que en una muestra de seis viajeros, dos prefieran permanecer con su grupo?
 - b. ¿De que en una muestra de seis viajeros, por lo menos dos prefieran permanecer con su grupo?
 - c. ¿De que en una muestra de 10 viajeros, ninguno prefiera permanecer con su grupo?
29. En San Francisco, 30% de los trabajadores emplean el transporte público (*USA Today*, 21 de diciembre de 2005).
 - a. ¿Cuál es la probabilidad de que en una muestra de 10 trabajadores exactamente tres empleen el transporte público?
 - b. ¿De que en una muestra de 10 trabajadores por lo menos tres empleen el transporte público?
30. Cuando una máquina nueva funciona adecuadamente, sólo 3% de los artículos producidos presentan algún defecto. Suponga que selecciona aleatoriamente dos piezas producidas con la nueva máquina y que busca el número de piezas defectuosas.
 - a. Describa las condiciones en las que éste será un experimento binomial.
 - b. Elabore un diagrama de árbol como el de la figura 5.3 en el que se muestre este problema como un experimento de dos ensayos.
 - c. ¿En cuántos resultados experimentales hay exactamente una pieza defectuosa?
 - d. Calcule las probabilidades de hallar ninguna pieza defectuosa, exactamente una pieza defectuosa y dos piezas defectuosas.
31. Nueve por ciento de los estudiantes tienen un balance en su tarjeta de crédito mayor a \$7000 (*Reader's Digest*, julio de 2002). Suponga que selecciona aleatoriamente 10 estudiantes para entrevistarlos respecto del uso de su tarjeta de crédito

Autoexamen

- a. ¿Es la selección de 10 estudiantes un experimento binomial? Explique.
 - b. ¿Cuál es la probabilidad de que dos de los estudiantes tengan un balance en su tarjeta de crédito superior a \$7000?
 - c. ¿De que ninguno tenga un balance en su tarjeta de crédito superior a \$7000?
 - d. ¿De que por lo menos tres tengan un balance en su tarjeta de crédito superior a \$7000?
32. Los radares militares y los sistemas para detección de misiles tienen por objeto advertir a un país de un ataque enemigo. Una cuestión de confiabilidad es si el sistema de detección será capaz de detectar un ataque y emitir un aviso. Suponga que la probabilidad de que un determinado sistema de detección detecte un ataque con misiles es 0.90. Use la distribución de probabilidad binomial para responder las preguntas siguientes.
- a. ¿Cuál es la probabilidad de que un solo sistema de detección detecte un ataque?
 - b. Si se instalan dos sistemas de detección en una misma área y los dos operan independientemente, ¿cuál es la probabilidad de que por lo menos uno de los sistemas detecte el ataque?
 - c. Si se instalan tres sistemas, ¿cuál es la probabilidad de que por lo menos uno de los sistemas detecte el ataque?
 - d. ¿Recomendaría que se usaran varios sistemas de detección? Explique.
33. Cincuenta por ciento de los estadounidenses creyeron que el país se encontraba en una recesión aun cuando en la economía no se habían observado dos trimestres seguidos con crecimiento negativo. (*BusinessWeek*, 30 de julio de 2001). Dada una muestra de 20 estadounidenses, calcule lo siguiente.
- a. Calcule la probabilidad de que exactamente 12 personas hayan creído que el país estaba en recesión.
 - b. De que no más de cinco personas hayan creído que el país estaba en recesión
 - c. ¿Cuántas personas esperaría usted que dijeran que el país estuvo en recesión?
 - d. Calcule la varianza y la desviación estándar del número de personas que creyeron que el país estuvo en recesión.
34. En una encuesta realizada por la Oficina de Censos de Estados Unidos se encontró que 25% de las personas de 25 años o más habían estudiado cuatro años en la universidad (*The New York Times Almanac*, 2006). Dada una muestra de 15 individuos de 25 años o más, conteste las preguntas siguientes.
- a. ¿Cuál es la probabilidad de que cuatro hayan estudiado cuatro años en la universidad?
 - b. ¿De que tres o más hayan estudiado cuatro años en la universidad?
35. En una universidad se encontró que 20% de los estudiantes no terminan el primer curso de estadística, al curso se inscriben 20 estudiantes.
- a. Calcule la probabilidad de que dos o menos no terminen.
 - b. De que cuatro, exactamente, no terminen.
 - c. De que más de tres no terminen.
 - d. ¿Cuál es el número esperado de estudiantes que no terminan?
36. En el caso particular de una variable aleatoria binomial, es factible calcular la varianza empleando la fórmula $\sigma^2 = np(1 - p)$. En el caso del problema de la tienda de ropa Martin Clothing Store, en donde $n = 3$ y $p = 0.3$, se encontró que $\sigma^2 = np(1 - p) = 3(0.3)(0.7) = 0.63$. Aplique la definición general de varianza para una variable aleatoria discreta, ecuación (5.5), y las probabilidades de la tabla 5.7 para comprobar que la varianza es 0.63
37. Veintitres por ciento de los automóviles no cuenta con un seguro (CNN, 23 de febrero de 2006). En un fin de semana determinado hay 35 automóviles que sufren un accidente.
- a. ¿Cuál es el número esperado de estos automóviles que no cuentan con un seguro?
 - b. ¿Cuál es la varianza y la desviación estándar?

5.5

Distribución de probabilidad de Poisson

En esta sección estudiará una variable aleatoria discreta que se suele usar para estimar el número de veces que sucede un hecho determinado (ocurrencias) en un intervalo de tiempo o de espacio. Por ejemplo, la variable de interés va desde el número de automóviles que llegan (llegadas) a un lavado de coches en una hora o el número de reparaciones necesarias en 10 millas de una autopista hasta el número de fugas en 100 millas de tubería. Si se satisfacen las condiciones si-

La distribución de probabilidad de Poisson suele emplearse para modelar las llegadas aleatorias a una línea de espera (fila).

guientes, el número de ocurrencias es una variable aleatoria discreta, descrita por la **distribución de probabilidad de Poisson**.

PROPIEDADES DE UN EXPERIMENTO DE POISSON

1. La probabilidad de ocurrencia es la misma para cualesquiera dos intervalos de la misma magnitud.
2. La ocurrencia o no-ocurrencia en cualquier intervalo es independiente de la ocurrencia o no-ocurrencia en cualquier otro intervalo.

La **función de probabilidad de Poisson** se define mediante la ecuación (5.11).

Simeon Poisson dio clases de matemáticas en la Ecole Polytechnique de París de 1802 a 1808. En 1837 publicó un trabajo titulado "Investigación sobre la probabilidad de veredictos en materia criminal y civil" en el que presenta un estudio sobre lo que después se conoció como distribución de Poisson.

FUNCIÓN DE PROBABILIDAD DE POISSON

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \quad (5.11)$$

en donde

$f(x)$ = probabilidad de x ocurrencias en un intervalo
 μ = valor esperado o número medio de ocurrencias en un intervalo
 $e = 2.71828$

Antes de considerar un ejemplo para ver cómo se usa la distribución de Poisson, observe que el número de ocurrencias x , no tiene límite superior. Ésta es una variable aleatoria discreta que toma los valores de una sucesión infinita de números ($x = 0, 1, 2, \dots$).

Un ejemplo considerando intervalos de tiempo

Suponga que desea saber el número de llegadas, en un lapso de 15 minutos, a la rampa del cajero automático de un banco. Si se puede suponer que la probabilidad de llegada de los automóviles es la misma en cualesquiera dos lapsos de la misma duración y si la llegada o no-llegada de un automóvil en cualquier lapso es independiente de la llegada o no-llegada de un automóvil en cualquier otro lapso, se puede aplicar la función de probabilidad de Poisson. Dichas condiciones se satisfacen y en un análisis de datos pasados encuentra que el número promedio de automóviles que llegan en un lapso de 15 minutos es 10; en este caso use la función de probabilidad siguiente.

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Aquí la variable aleatoria es x = número de automóviles que llegan en un lapso de 15 minutos.

Si la administración desea saber la probabilidad de que lleguen exactamente cinco automóviles en 15 minutos, $x = 5$, y se obtiene

$$\text{Probabilidad de que lleguen exactamente 5 automóviles en 15 minutos} = f(5) = \frac{10^5 e^{-10}}{5!} = 0.0378$$

Aunque esta probabilidad se obtuvo evaluando la función de probabilidad con $\mu = 10$ y $x = 5$, suele ser más fácil consultar una tabla de probabilidad de Poisson. Dichas tablas proporcionan las probabilidades para valores específicos de x y μ . La tabla 7 del apéndice B es una tabla de probabilidad de Poisson. Para mayor comodidad, en la tabla 5.9 se reproduce parte de la tabla 7 del apéndice B. Observe que para usar una tabla de probabilidades de Poisson se necesitan sólo

Los laboratorios Bell usaron la distribución de Poisson para modelar las llegadas de llamadas telefónicas.

TABLA 5.9 ALGUNOS VALORES DE LAS TABLAS DE PROBABILIDAD DE POISSON
EJEMPLO: $\mu = 10, x = 5; f(5) = .0378$

x	μ									
	9.1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10
0	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
1	0.0010	0.0009	0.0009	0.0008	0.0007	0.0007	0.0006	0.0005	0.0005	0.0005
2	0.0046	0.0043	0.0040	0.0037	0.0034	0.0031	0.0029	0.0027	0.0025	0.0023
3	0.0140	0.0131	0.0123	0.0115	0.0107	0.0100	0.0093	0.0087	0.0081	0.0076
4	0.0319	0.0302	0.0285	0.0269	0.0254	0.0240	0.0226	0.0213	0.0201	0.0189
5	0.0581	0.0555	0.0530	0.0506	0.0483	0.0460	0.0439	0.0418	0.0398	0.0378
6	0.0881	0.0851	0.0822	0.0793	0.0764	0.0736	0.0709	0.0682	0.0656	0.0631
7	0.1145	0.1118	0.1091	0.1064	0.1037	0.1010	0.0982	0.0955	0.0928	0.0901
8	0.1302	0.1286	0.1269	0.1251	0.1232	0.1212	0.1191	0.1170	0.1148	0.1126
9	0.1317	0.1315	0.1311	0.1306	0.1300	0.1293	0.1284	0.1274	0.1263	0.1251
10	0.1198	0.1210	0.1219	0.1228	0.1235	0.1241	0.1245	0.1249	0.1250	0.1251
11	0.0991	0.1012	0.1031	0.1049	0.1067	0.1083	0.1098	0.1112	0.1125	0.1137
12	0.0752	0.0776	0.0799	0.0822	0.0844	0.0866	0.0888	0.0908	0.0928	0.0948
13	0.0526	0.0549	0.0572	0.0594	0.0617	0.0640	0.0662	0.0685	0.0707	0.0729
14	0.0342	0.0361	0.0380	0.0399	0.0419	0.0439	0.0459	0.0479	0.0500	0.0521
15	0.0208	0.0221	0.0235	0.0250	0.0265	0.0281	0.0297	0.0313	0.0330	0.0347
16	0.0118	0.0127	0.0137	0.0147	0.0157	0.0168	0.0180	0.0192	0.0204	0.0217
17	0.0063	0.0069	0.0075	0.0081	0.0088	0.0095	0.0103	0.0111	0.0119	0.0128
18	0.0032	0.0035	0.0039	0.0042	0.0046	0.0051	0.0055	0.0060	0.0065	0.0071
19	0.0015	0.0017	0.0019	0.0021	0.0023	0.0026	0.0028	0.0031	0.0034	0.0037
20	0.0007	0.0008	0.0009	0.0010	0.0011	0.0012	0.0014	0.0015	0.0017	0.0019
21	0.0003	0.0003	0.0004	0.0004	0.0005	0.0006	0.0006	0.0007	0.0008	0.0009
22	0.0001	0.0001	0.0002	0.0002	0.0002	0.0002	0.0003	0.0003	0.0004	0.0004
23	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001

dos valores, x y μ . En la tabla 5.9 la probabilidad de cinco llegadas en un lapso de 15 minutos se obtiene localizando el valor que se encuentra en el renglón correspondiente a $x = 5$ y la columna correspondiente a $\mu = 10$. Así obtiene $f(5) = 0.0378$

La media de la distribución de Poisson en el ejemplo anterior fue $\mu = 10$ llegadas en un lapso de 15 minutos. Una propiedad de la distribución de Poisson es que la media y la varianza de la distribución son iguales. Por tanto, la varianza del número de llegadas en un lapso de 15 minutos es $\sigma^2 = 10$. La desviación estándar es $\sigma = \sqrt{10} = 3.16$.

En el ejemplo anterior se usó un lapso de 15 minutos, pero también se usan otros lapsos. Suponga que desea calcular la probabilidad de una llegada en un lapso de 3 minutos. Como 10 es el número esperado de llegadas en un lapso de 15 minutos: $10/15 = 2/3$ es el número esperado de llegadas en un lapso de un minuto y que $(2/3)(3 \text{ minutos}) = 2$ es el número esperado de llegadas en un lapso de 3 minutos. Entonces, la probabilidad de x llegadas en un lapso de 3 minutos con $\mu = 2$ está dada por la siguiente función de probabilidad de Poisson.

$$f(x) = \frac{2^x e^{-2}}{x!}$$

La probabilidad de una llegada en un lapso de 3 minutos se obtiene como sigue:

$$\text{Probabilidad de exactamente una llegada en 3 minutos} = f(1) = \frac{2^1 e^{-2}}{1!} = 0.2707$$

Una propiedad de la distribución de Poisson es que la media y la varianza son iguales.

Antes se calculó la probabilidad de cinco llegadas en un lapso de 15 minutos; se obtuvo 0.0378. Observe que la probabilidad de una llegada en un lapso de tres minutos (0.2707) no es la misma. Para calcular la probabilidad de Poisson en un lapso diferente, primero hay que convertir la llegada media al lapso que interesa y después calcular la probabilidad.

Un ejemplo considerando intervalos de longitud o de distancia

Ahora se da un ejemplo en el que no aparecen intervalos de tiempo y en el que se usa la distribución de Poisson. Asuma que le interesa la ocurrencia de una avería importante en una autopista un mes después de que ha sido repavimentada. Supondrá que la probabilidad de que haya una avería es la misma en cualesquiera dos tramos, de una misma longitud, de la autopista y que la ocurrencia o no-ocurrencia de una avería en un tramo es independiente de la ocurrencia o no-ocurrencia de una avería en cualquier otro tramo. Por tanto, emplea la distribución de Poisson.

También sabe que el promedio de averías importantes, un mes después de la repavimentación, son dos averías por milla. Desea determinar la probabilidad de que no haya ninguna avería en un determinado tramo de tres millas de autopista. Como lo que interesa es un intervalo cuya longitud es de tres millas, $\mu = (2 \text{ averías/milla})(3 \text{ millas}) = 6$ representa el número esperado de averías importantes en un tramo de tres millas de autopista. Mediante la ecuación (5.11), la probabilidad de que no haya ninguna avería importante es $f(0) = 6^0 e^{-6}/0! = 0.0025$. Por tanto, es poco probable que no haya ninguna avería importante en este tramo de tres millas. En efecto, este ejemplo indica que hay una probabilidad de $1 - 0.0025 = 0.9975$ de que haya por lo menos una avería importante en este tramo de tres millas de autopista.

Ejercicios

Métodos

38. Considere una distribución de Poisson con $\mu = 3$.
 - a. Dé la adecuada función de probabilidad de Poisson.
 - b. Calcule $f(2)$.
 - c. Calcule $f(1)$.
 - d. Calcule $P(x \geq 2)$.
39. Considere una distribución de Poisson en que la media es de dos ocurrencias por un periodo de tiempo.
 - a. Dé la adecuada función de probabilidad de Poisson.
 - b. ¿Cuál es el número esperado de ocurrencias en tres periodos de tiempo?
 - c. Dé la adecuada función de probabilidad de Poisson para determinar la probabilidad de x ocurrencias en tres lapsos.
 - d. Calcule la probabilidad de dos ocurrencias en un periodo de tiempo.
 - e. Calcule la probabilidad de seis ocurrencias en tres periodos de tiempo.
 - f. Calcule la probabilidad de cinco ocurrencias en dos periodos de tiempo.

Autoexamen

Aplicaciones

40. A la oficina de reservaciones de una aerolínea regional llegan 48 llamadas por hora.
 - a. Calcule la probabilidad de recibir cinco llamadas en un lapso de 5 minutos.
 - b. Estime la probabilidad de recibir exactamente 10 llamadas en un lapso de 15 minutos.
 - c. Suponga que no hay ninguna llamada en espera. Si el agente de viajes necesitará 5 minutos para la llamada que está atendiendo, ¿cuántas llamadas habrá en espera para cuando él termine? ¿Cuál es la probabilidad de que no haya ninguna llamada en espera?
 - d. Si en este momento no hay ninguna llamada, ¿cuál es la probabilidad de que el agente de viajes pueda tomar 3 minutos de descanso sin ser interrumpido por una llamada?

Autoexamen

41. Durante el periodo en que una universidad recibe inscripciones por teléfono, llegan llamadas a una velocidad de una cada dos minutos.
 - a. ¿Cuál es el número esperado de llamadas en una hora?
 - b. ¿Cuál es la probabilidad de que haya tres llamadas en cinco minutos?
 - c. ¿De que no haya llamadas en un lapso de cinco minutos?
42. En Estados Unidos, cada año, más de 50 millones de huéspedes se alojan en un “Bread and breakfast” (B&B). El sitio Web dedicado a los alojamientos tipo Bread and Breakfast en Estados Unidos (www.bestinns.net), que tiene un promedio aproximado de siete visitantes por minuto, permite a muchos B&B obtener huéspedes (*Time*, septiembre de 2001).
 - a. Calcule la probabilidad de que no haya ningún visitante al sitio Web en un lapso de un minuto.
 - b. De que haya dos o más visitantes al sitio Web en un lapso de un minuto.
 - c. De que haya uno o más visitantes al sitio Web en un lapso de 30 segundos.
 - d. De que haya cinco o más visitantes al sitio Web en un lapso de un minuto.
43. Los pasajeros de las aerolíneas llegan en forma aleatoria e independiente al mostrador de revisión de pasajeros. La tasa media de llegada es 10 pasajeros por minuto.
 - a. Calcule la probabilidad de que no llegue ningún pasajero en un lapso de un minuto.
 - b. Calcule la probabilidad de que lleguen tres o menos pasajeros en un lapso de un minuto.
 - c. De que no llegue ningún pasajero en un lapso de 15 segundos.
 - d. De que llegue por lo menos un pasajero en un lapso de 15 segundos.
44. Cada año ocurren en promedio 15 accidentes aéreos (*The World Almanac and Book of Facts*, 2004).
 - a. Calcule el número medio de accidentes aéreos por mes.
 - b. Calcule la probabilidad de que no haya ningún accidente en un mes.
 - c. De que haya exactamente un accidente en un mes.
 - d. De que haya más de un accidente en un mes.
45. El National Safety Council de Estados Unidos estima que los accidentes fuera del trabajo tienen para las empresas un costo de casi \$200 mil millones anuales en pérdida de productividad. Con base en estos datos, las empresas que tienen 50 empleados esperan tener por lo menos tres accidentes fuera del trabajo por año. Para estas empresas con 50 empleados, conteste las preguntas siguientes.
 - a. ¿Cuál es la probabilidad de que no haya ningún accidente fuera del trabajo en un año?
 - b. ¿De que haya por lo menos dos accidentes fuera del trabajo en un año?
 - c. ¿Cuál es el número esperado de accidentes fuera del trabajo en un lapso de seis meses?
 - d. ¿Cuál es la probabilidad de que no haya ningún accidente fuera del trabajo en los próximos seis meses?

5.6

Distribución de probabilidad hipergeométrica

La **distribución de probabilidad hipergeométrica** está estrechamente relacionada con la distribución binomial. Pero difieren en dos puntos: en la distribución hipergeométrica los ensayos no son independientes y la probabilidad de éxito varía de ensayo a ensayo.

En la notación usual en la distribución hipergeométrica, r denota el número de elementos considerados como éxitos que hay en una población de tamaño N , y $N - r$ denota el número de elementos considerados como fracasos que hay en dicha población. La **función de probabilidad hipergeométrica** se usa para calcular la probabilidad de que en una muestra aleatoria de n elementos, seleccionados sin reemplazo, se tengan x éxitos y $n - x$ fracasos. Para que se presente este resultado, debe tener x éxitos de los r éxitos que hay en la población y $n - x$ fracasos de los $N - r$ fracasos. La siguiente función de probabilidad hipergeométrica proporciona $f(x)$, la probabilidad de tener x éxitos en una muestra de tamaño n .

FUNCIÓN DE PROBABILIDAD HIPERGEOMÉTRICA

$$f(x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}} \quad \text{para } 0 \leq x \leq r \quad (5.12)$$

donde

$f(x)$ = probabilidad de x éxitos en n ensayos

n = número de ensayos

N = número de elementos en la población

r = número de elementos en la población considerados como éxitos

Observe que $\binom{N}{n}$ representa el número de maneras en que es posible tomar una muestra de tamaño n de una población de tamaño N ; $\binom{r}{x}$ representa el número de formas en que se toman x éxitos de un total de r éxitos que hay en la población, y $\binom{N-r}{n-x}$ representa el número de maneras en que se puede tomar $n-x$ fracasos de un total de $N-r$ que hay en la población.

Para ilustrar los cálculos que se emplean al usar la ecuación (5.12), considere la siguiente aplicación al control de calidad. Una empresa fabrica fusibles que empaca en cajas de 12 unidades cada una. Asuma que un inspector selecciona al azar tres de los 12 fusibles de una caja para inspeccionarlos. Si la caja contiene exactamente cinco fusibles defectuosos, ¿cuál es la probabilidad de que el inspector encuentre que uno de los tres fusibles está defectuoso? En esta aplicación $n = 3$ y $N = 12$. Si $r = 5$ fusibles defectuosos en la caja, la probabilidad de hallar $x = 1$ defectuoso es

$$f(1) = \frac{\binom{5}{1} \binom{7}{2}}{\binom{12}{3}} = \frac{\left(\frac{5!}{1!4!}\right) \left(\frac{7!}{2!5!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(5)(21)}{220} = 0.4773$$

Ahora suponga que desea conocer la probabilidad de hallar *por lo menos* un fusible defectuoso. La manera más sencilla de contestar es calcular primero la probabilidad de que el inspector no encuentre ningún fusible defectuoso. La probabilidad de $x = 0$ es

$$f(0) = \frac{\binom{5}{0} \binom{7}{3}}{\binom{12}{3}} = \frac{\left(\frac{5!}{0!5!}\right) \left(\frac{7!}{3!4!}\right)}{\left(\frac{12!}{3!9!}\right)} = \frac{(1)(35)}{220} = 0.1591$$

Si la probabilidad de cero fusibles defectuosos es $f(0) = 0.1591$, se concluye que la probabilidad de hallar por lo menos un fusible defectuoso debe ser $1 - 0.1591 = 0.8409$. Así, existe una probabilidad razonablemente alta de que el inspector encuentre por lo menos un fusible defectuoso.

La media y la varianza de una distribución hipergeométrica son las siguientes.

$$E(x) = \mu = n \left(\frac{r}{N} \right) \quad (5.13)$$

$$\text{Var}(x) = \sigma^2 = n \left(\frac{r}{N} \right) \left(1 - \frac{r}{N} \right) \left(\frac{N-n}{N-1} \right) \quad (5.14)$$

En el ejemplo anterior $n = 3$, $r = 5$ y $N = 12$. Por tanto, la media y la varianza del número de fusibles defectuosos es

$$\mu = n \left(\frac{r}{N} \right) = 3 \left(\frac{5}{12} \right) = 1.25$$

$$\sigma^2 = n \left(\frac{r}{N} \right) \left(1 - \frac{r}{N} \right) \left(\frac{N-n}{N-1} \right) = 3 \left(\frac{5}{12} \right) \left(1 - \frac{5}{12} \right) \left(\frac{12-3}{12-1} \right) = 0.60$$

La desviación estándar es $\sigma = \sqrt{0.60} = 0.77$.

NOTAS Y COMENTARIOS

Considere una distribución hipergeométrica con n ensayos. Sea $p = (r/N)$ la probabilidad de un éxito en el primer ensayo. Si el tamaño de la población es grande, el término $(N-n)/(N-1)$ de la ecuación (5.14) se aproxima a 1. Entonces, el valor esperado y la varianza se expresan como $E(x) = np$ y $\text{Var}(x) = np(1-p)$. Preste atención a que estas expresio-

nes son las mismas que se usan para calcular el valor esperado y la varianza en una distribución binomial, ecuaciones (5.9) y (5.10). Cuando el tamaño de la población es grande, se aproxima una distribución hipergeométrica mediante una distribución binomial con n ensayos y probabilidad de éxito $p = (r/N)$.

Ejercicios

Métodos

46. Suponga que $N = 10$ y $r = 3$. Calcule las probabilidades hipergeométricas correspondientes a los valores siguiente de n y x .
 - a. $n = 4, x = 1$.
 - b. $n = 2, x = 2$.
 - c. $n = 2, x = 0$.
 - d. $n = 4, x = 2$.
47. Suponga que $N = 15$ y $r = 4$. ¿Cuál es la probabilidad de $x = 3$ para $n = 10$?

Aplicaciones

48. En una encuesta realizada por Gallup Organization, se les preguntó a los interrogados, “Cuál es el deporte que prefieres ver”. Fútbol y básquetbol ocuparon el primero y segundo lugar de preferencia (www.gallup.com, 3 de enero de 2004). Si en un grupo de 10 individuos, siete prefieren fútbol y tres prefieren básquetbol. Se toma una muestra aleatoria de tres de estas personas.
 - a. ¿Cuál es la probabilidad de que exactamente dos prefieren el fútbol?
 - b. ¿De qué la mayoría (ya sean dos o tres) prefiere el fútbol?
49. Blackjack, o veintiuno, como se le suele llamar, es un popular juego de apuestas en los casinos de Las Vegas. A un jugador se le reparten dos cartas. Las figuras (sotas, reinas y reyes) y los 10 valen 10 puntos. Los ases valen 1 u 11. Una baraja de 52 cartas tiene 16 cartas que valen 10 (sotas, reinas, reyes y dieces) y cuatro ases.

Autoexamen

- a. ¿Cuál es la probabilidad de que las dos cartas repartidas sean ases o cartas que valgan 10 puntos?
 - b. ¿De que las dos cartas sean ases?
 - c. ¿De que las dos cartas valgan 10?
 - d. Un blackjack es una carta de 10 puntos y un as que suman 21. Use sus respuestas a los incisos a, b y c para determinar la probabilidad de que a un jugador se le reparta blackjack. (*Indicación:* El inciso c no es un problema hipergeométrico. Desarrolle su propio razonamiento lógico para combinar las probabilidades hipergeométricas de los incisos a, b y c para responder esta pregunta.)
50. Una empresa fabrica computadoras personales en dos fábricas, una en Texas y la otra en Hawai. La fábrica de Texas tiene 40 empleados; la fábrica de Hawai tiene 20 empleados. A una muestra aleatoria de 20 empleados se le pide que llene un cuestionario sobre prestaciones.
- a. ¿Cuál es la probabilidad de que ninguno de los empleados de la muestra trabaje en la fábrica de Hawai?
 - b. ¿De que uno de los empleados de la muestra trabaje en la fábrica de Hawai?
 - c. ¿De que dos o más de los empleados de la muestra trabajen en la fábrica de Hawai?
 - d. ¿De que nueve de los empleados de la muestra trabajen en la fábrica de Texas?
51. En una revista de encuestas se da información sobre la evaluación a los platillos, la decoración y el servicio de varios de los principales restaurantes de Estados Unidos. En 15 de los mejor evaluados restaurantes de Boston, el costo promedio de una cena, que incluye una bebida y la propina, es \$48.60. Usted va a ir en viaje de negocios a Boston y le gustaría cenar en tres de estos restaurantes. Su empresa le pagará máximo \$50 por cena. Sus conocidos en Boston le han informado que en una tercera parte de estos restaurantes una cena cuesta más de \$50. Suponga que escoge al azar tres de estos restaurantes para ir a cenar.
- a. ¿Cuál es la probabilidad de que el costo de ninguna de las cenas sea mayor a la cantidad que paga su empresa?
 - b. ¿De que el costo de una de las cenas sea mayor a la cantidad que paga su empresa?
 - c. ¿De que el costo de dos de las cenas sea mayor a la cantidad que paga su empresa?
 - d. ¿De que el costo de las tres cenas sea mayor a la cantidad que paga su empresa?
52. En un pedido de 10 artículos hay dos defectuosos y ocho no defectuosos. Para la inspección del pedido se tomará una muestra y se inspeccionará. Si se encuentra un artículo defectuoso todo el pedido de 10 artículos será devuelto.
- a. Si toma una muestra de tres artículos, ¿cuál es la probabilidad de que devuelva el pedido?
 - b. Si toma una muestra de cuatro artículos, ¿cuál es la probabilidad de que devuelva el pedido?
 - c. Si toma una muestra de cinco artículos, ¿cuál es la probabilidad de que devuelva el pedido?
 - d. Si la administración desea que la probabilidad de rechazar un pedido en el que haya dos artículos defectuosos y ocho no defectuosos sea 0.90, ¿de qué tamaño recomienda que sea la muestra?

Resumen

Una variable aleatoria da una descripción numérica de los resultados de un experimento. La distribución de probabilidad de una variable aleatoria describe cómo se reparten las probabilidades entre los valores que toma dicha variable. En toda variable aleatoria discreta, x , su distribución de probabilidad se define mediante una función de probabilidad, que se denota $f(x)$ y la cual da la probabilidad que corresponde a cada valor de la variable aleatoria. Una vez que se ha definido la función de probabilidad, es posible calcular el valor esperado, la varianza y la desviación estándar de la variable aleatoria.

La distribución binomial se usa para determinar la probabilidad de x éxitos en n ensayos, siempre que el experimento satisfaga las propiedades siguientes:

1. El experimento consista en una serie de n ensayos idénticos.
2. En cada ensayo haya dos resultados posibles, uno llamado éxito y el otro fracaso.
3. La probabilidad de un éxito no varíe de un ensayo a otro. Por tanto, la probabilidad de fracaso, $(1 - p)$, tampoco variará de un resultado a otro.
4. Los ensayos sean independientes.

Si se satisfacen estas cuatro propiedades, la probabilidad de x éxitos en n ensayos se determina usando la función de probabilidad binomial. También se presentaron las fórmulas para hallar la media y la varianza de una distribución binomial.

La distribución de Poisson se usa cuando se quiere obtener la probabilidad de x ocurrencias de un evento en un determinado intervalo de tiempo o de espacio. Para que se emplee la distribución de Poisson deben satisfacerse las condiciones siguientes:

1. La probabilidad de una ocurrencia del evento es la misma para cualesquier dos intervalos de la misma longitud.
2. La ocurrencia o no-ocurrencia del evento en un determinado intervalo es independiente de la ocurrencia o no-ocurrencia del evento en cualquier otro intervalo.

En la sección 5.6 se presentó la tercera distribución discreta de probabilidad presentada, la distribución hipergeométrica. Es como la binomial, que se usa para calcular la probabilidad de x éxitos en n ensayos, pero, a diferencia de ésta, la probabilidad de éxito sí varía de un ensayo a otro.

Glosario

Variable aleatoria Una descripción numérica del resultado de un experimento.

Variable aleatoria discreta Una variable aleatoria que puede asumir un número finito de valores o un número infinito de valores de una sucesión.

Variable aleatoria continua Ésta toma cualquier valor de un intervalo o de una colección de intervalos.

Distribución de probabilidad Descripción de cómo se distribuyen las probabilidades entre los valores de una variable aleatoria.

Función de probabilidad Se denota $f(x)$ y da la probabilidad de que x tome un determinado valor de una variable aleatoria.

Distribución de probabilidad uniforme discreta Distribución de probabilidad para la cual cada posible valor de la variable aleatoria tienen la misma probabilidad.

Valor esperado Medida de localización central de una variable aleatoria.

Varianza Medida de la variabilidad o dispersión de una variable aleatoria.

Desviación estándar Raíz cuadrada positiva de la varianza.

Experimento binomial Un experimento que tiene cuatro propiedades que se dan al principio de la sección 5.4.

Distribución de probabilidad binomial Distribución de probabilidad da la probabilidad de x éxitos en n ensayos de un experimento binomial.

Función de probabilidad binomial La función usada para calcular las probabilidades binomiales.

Distribución de probabilidad de Poisson Distribución de probabilidad da la probabilidad de x ocurrencias de un evento en un determinado intervalo de tiempo o de espacio.

Función de probabilidad de Poisson La función usada para calcular las probabilidades de Poisson.

Distribución de probabilidad hipergeométrica Distribución de probabilidad da la probabilidad de x éxitos en n ensayos a partir de una población en la que hay r éxitos y $N - r$ fracasos.

Función de probabilidad hipergeométrica La función usada para calcular probabilidades hipergeométricas

Fórmulas clave

Función de probabilidad uniforme discreta

$$f(x) = 1/n \quad (5.3)$$

Valor esperado en una variable aleatoria discreta

$$E(x) = \mu = \sum x f(x) \quad (5.4)$$

Varianza en una variable aleatoria discreta

$$\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 f(x) \quad (5.5)$$

Número de resultados experimentales en los que se encuentran exactamente x éxitos en n ensayos

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (5.6)$$

Función de probabilidad binomial

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)} \quad (5.8)$$

Valor esperado en una distribución binomial

$$E(x) = \mu = np \quad (5.9)$$

Varianza en una distribución binomial

$$\text{Var}(x) = \sigma^2 = np(1-p) \quad (5.10)$$

Función de probabilidad de Poisson

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \quad (5.11)$$

Función de probabilidad hipergeométrica

$$f(x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}} \quad \text{para } 0 \leq x \leq r \quad (5.12)$$

Valor esperado en la distribución hipergeométrica

$$E(x) = \mu = n \left(\frac{r}{N} \right) \quad (5.13)$$

Varianza en la distribución hipergeométrica

$$\text{Var}(x) = \sigma^2 = n \left(\frac{r}{N} \right) \left(1 - \frac{r}{N} \right) \left(\frac{N-n}{N-1} \right) \quad (5.14)$$

Ejercicios complementarios

53. El *Barron's* Big Money Poll preguntó a 131 gerentes de inversiones de Estados Unidos acerca de sus puntos de vista sobre las inversiones a corto plazo (*Barron's*, 28 de octubre de 2002). De acuerdo con las respuestas 4% se encontraban muy optimistas, 39 % se encontraban optimistas, 29% se encontraban neutrales, 21% se encontraban pesimistas y 7% se encontraban muy pesimistas. Sea x la variable aleatoria que refleje el grado de optimismo y que vaya desde $x = 1$ para muy pesimista hasta $x = 5$ para muy optimista.
- Elabore una distribución de probabilidad para el grado de optimismo de los gerentes de inversiones.
 - Calcule el valor esperado del grado de optimismo.
 - Calcule la varianza y la desviación estándar del grado de optimismo.
 - Haga un comentario sobre lo que le dicen sus resultados acerca del grado de optimismo y su variabilidad.
54. La American Association of Individual Investors publica una guía anual con los principales fondos mutualistas (*The Individual Investor's Guide to the Top Mutual Funds*, 22^a ed., American Association of Individual Investors, 2003). En la tabla 5.10 se presenta la clasificación de 29 fondos mutualistas de acuerdo con el riesgo.
- Sea x una variable que va desde $x = 1$ con el menor riesgo hasta el mayor riesgo con $x = 5$. Elabore una distribución de probabilidad para el nivel de riesgo.
 - ¿Cuál es el valor esperado y la varianza del nivel de riesgo?
 - Se encontró que 11 de éstos eran fondos de renta fija. De ellos siete se clasificaron como bajos y cuatro como abajo del promedio. Compare el riesgo de los fondos de renta fija con los 18 fondos de acciones.

TABLA 5.10 DE 29 FONDOS MUTUALISTAS

Número de fondos	Nivel de riesgo: categorías
Bajo	7
Bajo el promedio	6
Promedio	3
Sobre el promedio	6
Alto	7

55. Al hacer el presupuesto de gastos para el próximo año en una universidad, se obtuvieron los siguientes pronósticos de gastos (dados en millones de dólares) \$9, \$10, \$11, \$12 y \$13. Como no se sabe cuáles son los gastos actuales, a los gastos calculados se les asignaron las probabilidades 0.3, 0.2, 0.25, 0.05 y 0.2.
- Dé la distribución de probabilidad para estos pronósticos de gastos.
 - ¿Cuál es el valor esperado en estos pronósticos de gastos?
 - ¿Cuál es la varianza en el pronóstico de gastos para el año próximo?
 - Si las proyecciones de ingreso estiman que éste será de \$12 millones, ¿cómo será la situación financiera de la universidad?
56. En un estudio realizado por la Bureau of Transportation Statistics se encontró que, en promedio, la duración del recorrido de la casa al trabajo de una persona es de 26 minutos. También que 5% de las personas necesitan más de una hora para transportarse de su casa al trabajo.
- Si interroga a 20 de estas personas, ¿cuál es la probabilidad de que informen que necesitan más de una hora para ir de su casa al trabajo?
 - Si interroga a 20 de estas personas, ¿cuál es la probabilidad de que ninguna de ellas informe que necesita más de una hora para ir de su casa al trabajo?

