

# CAPÍTULO 3



## Estadística descriptiva: medidas numéricas

### CONTENIDO

LA ESTADÍSTICA EN LA PRÁCTICA:  
*SMALL FRY DESIGN*

#### 3.1 MEDIDAS DE LOCALIZACIÓN

Media  
Mediana  
Moda  
Percentiles  
Cuartiles

#### 3.2 MEDIDAS DE VARIABILIDAD

Rango  
Rango intercuartílico  
Varianza  
Desviación estándar  
Coeficiente de variación

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

Forma de la distribución  
Puntos  $z$   
Teorema de Chebyshev

Regla empírica

Detección de observaciones atípicas

#### 3.4 ANÁLISIS EXPLORATORIO DE DATOS

Resumen de cinco números  
Diagrama de caja

#### 3.5 MEDIDAS DE ASOCIACIÓN ENTRE DOS VARIABLES

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

#### 3.6 LA MEDIA PONDERADA Y EL EMPLEO DE DATOS AGRUPADOS

Media ponderada  
Datos agrupados

## LA ESTADÍSTICA *(en)* LA PRÁCTICA

### SMALL FRY DESIGN\*

SANTA ANA, CALIFORNIA

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

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

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

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

Media	40 días
Mediana	35 días
Moda	31 días

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



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

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

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

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

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

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

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

3.1

## Medidas de localización

### Media

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

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

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

#### MEDIA MUESTRAL

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

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

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

La letra griega  $\Sigma$  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.

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

Por tanto, para calcular la media muestral, escriba

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

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

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

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

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

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

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

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

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

#### MEDIA POBLACIONAL

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

## Mediana

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

#### MEDIANA

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

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

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

32 42 46 46 54

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

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

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

Como  $n = 12$  es par, se localizan los dos valores de en medio: 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

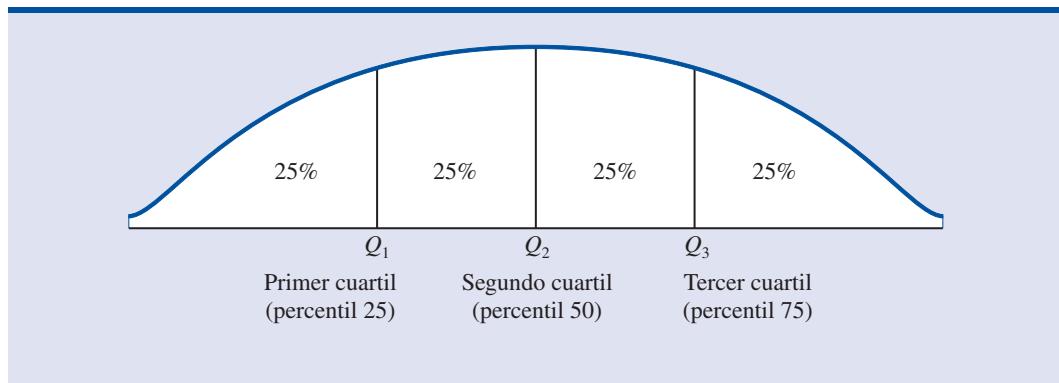
3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

**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:

$$Q_1 = \text{primer cuartil, o percentil 25}$$

$$Q_2 = \text{segundo cuartil, o percentil 50}$$

$$Q_3 = \text{tercer cuartil, o percentil 75}$$

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

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

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

Para hallar  $Q_1$ ,

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

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

Para hallar  $Q_3$ ,

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

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

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

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

(Mediana)

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

## NOTAS Y COMENTARIOS

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

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

## Ejercicios

### Método

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

- 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**

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

Fuente: *AAII Journal*, enero de 2003.

archivo  
en **CD**  
**Broker**

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.

## Autoexamen

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

87	59
80	110
47	83
73	79
50	50
93	66
72	115

Tiempos de espera para las SE en hospitales en equilibrio

60	39
54	32
18	56
29	26
45	37
34	38

- 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   
Disney

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

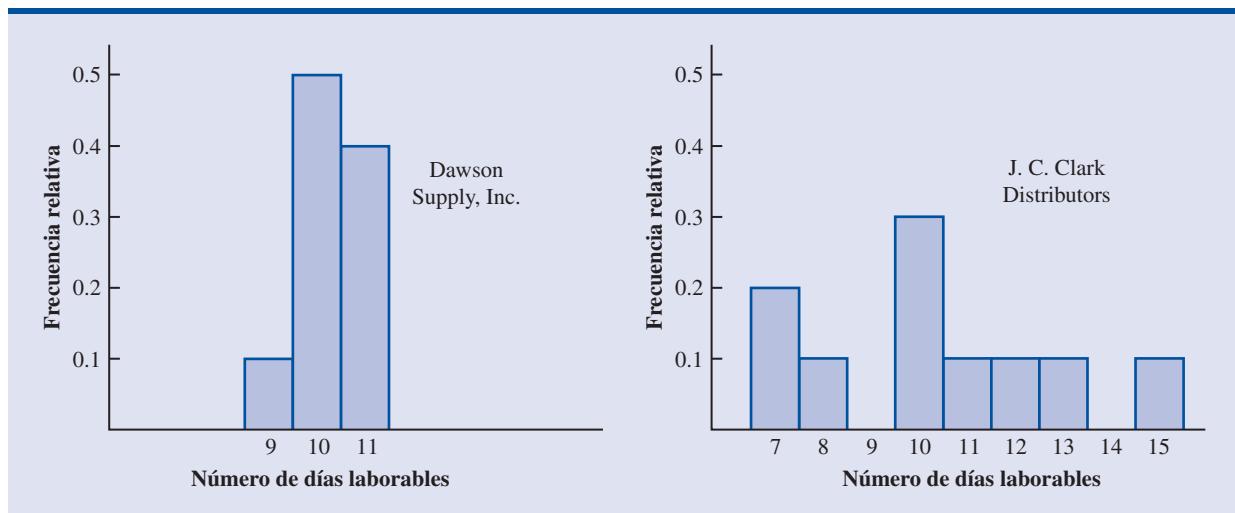
## 3.2 Medidas de variabilidad

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-

*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.*

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



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

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

## Rango

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

### RANGO

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

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

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

## Rango intercuartílico

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

## RANGO INTERCUARTÍLICO

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

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

## Varianza

La **varianza** es una medida de variabilidad que utiliza todos los datos. La varianza está basada en la diferencia entre el valor de cada observación ( $x_i$ ) y la media. A la diferencia entre cada valor  $x_i$  y la media ( $\bar{x}$  cuando se trata de una muestra,  $\mu$  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  $\sigma^2$ . En una población en la que hay  $N$  observaciones y la media poblacional es  $\mu$ , la varianza poblacional se define como sigue.

