CAPÍTULO 1

Datos y estadísticas

CONTENIDO

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 suscriptores 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*.

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

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.

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

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.

archivo en CD
BWS&P

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

Empresa	Bolsa de valores	Denominación abreviada Ticker	Posición en *BusinessWeek*	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	BMY	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.

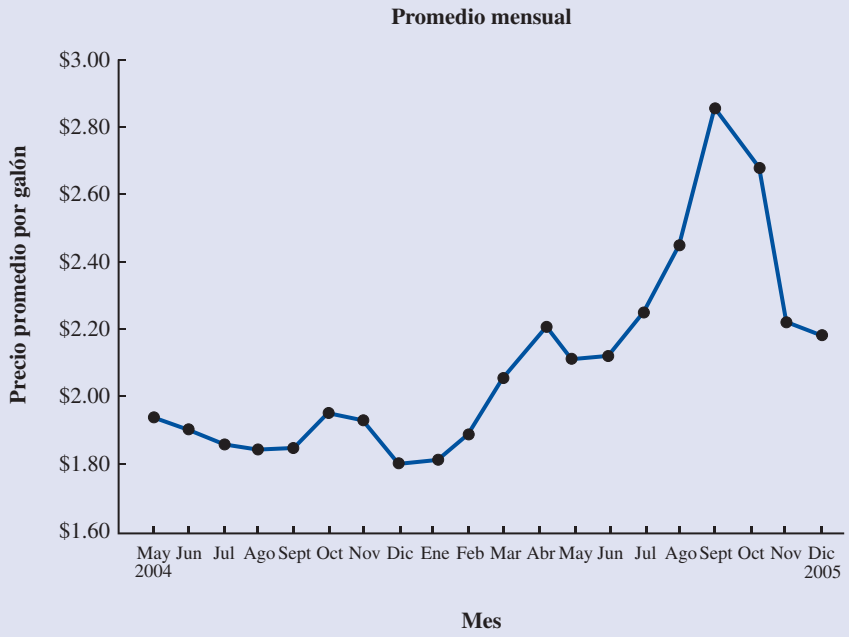
Por otro lado, las operaciones aritméticas sí tienen sentido en las variables cuantitativas. Por ejemplo, cuando se tienen variables cuantitativas, los datos se pueden sumar y luego dividir entre el número de observaciones para calcular el valor promedio. Este promedio suele ser útil y fácil de interpretar. 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



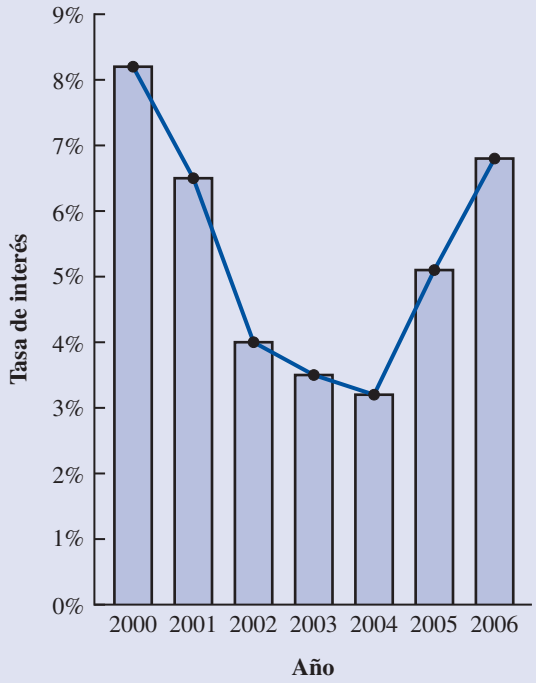
Fuente: U.S. Energy Information Administration, enero de 2006.

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.

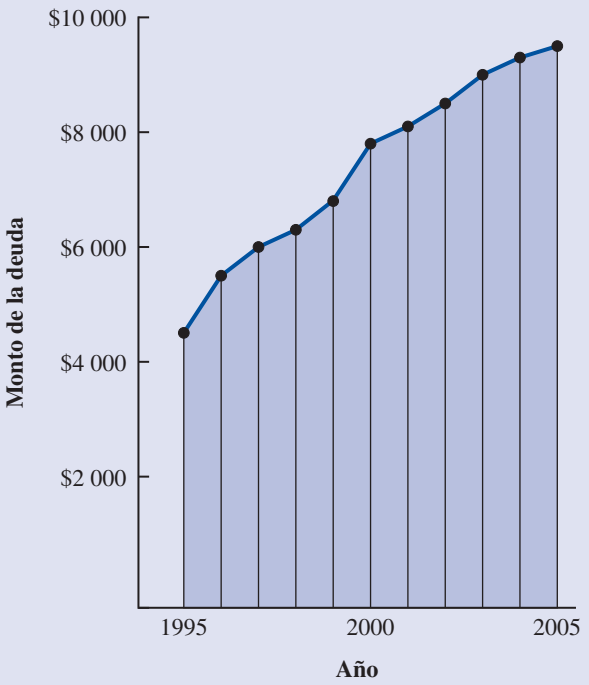
En la gráfica (B) se observa un inquietante aumento en el adeudo promedio por hogar en tarjetas de crédito durante un periodo de 10 años, de 1995 a 2005. Advierta cómo en la series de tiempo se nota un aumento anual casi constante en el adeudo promedio por hogar en tarjetas de crédito que va de $4500 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.

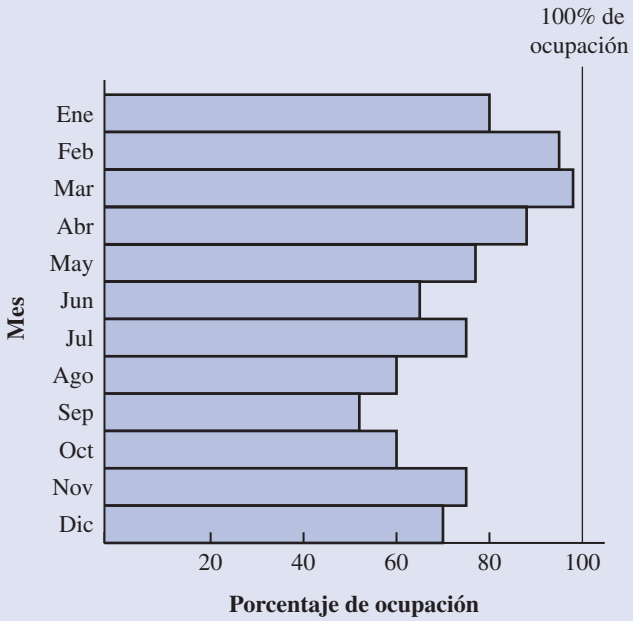
FIGURA 1.2 DIVERSAS GRÁFICAS DE DATOS DE SERIES DE TIEMPO



(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*.

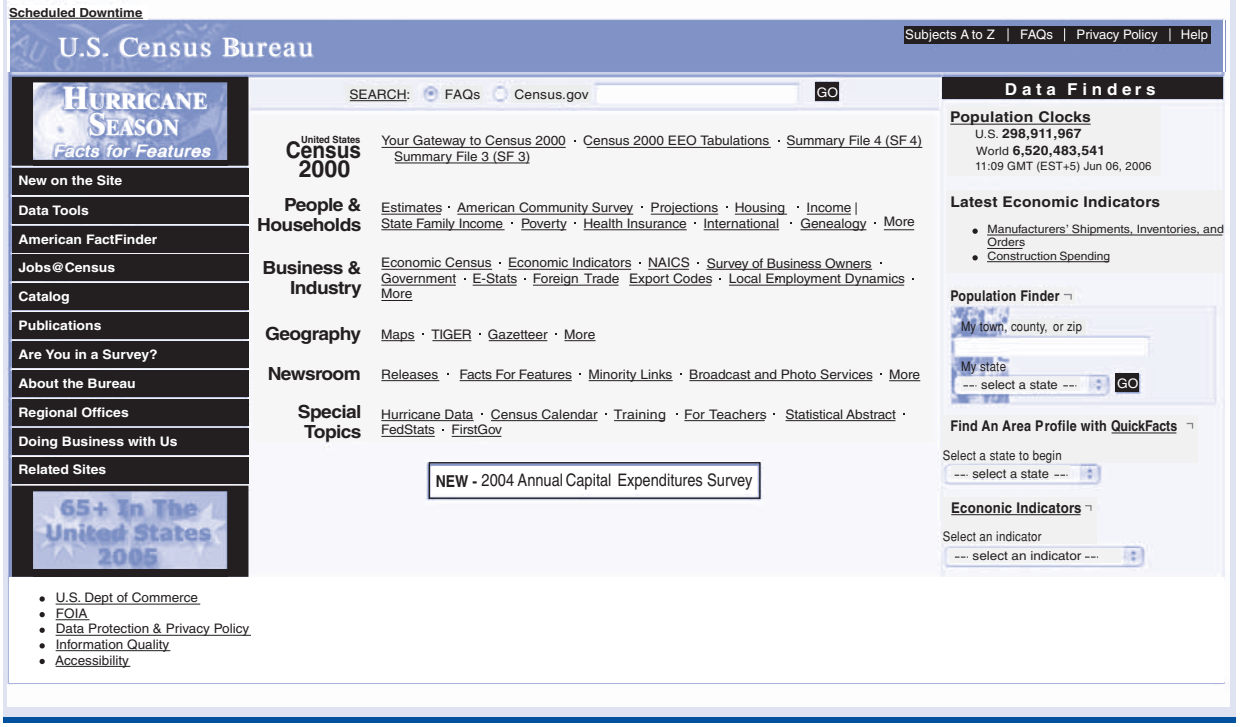
El mayor estudio estadístico experimental jamás realizado se cree que es el experimento del Servicio de Salud Pública para la vacuna Salk contra la polio. Se eligieron casi 2 millones de niños de 1o., 2o. y 3er. grados en Estados Unidos.