- c. Si en una empresa hay 2000 empleados, ¿cuál es el número esperado de empleados que necesita más de una hora para trasladarse de su casa al trabajo?
 - d. Si en una empresa hay 2000 empleados, ¿cuál es la varianza y la desviación estándar del número de empleados que necesitan más de una hora para trasladarse de su casa al trabajo.
57. Una empresa piensa entrevistar a los usuarios de Internet para saber cómo será recibida su página por los grupos de las distintas edades. De acuerdo con la Census Bureau, 40% de las personas entre 18 y 54 años y 12% de las personas de 55 años o más usan Internet.
- a. ¿Cuántas personas entre 18 y 54 años hay que contactar para hallar un número esperado de por lo menos 10 usuarios de Internet?
 - b. ¿Cuántas personas de 55 años o más hay que contactar para hallar un número esperado de por lo menos 10 usuarios de Internet?
 - c. Si se contacta el número de personas entre 18 y 54 años sugerido por el inciso a, ¿cuál es la desviación estándar del número que será usuario de Internet?
 - d. Si se contacta el número de personas de entre 55 años o más sugerido por el inciso b, ¿cuál es la desviación estándar del número de quienes serán usuarios de Internet?
58. Muchas empresas usan una técnica de control de calidad conocida como muestreo de aceptación para vigilar los pedidos que reciben de piezas, materia prima, etc. En la industria electrónica, los componentes se suelen recibir por lotes grandes. La inspección de una muestra de n componentes se considera como n ensayos de un experimento binomial. El resultado de la revisión de cada componente (ensayo) es que el componente sea clasificado como bueno o como defectuoso. Reynolds Electronics acepta el lote de un determinado distribuidor si los componentes defectuosos encontrados en el lote no son más de 1%. Suponga que se prueba una muestra aleatoria de cinco artículos del último lote recibido.
- a. Asuma que 1% del lote recibido está defectuoso. Calcule la probabilidad de que ningún elemento de la muestra esté defectuoso.
 - b. Admita que 1% del lote recibido está defectuoso. Calcule la probabilidad de que exactamente un elemento de la muestra esté defectuoso.
 - c. ¿Cuál es la probabilidad de encontrar uno o más artículos defectuosos si 1% del lote está defectuoso?
 - d. ¿Estaría usted tranquilo al aceptar el lote si se encuentra un artículo defectuoso? ¿Por qué sí o por qué no?
59. La tasa de desempleo es 4.1% (*Barron's*, 4 de septiembre de 2000). Suponga que selecciona aleatoriamente 100 personas empleables.
- a. ¿Cuál es el número esperado de personas que están desempleadas?
 - b. ¿Cuál es la varianza y la desviación estándar del número de personas que están desempleadas?
60. Un sondeo de Zogby encontró que de los estadounidenses para quienes la música es “muy importante” en su vida, 30% dice que su estación de radio “siempre” toca la clase de música que le gusta. Suponga que toma una muestra de 800 personas para quienes la música es muy importante en su vida.
- a. ¿Cuántas afirmarían que su estación de radio siempre toca la música que les gusta?
 - b. ¿Cuál es la desviación estándar del número de interrogados para quienes su estación de radio siempre toca la música que les gusta?
 - c. ¿Cuál es la desviación estándar del número de interrogados para quienes su estación de radio no siempre toca la música que les gusta?
61. A un lavado de coches los automóviles llegan en forma aleatoria e independiente; la probabilidad de una llegada es la misma en cualesquiera dos intervalos de la misma duración. La tasa de llegada media es 15 automóviles por hora. ¿Cuál es la probabilidad de que en una hora cualquiera de operación lleguen 20 o más automóviles?
62. En un proceso nuevo de producción automática hay en promedio 1.5 interrupciones por día. Debido al elevado costo de las interrupciones, los directivos están preocupados por la posibilidad de que en un día haya tres o más interrupciones. Suponga que las interrupciones se presentan en forma aleatoria, que la probabilidad de una interrupción es la misma en cualesquiera dos intervalos de una misma duración y que las interrupciones en un intervalo de tiempo son independientes de

las interrupciones en otro intervalo de tiempo. ¿Cuál es la probabilidad de que haya tres o más interrupciones en un día?

63. Un director regional responsable del desarrollo de los negocios en una determinada área está preocupado por el número de fracasos de pequeños negocios. Si en promedio fracasan 10 pequeños negocios por mes, ¿Cuál es la probabilidad de que exactamente cuatro pequeños negocios fracasen en un mes determinado? Suponga que la probabilidad de fracasos es la misma en cada dos meses que se tomen y que la ocurrencia o no-ocurrencia de fracasos en un determinado mes es independiente de la ocurrencia o no-ocurrencia de fracasos en cualquier otro mes
64. Las llegadas de los clientes a un banco son aleatorias e independientes; la probabilidad de una llegada en un lapso cualquiera de un minuto es la misma que la probabilidad de una llegada en otro lapso cualquiera de un minuto. Conteste las preguntas siguientes suponiendo que la tasa media de llegadas en un lapso de un minuto es tres clientes.
 - a. ¿Cuál es la probabilidad de exactamente tres llegadas en un minuto?
 - b. ¿Cuál es la probabilidad de por lo menos tres llegadas en un minuto?
65. Una baraja contiene 52 cartas, de las cuales cuatro son ases. ¿Cuál es la probabilidad de que en una repartición de cinco cartas haya:
 - a. Un par de ases?
 - b. Exactamente un as?
 - c. Ningún as?
 - d. Por lo menos un as?
66. En la semana que terminó el 16 de septiembre de 2001, Tiger Woods estuvo a la cabeza en ganancia de dinero en el PGA Tour, con una ganancia total de \$5 517 777. De los 10 principales jugadores en ganancias de dinero siete usaron pelotas de golf de la marca Titleist (www.pgatour.com). Suponga que toma al azar a dos de estos principales ganadores.
 - a. ¿Cuál es la probabilidad de que exactamente uno use una pelota de golf de la marca Titleist?
 - b. ¿De que los dos usen una pelota de golf de la marca Titleist?
 - c. ¿De que ninguno use una pelota de golf de la marca Titleist?

Apéndice 5.1 Distribuciones de probabilidad con Minitab

Los paquetes para estadística como Minitab ofrecen procedimientos relativamente fáciles y eficientes para calcular probabilidades binomiales. En este apéndice se muestra paso a paso el procedimiento para hallar las probabilidades binomiales del problema de la tienda de ropa Martin Clothing Store de la sección 5.4. Recuerde que las probabilidades binomiales deseadas son para $n = 10$ y $p = 0.30$. Antes de empezar con la rutina de Minitab, el usuario debe ingresar los valores deseados de la variable aleatoria en una columna de la hoja de cálculo. Aquí se han ingresado los valores 0, 1, 2, . . . , 10 en la columna 1 (véase la figura 5.5) para generar la distribución de probabilidad binomial completa. Los pasos para obtener las probabilidades binomiales deseadas usando Minitab son los siguientes.

Paso 1. Seleccionar el menú **Calc**

Paso 2. Elegir **Probability distributions**

Paso 3. Elegir **Binomial**

Paso 4. Cuando aparezca el cuadro de diálogo Binomial Distribution:

Seleccionar **Probability**

Ingresar 10 en el cuadro **Number of trials**

Ingresar 0.3 en el cuadro **Probability of succes**

Ingresar C1 en el cuadro **Input column**

Clic en **OK**

El resultado que da Minitab con las probabilidades binomiales aparecerá como se muestra en la figura 5.5.

De manera similar, Minitab proporciona probabilidades de Poisson e hipergeométricas. Por ejemplo, para calcular probabilidades de Poisson, las únicas diferencias están en el paso 3, en el que se deberá seleccionar la opción **Poisson** y en el paso 4, en el que se deberá ingresar **Mean** en lugar del número de ensayos y la probabilidad de éxito

Apéndice 5.2 Distribuciones de probabilidad discreta con Excel

Excel proporciona funciones para calcular las probabilidades de las distribuciones binomial, de Poisson e hipergeométrica tratadas en este capítulo. La función de Excel para calcular probabilidades binomiales es DISTR.BINOM. Esta función tiene cuatro argumentos: x (el número de éxitos), n (el número de ensayos), p (la probabilidad de éxito) y acumulado. Se usa FALSO como cuarto argumento (acumulado) si se quiere la probabilidad de x éxitos y VERDADERO se usa como cuarto argumento si se desea la probabilidad acumulada de x o menos éxitos. A continuación se muestra cómo calcular la probabilidad de 0 a 10 éxitos en el caso del problema de la tienda de ropa Martin Clothing Store de la sección 5.4 (véase figura 5.5).

A medida que se describe la elaboración de la hoja de cálculo consulte la figura 5.6; la hoja de cálculo con las fórmulas aparece en segundo plano y la hoja de cálculo con los valores en primer plano. En la celda B1 ingrese el número de ensayos (10), en la celda B2 la probabilidad de

FIGURA 5.6 HOJA DE CÁLCULO DE EXCEL PARA CALCULAR PROBABILIDADES BINOMIALES

	A	B	C	D
1	Number of Trials (n)	10		
2	Probability of Success (p)	0.3		
3				
4		x	$f(x)$	
5		0	=BINOMDIST(B5,\$B\$1,\$B\$2,FALSE)	
6		1	=BINOMDIST(B6,\$B\$1,\$B\$2,FALSE)	
7		2	=BINOMDIST(B7,\$B\$1,\$B\$2,FALSE)	
8		3	=BINOMDIST(B8,\$B\$1,\$B\$2,FALSE)	
9		4	=BINOMDIST(B9,\$B\$1,\$B\$2,FALSE)	
10		5	=BINOMDIST(B10,\$B\$1,\$B\$2,FALSE)	
11		6	=BINOMDIST(B11,\$B\$1,\$B\$2,FALSE)	
12		7	=BINOMDIST(B12,\$B\$1,\$B\$2,FALSE)	
13		8	=BINOMDIST(B13,\$B\$1,\$B\$2,FALSE)	
14		9	=BINOMDIST(B14,\$B\$1,\$B\$2,FALSE)	
15		10	=BINOMDIST(B15,\$B\$1,\$B\$2,FALSE)	
16				

	A	B	C	D
1	Number of Trials (n)	10		
2	Probability of Success (p)	0.3		
3				
4		x	$f(x)$	
5		0	0.0282	
6		1	0.1211	
7		2	0.2335	
8		3	0.2668	
9		4	0.2001	
10		5	0.1029	
11		6	0.0368	
12		7	0.0090	
13		8	0.0014	
14		9	0.0001	
15		10	0.0000	
16				

éxito y en las celdas B5:B15 los valores de la variable aleatoria. Con los pasos siguientes generará las probabilidades deseadas.

Paso 1. Usar la función DISTR.BINOM para calcular la probabilidad de $x = 0$ ingresando la fórmula siguiente en la celda C5:

=BINOMDIST(B5,\$B\$1,\$B\$2,FALSO)

Paso 2. Copiar la fórmula de la celda C5 en las celdas C6:C15.

La hoja de cálculo con los valores en la figura 5.6 muestra que las probabilidades obtenidas son las mismas que aparecen en la figura 5.5. Las probabilidades de Poisson e hipergeométrica se calculan de manera similar. Se emplean las funciones POISSON y DISTR.HIPERGEOM. La herramienta de Excel Insertar función puede ayudar al usuario a ingresar los argumentos adecuados para estas funciones (véase apéndice 2.2).

CAPÍTULO 6



Distribuciones de probabilidad continua

CONTENIDO

LA ESTADÍSTICA EN

LA PRÁCTICA:

PROCTER & GAMBLE

6.1 DISTRIBUCIÓN DE PROBABILIDAD UNIFORME

Áreas como medida de probabilidad

6.2 DISTRIBUCIÓN DE PROBABILIDAD NORMAL

Curva normal

Distribución de probabilidad normal estándar

Cálculo de probabilidades en cualquier distribución de probabilidad normal

El problema de la empresa Gear Tire

6.3 APROXIMACIÓN NORMAL DE LAS PROBABILIDADES BINOMIALES

6.4 DISTRIBUCIÓN DE PROBABILIDAD EXPONENCIAL

Cálculo de probabilidades para la distribución exponencial

Relación entre la distribución de Poisson y la exponencial



LA ESTADÍSTICA *en* LA PRÁCTICA

PROCTER & GAMBLE* CINCINNATI, OHIO

Procter & Gamble (P&G) produce y comercializa productos como detergentes, pañales desechables, productos farmacéuticos que no requieren receta, dentífricos, jabones de tocador y toallas de papel. En todo el mundo P&G tiene la marca líder en más categorías que cualquiera otra empresa de productos de consumo. Desde su fusión con Gillette, P&G también comercializa rasuradoras, navajas para afeitar y muchos otros productos para el cuidado personal.

Al ser uno de los líderes en aplicación de los métodos estadísticos para la toma de decisiones, P&G emplea personas con diversas formaciones académicas: ingenieros, especialistas en estadística, en investigación de operaciones y en negocios. Las principales tecnologías cuantitativas en las que estos profesionistas aplican sus conocimientos son decisiones probabilísticas y análisis de riesgos, simulación avanzada, mejoramiento de la calidad y métodos cuantitativos (por ejemplo, programación lineal, análisis de regresión, análisis de probabilidad).

La División de Productos Químicos para la Industria de P&G es una de las principales proveedoras de alcoholes grasos obtenidos de sustancias naturales, como el aceite de coco, y de derivados del petróleo. La división deseaba saber qué riesgos económicos y cuáles oportunidades existen para la expansión de sus instalaciones dedicadas a la producción de alcoholes grasos; por tanto, solicitó la ayuda de los expertos de P&G en decisiones probabilísticas y en análisis de riesgos. Después de estructurar y modelar el problema, los expertos determinaron que la clave para la rentabilidad era la diferencia entre los costos de las materias primas provenientes del petróleo y del coco. Los costos futuros no se podrían saber, pero los analistas los calcularon mediante las siguientes variables aleatorias continuas.

x = precio del aceite de coco por libra
de alcoholes grasos

y

y = precio de la materia prima proveniente
del petróleo por libra de alcoholes grasos



Algunos de los muchos productos de Procter & Gamble son bien conocidos. © AFP/Getty Images.

Como la clave de la rentabilidad era la diferencia entre estas dos variables aleatorias, se empleó una tercera variable aleatoria para el análisis $d = x - y$. Para determinar las distribuciones de probabilidad de x y y entrevistaron a varios expertos. Después, esta información se empleó para elaborar una distribución de probabilidad de la diferencia entre los precios d . En esta distribución de probabilidad continua se encontró que la probabilidad de que la diferencia entre los precios fuera \$0.0655 o menos, era 0.90 y que la probabilidad de que la diferencia entre los precios fuera \$0.035 o menos era 0.50. Además, la probabilidad de que la diferencia fuera \$0.0045 o menos era sólo 0.10.[†]

La dirección de esta división pensó que la clave para alcanzar un consenso estaba en poder cuantificar el impacto de las diferencias entre los precios de las materias primas. Las probabilidades obtenidas se usaron en un análisis sensible a la diferencia entre los precios de las materias primas. Este análisis arrojó suficiente información como para sustentar una recomendación para los directivos.

Usar variables aleatorias continuas y sus distribuciones de probabilidad ayudó a P&G a analizar los riesgos económicos relacionados con su producción de alcoholes grasos. En este capítulo el lector conocerá las variables aleatorias continuas y sus distribuciones de probabilidad, entre ellas una de las distribuciones de probabilidad más importantes en la estadística, la distribución normal.

*Los autores agradecen a Joel Kahn de P&G por proporcionar este artículo para *La estadística en la práctica*.

[†]Las diferencias de precios dadas aquí están modificadas para proteger los datos.

En el capítulo anterior se estudiaron las variables aleatorias discretas y sus distribuciones de probabilidad. En este capítulo se tratan las variables aleatorias continuas. En específico verá tres distribuciones de probabilidad continua: la uniforme, la normal y la exponencial.

Una diferencia fundamental entre las variables aleatorias discretas y las variables aleatorias continuas es cómo se calculan las probabilidades. En las variables aleatorias discretas la función de probabilidad $f(x)$ da la probabilidad de que la variable aleatoria tome un valor determinado. En las variables aleatorias continuas, la contraparte de la función de probabilidad es la **función de densidad de probabilidad**, que también se denota $f(x)$. La diferencia está en que la función de densidad de probabilidad no da probabilidades directamente. Si no que el área bajo la curva de $f(x)$ que corresponde a un intervalo determinado proporciona la probabilidad de que la variable aleatoria tome uno de los valores de ese intervalo. De manera que cuando se calculan probabilidades de variables aleatorias continuas se calcula la probabilidad de que la variable aleatoria tome alguno de los valores dentro de un intervalo.

Como en cualquier punto determinado el área bajo la gráfica de $f(x)$ es cero, una de las consecuencias de la definición de la probabilidad de una variable aleatoria continua es que la probabilidad de cualquier valor determinado de la variable aleatoria es cero. Estos conceptos se demuestran en la sección 6.1 con una variable que tiene una distribución uniforme.

Gran parte del capítulo se dedica a describir y mostrar aplicaciones de la distribución normal. La distribución normal es muy importante por tener muchas aplicaciones y un amplio uso en la inferencia estadística. El capítulo concluye con el estudio de la distribución exponencial. La distribución exponencial es útil en aplicaciones en las que intervienen factores como tiempos de espera y tiempos de servicios.

6.1

Distribución de probabilidad uniforme

Siempre que una probabilidad sea proporcional a la longitud del intervalo, la variable aleatoria estará distribuida uniformemente.

Considere una variable aleatoria x que representa el tiempo de vuelo de un avión que viaja de Chicago a Nueva York. Suponga que el tiempo de vuelo es cualquier valor en el intervalo de 120 minutos a 140 minutos. Dado que la variable aleatoria x toma cualquier valor en este intervalo, x es una variable aleatoria continua y no una variable aleatoria discreta. Admita que cuenta con datos suficientes como para concluir que la probabilidad de que el tiempo de vuelo esté en cualquier intervalo de 1 minuto es el mismo que la probabilidad de que el tiempo de vuelo esté en cualquier otro intervalo de 1 minuto dentro del intervalo que va de 120 a 140 minutos. Como cualquier intervalo de 1 minuto es igual de probable, se dice que la variable aleatoria x tiene una **distribución de probabilidad uniforme**. La función de densidad de probabilidad que define la distribución uniforme de la variable aleatoria tiempo de vuelo, es

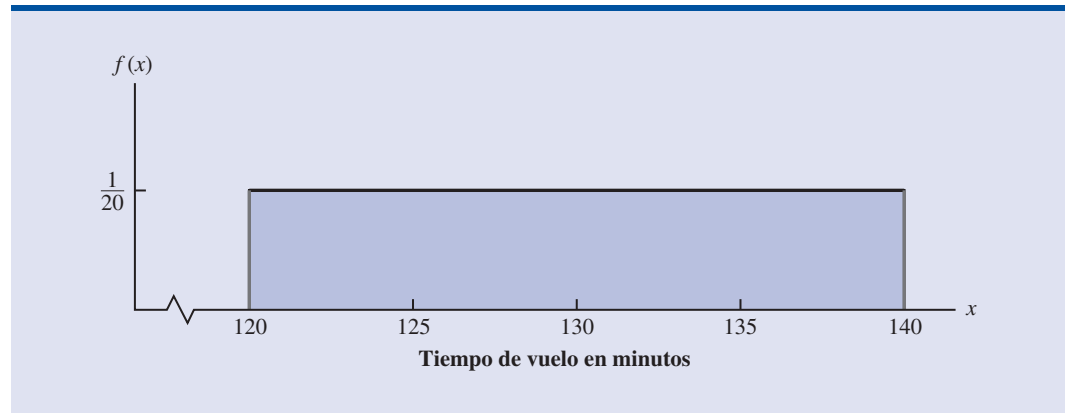
$$f(x) = \begin{cases} 1/20 & \text{para } 120 \leq x \leq 140 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La figura 6.1 es una gráfica de esta función de densidad de probabilidad. En general, la función de densidad de probabilidad uniforme de una variable aleatoria x se define mediante la fórmula siguiente.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD UNIFORME

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{en cualquier otro caso} \end{cases} \quad (6.1)$$

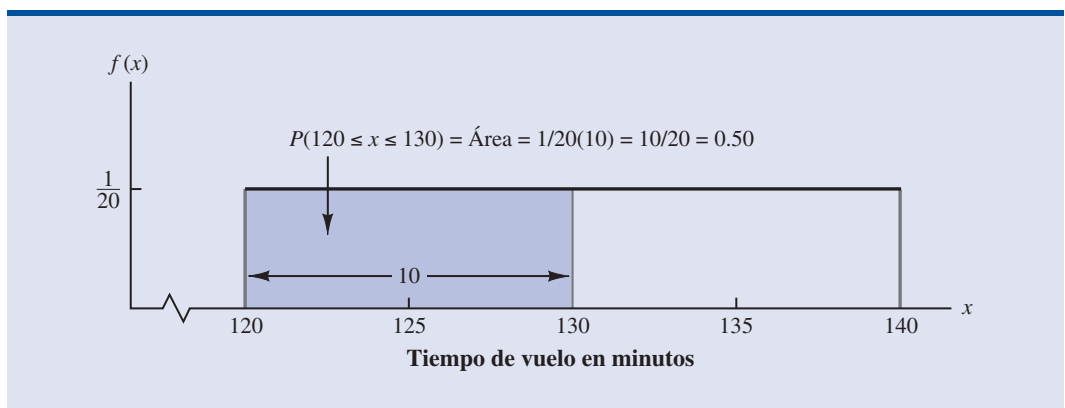
En el caso de la variable aleatoria tiempo de vuelo, $a = 120$ y $b = 140$.

FIGURA 6.1 DISTRIBUCIÓN DE PROBABILIDAD UNIFORME PARA EL TIEMPO DE VUELO

Como se hizo notar en la introducción, en el caso de una variable aleatoria continua, sólo se considera la probabilidad en términos de la posibilidad de que la variable aleatoria tome un valor dentro de un determinado intervalo. En el ejemplo del tiempo de vuelo, una pregunta aceptable acerca de una probabilidad es: ¿Cuál es la probabilidad de que el tiempo de vuelo se encuentre entre 120 y 130 minutos? Es decir, ¿cuál es $P(120 \leq x \leq 130)$? Como el tiempo de vuelo debe estar entre 120 y 140 minutos y como se ha dicho que la probabilidad es uniforme en este intervalo, es factible decir que $P(120 \leq x \leq 130) = 0.50$. En la sección siguiente se muestra que esta probabilidad se calcula como el área bajo la gráfica de $f(x)$ desde 120 hasta 130 (véase figura 6.2)

Áreas como medida de probabilidad

Ahora una observación acerca de la gráfica de la figura 6.2. Considere el área bajo la gráfica de $f(x)$ en el intervalo que va de 120 a 130. Esta área es rectangular y el área de un rectángulo es simplemente el ancho multiplicado por la altura. Si el ancho del intervalo es igual a $130 - 120 = 10$ y la altura es igual al valor de la función de densidad de probabilidad $f(x) = 1/20$, se tiene, $\text{área} = \text{ancho} \times \text{alto} = 10(1/20) = 10(1/20) = 10/20 = 0.50$.

FIGURA 6.2 EL ÁREA PROPORCIONA LA PROBABILIDAD DE QUE EL TIEMPO DE VUELO ESTÉ ENTRE 120 Y 130 MINUTOS.

¿Qué observación se puede hacer acerca de la área bajo la curva de $f(x)$ y la probabilidad? ¡Son idénticas! En efecto, esta observación es correcta y válida para todas las variables aleatorias continuas. Una vez que se ha dado la función de densidad de probabilidad $f(x)$, la probabilidad de que x tome un valor entre algún valor menor x_1 y otro valor mayor x_2 se encuentra calculando el área bajo la gráfica de $f(x)$ y sobre el intervalo de x_1 a x_2 .

Dada la distribución uniforme del tiempo de vuelo y usando la interpretación de área como probabilidad es posible contestar cualquier pregunta acerca de la probabilidad de los tiempos de vuelo. Por ejemplo, ¿cuál es la probabilidad de un tiempo de vuelo entre 128 y 136 minutos? El ancho del intervalo es $136 - 128 = 8$. Como la altura uniforme de $f(x) = 1/120$, se ve que $P(128 \leq x \leq 136) = 8(1/20) = 0.40$.

Observe que $P(120 \leq x \leq 140) = 20(1/20) = 1$; es decir, el área total bajo la gráfica de $f(x)$ es igual a 1. Esta propiedad es válida para todas las distribuciones de probabilidad continua y es el análogo de la condición de que la suma de las probabilidades debe ser igual a 1 en el caso de una función de probabilidad discreta.

Dos diferencias importantes sobresalen entre el tratamiento de una variable aleatoria continua y el tratamiento de una variable aleatoria discreta.

1. Ya no se habla de la probabilidad de que una variable aleatoria tome un determinado valor. Se habla de la probabilidad de que una variable aleatoria tome un valor dentro de un intervalo dado.
2. La probabilidad de que una variable aleatoria continua tome un valor dentro de un determinado intervalo que va de x_1 a x_2 se define como el área bajo la gráfica de la función de densidad de probabilidad entre x_1 y x_2 . Como un solo punto es un intervalo cuyo ancho es cero, esto implica que la probabilidad de que una variable aleatoria continua tome un valor exacto, cualquiera, es cero. Esto también significa que en cualquier intervalo la probabilidad de que una variable aleatoria continua tome un valor es la misma, ya sea que se incluyan o no los extremos del intervalo.

Para ver que la probabilidad de un solo punto es 0, consulte la figura 6.2 y calcule la probabilidad de un solo punto, por ejemplo, $x = 125$. $P(x = 125) = P(125 \leq x \leq 125) = 0(1/20) = 0$.

El cálculo del valor esperado y de la varianza de una variable aleatoria continua es análogo al de una variable aleatoria discreta. Sin embargo, como en este caso interviene el cálculo integral la deducción de estas fórmulas queda para cursos más avanzados.

En el caso de la distribución de probabilidad continua uniforme presentada en esta sección, las fórmulas para el valor esperado y para la varianza son

$$E(x) = \frac{a + b}{2}$$

$$\text{Var}(x) = \frac{(b - a)^2}{12}$$

En estas fórmulas a es el menor valor y b es el mayor valor que toma la variable aleatoria.

Al aplicar estas fórmulas a la distribución uniforme de los tiempos de vuelo de Chicago a Nueva York, se obtiene,

$$E(x) = \frac{(120 + 140)}{2} = 130$$

$$\text{Var}(x) = \frac{(140 - 120)^2}{12} = 33.33$$

La desviación estándar de los tiempos de vuelo se encuentra sacando la raíz cuadrada de la varianza. Por tanto, $\sigma = 5.77$ minutos.

NOTAS Y COMENTARIOS

Para ver más claramente por qué la altura de una función de densidad de probabilidad no es una probabilidad, considere la variable aleatoria cuya distribución de probabilidad uniforme es la siguiente.

$$f(x) = \begin{cases} 2 & \text{para } 0 \leq x \leq 0.5 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

La altura de la función de densidad de probabilidad, $f(x)$ es 2 para todos los valores de x entre 0 y 0.5. Pero se sabe que las probabilidades nunca pueden ser mayores a 1. Por tanto, $f(x)$ no se interpreta como la probabilidad de x .

Ejercicios

Métodos

Autoexamen

- La variable aleatoria x está distribuida uniformemente entre 1.0 y 1.5.
 - Dé la gráfica de la función de densidad de probabilidad.
 - Calcule $P(x = 1.25)$.
 - Calcule $P(1.0 \leq x \leq 1.25)$.
 - Calcule $P(1.20 < x < 1.5)$.
- La variable aleatoria x está distribuida uniformemente entre 10 y 20.
 - Dé la gráfica de la función de densidad de probabilidad.
 - Calcule $P(x < 15)$.
 - Calcule $P(12 \leq x \leq 18)$.
 - Calcule $E(x)$.
 - Calcule $\text{Var}(x)$.

Aplicaciones

Autoexamen

- En su vuelo de Cincinnati a Tampa, Delta Airlines da como tiempo de vuelo 2 horas, 5 minutos. En realidad los tiempos de vuelo están distribuidos uniformemente entre 2 horas y 2 horas, 20 minutos.
 - Dé la gráfica de la función de densidad de probabilidad del tiempo de vuelo.
 - ¿Cuál es la probabilidad de que un vuelo no se retrase más de 5 minutos?
 - ¿De que un vuelo no se retrase más de 10 minutos?
 - ¿Cuál es el tiempo de vuelo esperado?
- La mayoría de los lenguajes de computadora tienen una función para generar números aleatorios. En Excel, la función ALEATORIO se usa para generar números aleatorios entre 0 y 1. Si x denota un número aleatorio generado mediante ALEATORIO, entonces x es una variable aleatoria continua, cuya función de densidad de probabilidad es la siguiente.

$$f(x) = \begin{cases} 1 & \text{para } 0 \leq x \leq 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

- Haga la gráfica de la función de densidad de probabilidad.
- ¿Cuál es la probabilidad de generar un número aleatorio entre 0.25 y 0.75?
- ¿De generar un número aleatorio menor o igual que 0.30?
- ¿De generar un número aleatorio mayor o igual que 0.60?
- Genere 50 números aleatorios ingresando = ALEATORIO() en 50 celdas de una hoja de cálculo de Excel.
- Calcule la media y la desviación estándar de los números del inciso e.

5. La *driving distance* de los 100 mejores golfistas del Tour PGA está entre 284.7 y 310.6 yardas (*Golfweek*, 29 de marzo de 2003). Suponga que las *driving distance* de estos golfistas se encuentran uniformemente distribuidas en este intervalo.
 - a. Dé una expresión matemática de la función de densidad de probabilidad correspondiente a estas *driving distance*
 - b. ¿Cuál es la probabilidad de que la *driving distance* de uno de estos golfistas sea menor que 290 yardas?
 - c. ¿De que la *driving distance* de uno de estos golfistas sea por lo menos de 300 yardas?
 - d. ¿De que la *driving distance* de uno de estos golfistas esté entre 290 y 305 yardas?
 - e. ¿Cuántos de estos jugadores lanzan la pelota por lo menos a 290 yardas?
6. En las botellas de un detergente líquido se indica que el contenido es de 12 onzas por botella. En la operación de producción se llenan las botellas uniformemente de acuerdo con la siguiente función de densidad de probabilidad.

$$f(x) = \begin{cases} 8 & \text{para } 11.975 \leq x \leq 12.100 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

- a. ¿Cuál es la probabilidad de que el contenido de una botella esté entre 12 y 12.05 onzas?
 - b. ¿De que el contenido de una botella sea 12.02 onzas o más?
 - c. En el control de calidad se acepta que una botella sea llenada con más o menos 0.02 onzas de lo indicado en la etiqueta. ¿Cuál es la probabilidad de que una de las botellas de detergente no satisfaga estos estándares?
7. Suponga que quiere comprar un terreno y sabe que también hay otros compradores interesados.* El vendedor revela que aceptará la oferta mayor que sea superior a \$10 000. Si la oferta del competidor x es una variable aleatoria que está uniformemente distribuida entre \$10 000 y \$15 000.
 - a. Asuma que usted ofrece \$12 000. ¿Cuál es la probabilidad de que su oferta sea aceptada?
 - b. Si usted ofrece \$14 000. ¿Cuál es la probabilidad de que su oferta sea aceptada?
 - c. ¿Cuál es la cantidad que deberá ofrecer para maximizar la probabilidad de obtener la propiedad?
 - d. Suponga que conoce a quien está dispuesto a pagar \$16 000 por la propiedad. ¿Consideraría la posibilidad de ofrecer una cantidad menor que la del inciso c?

6.2

Distribución de probabilidad normal

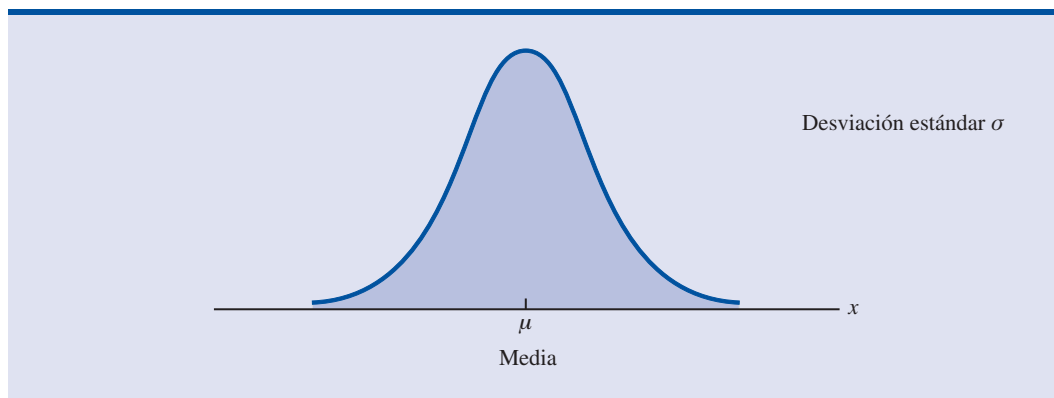
Abraham de Moivre, un matemático francés, publicó en 1733 Doctrina de las posibilidades. De Moivre dedujo la distribución normal.

La distribución de probabilidad más usada para describir variables aleatorias continuas es la **distribución de probabilidad normal**. La distribución normal tiene gran cantidad de aplicaciones prácticas, en las cuales la variable aleatoria puede ser el peso o la estatura de las personas, puntuaciones de exámenes, resultados de mediciones científicas, precipitación pluvial u otras cantidades similares. La distribución normal también tiene una importante aplicación en inferencia estadística, tema principal del resto de este libro. En estas aplicaciones, la distribución normal describe qué tan probables son los resultados obtenidos de un muestreo

Curva normal

En la figura 6.3 aparece la forma de la distribución normal, una curva normal en forma de campana. A continuación se presenta la función de densidad de probabilidad que define la curva en forma de campana de la distribución normal.

* Este ejercicio está basado en un problema sugerido por el profesor Roger Myerson de la Universidad de Northwestern.

FIGURA 6.3 CURVA EN FORMA DE CAMPANA DE UNA DISTRIBUCIÓN NORMAL**FUNCIÓN DE DENSIDAD DE PROBABILIDAD NORMAL**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (6.2)$$

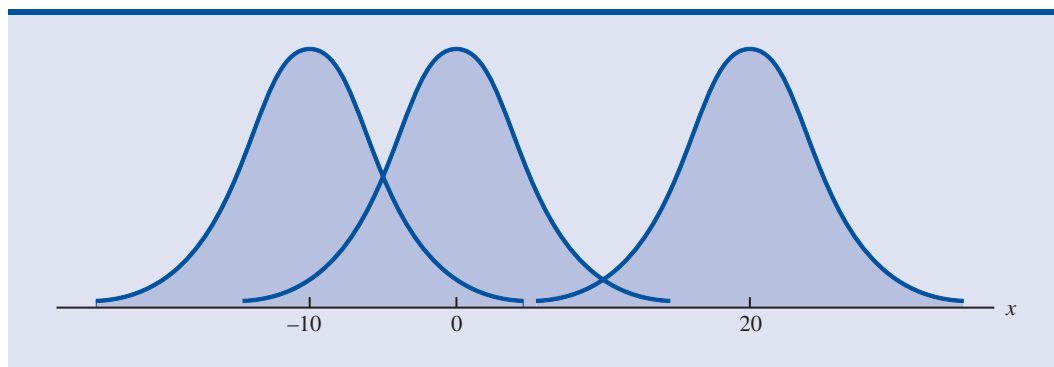
donde

 μ = media σ = desviación estándar π = 3.14159 e = 2.71828

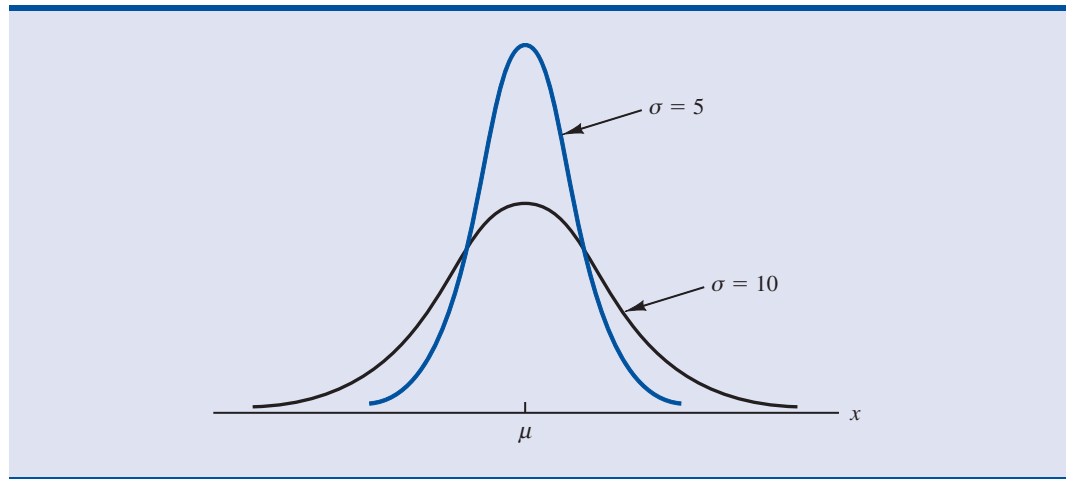
Las siguientes son observaciones importantes acerca de las características de las distribuciones normales.

La curva normal tiene dos parámetros, μ y σ . Estos parámetros determinan la localización y la forma de la distribución normal.

1. Toda la familia de distribuciones normales se diferencia por medio de dos parámetros: la media μ y la desviación estándar σ .
2. El punto más alto de una curva normal se encuentra sobre la media, la cual coincide con la mediana y la moda.
3. La media de una distribución normal puede tener cualquier valor: negativo, positivo o cero. A continuación se muestran tres distribuciones normales que tienen la misma desviación estándar, pero diferentes medias. (-10 , 0 y 20).



4. La distribución normal es simétrica, siendo la forma de la curva normal al lado izquierdo de la media, la imagen especular de la forma al lado derecho de la media. Las colas de la curva normal se extienden al infinito en ambas direcciones y en teoría jamás tocan el eje horizontal. Dado que es simétrica, la distribución normal no es sesgada; su sesgo es cero.
5. La desviación estándar determina qué tan plana y ancha es la curva normal. Desviaciones estándar grandes corresponden a curvas más planas y más anchas, lo cual indica mayor variabilidad en los datos. A continuación se muestran dos curvas normales que tienen la misma media pero distintas desviaciones estándar.