## VARIANZA POBLACIONAL

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

En la mayor parte de las aplicaciones de la estadística, los datos a analizar provienen de una muestra. Cuando se calcula la varianza muestral, lo que interesa es estimar la varianza poblacional  $\sigma^2$ . 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  $\sigma^2$ .*

## VARIANZA MUESTRAL

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

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

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

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

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

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

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

bien 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)<sup>2</sup>. Las unidades al cuadrado de la varianza dificultan la comprensión e interpretación intuitiva de los valores numéricos de la varianza. Aquí lo recomendable es entender la varianza como una medida útil para comparar la variabilidad de dos o más variables. Al comparar variables, la que tiene la varianza mayor, muestra más variabilidad. Otra interpretación del valor de la varianza suele ser innecesaria.

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

**TABLA 3.4** CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS DE LOS SUELdos INICIALES

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

Empleando la ecuación (3.5),

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

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

## Desviación estándar

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

### DESVIACIÓN ESTÁNDAR

*La desviación estándar muestral  $s$  es el estimador de la desviación estándar poblacional  $\sigma$ .*

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

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

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

¿Qué se gana con convertir la varianza en la correspondiente desviación estándar? Recuerde de 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)<sup>2</sup>. 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

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

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

## NOTAS Y COMENTARIOS

1. Los paquetes de software para estadística y las hojas de cálculo sirven para buscar los estadísticos descriptivos presentados en este capítulo. Una vez que los datos se han ingresado en una hoja de cálculo, basta emplear unos cuantos mandos 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 (*Business-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

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

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

## 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**

**Autoexamen**

### Aplicaciones

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

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

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

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?

22. La Asociación Estadounidense de Inversionistas Individuales realiza cada año una investigación sobre los corredores de bolsa con descuento (*AAII Journal*, enero de 2003). En la tabla 3.2 se muestran las comisiones que cobran 24 corredores de bolsa con descuento por dos tipos de transacciones: transacción con ayuda del corredor de 100 acciones a \$50 la acción y transacción en línea de 500 acciones a \$50 la acción.
- Calcule el rango y el rango intercuartílico en cada tipo de transacción.
  - Calcule la varianza y la desviación estándar en cada tipo de transacción.
  - Calcule el coeficiente de variación en cada tipo de transacción.
  - Compare la variabilidad en el costo que hay en los dos tipos de transacciones
24. Las puntuaciones de un jugador de golf en el 2005 y 2006 son las siguientes:
- |      |    |    |    |    |    |    |    |    |
|------|----|----|----|----|----|----|----|----|
| 2005 | 74 | 78 | 79 | 77 | 75 | 73 | 75 | 77 |
| 2006 | 71 | 70 | 75 | 77 | 85 | 80 | 71 | 79 |
- Use la media y la desviación estándar para evaluar a este jugador de golf en estos dos años.
  - ¿Cuál es la principal diferencia en su desempeño en estos dos años? ¿Se puede ver algún progreso en sus puntuaciones del 2006?, ¿cuál?
24. Los siguientes son los tiempos que hicieron los velocistas de los equipos de pista y campo de una universidad en un cuarto de milla y en una milla (los tiempos están en minutos).

*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

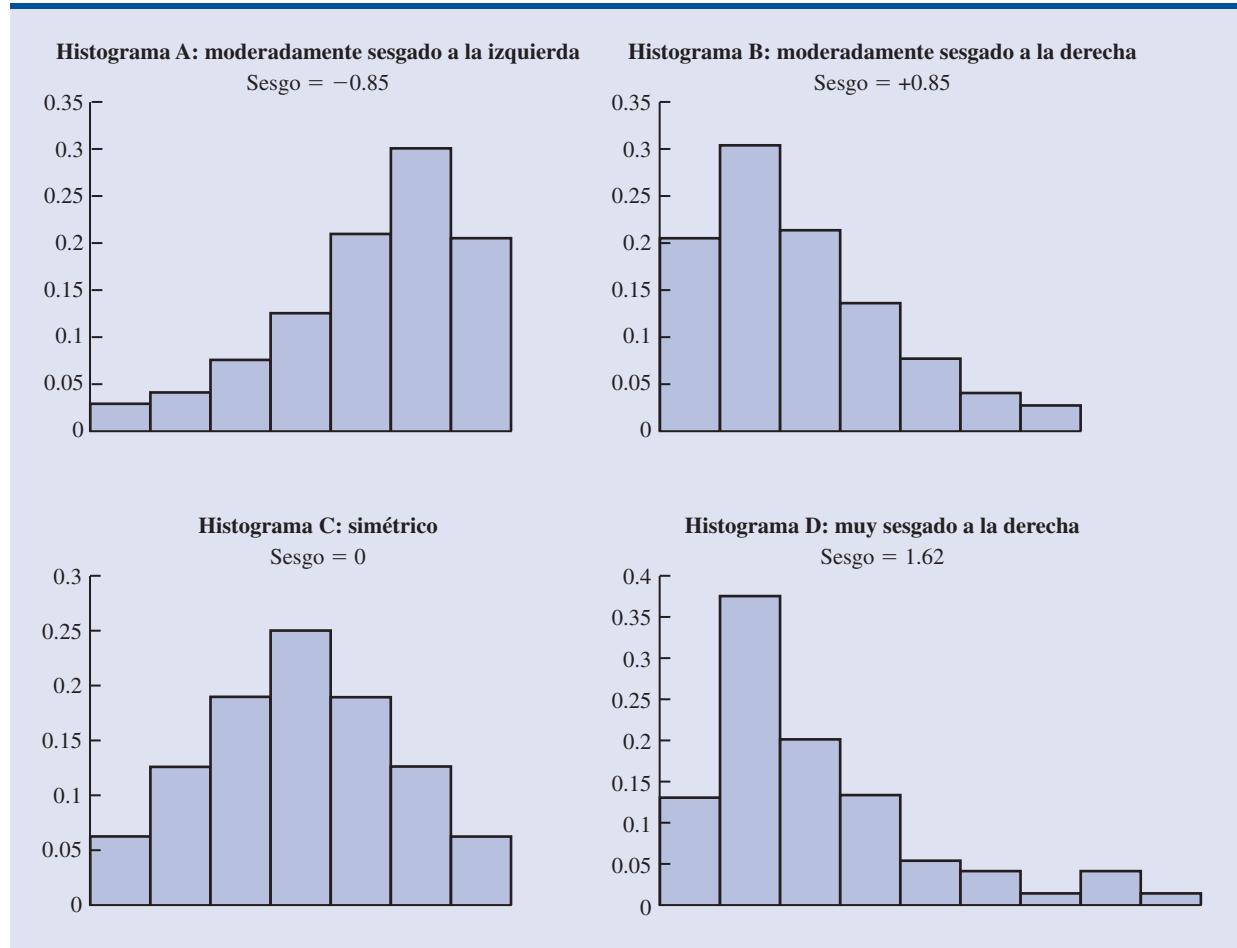
Se han descrito ya varias medidas de localización y de variabilidad de los datos. Además de estas medidas se necesita una medida de la forma de la distribución. En el capítulo 2 se vio que un histograma es una representación gráfica que muestra la forma de una distribución. Una medida numérica importante 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  $x_1, x_2, \dots, x_n$ . Suponga además que ya determinó la media muestral, que es  $\bar{x}$  y la desviación estándar muestral, que es  $s$ . Para cada valor  $x_i$  existe otro valor llamado **punto z**. La ecuación (3.9) permite calcular el punto  $z$  correspondiente a cada  $x_i$ .