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-

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

The LOBSTER Pot
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	❑	❑	❑	❑
Amabilidad en el servicio	❑	❑	❑	❑
Prontitud en el servicio	❑	❑	❑	❑
Limpieza	❑	❑	❑	❑
Gestión	❑	❑	❑	❑

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 dividien-

FIGURA 1.5 GRÁFICA DE BARRAS DE LA VARIABLE BOLSA DE VALORES

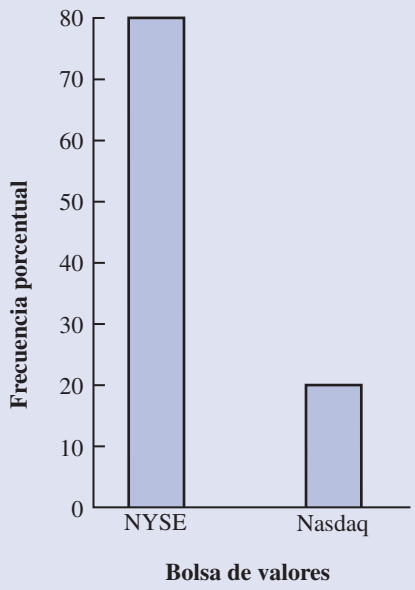
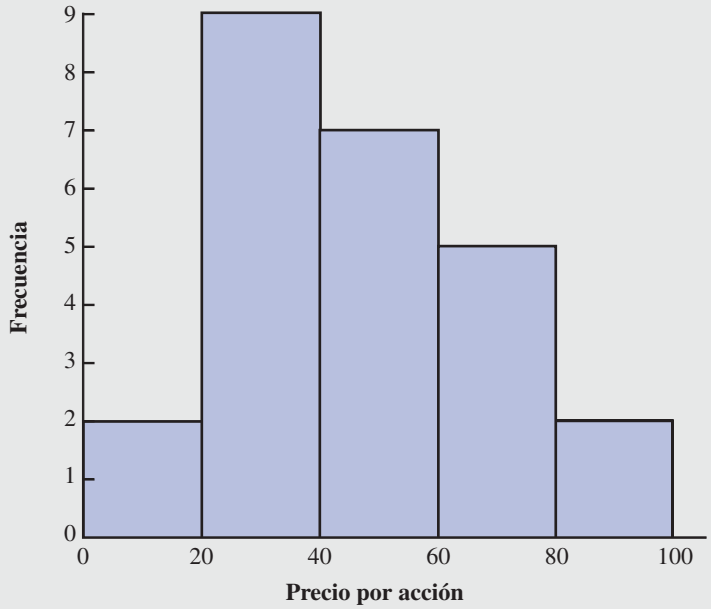


FIGURA 1.6 HISTOGRAMA DE LOS PRECIOS POR ACCIÓN DE 25 ACCIONES S&P



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.

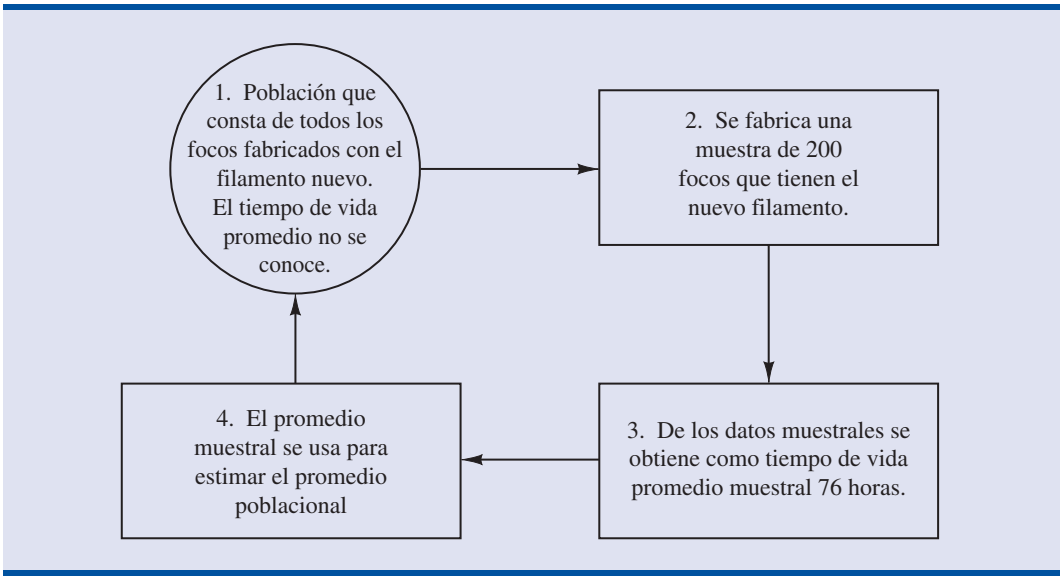
Siempre que un estadístico usa una muestra para estimar una característica poblacional que interesa, suele proporcionar información acerca de la calidad o precisión de la estimación. En el ejemplo de Norris, el estadístico puede informar que la estimación puntual del tiempo de vida promedio de la población 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.

archivo en CD
Norris

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.

Ejercicios complementarios

1. Describa la diferencia entre estadística como dato numérico y estadística como disciplina o campo de estudio.

Autoexamen

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?

Autoexamen

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

archivo en CD
Hotel

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

archivo en CD Minisystems

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 Nacional 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?

a. ¿Cuál es su edad?
b. ¿Es usted hombre o mujer?
c. ¿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.
d. ¿Cuánto tiempo hace que tiene su trabajo o cargo actual?
e. ¿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.

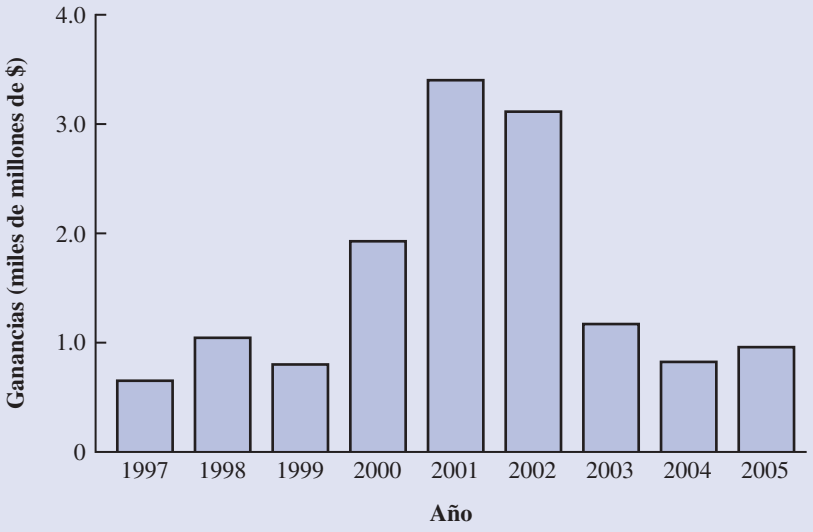
11. Diga de cada una de las variables siguientes si es cualitativa o cuantitativa e indique la escala de medición a la que pertenece.
    a. Ventas anuales.
    b. Tamaño de los refrescos (pequeño, mediano, grande).
    c. Clasificación como empleado (GS 1 a GS 18).
    d. Ganancia por acción.
    e. Modo de pago (al contado, cheque, tarjeta de crédito).
12. 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

    a. ¿Cuál es la población que se estudia?
    b. ¿El uso de un cuestionario es una buena manera de tener información de los pasajeros en los vuelos de llegada?
    c. Diga de cada una de las cuatro preguntas si los datos que suministra son cualitativos o cuantitativos.

Autoexamen

13. 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 prometedor 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

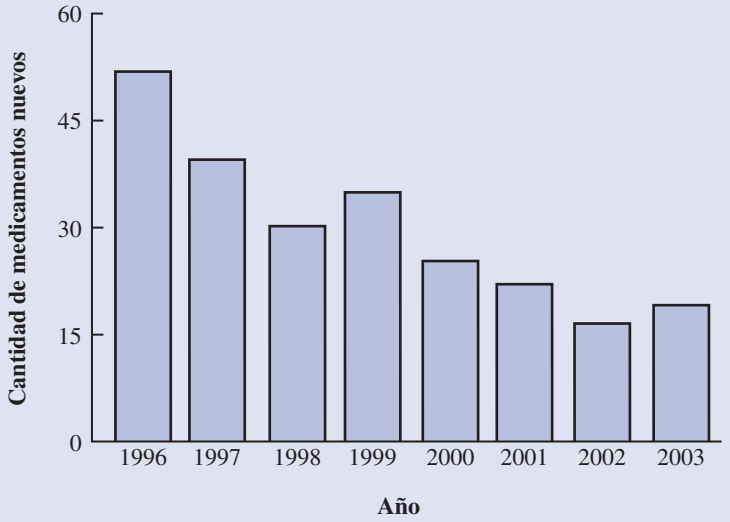
a. Haga una gráfica de series de tiempo para los años 2004 a 2007 en la que se observe la cantidad de vehículos fabricados por cada empresa. Muestre las series de tiempo de los cuatro fabricantes en la misma gráfica.
b. General Motors ha sido sin discusión el principal fabricante de automóviles desde 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 anormalidades 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 anormalidades 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 presentan anormalidades 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 anormalidades en los tejidos?
    e. Estudios médicos a menudo utilizan muestras grandes (en este caso, 3980). ¿Por qué?
22. En otoño de 2003, Arnold Schwarzeneger disputó al gobernador Gray Davis la gobernatura de California. En una encuesta realizada entre los votantes registrados se encontró que Arnold Schwarzeneger 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 en 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

archivo en CD
Shadow02

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.
    a. ¿Cuántas variables hay en este conjunto de datos?
    b. ¿Qué variables son cualitativas y cuáles son cuantitativas?
    c. 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.
    d. 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.
    e. ¿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*

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.

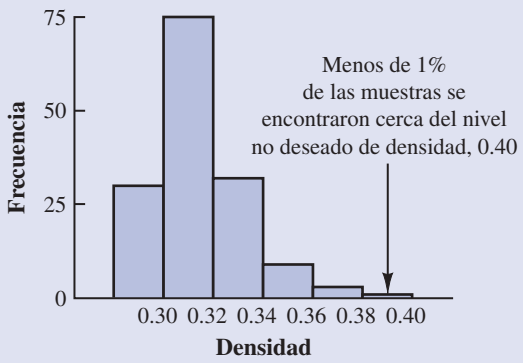
En este capítulo se estudiarán métodos tabulares y gráficos de la estadística descriptiva como distribuciones de frecuencia, gráficas de barras, histogramas, diagramas de tallo y hoja, tabulaciones cruzadas y otros. El objeto de estos métodos es resumir los datos de manera que sean entendibles e interpretables con facilidad.

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

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 sobreponen).

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

archivo en CD
SoftDrink

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 disyuntas. 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

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

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.