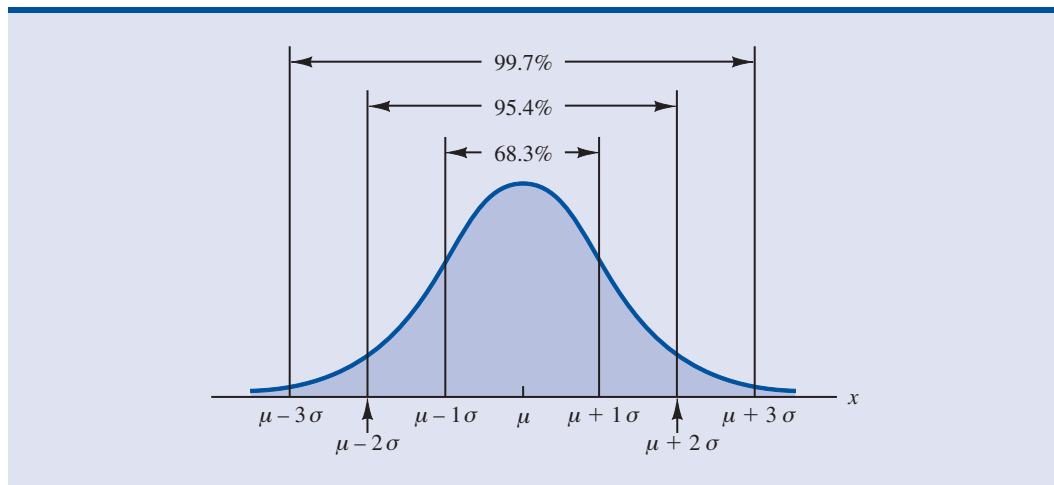
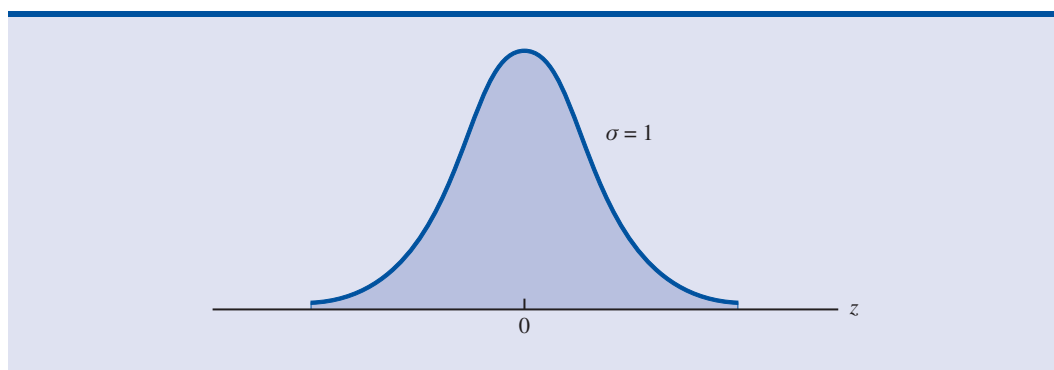
6. Las probabilidades correspondientes a la variable aleatoria normal se dan mediante áreas bajo la curva normal. Toda el área bajo la curva de una distribución normal es 1. Como esta distribución es simétrica, el área bajo la curva y a la izquierda de la media es 0.50 y el área bajo la curva y a la derecha de la media es 0.50.
7. Los porcentajes de los valores que se encuentran en algunos intervalos comúnmente usados son:
 - a. 68.3% de los valores de una variable aleatoria normal se encuentran más o menos una desviación estándar de la media.
 - b. 95.4% de los valores de una variable aleatoria normal se encuentran más o menos dos desviaciones estándar de la media.
 - c. 99.7% de los valores de una variable aleatoria normal se encuentran más o menos tres desviaciones estándar de la media.

Estos porcentajes son la base de la regla empírica que se presentó en la sección 3.3.

En la figura 6.4 aparece una gráfica de las propiedades a, b y c.

Distribución de probabilidad normal estándar

Una variable aleatoria que tiene una distribución normal con una media cero y desviación estándar de uno tiene una **distribución normal estándar**. Para designar esta variable aleatoria normal se suele usar la letra z . La figura 6.5 es la gráfica de la distribución normal estándar. Esta distribución tiene el mismo aspecto general que cualquier otra distribución normal, pero tiene las propiedades especiales, $\mu = 0$ y $\sigma = 1$.

FIGURA 6.4 ÁREAS BAJO LA CURVA DE CUALQUIER DISTRIBUCIÓN NORMAL**FIGURA 6.5** DISTRIBUCIÓN NORMAL ESTÁNDAR

Dado que $\mu = 0$ y $\sigma = 1$, la fórmula de la función de densidad de probabilidad normal estándar es una versión más simple de la ecuación (6.2).

FUNCIÓN DE DENSIDAD NORMAL ESTÁNDAR

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Como ocurre con otras variables aleatorias continuas, los cálculos de la probabilidad en cualquier distribución normal se hacen calculando el área bajo la gráfica de la función de densidad de probabilidad. Por tanto, para hallar la probabilidad de que una variable aleatoria normal esté dentro de un determinado intervalo, se tiene que calcular el área que se encuentra bajo la curva normal y sobre ese intervalo.

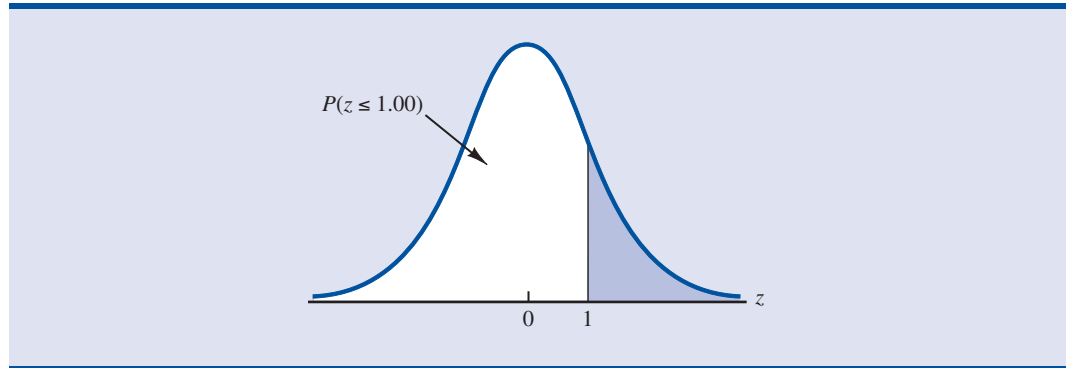
Para la distribución normal estándar ya se encuentran calculadas las áreas bajo la curva normal y se cuenta con tablas que dan estas áreas y que se usan para calcular las probabilidades. Estas tablas se encuentran en los forros interiores al inicio del libro. La tabla del forro izquierdo contiene áreas, o probabilidades acumuladas, correspondientes a valores de z menores o iguales a la media, cero. La tabla siguiente contiene áreas, o probabilidades acumuladas, correspondientes a valores de z mayores o iguales a la media de cero.

En la función de densidad de probabilidad normal, la altura de la curva varía y para calcular las áreas que representan las probabilidades se requiere de matemáticas más avanzadas.

Los tres tipos de probabilidades que se necesitan calcular son: (1) la probabilidad de que la variable aleatoria normal estándar z sea menor o igual que un valor dado; (2) la probabilidad de que z esté entre dos valores dados, y (3) la probabilidad de que z sea mayor o igual que un valor dado. Para mostrar el uso de las tablas de probabilidad acumulada de la distribución normal estándar en el cálculo de estos tres tipos de probabilidades, se consideran algunos ejemplos.

Debido a que la variable aleatoria normal estándar es continua, $P(z \leq 1.00) = P(z < 1.00)$.

Se empieza por mostrar cómo se calcula la probabilidad de que z sea menor o igual a 1.00; es decir $P(z \leq 1.00)$. Esta probabilidad acumulada es el área bajo la curva normal a la izquierda de $z = 1.00$ como se muestra en la gráfica siguiente.

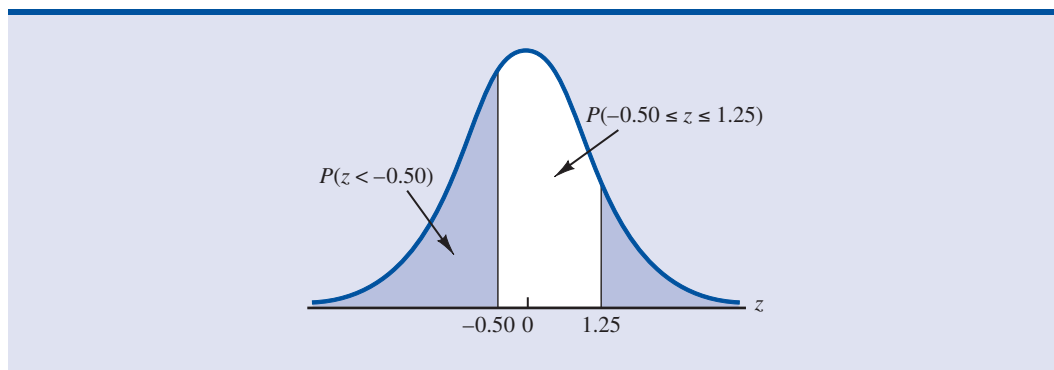


Consulte la página del lado derecho de la tabla de la distribución de probabilidad normal estándar que se encuentra dentro de la cubierta frontal del libro. Esta probabilidad acumulada correspondiente a $z = 1.00$ es el valor que en la tabla se localiza en la intersección del renglón cuyo encabezado es 1.0 y la columna cuyo encabezado es 0.00. Primero localice 1.0 en la columna del extremo izquierdo de la tabla y después localice 0.00 en el renglón en la parte superior de la tabla. Observe que en el interior de la tabla, el renglón 1.0 y la columna 0.00 se cruzan en el valor 0.8413; por tanto, $P(z \leq 1.00) = 0.8413$. Estos pasos se muestran en el extracto siguiente de las tablas de probabilidad.

z	0.00	0.01	0.02
.			
.			
.			
0.9	0.8159	0.8186	0.8212
1.0	0.8413	0.8438	0.8461
1.1	0.8643	0.8665	0.8686
1.2	0.8849	0.8869	0.8888
.			
.			
.			

$P(z \leq 1.00)$

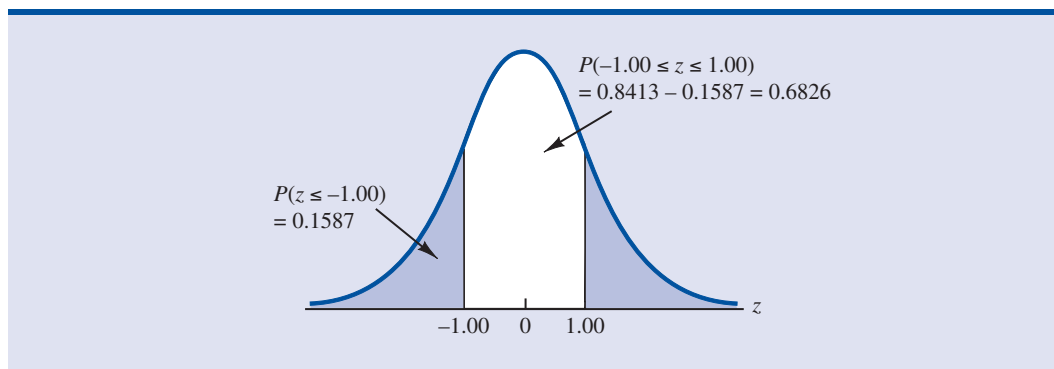
Para ilustrar el segundo tipo de cálculo de una probabilidad se muestra cómo calcular la probabilidad de que z esté en el intervalo entre -0.50 y 1.25 ; esto es, $P(-.50 \leq z \leq 1.25)$. En la gráfica siguiente se muestra esta área o probabilidad.



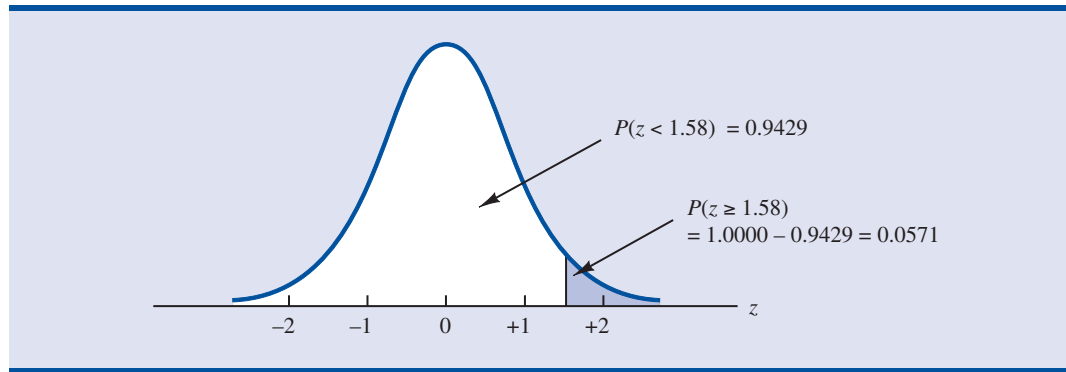
Para calcular esta probabilidad son necesarios tres pasos. Primero, se encuentra el área bajo la curva normal a la izquierda de $z = 1.25$. Segundo, se encuentra el área bajo la curva normal a la izquierda de $z = -0.50$. Por último, se resta el área a la izquierda de $z = -0.50$ del área a la izquierda de $z = 1.25$ y se encuentra, $P(-0.50 \leq z \leq 1.25)$.

Para encontrar el área bajo la curva normal a la izquierda de $z = 1.25$, primero se localiza en la tabla de probabilidad normal estándar el renglón 1.2 y después se avanza por ese renglón hasta la columna 0.05. Como el valor que aparece en el renglón 1.2 columna 0.05 es 0.8944, $P(z \leq 1.25) = 0.8944$. De manera similar, para encontrar el área bajo la curva a la izquierda de $z = -0.50$ se usa el forro izquierdo de la tabla para localizar el valor en el renglón -0.5 columna 0.00; como el valor que se encuentra es 0.3085, $P(z \leq -0.50) = 0.3085$. Por tanto, $P(-0.50 \leq z \leq 1.25) = P(z \leq 1.25) - P(z \leq -0.50) = 0.8944 - 0.3085 = 0.5859$.

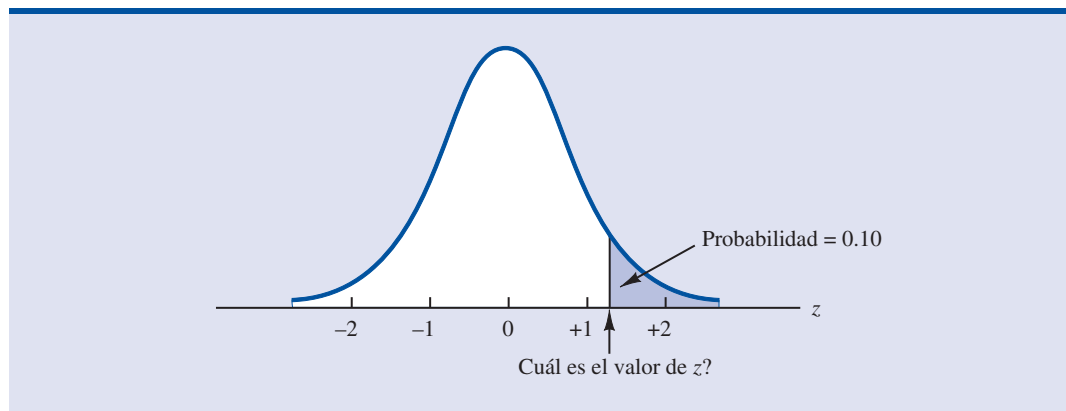
A continuación se presenta otro ejemplo para calcular la probabilidad de que z esté en el intervalo entre dos valores dados. Con frecuencia se desea calcular la probabilidad de que una variable aleatoria normal tome un valor dentro de cierto número de desviaciones estándar respecto a la media. Suponga que desea calcular la probabilidad de que la variable aleatoria normal estándar se encuentre a no más de una desviación estándar de la media; es decir, $P(-1.00 \leq z \leq 1.00)$. Para calcular esta probabilidad tiene que hallar el área bajo la curva entre -1.00 y 1.00 . Antes encontró que $P(z \leq 1.00) = 0.8413$. Si va al forro izquierdo, encontrará que el área bajo la curva a la izquierda de $z = -1.00$ es 0.1587, de manera que $P(z \leq -1.00) = 0.1587$. Por tanto, $P(-1.00 \leq z \leq 1.00) = P(z \leq 1.00) - P(z \leq -1.00) = 0.8413 - 0.1587 = 0.6826$. Esta probabilidad se muestra en forma gráfica en la figura siguiente.



Para ilustrar cómo se calcula el tercer tipo de probabilidad, suponga que desea calcular la probabilidad de tener un valor z por lo menos igual a 1.58; es decir, $P(z \geq 1.58)$. El valor en el renglón $z = 1.5$, columna 0.08 de la tabla normal acumulada es 0.9429; por tanto, $P(z < 1.58) = 0.9429$. Pero, como toda el área bajo la curva normal es 1, $P(z \geq 1.58) = 1 - 0.9429 = 0.0571$. En la figura siguiente se muestra esta probabilidad.



En los ejemplos anteriores se muestra cómo calcular probabilidades dados determinados valores de z . En algunas situaciones se da una probabilidad y se trata de hacer lo contrario, encontrar el correspondiente valor de z . Suponga que desea hallar un valor z tal que la probabilidad de obtener un valor z mayor sea 0.10. En la figura siguiente se muestra en forma gráfica esta situación.



Dada una probabilidad, se puede usar la tabla normal estándar para encontrar el correspondiente valor de z .

Este problema es la situación contraria a la presentada en los ejemplos anteriores, en ellos se dio el valor z y se halló la probabilidad o área correspondiente. En este ejemplo se da la probabilidad, o el área, y se pide hallar el valor correspondiente de z . Para esto se usa la tabla de probabilidad normal estándar de una manera un poco diferente.

Recuerde que la tabla de probabilidad normal estándar da el área bajo la curva a la izquierda de un determinado valor z . Se ha recibido la información de que el área en la cola superior (derecha) de la curva es 0.10. Por tanto, el área bajo la curva a la izquierda del valor desconocido de z debe ser 0.9000. Al recorrer el cuerpo de la tabla, se encuentra que 0.8997 es la probabilidad acumulada más cercana a 0.9000. A continuación se reproduce la sección de la tabla en la que se encuentra este resultado.

z	0.06	0.07	0.08	0.09
.				
.				
.				
1.0	0.8554	0.8577	0.8599	0.8621
1.1	0.8770	0.8790	0.8810	0.8830
1.2	0.8962	0.8980	0.8997	0.9015
1.3	0.9131	0.9147	0.9162	0.9177
1.4	0.9279	0.9292	0.9306	0.9319
.				
.				
.				

Probabilidad acumulada más cercana a 0.9000

Al leer el valor de z en la columna del extremo izquierdo y en el renglón superior de la tabla, se encuentra que el valor de z es 1.28. De manera que un área de aproximadamente 0.9000 (en realidad de 0.8997) es la que se encuentra a la izquierda de $z = 1.28$.* En términos de la pregunta originalmente planteada, 0.10 es la probabilidad aproximada de que z sea mayor que 1.28.

Estos ejemplos ilustran que la tabla de probabilidades acumuladas para la distribución de probabilidad normal estándar se puede usar para hallar probabilidades correspondientes a valores de la variable aleatoria normal estándar z . Es posible hacer dos tipos de preguntas. En el primer tipo de pregunta se dan valores, o un valor de z , y se pide usar la tabla para determinar el área o probabilidad correspondiente. En el segundo tipo de pregunta se da un área, o probabilidad, y se pide usar la tabla para encontrar el correspondiente valor de z . Por tanto, se necesita tener flexibilidad para usar la tabla de probabilidad normal estándar para responder la pregunta deseada. En la mayoría de los casos, hacer un bosquejo de la gráfica de la distribución de probabilidad normal estándar y sombrear el área deseada será una ayuda para visualizar la situación y encontrar la respuesta correcta.

Cálculo de probabilidades en cualquier distribución de probabilidad normal

La razón por la cual la distribución normal estándar se ha visto de manera tan amplia es que todas las distribuciones normales son calculadas mediante la distribución normal estándar. Esto es, cuando distribución normal con una media μ cualquiera y una desviación estándar σ cualquiera, las preguntas sobre las probabilidades en esta distribución se responden pasando primero a la distribución normal estándar. Use las tablas de probabilidad normal estándar y los valores apropiados de z para hallar las probabilidades deseadas. A continuación se da la fórmula que se emplea para convertir cualquier variable aleatoria x con media μ y desviación estándar σ en la variable aleatoria normal estándar z .

La fórmula para la variable aleatoria normal estándar es semejante a la fórmula que se dio en el capítulo 3 para los puntos z de un conjunto de datos.

CONVERSIÓN A LA VARIABLE ALEATORIA NORMAL ESTÁNDAR

$$z = \frac{x - \mu}{\sigma} \quad (6.3)$$

* Se podía haber hecho una interpolación en el cuerpo de la tabla para obtener una aproximación más exacta al valor z que corresponde al área 0.9000. Al hacerlo en busca de un lugar decimal más preciso se obtiene 1.282. Sin embargo, en la mayor parte de las situaciones prácticas, es suficiente con la precisión obtenida usando el valor más cercano al valor deseado que da la tabla.

Un valor x igual a su media μ da como resultado $z = (\mu - \mu)/\sigma = 0$. De manera que un valor x igual a su media corresponde a $z = 0$. Ahora suponga que x se encuentra una desviación estándar arriba de su media. Es decir, $x = \mu + \sigma$. Aplicando la ecuación (6.3) el valor correspondiente es $z = [(\mu + \sigma) - \mu]/\sigma = \sigma/\sigma = 1$. Así que un valor de x que es una desviación estándar mayor que su media corresponde a $z = 1$. En otras palabras, z se interpreta como el número de desviaciones estándar a las que está una variable aleatoria x de su media μ .

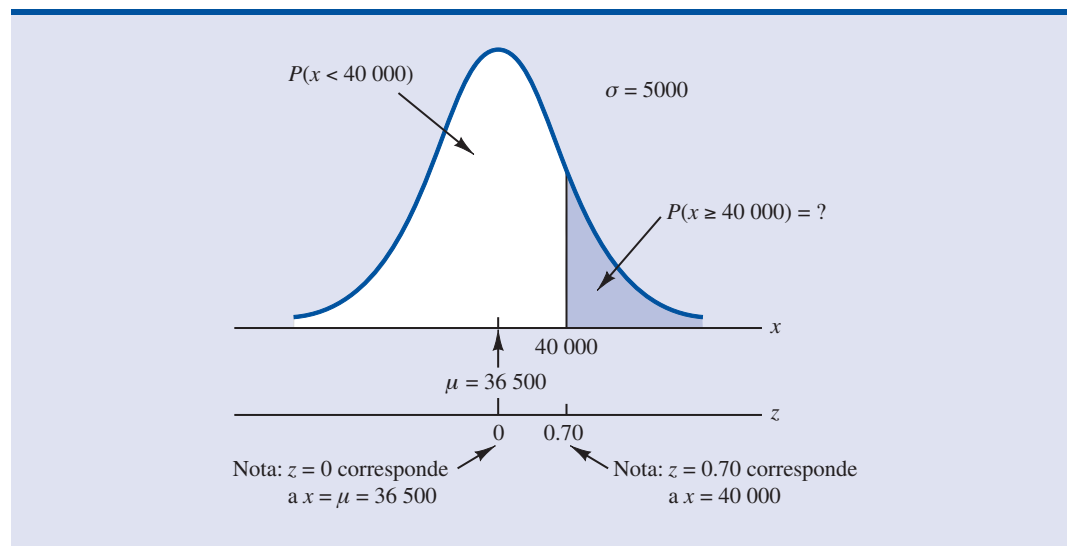
Para ver cómo esta distribución permite calcular probabilidades en cualquier distribución normal, admita que tiene una distribución en la que $\mu = 10$ y $\sigma = 2$. ¿Cuál es la probabilidad de que la variable aleatoria x esté entre 10 y 14? Empleando la ecuación (6.3) se ve que para $z = (x - \mu)/\sigma = (10 - 10)/2 = 0$ y que para $x = 14$, $z = (14 - 10)/2 = 4/2 = 2$. Así, la respuesta a la pregunta acerca de la probabilidad de que x esté entre 10 y 14 está dada por la probabilidad equivalente de que z esté entre 0 y 2 en la distribución normal estándar. En otras palabras, la probabilidad que se está buscando es que la variable aleatoria x esté entre su media y dos desviaciones estándar arriba de la media. Usando $z = 2$ y la tabla de probabilidad normal estándar del forro interior, se ve que $P(z \leq 2) = 0.9772$. Como $P(z \leq 0) = 0.5000$, se tiene que $P(0.00 \leq z \leq 2.00) = P(z \leq 2) - P(z \leq 0) = 0.9772 - 0.5000 = 0.4772$. Por tanto, la probabilidad de que x esté entre 10 y 14 es 0.4772.

El problema de la empresa Grear Tire

Para una aplicación de la distribución de probabilidad normal, suponga que Grear Tire Company ha fabricado un nuevo neumático que será vendido por una cadena nacional de tiendas de descuento. Como este neumático es un producto nuevo, los directivos de Grear piensan que la garantía de duración será un factor importante en la aceptación del neumático. Antes de finalizar la póliza de garantía, los directivos necesitan información probabilística acerca de x = duración del neumático en número de millas.

De acuerdo con las pruebas realizadas al neumático, los ingenieros de Grear estiman que la duración media en millas es $\mu = 36\,500$ millas y que la desviación estándar es $\sigma = 5\,000$. Además, los datos recogidos indican que es razonable suponer una distribución normal. ¿Qué porcentaje de los neumáticos se espera que duren más de 40 000 millas? En otras palabras, ¿cuál es la probabilidad de que la duración de los neumáticos sea superior a 40 000? Esta pregunta se responde hallando el área de la región sombreada que se observa en la gráfica de la figura 6.6.

FIGURA 6.6 DISTRIBUCIÓN DE DURACIÓN EN MILLAS PARA GREAR TIRE COMPANY



Para $x = 40\,000$, se tiene

$$z = \frac{x - \mu}{\sigma} = \frac{40\,000 - 36\,500}{5\,000} = \frac{3\,500}{5\,000} = 0.70$$

Observe que en la parte inferior de la figura 6.6 el valor $x = 40\,000$ en la distribución normal de Grear Tire corresponde a $z = 0.70$ en la distribución normal estándar. Mediante la tabla de probabilidad normal estándar se encuentra que el área bajo la curva normal estándar a la izquierda de $z = 0.70$ es 0.7580. De manera que $1.000 - 0.7580 = 0.2420$ es la probabilidad de que z sea mayor a 0.70 y por tanto de que x sea mayor a 40 000. Entonces 24.2% de los neumáticos durará más de 40 000 millas.

Ahora suponga que Grear está considerando una garantía que dé un descuento en la sustitución del neumático original si éste no dura lo que asegura la garantía. ¿Cuál deberá ser la duración en millas especificada en la garantía si Grear desea que no más de 10% de los neumáticos alcancen la garantía? En la figura 6.7 se plantea esta pregunta en forma gráfica.

De acuerdo con la figura 6.7, el área bajo la curva a la izquierda de la cantidad desconocida de millas para la garantía debe ser 0.10. De manera que primero se debe encontrar el valor de z que deja un área de 0.10 en el extremo de la cola izquierda de la distribución normal estándar. Según la tabla de probabilidad normal estándar $z = -1.28$ deja un área de 0.10 en el extremo de la cola izquierda. Por tanto, $z = -1.28$ es el valor de la variable aleatoria normal estándar que corresponde a las millas de duración deseadas para la garantía en la distribución normal de Grear Tire. Para hallar el valor de x que corresponde a $z = -1.28$, se tiene

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} = -1.28 \\ x - \mu &= -1.28\sigma \\ x &= \mu - 1.28\sigma \end{aligned}$$

Las millas de garantía que se desean encontrar están a 1.28 desviaciones estándar abajo de la media. Por tanto, $x = \mu - 1.28\sigma$.

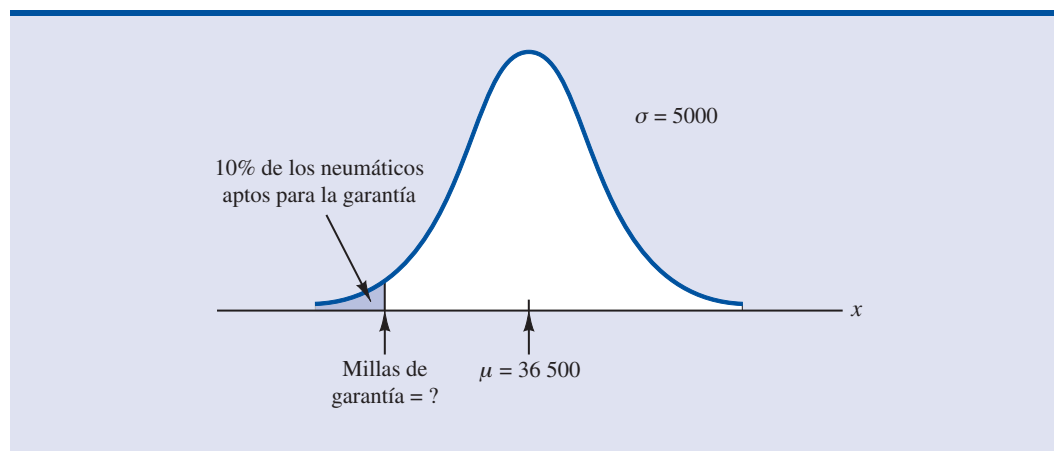
Como $\mu = 36\,500$ y $s = 5\,000$,

$$x = 36\,500 - 1.28(5\,000) = 30\,100$$

Al establecer una garantía a partir de 30 000 millas, el porcentaje real apto para la garantía será 9.68%.

Por tanto, una garantía de 30 100 millas cumplirá con el requerimiento de que aproximadamente 10% de los neumáticos sean aptos para la garantía. Con esta información, quizá la empresa establezca una garantía de 30 000 millas.

FIGURA 6.7 GARANTÍA DE GREAR



Nuevamente, se observa la importancia de las distribuciones de probabilidad en el suministro de información para la toma de decisiones. A saber, una vez que la distribución de probabilidad es establecida para una aplicación en particular, puede ser usada para obtener información probabilística acerca del problema. La probabilidad no recomienda directamente una decisión, pero suministra información que ayuda a tomarla entendiendo mejor los riesgos y la incertidumbre asociados al problema. Finalmente, esta información ayuda a enriquecer una buena decisión.

Ejercicios

Métodos

8. Usando como guía la figura 6.4, dibuje la curva normal de la variable aleatoria x cuya media es $\mu = 100$ con desviación estándar de $\sigma = 10$. Indique en el eje horizontal los valores 70, 80, 90, 100, 110, 120 y 130.
9. Una variable aleatoria es normalmente distribuida con media $\mu = 50$ y desviación estándar $\sigma = 5$.
 - a. Dibuje la curva normal de la función de densidad de probabilidad. En el eje horizontal dé los valores 35, 40, 45, 50, 55, 60 y 65. En la figura 6.4 se observa que la curva normal casi toca el eje horizontal en los puntos que se encuentran tres desviaciones estándar arriba de la media y tres desviaciones estándar debajo de la media (en este caso en 35 y 65).
 - b. ¿Cuál es la probabilidad de que la variable aleatoria tome un valor entre 45 y 55?
 - c. ¿De qué la variable aleatoria tome un valor entre 40 y 60?
10. Dibuje la gráfica de la distribución normal estándar. Etiquete el eje horizontal con los valores -3 , -2 , -1 , 0 , 1 , 2 y 3 . Después use la tabla de probabilidades de la distribución normal estándar que se encuentra en el forro interior del libro para calcular las probabilidades siguientes.
 - a. $P(z \leq 1.5)$
 - b. $P(z \leq 1)$
 - c. $P(1 \leq z \leq 1.5)$
 - d. $P(0 < z < 2.5)$
11. Dado que z es la variable normal estándar, calcule las probabilidades siguientes.
 - a. $P(z \leq -1.0)$
 - b. $P(z \geq -1)$
 - c. $P(z \geq -1.5)$
 - d. $P(-2.5 \leq z)$
 - e. $P(-3 < z \leq 0)$
12. Dado que z es la variable normal estándar, calcule las probabilidades siguientes.
 - a. $P(0 \leq z \leq 0.83)$
 - b. $P(-1.57 \leq z \leq 0)$
 - c. $P(z > 0.44)$
 - d. $P(z \geq -0.23)$
 - e. $P(z < 1.20)$
 - f. $P(z \leq -0.71)$
13. Dado que z es la variable normal estándar, calcule las probabilidades siguientes.
 - a. $P(-1.98 \leq z \leq 0.49)$
 - b. $P(0.52 \leq z \leq 1.22)$
 - c. $P(-1.75 \leq z \leq -1.04)$
14. Dado que z es la variable normal estándar, encuentre z en cada una de las situaciones siguientes.
 - a. El área a la izquierda de z es 0.9750.
 - b. El área entre 0 y z es 0.4750.
 - c. El área a la izquierda de z es 0.7291.
 - d. El área a la derecha de z es 0.1314.
 - e. El área a la izquierda de z es 0.6700.
 - f. El área a la derecha de z es 0.3300.

Autoexamen

15. Dado que z es la variable normal estándar, halle z en cada una de las situaciones siguientes.
 - a. El área a la izquierda de z es 0.2119
 - b. El área entre $-z$ y z es 0.9030.
 - c. El área entre $-z$ y z es 0.2052.
 - d. El área a la izquierda de z es 0.9948.
 - e. El área a la derecha de z es 0.6915.
16. Dado que z es la variable normal estándar, encuentre z en cada una de las situaciones siguientes.
 - a. El área a la derecha de z es 0.01
 - b. El área a la derecha de z es 0.025.
 - c. El área a la derecha de z es 0.05.
 - d. El área a la derecha de z es 0.10.

Aplicaciones

Autoexamen

17. Una persona con una buena historia crediticia tiene una deuda promedio de \$15 015 (*BusinessWeek*, 20 de marzo de 2006). Suponga que la desviación estándar es de \$3 540 y que los montos de las deudas están distribuidos normalmente.
 - a. ¿Cuál es la probabilidad de que la deuda de una persona con buena historia crediticia sea mayor a \$18 000?
 - b. ¿De que la deuda de una persona con buena historia crediticia sea de menos de \$10 000?
 - c. ¿De que la deuda de una persona con buena historia crediticia esté entre \$12 000 y \$18 000?
 - d. ¿De que la deuda de una persona con buena historia crediticia sea mayor a \$14 000?
18. El precio promedio de las acciones que pertenecen a S&P500 es de \$30 y la desviación estándar es \$8.20 (*BusinessWeek*, Special Annual Issue, primavera de 2003). Suponga que los precios de las acciones están distribuidos normalmente.
 - a. ¿Cuál es la probabilidad de que el precio de las acciones de una empresa sea por lo menos de \$40?
 - b. ¿De que el precio de las acciones de una empresa no sea mayor a \$20?
 - c. ¿De cuánto deben ser los precios de las acciones de una empresa para que esté entre las 10% mejores?
19. La cantidad promedio de precipitación pluvial en Dallas, Texas, durante el mes de abril es 3.5 pulgadas (*The World Almanac*, 2000). Suponga que se puede usar una distribución normal y que la desviación estándar es 0.8 pulgadas.
 - a. ¿Qué porcentaje del tiempo la precipitación pluvial en abril es mayor que 5 pulgadas?
 - b. ¿Qué porcentaje del tiempo la precipitación pluvial en abril es menor que 3 pulgadas?
 - c. Un mes se considera como extremadamente húmedo si la precipitación pluvial es 10% superior para ese mes. ¿Cuánta debe ser la precipitación pluvial en abril para que sea considerado un mes extremadamente húmedo?
20. En enero de 2003 un empleado estadounidense pasaba, en promedio, 77 horas conectado a Internet durante las horas de trabajo (CNBC, 15 de marzo de 2003). Suponga que la media poblacional es 77 horas, tiempos que están distribuidos normalmente y que la desviación estándar es 20 horas.
 - a. ¿Cuál es la probabilidad de que en enero de 2003 un empleado seleccionado aleatoriamente haya pasado menos de 50 horas conectado a Internet?
 - b. ¿Qué porcentaje de los empleados pasó en enero de 2003 más de 100 horas conectado a Internet?
 - c. Un usuario es clasificado como intensivo si se encuentra en el 20% superior de uso. ¿Cuántas horas tiene un empleado que haber estado conectado a Internet en enero de 2003 para que se le considerara un usuario intensivo?
21. La puntuación de una persona en una prueba de IQ debe estar en el 2% superior para que sea clasificado como miembro del grupo Mensa, la sociedad internacional de IQ elevado (*U.S. Airways Attaché*, septiembre de 2000). Si las puntuaciones de IQ tienen una distribución normal con una media de 100 y desviación estándar de 15, ¿cuál debe ser la puntuación de una persona para que se le considere miembro del grupo Mensa?

22. La tasa de remuneración media por hora para administrativos financieros en una determinada región es \$32.62 y la desviación estándar es \$2.32 (Bureau of Labor Statistics, septiembre de 2005). Suponga que estas tasas de remuneración están distribuidas normalmente.
- ¿Cuál es la probabilidad de que un directivo financiero tenga una remuneración entre \$30 y \$35 por hora?
 - ¿Qué tan alta debe ser la remuneración por hora para que un directivo financiero tenga un pago 10% superior?
 - ¿Cuál es la probabilidad de que la remuneración por hora de un directivo financiero sea menos de \$28 por hora?
23. El tiempo necesario para hacer un examen final en un determinado curso de una universidad tiene una distribución normal cuya media es 80 minutos con desviación estándar de 10 minutos. Conteste las preguntas siguientes
- ¿Cuál es la probabilidad de terminar el examen en una hora o menos?
 - ¿Cuál es la probabilidad de que un estudiante termine el examen en más de 60 minutos pero en menos de 75 minutos?
 - Suponga que en la clase hay 60 estudiantes y que el tiempo para resolver el examen es de 90 minutos. ¿Cuántos estudiantes piensa usted que no podrán terminar el examen en este tiempo?
24. El volumen de negociaciones en la Bolsa de Nueva York es más intenso en la primera media hora (en la mañana temprano) y la última media hora (al final de la tarde) de un día de trabajo. A continuación se presentan los volúmenes (en millones de acciones) de 13 días de enero y febrero.