### PUNTO $z$

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

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 $z$ ( $\frac{x_i - \bar{x}}{s}$ )
46	2	$2/8 = 0.25$
54	10	$10/8 = 1.25$
42	-2	$-2/8 = -0.25$
46	2	$2/8 = 0.25$
32	-12	$-12/8 = -1.50$

### TEOREMA DE CHEBYSHEV

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

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

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

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

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

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

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

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

### Regla empírica

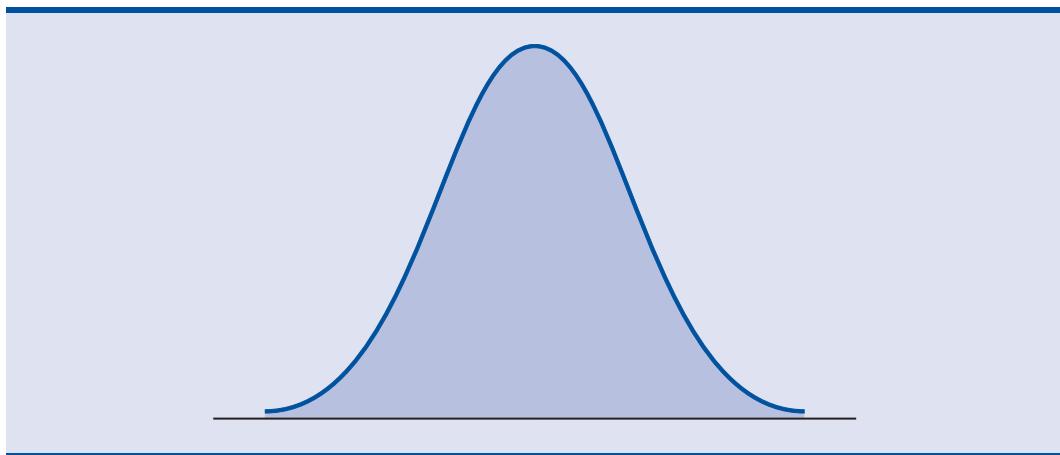
*En el teorema de Chebyshev se requiere que  $z > 1$ , pero  $z$  no tiene que ser entero.*

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

### 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 quizás 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 quizás es un dato con un valor inusual, anotado correctamente y que sí pertenece al conjunto de datos. En tal caso debe conservarse.

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

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

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

### NOTAS Y COMENTARIOS

1. El teorema de Chebyshev es aplicable a cualquier conjunto de datos y se usa para 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?
27. Considere una muestra en que la media es 30 y la desviación estándar es 5. Utilice el teorema de Chebyshev para determinar el porcentaje de los datos que se encuentra dentro de cada uno de los rangos siguientes.
  - a. 20 a 40
  - b. 15 a 45
  - c. 22 a 38
  - d. 18 a 42
  - e. 12 a 48
28. Suponga datos que tienen una distribución en forma de campana cuya media es 30 y desviación estándar 5. Utilice la regla empírica para determinar el porcentaje de los datos que se encuentra dentro de cada uno de los rangos siguientes.
  - a. 20 a 40
  - b. 15 a 45
  - c. 25 a 35

### Aplicaciones



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

- ¿Qué porcentaje de los estudiantes obtuvo una puntuación superior a 607?
  - ¿Qué porcentaje de los estudiantes obtuvo una puntuación superior a 707?
  - ¿Qué porcentaje de los estudiantes obtuvo una puntuación entre 407 y 507?
  - ¿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.
- ¿Cuál es el punto  $z$  de un cobertizo cuyo precio es de \$2300?
  - ¿Cuál es el punto  $z$  de un cobertizo cuyo precio es de \$4900?
  - Interprete los valores  $z$  de los incisos a y b. Diga si alguno de ellos debe ser considerado como una observación atípica.
  - 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

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

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

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

- 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](http://www.audioreview.com)). La escala de evaluación es de 1 a 5, siendo 5 lo mejor.