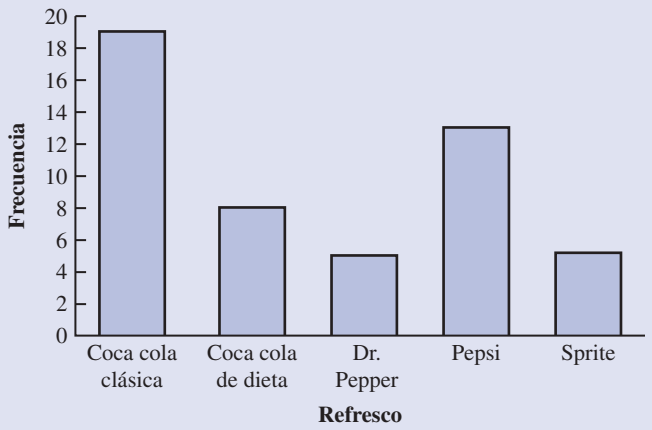
Gráficas de barra y gráficas de pastel

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

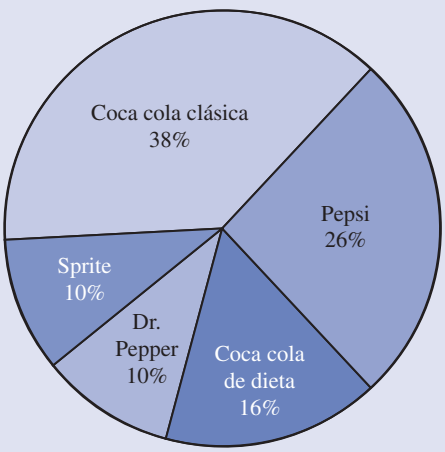


En el control de calidad, las gráficas de barras se usan para identificar las principales causas de problemas. Las graficas se acomodan en orden de alturas descendentes de izquierda a derecha colocando primero la causa de frecuencia más común en primer lugar. A esta gráfica de barras se le llama diagrama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

Autoexamen

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.

Aplicaciones

archivo en CD
TVMedia

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

a. ¿Estos datos son cualitativos o cuantitativos?
b. Proporcione las distribuciones de frecuencia y de frecuencia relativa.
c. Construya una gráfica de barras y una gráfica de pastel.
d. 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.

archivo en CD
Names

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:
a. Distribuciones de frecuencia relativa y porcentual.
b. Una gráfica de barras.
c. Una gráfica de pastel.
d. 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.

archivo en CD
Networks

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

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

b. ¿Cuál o cuáles cadenas de televisión han presentado los programas de mayor índice de audiencia? Compare los desempeños de ABC, CBS y NBC.

Autoexamen

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 fiel (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:

archivo en CD
CEOs

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

a. Diga por qué son cualitativos estos datos.
b. Dé una distribución de frecuencias y una distribución de frecuencia relativa.
c. Dé una gráfica de barras.
d. 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

1. Determinar el número de clases disyuntas.
2. Determinar el ancho de cada clase
3. Determinar los límites de clase.

archivo en CD
Audit

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.

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

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.

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

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

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.

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.

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 disyuntas. 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.

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

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 33–34. 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

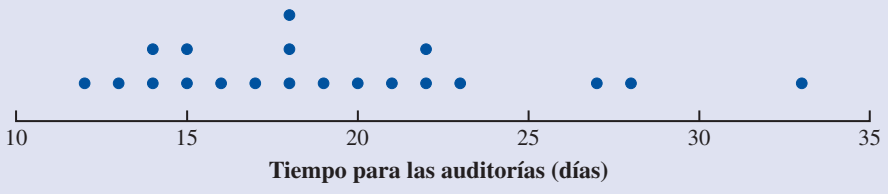
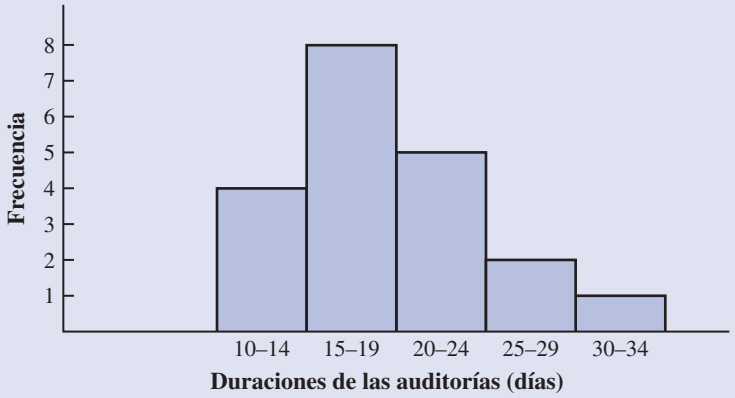


FIGURA 2.4 HISTOGRAMA DE LOS DATOS DE LAS DURACIONES DE LAS AUDITORÍAS



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.

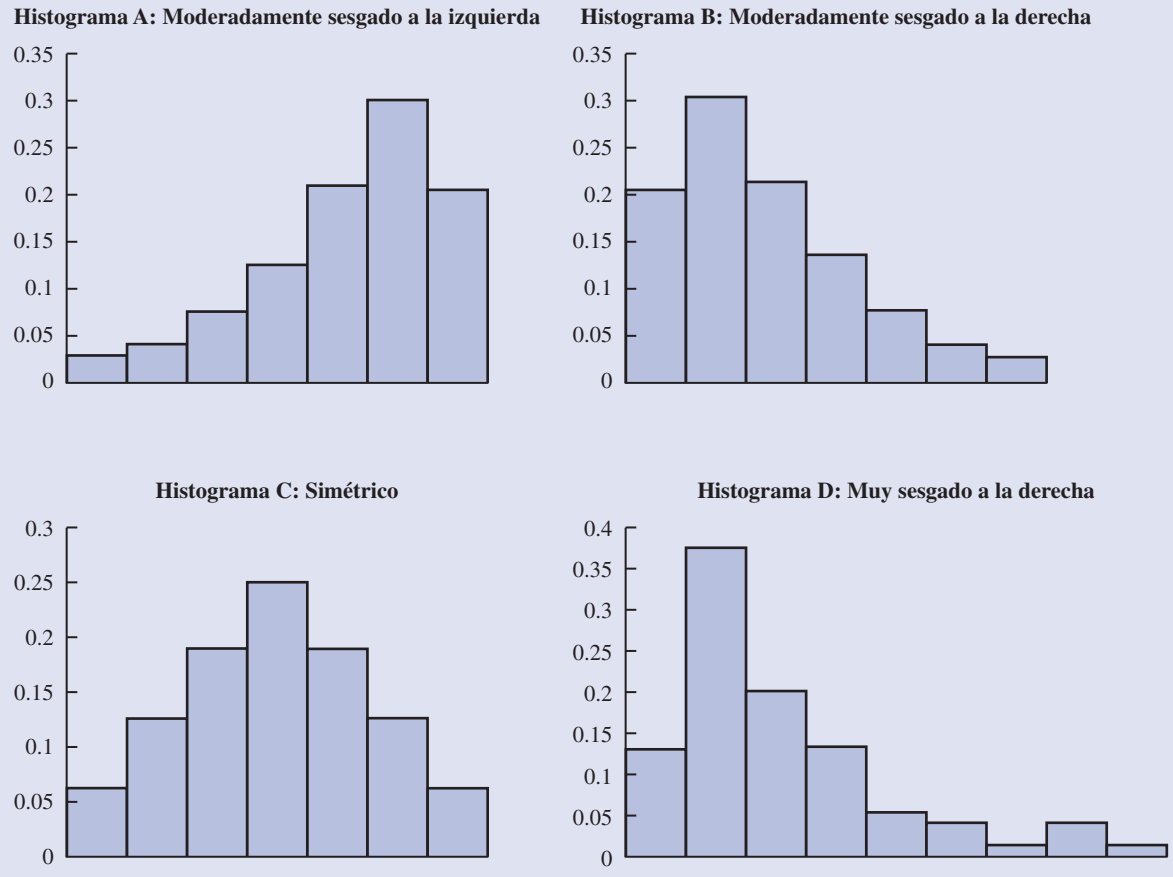
Uno de los usos más importantes de un histograma es proveer información acerca de la forma de la distribución. En la figura 2.5 se muestran cuatro histogramas construidos a partir de distribuciones de frecuencia relativa. En el histograma A se muestra un conjunto de datos moderadamente sesgado a la izquierda. 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 ditribución 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.

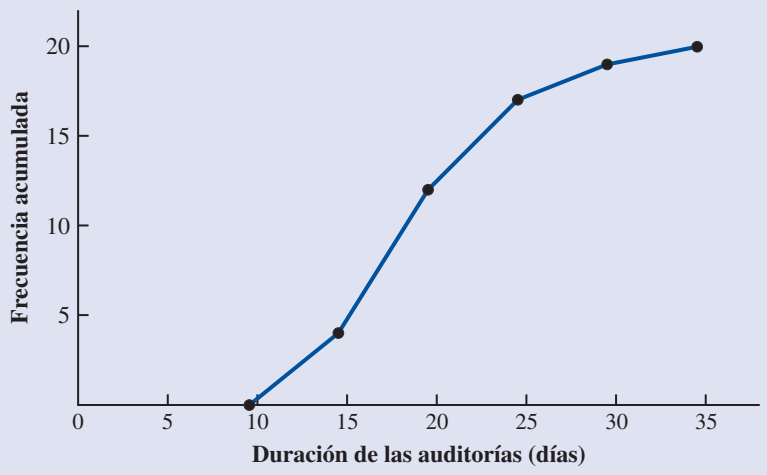
Por último, se tiene que la **distribución de frecuencias relativas acumuladas** indica la proporción de todos los datos que tienen valores menores o iguales al límite superior de cada clase, y la **distribución de frecuencias porcentuales acumuladas** indica el porcentaje de todos los datos que tienen 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

archivo en CD
Frequency

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.

Autoexamen

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

a. Construya un diagrama de punto.
b. Elabore una distribución de frecuencia.
c. 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.
a. Muestre la distribución de la frecuencia.
b. Exprese la distribución de la frecuencia relativa.
c. Muestre la distribución de frecuencia acumulada.
d. Presente la distribución de frecuencia relativa acumulada.
e. ¿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

a. Con los datos del ingreso anual elabore un histograma. ¿Qué evidencia de sesgo observa? ¿Es razonable este sesgo? Explique.
b. Con los datos de las calificaciones elabore un histograma. ¿Qué evidencia de sesgo observa? Explique.
c. 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).

archivo en CD
PriceShare

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

a. Con estos datos elabore una distribución de frecuencia.
b. 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.
c. ¿Cuáles son las acciones que tienen el precio más alto y el más bajo?
d. 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.

archivo en CD
Holiday

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

a. ¿Cuál es la menor cantidad gastada en vacaciones? ¿Cuál la mayor?
b. Use $250 como amplitud de clase para elaborar con estos datos una distribución de frecuencia y una distribución de frecuencia porcentual.
c. Elabore un histograma y comente la forma de la distribución.
d. ¿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:
a. Una distribución de frecuencia (con las clases 1–5, 6–10, 11–15, 16–20, etc.)
b. Una distribución de frecuencia relativa
c. 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).

archivo en CD
Concerts

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.

archivo en CD
Computer

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

archivo en CD
ApTest

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 5 4
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.