214	163	265	194	180
202	198	212	201	
174	171	211	211	

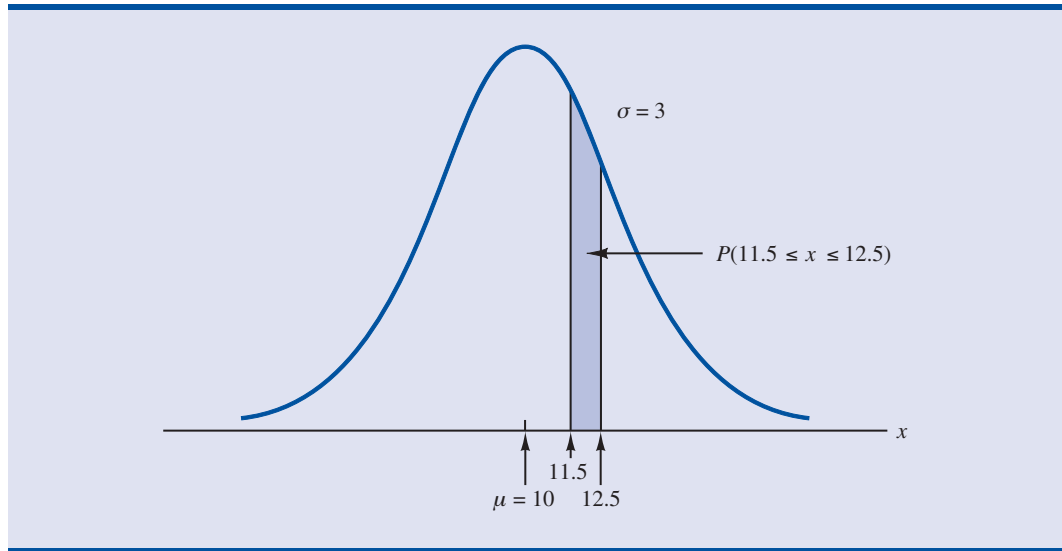
- La distribución de probabilidad de los volúmenes de negociaciones es aproximadamente normal.
- Calcule la media y la desviación estándar a usar como estimaciones de la media y de la desviación estándar de la población.
 - ¿Cuál es la probabilidad de que, en un día elegido al azar, el volumen de negociaciones en la mañana temprano sea superior a 180 millones de acciones?
 - ¿Cuál es la probabilidad de que, en un día elegido al azar, el volumen de negociaciones en la mañana temprano sea superior a 230 millones de acciones?
 - ¿Cuántas acciones deberán ser negociadas para que el volumen de negociaciones en la mañana temprano de un día determinado pertenezca al 5% de los días de mayor movimiento?
25. De acuerdo con la Sleep Foundation, en promedio se duermen 6.8 horas por noche. Suponga que la desviación estándar es 0.6 horas y que la distribución de probabilidad es normal.
- ¿Cuál es la probabilidad de que una persona seleccionada al azar duerma más de ocho horas?
 - ¿De que una persona tomada aleatoriamente duerma seis horas o menos?
 - Los médicos aconsejan dormir entre siete y nueve horas por noche. ¿Qué porcentaje de la población duerme esta cantidad?

6.3

Aproximación normal de las probabilidades binomiales

En la sección 5.4 se presentó la distribución binomial discreta. Recuerde que un experimento binomial consiste en una serie de n ensayos idénticos e independientes, habiendo para cada ensayo dos resultados posibles, éxito o fracaso. La probabilidad de éxito en un ensayo es la misma que en cualquier otro de los ensayos y se denota p . La variable aleatoria binomial es el número de éxitos en n ensayos y lo que se quiere saber es la probabilidad de x éxitos en n ensayos.

FIGURA 6.8 APROXIMACIÓN NORMAL A UNA PROBABILIDAD BINOMIAL
DISTRIBUCIÓN EN LA QUE $n = 100$ Y $p = 0.10$ MOSTRANDO LA
PROBABILIDAD DE 12 ERRORES



La evaluación de una función de probabilidad binomial, a mano o con una calculadora, se dificulta cuando el número de ensayos es muy grande. En los casos en que $np \geq 5$ y $n(1 - p) \geq 5$, la distribución normal proporciona una aproximación a las probabilidades binomiales que es fácil de usar. Cuando se usa la aproximación normal a la binomial, en la definición de la curva normal $\mu = np$ y $\sigma = \sqrt{np(1 - p)}$.

Para ilustrar la aproximación normal a la binomial, suponga que una empresa sabe por experiencia que 10% de sus facturas tienen algún error. Toma una muestra de 100 facturas y desea calcular la probabilidad de que 12 de estas facturas contengan algún error. Es decir, quiere hallar la probabilidad binomial de 12 éxitos en 100 ensayos. Aplicando la aproximación normal a este caso se tiene, $\mu = np = (100)(0.1) = 10$ y $\sigma = \sqrt{np(1 - p)} = \sqrt{(100)(0.1)(0.9)} = 3$. En la figura 6.8 se muestra la distribución normal con $\mu = 10$ y $\sigma = 3$.

Recuerde que en una distribución de probabilidad continua las probabilidades se calculan como áreas bajo la curva de la función de densidad de probabilidad. En consecuencia, la probabilidad que tiene un solo valor de la variable aleatoria es cero. Por tanto, para aproximar la probabilidad binomial de 12 éxitos se calcula el área correspondiente bajo la curva normal entre 11.5 y 12.5. Al 0.5 que se suma y se resta al 12 se le conoce como **factor de corrección por continuidad**. Este factor se introduce debido a que se está usando una distribución continua para aproximar una distribución discreta. Así, $P(x = 12)$ de una distribución binomial *discreta* se aproxima mediante $P(11.5 \leq x \leq 12.5)$ en la distribución normal *continua*.

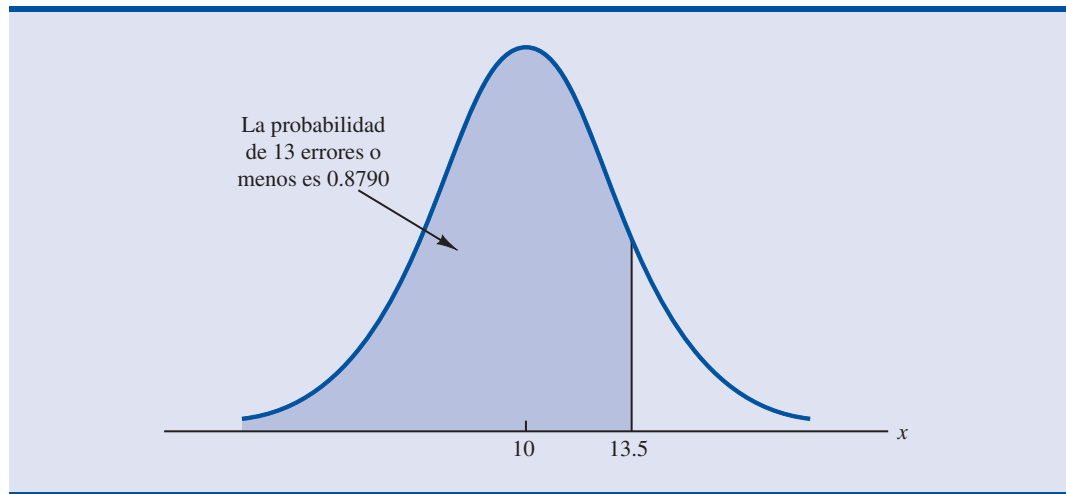
Convirtiendo la distribución normal estándar para calcular $P(11.5 \leq x \leq 12.5)$, se tiene

$$z = \frac{x - \mu}{\sigma} = \frac{12.5 - 10.0}{3} = 0.83 \quad \text{para } x = 12.5$$

y

$$z = \frac{x - \mu}{\sigma} = \frac{11.5 - 10.0}{3} = 0.50 \quad \text{para } x = 11.5$$

FIGURA 6.9 APROXIMACIÓN NORMAL A UNA PROBABILIDAD BINOMIAL: DISTRIBUCIÓN EN LA QUE $n = 100$ Y $p = 0.10$ MUESTRAN LA PROBABILIDAD DE 13 ERRORES O MENOS



En la tabla de la probabilidad normal estándar aparece que el área bajo la curva (figura 6.8) a la izquierda de 12.5 es 0.7967. De manera similar, el área bajo la curva a la izquierda de 11.5 es 0.6915. Por tanto, el área entre 11.5 y 12.5 es $0.7967 - 0.6915 = 0.1052$. El cálculo normal de la probabilidad de 12 éxitos en 100 ensayos es 0.1052.

Para tener un ejemplo más, suponga que se quiere calcular la probabilidad de 13 o menos facturas con errores en una muestra de 100 facturas. En la figura 6.9 se muestra el área bajo la curva que aproxima esta probabilidad. Observe que debido al uso del factor de continuidad el valor que se usa para calcular esta probabilidad es 13.5. El valor z que corresponde a 13.5 es

$$z = \frac{13.5 - 10.0}{3.0} = 1.17$$

En la tabla de probabilidad normal estándar se observa que el área bajo la curva normal estándar y a la izquierda de $z = 1.17$ es 0.8790. El área bajo la curva normal que aproxima la probabilidad de 13 o menos facturas con errores es la porción sombreada que se observa en la gráfica de la figura 6.9.

Ejercicios

Métodos

Autoexamen

26. En una distribución de probabilidad binomial con $p = 0.20$ y $n = 100$.
 - a. ¿Cuál es la media y la desviación estándar?
 - b. ¿En esta situación las probabilidades binomiales pueden ser aproximadas por la distribución de probabilidad normal? Explique.
 - c. ¿Cuál es la probabilidad de exactamente 24 éxitos?
 - d. ¿Cuál es la probabilidad de 18 a 22 éxitos?
 - e. ¿Cuál es la probabilidad de 15 o menos éxitos?
27. Suponga que se tiene una distribución de probabilidad binomial en la que $p = 0.60$ y $n = 200$.
 - a. ¿Cuál es la media y la desviación estándar?
 - b. ¿En esta situación las probabilidades binomiales puedan ser aproximadas por la distribución de probabilidad normal? Explique.
 - c. ¿Cuál es la probabilidad de 100 a 110 éxitos?

- d. ¿Cuál es la probabilidad de 130 o más éxitos?
- e. ¿Cuál es la ventaja de usar la distribución de probabilidad normal para aproximar las probabilidades binomiales? Use el inciso d para explicar las ventajas.

Aplicaciones

Autoexamen

28. El presidente Bush propuso eliminar los impuestos sobre los dividendos que pagan los accionistas debido a que esto resulta en un doble pago de impuestos. Las ganancias que se usan para pagar los dividendos ya han sido grabadas. En un sondeo sobre este tema se encontró que 47% de los estadounidenses estaban a favor de esta propuesta. La posición de los partidos políticos era 64% de los republicanos y 29% de los demócratas a favor de la propuesta (*Investor's Business Daily*, 13 de enero de 2003). Suponga que 250 estadounidenses se reúnen para una conferencia acerca de la propuesta.
 - a. ¿Cuál es la probabilidad de que por lo menos la mitad del grupo esté a favor de la propuesta?
 - b. Más tarde se enteró de que en el grupo hay 150 republicanos y 100 demócratas. Ahora, ¿cuál es su estimación del número esperado a favor de la propuesta?
 - c. Ahora que conoce la composición del grupo, ¿cree que un conferencista a favor de la propuesta sea mejor recibido que uno que esté en contra de la propuesta?
29. La tasa de desempleo es de 5.8% (Bureau of Labor Statistics, www.bls.gov, 3 de abril de 2003). Suponga que se seleccionan aleatoriamente 100 personas que se pueden emplear.
 - a. ¿Cuál es el número esperado de quienes están desempleados?
 - b. ¿Cuál es la varianza y la desviación estándar del número de los que están desempleados?
 - c. ¿Cuál es la probabilidad de que exactamente seis estén desempleados?
 - d. ¿Cuál es la probabilidad de que por lo menos cuatro estén desempleados?
30. Cuando usted firma un contrato para una tarjeta de crédito, ¿lee cuidadosamente el contrato? En un sondeo FindLaw.com le preguntó a las personas “¿Qué tan cuidadosamente lee usted un contrato para una tarjeta de crédito?” Los hallazgos fueron que 44% leen cada palabra, 33% leen lo suficiente para entender el contrato, 11% sólo le echa una mirada y 4% no lo leen en absoluto.
 - a. En una muestra de 500 personas ¿cuántas esperaría usted que respondan que leen cada palabra de un contrato para una tarjeta de crédito?
 - b. En una muestra de 500 personas ¿cuál es la probabilidad de que 200 o menos digan que leen cada palabra de un contrato para una tarjeta de crédito?
 - c. En una muestra de 500 personas ¿cuál es la probabilidad de que por lo menos 15 digan que no leen en absoluto un contrato para una tarjeta de crédito?
31. El Myrtle Beach hotel tiene 120 habitaciones. En los meses de primavera su ocupación es de 75%.
 - a. ¿Cuál es la probabilidad de que por lo menos la mitad de las habitaciones estén ocupadas en un día dado?
 - b. ¿De que 100 o más de las habitaciones estén ocupadas en un día dado?
 - c. ¿De que 80 o menos de las habitaciones estén ocupadas en un día dado?

6.4

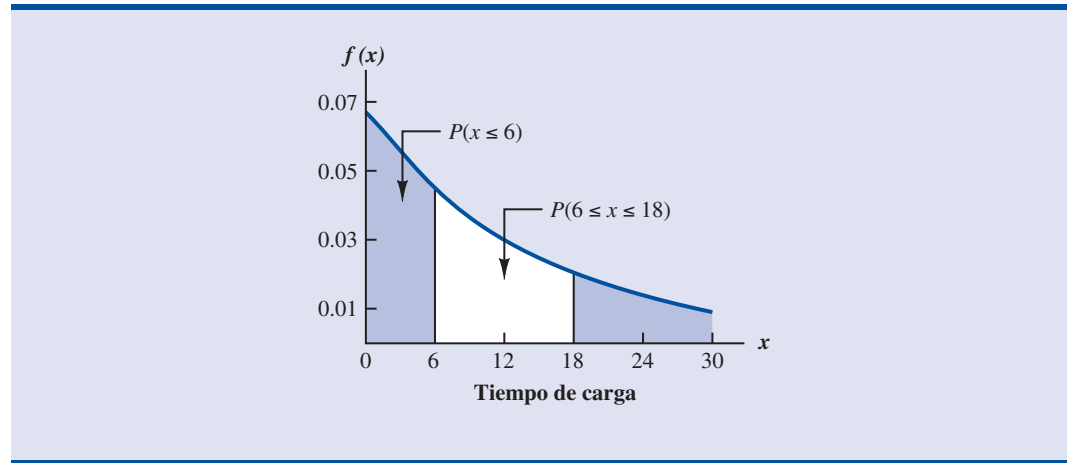
Distribución de probabilidad exponencial

La **distribución de probabilidad exponencial** se aplica a variables como las llegadas de automóviles a un lavado de coches, los tiempos requeridos para cargar un camión, las distancias entre dos averías en una carretera, etc. A continuación se da la función de densidad de probabilidad exponencial.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD EXPONENCIAL

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{para } x \geq 0, \mu > 0 \quad (6.4)$$

donde μ = valor esperado o media

FIGURA 6.10 DISTRIBUCIÓN EXPONENCIAL PARA EL EJEMPLO DEL ÁREA DE CARGA

Como ejemplo de la distribución exponencial, suponga que x representa el tiempo que se necesita para cargar un camión en un área de carga, y que este tiempo de carga sigue una distribución exponencial. Si el tiempo de carga medio o promedio es 15 minutos ($\mu = 15$), la función de densidad de probabilidad apropiada para x es

$$f(x) = \frac{1}{15} e^{-x/15}$$

La figura 6.10 es la gráfica de esta función de densidad de probabilidad.

Cálculo de probabilidades en la distribución exponencial

Como ocurre con cualquier distribución de probabilidad continua, el área bajo la curva correspondiendo a un intervalo da la probabilidad de que la variable aleatoria tome algún valor en ese intervalo. En el ejemplo del área de carga, la probabilidad de que cargar un camión necesite 6 minutos o menos $P(x \leq 6)$ está definida como el área bajo la curva de la figura 10.6 que va desde $x = 0$ hasta $x = 6$. De manera similar, la probabilidad de que el tiempo de carga sean 18 minutos o menos $P(x \leq 18)$ es el área bajo la curva desde $x = 0$ hasta $x = 18$. Observe también que la probabilidad de que el tiempo de carga esté entre 6 y 18 minutos $P(6 \leq x \leq 18)$ corresponde al área bajo la curva desde $x = 6$ hasta $x = 18$.

Para calcular probabilidades exponenciales como las que se acaban de describir, se usa la fórmula siguiente. Esta fórmula aporta la probabilidad acumulada de obtener un valor de la variable aleatoria exponencial que sea menor o igual que algún valor específico denotado por x_0 .

DISTRIBUCIÓN EXPONENCIAL: PROBABILIDADES ACUMULADAS

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \quad (6.5)$$

En el ejemplo del área de carga, x = tiempo de carga en minutos y $\mu = 15$ minutos. A partir de la ecuación (6.5)

$$P(x \leq x_0) = 1 - e^{-x_0/15}$$

Por tanto, la probabilidad de que cargar un camión requiera 6 minutos o menos es

$$P(x \leq 6) = 1 - e^{-6/15} = 0.3297$$

En aplicaciones de colas de espera, la distribución exponencial suele emplearse para tiempos de servicio.

Con la ecuación (6.5) se calcula la probabilidad de que cargar un camión requiera 18 minutos o menos.

$$P(x \leq 18) = 1 - e^{-18/15} = 0.6988$$

De manera que la probabilidad de que para cargar un camión se necesiten entre 6 y 18 minutos es igual a $0.6988 - 0.3297 = 0.3691$. Probabilidades para cualquier otro intervalo se calculan de manera semejante.

La distribución exponencial tiene la propiedad de que la media y la desviación estándar son iguales.

En el ejemplo anterior el tiempo medio para cargar un camión fue $\mu = 15$ minutos. La distribución exponencial tiene la propiedad de que la media de la distribución y la desviación estándar de la distribución son iguales. Por tanto, la desviación estándar del tiempo que se necesita para cargar un camión es $\sigma = 15$ minutos y la varianza es $\sigma^2 = (15)^2 = 225$.

Relación entre la distribución de Poisson y la exponencial

En la sección 5.5 se presentó la distribución de probabilidad de Poisson como una distribución de probabilidad discreta que se usa para examinar el número de ocurrencias de un evento en un determinado intervalo de tiempo o de espacio. Recuerde que la función de probabilidad de Poisson es

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

donde

μ = valor esperado o número medio de ocurrencias
en un determinado intervalo

Si las llegadas siguen una distribución de Poisson, el tiempo entre las llegadas debe seguir una distribución exponencial.

La distribución de probabilidad exponencial continua está relacionada con la distribución discreta de Poisson. Si la distribución de Poisson da una descripción del número de ocurrencias por intervalo, la distribución exponencial aporta una descripción de la longitud de los intervalos entre las ocurrencias.

Para ilustrar esta relación, suponga que el número de automóviles que llegan a un lavado de coches durante una hora se describe mediante la distribución de probabilidad de Poisson, con una media de 10 automóviles por hora. La función de probabilidad de Poisson que da la probabilidad de x llegadas por hora es

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Dado que el número promedio de llegadas es 10 automóviles por hora, el tiempo promedio entre las llegadas de los automóviles es

$$\frac{1 \text{ hora}}{10 \text{ automóviles}} = 0.1 \text{ hora/automóvil}$$

Entonces, la distribución exponencial que describe el tiempo entre las llegadas tiene una media de $\mu = 0.1$ por automóvil; la función de densidad de probabilidad exponencial es

$$f(x) = \frac{1}{0.1} e^{-x/0.1} = 10e^{-10x}$$

NOTAS Y COMENTARIOS

Como se observa en la figura 6.10, la distribución exponencial es sesgada a la derecha. En efecto, la medida del sesgo en la distribución exponencial es

2. La distribución exponencial da una idea clara de cómo es una distribución sesgada.

Ejercicios

Métodos

32. Considere la siguiente función de densidad de probabilidad exponencial.

$$f(x) = \frac{1}{8} e^{-x/8} \quad \text{para } x \geq 0$$

- Halle $P(x \leq 6)$.
- Encuentre $P(x \leq 4)$.
- Halle $P(x \geq 6)$.
- Encuentre $P(4 \leq x \leq 6)$.

33. Considere la siguiente función de densidad de probabilidad exponencial.

$$f(x) = \frac{1}{3} e^{-x/3} \quad \text{para } x \geq 0$$

- Dé la fórmula para hallar $P(x \leq x_0)$.
- Halle $P(x \leq 2)$.
- Encuentre $P(x \geq 3)$.
- Halle $P(x \leq 5)$.
- Halle $P(2 \leq x \leq 5)$.

Aplicaciones

34. El tiempo requerido para pasar por la inspección en los aeropuertos puede ser molesto para los pasajeros. El tiempo medio de espera en los periodos pico en el Cincinnati/Northern Kentucky International Airport es de 12.1 minutos (*The Cincinnati Enquirer*, 2 de febrero de 2006). Suponga que los tiempos para pasar por la inspección de seguridad tienen una distribución exponencial.

- ¿Cuál es la probabilidad de que durante los periodos pico se requieran 10 minutos para pasar la inspección de seguridad?
- ¿De qué durante los periodos pico se requieran más de 20 minutos para pasar la inspección de seguridad?
- ¿De qué durante los periodos pico se requieran entre 10 y 20 minutos para pasar la inspección de seguridad?
- Son las 8 de la mañana (periodo pico) y usted se acaba de formar en la fila para la inspección de seguridad. Para alcanzar su avión, tiene que estar en la puerta de arribo en no más de 30 minutos. Si necesitara 12 minutos una vez pasada la inspección de seguridad para llegar a la puerta de arribo, ¿cuál es la probabilidad de que pierda el avión?

35. Los tiempos entre las llegadas de vehículos a un determinado entronque siguen una distribución de probabilidad exponencial cuya media es 12 segundos.

- Dibuje esta distribución de probabilidad exponencial.
- ¿Cuál es la probabilidad de que los tiempos de llegada entre vehículos sean 12 segundos o menos?

- c. ¿Cuál es la probabilidad de que los tiempos de llegada entre vehículos sean 6 segundos o menos?
 - d. ¿Cuál es la probabilidad de 30 o más segundos entre los tiempos de llegada?
36. El tiempo de vida (en hora) de un dispositivo electrónico es una variable aleatoria que tiene la siguiente función de densidad de probabilidad exponencial.

$$f(x) = \frac{1}{50} e^{-x/50} \quad \text{para } x \geq 0$$

- a. ¿Cuál es la media del tiempo de vida de este dispositivo?
 - b. ¿Cuál es la probabilidad de que el dispositivo tenga una falla en las primeras 25 horas de funcionamiento?
 - c. ¿Cuál es la probabilidad de que el dispositivo funcione 100 o más horas sin fallar?
37. Sparagowsky & Associates hace un estudio sobre los tiempos necesarios para atender a un cliente en la ventanilla de su automóvil en los restaurantes de comida rápida. En McDonald's el tiempo medio para atender a un cliente fue 2.78 minutos (*The Cincinnati Enquirer*, 9 de julio de 2000). Tiempos de servicio como los de estos restaurantes suelen seguir una distribución exponencial.
- a. ¿Cuál es la probabilidad de que el tiempo para atender a un cliente sea menor que 2 minutos?
 - b. ¿De que el tiempo para atender a un cliente sean más de 5 minutos?
 - c. ¿De que el tiempo para atender a un cliente sean más de 2.78 minutos?
38. ¿Las interrupciones durante su trabajo reducen su productividad? De acuerdo con un estudio realizado por la University of California–Irvine, las personas de negocios son interrumpidas aproximadamente 51/2 veces por hora (*Fortune*, 20 de marzo de 2006). Suponga que el número de interrupciones sigue una distribución de probabilidad de Poisson.
- a. Dé la distribución de probabilidad para el tiempo entre las interrupciones.
 - b. ¿Cuál es la probabilidad de que una persona de negocios no tenga ninguna interrupción en 15 minutos?
 - c. ¿Cuál es la probabilidad de que la siguiente interrupción a una determinada persona de negocios ocurra en no más de 10 minutos?

Resumen

En este capítulo se amplía el estudio de las distribuciones de probabilidad al caso de las variables aleatorias continuas. La principal diferencia conceptual entre distribuciones de probabilidades discretas y continuas está en el método para calcular las probabilidades. En el caso de distribuciones discretas la función de probabilidad $f(x)$ da la probabilidad de que la variable aleatoria x tome diversos valores. En el caso de las distribuciones continuas, la función de densidad de probabilidad $f(x)$ no da directamente valores de probabilidad. Aquí, las probabilidades están dadas por áreas bajo la curva o gráfica de la función de densidad de probabilidad $f(x)$. Como el área bajo la curva sobre un solo punto es 0, se observa que en una variable aleatoria continua la probabilidad de cualquier valor particular es 0.

Se vieron a detalle tres distribuciones de probabilidad continua: la uniforme, la normal y la exponencial. La distribución normal es muy empleada en la inferencia estadística y será muy empleada en lo que resta del libro.

Glosario

Función de densidad de probabilidad Función que se usa para calcular probabilidades de una variable aleatoria continua. El área bajo la gráfica de una función de densidad de probabilidad y sobre un intervalo representa probabilidad.

Distribución de probabilidad uniforme Distribución de probabilidad continua en la cual la probabilidad de que una variable aleatoria tome un valor en cualquier intervalo es igual para intervalos de igual longitud.

Distribución de probabilidad normal Una distribución de probabilidad continua. Su función de densidad de probabilidad tiene forma de campana y está determinada por la media μ y la desviación estándar σ .

Distribución de probabilidad normal estándar Distribución normal en la cual la media es cero y la desviación estándar es uno.

Factor de corrección por continuidad Valor de 0.5 que se suma o resta al valor de x cuando se usa una distribución normal continua para aproximar una distribución binomial discreta.

Distribución de probabilidad exponencial Una distribución de probabilidad continua útil para calcular probabilidades acerca del tiempo que se necesita para realizar una tarea.

Fórmulas clave

Función de densidad de probabilidad uniforme

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{si no es así} \end{cases} \quad (6.1)$$

Función de densidad de probabilidad normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad (6.2)$$

Conversión a la variable aleatoria normal estándar

$$z = \frac{x - \mu}{\sigma} \quad (6.3)$$

Función de densidad de probabilidad exponencial

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{para } x \geq 0, \mu > 0 \quad (6.4)$$

Distribución exponencial: probabilidades acumuladas

$$P(x \leq x_0) = 1 - e^{-x_0/\mu} \quad (6.5)$$

Ejercicios complementarios

39. Una ejecutiva de negocios se va a mudar de Chicago a Atlanta y necesita vender rápidamente su casa en Chicago. Un empleado le ofrece comprársela en \$210 000, pero la oferta expira al final de esa semana. En este momento la ejecutiva no tiene otra oferta mejor, pero piensa que puede dejar la casa en el mercado un mes más. De acuerdo con las pláticas que ha tenido con su agente inmobiliario la ejecutiva cree que los precios que pueden ofrecerle dejando la casa un mes más en el mercado están distribuidos uniformemente entre \$200 000 y \$225 000.
- Si deja la casa en el mercado un mes más, ¿cuál es la expresión matemática para la función de densidad de probabilidad de los precios de venta que le sean ofrecidos?
 - Si la deja en el mercado un mes más, ¿cuál es la probabilidad de que venda la casa en por lo menos \$215 000?
 - Si la deja en el mercado un mes más, ¿cuál es la probabilidad de que venda la casa en menos de \$210 000?
 - ¿Deberá dejar la ejecutiva su casa en el mercado un mes más? ¿Por qué sí o por qué no?

40. La U.S. Bureau of Labor Statistics informa que el gasto promedio anual en alimentos y bebidas de una familia es \$5700 (*Money*, diciembre de 2003). Suponga que los gastos anuales en alimentos y bebidas están distribuidos en forma normal y que la desviación estándar es \$1500.
 - a. ¿Cuál es el intervalo en que se encuentran los gastos de 10% de las familias que tienen los menores gastos anuales en alimentos y bebidas?
 - b. ¿Qué porcentaje de las familias gasta más de \$7000 anualmente en alimentos y bebidas?
 - c. ¿Cuál es el intervalo en el que se encuentran los gastos de 5% de las familias que tienen los gastos más altos en alimentos y bebidas?
41. Motorola usa la distribución normal para determinar la probabilidad de defectos y el número de defectos esperados en un proceso de producción. Suponga que en un proceso de producción el peso promedio de los artículos producidos es 10 onzas. Calcule la probabilidad de un defecto y el número esperado de defectos en 100 unidades producidas en las situaciones siguientes.
 - a. La desviación estándar del proceso es 0.15 y los límites para el proceso se han fijado en más o menos una desviación estándar. Las unidades que pesen más de 9.85 o menos de 10.15 onzas se clasifican como defectuosas.
 - b. Después de hacer mejoras al proceso de producción, la desviación estándar se reduce a 0.05. Asuma que se siguen usando los mismos límites para el proceso; artículos que pesen menos de 9.85 o más de 10.15 onzas se clasifican como defectuosos.
 - c. ¿Cuál es la ventaja de haber reducido la variación en el proceso de producción, haciendo que los límites se encuentren a un número mayor de desviaciones estándar de la media?
42. El promedio anual de gastos de una familia estadounidense en transporte diario es \$6312 (*Money*, agosto de 2001). Suponga que dicha cantidad está distribuida normalmente.
 - a. Si sabe que 5% de las familias estadounidenses gastan menos de \$1000 en el transporte diario. ¿Cuál es la desviación estándar en esta cantidad de gasto?
 - b. ¿Cuál es la probabilidad de que un hogar gaste entre \$4000 y \$6000?
 - c. ¿En que intervalo se encuentran los gastos de las familias que constituyen 3% de las familias con los gastos más altos en transporte?
43. *Condé Nast Traveler* publica la lista de oro de los mejores hoteles en todo el mundo. Broadmoor Hotel en Colorado Springs tiene 700 habitaciones y estuvo en la lista de oro en 2004 (*Condé Nast Traveler*, enero de 2004). El grupo encargado del marketing de este hotel pronostica una demanda media de 670 habitaciones para el próximo fin de semana. Suponga que la demanda para el próximo fin de semana está distribuida normalmente y que la desviación estándar es 30.
 - a. ¿Cuál la probabilidad de que se ocupen todas las habitaciones del hotel?
 - b. ¿Cuál la probabilidad de que se ocupen 50 o más habitaciones del hotel?
 - c. ¿Recomendaría al hotel hacer una promoción para aumentar la demanda? ¿Qué consideraciones serían importantes?
44. Ward Doering Auto Sales está pensando en ofrecer un contrato especial de servicio que cubra todos los costos de servicio de los automóviles vendidos. De acuerdo con la experiencia, el director de la empresa estima que los costos anuales de servicio están distribuidos casi normalmente con una media de \$150 y una desviación estándar de \$25.
 - a. Si la empresa ofrece a los clientes el contrato de servicio por una cantidad anual de \$200, ¿cuál es la probabilidad de que el costo de un servicio sea mayor a los \$200 del precio del contrato?
 - b. ¿Cuál es la ganancia esperada por la empresa en estos contratos de servicio?
45. ¿La falta de sueño es causa de accidentes de tráfico de consecuencias fatales? En un estudio se encontró que el número promedio por año de accidentes de tráfico con consecuencias fatales ocasionados por conductores somnolientos es 1550 (*BusinessWeek*, 26 de enero de 2004). Suponga que el número promedio anual de accidentes de tráfico de consecuencias fatales está distribuido normalmente con una desviación estándar de 300.
 - a. ¿Cuál es la probabilidad de que haya menos de 1000 accidentes fatales en un año?
 - b. ¿De que el número anual de accidentes fatales esté entre 1000 y 2000?
 - c. Para que un año se encuentre en el 5% superior en número de accidentes fatales, cuántos de éstos tendrán que ocurrir?

46. Suponga que las puntuaciones obtenidas en el examen de admisión a una universidad están distribuidas en forma normal con una media de 450 y una desviación estándar de 100.
- ¿Qué porcentaje de las personas que hacen el examen tendrá una puntuación entre 400 y 500?
 - Si la puntuación que obtiene un estudiante es 630. ¿Qué porcentaje de los estudiantes que hacen el examen tendrá una puntuación mayor? ¿Qué porcentaje tendrá una puntuación menor?
 - Si la universidad no admite estudiantes que obtengan una puntuación menor a 480, ¿qué porcentaje de los estudiantes que hacen el examen podrá ser aceptado?
47. De acuerdo con *Adversiting Age*, el salario base promedio de las mujeres que trabajan como publicistas es superior al salario base promedio de los hombres. El salario base promedio de las mujeres es \$67 000 y el salario base promedio de los hombres es \$65 500 (*Working Woman*, julio/agosto de 2000). Suponga que los salarios están distribuidos normalmente con una desviación estándar de \$7000 tanto para hombres como para mujeres.
- ¿Cuál es la probabilidad de que una mujer tenga un salario mayor que \$75 000?
 - ¿De que un hombre tenga un salario mayor que \$75 000?
 - ¿De que una mujer tenga un salario mayor que \$50 000?
 - ¿Cuánto tendrá que ganar una mujer para tener un salario mayor que 99% de los hombres?
48. Una máquina llena recipientes con un determinado producto. De acuerdo con datos anteriores se sabe que la desviación estándar en los pesos rellenos es 0.6 onzas. Si sólo 2% de los recipientes llenados tienen menos de 18 onzas, ¿cuál es el peso medio de llenado de la máquina? Es decir, a cuánto es igual μ ? Suponga que los pesos llenados tienen una distribución normal.
49. Considere un examen de opción múltiple con 50 preguntas. Para cada pregunta hay cuatro respuestas posibles. Suponga que un estudiante que ha hecho las tareas y asistido a clase tiene 75% de probabilidad de contestar correctamente las preguntas.
- Para obtener A de calificación, un estudiante tiene que contestar correctamente 43 o más preguntas. ¿Qué porcentaje de los estudiantes que hicieron las tareas y asistieron a clase obtendrá A de calificación?
 - Para obtener C de calificación, un estudiante tiene que contestar correctamente de 35 a 39 preguntas. ¿Qué porcentaje de los estudiantes que hicieron las tareas y asistieron a clases obtendrá C de calificación?
 - Para aprobar el examen, un estudiante tiene que contestar correctamente 30 preguntas o más. ¿Qué porcentaje de los estudiantes que hicieron las tareas y asistieron a clases pasará el examen?
 - Suponga que un estudiante no asistió a clases ni hizo las tareas. Además, dicho estudiante sólo tratará de adivinar las respuestas a las preguntas. ¿Cuál es la probabilidad de que el estudiante conteste correctamente 30 o más preguntas y pase el examen?
50. En Las Vegas un jugador de blackjack se entera de que la casa proporcionará una habitación gratis a quien juegue cuatro horas con un promedio de apuesta de \$50. La estrategia del jugador tiene una probabilidad de ganar en cualquier mano de 0.49 y el jugador sabe que hay 60 manos por hora. Suponga que el jugador juega durante cuatro horas con una apuesta de \$50 por mano.
- ¿Cuál es la ganancia esperada del jugador?
 - ¿Cuál es la probabilidad de que el jugador pierda \$1000 o más?
 - ¿Cuál es la probabilidad de que el jugador gane?
 - Si el jugador empieza con \$1500. ¿Cuál es la probabilidad de que el jugador se vaya a la bancarrota?
51. El tiempo, en minutos, que un estudiante usa una terminal de computadora en el centro de cálculo de una universidad sigue una distribución de probabilidad exponencial con una media de 36 minutos. Suponga que un estudiante llega a una terminal precisamente en el momento en que otro estudiante quería usar la terminal.
- ¿Cuál es la probabilidad de que el segundo estudiante tenga que esperar 15 minutos o menos?
 - ¿De que el segundo estudiante tenga que esperar entre 15 y 45 minutos?
 - ¿Cuál la probabilidad de que el segundo estudiante tenga que esperar una hora o más?
52. El sitio Web de Bed and Breakfast Inns of North America (www.cimarron.net) recibe aproximadamente siete visitas por minuto (*Time*, septiembre de 2001). Suponga que el número de visitantes por minuto sigue una distribución de probabilidad de Poisson.

- a. ¿Cuál es el tiempo medio entre las visitas a este sitio de la Web?
 - b. Muestre la función de densidad de probabilidad exponencial para los tiempos entre las visitas a este sitio.
 - c. ¿Cuál es la probabilidad de que nadie visite este sitio en un lapso de 1 minuto?
 - d. ¿Cuál es la probabilidad de que nadie visite este sitio en un lapso de 12 minutos?
53. En la ciudad de Nueva York el tiempo de recorrido promedio al trabajo es de 36.5 minutos.
- a. Suponga que la distribución de probabilidad exponencial es aplicable y muestre la función de densidad de probabilidad para las duraciones de los recorridos al trabajo en Nueva York.
 - b. ¿Cuál es la probabilidad de que un neoyorquino típico necesite entre 20 y 40 minutos para transportarse a su trabajo?
 - c. ¿De que un neoyorquino típico necesite más de 40 minutos para transportarse a su trabajo?
54. El tiempo (en minutos) entre dos llamadas telefónicas en la oficina de solicitud de servicios de una aseguradora tiene la siguiente distribución de probabilidad exponencial.

$$f(x) = 0.50e^{-0.50x} \quad \text{para } x \geq 0$$

- a. ¿Cuál es el tiempo medio entre las llamadas telefónicas?
- b. ¿Cuál es la probabilidad de que pasen 30 segundos o menos entre llamadas telefónicas?
- c. ¿De que pase 1 minuto o menos entre las llamadas telefónicas?
- d. ¿Cuál es la probabilidad de que pasen 5 minutos o más sin que haya llamadas telefónicas?