Sistema de sonido	Evaluación	Sistema de sonido	Evaluación
Infinity Kappa 6.1	4.00	ACI Sapphire III	4.67
Allison One	4.12	Bose 501 Series	2.14
Cambridge Ensemble II	3.82	DCM KX-212	4.09
Dynaudio Contour 1.3	4.00	Eosone RSF1000	4.17
Hsu Rsch. HRSW12V	4.56	Joseph Audio RM7si	4.88
Legacy Audio Focus	4.32	Martin Logan 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 Omni 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
$Q_1 = 3465$						$Q_2 = 3505$					$Q_3 = 3600$			

(Mediana)

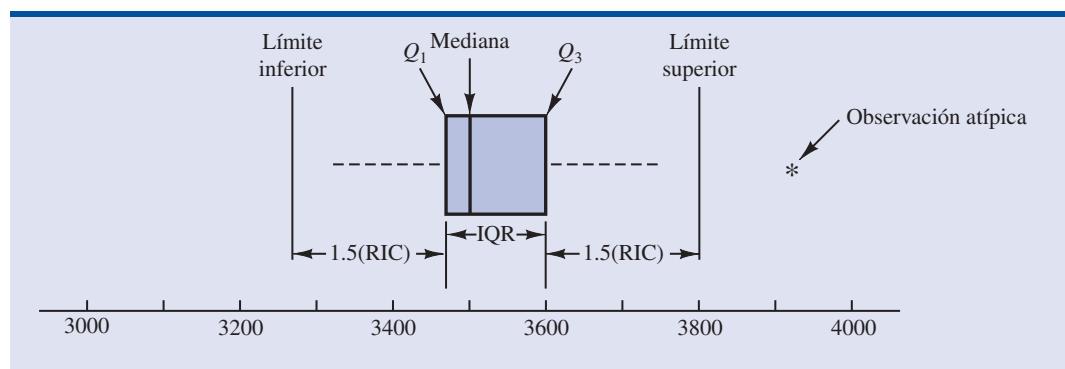
La media, que es 3505 y los cuartiles  $Q_1 = 3465$  y  $Q_3 = 3600$  se calcularon ya en la sección 3.1. Si revisa los datos encontrará que el valor menor es 3310 y el valor mayor es 3925. Así, el resumen de cinco números correspondiente a los datos de los salarios iniciales es 3310, 3465, 3505, 3600,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

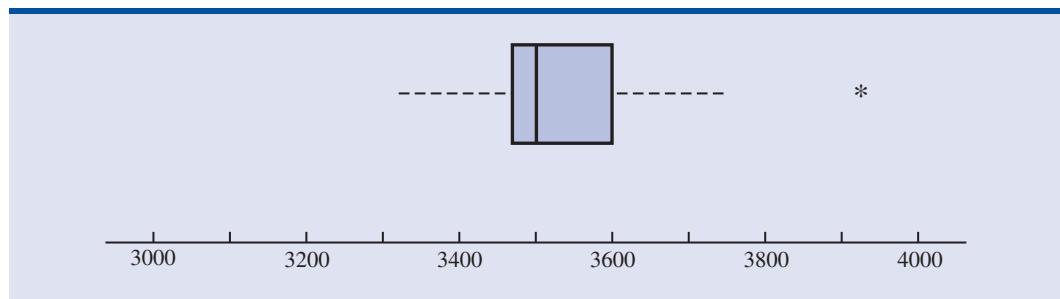
**FIGURA 3.5** DIAGRAMA DE CAJA DE LOS SALARIOS INICIALES, EN EL QUE SE MUESTRAN LAS LÍNEAS QUE INDICAN LOS LÍMITES INFERIOR Y SUPERIOR



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.

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

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



### NOTAS Y COMENTARIOS

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

### Ejercicios

#### Métodos

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

#### Aplicaciones

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

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

- 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.
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 ejecutivos.

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

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

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



**TABLA 3.6** RENDIMIENTOS PORCENTUALES ANUALES EN FONDOS MUTUALISTAS

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

## 3.5

## Medidas de la asociación entre dos variables

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

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

### Covarianza

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

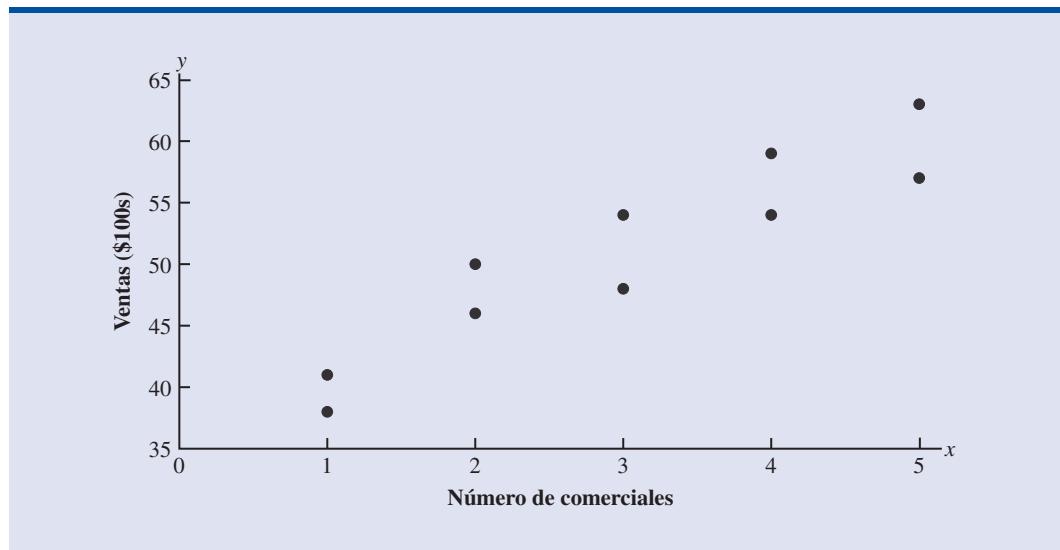
#### COVARIANZA MUESTRAL

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

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

**TABLA 3.7** DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

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

**FIGURA 3.7** DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

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

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

**TABLA 3.8** CÁLCULO DE LA COVARIANZA MUESTRAL

$x_i$	$y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
2	50	-1	-1	1
5	57	2	6	12
1	41	-2	-10	20
3	54	0	3	0
4	54	1	3	3
1	38	-2	-13	26
5	63	2	12	24
3	48	0	-3	0
4	59	1	8	8
2	46	-1	-5	5
Totales	30	0	0	99
$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$				

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

#### COVARIANZA POBLACIONAL

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

En la ecuación (3.11)  $\mu_x$  se usa para denotar la media poblacional de la variable  $x$  y  $\mu_y$  para denotar la media poblacional de la variable  $y$ . La covarianza  $\sigma_{xy}$  está definida para una población de tamaño  $N$ .