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

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

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

Autoexamen

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

Autoexamen

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 (*AAII 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.

archivo en CD
Broker

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

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			

a. Con estos datos elabore un diagrama de tallo y hojas.
b. 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.

archivo en CD
Marathon

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

a. Realice un diagrama expandido de tallo y hojas.
b. ¿En qué grupo de edad hubo más participantes?
c. ¿Qué edad se presenta con más frecuencia?
d. 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

archivo en CD
Restaurant

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

	Precio				
Calidad	**$10 –19**	**$20 –29**	**$30 –39**	**$40 – 49**	**Total**
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

	Precio				
Calidad	**$10–19**	**$20–29**	**$30–39**	**$40–49**	**Total**
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.

	Juez		
Sentencia	**Luckett**	**Kendall**	**Total**
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		
Sentencia	**Tribunal de primera instancia**	**Tribunal municipal**	**Total**
Ratificada	29 (91%)	100 (85%)	129
Revocada	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

	Juez Kendall		
Sentencia	**Tribunal de primera instancia**	**Tribunal municipal**	**Total**
Ratificada	90 (90%)	20 (80%)	110
Revocada	10 (10%)	5 (20%)	15
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

archivo en CD
Stereo

Semana	Número de comerciales x	Ventas (en cientos de dólares) 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

*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

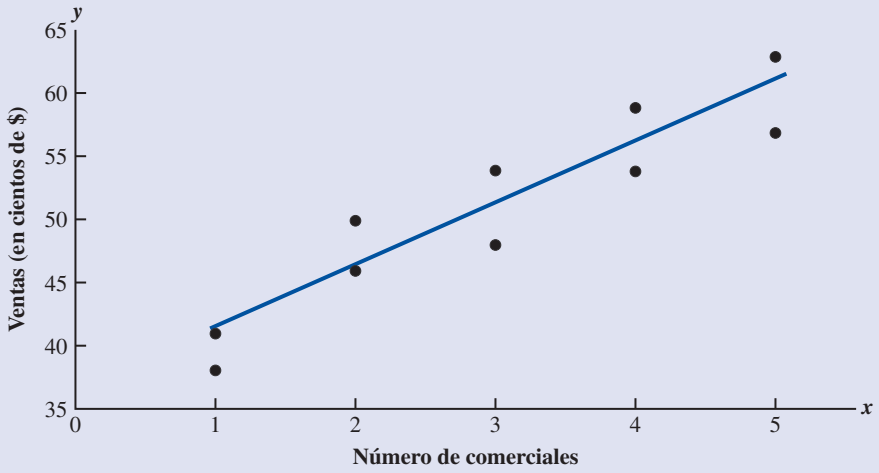
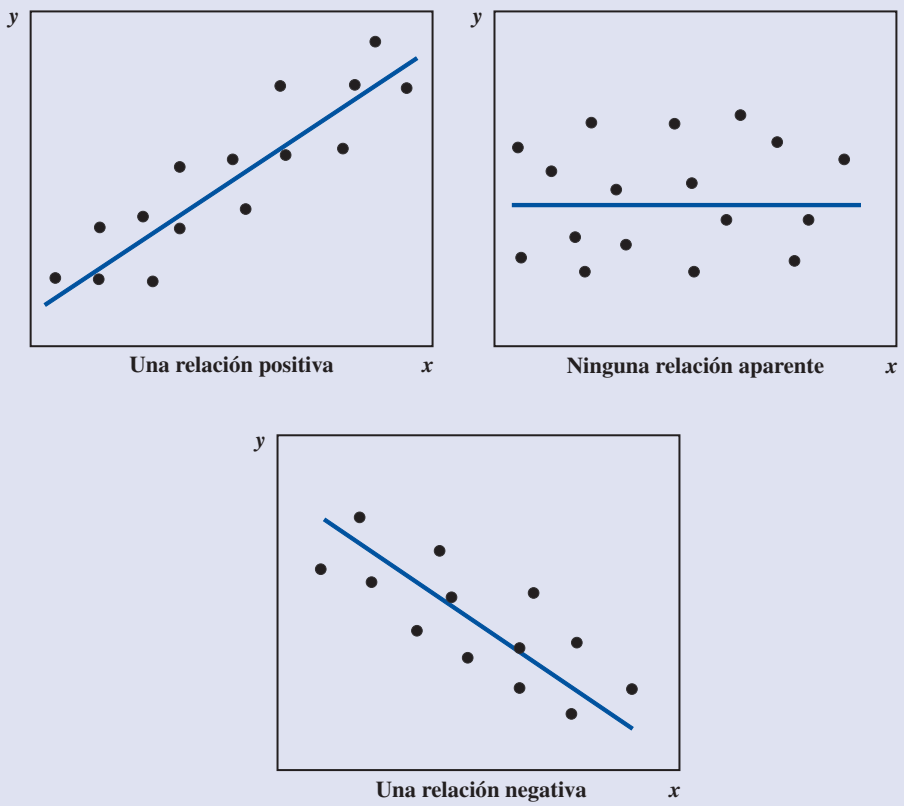


FIGURA 2.8 TIPOS DE RELACIÓN QUE APARECEN EN LOS DIAGRAMAS DE DISPERSIÓN



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

Autoexamen

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.

archivo en CD
Crosstab

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

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

Autoexamen

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

archivo en CD
Scatter

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

a. Elabore un diagrama de dispersión para la relación entre x y y.
b. ¿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*).

	Ingreso por familia (en miles de dólares)					
Nivel de estudios	**Menos de 25**	**25.0–49.9**	**50.0–74.9**	**75.0–99.9**	**100 o más**	**Total**
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

a. 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.
b. ¿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?
c. 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.
a. Calcule los porcentajes e identifique las distribuciones de frecuencia porcentual. ¿Qué porcentaje de jefes de familia no terminó la secundaria?
b. ¿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?
c. 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

	Condición de los *greens*	
Hándicap	**Demasiado rápido**	**Bien**
Menos de 15	10	40
15 o más	25	25

Jugadoras

	Condición de los greens	
Hándicap	**Demasiado rápido**	**Bien**
Menos de 15	1	9
15 o más	39	51

a. 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.

archivo en CD IBD

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.

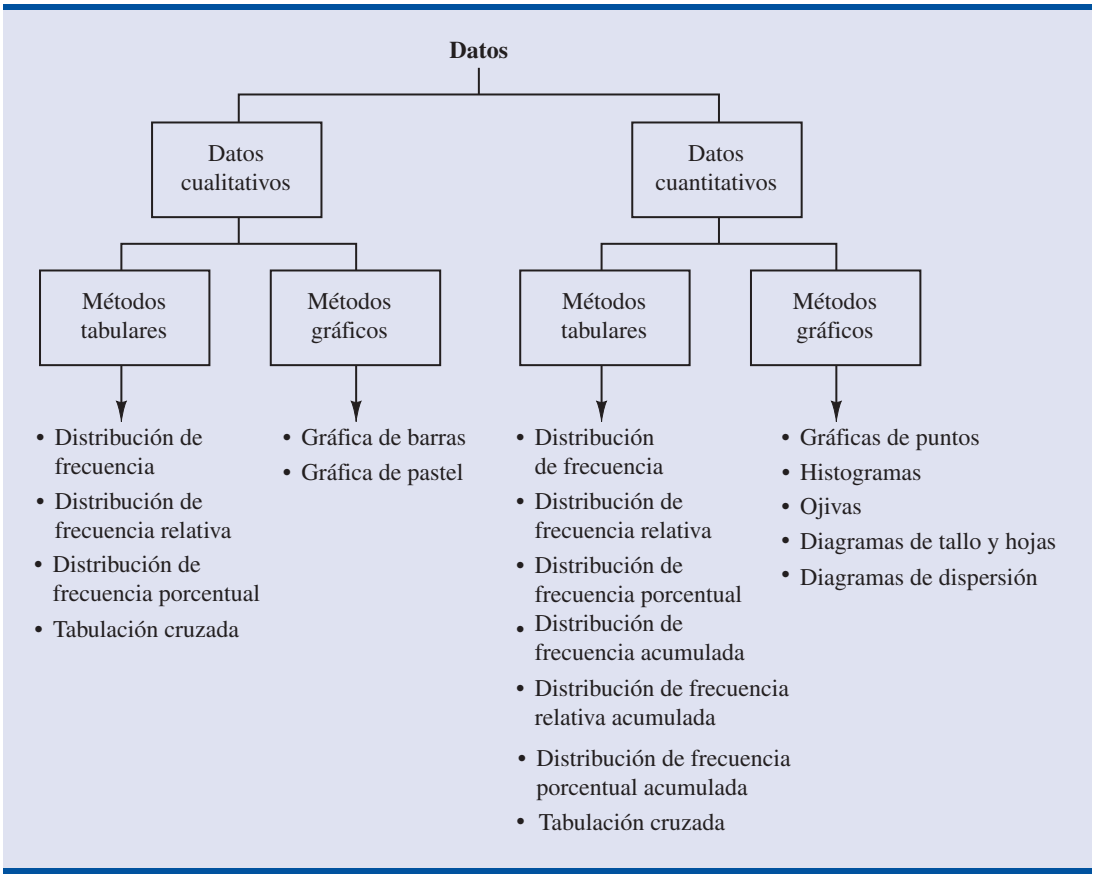
Cuando se tienen grandes conjuntos de datos es indispensable usar paquetes de software para la elaboración de resúmenes tabulares o gráficos de los datos. En los dos apéndices de este capítulo se explica el uso de Minitab y de Excel con tal propósito.

archivo en CD
NFL

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 McDougle	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

$$\frac{\text{Frecuencia de la clase}}{n} \tag{2.1}$$

Ancho aproximado de clase

$$\frac{\text{Dato mayor} - \text{Dato menor}}{\text{Número de clases}} \tag{2.2}$$

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.

archivo en CD
AutoData

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.

archivo en CD
Major

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):

archivo en CD
Golf

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 (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.

archivo en CD
DivYield

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.