Caso problema Specialty Toys

Specialty Toys, Inc. vende una gran variedad de nuevos e innovadores juguetes para niños. Los directivos saben que la época prenavideña es la mejor oportunidad para la introducción de un nuevo juguete, en esta época muchas personas buscan cosas novedosas para los regalos navideños. Cuando Specialty descubre un nuevo juguete con un buen potencial de mercado, elige alguna fecha en octubre para su lanzamiento.

Para contar con los juguetes en octubre Specialty hace los pedidos a sus proveedores en junio o julio de cada año. La demanda de juguetes para niños puede ser muy volátil. Si un nuevo juguete se pone de moda, la posibilidad de que se agote suele incrementar la demanda hasta niveles altos y se pueden obtener grandes ganancias. Sin embargo, un nuevo juguete también puede fracasar dejando a Specialty con un gran inventario que debe vender a precios reducidos. La interrogante más importante que enfrenta la empresa es decidir cuántas unidades comprar de un juguete nuevo para satisfacer la demanda. Si compra muy pocos, perderá ventas; si compra demasiados, las ganancias se reducirán por los precios bajos que tendrá que ofrecer en una liquidación.

En la próxima temporada Specialty desea introducir un juguete nuevo que se llama *El osito pronosticador del clima*. Esta variación de un osito de peluche que habla es fabricada por una empresa en Taiwan. Cuando un niño oprime la mano del osito, éste empieza a hablar. El osito tiene un barómetro que le ayuda, de acuerdo con el estado del tiempo, a elegir una de cinco frases que pronostican el estado del tiempo. Las frases van desde “¡Parece que es un bonito día! Que te diviertas” hasta “Parece que va a llover. No se te olvide llevar tu paraguas”. Pruebas realizadas con el producto indican que, aunque no es preciso, sus pronósticos del tiempo son sorprendentemente buenos. Varios de los directivos de Specialty opinan que los pronósticos del tiempo del osito son tan buenos como muchos de los pronósticos del tiempo que se dan en televisión.

Como ocurre con todos los productos, Specialty se enfrenta a la pregunta de cuántos ositos ordenar para la temporada siguiente. Las cantidades que sugieren los directivos son 15 000, 18 000, 24 000 o 28 000 unidades. El intervalo tan amplio en que se encuentran estas cantidades indica una considerable discrepancia en lo que se refiere al potencial de mercado. El equipo de directivos le solicita a usted un análisis de las probabilidades de terminar el inventario de acuerdo con diversas cantidades a comprar, así como una estimación del potencial de ganancias y su ayuda para hacer una recomendación de la cantidad que se debe comprar. Specialty espera vender *El osito pronosticador del clima* a \$24 con base en un costo de \$16 por unidad. Si hay inventario sobrante después de la temporada de las fiestas decembrinas, Specialty venderá las unidades res-

tantes a \$5 cada una. Después de revisar las ventas anteriores de productos semejantes, los expertos de Specialty pronostican una demanda esperada de 20 000 unidades y 0.95 de probabilidad de que la demanda esté entre 10 000 y 30 000.

Informe administrativo

Elabore un informe sobre los puntos siguientes y recomiende la cantidad a comprar de *El osito pronosticador del clima*.

1. Use los pronósticos de ventas para describir una distribución de probabilidad normal que pueda servir para aproximar la distribución de la demanda. Dibuje la distribución y dé su media y su desviación estándar.
2. Calcule la probabilidad de terminar el inventario de acuerdo con las cantidades a comprar sugeridas por los miembros del equipo de directivos.
3. Calcule las ganancias proyectadas de acuerdo con las cantidades a comprar sugeridas por los miembros del equipo de directivos bajo tres escenarios: el peor de los casos, en el cual se venderán 10 000 unidades, en el caso más probable, en el cual se venderán 20 000 unidades y en el mejor de los casos en el cual se venderán 30 000 unidades.
4. Uno de los directivos de Specialty encuentra que el potencial de ganancia es tan bueno que la cantidad a comprar debe tener 70% de posibilidades de satisfacer la demanda y 30% de posibilidades de quedarse sin mercancía. De acuerdo con esto, ¿qué cantidad debe comprarse y cuál es la ganancia proyectada bajo cada uno de los tres escenarios?
5. Dé su propia recomendación sobre la cantidad que debe comprarse y muestre las proyecciones de ganancia correspondientes. Fundamente su recomendación.

Apéndice 6.1 Distribuciones de probabilidad continua con Minitab

Para demostrar el procedimiento de Minitab para el cálculo de probabilidades continuas se retomará el problema de la empresa Grear Tire, en el que la duración de los neumáticos en millas se describió mediante una distribución normal en la que $\mu = 36\,500$ y $\sigma = 5000$. Una de las preguntas que se plantearon fue: ¿cuál es la probabilidad de que los neumáticos duren más de 40 000 millas?

Para distribuciones de probabilidad continua, Minitab proporciona probabilidades acumuladas; es decir, Minitab da la probabilidad de que la variable aleatoria tome un valor menor o igual que una constante específica. En el caso de la pregunta sobre la duración de los neumáticos, Minitab se puede usar para determinar la probabilidad acumulada de que un neumático dure 40 000 millas o menos. (En este caso la constante específica es 40 000.) Una vez que se tiene la probabilidad acumulada que proporciona Minitab, es necesario restar esta probabilidad de 1 para determinar la probabilidad de que el neumático dure más de 40 000 millas.

Para que Minitab calcule una probabilidad, es necesario ingresar la constante específica en una de las columnas de la hoja de cálculo. En este caso se introduce la constante específica 40 000 en la columna C1 de la hoja de cálculo de Minitab. A continuación se presentan los pasos necesarios para que Minitab calcule la probabilidad acumulada de que la variable aleatoria normal tome valores menores o iguales que 40 000.

Paso 1. Seleccionar el menú **Calc**

Paso 2. Elegir **Probability Distributions**

Paso 3. Elegir **Normal**

Paso 4. Cuando aparezca el cuadro de diálogo Normal Distribution:

Seleccionar **Cumulative probability**

Ingresar 36 500 en el cuadro **Mean**

Ingresar 5 000 en el cuadro **Standard deviation**

Ingresar C1 en el cuadro **Input column** (la columna que contiene 40 000)

Clic en **OK**

Después de que el usuario hace clic en **OK**, Minitab da la probabilidad acumulada de que la variable aleatoria normal tome un valor menor o igual que 40 000. Minitab indica que esta probabilidad es 0.7580. Como lo que interesa es la probabilidad de que el neumático dure más de 40 000, la probabilidad buscada es $1 - 0.7580 = 0.2420$.

Otra pregunta en el problema de la empresa Grear Tire fue: ¿cuál es la duración en millas que la empresa debe establecer en la garantía de manera que en no más de 10% de los neumáticos se tenga que pagar la garantía? En este caso se da una probabilidad y se quiere hallar el valor correspondiente de la variable aleatoria. Minitab usa una rutina de cálculo inverso para hallar el valor de la variable aleatoria que corresponde a la probabilidad acumulada dada. Primero, se ingresa la probabilidad acumulada en la hoja de cálculo de Minitab (por ejemplo en C1). En este caso la probabilidad acumulada es 0.10. Después, los tres primeros pasos del procedimiento de Minitab son los dados antes. En el paso 4 se selecciona **Inverse cumulative probability** en lugar de **Cumulative probability** y se realiza la parte restante de este paso. Minitab da entonces 30 092 millas para la duración en la garantía.

Minitab también calcula las probabilidades de otras distribuciones de probabilidad continua, entre las que se encuentra la distribución de probabilidad exponencial. Para calcular probabilidades exponenciales se sigue el procedimiento antes dado para la distribución de probabilidad normal eligiendo la opción **Exponential** en el paso 3. El paso 4 es igual, salvo que no es necesario ingresar la desviación estándar. Minitab da los resultados de probabilidades acumuladas o probabilidades acumuladas inversas en la misma forma que se describió para la distribución de probabilidad normal.

Apéndice 6.2 Distribuciones de probabilidad continua con Excel

Excel permite calcular las probabilidades de varias distribuciones de probabilidad continuas. Entre las que se encuentran las distribuciones de probabilidad normal y exponencial. En este apéndice, se describe cómo usar Excel para calcular probabilidades en cualquier distribución normal. El procedimiento para la exponencial y para las otras distribuciones continuas es semejante al descrito aquí para la distribución normal.

Recuerde el problema de la empresa Grear Tire, la duración de los neumáticos en millas se describe mediante una distribución normal con media $\mu = 36\,500$ y $\sigma = 5000$. Suponga que se desea conocer la probabilidad de que un neumático dure más de 40 000 millas.

La función de Excel **DISTR.NORM.** suministra probabilidades acumuladas de una distribución normal. La forma general de la función es **DISTR.NORM** (*x*,media,desv_estándar,acum). En el cuarto argumento se especifica **VERDADERO** si se desea una probabilidad acumulada. De esta manera, para calcular la probabilidad acumulada de que la duración de un neumático sea menor o igual que 40 000 millas se ingresará la fórmula siguiente en cualquier celda de la hoja de cálculo Excel:

=DISTR.NORM(40000,36500,5000,VERDADERO)

En este momento, en la celda en que se ingresó la fórmula aparecerá 0.7580, indicando que la probabilidad de que la duración del neumático sea 40 000 millas es 0.7580. Por tanto, la probabilidad de que un neumático dure más de 40 000 millas es $1 - 0.7580 = 0.2420$.

La función de Excel **DISTR.NORM.INV.** usa un cálculo inverso para hallar el valor de *x* que corresponde a una probabilidad acumulada dada. Por ejemplo, si se desea hallar la duración que Grear debe ofrecer en su garantía de manera que no más de 10% de los neumáticos sean aptos para solicitar la garantía. Se ingresará en cualquier celda de la hoja de cálculo de Excel la fórmula siguiente:

=DISTR.NORM.INV(.1,36500,5000)

En este momento, en la celda en la que se ingresó la fórmula aparecerá 30092, indicando que la probabilidad de que un neumático dure 30 092 millas es 0.10.

La función de Excel para calcular probabilidades exponenciales es **DISTR.EXP.** Usar esta función es muy sencillo. Pero si se necesita ayuda para especificar los argumentos adecuados, se puede usar la herramienta Insertar Función de Excel (véase apéndice E).

CAPÍTULO 7



Muestreo y distribuciones muestrales

CONTENIDO

LA ESTADÍSTICA EN
LA PRÁCTICA: MEADWESTVACO
CORPORATION

7.1 EL PROBLEMA DE
MUESTREO DE
ELECTRONICS ASSOCIATES

7.2 MUESTREO ALEATORIO
SIMPLE
Muestreo de una población finita
Muestreo de una población
infinita

7.3 ESTIMACIÓN PUNTUAL

7.4 INTRODUCCIÓN A LAS
DISTRIBUCIONES
MUESTRALES

7.5 DISTRIBUCIÓN
MUESTRAL DE \bar{x}
Valor esperado de \bar{x}
Desviación estándar de \bar{x}
Forma de la distribución
muestral de \bar{x}
Distribución muestral
de \bar{x} en el problema EAI
Valor práctico de la distribución
muestral de \bar{x}

Relación entre el tamaño
de la muestra y la distribución
muestral de \bar{x}

7.6 DISTRIBUCIÓN
MUESTRAL DE \bar{p}
Valor esperado de \bar{p}
Desviación estándar de \bar{p}
Forma de la distribución
muestral de \bar{p}
Valor práctico de la distribución
muestral de \bar{p}

7.7 PROPIEDADES DE LOS
ESTIMADORES PUNTUALES
Insegadez
Eficiencia
Consistencia

7.8 OTROS MÉTODOS
DE MUESTREO
Muestreo aleatorio estratificado
Muestreo por conglomerados
Muestreo sistemático
Muestreo de conveniencia
Muestreo subjetivo

LA ESTADÍSTICA *en* LA PRÁCTICA

MEADWESTVACO CORPORATION*

STAMFORD, CONNECTICUT

MeadWestvaco Corporation, líder mundial en la producción de embalajes y papeles especiales, productos de consumo y de oficina y de sustancias químicas especiales, emplea a más de 30 000 personas. Opera a nivel mundial en 29 países y atiende a clientes localizados en 100 países. MeadWestvaco tiene una posición líder en la producción de papel, con una capacidad de 1.8 millones de toneladas anuales. Entre los productos de la empresa se encuentran papel para libros de texto, papel para revistas, sistemas de embalaje para bebidas y productos de oficina. Los consultores internos de MeadWestvaco usan el muestreo para obtener diversas informaciones que permiten a la empresa ganar productividad y seguir siendo competitiva.

Por ejemplo, MeadWestvaco posee bosques que le proporcionan los árboles, o la materia prima, para muchos de los productos de la empresa. Los directivos necesitan información confiable y precisa acerca de los bosques maderables para evaluar las posibilidades de satisfacción de las futuras necesidades de materia prima. ¿Cuál es el volumen actual de los bosques? ¿Cuál ha sido el crecimiento de los bosques? ¿Cuál es el crecimiento proyectado de los bosques? Las respuestas a estas preguntas permiten a los directivos de la empresa elaborar los planes para el futuro, tales como planes a largo plazo y calendarios para la poda de árboles.

¿Cómo recolecta MeadWestvaco la información que necesita acerca de los amplios bosques que requiere? Los datos que obtiene de puntos muestrales en los bosques son la base para contar con información acerca de la población de árboles propiedad de la empresa. Para localizar estos puntos muestrales, primero se dividen los bosques en tres secciones de acuerdo con la localización y tipo de árboles. Mediante mapas y números aleatorios los analistas de MeadWestvaco identifican puntos muestrales aleatorios de 1/5 a 1/7 acres en cada sección del bosque. Los ingenie-



El muestreo aleatorio de sus bosques permite a MeadWestvaco satisfacer necesidades futuras de materia prima. © Walter Hodges/Corbis.

ros forestales de MeadWestvaco recogen los datos de estos puntos muestrales para obtener información acerca de la población forestal.

También participan en el proceso de campo de la recolección de datos. Con periodicidad, en equipos de dos personas, recolectan la información de cada árbol en todos los puntos muestrales. Los datos muestrales se ingresan en el sistema computacional de inventario forestal continuo (IFC) de la empresa. Los informes obtenidos del sistema IFC contienen información de distribuciones de frecuencia con estadísticos sobre los tipos de árboles, volumen de los bosques, tasas de crecimiento anteriores y crecimiento y volumen proyectados para el futuro. El muestreo y las correspondientes informaciones estadísticas de los datos muestrales proporcionan la información esencial para la adecuada administración de los bosques de MeadWestvaco.

En este capítulo se estudiará el muestreo aleatorio simple y el proceso de selección de muestras. Se verá también cómo se usan estadísticos como la media muestral y la proporción muestral para estimar la media de la población y la proporción de la población.

*Los autores agradecen al doctor Edgard P. Winkofsky por proporcionar la información para *La estadística en la práctica*.

En el capítulo 1 se definieron los términos población y muestra. Estas definiciones se retoman aquí.

1. Una *población* es el conjunto de todos los elementos que interesan en un estudio.
2. Una *muestra* es un subconjunto de la población.

A las características numéricas de una población, como la media y la desviación estándar, se les llama **parámetros**. El principal propósito de la inferencia estadística es hacer estimaciones y pruebas de hipótesis acerca de los parámetros poblacionales usando la información que propor-

ciona una muestra. Para empezar, se presentan dos situaciones en las que a partir de muestras se obtienen estimaciones de parámetros poblacionales.

1. Un fabricante de neumáticos elabora un nuevo modelo que tendrá mayor duración que los actuales neumáticos de la empresa. Para estimar la duración media, en millas, el fabricante selecciona una muestra de 120 neumáticos nuevos para probarlos. De los resultados de esta prueba se obtiene una duración media de 36 500 millas. Por tanto, una estimación de la duración media, en millas, de la población de nuevos neumáticos es 36 500 millas.
2. Los miembros de un partido político deseaban apoyar a un determinado candidato para senador, y los dirigentes del partido deseaban tener una estimación de la proporción de votantes registrados que podían estar a favor del candidato. El tiempo y el costo de preguntar a cada uno de los individuos de la población de votantes registrados eran prohibitivos. Por tanto, se seleccionó una muestra de 400 votantes registrados; 160 de los 400 votantes indicaron estar a favor del candidato. Una estimación de la proporción de la población de votantes registrados a favor del candidato es $160/400 = 0.40$.

Estos dos ejemplos ilustran algunas de las razones por las que se usan muestras. Observe que en el ejemplo de los neumáticos, obtener datos sobre su tiempo de duración implica usarlos hasta que se acaben. Es claro que no es posible probar toda la población de neumáticos; una muestra es la única manera factible de obtener los datos de duración deseados. En el ejemplo del candidato, preguntar a cada uno de los votantes registrados es, en teoría, posible, pero el tiempo y el costo para hacerlo son prohibitivos; de manera que se prefiere una muestra de los votantes registrados.

Es importante darse cuenta de que los resultados muestrales sólo proporcionan una *estimación* de los valores de las características de la población. No se espera que la media muestral de 36 500 millas sea exactamente igual al millaje medio de todos los neumáticos de la población, tampoco que 0.40, o 40% de la población de los votantes registrados esté a favor del candidato. La razón es simple, la muestra sólo contiene una parte de la población. Con métodos de muestreo adecuados, los resultados muestrales proporcionarán estimaciones “buenas” de los parámetros poblacionales. Pero ¿cuán buenos puede esperarse que sean los resultados muestrales? Por fortuna, existen procedimientos estadísticos para responder esta pregunta.

En este capítulo se enseña cómo emplear el muestreo aleatorio simple para seleccionar una muestra de una población. Después, cómo usar una muestra aleatoria simple para calcular estimaciones de una media poblacional, de una desviación estándar poblacional y de una proporción poblacional. Además, también se presenta el importante concepto de distribución muestral. Como verá, el conocimiento de la distribución muestral adecuada permite decir qué tan cerca se encuentran las estimaciones muestrales de los correspondientes parámetros poblacionales. En la última sección se estudian alternativas al muestreo aleatorio simple, usadas con frecuencia en la práctica.

Una media muestral suministra una estimación de la media poblacional y una proporción muestral suministra una estimación de la proporción poblacional. Con dichas estimaciones puede esperarse un cierto error de estimación. Este capítulo enseña las bases para estimar cuán grande puede ser ese error.

7.1

El problema de muestreo de Electronics Associates

Al director de personal de Electronics Associates, Inc. (EAI), se le ha encargado la tarea de elaborar un perfil de los 2500 administradores de la empresa. Las características a determinar son el sueldo medio anual de los administradores y la proporción de administradores que ha terminado el programa de capacitación de la empresa.

Con los 2500 administradores de la empresa como la población para este estudio, es posible hallar el sueldo anual y la situación respecto al programa de capacitación de cada persona al consultar los archivos del personal. El archivo con los datos que contiene esta información para cada uno de los 2500 administradores que forman la población se encuentra en el disco compacto que se distribuye con el libro.

Con los datos de EAI y las fórmulas presentadas en el capítulo 3, se calcula la media poblacional y la desviación estándar poblacional de los salarios anuales.

Media poblacional: $\mu = \$51\,800$

Desviación estándar poblacional: $\sigma = \$4000$



Algunos de los costos de recopilar información de una muestra son sustancialmente menores que hacerlo de una población; especialmente cuando se deben realizar entrevistas personales para recopilar la información.

Los datos sobre la situación del programa de capacitación muestran que 1500 de los 2500 administradores han terminado el programa de capacitación. Si p denota la proporción de la población que ha terminado el programa de capacitación, se tiene que $p = 1500/2500 = 0.60$. La media poblacional de los sueldos anuales ($\mu = \$51\,800$), la desviación estándar poblacional de los sueldos anuales ($\sigma = \$4000$) y la proporción poblacional de quienes han terminado el programa de capacitación ($p = 0.60$) son parámetros de la población de administradores de EAI.

Ahora suponga que la información necesaria sobre todos los administradores de EAI no esté disponible en la base de datos de la empresa. La pregunta que se considera ahora es: ¿cómo puede obtener el director de personal de la empresa, estimaciones de los parámetros poblacionales usando una muestra de los administradores, en lugar de usar a los 2500 administradores de la población. Asuma que se va a emplear una muestra de 30 administradores. Es obvio que el tiempo y el costo de la elaboración de un perfil será mucho menor usando 30 administradores que la población entera. Si el director de personal tuviera la certeza de que una muestra de 30 administradores proporciona la información adecuada acerca de la población de los 2500 administradores, preferiría trabajar con una muestra que hacerlo con toda la población. Para explorar la posibilidad de usar una muestra para el estudio de EAI, primero se considerará cómo determinar una muestra de 30 administradores.

7.2

Muestreo aleatorio simple

Para seleccionar una muestra de una población hay diversos métodos; uno de los más comunes es el **muestreo aleatorio simple**. La definición de muestreo aleatorio simple y del proceso de seleccionar una muestra aleatoria simple dependen de si la población es *finita* o *infinita*. Como el problema de muestreo de EAI tiene una población finita de 2500 administradores, primero se considera el muestreo de una población finita.

Muestreo de una población finita

Una muestra aleatoria simple de tamaño n de una población finita de tamaño N se define como sigue.

MUESTREO ALEATORIO SIMPLE (POBLACIÓN FINITA)

Una muestra aleatoria simple de tamaño n de una población finita de tamaño N es una muestra seleccionada de manera que cada posible muestra de tamaño n tenga la misma probabilidad de ser seleccionada.

Un procedimiento para seleccionar una muestra aleatoria simple de una población finita es elegir los elementos para la muestra de uno en uno, de manera que, en cada paso, cada uno de los elementos que quedan en la población tenga la misma probabilidad de ser seleccionado. Al seleccionar n elementos de esta manera, será satisfecha la definición de muestra aleatoria simple seleccionada de una población finita.

Para seleccionar una muestra aleatoria simple de la población finita de administradores de EAI, primero se le asigna a cada administrador un número. Por ejemplo, se les asignan los números del 1 al 2500 en el orden en que aparecen sus nombres en el archivo de personal de EAI. A continuación se consulta la tabla de dígitos aleatorios que se muestran en la tabla 7.1. Al consultar el primer renglón de la tabla se da cuenta que cada dígito, 6, 3, 2, ... es un dígito aleatorio con la misma oportunidad de aparecer que cualquier otro. Como el número mayor en la lista de la población de administradores de EAI, 2500, tiene cuatro dígitos, se seleccionarán números aleatorios de la tabla en conjuntos o grupos de cuatro dígitos. Aun cuando para la selección de números aleatorios se puede empezar en cualquier lugar de la tabla y avanzar sistemáticamente en una de las cuatro direcciones, aquí se usará el primer renglón de la tabla 7.1 y se avanzará de izquierda a derecha. Los primeros 7 números aleatorios de cuatro dígitos son

6327 1599 8671 7445 1102 1514 1807

Los números aleatorios en la tabla aparecen en grupos de cinco para facilitar su lectura.

Los números aleatorios generados por computadora también sirven para realizar el proceso de selección de una muestra aleatoria. Excel proporciona una función para generar números aleatorios en sus hojas de cálculo.

Los números aleatorios en la tabla aparecen en grupos de cinco para facilitar su lectura.

TABLA 7.1 NÚMEROS ALEATORIOS

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289

Como los números de la tabla son aleatorios, estos números de cuatro dígitos son todos igualmente posibles. Ahora se pueden usar estos números aleatorios de cuatro dígitos para darle a cada uno de los administradores que constituyen la población la misma oportunidad de ser incluido en la muestra aleatoria. El primer número, 6327, es mayor que 2500. No corresponde a ninguno de los administradores numerados que forman la población y por tanto se descarta. El segundo número, 1599, está entre 1 y 2500. Por tanto, el primer administrador seleccionado para la muestra aleatoria es el administrador que tiene el número 1599 en la lista de los administradores de EAI. Siguiendo este proceso, se ignoran los números 8671 y 7445 antes de identificar a los administradores con los números 1102, 1514 y 1807 e incluirlos en la muestra aleatoria. Este proceso sigue hasta que se tiene la muestra aleatoria de 30 administrativos de EAI.

Al realizar este proceso para la selección de una muestra aleatoria simple, es posible que un número que ya haya sido usado se encuentre de nuevo en la tabla antes de completar la muestra de los 30 administradores de EAI. Como no se quiere seleccionar a un administrador más de una vez, cualquier número aleatorio que ya ha sido usado se ignora, porque el administrador correspondiente ya se ha incluido en la muestra. A este tipo de selección se le conoce como **muestreo sin reemplazo**. Cuando se selecciona una muestra en la que se acepten números aleatorios ya usados y los administradores correspondientes sean incluidos dos o más veces, se está **muestreando con reemplazo**. Muestrear con reemplazo es una forma válida de identificar una muestra aleatoria simple. Sin embargo, el muestreo sin reemplazo es el procedimiento de muestreo más usado. Cuando se habla de muestreo aleatorio simple, se asumirá que el muestreo es sin reemplazo.

Muestreo de una población infinita

En algunas situaciones la población o bien es infinita o tan grande que, para fines prácticos, se considera infinita. Por ejemplo, suponga que un restaurante de comida rápida desea obtener el

En la práctica, la población en estudio se considera infinita si se tiene un proceso continuo en el que sea imposible contar o enumerar cada uno de los elementos de la población.

perfil de su clientela seleccionando una muestra aleatoria de los mismos y pidiéndole a cada cliente que llene un breve cuestionario. En tales situaciones, el proceso continuo de clientes que visitan el restaurante puede verse como que los clientes provienen de una población infinita. La definición de muestra aleatoria simple tomada de una población infinita es la siguiente

MUESTRA ALEATORIA SIMPLE (POBLACIÓN INFINITA)

Una muestra aleatoria simple de una población infinita es una muestra seleccionada de manera que se satisfagan las condiciones siguientes.

1. Cada uno de los elementos seleccionados proviene de la población.
2. Cada elemento se selecciona independientemente.

En poblaciones infinitas un procedimiento para la selección de una muestra debe ser concebido especialmente para cada situación, de manera que permita seleccionar los elementos de manera independiente y evitar así un sesgo en la selección que dé mayores probabilidades de selección a ciertos tipos de elementos.

En poblaciones infinitas un procedimiento para la selección de una muestra debe ser concebido especialmente para cada situación, de manera que permita seleccionar los elementos de manera independiente y evitar así un sesgo en la selección que dé mayores probabilidades de selección a ciertos tipos de elementos. En el ejemplo de la selección de una muestra aleatoria simple entre los clientes de un restaurante de comida rápida, el primer requerimiento es satisfecho por cualquier cliente que entra en el restaurante. El segundo requerimiento es satisfecho seleccionando a los clientes de manera independiente. El objetivo del segundo requerimiento es evitar sesgos de selección. Habría un sesgo de selección, por ejemplo, si cinco clientes consecutivos que se seleccionaran fueran amigos. Es de esperar que estos clientes tengan perfiles semejantes. Dichos sesgos se evitan haciendo que la selección de un cliente no influya en la selección de cualquier otro cliente. En otras palabras, los clientes deben ser seleccionados de manera independiente.

McDonald's, el restaurante líder en comida rápida, realizó un muestreo aleatorio simple precisamente en una situación así. El procedimiento de muestreo se basó en el hecho de que algunos clientes presentaban cupones de descuento. Cada vez que un cliente presentaba un cupón de descuento, al siguiente cliente que se atendía se le pedía que llenara un cuestionario sobre el perfil del cliente. Como los clientes que llegaban al restaurante presentaban cupones de descuento aleatoria e independientemente, este plan de muestreo garantizaba que los clientes fueran seleccionados de manera independiente. Por tanto, los dos requerimientos para un muestreo aleatorio simple de una población infinita fueron satisfechos.

Las poblaciones infinitas suelen asociarse con un proceso que opera continuamente a lo largo del tiempo. Por ejemplo, partes fabricadas en una línea de producción, transacciones en un banco, llamadas que llegan a un centro de asesoría técnica y clientes que entran en las tiendas son considerados como provenientes de una población infinita. En tales casos un procedimiento de muestreo creativo garantiza que no haya sesgos de selección y que los elementos de la muestra sean seleccionados en forma independiente.

NOTAS Y COMENTARIOS

1. El número de muestras aleatorias simples distintas de tamaño n que pueden seleccionarse de una población finita de tamaño N es

$$\frac{N!}{n!(N - n)!}$$

En esta fórmula $N!$ y $n!$ son factoriales, vistos en el capítulo 4. Al usar esta expresión con los

datos del problema de EAI, en el que $N = 2500$ y $n = 30$, se ve que se pueden tomar 2.75×10^{69} muestras aleatorias simples distintas de 30 administradores de EAI.

2. Para tomar una muestra aleatoria pueden emplearse paquetes de software. En los apéndices del capítulo se muestra cómo usar Minitab y Excel para seleccionar una muestra aleatoria simple de una población finita

Ejercicios

Método

Autoexamen

1. Dada una población finita que tiene cinco elementos A, B, C, D y E seleccione 10 muestras aleatorias simples de tamaño 2.
 - a. Enumere las 10 muestras empezando con AB, AC y así en lo sucesivo.
 - b. Usando el muestreo aleatorio simple, ¿cuál es la probabilidad que tiene cada muestra de tamaño 2 de ser seleccionada?
 - c. Si el número aleatorio 1 corresponde a A, el número 2 corresponde a B y así en lo sucesivo. Enliste la muestra aleatoria de tamaño 2 que será seleccionada al usar los números aleatorios 8 0 5 7 5 3 2.
2. Suponga que una población finita tiene 350 elementos. A partir de los últimos tres dígitos de cada uno de los siguientes números aleatorios de cinco dígitos (por ejemplo: 601, 022, 448,...), determine los primeros cuatro elementos que se seleccionarán para una muestra aleatoria simple.

98601 73022 83448 02147 34229 27553 84147 93289 14209

Aplicaciones

Autoexamen

3. *Fortune* publicó datos sobre ventas, valor del activo, valor de mercado y ganancias por acción de las 500 corporaciones industriales más grandes de Estados Unidos (*Fortune* 500, 2003). Suponga que usted desea seleccionar una muestra aleatoria simple de 10 corporaciones de la lista *Fortune* 500. Use los tres últimos dígitos de la columna 9 de la tabla 7.1, empezando con 554. Leyendo hacia abajo por esa columna, identifique los números de las 10 corporaciones que se tomarán para la muestra.
4. A continuación se presentan las 10 acciones más activas en la Bolsa de Nueva York del 6 de marzo del 2006 (*The Wall Street Journal*, 7 de marzo, 2006).

AT&T	Lucent	Nortel	Qwest	Bell South
Pfizer	Texas Instruments	Gen. Elect.	iShrMSJpn	LSI Logic

Las autoridades decidieron investigar las prácticas de negociación usando una muestra de tres de estas acciones.

- a. Empezando en el primer dígito aleatorio de la columna seis de la tabla 7.1, lea los números descendiendo por esa columna para seleccionar una muestra aleatoria simple de tres acciones para las autoridades.
 - b. Con la información dada en la primera nota y comentario, determine cuántas muestras aleatorias simples diferentes de tamaño 3 pueden seleccionarse de una lista de 10 acciones.
5. Una organización de estudiantes desean estimar la proporción de estudiantes que están a favor de una disposición de la escuela. Se cuenta con una lista con los nombres y direcciones de los 645 estudiantes inscritos el presente trimestre. Tomando números aleatorios de tres dígitos del renglón 10 de la tabla 7.1 y avanzando por ese renglón de izquierda a derecha, determine los 10 primeros estudiantes que serán seleccionados usando un muestreo aleatorio simple. Los números aleatorios de tres dígitos empiezan con 816, 283 y 610.
 6. El *County and City Data Book* del Census Bureau cuenta con información de los 3139 condados de Estados Unidos. Suponga que para un estudio nacional se recogerán datos de 30 condados seleccionados aleatoriamente. De la última columna de la tabla 7.1 extraiga números aleatorios de cuatro dígitos para determinar los primeros cinco condados seleccionados para la muestra. Ignore los primeros dígitos y empiece con los números aleatorios de cuatro dígitos 9945, 8364, 5702 y así sucesivamente.
 7. Suponga que se va a tomar una muestra aleatoria simple de 12 de los 372 médicos de una determinada ciudad. Una organización médica le proporciona los nombres de los médicos. De la tabla

7.1 use la columna ocho de números aleatorios de cinco dígitos para determinar cuáles serán los 12 médicos para la muestra. Ignore los primeros dos dígitos de cada grupo de cinco dígitos de números aleatorios. Este proceso empieza con el número aleatorio 108 y continúa descendiendo por la columna de números aleatorios.

8. La lista siguiente proporciona los 25 mejores equipos de futbol de la NCAA en la temporada del 2002 (*NCAA News*, 4 de enero de 2003). De la tabla 7.1 use la novena columna que empieza con 13 554, para seleccionar una muestra aleatoria simple de seis equipos de futbol. Empiece con el equipo 13 y use los primeros dos dígitos de cada renglón de la novena columna para el proceso de selección. ¿Cuáles son los seis equipos de futbol seleccionados para la muestra aleatoria simple?

1. Ohio State	14. Virginia Tech
2. Miami	15. Penn State
3. Georgia	16. Auburn
4. Southern California	17. Notre Dame
5. Oklahoma	18. Pittsburgh
6. Kansas State	19. Marshall
7. Texas	20. West Virginia
8. Iowa	21. Colorado
9. Michigan	22. TCU
10. Washington State	23. Florida State
11. North Carolina State	24. Florida
12. Boise State	25. Virginia
13. Maryland	

9. *The Wall Street Journal* proporciona el valor de activo neto, el rendimiento porcentual en lo que va del año y el rendimiento porcentual en tres años de 555 fondos mutualistas (*The Wall Street Journal*, 25 de abril de 2003). Suponga que se va a usar una muestra aleatoria simple de 12 de estos 555 fondos mutualistas para un estudio acerca de su tamaño y desempeño. Use la cuarta columna de números aleatorios en la tabla 7.1 empezando con el número 51102, para seleccionar la muestra aleatoria simple de 12 fondos mutualistas. Empiece con el fondo 102 y use los últimos tres dígitos de cada renglón de la cuarta columna para el proceso de selección. ¿Cuáles son los números de los 12 fondos mutualistas en esta muestra aleatoria simple?
10. Indique si las poblaciones siguientes se consideran finitas o infinitas.
- Todos los votantes registrados en el estado de California.
 - Todos los equipos de televisión que pueden ser producidos en una determinada fábrica.
 - Todas las órdenes que pueden ser procesadas por Allentown, Pensilvania, planta de TV-M Company.
 - Todas las llamadas de emergencia que pueden ser recibidas en una estación de policía.
 - Todas las piezas producidas por Fibercon, Inc., en el segundo turno el 17 de mayo.

7.3

Estimación puntual

Una vez descrito cómo seleccionar una muestra aleatoria simple, se vuelve al problema de EAI. En la tabla 7.2 se presenta una muestra aleatoria simple de 30 administradores con sus respectivos datos de sueldo anual y de participación en el programa de capacitación. La notación x_1 , x_2 , etc., se usa para denotar el sueldo anual del primer administrador de la muestra, del segundo, y así sucesivamente. La participación en el programa de capacitación se indica por un Sí en la columna programa de entrenamiento.

Para estimar el valor de un parámetro poblacional, la característica correspondiente se calcula con los datos de la muestra, a lo que se le conoce como **estadístico muestral**. Por ejemplo, para estimar la media poblacional μ y la desviación estándar poblacional σ de los salarios anuales de los administradores de EAI, se emplean los datos de la tabla 7.2 y se calculan los es-

TABLA 7.2 SALARIOS ANUALES Y SITUACIÓN RESPECTO AL PROGRAMA DE CAPACITACIÓN DE LOS ADMINISTRADORES PERTENECIENTES A UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES DE EAI

Salario anual	Programa de capacitación	Salario anual (\$)	Programa de capacitación
$x_1 = 49\,094.30$	Sí	$x_{16} = 51\,766.00$	Sí
$x_2 = 53\,263.90$	Sí	$x_{17} = 52\,541.30$	No
$x_3 = 49\,643.50$	Sí	$x_{18} = 44\,980.00$	Sí
$x_4 = 49\,894.90$	Sí	$x_{19} = 51\,932.60$	Sí
$x_5 = 47\,621.60$	No	$x_{20} = 52\,973.00$	Sí
$x_6 = 55\,924.00$	Sí	$x_{21} = 45\,120.90$	Sí
$x_7 = 49\,092.30$	Sí	$x_{22} = 51\,753.00$	Sí
$x_8 = 51\,404.40$	Sí	$x_{23} = 54\,391.80$	No
$x_9 = 50\,957.70$	Sí	$x_{24} = 50\,164.20$	No
$x_{10} = 55\,109.70$	Sí	$x_{25} = 52\,973.60$	No
$x_{11} = 45\,922.60$	Sí	$x_{26} = 50\,241.30$	No
$x_{12} = 57\,268.40$	No	$x_{27} = 52\,793.90$	No
$x_{13} = 55\,688.80$	Sí	$x_{28} = 50\,979.40$	Sí
$x_{14} = 51\,564.70$	No	$x_{29} = 55\,860.90$	Sí
$x_{15} = 56\,188.20$	No	$x_{30} = 57\,309.10$	No

estadísticos muestrales correspondientes; media muestral \bar{x} y desviación estándar muestral s . Con las fórmulas para la media muestral y la desviación estándar muestral presentadas en el capítulo 3 se obtiene que la media muestral es

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1\,554\,420}{30} = \$51\,814$$

y la desviación estándar muestral es

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{325\,009\,260}{29}} = \$3\,348$$

Para estimar p , la proporción de administradores que han terminado el programa de capacitación, se usa la proporción muestral correspondiente \bar{p} . Sea x el número de administradores de la muestra que han terminado el programa de capacitación. De acuerdo con la tabla 7.2, $x = 19$. Por tanto, como el tamaño de la muestra es $n = 30$, la proporción muestral es

$$\bar{p} = \frac{x}{n} = \frac{19}{30} = 0.63$$

Al hacer los cálculos anteriores, se lleva a cabo el proceso estadístico conocido como *estimación puntual*. A la media muestral \bar{x} se le conoce como el **estimador puntual** de la media poblacional μ , a la desviación estándar muestral s como el estimador puntual de la desviación estándar poblacional σ y a la proporción muestral \bar{p} como el estimador puntual de la proporción poblacional p . Al valor numérico obtenido de \bar{x} , s , o \bar{p} se les conoce como **estimaciones puntuales**. Así, en la muestra aleatoria simple de 30 administradores de EAI que se presenta en la tabla 7.2, \$51 814 es la estimación puntual de μ , \$3 348 es la estimación puntual de σ y 0.63 es la estimación puntual de p . En la tabla 7.3 se resumen los resultados muestrales y se comparan las estimaciones puntuales con los valores de los parámetros poblacionales.