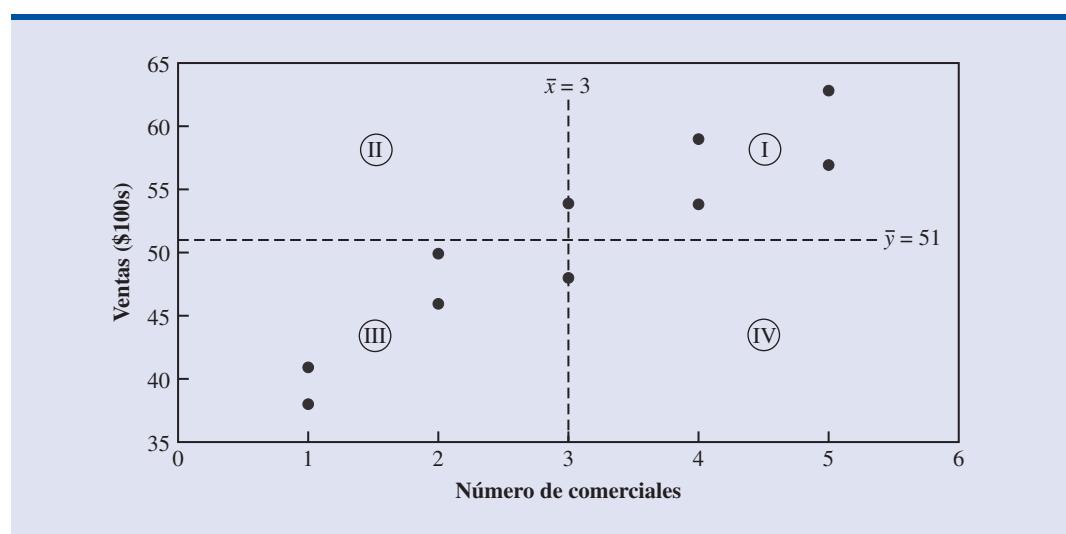
#### Interpretación de la covarianza

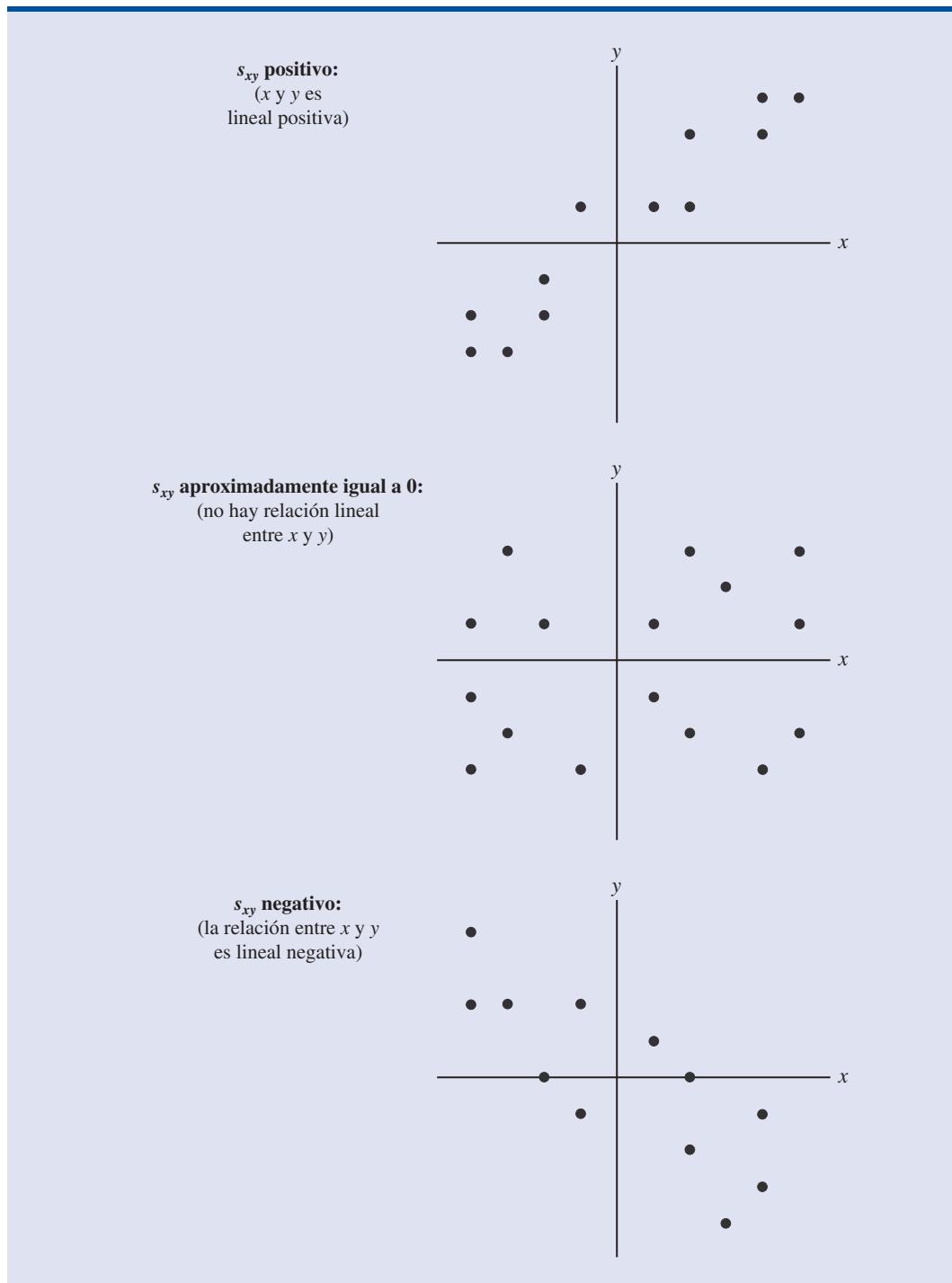
Para ayudar a la interpretación de la covarianza muestral, considere la figura 3.8; presenta el mismo diagrama de dispersión de la figura 3.7 pero con una línea vertical punteada en  $\bar{x} = 3$  y una línea horizontal punteada en  $\bar{y} = 51$ . Estas líneas dividen a la gráfica en cuatro cuadrantes.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

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

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

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



**FIGURA 3.9** INTERPRETACIÓN DE LA COVARIANZA MUESTRAL

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

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

## Coeficiente de correlación

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

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

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

donde

$r_{xy}$  = coeficiente de correlación muestral

$s_{xy}$  = covarianza muestral

$s_x$  = desviación estándar muestral de  $x$

$s_y$  = desviación estándar muestral de  $y$

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

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

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

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

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

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

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

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

*El coeficiente de correlación muestral  $r_{xy}$  proporciona un estimador del coeficiente de correlación poblacional  $\rho_{xy}$ .*