archivo en CD
SATScores

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

archivo en CD
Shadow

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.

a. Con los datos de bolsa de cambio haga una distribución de frecuencia y otra de frecuencia relativa. ¿Cuál tiene más acciones fantasma?
b. 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).

archivo en CD
Population

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		

a. Elabore una distribución de frecuencia, una de frecuencia porcentual y un histograma. Use como ancho de clase 2.5 millones.
b. Explique el sesgo de la distribución.
c. ¿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

a. Dé un diagrama de tallo y hojas.
b. Indique cuáles son las ventas anuales menores, mayores e intermedias.
c. ¿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).

archivo en CD
CityTemp

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

a. Con las temperaturas altas elabore un diagrama de tallo y hojas.
b. Con las temperaturas bajas elabore un diagrama de tallo y hojas.
c. Compare los dos diagramas y haga comentarios acerca de las diferencias entre las temperaturas más altas y las más bajas.
d. 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.
a. Elabore un diagrama de dispersión que muestre la relación entre las dos variables, temperatura más alta y temperatura más baja.
b. 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

archivo en CD
OccupSat

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.

a. Dé una tabulación cruzada para ocupación y satisfacción en el trabajo.
b. En la tabulación cruzada del inciso a calcule los porcentajes de renglones.
c. ¿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).

archivo en CD
RevEmps

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

a. Haga un diagrama de dispersión para mostrar la relación entre las variables ingreso y empleados.
b. 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*.

archivo en CD
Fortune

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

archivo en CD
PelicanStores

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 Propietary 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

archivo en CD
Movies

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

archivo en CD
Audit

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

archivo en CD
Audit

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

archivo en CD
ApTest

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

archivo en CD
Stereo

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

archivo en CD
Restaurant

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.

archivo en CD
SoftDrink

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					

Nota: Los renglones 11–44 están ocultos.

	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					

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

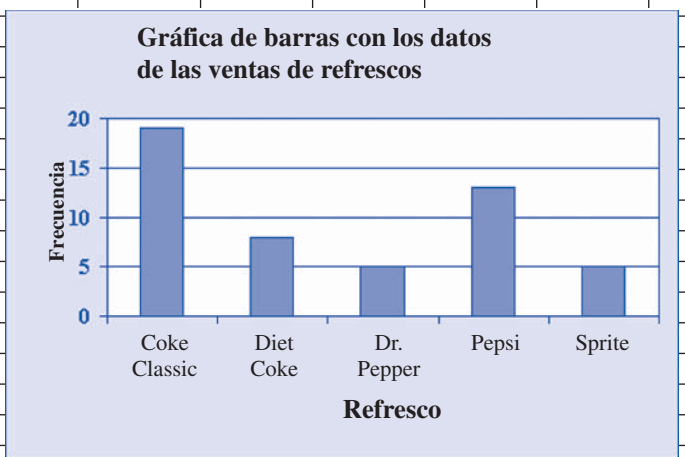
archivo en CD
SoftDrink

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 gráficas. Los pasos a seguir son:

Paso 1. Seleccionar las celdas C1:D6
Paso 2. Hacer clic en el botón **Asistente para gráficas** 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

	A	B	C	D	E	F	G	H	I
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									
53									
54									
55									
56									
57									



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.

archivo en CD
Audit

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.

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

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.

*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

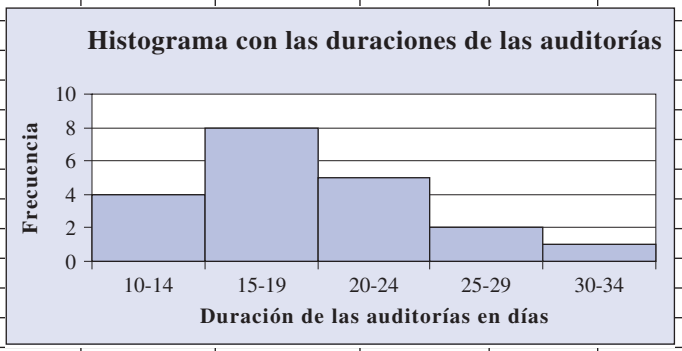
Paso 2. Hacer clic en el botón **Asistente para gráficas** de la barra de herramientas estándar (o seleccionar el menú **Insertar** y elegir la opción **Gráfico**)

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

	A	B	C	D	E	F	G
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						
22							



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.

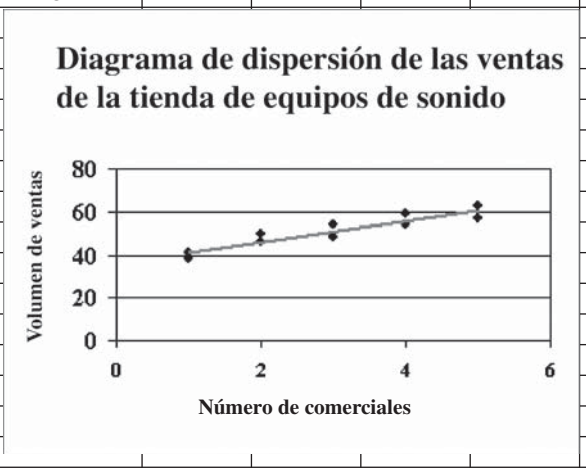
Paso 2. Cuando aparezca el cuadro de diálogo Formato de punto de datos:
Seleccionar la pestaña **Opciones**
Ingresar 0 en el cuadro **Ancho del rango**
Hacer clic en **Aceptar**

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

	A	B	C	D	E	F	G	H
1	Semana	Cantidad de comerciales	Volumen de ventas					
2	1	2	50					
3	2	5	57					
4	3	1	41					
5	4	3	54					
6	5	4	54					
7	6	1						
8	7	5						
9	8	3						
10	9	4						
11	10	2						
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								



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

archivo en CD
Stereo

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 las celda B1:C11

Paso 2. Hacer clic en el botón **Asistente para gráficas** 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

archivo en CD
Restaurant

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

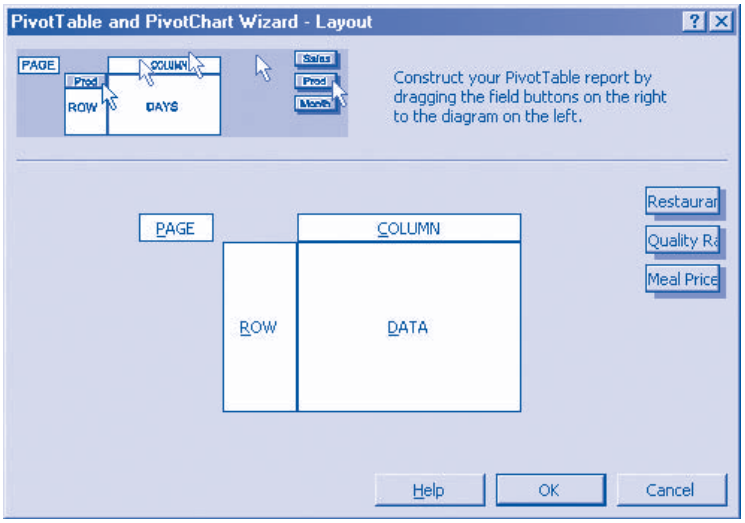


FIGURA 2.17 ASISTENTE PARA TABLAS Y GRÁFICOS DINÁMICOS: DISEÑO

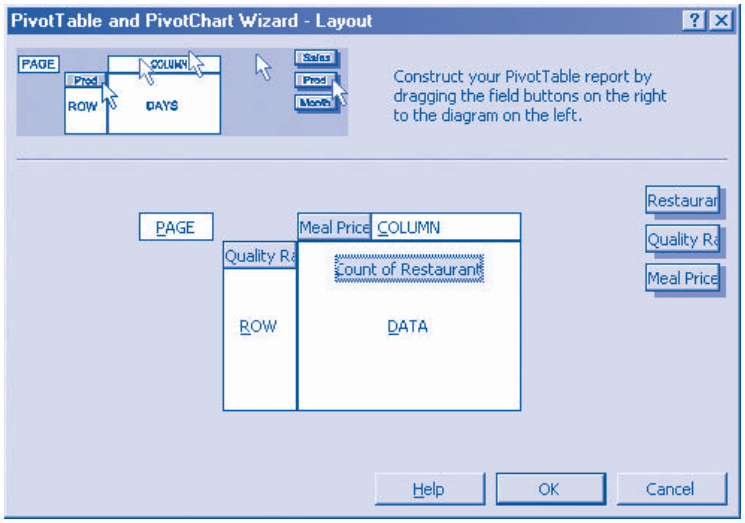


FIGURA 2.18 RESULTADO INICIAL DEL INFORME DE TABLA DINÁMICA (LAS COLUMNAS D:AK ESTÁN OCULTAS)

	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							

PivotTable
PivotTable ▾

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*
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.

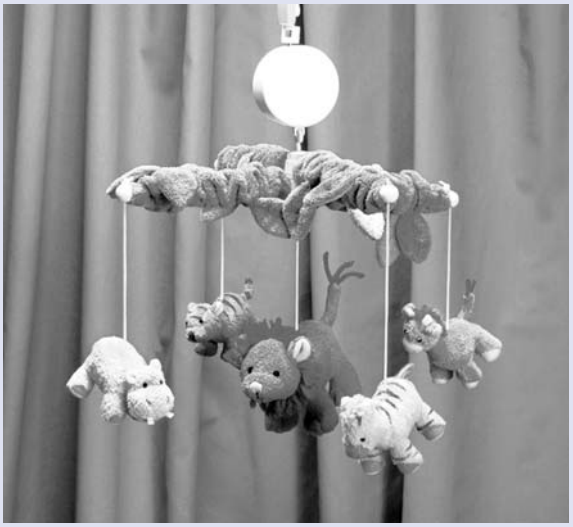
Small Fry Design emplea representantes independientes para la venta de sus productos a tiendas de mobiliario para niños, tiendas de accesorios y ropa para niños, tiendas de regalos, tiendas exclusivas de departamentos e importantes empresas de ventas por catálogo. En la actualidad los productos de Small 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

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad (3.1)$$

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

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

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 54 42 46 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.

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

Por tanto, para calcular la media muestral, escriba

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

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

archivo en CD
StartSalary

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

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

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$
$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$
$$= \frac{42\,480}{12} = 3540$$

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

$$\mu = \frac{\Sigma x_i}{N} \qquad \textbf{(3.2)}$$

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 replantea así:

MEDIANA

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

a. Si el número de observaciones es impar, la mediana es el valor de enmedio.
b. 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

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

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

$$3310 \quad 3355 \quad 3450 \quad 3480 \quad 3480 \quad \underbrace{3490 \quad 3520}_{\text{Los dos valores de en medio}} \quad 3540 \quad 3550 \quad 3650 \quad 3730 \quad 3925$$

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

Seguir estos pasos facilita el cálculo de los percentiles.