TABLA 7.3 INFORMACIÓN DE LAS ESTIMACIONES PUNTUALES OBTENIDAS DE UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES DE EAI

Parámetro poblacional	Valor del parámetro	Estimador puntual	Estimación puntual
μ = Media poblacional de los salarios anuales	\$51 800	\bar{x} = Media muestral de los salarios anuales	\$51 814
σ = Desviación estándar poblacional de los salarios anuales	\$4 000	s = Desviación estándar muestral de los salarios anuales	\$3 348
p = Proporción poblacional que ha terminado el programa de capacitación	0.60	\bar{p} = Proporción muestral que ha terminado el programa de capacitación	0.63

Como se observa en la tabla 7.3, las estimaciones puntuales difieren un poco de los correspondientes parámetros poblacionales. Estas diferencias son de esperarse ya que para elaborar las estimaciones muestrales se usa una muestra, y no un censo de toda la población. En el capítulo siguiente se verá cómo elaborar un intervalo de estimación para tener información acerca de qué tan cerca está la estimación muestral del parámetro poblacional.

Ejercicios

Métodos

11. Los datos siguientes provienen de una muestra aleatoria simple.

5	8	10	7	10	14
---	---	----	---	----	----

 - a. ¿Cuál es la estimación puntual de la media poblacional?
 - b. ¿Cuál es la estimación puntual de la desviación estándar poblacional?
12. Como respuestas a una pregunta de una encuesta a 150 individuos de una muestra se obtuvieron 75 Sí, 55 No y 20 individuos no dieron su opinión.
 - a. ¿Cuál es la estimación puntual de la proporción de la población que responde Sí?
 - b. ¿Cuál es la estimación puntual de la proporción de la población que responde No?

Aplicaciones

13. La siguiente información son datos obtenidos en una muestra aleatoria de las ventas de 5 meses:

Mes:	1	2	3	4	5
Unidades vendidas:	94	100	85	94	92

 - a. Calcule una estimación puntual de la media poblacional del número medio de unidades vendidas por mes.
 - b. Calcule una estimación puntual de la desviación estándar del número de unidades vendidas por mes.
14. *BusinessWeek* publicó información sobre 283 fondos mutualistas (*BusinessWeek* 26 de enero de 2004). En el conjunto de datos MutualFunds se encuentra una muestra de 40 de estos fondos. Use este conjunto de datos para hacer lo que se pide en los incisos siguientes.
 - a. Calcule una estimación puntual de la proporción de fondos de inversión de *BusinessWeek* que son fondos de cargo.
 - b. Calcule una estimación puntual de la proporción de fondos clasificados como de alto riesgo.
 - c. Calcule una estimación puntual de la proporción de fondos con una puntuación abajo del promedio para el riesgo.
15. Muchos de los medicamentos empleados en el tratamiento del cáncer son costosos. *BusinessWeek* informó de los costos de los tratamientos con Herceptin, un medicamento para tratar el cáncer de

Autoexamen

Autoexamen

archivo
en
MutualFund

CD

mama (*BusinessWeek*, 30 de enero de 2006). Los siguientes son los costos de tratamientos con Herceptin en una muestra aleatoria de 10 pacientes.

4376	5578	2717	4920	4495
4798	6446	4119	4237	3814

- Calcule una estimación puntual del costo medio de un tratamiento con Herceptin.
 - Calcule una estimación puntual de la desviación estándar en los costos de los tratamientos con Herceptin.
- En una muestra de 50 empresas de *Fortune* 500, 5 se encontraban en Nueva York, 6 en California, 2 en Minesota y 1 en Wisconsin.
 - Dé una estimación de la proporción de empresas de *Fortune* 500 que se encuentran en Nueva York.
 - Dé una estimación del número de empresas de *Fortune* 500 que se encuentran en Minesota.
 - Dé una estimación de la proporción de empresas de *Fortune* 500 que no se encuentran en ninguno de estos estados.
 - La American Association of Individuals Investors (AAII) hace sondeos semanales entre sus suscriptores para determinar cuántos se muestran optimistas, pesimistas o indiferentes respecto al mercado de acciones a corto plazo. Sus hallazgos en la semana que terminó el 2 de marzo de 2006 son consistentes con los resultados muestrales siguientes (www.aaii.com).

Optimistas 409 Indiferentes 299 Pesimistas 291

Dé una estimación puntual de los parámetros poblacionales siguientes.

- Proporción de suscriptores de AAII optimistas respecto al mercado de acciones.
- Proporción de suscriptores de AAII indiferentes respecto al mercado de acciones.
- Proporción de suscriptores de AAII pesimistas respecto al mercado de acciones.

7.4

Introducción a las distribuciones muestrales

En la sección anterior se dijo que la media muestral \bar{x} es el estimador puntual de la media poblacional μ y que la proporción muestral \bar{p} es el estimador puntual de la proporción poblacional p . En la muestra aleatoria simple de los 30 administradores de EAI que se presenta en la tabla 7.2, la estimación puntual de μ es $\bar{x} = \$51\,814$ y la estimación puntual de p es $\bar{p} = 0.63$. Suponga que se selecciona otra muestra aleatoria simple de 30 administradores de EAI y se obtienen las estimaciones puntuales siguientes:

Media muestral: $\bar{x} = \$52\,670$

Proporción muestral: $\bar{p} = 0.70$

Observe que se obtuvieron valores diferentes de \bar{x} y de \bar{p} . En efecto, otra muestra aleatoria simple de 30 administradores de EAI no se puede esperar que dé las mismas estimaciones puntuales que la primera muestra.

Ahora suponga que el proceso de seleccionar una muestra aleatoria simple de 30 administradores se repite una y otra y otra vez y que cada vez se calculan los valores de \bar{x} y de \bar{p} . En la tabla 7.4 se muestra una parte de los resultados obtenidos en 500 muestras aleatorias simples y en la tabla 7.5 las distribuciones de frecuencias y distribuciones de frecuencias relativas de los valores de las 500 \bar{x} . En la figura 7.1 se muestra el histograma de las frecuencias de los valores de \bar{x} .

En el capítulo 5 se definió una variable aleatoria como una descripción numérica del resultado de un experimento. Si el proceso de seleccionar una muestra aleatoria simple se considera como un experimento, la media muestral \bar{x} es el valor numérico del resultado de ese experimento. Por tanto, la media muestral es una variable aleatoria. Entonces, como ocurre con otras variables aleatorias, \bar{x} tiene una media o valor esperado, una desviación estándar y una distribución

Poder entender el material de los capítulos siguientes depende de entender y usar las distribuciones muestrales que se presentan en este capítulo.

TABLA 7.4 VALORES DE \bar{x} Y DE \bar{p} OBTENIDOS EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES DE EAI CADA UNA

Muestra número	Media muestral (\bar{x})	Proporción muestral (\bar{p})
1	51 814	0.63
2	52 670	0.70
3	51 780	0.67
4	51 588	0.53
·	·	·
·	·	·
·	·	·
500	51 752	0.50

de probabilidad. Como los distintos valores que toma \bar{x} son resultado de distintas muestras aleatorias simples, a la distribución de probabilidad de \bar{x} se le conoce como **distribución muestral** de \bar{x} . Conocer esta distribución muestral y sus propiedades permitirá hacer declaraciones de probabilidad de qué tan cerca está la media muestral \bar{x} de la media poblacional μ .

De regreso a la figura 7.1, se necesitaría enumerar todas las muestras posibles de 30 administradores y calcular cada una de las medias muestrales para determinar totalmente la distribución muestral de \bar{x} . Sin embargo, el histograma de 500 valores \bar{x} da una aproximación a esta distribución muestral. En esta aproximación se observa la apariencia de curva de campana de esta distribución. Además, la mayor concentración de valores de \bar{x} y la media de los 500 valores de \bar{x} se encuentran cerca de la media poblacional $\mu = \$51\ 800$. En la sección siguiente se describirán más detalladamente las propiedades de la distribución muestral de \bar{x} .

Los 500 valores de las proporciones muestrales \bar{p} se resumen en el histograma de frecuencias relativas de la figura 7.2. Como ocurre con \bar{x} , \bar{p} es una variable aleatoria. Si se tomara cada muestra posible de tamaño 30 y para cada muestra se calculara el valor \bar{p} , la distribución de probabilidad que se obtuviera sería la distribución muestral de \bar{p} . En la figura 7.2, el histograma de frecuencias relativas de los 500 valores muestrales da una idea general de la apariencia de la distribución muestral de \bar{p} .

En la práctica sólo se selecciona una muestra aleatoria simple de la población. En esta sección el proceso de muestreo se repitió 500 veces para ilustrar que es posible tomar muchas mues-

TABLA 7.5 DISTRIBUCIÓN DE FRECUENCIAS DE \bar{x} EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES DE EAI CADA UNA

Salario anual medio (\$)	Frecuencia	Frecuencia relativa
49 500.00–49 999.99	2	0.004
50 000.00–50 499.99	16	0.032
50 500.00–50 999.99	52	0.104
51 000.00–51 499.99	101	0.202
51 500.00–51 999.99	133	0.266
52 000.00–52 499.99	110	0.220
52 500.00–52 999.99	54	0.108
53 000.00–53 499.99	26	0.052
53 500.00–53 999.99	6	0.012
	Totales 500	1.000

FIGURA 7.1 HISTOGRAMA DE LAS FRECUENCIAS RELATIVAS DE LOS VALORES DE \bar{x} OBTENIDOS EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES CADA UNA

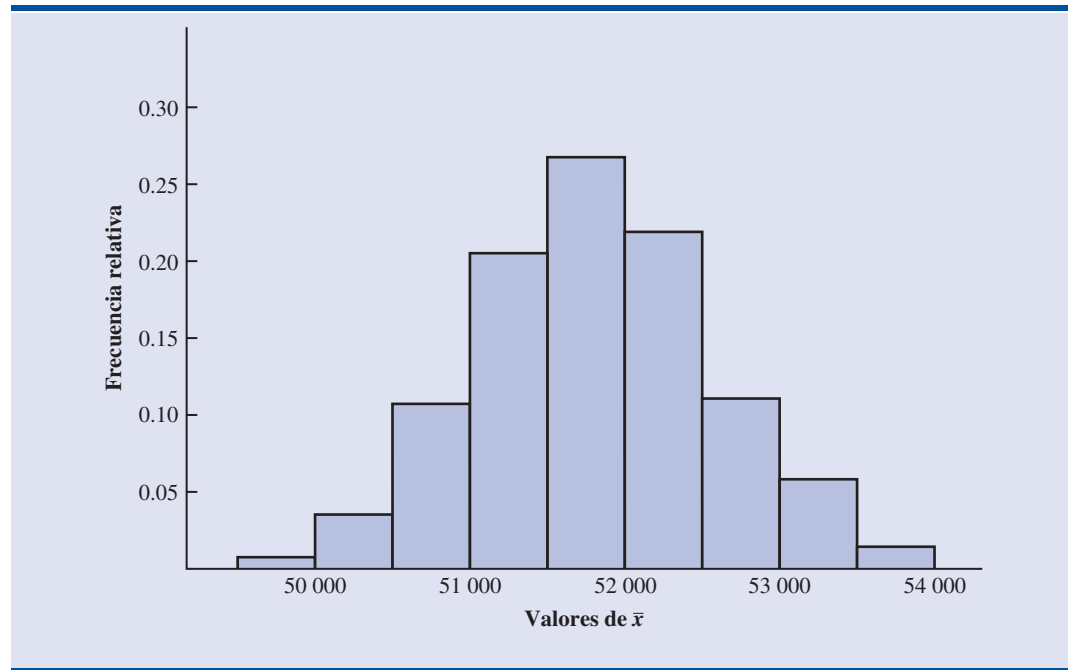
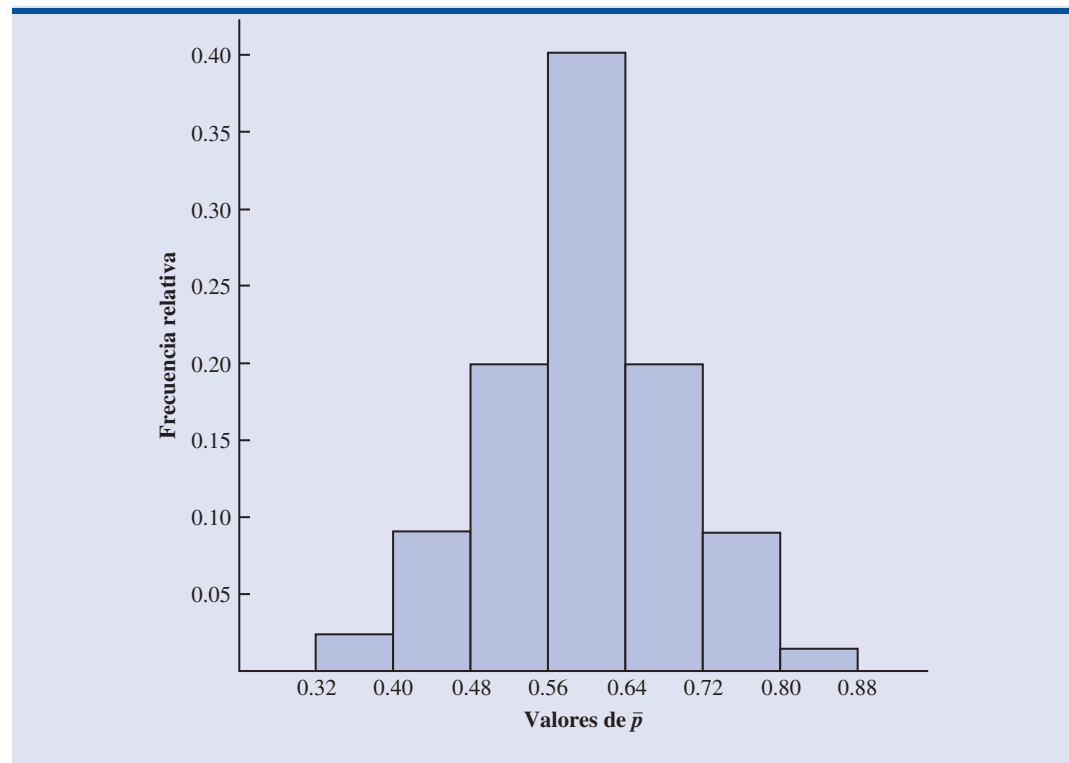


FIGURA 7.2 HISTOGRAMA DE LAS FRECUENCIAS RELATIVAS DE LOS VALORES DE \bar{p} OBTENIDOS EN 500 MUESTRAS ALEATORIAS SIMPLES DE 30 ADMINISTRADORES CADA UNA



tras diferentes y que diferentes muestras darán valores distintos de los estadísticos muestrales \bar{x} y \bar{p} . A la distribución muestral de cualquier estadístico determinado se le llama distribución muestral del estadístico. En la sección 7.5 se presentan las características de la distribución muestral de \bar{x} . En la sección 7.6 se muestran las características de la distribución muestral de \bar{p} .

7.5

Distribución muestral de \bar{x}

En la sección anterior se dijo que la media muestral \bar{x} es una variable aleatoria y que a su distribución de probabilidad se le llama distribución muestral de \bar{x} .

DISTRIBUCIÓN MUESTRAL DE \bar{x}

La distribución muestral de \bar{x} es la distribución de probabilidad de todos los valores de la media muestral \bar{x} .

En esta sección se describen las propiedades de la distribución muestral de \bar{x} . Como ocurre con otras distribuciones de probabilidad estudiadas, la distribución muestral de \bar{x} tiene un valor esperado, una desviación estándar y una forma característica. Para empezar se considerará la media de todos los valores de \bar{x} , a la que se conoce como valor esperado de \bar{x} .

Valor esperado de \bar{x}

En el problema de muestreo de EAI se vio que en distintas muestras aleatorias simples se obtienen valores diferentes para la media muestral \bar{x} . Como la variable aleatoria \bar{x} puede tener muchos valores diferentes, suele ser de interés conocer la media de todos los valores de \bar{x} que se obtienen con diferentes muestras aleatorias simples. La media de la variable aleatoria \bar{x} es el valor esperado de \bar{x} . Sea $E(\bar{x})$ el valor esperado de \bar{x} y μ la media de la población de la que se selecciona una muestra aleatoria simple. Se puede demostrar que cuando se emplea el muestreo aleatorio simple, $E(\bar{x})$ y μ son iguales.

VALOR ESPERADO DE \bar{x}

$$E(\bar{x}) = \mu \quad (7.1)$$

donde

$$\begin{aligned} E(\bar{x}) &= \text{valor esperado de } \bar{x} \\ \mu &= \text{media poblacional} \end{aligned}$$

El valor esperado de \bar{x} es igual a la media de la población de la que se tomó la muestra.

Esto enseña que usando el muestreo aleatorio simple, el valor esperado o media de la distribución muestral de \bar{x} es igual a la media de la población. En la sección 7.1 se vio que el sueldo anual medio de los administradores de EAI es $\mu = \$51\,800$. Por tanto, de acuerdo con la ecuación (7.1), la media de todas las medias muestrales en el estudio de EAI es también \$51 800.

Cuando el valor esperado de un estimador puntual es igual al parámetro poblacional, se dice que el estimador puntual es **insesgado**. Por tanto, la ecuación (7.1) muestra que \bar{x} es un estimador insesgado de la media poblacional μ .

Desviación estándar de \bar{x}

Ahora se definirá la desviación estándar de la distribución muestral de \bar{x} . Se empleará la notación siguiente.

$\sigma_{\bar{x}}$ = desviación estándar de \bar{x}

σ = desviación estándar de la población

n = tamaño de la muestra

N = tamaño de la población

Es posible demostrar que usando el muestreo aleatorio simple, la desviación estándar de \bar{x} depende de si la población es finita o infinita. Las dos fórmulas para la desviación estándar son las siguientes.

DESVIACIÓN ESTÁNDAR DE \bar{x}

<i>Población finita</i>	<i>Población infinita</i>
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \left(\frac{\sigma}{\sqrt{n}} \right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

(7.2)

Al comparar las dos fórmulas se ve que el factor $\sqrt{(N-n)/(N-1)}$ se requiere cuando la población es finita, pero no cuando es infinita. A este factor se le conoce como **factor de corrección para una población finita**. En muchas de las situaciones prácticas de muestreo, se encuentra que aunque la población sea finita, es “grande”, mientras que el tamaño de la muestra es “pequeño”. En estos casos el factor de corrección para una población finita $\sqrt{(N-n)/(N-1)}$ es casi igual a 1. Por tanto, la diferencia entre el valor de la desviación estándar de \bar{x} en el caso de poblaciones finita o infinitas se vuelve despreciable. Entonces, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ es una buena aproximación a la desviación estándar de \bar{x} , aun cuando la población sea finita. Esta observación lleva al siguiente lineamiento, o regla general, para calcular la desviación estándar de \bar{x} .

USO DE LA EXPRESIÓN SIGUIENTE PARA CALCULAR LA DESVIACIÓN ESTÁNDAR DE \bar{x}

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{siempre que} \quad \text{color: blue; font-weight: bold;">(7.3)$$

siempre que

1. La población sea infinita; o
2. La población sea finita y el tamaño de la muestra sea menor o igual a 5% del tamaño de la población; es decir, $n/N \leq 0.05$.

En el problema 21 se muestra que cuando $n/N \leq 0.05$, el factor de corrección para una población finita tiene poco efecto en el valor de $\sigma_{\bar{x}}$.

En los casos en que $n/N > 0.05$, para calcular $\sigma_{\bar{x}}$ deberá usarse la versión para poblaciones finitas de la fórmula (7.2). En este libro, a menos que se indique otra cosa, se supondrá que el tamaño de la población es “grande”, $n/N \leq 0.05$, y se usará la expresión (7.3) para calcular $\sigma_{\bar{x}}$.

El término error estándar se usa en la inferencia estadística para referirse a la desviación estándar de un estimador puntual.

Para calcular $\sigma_{\bar{x}}$ se necesita conocer σ , la desviación estándar de la población. Para subrayar, aún más, la diferencia entre $\sigma_{\bar{x}}$ y σ , a la desviación estándar de \bar{x} , $\sigma_{\bar{x}}$ se le llama **error estándar** de la media. En general, el término *error estándar* se refiere a la desviación estándar de un estimador puntual. Más adelante se verá que el valor del error estándar de la media ayuda a determinar qué tan lejos puede estar la media muestral de la media poblacional. Ahora, de nuevo con el ejemplo de EAI se calcula el error estándar de la media correspondiente a las muestras aleatorias simples de 30 administradores de EAI.

En la sección 7.1 se halló que la desviación estándar de los sueldos anuales en la población de los 2500 administradores de EAI era $\sigma = 4000$. En este caso la población es finita, $N = 2500$. Sin embargo, como el tamaño de la muestra es 30, se tiene $n/N = 30/2500 = 0.012$. Como el tamaño de la muestra es menor que 5% del tamaño de la población, se puede ignorar el factor de corrección para una población finita y usar la ecuación (7.3) para calcular el error estándar.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{30}} = 730.3$$

Forma de la distribución muestral de \bar{x}

Los resultados anteriores respecto al valor esperado y a la desviación estándar en la distribución muestral de \bar{x} son aplicables a cualquier población. El paso final en la identificación de las características de la distribución muestral de \bar{x} es determinar la forma de la distribución muestral. Se considerarán dos casos: 1. La población tiene distribución normal, y 2. La población no tiene distribución normal.

La población tiene distribución normal. En muchas situaciones es razonable suponer que la población de la que se seleccionó la muestra aleatoria simple tenga distribución normal o casi normal. Cuando la población tiene distribución normal, la distribución muestral de \bar{x} está distribuida normalmente sea cual sea el tamaño de la muestra.

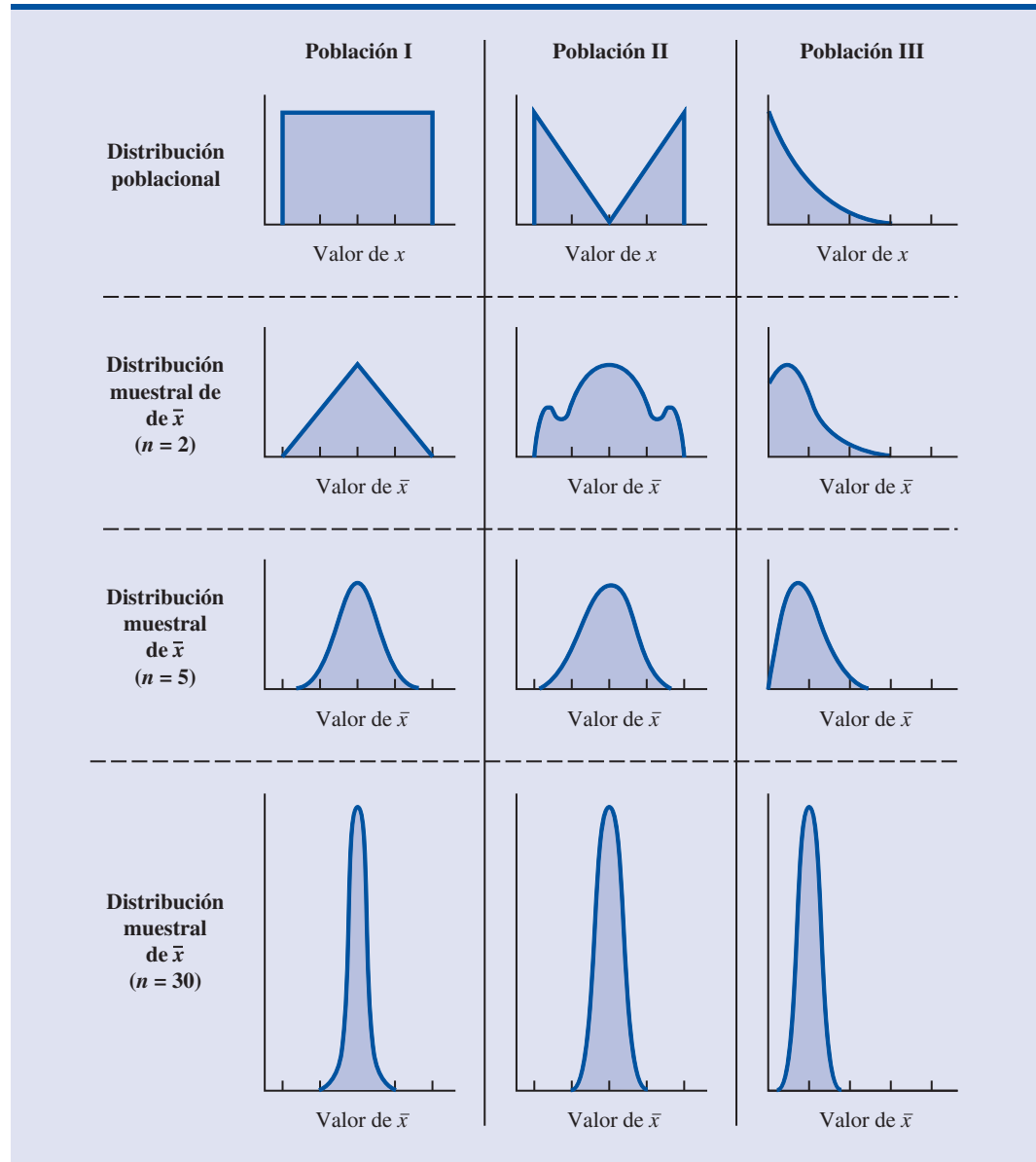
La población no tiene distribución normal. Cuando la población de la que se tomó la muestra aleatoria simple no tiene distribución normal, el **teorema del límite central** ayuda a determinar la forma de la distribución muestral de \bar{x} . El enunciado del teorema del límite central aplicado a la distribución muestral de \bar{x} dice lo siguiente.

TEOREMA DEL LÍMITE CENTRAL

Cuando se seleccionan muestras aleatorias simples de tamaño n de una población, la distribución muestral de la media muestral \bar{x} puede aproximarse mediante una distribución normal a medida que el tamaño de la muestra se hace grande.

En la figura 7.3 se muestra cómo funciona el teorema del límite central en tres poblaciones diferentes; cada columna se refiere a una de las poblaciones. En el primer renglón de la figura se muestra que ninguna de las tres poblaciones está distribuida normalmente. La población I tiene una distribución uniforme. A la población II se le conoce como distribución en forma de orejas de conejo. Esta distribución es simétrica, pero los valores más probables se encuentran en las colas de la distribución. La forma de la población III se parece a una distribución exponencial; es sesgada a la derecha.

En los tres renglones siguientes de la figura 7.3 se muestran las formas de las distribuciones muestrales para tamaños de muestras $n = 2$, $n = 5$ y $n = 30$. Cuando el tamaño de la muestra es 2, se observa que cada distribución muestral tiene una forma diferente a la distribución poblacional correspondiente. Con muestras de tamaño 5, se observa que las formas de las distribuciones muestrales en los casos de las poblaciones I y II empiezan a parecerse a la forma de una distribución normal. En el caso de la población III, aun cuando la forma de la distribución muestral empieza a ser parecida a una distribución normal, todavía se observa cierto sesgo a la derecha.

FIGURA 7.3 ILUSTRACIÓN DEL TEOREMA DEL LÍMITE CENTRAL CON TRES POBLACIONES

Por último, para muestras de tamaño 30, las formas de cada una de las tres distribuciones muestrales es aproximadamente normal.

Desde el punto de vista de la práctica, será conveniente saber qué tan grande necesita ser el tamaño de la muestra para que aplique el teorema del límite central y pueda suponer que la forma de la distribución muestral es aproximadamente normal. En las investigaciones estadísticas se ha estudiado este problema en distribuciones muestrales de \bar{x} de muy diversas poblaciones y para muy diversos tamaños de muestras. Lo que se acostumbra hacer en la práctica es suponer que, en la mayor parte de las aplicaciones, la distribución muestral de \bar{x} se puede aproximar mediante una distribución normal siempre que la muestra sea de tamaño 30 o mayor. En los casos en que la población es muy sesgada o existen observaciones atípicas, pueden necesitarse muestras de tamaño 50. Por último, si la población es discreta, el tamaño de muestra necesario para la aproximación normal suele depender de la proporción poblacional. Más acerca de este tema se dirá cuando se estudie la distribución muestral de \bar{p} en la sección 7.6.

Distribución muestral de \bar{x} en el problema EAI

En el problema de EAI, para el que ya previamente se mostró que $E(\bar{x}) = \$51\,800$ y $\sigma_{\bar{x}} = 730.3$, no se cuenta con ninguna información acerca de la distribución de la población; puede estar o no distribuida normalmente. Si la población tiene una distribución normal, la distribución muestral de \bar{x} estará distribuida normalmente. Si la población no tiene una distribución normal, la muestra aleatoria simple de 30 administradores y el teorema del límite central permiten concluir que la distribución muestral de \bar{x} puede aproximarse mediante una distribución normal. En cualquiera de los casos, se concluye que la distribución muestral de \bar{x} se describe mediante una distribución normal como la que se muestra en la figura 7.4.

Valor práctico de la distribución muestral de \bar{x}

Siempre que se seleccione una muestra aleatoria simple y se use el valor de la media muestral para estimar el valor de la media poblacional μ , no se podrá esperar que la media muestral sea exactamente igual a la media poblacional. La razón práctica por la que interesa la distribución muestral de \bar{x} es que se puede usar para proporcionar información probabilística acerca de la diferencia entre la media muestral y la media poblacional. Para demostrar este uso, se retomará el problema de EAI.

Suponga que el director de personal cree que la media muestral será una estimación aceptable de la media poblacional si la primera está a más o menos de \$500 de la media poblacional. Sin embargo, no es posible garantizar que la media muestral esté a más o menos de \$500 de la media poblacional. En efecto, en la tabla 7.5 y en la figura 7.1 se observa que algunas de las 500 medias muestrales difieren en más de \$2000 de la media poblacional. Entonces hay que pensar en el requerimiento del director de personal en términos de probabilidad. Es decir, al director de personal le interesa la interrogante siguiente: “¿Cuál es la probabilidad de que la media muestral obtenida usando una muestra aleatoria simple de 30 administradores de EAI, se encuentre a más o menos de \$500 de la media poblacional?”

Como ya se han identificado las propiedades de la distribución muestral de \bar{x} (véase figura 7.4), se usará esta distribución para contestar esta interrogante probabilística. Observe la distribución muestral de \bar{x} que se muestra nuevamente en la figura 7.5. Como la media poblacional es \$51 800, el director de personal desea saber cuál es la probabilidad de que \bar{x} esté entre \$51 300 y \$52 300. Esta probabilidad corresponde al área sombreada de la distribución muestral que apa-

FIGURA 7.4 DISTRIBUCIÓN MUESTRAL DE \bar{x} PARA EL SUELDO ANUAL EN UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES

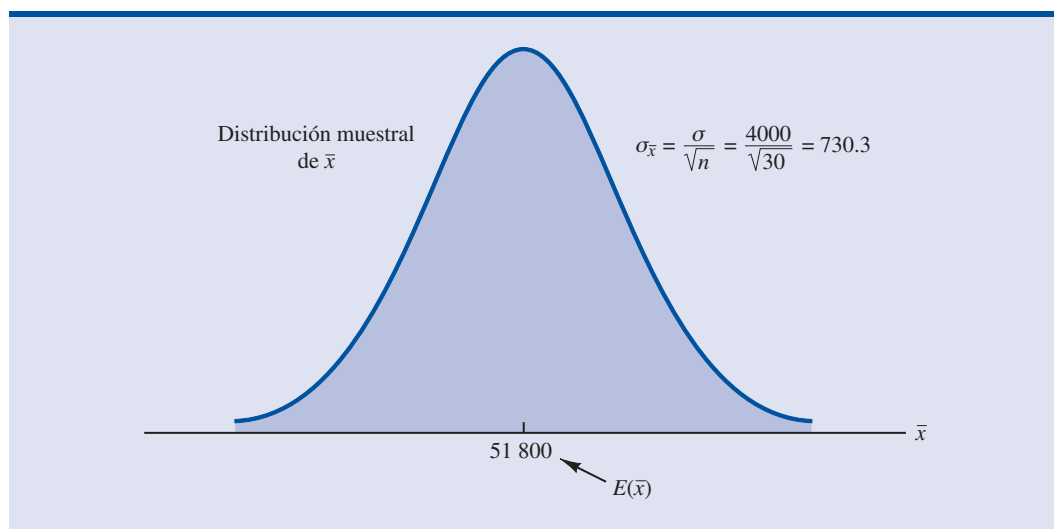
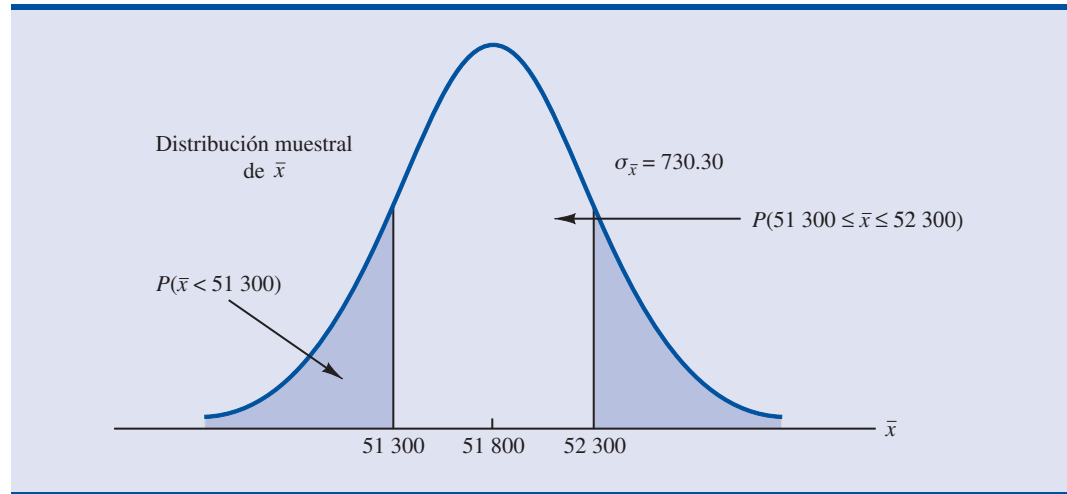


FIGURA 7.5 PROBABILIDAD DE QUE UNA MEDIA MUESTRAL DE UNA MUESTRA ALEATORIA SIMPLE DE 30 ADMINISTRADORES DE EAI SE ENCUENTRE ENTRE LOS \$500 DE LA MEDIA POBLACIONAL



rece en la figura 7.5. Como la distribución muestral está distribuida normalmente y su media es \$51 800 y el error estándar de la media es 730.3, se usa la tabla de probabilidad normal estándar para hallar el área o probabilidad.

Primero se calcula el valor de z en el extremo superior de este intervalo (52 300) y se usa la tabla para hallar el área bajo la curva a la izquierda de ese punto (área hacia la cola izquierda). Después se calcula el valor de z en el extremo inferior de este intervalo (51 300) y se usa la tabla para hallar el área bajo la curva a la izquierda de este punto (otra área hacia la cola izquierda). Al restar la segunda área de la primera, se obtiene la probabilidad buscada.

En $\bar{x} = 52\,300$, se tiene

$$z = \frac{52\,300 - 51\,800}{730.30} = 0.68$$

En la tabla de probabilidad normal estándar la probabilidad acumulada (área a la izquierda de $z = 0.68$) es 0.7517.

En $\bar{x} = 51\,300$, se tiene

$$z = \frac{51\,300 - 51\,800}{730.30} = -0.68$$

El área bajo la curva a la izquierda de $z = -0.68$ es 0.2483. Por tanto, $P(51\,300 \leq \bar{x} \leq 52\,300) = P(z \leq 0.68) - P(z < -0.68) = 0.7517 - 0.2483 = 0.5034$.

Estos cálculos indican que hay una probabilidad de 0.5034 de que con una muestra aleatoria simple de 30 administradores de EAI se obtenga una media muestral \bar{x} que esté a más o menos de \$500 de la media poblacional. Por tanto, la probabilidad de que la diferencia entre \bar{x} y $\mu = \$51,800$ sea superior a \$500 es $1 - 0.5034 = 0.4966$. En otras palabras, una muestra aleatoria simple de 30 administradores de EAI tiene aproximadamente 50/50 oportunidades de tener una media muestral que no difiera de la media poblacional en más de los aceptables \$500. Quizá deba pensarse en una muestra de tamaño mayor. Se explorará esta posibilidad considerando la relación entre el tamaño de la muestra y la distribución muestral de \bar{x} .

La distribución muestral de \bar{x} se usa para obtener información probabilística acerca de qué tan cerca se encuentra la media muestral \bar{x} de la media poblacional μ .