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

donde

$\rho_{xy}$  = coeficiente de correlación poblacional

$\sigma_{xy}$  = covarianza poblacional

$\sigma_x$  = desviación estándar poblacional de  $x$

$\sigma_y$  = desviación estándar poblacional de  $y$

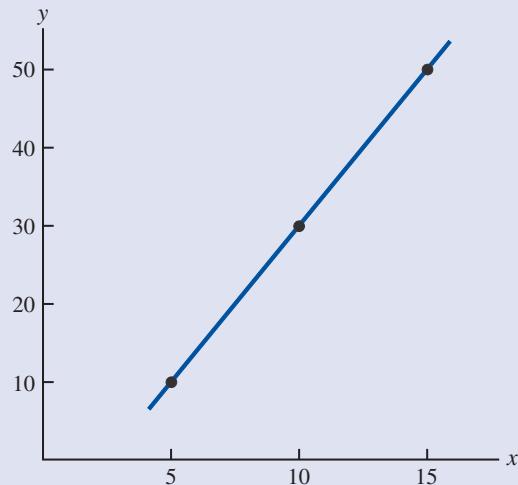
El coeficiente de correlación muestral  $r_{xy}$  proporciona un estimador del coeficiente de correlación poblacional  $\rho_{xy}$ .

### 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
Total	30	90	50	0	800	200
		$\bar{x} = 10$	$\bar{y} = 30$			

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

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

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

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

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

El coeficiente de correlación va desde  $-1$  hasta  $+1$ . Los valores cercanos a  $-1$  o a  $+1$  corresponden a una relación lineal fuerte. Entre más cercano a cero sea el valor de la correlación, más débil es la relación lineal.

De manera que el valor del coeficiente de correlación muestral es 1.

En general, puede demostrar que si todos los valores del conjunto de datos caen en una línea recta con pendiente positiva, el coeficiente de correlación será  $+1$ ; es decir, un coeficiente de correlación de  $+1$  corresponde a una relación lineal positiva perfecta entre  $x$  y  $y$ . Por otra parte, si los puntos del conjunto de datos caen sobre una línea recta con pendiente negativa, el coeficiente de correlación muestral será  $-1$ ; un coeficiente de correlación de  $-1$  corresponde a una relación lineal negativa perfecta entre  $x$  y  $y$ .

Suponga ahora que un conjunto de datos muestra una relación lineal positiva entre  $x$  y  $y$ , pero que la relación no es perfecta. El valor de  $r_{xy}$  será menor a 1, indicando que no todos los puntos del diagrama de dispersión se encuentran en una línea recta. Entre más se desvían 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, quizás 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).

<b>Rating</b>	19	17	17	14	16	12	15	12	13
<b>Share</b>	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.

<b>Velocidad</b>	30	50	40	55	30	25	60	25	50	55
<b>Rendimiento</b>	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).
- |             |         |      |      |       |      |       |      |      |      |       |
|-------------|---------|------|------|-------|------|-------|------|------|------|-------|
| 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 |

archivo  
en **CD**  
StockMarket

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

- ¿Cuál es la media muestral de las temperaturas diarias más elevadas?
- ¿Cuál es la media muestral de las temperaturas diarias más bajas?
- ¿Cuál es la correlación entre temperaturas más elevadas y temperaturas más bajas?

### 3.6

## La media ponderada y el empleo de datos agrupados

En la sección 3.1 se presentó la media como una de las medidas más importantes de localización central. La fórmula para la media de una muestra en la que hay  $n$  observaciones se escribe como sigue.

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

En esta fórmula, a cada  $x_i$  se le da la misma importancia o el mismo peso. Aunque esto es lo más común, en algunas situaciones la media se calcula dando a cada observación un peso que refleja su importancia. A una media calculada de esta manera se le llama **media ponderada**.

### Media ponderada

La media ponderada se calcula:

#### MEDIA PONDERADA

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad (3.15)$$

donde

$x_i$  = valor de la observación  $i$

$w_i$  = peso para la observación  $i$

Si los datos provienen de una muestra, la ecuación (3.15) proporciona la media ponderada muestral. Si son de una población,  $\mu$  se sustituye por  $\bar{x}$  en la ecuación (3.15) y se obtiene la media ponderada poblacional.

Como ejemplo de la necesidad de la media ponderada muestral, considere la muestra siguiente de cinco compras de materia prima realizadas en los últimos tres meses.

Compra	Costo por libra (\$)	Número de libras
1	3.00	1200
2	3.40	500
3	2.80	2750
4	2.90	1000
5	3.25	800

Observe que el costo por libra varía desde \$2.80 hasta \$3.40 y la cantidad comprada varía desde 500 hasta 2 750 libras. Suponga que el administrador quiere información sobre el costo medio por libra de la materia prima. Como las cantidades compradas varían, es necesario emplear la fórmula para la media ponderada. Los valores de los datos de los cinco costos por libra son  $x_1 = 3.00$ ,  $x_2 = 3.40$ ,  $x_3 = 2.80$ ,  $x_4 = 2.90$ , y  $x_5 = 3.25$ . El costo medio ponderado por libra se ob-

tiene ponderando cada costo con su cantidad correspondiente. Por ejemplo, los pesos (de ponderación) son  $w_1 = 1200$ ,  $w_2 = 500$ ,  $w_3 = 2750$ ,  $w_4 = 1000$  y  $w_5 = 800$ . De acuerdo con la ecuación (3.15) la media ponderada se calcula:

$$\begin{aligned}\bar{x} &= \frac{1200(3.00) + 500(3.40) + 2750(2.80) + 1000(2.90) + 800(3.25)}{1200 + 500 + 2750 + 1000 + 800} \\ &= \frac{18\,500}{6250} = 2.96\end{aligned}$$

Así, los cálculos de la media ponderada indican que el costo medio por libra de materia prima es \$2.96. Observe que si hubiera usado la ecuación (3.14) en lugar de la fórmula para la media ponderada, hubiera obtenido resultados engañosos. En ese caso la media de los valores de los cinco costos por libra sería  $(3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \$3.07$ , valor que exagera el costo medio real por libra comprada.