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 $i + 1$.

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

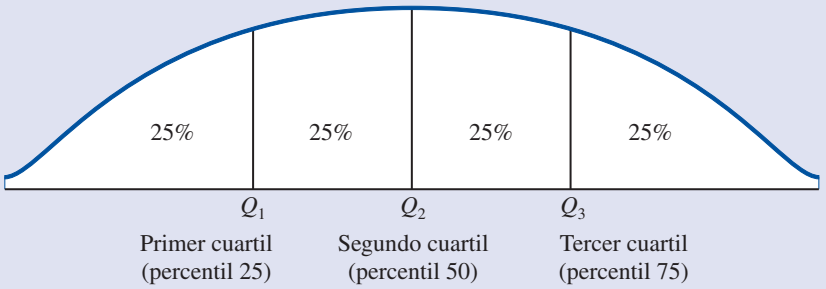
Paso 2.

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

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

$Q_1 = 3465$ $Q_2 = 3505$ (Mediana) $Q_3 = 3600$

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 obtiene 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

1. Los valores de los datos en una muestra son 10, 20, 12, 17 y 16. Calcule la media y la mediana.
2. Los datos en una muestra son 10, 20, 21, 17, 16 y 25. Calcule la media y la mediana.
3. Los valores en una muestra son 27, 25, 20, 15, 30, 34, 28 y 25. Calcule los percentiles 20, 25, 65 y 75
4. 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.

Autoexamen

Aplicaciones

5. 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

archivo en CD
Hotels

a. ¿Cuál es la media en el precio de estas habitaciones?
b. ¿Cuál es la mediana en el precio de estas habitaciones?
c. ¿Cuál es la moda?
d. ¿Cuál es el primer cuartil?
e. ¿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):

archivo en CD
BASalary

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

a. Para cada uno de los grupos de sueldos iniciales calcule moda, mediana y media.
b. Para cada uno de los grupos de sueldos iniciales calcule el primer y el tercer cuartil.
c. 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.
a. 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.
b. Calcule la media, mediana y moda de las comisiones que se cobran por una transacción en línea de 500 acciones a $50 por acción.
c. ¿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?
d. ¿Está relacionado el costo de la transacción con el monto de la transacción?

TABLA 3.2 COMISIONES QUE COBRAN LOS CORREDORES DE BOLSA

archivo en CD
Broker

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

   a. Calcule la media y la moda.
   b. 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.
   c. Calcule el primer y el tercer cuartil.
   d. 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

   a. ¿Cuál es la media de los minutos de uso por mes?
   b. ¿Cuál es la mediana de los minutos de uso por mes?
   c. ¿Cuál es el percentil 85?
   d. 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		

   a. Calcule la media y la mediana de estos tiempos de espera en los hospitales a toda capacidad.
   b. Calcule la media y la mediana de estos tiempos de espera en los hospitales en equilibrio.
   c. 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.

archivo en CD
Disney

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

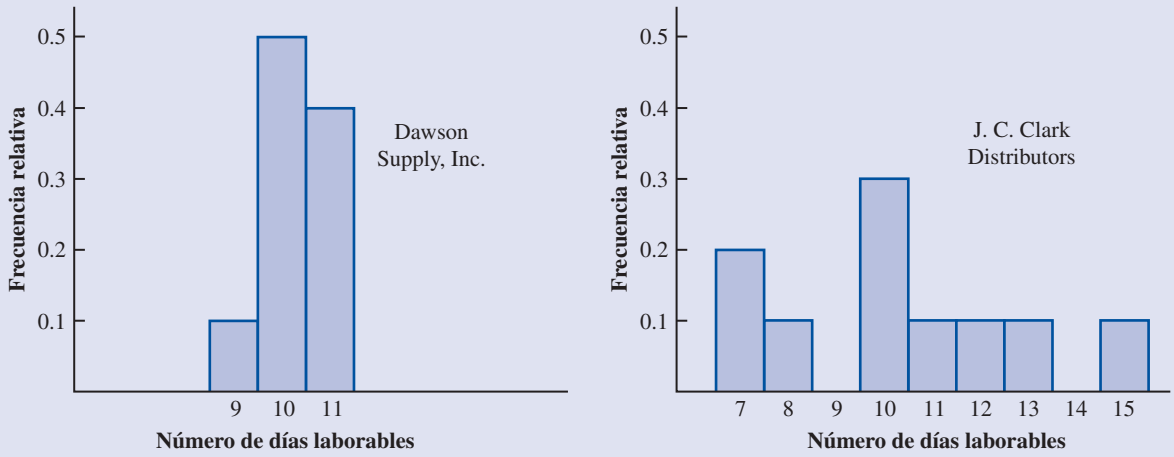
3.2 Medidas de variabilidad

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

$$\text{IQR} = Q_3 - Q_1 \tag{3.3}$$

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

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

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

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

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 $\Sigma(x_i - \bar{x})^2 = 256$. Por tanto, siendo $n - 1 = 4$, la varianza muestral es

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

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 $\Sigma(x_i - \bar{x}) = 0$. 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

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

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

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.

Coeficiente 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

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

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 (*Bussines-Week,* 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

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

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

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.

Autoexamen

Aplicaciones

Autoexamen

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?

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

   a. Calcule la media, la varianza y la desviación estándar de estas tarifas.
   b. 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.
   a. Calcule el rango y el rango intercuartílico.
   b. Calcule la varianza muestral y la desviación estándar muestral.
   c. 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

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

archivo en CD
Broker

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.
    a. Calcule el rango y el rango intercuartílico en cada tipo de transacción.
    b. Calcule la varianza y la desviación estándar en cada tipo de transacción.
    c. Calcule el coeficiente de variación en cada tipo de transacción.
    d. 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

    a. Use la media y la desviación estándar para evaluar a este jugador de golf en estos dos años.
    b. ¿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).

Tiempos en un cuarto de milla:	0.92	0.98	1.04	0.90	0.99
Tiempos en una milla:	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**.

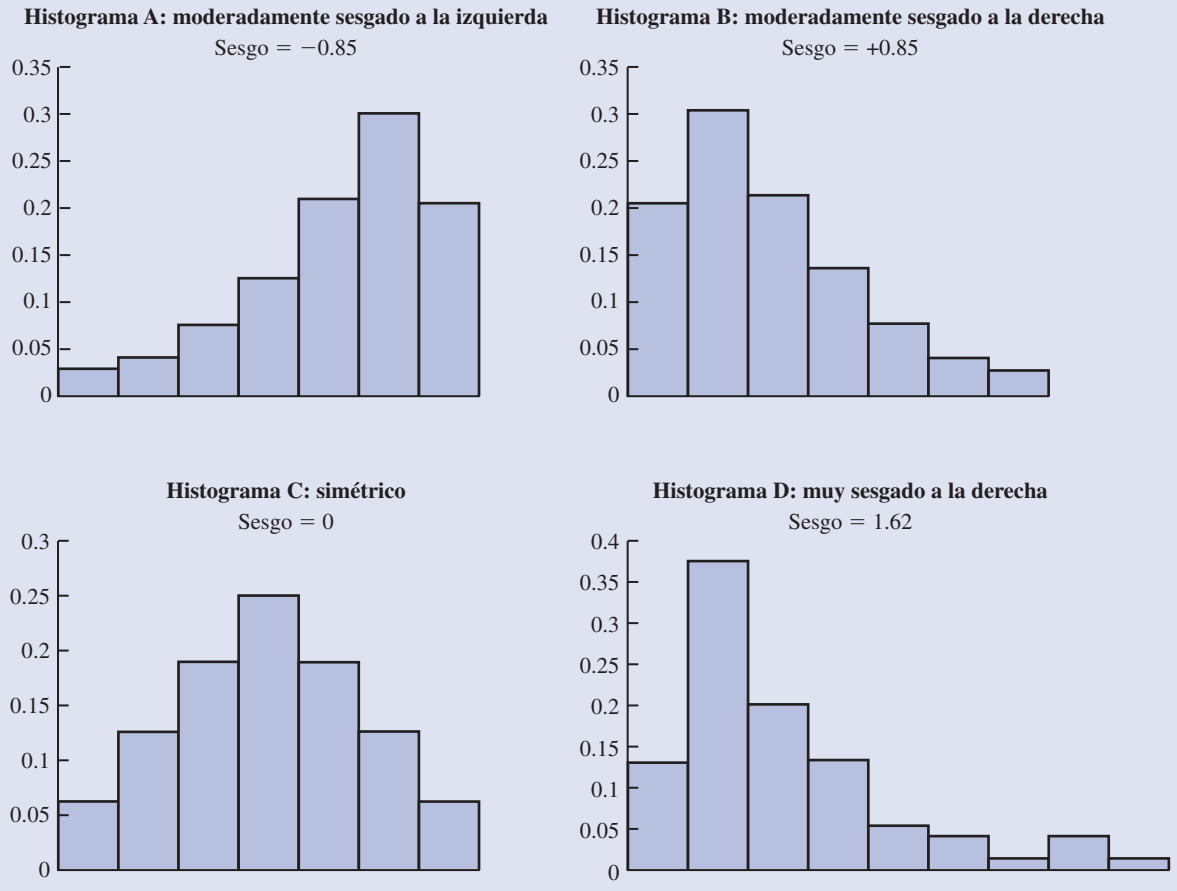
Forma de la distribución

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 $x_1, x_2, \ldots, x_n$. 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

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (3.9)$$

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

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

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

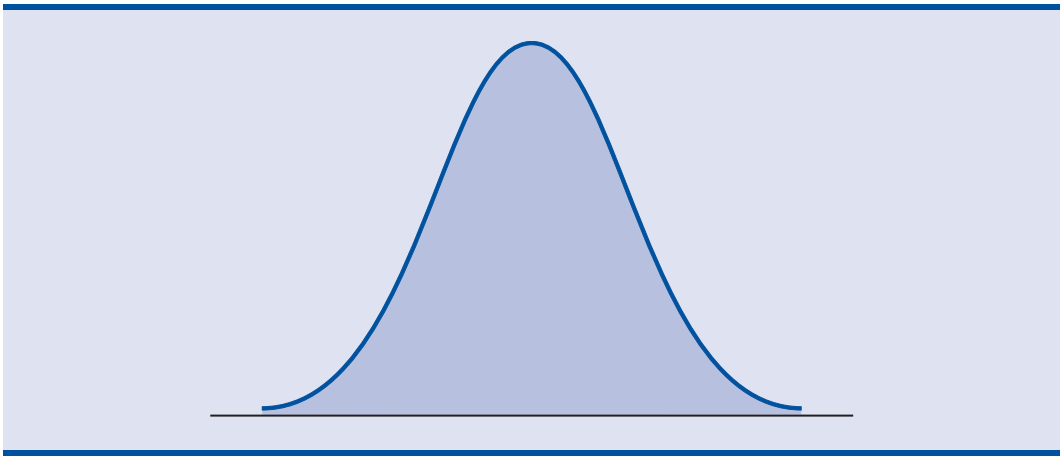
Una de las ventajas del teorema de Chebyshev es que se aplica a cualquier conjunto de datos, sin importar la forma de la distribución de los datos. En efecto se usa para cualquiera de las distribuciones de la figura 3.3. Sin embargo, en muchas aplicaciones prácticas los datos muestran una distribución 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.