Relación entre el tamaño de la muestra y la distribución muestral de \bar{x}

Suponga que en el problema de muestreo de EAI se toma una muestra aleatoria simple de 100 administradores en lugar de los 30 considerados. La intuición indica que teniendo más datos proporcionados por una muestra mayor, la media muestral basada en $n = 100$ proporcionará una mejor estimación de la media poblacional que una media muestral basada en $n = 30$. Para ver cuánto es mejor, se considerará la relación entre el tamaño de la muestra y la distribución muestral de \bar{x} .

Primero observe que $E(\bar{x}) = \mu$ independientemente del tamaño de la muestra. Entonces, la media de todos los valores posibles de \bar{x} es igual a la media poblacional μ independientemente del tamaño n de la muestra. Pero, el error estándar de la media, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, está relacionado con la raíz cuadrada del tamaño de la muestra. Siempre que el tamaño de la muestra aumente, el error estándar de la media $\sigma_{\bar{x}}$ disminuirá. Con $n = 30$, el error estándar de la media en el problema de EAI es 730.3. Sin embargo, aumentando el tamaño de la muestra $n = 100$, el error estándar de la media disminuye a

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4000}{\sqrt{100}} = 400$$

En la figura 7.6 se muestran las distribuciones muestrales de \bar{x} correspondientes a $n = 30$ y a $n = 100$. Como la distribución muestral con $n = 100$ tiene un error estándar más pequeño, habrá menos variación entre los valores de \bar{x} y éstos tenderán a estar más cerca de la media poblacional que los valores de \bar{x} con $n = 30$.

La distribución muestral de \bar{x} , en el caso $n = 100$, puede emplearse para calcular la probabilidad de que una muestra aleatoria simple de 100 administradores de EAI dé una media muestral que no difiera de los \$500 de la media poblacional. Como la distribución muestral es normal y su media es \$51 800 y el error estándar de la media es 400, se emplea la tabla de probabilidad normal estándar para hallar el área o la probabilidad.

Para $\bar{x} = 52\,300$ (véase figura 7.7) se tiene

$$z = \frac{52\,300 - 51\,800}{400} = 1.25$$

FIGURA 7.6 COMPARACIÓN ENTRE LAS DISTRIBUCIONES MUESTRALES DE \bar{x} CON MUESTRAS ALEATORIAS SIMPLES DE TAMAÑO $n = 30$ ADMINISTRADORES DE EAI Y CON MUESTRAS DE TAMAÑO $n = 100$ ADMINISTRADORES

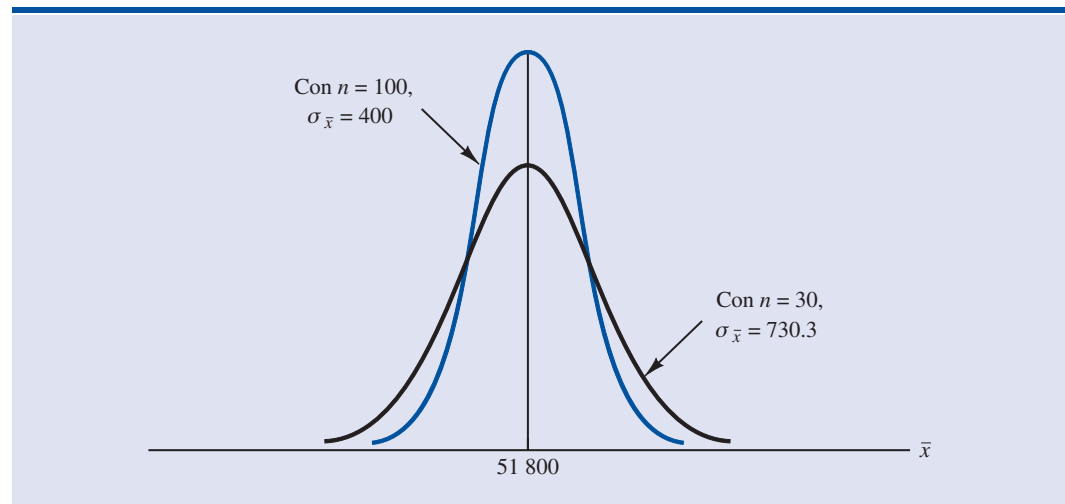
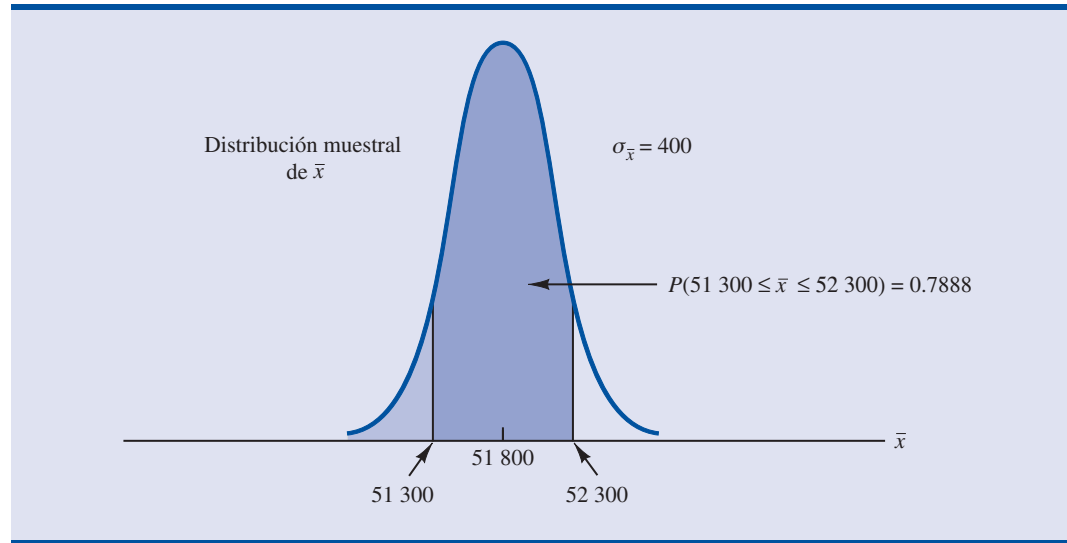


FIGURA 7.7 PROBABILIDAD DE QUE LA MEDIA MUESTRAL NO DIFIERA EN MÁS DE \$500 DE LA MEDIA POBLACIONAL USANDO UNA MUESTRA ALEATORIA SIMPLE DE 100 ADMINISTRADORES DE EAI



En la tabla de probabilidad normal estándar se encuentra que la probabilidad acumulada correspondiente a $z = 1.25$ es 0.8944.

Para $\bar{x} = 51\,300$, se tiene

$$z = \frac{51\,300 - 51\,800}{400} = -1.25$$

La probabilidad acumulada correspondiente a $z = -1.25$ es 0.1056. Por tanto, $P(51\,300 \leq \bar{x} \leq 52\,300) = P(z \leq 1.25) - P(z \leq -1.25) = 0.8944 - 0.1056 = 0.7888$. Entonces, aumentando el tamaño de la muestra de 30 a 100 administradores de EAI, la probabilidad de obtener una muestra aleatoria simple que esté entre los \$500 de la media poblacional aumenta de 0.5034 a 0.7888.

Aquí, el punto importante es que cuando aumenta el tamaño de la muestra, el error estándar de la media disminuye. Así, una muestra de mayor tamaño proporciona mayor probabilidad de que la media muestral esté dentro de una distancia determinada de la media poblacional.

NOTAS Y COMENTARIOS

1. Al presentar la distribución muestral de \bar{x} para el problema de EAI, se aprovechó que se conocían la media poblacional $\mu = 51\,800$ y la desviación estándar poblacional $\sigma = 4000$. Sin embargo, lo usual es que los valores de la media poblacional μ y de la desviación estándar poblacional σ , que se necesitan para determinar la distribución muestral de \bar{x} , no se conozcan. En el capítulo 8 se verá cómo se usan la media muestral \bar{x} y la desviación estándar muestral s cuando no se conocen μ y σ .
2. La demostración del teorema del límite central requiere observaciones independientes en la muestra. Esta condición se satisface cuando se trata de poblaciones infinitas y cuando se trata de poblaciones finitas, si el muestreo se hace con reemplazo. Aunque el teorema del límite central no se refiere directamente a muestreos sin reemplazo de poblaciones finitas, se aplican los hallazgos del teorema del límite central cuando la población es de tamaño grande.

Ejercicios

Métodos

18. La media de una población es 200 y su desviación estándar es 50. Se va a tomar una muestra aleatoria simple de tamaño 100 y se usará la media muestral para estimar la media poblacional.
 - a. ¿Cuál es el valor esperado de \bar{x} ?
 - b. ¿Cuál es la desviación estándar de \bar{x} ?
 - c. Muestre la distribución muestral de \bar{x} .
 - d. ¿Qué muestra la distribución muestral de \bar{x} ?
19. La media de una población es 200 y su desviación estándar es 50. Suponga que se selecciona una variable aleatoria simple de tamaño 100 y se usa \bar{x} para estimar μ .
 - a. ¿Cuál es la probabilidad de que la diferencia entre la media muestral y la media poblacional no sea mayor que ± 5 ?
 - b. ¿De qué la diferencia entre la media muestral y la media poblacional no sea mayor que ± 10 ?
20. Suponga que la desviación estándar poblacional es $\sigma = 25$. Calcule el error estándar de la media, $\sigma_{\bar{x}}$, con muestras de tamaño 50, 100, 150 y 200. ¿Qué puede decir acerca del error estándar de la media conforme el tamaño de la muestra aumenta?
21. Suponga que de una población en la que $\sigma = 10$ se toma una muestra aleatoria simple de tamaño 50. Halle el valor del error estándar de la media en cada uno de los casos siguientes (si es necesario use el factor de corrección para una población finita).
 - a. El tamaño de la población es infinito.
 - b. El tamaño de la población es $N = 50\,000$.
 - c. El tamaño de la población es $N = 5\,000$.
 - d. El tamaño de la población es $N = 500$.

Aplicaciones

22. Regrese al problema de los administradores de EAI. Suponga que se usa una muestra aleatoria simple de 60 administradores.
 - a. Dibuje la distribución muestral de \bar{x} si se emplean muestras aleatorias simples de tamaño 60.
 - b. ¿Qué pasa con la distribución muestral de \bar{x} si se usan muestras aleatorias simples de tamaño 120?
 - c. ¿Qué puede decir acerca de lo que le pasa a la distribución muestral de \bar{x} conforme el tamaño de la muestra aumenta? ¿Parece ser lógica esta generalización? Explique.
23. En el problema de EAI (véase figura 7.5), se mostró que con $n = 30$, la probabilidad de que la media muestral no difiriera más de \$500 de la media poblacional era 0.5034.
 - a. ¿Cuál es la probabilidad de que la media muestral no difiera más de \$500 de la media poblacional si se usa una muestra de tamaño 60?
 - b. Responda el inciso a si el tamaño de la muestra es 120.
24. El costo medio de la colegiatura en una universidad estatal de Estados Unidos es \$4260 anuales. Considere este valor como media poblacional y asuma que la desviación estándar poblacional es $\sigma = \$900$. Suponga que selecciona una muestra aleatoria de 50 universidades.
 - a. Presente la distribución muestral de \bar{x} como media muestral de la colegiatura en las 50 universidades.
 - b. ¿Cuál es la probabilidad de que la muestra aleatoria simple proporcione una media muestral que no difiera de la media poblacional en más de \$250?
 - c. ¿Cuál es la probabilidad de que la muestra aleatoria simple proporcione una media muestral que no difiera de la media poblacional en más de \$100?
25. El College Board American College Testing Program informa que en el examen de admisión a las universidades, a nivel nacional, la media poblacional de las puntuaciones que se obtienen es $\mu = 1020$ (*The World Almanac 2003*). Suponga que la desviación estándar poblacional es $\sigma = 100$.

Autoexamen

Autoexamen

- a. ¿Cuál es la probabilidad de que en una muestra aleatoria de 75 estudiantes la media muestral de las puntuaciones no difiera más de 10 puntos de la media poblacional?
 - b. ¿Cuál es la probabilidad de que en una muestra aleatoria de 75 estudiantes la media muestral de las puntuaciones no difiera más de 20 puntos de la media poblacional?
26. El costo medio anual de un seguro para automóvil es de \$939 (*CNBC*, 23 de febrero de 2006). Suponga que la desviación estándar es $\sigma = \$245$.
- a. ¿Cuál es la probabilidad de que una muestra aleatoria simple de pólizas de seguros de automóvil la media muestral no difiera más de \$25 de la media poblacional si el tamaño de la muestra es 30, 50, 100 y 400?
 - b. ¿Qué ventaja tiene una muestra grande cuando se quiere estimar la media poblacional?
27. *BusinessWeek* realizó una encuesta entre los estudiantes que terminaban sus estudios en los 30 programas de una maestría (*BusinessWeek*, 22 de septiembre de 2003). De acuerdo con esta encuesta el salario medio anual de una mujer y de un hombre 10 años después de terminar sus estudios es \$117 000 y \$168 000, respectivamente. Suponga que la desviación estándar entre los salarios de las mujeres es \$25 000 y entre los salarios de los hombres es \$40 000.
- a. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 40 hombres la media muestral no difiera más de \$10 000 de la media poblacional de \$168 000?
 - b. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 40 mujeres la media muestral no difiera más de \$10 000 de la media poblacional de \$117 000?
 - c. ¿En cuál de los dos casos, inciso a o inciso b, hay más probabilidad de obtener una media muestral que no difiera en más de \$10 000 de la media poblacional? ¿Por qué?
 - d. ¿Cuál es la probabilidad de que en una muestra aleatoria simple de 100 hombres, la media muestral no difiera en más de \$4000 de la media poblacional?
28. Un hombre golfista tiene una puntuación promedio de 95 y una mujer de 106 (*Golf Digest*, abril de 2006). Considere estos valores como medias poblacionales de los hombres y de las mujeres y suponga que la desviación estándar poblacional es $\sigma = 14$ golpes en ambos casos. Se tomará una muestra aleatoria simple de 40 golfistas hombres y otra de 45 mujeres golfistas.
- a. Dé la distribución muestral de \bar{x} correspondiente a los hombres golfistas.
 - b. ¿Cuál es la probabilidad de que, en el caso de los hombres golfistas, la media muestral no difiera en más de 3 golpes de la media poblacional?
 - c. ¿Cuál es la probabilidad de que, en el caso de las mujeres golfistas, la media muestral no difiera en más de 3 golpes de la media poblacional?
 - d. ¿En cuál de los casos, inciso a o inciso b, es mayor la probabilidad de que la media muestral no difiera en más de 3 golpes de la media poblacional? ¿Por qué?
29. En el norte de Kentucky (*The Cincinnati Enquirer*, 21 de enero de 2006) el precio promedio de la gasolina sin plomo era \$2.34. Use este precio como media poblacional y suponga que la desviación estándar poblacional es \$0.20.
- a. ¿Cuál es la probabilidad de que el precio medio en una muestra de 30 gasolineras no difiera en más de \$0.30 de la media poblacional?
 - b. ¿Cuál es la probabilidad de que el precio medio en una muestra de 50 gasolineras no difiera en más de \$0.30 de la media poblacional?
 - c. ¿Cuál es la probabilidad de que el precio medio en una muestra de 100 gasolineras no difiera en más de \$0.30 de la media poblacional?
 - d. ¿Recomendaría usted alguno de los tamaños muestrales de los incisos a, b o c para que la probabilidad de que el precio muestral no difiriera en más de \$0.30 de la media muestral fuera \$0.95?
30. Para estimar la edad media de una población de 4000 empleados se toma una muestra de 40 empleados.
- a. ¿Usted usaría el factor de corrección para una población finita en el cálculo del error estándar de la media? Explique.

- b. Si la desviación estándar poblacional es $\sigma = 8.2$ años, calcule el error estándar con y sin el factor de corrección para una población finita. ¿Cuál es la base para ignorar el factor de corrección para la población finita, si $n/N \leq 0.05$?
- c. ¿Cuál es la probabilidad de que la media muestral de las edades de los empleados no difiera en más de ± 2 años de la media poblacional de las edades?

7.6

Distribución muestral de \bar{p}

La proporción muestral \bar{p} es el estimador puntual de la proporción poblacional p . La fórmula para calcular la proporción muestral es

$$\bar{p} = \frac{x}{n}$$

donde

x = número de elementos de la muestra que poseen la característica de interés

n = tamaño de la muestra

Como se indicó en la sección 7.4, la proporción muestral \bar{p} es una variable aleatoria y su distribución de probabilidad se conoce como distribución muestral de \bar{p} .

DISTRIBUCIÓN MUESTRAL DE \bar{p}

La distribución muestral de \bar{p} es la distribución de probabilidad de todos los posibles valores de la proporción muestral \bar{p} .

Para determinar qué tan cerca está la proporción muestral \bar{p} de la proporción poblacional p , se necesita entender las propiedades de la distribución muestral de \bar{p} : el valor esperado de \bar{p} , la desviación estándar de \bar{p} y la forma de la distribución muestral de \bar{p} .

Valor esperado de \bar{p}

El valor esperado de \bar{p} , la media de todos los posibles valores de \bar{p} , es igual a la proporción poblacional p .

VALOR ESPERADO DE \bar{p}

$$E(\bar{p}) = p \quad (7.4)$$

donde

$$E(\bar{p}) = \text{valor esperado de } \bar{p}$$

$$p = \text{proporción poblacional}$$

Como $E(\bar{p}) = p$, \bar{p} es un estimador insesgado de p . Recuerde que en la sección 7.1 se encontró que en la población de EAI $p = 0.60$, siendo p la proporción de la población de administradores que han participado en el programa de capacitación de la empresa. Por tanto, el valor esperado de \bar{p} en el problema de muestreo de EAI es 0.60.

Desviación estándar de \bar{p}

Como en el caso de la desviación estándar de \bar{x} la desviación estándar de \bar{p} obedece a si la población es finita o infinita. Las dos fórmulas para calcular la desviación estándar de \bar{p} se presentan a continuación.

DESVIACIÓN ESTÁNDAR DE \bar{p}

<i>Población finita</i>	<i>Población infinita</i>	
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$	(7.5)

Al comparar las dos fórmulas (7.5) se aprecia que la única diferencia es el uso del factor de corrección para una población finita $\sqrt{(N-n)/(N-1)}$.

Como en el caso de la media poblacional \bar{x} , la diferencia entre las expresiones para una población finita y para una infinita es despreciable si el tamaño de la población finita es grande en comparación con el tamaño de la muestra. Se seguirá la misma regla recomendada para la media poblacional. Es decir, si la población es finita y $n/N \leq 0.05$ se usará $\sigma_{\bar{p}} = \sqrt{p(1-p)/n}$. Pero, si la población es finita y $n/N > 0.05$, entonces deberá usar el factor de corrección para una población finita. También, a menos que se especifique otra cosa, en este libro se supondrá que el tamaño de la población es grande en comparación al tamaño de la muestra y por tanto, el factor de corrección para una población finita no será necesario.

En la sección 7.5 se usó el error estándar de la media para referirse a la desviación estándar de \bar{x} . Se dijo que en general el término error estándar se refiere a la desviación estándar de un estimador puntual. Así, en el caso de proporciones se usa *el error estándar de la proporción* para referirse a la desviación estándar de \bar{p} . Ahora se vuelve al ejemplo de EAI para calcular el error estándar de la proporción en la muestra aleatoria simple de los 30 administradores de EAI.

En el estudio de EAI se sabe que la proporción poblacional de administradores que han participado en el programa de capacitación es $p = 0.60$. Como $n/N = 30/2\,500 = 0.012$ se puede ignorar el factor de corrección para una población finita al calcular el error estándar de la proporción. En la muestra aleatoria simple de 30 administradores, $\sigma_{\bar{p}}$ es

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.60(1-0.60)}{30}} = 0.0894$$

Forma de la distribución muestral de \bar{p}

Ahora que se conoce la media y la desviación estándar de la distribución muestral de \bar{p} , el último paso es determinar la forma de la distribución muestral. La proporción muestral es $\bar{p} = x/n$. En una muestra aleatoria simple de una población grande, el valor de x es una variable aleatoria binomial que indica el número de los elementos de la muestra que tienen la característica de interés. Como n es una constante, la probabilidad de x/n es la misma que la probabilidad de x , lo cual significa que la distribución muestral de \bar{p} también es una distribución de probabilidad discreta y que la probabilidad de cada x/n es la misma que la probabilidad de x .

En el capítulo 6 se mostró que una distribución binomial se aproxima mediante una distribución normal siempre que el tamaño de la muestra sea lo suficientemente grande para satisfacer las dos condiciones siguientes.

$$np \geq 5 \quad \text{y} \quad n(1 - p) \geq 5$$

Suponiendo que se satisfagan estas dos condiciones, la distribución de probabilidad de x en la proporción muestral, $\bar{p} = x/n$, puede aproximarse por medio de una distribución normal. Y como n es una constante, la distribución muestral de \bar{p} también se aproxima mediante una distribución normal. Esta aproximación se formula como sigue:

La distribución muestral de \bar{p} se aproxima mediante una distribución normal siempre que $np \geq 5$ y $n(1 - p) \geq 5$.

En las aplicaciones prácticas, cuando se requiere una estimación de la proporción poblacional, casi siempre se encuentra que el tamaño de la muestra es suficientemente grande para poder usar la aproximación normal para la distribución muestral de \bar{p} .

Recuerde que en el problema de muestreo de EAI la proporción poblacional de administradores que han participado en el programa de capacitación es $p = 0.60$. Con una muestra aleatoria simple de tamaño 30, se tiene $np = 30(0.60) = 18$ y $n(1 - p) = 30(0.40) = 12$. Por tanto, la distribución muestral de \bar{p} se calcula mediante la distribución normal que se muestra en la figura 7.8.

Valor práctico de la distribución muestral de \bar{p}

El valor práctico de la distribución muestral de \bar{p} es que permite obtener información probabilística acerca de la diferencia entre la proporción muestral y la proporción poblacional. Por ejemplo, en el problema de EAI, el director de personal desea saber cuál es la probabilidad de obtener un valor de \bar{p} que no difiera en más de 0.05 de la proporción poblacional de los administradores de EAI que han participado en el programa de capacitación. Es decir, ¿cuál es la probabilidad de tener una muestra en la que \bar{p} esté entre 0.55 y 0.65? El área sombreada de la figura 7.9 corres-

FIGURA 7.8 DISTRIBUCIÓN MUESTRAL DE \bar{p} , PROPORCIÓN DE ADMINISTRADORES QUE HAN PARTICIPADO EN EL PROGRAMA DE CAPACITACIÓN DE EAI

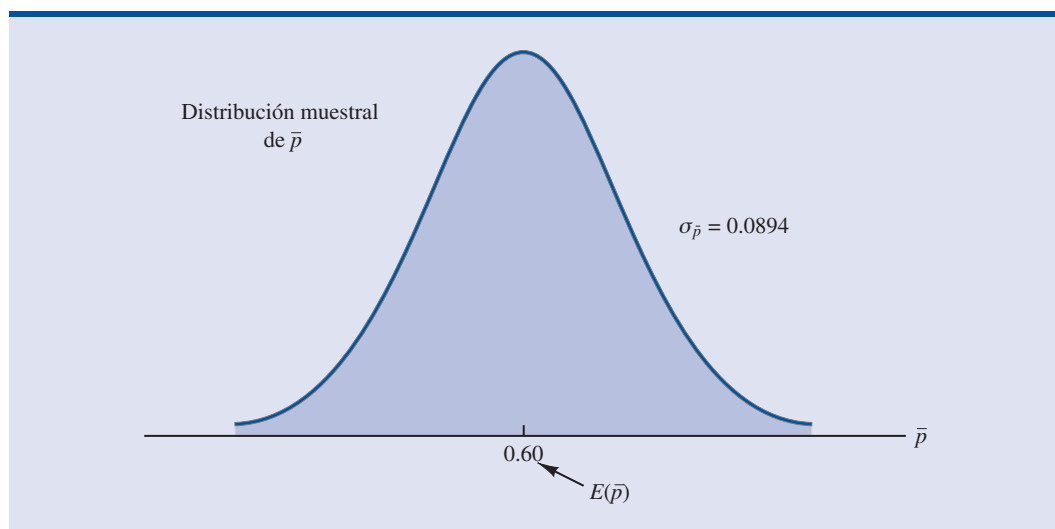
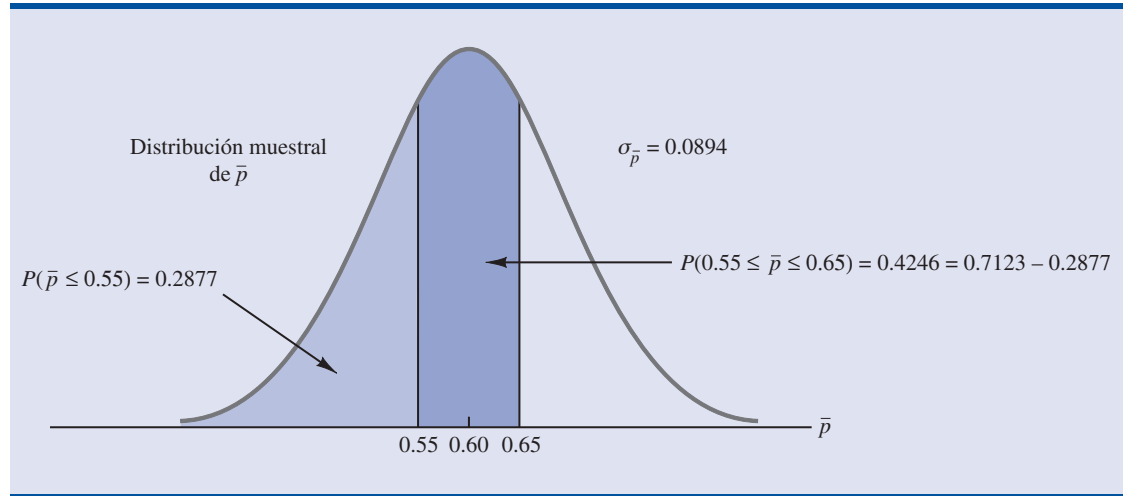


FIGURA 7.9 PROBABILIDAD DE QUE \bar{p} ESTÉ ENTRE 0.55 Y 0.65

ponde a esta probabilidad. A partir de que la distribución muestral de \bar{p} se aproxima mediante una distribución normal con media 0.60 y error estándar de la proporción $\sigma_{\bar{p}} = 0.0894$, se encuentra que la variable aleatoria normal estándar correspondiente a $\bar{p} = 0.65$ tiene el valor $z = (0.65 - 0.60)/0.0894 = 0.56$. En la tabla de probabilidad normal estándar aparece que la probabilidad acumulada que corresponde a $z = 0.56$ es 0.7123. De manera similar para $\bar{p} = 0.55$, se encuentra que $z = (0.55 - 0.60)/0.0894 = -0.56$. En la misma tabla y correspondiente a $z = -0.56$ es 0.2877. De esta manera, la probabilidad de seleccionar una muestra en la cual el valor de \bar{p} no difiera más de 0.05 de la proporción poblacional p está dada por $0.7123 - 0.2877 = 0.4246$.

Si se aumenta el tamaño de la muestra a $n = 100$, el error estándar de la proporción se convierte en

$$\sigma_{\bar{p}} = \sqrt{\frac{0.60(1 - 0.60)}{100}} = 0.049$$

Con una muestra de 100 administradores de EAI, se calcula ahora la probabilidad de que la proporción muestral tenga un valor que no difiera en más de 0.05 de la proporción poblacional. Como la distribución muestral es aproximadamente normal, con media 0.60 y desviación estándar 0.049, se puede usar la tabla de probabilidad normal estándar para hallar el área o probabilidad. Para $\bar{p} = 0.65$, se tiene $z = (0.65 - 0.60)/0.049 = 1.02$. La tabla de probabilidad normal estándar arroja que la probabilidad acumulada correspondiente a $z = 1.02$ es 0.8461. De manera similar, para $\bar{p} = 0.55$, se tiene que $z = (0.55 - 0.60)/0.049 = -1.02$. Se encuentra que la probabilidad acumulada correspondiente a $z = -1.02$ es 0.1539. Por tanto, si el tamaño de la muestra aumenta de 30 a 100, la probabilidad de que la proporción muestral \bar{p} no difiera en más de 0.05 de la proporción poblacional aumenta a $0.8461 - 0.1539 = 0.6922$.

Ejercicios

Métodos

31. De una muestra aleatoria de tamaño 100 de una población en la que $p = 0.40$.
 - a. ¿Cuál es el valor esperado de \bar{p} ?
 - b. ¿Cuál es el error estándar de \bar{p} ?

Autoexamen

- c. Exprese la distribución muestral de \bar{p} .
 - d. ¿Qué indica la distribución muestral de \bar{p} ?
32. Una proporción poblacional es 0.40. Se toma una muestra aleatoria de tamaño 200 y la proporción muestral \bar{p} se usa para estimar la proporción poblacional.
 - a. ¿Cuál es la probabilidad de que la proporción muestral esté entre ± 0.03 de la proporción poblacional?
 - b. ¿De que la proporción muestral esté entre ± 0.05 de la proporción poblacional?
 33. Suponga que la proporción poblacional es 0.55. Calcule el error estándar de la proporción, $\sigma_{\bar{p}}$, para los tamaños de muestra 100, 200, 500 y 1000. ¿Qué puede decir acerca del tamaño del error estándar a medida que el tamaño de la muestra aumenta?
 34. La proporción poblacional es 0.30. ¿Cuál es la probabilidad de que las proporciones muestral y poblacional esté entre ± 0.04 con los tamaños de muestra siguientes?
 - a. $n = 100$
 - b. $n = 200$
 - c. $n = 500$
 - d. $n = 1000$
 - e. ¿Qué ventaja tiene un tamaño grande de muestra?

Aplicaciones

Autoexamen

35. El director de una empresa piensa que 30% de los pedidos provienen de nuevos compradores. Para ver la proporción de nuevos compradores se usará una muestra aleatoria simple de 100 pedidos.
 - a. Suponga que el director está en lo cierto y que $p = 0.30$. ¿Cuál es la distribución muestral de \bar{p} en este estudio?
 - b. ¿Cuál es la probabilidad de que la proporción muestral de \bar{p} esté entre 0.20 y 0.40?
 - c. ¿Cuál es la probabilidad que la proporción muestral de \bar{p} esté entre 0.25 y 0.35?
36. *The Cincinnati Enquirer* informa que en Estados Unidos 66% de los adultos y 87% de los jóvenes entre 12 y 17 años usan Internet (*The Cincinnati Enquirer*, 7 de febrero de 2007). Considere estos datos como proporciones poblacionales y suponga que se usará una muestra de 300 adultos y 300 jóvenes para obtener información respecto de su opinión acerca de la seguridad en Internet.
 - a. Muestre la distribución muestral de \bar{p} , siendo \bar{p} la proporción muestral de adultos que usan Internet.
 - b. ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional de adultos que usan Internet no sea mayor que ± 0.04 ?
 - c. ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional de jóvenes que usan Internet no sea mayor que ± 0.04 ?
 - d. ¿Son diferentes las probabilidades del inciso b y del inciso c? Si es así, ¿por qué?
 - e. Responda al inciso b en el caso de que el tamaño de la muestra sea 600. ¿Es menor la probabilidad? ¿Por qué?
37. Los sondeos de *Time/CNN* entre los votantes siguieron la opinión del público respecto de los candidatos presidenciales en las votaciones del 2000. En uno de estos sondeos Yankelovich Partners empleó una muestra de 589 probables votantes (*Time*, 26 de junio de 2000). Suponga que la proporción poblacional a favor de un determinado candidato a la presidencia haya sido $p = 0.50$. Sea \bar{p} la proporción muestral en los posibles votantes que está a favor de ese candidato a la presidencia.
 - a. Muestre la distribución muestral de \bar{p} .
 - b. ¿Cuál es la probabilidad de que los sondeos de *Time/CNN* indiquen que la diferencia entre las proporciones muestral y poblacional en uno de estos sondeos no sea mayor que ± 0.04 ?
 - c. ¿Cuál es la probabilidad de que los sondeos de *Time/CNN* indiquen que la diferencia entre las proporciones muestral y poblacional en uno de estos sondeos no sea mayor que ± 0.03 ?
 - d. ¿Cuál es la probabilidad de que los sondeos de *Time/CNN* indiquen que la diferencia entre las proporciones muestral y poblacional en uno de estos sondeos no sea mayor que ± 0.02 ?

38. Roper ASW realizó una encuesta para obtener información acerca de la opinión de los estadounidenses respecto al dinero y la felicidad (*Money*, octubre de 2003). Cincuenta y seis por ciento de los entrevistados dijo revisar el estado de su bloc de cheques por lo menos una vez al mes.
 - a. Suponga que se toma una muestra de 400 estadounidenses adultos. Indique la distribución muestral de la proporción de adultos que revisan el estado de su bloc de cheques por lo menos una vez al mes.
 - b. ¿Cuál es la probabilidad de que la diferencia entre la proporción muestral y la proporción poblacional no sea mayor que ± 0.02 ?
 - c. ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.04 ?
39. El *Democrat and Chronicle* informa que 25% de los vuelos que llegaron al aeropuerto de San Diego en los primeros cinco meses de 2001, arribaron con retraso (*Democrat and Chronicle*, 23 de julio de 2001). Suponga que la proporción poblacional sea $p = 0.25$.
 - a. Muestre la distribución muestral de \bar{p} , la proporción de vuelos retrasados en una muestra de 1 000 vuelos.
 - b. ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.03 , si el tamaño de la muestra es 1000?
 - c. Responda el inciso b con una muestra de 500 vuelos.
40. The Grocery Manufacturers of America informa que 76% de los consumidores leen los ingredientes que se enumeran en la etiqueta de un producto. Suponga que la proporción poblacional es $p = 0.76$ y que de la población de consumidores se selecciona una muestra de 400 consumidores.
 - a. Exprese la distribución muestral de la proporción muestral \bar{p} , si \bar{p} es la proporción de consumidores de la muestra que lee los ingredientes que se enumeran en la etiqueta.
 - b. ¿Cuál es la probabilidad de que la diferencia entre las proporciones muestral y poblacional no sea mayor que ± 0.03 ?
 - c. Conteste el inciso b si el tamaño de la muestra es 750 consumidores.
41. El Food Marketing Institute informa que 17% de los hogares gastan más de \$100 en productos de abarrotes. Suponga que la proporción poblacional es $p = 0.17$ y que de la población se toma una muestra aleatoria simple de 800 hogares.
 - a. Exprese la distribución muestral de \bar{p} , la proporción muestral de hogares que gastan más de \$100 semanales en abarrotes.
 - b. ¿Cuál es la probabilidad de que la proporción poblacional no difiera en más de 0.02 de la proporción muestral?
 - c. Conteste el inciso b en el caso de que el tamaño de la muestra sea 1600 hogares.

7.7

Propiedades de los estimadores puntuales

En este capítulo se ha mostrado que los estadísticos muestrales, como la media muestral \bar{x} , la desviación estándar muestral s y la proporción muestral \bar{p} sirven como estimadores puntuales de sus correspondientes parámetros poblacionales, μ , σ y p . Resulta interesante que cada uno de estos estadísticos muestrales sean los estimadores puntuales de sus correspondientes parámetros poblacionales. Sin embargo, antes de usar un estadístico muestral como estimador puntual, se verifica si el estimador puntual tiene ciertas propiedades que corresponden a un buen estimador puntual. En esta sección se estudian las propiedades que deben tener los buenos estimadores puntuales: insesgadez, eficiencia y consistencia.

Como hay distintos estadísticos muestrales que se usan como estimadores puntuales de sus correspondientes parámetros poblacionales, en esta sección se usará la notación general siguiente.

θ = el parámetro poblacional de interés

$\hat{\theta}$ = el estadístico muestral o estimador puntual de θ

En esta notación θ es la letra griega theta y la notación $\hat{\theta}$ se lee “theta sombrero”. En general, θ representa cualquier parámetro poblacional como, por ejemplo, la media poblacional, la desvia-

ción estándar poblacional, la proporción poblacional, etc.; $\hat{\theta}$ representa el correspondiente estadístico muestral, por ejemplo, la media muestral, la desviación estándar muestral y la proporción muestral.

Insesgadez

Si el valor esperado del estadístico muestral es igual al parámetro poblacional que se estudia, se dice que el estadístico muestral es un *estimador insesgado* del parámetro poblacional.

INSESGADEZ

El estadístico muestral $\hat{\theta}$ es un estimado insesgado del parámetro poblacional θ si

$$E(\hat{\theta}) = \theta$$

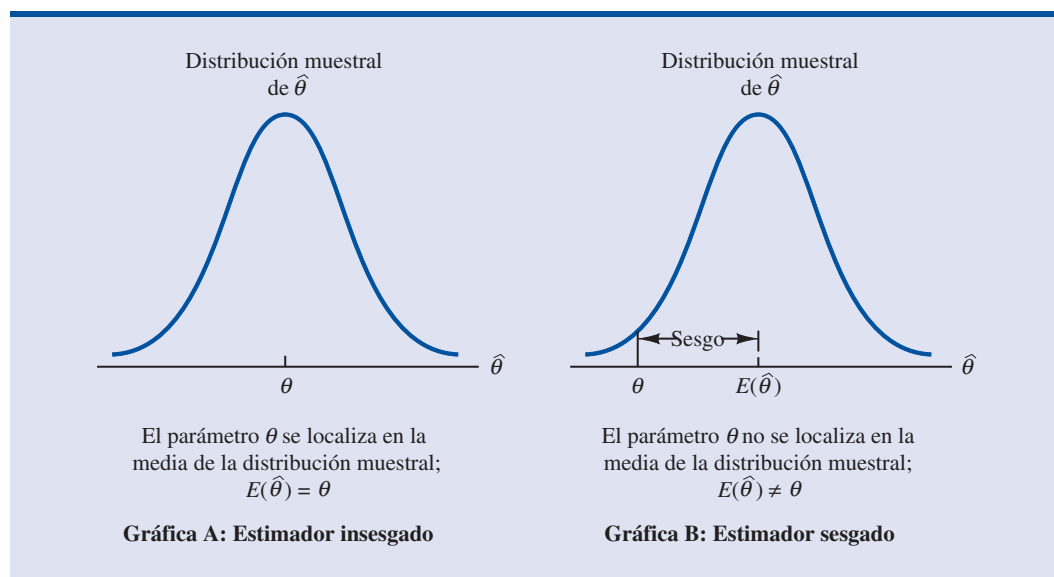
donde

$$E(\hat{\theta}) = \text{valor esperado del estadístico muestral } \hat{\theta}$$

Por tanto, el valor esperado, o media, de todos los posibles valores de un estadístico muestral insesgado es igual al parámetro poblacional que se estudia.