La selección de las ponderaciones para el cálculo de una determinada media ponderada dependen de la aplicación. Un ejemplo muy conocido por los estudiantes es el promedio de las calificaciones (en Estados Unidos). En este caso los valores de los datos son 4 que corresponde a A, 3 que corresponde a B, 2 que corresponde a C, 1 que corresponde a D y 0 que corresponde a F. Los pesos son los créditos por hora de cada materia. El ejercicio 54 al final de esta sección es un ejemplo del cálculo de esta media ponderada. En otros cálculos de la media ponderada se emplean como pesos cantidades como libras, dólares o volumen. En cualquier caso, si la importancia de las observaciones varía, el analista debe elegir los pesos que mejor reflejen la relevancia de cada observación en la determinación de la media.

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

## Datos agrupados

En la mayor parte de los casos, las medidas de localización y variabilidad se calculan mediante los valores individuales de los datos. Sin embargo, otras veces sólo se tienen datos agrupados o datos en una distribución de frecuencias. En la argumentación siguiente se muestra cómo usar la fórmula de la media ponderada para obtener aproximaciones a la media, la varianza y la desviación estándar de **datos agrupados**.

En la sección 2.2 se presentó una distribución de las duraciones en días en una muestra de auditorías de fin de año de una empresa pequeña de contadores públicos. La distribución de frecuencias de las duraciones de las auditorías que se obtuvo de una muestra de 20 clientes se presenta de nuevo en la tabla 3.11. Con base en esta distribución de frecuencias, ¿cuál es la media muestral de la duración de las auditorías?

Para calcular la media usando datos agrupados, considere el punto medio de cada clase como representativo de los elementos de esa clase. Si  $M_i$  denota el punto medio de la clase  $i$  y  $f_i$  denota la frecuencia de la clase  $i$ . Entonces la fórmula para la media ponderada (3.15) se usa con los valores de los datos denotados por  $M_i$  y los pesos dados por las frecuencias  $f_i$ . En este caso, el denominador de la ecuación (3.15) es la suma de las frecuencias, que es el tamaño de la muestra  $n$ .

**TABLA 3.11** DISTRIBUCIÓN DE FRECUENCIAS DE LAS DURACIONES DE LAS AUDITORÍAS

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

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

#### MEDIA MUESTRAL DE DATOS AGRUPADOS

$$\bar{x} = \frac{\sum f_i M_i}{n} \quad (3.16)$$

donde

$M_i$  = punto medio de la clase  $i$

$f_i$  = frecuencia de la clase  $i$

$n$  = tamaño de la muestra

Como el punto medio de clase,  $M_i$ , se encuentra a la mitad entre los límites de clase, en tabla 3.11 el punto medio de la primera clase, 10–14, es  $(10 + 14)/2 = 12$ . En la tabla 3.12 se presentan los cinco puntos medios de clase y los cálculos de la media ponderada de los datos de la duración de las auditorías. Como puede ver, la media muestral de la duración de las auditorías es 19 días.

Para calcular la varianza de datos agrupados se emplea una versión ligeramente modificada de la fórmula para la varianza dada en la ecuación (3.5). En la ecuación (3.5) los cuadrados de las desviaciones de los datos respecto a la media muestral se escribieron como  $(x_i - \bar{x})^2$ . Pero cuando se tienen datos agrupados no se conocen los valores. En este caso, se considera el punto medio de clase,  $M_i$ , como representativo de los valores  $x_i$  de la clase correspondiente. Por tanto, los cuadrados de las desviaciones respecto a la media  $(x_i - \bar{x})^2$  son sustituidos por  $(M_i - \bar{x})^2$ . Entonces, igual que en el cálculo de la media muestral de datos agrupados, pondera cada valor por la frecuencia de la clase,  $f_i$ . La suma de los cuadrados de las desviaciones respecto a la media de todos los datos se aproxima mediante  $\sum f_i(M_i - \bar{x})^2$ . En el denominador aparece el término  $n - 1$  en lugar de  $n$ , con objeto de hacer que la varianza muestral sea un estimador de la varianza poblacional. Por consiguiente, la fórmula usada para obtener la varianza muestral de datos agrupados es:

#### VARIANZA MUESTRAL PARA DATOS AGRUPADOS

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

**TABLA 3.12** CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS AGRUPADOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (días)	Punto medio de clase ( $M_i$ )	Frecuencia ( $f_i$ )	$f_i M_i$
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
			380
Media muestral $\bar{x} = \frac{\sum f_i M_i}{n} = \frac{380}{20} = 19$ días			

**TABLA 3.13** CÁLCULO DE LA VARIANZA MUESTRAL CON LOS DATOS AGRUPADOS DE LAS DURACIONES DE LAS AUDITORÍAS

Duración de la auditoría (días)	Punto medio de clase ( $M_i$ )	Frecuencia ( $f_i$ )	Desviación ( $M_i - \bar{x}$ )	Cuadrado de la desviación ( $(M_i - \bar{x})^2$ )	$f_i(M_i - \bar{x})^2$
10–14	12	4	-7	49	196
15–19	17	8	-2	4	32
20–24	22	5	3	9	45
25–29	27	2	8	64	128
30–34	32	1	13	169	169
		20			570
				$\Sigma f_i(M_i - \bar{x})^2$	
				$\text{Varianza muestral } s^2 = \frac{\Sigma f_i(M_i - \bar{x})^2}{n - 1} = \frac{570}{19} = 30$	

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

$$\mu = \frac{\sum f_i M_i}{N} \quad (3.18)$$

#### VARIANZA POBLACIONAL DE DATOS AGRUPADOS

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

#### 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.
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



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	<u>350</u>
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.

	Estadístico muestral	Parámetro poblacional
Media	$\bar{x}$	$\mu$
Varianza	$s^2$	$\sigma^2$
Desviación estándar	$s$	$\sigma$
Covarianza	$s_{xy}$	$\sigma_{xy}$
Correlación	$r_{xy}$	$\rho_{xy}$

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  $\mu$ , la varianza poblacional,  $\sigma^2$  y la desviación estándar poblacional,  $\sigma$ ).