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.

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.

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.

NOTAS Y COMENTARIOS

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

Autoexamen

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

Autoexamen

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.

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 Vilma 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).

archivo en CD
NCAA

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

a. Calcule la media y la desviación estándar de los puntos obtenidos por los equipos ganadores.
b. 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.
c. 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.

archivo en CD
Speakers

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 Aerius	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

a. Calcule la media y la mediana.
b. Aproxime el primer y el tercer cuartil.
c. Estime la desviación estándar.
d. El sesgo de estos datos es -1.67. Comente la forma de esta distribución.
e. Calcule los puntos z correspondientes a Allison One y a Ommi Audio
f. ¿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.

1. El valor menor.
2. El primer cuartil (Q_1).
3. 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.

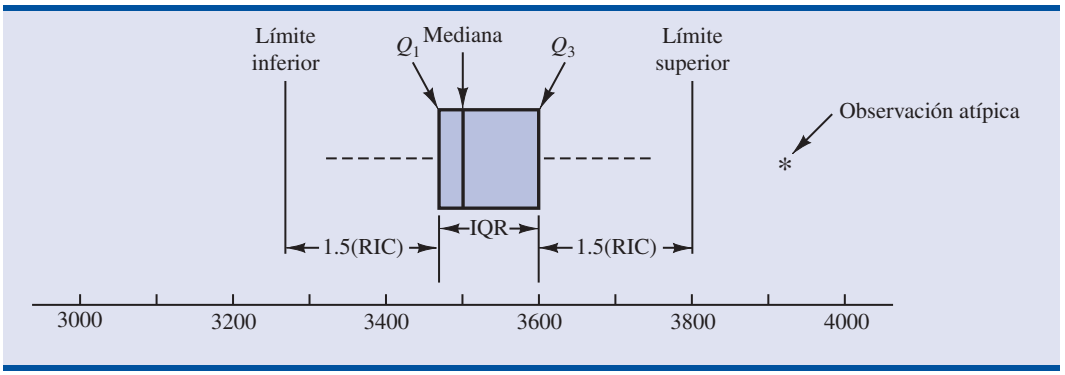
Diagrama de caja

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.

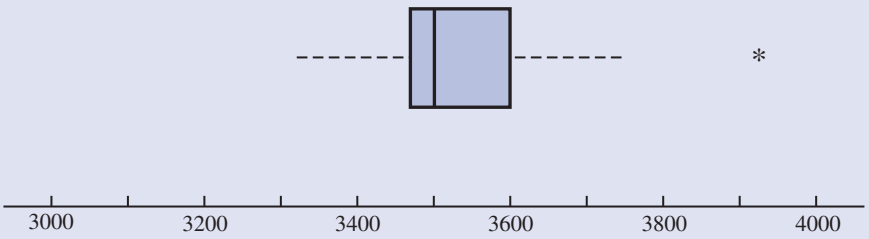
Los diagramas de caja proporcionan otra manera de identificar observaciones atípicas. Pero no necesariamente se identifican los mismos valores que los correspondientes a un punto z menor que −3 o mayor que +3. Puede emplear cualquiera de estos procedimientos, o los dos.

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?

Autoexamen

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

archivo en CD Property

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

a. Muestre el resumen de cinco números.
b. Calcule los límites inferior y superior.
c. 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.
d. La segunda propiedad más cara que aparece en la lista es de $3 100 000, ¿debe considerarse como valor atípico? Explique.
e. Dibuje el diagrama de caja.

Autoexamen

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			

a. Proporcione el resumen de cinco números.
b. Calcule los límites superior e inferior.
c. ¿Hay alguna observación atípica en estos datos?
d. 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?
e. 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).

archivo en CD
Baseball

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

a. ¿Cuál es la mediana de la nómina?
b. Proporcione el resumen de cinco números.
c. ¿Es una observación atípica la nómina de $208 millones de los Yankees de Nueva York? Explique.
d. 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.

archivo en CD
Mutual

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

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \qquad \textbf{(3.10)}$$

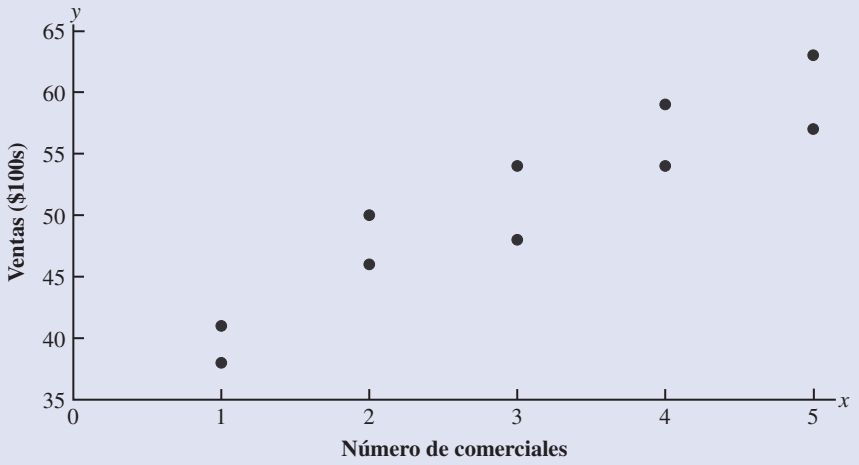
Esta fórmula aparea cada x_i con una y_i. Después se suman los productos obtenidos al multiplicar la desviación de cada x_i de su media muestral $\bar{x}$ por la desviación de la y_i correspondiente de su media muestral $\bar{y}$; esta suma se divide entre $n - 1$.

archivo en CD
Stereo

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 $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. 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

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

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

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

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

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

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.

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

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.

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

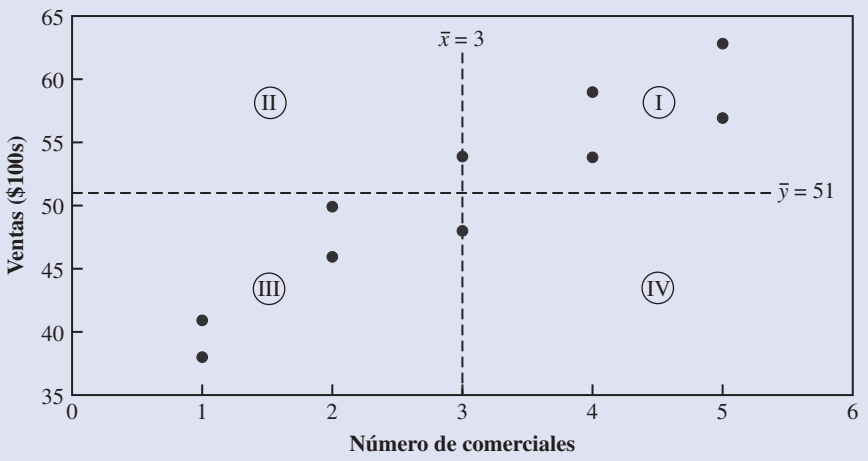
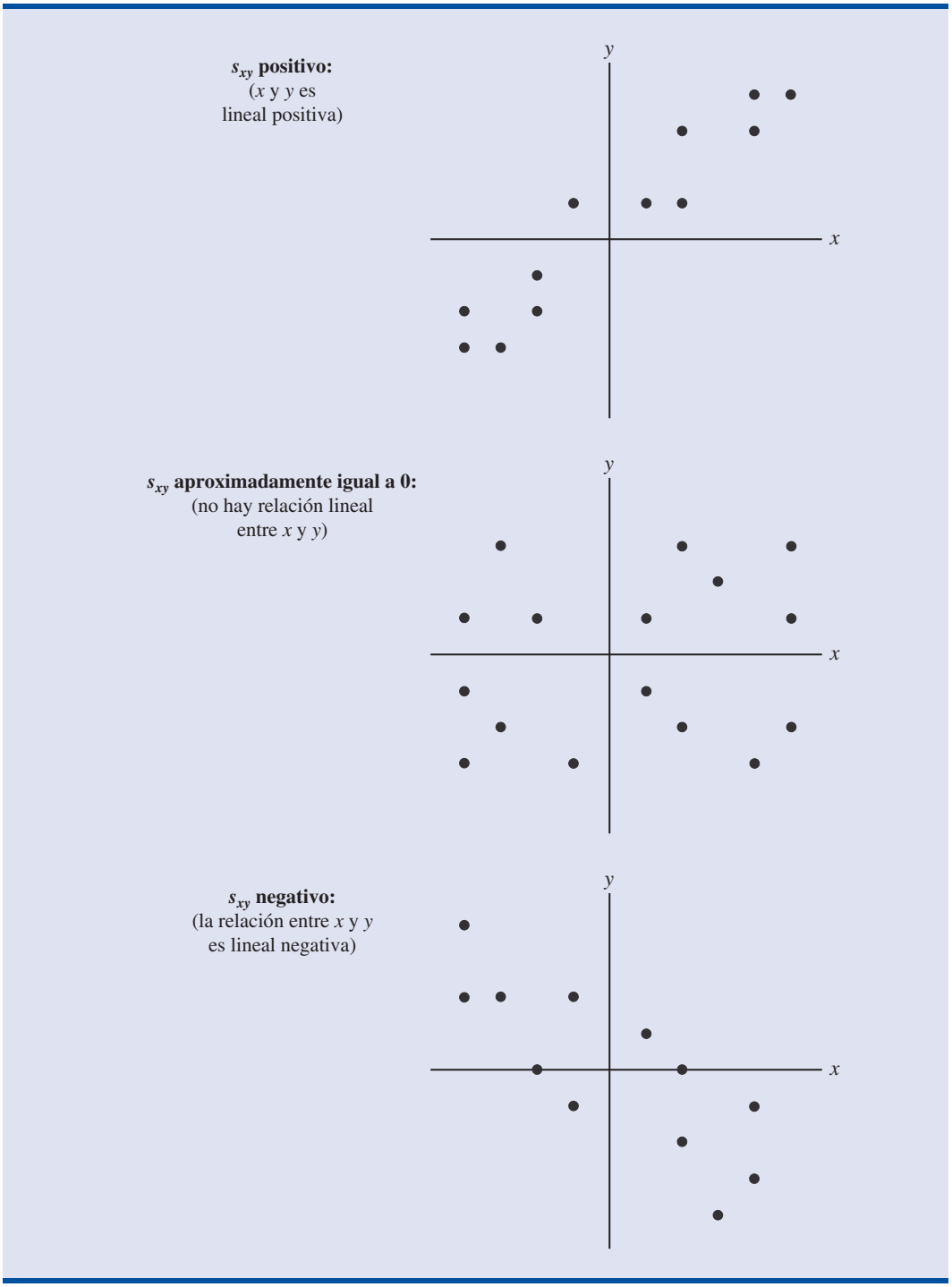


FIGURA 3.9 INTERPRETACIÓN DE LA COVARIANZA MUESTRAL



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 $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ 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

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

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.