En la figura 7.10 se muestran los casos de los estimadores puntuales sesgado e insesgado. En la figura en que se muestra el estimador insesgado, la media de la distribución muestral es igual al valor del parámetro poblacional. En este caso los errores de estimación se equilibran, ya que algunas veces el valor del estimador puntual $\hat{\theta}$ puede ser menor que θ y otras veces sea mayor que θ . En el caso del estimador sesgado, la media de la distribución muestral es menor o mayor que el valor del parámetro poblacional. En la gráfica B de la figura 7.10, $E(\hat{\theta})$ es mayor que θ ; así, la probabilidad de que los estadísticos muestrales sobreestimen el valor del parámetro poblacional es grande. En la figura se muestra la amplitud de este sesgo.

FIGURA 7.10 EJEMPLOS DE ESTIMADORES PUNTUALES SESGADO E INSESGADO



Al estudiar las distribuciones muestrales de la media muestral y de la proporción muestral, se vio que $E(\bar{x}) = \mu$ y que $E(\bar{p}) = p$. Por tanto, \bar{x} y \bar{p} son estimadores insesgados de sus correspondientes parámetros poblacionales μ y p .

En el caso de la desviación estándar muestral s y de la varianza muestral s^2 , se puede mostrar que $E(s^2) = \sigma^2$. Por tanto, se concluye que la varianza muestral s^2 es un estimador insesgado de la varianza poblacional σ^2 . En efecto, en el capítulo 3, cuando se presentaron las fórmulas para la varianza muestral y la desviación estándar muestral en el denominador se usó $n - 1$ en lugar de n para que la varianza muestral fuera un estimado insesgado de la varianza poblacional.

Eficiencia

Suponga que se usa una muestra aleatoria simple de n elementos para obtener dos estimadores puntuales insesgados de un mismo parámetro poblacional. En estas circunstancias preferirá usar el estimador puntual que tenga el menor error estándar, ya que dicho estimador tenderá a dar estimaciones más cercanas al parámetro poblacional. Se dice que el estimador puntual con menor error estándar tiene mayor **eficiencia relativa** que los otros.

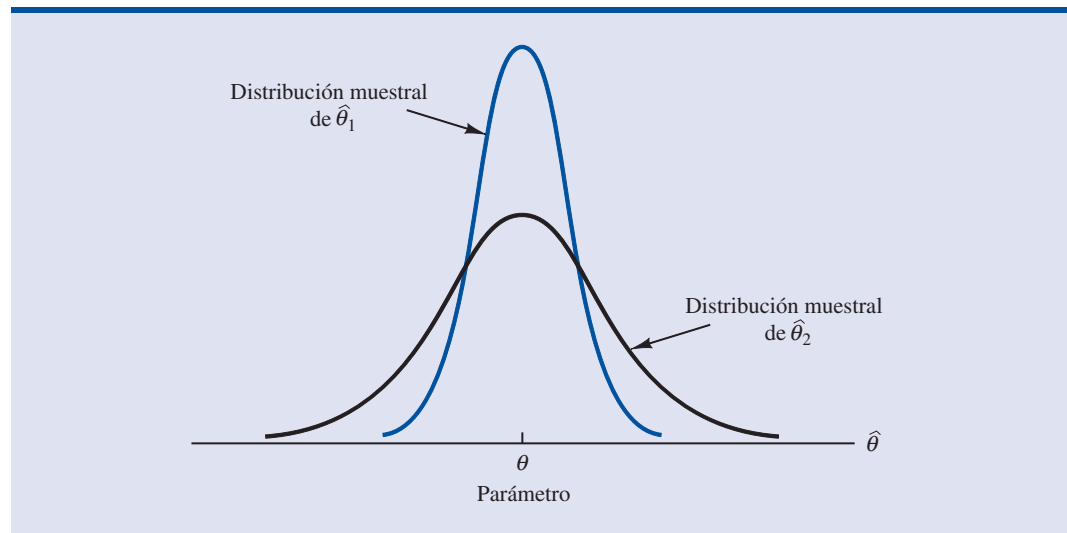
En la figura 7.11 se presentan las distribuciones muestrales de dos estimadores puntuales insesgados, $\hat{\theta}_1$ y $\hat{\theta}_2$. Observe que el error estándar de $\hat{\theta}_1$ es menor que el error estándar de $\hat{\theta}_2$; por tanto, los valores de $\hat{\theta}_1$ tienen más posibilidades de estar cerca del parámetro θ que los valores de $\hat{\theta}_2$. Como el error estándar del estimado puntual $\hat{\theta}_1$ es menor que el error estándar del estimado puntual $\hat{\theta}_2$, $\hat{\theta}_1$ es relativamente más eficiente que $\hat{\theta}_2$ y se prefiere como estimador puntual.

Consistencia

La tercera propiedad relacionada con un buen estimador puntual es la **consistencia**. Dicho de manera sencilla, un estimador puntual es consistente si el valor del estimador puntual tiende a estar más cerca del parámetro poblacional a medida que el tamaño de la muestra aumenta. En otras palabras, una muestra grande tiende a proporcionar mejor estimación puntual que una pequeña. Observe que en el caso de la media muestral \bar{x} , el error estándar de \bar{x} está dado por $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Puesto que $\sigma_{\bar{x}}$ está vinculado con el tamaño de la muestra, de manera que muestras mayores dan

Cuando se muestrean poblaciones normales, el error estándar de la media muestral es menor que el error estándar de la mediana muestral. Por tanto, la media muestral es más eficiente que la mediana muestral.

FIGURA 7.11 DISTRIBUCIONES MUESTRALES DE DOS ESTIMADORES PUNTUALES INSESADOS



valores menores de $\sigma_{\bar{x}}$, entonces muestras de tamaño grande tienden a proporcionar estimadores puntuales más cercanos a la media poblacional μ . Mediante un razonamiento similar, concluya que la proporción muestral \bar{p} es un estimador consistente de la proporción poblacional p .

NOTAS Y COMENTARIOS

En el capítulo 3 se dijo que la media y la mediana son dos medidas de localización central. En este capítulo sólo se estudió la media. La razón es que cuando se muestrea de una población normal, en la cual la media y la mediana poblacionales son idénticas, el error estándar de la mediana es cerca de 25% mayor que el error estándar de la media. Re-

cuerde que en el problema de EAI con $n = 30$, el error estándar de la media fue $\sigma_{\bar{x}} = 730.3$. El error estándar de la mediana en este problema será $1.25 \times (730.7) = 913$. Por tanto, la media muestral es más eficiente y tendrá más probabilidad de estar dentro de una determinada distancia de la media poblacional.

7.8

Otros métodos de muestreo

Esta sección proporciona una breve introducción a otros métodos de muestreo distintos al muestreo aleatorio simple.

Se describió el procedimiento de muestreo aleatorio simple y se estudiaron las propiedades de las distribuciones muestrales de \bar{x} y de \bar{p} cuando se usa el muestreo aleatorio simple. Sin embargo, el muestreo aleatorio simple no es el único método de muestreo que existe. Hay otros métodos como el muestreo aleatorio estratificado, el muestreo por conglomerados y el muestreo sistemático que, en ciertas situaciones, tienen ventajas sobre el muestreo aleatorio simple. En esta sección se introducen brevemente estos métodos de muestreo. En el capítulo 22 que se encuentra en el CD que se distribuye con el texto se estudian estos métodos de muestreo con más detenimiento.

Muestreo aleatorio estratificado

El muestreo aleatorio estratificado funciona mejor cuando la varianza entre los elementos de cada estrato es relativamente pequeña.

En el **muestreo aleatorio estratificado** los elementos de la población primero se dividen en grupos, a los que se les llama *estratos*, de manera que cada elemento pertenezca a uno y sólo un estrato. La base para la formación de los estratos, que puede ser departamento, edad, tipo de industria, etc., está a discreción de la persona que diseña la muestra. Sin embargo, se obtienen mejores resultados cuando los elementos que forman un estrato son lo más parecido posible. La figura 7.12 es un diagrama de una población dividida en H estratos.

Una vez formados los estratos, se toma una muestra aleatoria simple de cada estrato. Existen fórmulas para combinar los resultados de las muestras de los varios estratos en una estimación

FIGURA 7.12 DIAGRAMA DE UN MUESTREO ALEATORIO ESTRATIFICADO

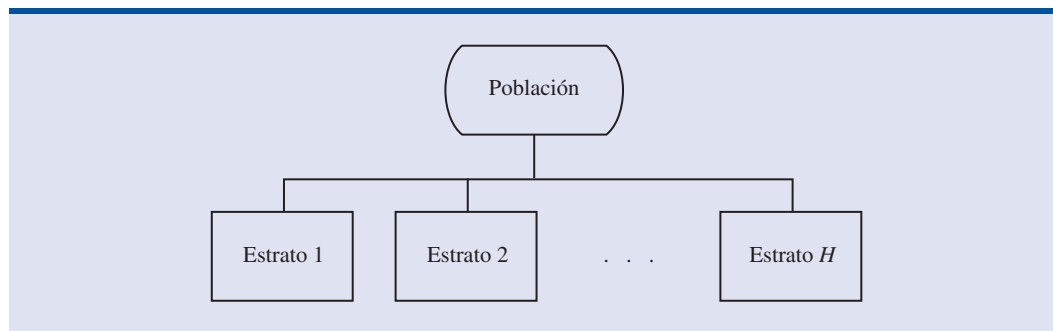
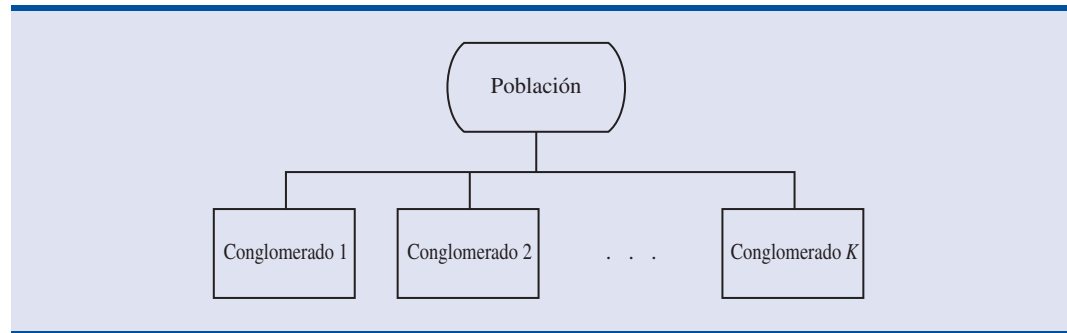


FIGURA 7.13 DIAGRAMA DEL MUESTREO POR CONGLOMERADOS

del parámetro poblacional de interés. El valor del muestreo aleatorio estratificado depende de qué tan homogéneos sean los elementos dentro de cada estrato. Si los elementos de un estrato son homogéneos, el estrato tendrá una varianza pequeña. Por tanto, con muestras relativamente pequeñas de los estratos se obtienen buenas estimaciones de las características de los estratos. Si los estratos son homogéneos, el muestreo aleatorio estratificado, proporciona resultados tan precisos como los de un muestreo aleatorio simple, pero con una muestra de tamaño total menor.

Muestreo por conglomerados

El muestreo por conglomerados funciona mejor cuando cada conglomerado proporciona una representación a menor escala de la población.

En el **muestreo por conglomerados** los elementos de la muestra primero se dividen en grupos separados, llamados *conglomerados*. Cada elemento de la población pertenece a uno y sólo un conglomerado (véase figura 7.13). Se toma una muestra aleatoria simple de los conglomerados. La muestra está formada por todos los elementos dentro de cada uno de los conglomerados que forman la muestra. El muestreo por conglomerados tiende a proporcionar mejores resultados cuando los elementos dentro de los conglomerados no son semejantes. Lo ideal es que cada conglomerado sea una representación, a pequeña escala, de la población. Si todos los conglomerados son semejantes en este aspecto, tomando en la muestra un número pequeño de conglomerados se obtendrá una buena estimación de los parámetros poblacionales.

Una de las principales aplicaciones del muestreo por conglomerados es el muestreo de áreas, en el que los conglomerados son las manzanas de una ciudad u otras áreas bien definidas. El muestreo por conglomerados requiere, por lo general, tamaños de muestra mayores que los requeridos en el muestreo aleatorio simple o en el muestreo aleatorio estratificado. Sin embargo, es posible reducir costos debido a que cuando se envía a un entrevistador a uno de los conglomerados de la muestra (por ejemplo, a una manzana de una ciudad), es posible obtener muchas observaciones en poco tiempo. Por tanto, se obtiene una muestra de tamaño grande a un costo significativamente menor.

Muestreo sistemático

Para ciertos muestreos, en especial en aquellos con poblaciones grandes, se necesita mucho tiempo para tomar una muestra aleatoria simple (hallando primero los números aleatorios y después contando y recorriendo toda una lista de la población hasta encontrar los elementos correspondientes). Una alternativa al muestreo aleatorio simple es el **muestreo sistemático**. Por ejemplo, si se quiere una muestra de tamaño 50 de una población que tiene 5000 elementos, se muestrea uno de cada $5000/50 = 100$ elementos de la población. En este caso, un muestreo sistemático consiste en seleccionar en forma aleatoria uno de los primeros elementos de la lista de la población. Los otros elementos se identifican contando a partir del primer elemento 100 elementos para tomar el elemento que tenga la posición 100 en la lista de la población, a partir de este elemento se cuentan otros 100 y así se continúa. Por lo general, de esta manera es más fácil de identificar la muestra de 50 que si se usara el muestreo aleatorio simple. Como el primer elemento que se selecciona es elegido en forma aleatoria, se supone que una muestra sistemática tiene las

propiedades de una muestra aleatoria simple. Esta suposición es aplicable, en especial, cuando la lista de los elementos de la población es un orden aleatorio de los elementos.

Muestreo de conveniencia

Los métodos de muestreo hasta ahora vistos se conocen como técnicas *probabilísticas de muestreo*. Los elementos seleccionados de una población tienen una probabilidad conocida de ser incluidos en la muestra. La ventaja del muestreo probabilístico es que, por lo general, se identifica la distribución muestral del estadístico muestral correspondiente. Para determinar las propiedades de la distribución muestral se usan las fórmulas presentadas en este capítulo para el muestreo aleatorio simple. La distribución muestral permite hacer afirmaciones probabilísticas acerca del error al usar los resultados muestrales para hacer inferencias acerca de la población.

El **muestreo de conveniencia** es una técnica de *muestreo no probabilístico*. Como el nombre lo indica, la muestra se determina por conveniencia. Los elementos se incluyen en la muestra sin que haya una probabilidad previamente especificada o conocida de que sean incluidos en la muestra. Por ejemplo, un profesor que realiza una investigación en una universidad puede usar estudiantes voluntarios para que constituyan una muestra; ¿la razón para elegirlos? simple, los tiene al alcance y participarán como sujetos a un costo bajo o sin costo. De manera similar, un inspector puede muestrear un cargamento de naranjas seleccionando al azar naranjas de varias de las cajas. Marcar cada naranja y usar un método probabilístico de muestreo puede no resultar práctico. Muestras como los paneles de voluntarios en investigaciones sobre los consumidores son también muestras de conveniencia.

Dichas muestras tienen la ventaja de que es relativamente fácil seleccionar la muestra y recoger los datos; sin embargo, es imposible evaluar la “bondad” de la muestra en términos de su representatividad de la población. Una muestra de conveniencia puede o no dar buenos resultados. Algunas veces los investigadores aplican los métodos estadísticos propios de muestras probabilísticas a las muestras de conveniencia, con el argumento de que la muestra de conveniencia se trata como si fuera una muestra probabilística. Sin embargo, estos argumentos no tienen fundamento y se debe tener cuidado al interpretar los resultados de muestreos de conveniencia que han sido usados para hacer inferencias acerca de la población.

Muestreo subjetivo

Otra técnica de muestreo no probabilística es el muestreo subjetivo. En este método la persona que más sabe sobre un asunto selecciona elementos de la población que considera los más representativos de la población. Este método suele ser una manera relativamente fácil de seleccionar una muestra. Por ejemplo, un reportero puede seleccionar dos o tres senadores considerando que estos senadores reflejan la opinión general de todos los senadores. Sin embargo, la calidad de los resultados muestrales depende de la persona que selecciona la muestra. Aquí también hay que tener mucho cuidado al hacer inferencias acerca de las poblaciones a partir de muestreos subjetivos.

NOTAS Y COMENTARIOS

Se recomienda usar métodos de muestreo probabilístico: muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerados o muestreo sistemático. Si se usan estos métodos existen fórmulas para evaluar la “bondad” de los resultados muestrales en términos de la cercanía de

los resultados a los parámetros poblacionales que se estiman. Con los muestreos de conveniencia o con los muestreos subjetivos no se puede estimar la bondad de los resultados. Por tanto, debe tenerse mucho cuidado al interpretar resultados basados en métodos de muestreo no probabilístico.

Resumen

En este capítulo se presentaron los conceptos de muestreo aleatorio simple y de distribución muestral. Se mostró cómo seleccionar una muestra aleatoria simple y la forma de usar los datos recolectados de la muestra para obtener estimadores puntuales de los parámetros poblacionales. Ya que distintas muestras aleatorias simples dan valores diferentes de los estimadores puntuales, los estimadores puntuales como \bar{x} y \bar{p} son variables aleatorias. A la distribución de probabilidad de una variable aleatoria de este tipo se le conoce como distribución muestral. En particular, se describieron la distribución muestral de la media muestral \bar{x} y la distribución muestral de la proporción muestral \bar{p} .

Al estudiar las características de las distribuciones muestrales de \bar{x} y de \bar{p} , se vio que $E(\bar{x}) = \mu$ y que $E(\bar{p}) = p$. Después de dar las fórmulas para la desviación estándar o error estándar de dichos estimadores, se describieron las condiciones necesarias para que las distribuciones muestrales de \bar{x} y de \bar{p} sigan una distribución normal. Otros métodos de muestreo fueron el muestreo aleatorio estratificado, el muestreo por conglomerados, el muestreo sistemático, el muestreo por conveniencia y el muestreo subjetivo.

Glosario

Parámetro Característica numérica de una población, por ejemplo, la media poblacional μ , la desviación estándar poblacional σ , la proporción poblacional p , etcétera.

Muestreo aleatorio simple Poblaciones finitas: muestra seleccionada de manera que cada una de las muestras de tamaño n tenga la misma probabilidad de ser seleccionada. Poblaciones infinitas: muestra seleccionada de manera que todos los elementos provengan de la misma población y los elementos se seleccionen de manera independiente.

Muestreo sin reemplazo Una vez que un elemento ha sido incluido en la muestra, se retira de la población y ya no se selecciona una vez más.

Muestreo con reemplazo Una vez que un elemento se ha incluido en la muestra, se regresa a la población. Un elemento ya seleccionado para la muestra puede ser seleccionado nuevamente y puede aparecer más de una vez en la muestra.

Estadístico muestral Característica muestral, por ejemplo, la media muestral \bar{x} , la desviación estándar muestral s , la proporción muestral \bar{p} , etc. El valor del estadístico muestral se usa para estimar el valor del correspondiente parámetro poblacional.

Estimador puntual Un estadístico muestral como \bar{x} , s , o \bar{p} que proporciona una estimación puntual del parámetro poblacional correspondiente.

Estimación puntual Valor de un estimador que se usa en una situación particular como estimación del parámetro poblacional.

Distribución muestral Distribución de probabilidad que consta de todos los posibles valores de un estadístico muestral.

Insesgado Propiedad de un estimador que consiste en que el valor esperado del estimador puntual es igual al parámetro poblacional que estima.

Factor de corrección para una población finita Es el factor $\sqrt{(N - n)/(N - 1)}$ que se usa en las fórmulas de $\sigma_{\bar{x}}$ y $\sigma_{\bar{p}}$ siempre que se muestrea de una población finita y no de una población infinita. Sin embargo, hay una regla generalmente aceptada, ignorar el factor de corrección en una población finita siempre que $n/N \leq 0.05$.

Error estándar La desviación estándar de un estimador puntual.

Teorema del límite central Permite usar la distribución de probabilidad normal para aproximar la distribución muestral de \bar{x} siempre que la muestra sea grande.

Eficiencia relativa Dados dos estimadores puntuales insesgados de un mismo parámetro poblacional, el estimador puntual que tenga menor error estándar será más eficiente.

Consistencia Propiedad de un estimador puntual que está presente siempre que muestras más grandes dan estimaciones puntuales más cercanas al parámetro poblacional.

Muestreo aleatorio estratificado Método probabilístico en el que primero se divide la población en estratos y después se toma una muestra aleatoria simple de cada estrato.

Muestreo por conglomerados Método probabilístico en el que primero se divide la población en conglomerados y después se toma una muestra aleatoria de los conglomerados.

Muestreo sistemático Método probabilístico en el que primero se selecciona uno de los primeros k elementos de una población y después se selecciona cada k -ésimo elemento de la población.

Muestreo de conveniencia Método no-probabilístico en el que la selección de los elementos para la muestra es de acuerdo con la conveniencia.

Muestreo subjetivo Método no-probabilístico en el que la selección de los elementos para la muestra es de acuerdo con la opinión de la persona que hace el estudio.

Fórmulas clave

Valor esperado de \bar{x}

$$E(\bar{x}) = \mu \quad (7.1)$$

Desviación estándar de \bar{x} (error estándar)

<i>Población finita</i>	<i>Población infinita</i>
$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \left(\frac{\sigma}{\sqrt{n}} \right)$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

(7.2)

Valor esperado de \bar{p}

$$E(\bar{p}) = p \quad (7.4)$$

Desviación estándar de \bar{p} (error estándar)

<i>Población finita</i>	<i>Población infinita</i>
$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \sqrt{\frac{p(1-p)}{n}}$	$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$

(7.5)

Ejercicios complementarios

42. *BusinessWeek's* Corporate Scoreboard proporciona datos trimestrales sobre distintos aspectos de las acciones de 899 empresas (*BusinessWeek*, 14 de agosto de 2000). Las empresas son numeradas del 1 al 899 en el orden en que aparecen en la lista del Corporate Scoreboard. Remítase a la parte inferior de la segunda columna de dígitos aleatorios de la tabla 7.1, ignore los dos primeros dígitos de cada conjunto de números, use números de tres dígitos, empiece con el 112, lea hacia arriba de la columna para determinar las ocho primeras empresas a incluir en una muestra aleatoria simple.
43. Los estadounidenses están cada vez más preocupados por el aumento en los costos de Medicare. En 1990 el promedio de gastos anuales de un derechohabiente de Medicare era \$3267; en el 2003 el promedio de gastos anuales de un derechohabiente de Medicare era \$6883 (*Money*, otoño de

- 2003). Suponga que usted contrata a una empresa consultora para tomar una muestra de 50 de los derechohabientes de Medicare en 2003 con objeto de investigar los gastos. Asuma que la desviación estándar en los gastos de Medicare en 2003 haya sido de \$2000.
- Muestre la distribución muestral de la media, en muestras de tamaño cincuenta, de los gastos de derechohabientes de Medicare en 2003.
 - ¿Cuál es la probabilidad de que la media muestral no se aleje más de $\pm \$300$ de la media poblacional?
 - ¿Cuál es la probabilidad de que la media muestral sea mayor que \$7500? Si la empresa que contrató le dice que la media muestral en los derechohabientes que entrevistó es \$7500, ¿dudaría que la empresa contratada hubiera hecho un muestreo adecuado? ¿Por qué sí o por qué no?
44. *BusinessWeek* encuesta a ex alumnos de administración 10 años después de terminados sus estudios (*BusinessWeek*, 22 de septiembre de 2003). Uno de los hallazgos fue que gastan en promedio \$115.50 semanales en comidas sociales. A usted se le pide que realice un estudio con una muestra de 40 de estos ex alumnos.
- Muestre la distribución muestral de \bar{x} , la media muestral de los gastos de 40 ex alumnos.
 - ¿Cuál es la probabilidad de que la media muestral no se aleje en más o menos de \$10 de la media poblacional?
 - Suponga que encuentra una media muestral de \$100. ¿Cuál es la probabilidad de hallar una media muestral de \$100 o menos? ¿Consideraría que los ex alumnos de esta muestra son un grupo inusual respecto a estos gastos? ¿Por qué sí o por qué no?
45. El tiempo promedio que un estadounidense ve televisión es 15 horas por semana (*Money*, noviembre de 2003). Suponga que se toma una muestra de 60 estadounidenses para investigar con más detalle sus hábitos a este respecto. Asuma que la desviación estándar poblacional en las horas de televisión semanales es $\sigma = 4$ horas.
- ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de 1 hora de la media poblacional?
 - ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de 45 minutos de la media poblacional?
46. En Indiana el salario anual promedio de un empleado del gobierno federal es \$41 979 (*The World Almanac*, 2001). Use esta cifra como media poblacional y suponga que la desviación estándar poblacional es $\sigma = \$5000$. Suponga que se selecciona una muestra de 50 de estos empleados del gobierno federal.
- ¿Cuál es el valor del error estándar de la media?
 - ¿Cuál es la probabilidad de que la media muestral sea mayor que \$41 979?
 - ¿Cuál es la probabilidad de que la media muestral no se aleje más o menos de \$1000 de la media poblacional?
 - ¿Qué tanto variaría la probabilidad del inciso c si el tamaño de la muestra se aumentara a 100?
47. Tres empresas llevan inventarios de distintos tamaños. El inventario de la empresa A contiene 2000 artículos, el inventario de la empresa B tiene 5000 artículos y el inventario de la empresa C 10 000. La desviación estándar poblacional de los costos de los artículos en los inventarios de estas empresas es $\sigma = 144$. Un consultor de estadística recomienda que cada empresa tome una muestra de 50 artículos de su inventario para obtener una estimación estadística válida del costo promedio por artículo. Los administradores de la empresa más pequeña opinan que como su población es menor se podrá hacer la estimación con una muestra mucho más pequeña de la que se requiere para la empresa más grande. Sin embargo, el consultor opina que para tener el mismo error estándar y, por tanto, la misma precisión en los resultados muestrales, todas las empresas deberán emplear el mismo tamaño de muestra, sin importar el tamaño de la población.
- Con el factor de corrección para una población finita, calcule el error estándar de cada una de las tres empresas para un tamaño de muestra de 50.
 - ¿Cuál es la probabilidad en cada empresa de que la media muestral \bar{x} esté a no más de ± 25 de la media poblacional μ ?

48. Un investigador informa sobre sus resultados diciendo que el error estándar de la media es 20. La desviación estándar poblacional es 500.
 - a. ¿De qué tamaño fue la muestra usada en esta investigación?
 - b. ¿Cuál es la probabilidad de que la estimación puntual esté a no más de ± 25 de la media poblacional?
49. Un inspector de control de calidad vigila periódicamente un proceso de producción. El inspector selecciona muestras aleatorias simples de artículos ya terminados y calcula la media muestral del peso del producto \bar{x} . Si en un periodo largo se encuentra que 5% de los valores de \bar{x} son mayores que 2.1 libras y 5% son menores que 1.9 libras. ¿Cuáles son la media y la desviación estándar de la población de los productos elaborados en este proceso?
50. Cerca de 28% de las empresas tienen como propietario a una mujer (*The Cincinnati Enquirer*, 26 de enero de 2006). Responda estas preguntas con base en una muestra de 240 empresas.
 - a. Muestre la distribución muestral de \bar{p} , la proporción muestral de las empresas propiedad de una mujer.
 - b. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.04 de la proporción poblacional?
 - c. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.02 de la proporción poblacional?
51. Una empresa de investigación de mercado realiza encuestas telefónicas con una tasa de respuesta de 40%, de acuerdo con la experiencia. ¿Cuál es la probabilidad de que en una muestra de 400 números telefónicos 150 personas cooperen y respondan las preguntas? En otras palabras, ¿cuál es la probabilidad de que la proporción muestral sea al menos $150/400 = 0.375$?
52. Los publicistas contratan proveedores de servicios de Internet y motores de búsqueda para poner su publicidad en los sitios Web. Pagan una cuota de acuerdo con el número de clientes potenciales que hacen clic en su publicidad. Por desgracia, el fraude por clic —la práctica de hacer clic en una publicidad con el solo objeto de aumentar las ganancias— se ha convertido en un problema. Cuarenta por ciento de los publicistas se quejan de haber sido víctima de fraude por clic (*BusinessWeek*, 13 de marzo de 2006). Suponga que se toma una muestra aleatoria de 380 publicistas con objeto de tener más información acerca de cómo son afectados por este fraude por clic.
 - a. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.04 de la proporción poblacional?
 - b. ¿Cuál es la probabilidad de que la proporción muestral sea mayor que 0.45?
53. La proporción de personas aseguradas con una compañía de seguros para automóviles que tienen una multa de tráfico en el periodo de un año es 0.15
 - a. Indique la distribución muestral de \bar{p} si se emplea una muestra aleatoria de 150 asegurados para determinar la proporción de quienes han tenido por lo menos una multa en un año.
 - b. ¿Cuál es la probabilidad de que la proporción muestral esté a no más de ± 0.03 de la proporción poblacional?
54. Lori Jeffrey es un exitoso representante de ventas de libros universitarios, tiene éxito en sus recomendaciones de libros en 25% de sus llamadas. Considere sus llamadas de ventas de un mes como muestra de todas sus posibles llamadas, suponga que en el análisis estadístico de los datos se encuentra que el error estándar de la proporción es 0.0625.
 - a. ¿De qué tamaño fue la muestra que se usó en el análisis? Es decir, ¿cuántas llamadas hizo Lori Jeffrey en ese mes?
 - b. Sea \bar{p} la proporción muestral de éxitos en sus recomendaciones de libros en ese mes. Muestre la distribución muestral de \bar{p} .
 - c. Mediante la distribución muestral de \bar{p} , calcule la probabilidad de que el vendedor tenga éxito en 30% o más de las llamadas de ventas en el lapso de un mes.

Apéndice 7.1 Valor esperado y desviación estándar de \bar{x}

En este apéndice se presentan las bases matemáticas de las expresiones $E(\bar{x})$, valor esperado de \bar{x} , ecuación (7.1), y $\sigma_{\bar{x}}$, desviación estándar de \bar{x} , ecuación (7.2).

Valor esperado de \bar{x}

Se tiene una población que tiene media μ y varianza σ^2 . Se selecciona una muestra aleatoria de tamaño n cuyas observaciones se denotan x_1, x_2, \dots, x_n . La media muestral \bar{x} se calcula como sigue.

$$\bar{x} = \frac{\sum x_i}{n}$$

Si se repiten los muestreos aleatorios de tamaño n , \bar{x} será una variable aleatoria que tomará diferentes valores dependiendo de los n elementos que formen la muestra. El valor esperado de la variable aleatoria \bar{x} es la media de todos los posibles valores \bar{x} .

$$\begin{aligned} \text{Media de } \bar{x} &= E(\bar{x}) = E\left(\frac{\sum x_i}{n}\right) \\ &= \frac{1}{n}[E(x_1) + E(x_2) + \dots + E(x_n)] \\ &= \frac{1}{n}[E(x_1) + E(x_2) + \dots + E(x_n)] \end{aligned}$$

Para cada x_i se tiene $E(x_i) = \mu$; por tanto,

$$\begin{aligned} E(\bar{x}) &= \frac{1}{n}(\mu + \mu + \dots + \mu) \\ &= \frac{1}{n}(n\mu) = \mu \end{aligned}$$

Este resultado indica que la media de todos los posibles valores de \bar{x} es igual a la media poblacional μ . Es decir $E(\bar{x}) = \mu$.

Desviación estándar de \bar{x}

Se tiene, de nuevo, una población con media μ y varianza σ^2 y una media muestral dada por

$$\bar{x} = \frac{\sum x_i}{n}$$

Se sabe que \bar{x} es una variable aleatoria que toma distintos valores en distintas muestras aleatorias de tamaño n , dependiendo de los elementos que constituyen la muestra. Lo que sigue es una deducción de la fórmula para la desviación estándar de los valores de \bar{x} , $\sigma_{\bar{x}}$, en el caso en el que la población sea infinita. La deducción de la fórmula para $\sigma_{\bar{x}}$ cuando la población es finita y el muestreo se hace sin reemplazo es más complicada y queda fuera de los alcances de este texto.

De regreso al caso de una población infinita, recuerde que una muestra aleatoria simple de una población infinita, consta de observaciones x_1, x_2, \dots, x_n que son independientes. Las dos expresiones siguientes son fórmulas generales para la varianza de una variable aleatoria.

$$\text{Var}(ax) = a^2 \text{Var}(x)$$

donde a es una constante y x es una variable aleatoria, y

$$\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y)$$

donde x y y son variables aleatorias *independientes*. Usando las ecuaciones anteriores, se puede deducir la fórmula para la varianza de la variable \bar{x} como sigue.

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

Entonces, como $1/n$ es una constante, se tiene

$$\begin{aligned} \text{Var}(\bar{x}) &= \left(\frac{1}{n}\right)^2 \text{Var}(\sum x_i) \\ &= \left(\frac{1}{n}\right)^2 \text{Var}(x_1 + x_2 + \dots + x_n) \end{aligned}$$

En el caso de una población infinita, las variables aleatorias x_1, x_2, \dots, x_n son independientes, lo que nos permite escribir

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 [\text{Var}(x_1) + \text{Var}(x_2) + \dots + \text{Var}(x_n)]$$

Para toda x_i , se tiene $\text{Var}(x_i) = \sigma^2$; por tanto se tiene

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \dots + \sigma^2)$$

Como en esta expresión hay n valores σ^2 , se tiene

$$\text{Var}(\bar{x}) = \left(\frac{1}{n}\right)^2 (n\sigma^2) = \frac{\sigma^2}{n}$$

Sacando ahora la raíz cuadrada, se obtiene la fórmula de la desviación estándar de \bar{x} .

$$\sigma_{\bar{x}} = \sqrt{\text{Var}(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

Apéndice 7.2 Muestreo aleatorio con Minitab

Si en un archivo se encuentra una lista con los elementos de una población, se puede usar Minitab para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna 1 del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac-The Millenium Edition 2000*). La columna 2 contiene

TABLA 7.6 PUNTUACIÓN GENERAL PARA LAS PRIMERAS 10 ÁREAS METROPOLITANAS EN EL CONJUNTO DE DATOS METAREAS

Área metropolitana	Puntuación
Albany, NY	64.18
Albuquerque, NM	66.16
Appleton, WI	60.56
Atlanta, GA	69.97
Austin, TX	71.48
Baltimore, MD	69.75
Birmingham, AL	69.59
Boise City, ID	68.36
Boston, MA	68.99
Buffalo, NY	66.10

la puntuación general dada a cada área. En la tabla 7.6 se presentan las primeras 10 áreas metropolitanas con sus puntuaciones correspondientes.

Suponga que pretende seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio sobre el costo de la vida en Estados Unidos y Canadá. Para seleccionar la muestra aleatoria se siguen los pasos que se indica a continuación.

Paso 1. Seleccionar el menú desplegable **Calc**

Paso 2. Elegir **Random Data**

Paso 3. Elegir **Sample From Columns**

Paso 4. Cuando aparezca el cuadro de diálogo **Sample From Columns:**

Ingresar 30 en el cuadro **Sample**

Ingresar C1 C2 en el cuadro que se encuentra debajo

Ingresar C3 C4 en el cuadro **Store samples in**

Paso 5. Hacer clic en **OK**

La muestra aleatoria con las 30 áreas metropolitanas aparece en las columnas C3 y C4.

Apéndice 7.3 Muestreo aleatorio con Excel

Si en un archivo se encuentra una lista con los elementos de una población, Excel se podrá usar para seleccionar una muestra aleatoria simple. Por ejemplo, en la columna A del conjunto de datos MetAreas se proporciona una lista de las 100 principales áreas metropolitanas de Estados Unidos y Canadá (*Places Rated Almanac-The Millenium Edition 2000*).

La columna B contiene la puntuación general dada a cada área. En la tabla 7.6 se presentan las primeras 10 áreas metropolitanas con sus puntuaciones correspondientes.

Suponga que quiere seleccionar una muestra aleatoria simple de 30 áreas metropolitanas con objeto de hacer un estudio sobre el costo de la vida en Estados Unidos y Canadá.

Los renglones de cualquier conjunto de datos en Excel se pueden colocar en orden aleatorio agregando una columna al conjunto de datos y llenando la columna con números aleatorios mediante la función = ALEATORIO(); después con la herramienta de Excel Orden ascendente aplicada a la columna de números aleatorios, los renglones del conjunto de datos quedarán reordenados aleatoriamente. La muestra aleatoria de tamaño n aparecerá en los n primeros renglones del conjunto de datos reordenado.

En el conjunto de datos MetAreas, los encabezados aparecen en el renglón 1 y las 100 áreas metropolitanas se encuentran en los renglones 2 a 101. Para seleccionar una muestra aleatoria de 30 áreas metropolitanas siga los pasos siguientes.

Paso 1. Ingresar = ALEATORIO() en la celda C2.

Paso 2. Copiar la celda C2 a las celdas C3:C101

Paso 3. Seleccionar cualquier celda de la columna C

Paso 4. Clic en el botón **Orden ascendente** de la barra de herramientas.

La muestra aleatoria con 30 áreas metropolitanas aparecerá en los renglones 2 a 31 del conjunto de datos reordenado. Los números aleatorios de la columna C ya no se necesitan y pueden borrarse si se desea.