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

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

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

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

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

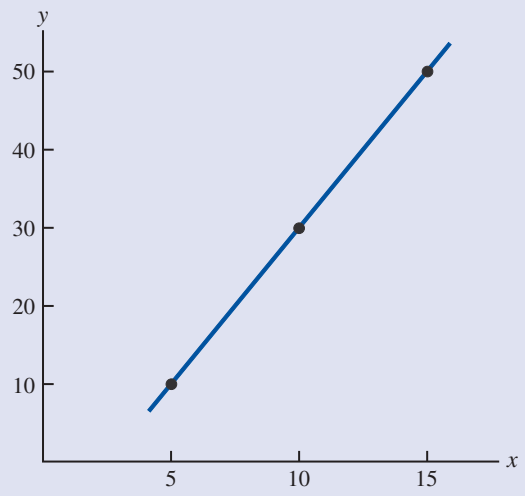


TABLA 3.9 CÁLCULOS PARA OBTENER EL COEFICIENTE DE CORRELACIÓN MUESTRAL

	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

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

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

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

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

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

a. Elabore un diagrama de dispersión con x en el eje horizontal.
b. ¿Qué indica el diagrama de dispersión elaborado en el inciso a respecto a la relación entre las dos variables?
c. Calcule e interprete la covarianza muestral.
d. 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

a. Elabore un diagrama de dispersión con estas variables.
b. ¿Qué indica este diagrama de dispersión respecto de la relación entre x y y?
c. Calcule e interprete la covarianza muestral.
d. 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

a. Elabore un diagrama de dispersión con los *ratings* en el eje horizontal.
b. ¿Cuál es la relación entre *rating* y *share*? Explique.
c. Calcule e interprete la covarianza muestral.
d. 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

archivo en CD
PCs

Notebook	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

a. Calcule el coeficiente de correlación muestral.
b. ¿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).

archivo en CD
StockMarket

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

a. Muestre el diagrama de dispersión.
b. Calcule el coeficiente de correlación muestral de estos datos.
c. 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.)

archivo en CD
Temperature

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

a. ¿Cuál es la media muestral de las temperaturas diarias más elevadas?
b. ¿Cuál es la media muestral de las temperaturas diarias más bajas?
c. ¿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.

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (3.14)$$

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

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \qquad (3.15)$$

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:

$$\bar{x} = \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800}$$

$$= \frac{18\,500}{6250} = 2.96$$

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.

El cálculo de las calificaciones es un buen ejemplo del uso de la media ponderada.

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.

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, $\Sigma f_i = n$. De manera que la ecuación para la media muestral de datos agrupados es la siguiente:

MEDIA MUESTRAL DE DATOS AGRUPADOS

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

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 $\Sigma f_i(M_i - \bar{x})^2$. 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

$$s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

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
		20	380

Media muestral $\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19$ 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
					$\Sigma f_i(M_i - \bar{x})^2$

$$\text{Varianza muestral } s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$$

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

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

VARIANZA POBLACIONAL DE DATOS AGRUPADOS

$$\sigma^2 = \frac{\Sigma f_i(M_i - \mu)^2}{N} \tag{3.19}$$

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

a. Calcule la media ponderada.
b. 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.

Autoexamen

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

a. Calcule la media muestral.
b. Calcule la varianza muestral y la desviación estándar muestral.

Aplicaciones

Autoexamen

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.
a. Calcule el promedio de calificaciones de este estudiante.
b. 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

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.
Coeficiente 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

$$\bar{x} = \frac{\Sigma x_i}{n} \tag{3.1}$$

Media poblacional

$$\mu = \frac{\Sigma x_i}{N} \tag{3.2}$$

Rango intercuartílico

$$\text{RIC} = Q_3 - Q_1 \tag{3.3}$$

Varianza poblacional

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

Varianza muestral

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

Desviación estándar

$$\text{Desviación estándar muestral} = s = \sqrt{s^2} \tag{3.6}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2} \tag{3.7}$$

Coeficiente de variación

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

Punto z

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

Covarianza muestral

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{3.10}$$

Covarianza poblacional

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

Coeficiente de correlación del producto–momento de Pearson: datos muestrales

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{3.12}$$

Coeficiente de correlación del producto–momento de Pearson: datos poblacionales

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.13}$$

Media ponderada

$$\bar{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} \tag{3.15}$$

Media muestral de datos agrupados

$$\bar{x} = \frac{\Sigma f_i M_i}{n} \tag{3.16}$$

Varianza muestral de datos agrupados

$$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n - 1} \tag{3.17}$$

Media poblacional de datos agrupados

$$\mu = \frac{\Sigma f_i M_i}{N} \tag{3.18}$$

Varianza poblacional de datos agrupados

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} \tag{3.19}$$

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.

archivo en CD
Visa

236	1710	1351	825	7450
316	4135	1333	1584	387
991	3396	170	1428	1688

a. Calcule la media y la mediana.
b. Calcule el primero y tercer cuartil.
c. Calcule el rango y el rango intercuartílico.
d. Calcule la varianza y la desviación estándar.
e. 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?
f. ¿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.

archivo en CD
Ages

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		

a. Determine la mediana en la edad de hombres y mujeres al contraer el primer matrimonio.
b. Calcule el primer y tercer cuartil tanto en los hombres como en las mujeres.
c. 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

a. ¿Cuáles son la media y mediana de los rendimientos de dividendos?
b. ¿Cuál es la varianza y la desviación estándar?
c. ¿Qué empresa proporciona el mayor rendimiento de dividendos?
d. ¿Cuál es el punto z correspondiente a McDonalds? Interprete este punto z.
e. ¿Cuál es el punto z de General Motors? Interprete este punto z.
f. 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

a. Entre los estudiantes que toman un préstamo estudiantil, ¿cuál es la mediana en la deuda que tienen una vez terminados sus estudios?
b. ¿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.

archivo en CD
Penalty

820 270 450 1010 890 700 1350 350 300 1200
390 730 2040 230 640 350 420 270 370 620

a. ¿Cuál es la media en multas?
b. ¿Cuál es la desviación estándar?
c. ¿Es una observación atípica la multa más alta, $2040?
d. ¿Cuáles son algunas de las ventajas que tienen los propietarios de los negocios pequeños al contratar una empresa de servicio de pago de nomina 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

a. Calcule la media muestral en el tiempo que se necesita con cada transporte.
b. Calcule la desviación estándar para cada transporte.
c. De acuerdo con los resultados en los incisos a y b ¿cuál será el medio de transporte preferido? Explique.
d. 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.

archivo en CD
Homes

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

a. ¿Cuál es la mediana muestral de los precios de las casas?
b. 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?
c. ¿Cuáles son el primer y tercer cuartiles de los datos muestrales?
d. Dé el resumen de cinco números para los precios de las casas.
e. ¿Existe alguna observación atípica en los datos?
f. ¿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.

archivo en CD
Beer

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
	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 Propietary 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.

archivo en CD
PelicanStores

TABLA 3.14 MUESTRA 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
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

archivo en CD
Movies

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
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

Caso problema 3 Las escuelas de negocios de Asia–Pacífico

archivo en CD
Asian

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	Colegiatura para estudiantes locales ($)	Colegiatura para estudiantes de fuera ($)	Edad	% de extranjeros	GMAT	Examen de inglés	Experiencia laboral	Salario inicial ($)
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 (Niigata)	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.

$$\begin{aligned}\text{Rango} &= \text{Máximo} - \text{Mínimo}\\ \text{RIC} &= Q_3 - Q_1\\ \text{Varianza} &= (\text{StDev})^2\\ \text{Coeficiente de variación} &= (\text{StDev/Media}) \times 100\end{aligned}$$

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.

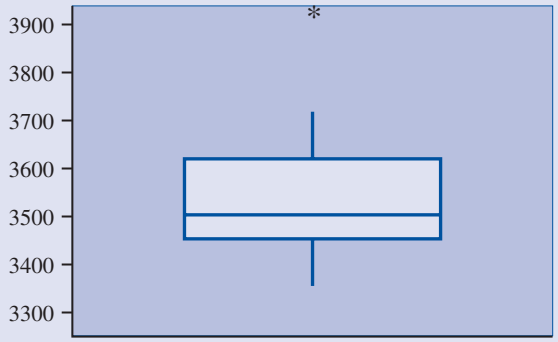
FIGURA 3.11 ESTADÍSTICOS DESCRIPTIVOS Y DIAGRAMA DE CAJA PROPORCIONADOS POR MINITAB

Panel A: Descriptive Statistics

N	N*	Mean	SEMean	StDev
12	0	3540.0	47.8	165.7

Minimum	Q1	Median	Q3	Maximum
3310.0	3457.5	3505.0	3625.0	3925.0

Panel B: Box Plot



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 al columna C2 de la hoja de cálculo de Minitab. Para generar los estadísticos descriptivos realice los pasos siguientes:

archivo en CD
StartSalary

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.

archivo en CD
Stereo

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

archivo en CD
StartSalary

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:

$$=\text{AVERAGE(B2:B13)}$$

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.

archivo en CD
Stereo

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						

archivo en CD
StartSalary

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.