**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

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

### Media poblacional

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

### Rango intercuartílico

$$RIC = Q_3 - Q_1 \quad (3.3)$$

### Varianza poblacional

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

### Varianza muestral

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

### Desviación estándar

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

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

**Coefficiente de variación**

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

**Punto  $z$** 

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

**Covarianza muestral**

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

**Covarianza poblacional**

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

**Coefficiente de correlación del producto–momento de Pearson: datos muestrales**

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

**Coefficiente de correlación del producto–momento de Pearson: datos poblacionales**

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

**Media ponderada**

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad (3.15)$$

**Media muestral de datos agrupados**

$$\bar{x} = \frac{\sum f_i M_i}{n} \quad (3.16)$$

**Varianza muestral de datos agrupados**

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

**Media poblacional de datos agrupados**

$$\mu = \frac{\sum f_i M_i}{N} \quad (3.18)$$

**Varianza poblacional de datos agrupados**

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

### Ejercicios complementarios

58. De acuerdo con 2003 Annual Consumer Spending Survey, el cargo promedio mensual a una tarjeta de crédito Bank of America Visa fue de \$1838 (*U.S. Airways Attaché Magazine*, diciembre de 2003). En una muestra de cargos mensuales a tarjetas de crédito los datos obtenidos son los siguientes.

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](http://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

- Entre los estudiantes que toman un préstamo estudiantil, ¿cuál es la mediana en la deuda que tienen una vez terminados sus estudios?
  - ¿Cuál es la varianza y cuál la desviación estándar?
62. Los propietarios de negocios pequeños suelen contratar a empresas con servicio de nómina para que se encarguen del pago de sus empleados. Las razones son que encuentran regulaciones complicadas para el pago de impuestos y que las multas por errores en los impuestos de los empleados son elevadas. De acuerdo con el Internal Revenue Service, 26% de las declaraciones de impuestos de los empleados contienen errores que ocasionan multas a los dueños. (*The Wall Street Journal*, 30 de enero de 2006). La siguiente es una muestra de 20 multas a propietarios de negocios pequeños.

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

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

- Calcule la media muestral en el tiempo que se necesita con cada transporte.
  - Calcule la desviación estándar para cada transporte.
  - De acuerdo con los resultados en los incisos a y b, ¿cuál será el medio de transporte preferido? Explique.
  - Para cada medio de transporte elabore un diagrama de caja. ¿Se confirma la conclusión que dio en el inciso c mediante una comparación de los diagramas de caja?
64. La National Association of Realtors informa sobre la mediana en el precio de una casa en Estados Unidos y sobre el aumento de esta mediana en los últimos cinco años. Use la muestra de precios de casas para responder a las preguntas siguientes.

995.9	48.8	175.0	263.5	298.0	218.9	209.0
628.3	111.0	212.9	92.6	2325.0	958.0	212.5

- ¿Cuál es la mediana muestral de los precios de las casas?
  - En enero del 2001 la National Association of Realtors informó que la mediana en el precio de una casa en Estados Unidos era \$139 300. ¿Cuál ha sido el incremento porcentual de la mediana en el precio de una casa en cinco años?
  - ¿Cuáles son el primer y tercer cuartiles de los datos muestrales?
  - Dé el resumen de cinco números para los precios de las casas.
  - ¿Existe alguna observación atípica en los datos?
  - ¿En la muestra cuál es la media en el precio de una casa? ¿Por qué prefiere la National Association of Realtors usar en sus informes la mediana en el precio de las casas?
65. Los datos siguientes son los gastos en publicidad (en millones de dólares) y los envíos en millones de barriles (bbls.) de las 10 principales marcas de cerveza.



Marca	Gastos en publicidad (millones de dólares)	Despachos en bbls (millones)
Budweiser	120.0	36.3
Bud Light	68.7	20.7
Miller Lite	100.1	15.9
Coors Light	76.6	13.2
Busch	8.7	8.1
Natural Light	0.1	7.1
Miller Genuine Draft	21.5	5.6
Miller High Life	1.4	4.4
Busch Lite	5.3	4.3
Milwaukee's Best	1.7	4.3

- a. ¿Cuál es la covarianza muestral? ¿Indica que hay una relación positiva o negativa?
- b. ¿Cuál es el coeficiente de correlación?
66. *Road & Track* proporciona la muestra siguiente de desgaste en llantas y la capacidad de carga máxima de llantas de automóviles.

Desgaste en llantas	Capacidad de carga máxima
75	853
82	1047
85	1135
87	1201
88	1235
91	1356
92	1389
93	1433
105	2039

- a. Con estos datos elabore un diagrama de dispersión en el que el desgaste ocupe el eje *x*.
- b. Calcule el coeficiente de correlación muestral. ¿Qué indica el coeficiente de correlación muestral acerca de la relación entre el desgaste y la capacidad de carga máxima?
67. Los datos siguientes presentan el seguimiento de la rentabilidad primaria por acción durante 52 semanas y los valores contables reportados por 10 empresas (*The Wall Street Journal*, 13 de marzo de 2000).

Empresa	Valor contable	Rentabilidad
Am Elec	25.21	2.69
Columbia En	23.20	3.01
Con Ed	25.19	3.13
Duke Energy	20.17	2.25
Edison Int'l	13.55	1.79
Enron Cp.	7.44	1.27
Peco	13.61	3.15
Pub Sv Ent	21.86	3.29
Southn Co.	8.77	1.86
Unicom	23.22	2.74

- a. Elabore un diagrama de dispersión, que los valores contables ocupen el eje  $x$ .
- b. Calcule el coeficiente de correlación muestral. ¿Qué indica este coeficiente acerca de la relación entre la rentabilidad por acción y el valor contable?
68. Una técnica de pronóstico conocida como promedios móviles emplea el promedio o la media de los  $n$  períodos más recientes para pronosticar el valor siguiente en los datos de una serie de tiempo. En un promedio móvil de tres períodos, se usan los datos de los tres períodos 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 Proprietary card se refiere a pagos realizados con tarjeta de crédito de National Clothing. A los clientes que hicieron compras con un cupón de descuento se les denomina aquí promocionales y a quienes hicieron sus compras sin emplear cupón de descuento se les denomina regulares. Como a los clientes de las tiendas Pelican no se les enviaron cupones promocionales, los directivos consideran que las ventas hechas a las personas que presentaron un cupón de descuento son ventas que de otro modo no se hubieran realizado. Es obvio que Pelican espera que los clientes promocionales continúen comprando en sus tiendas.

La mayor parte de las variables que aparecen en la tabla 3.14 se explican por sí mismas, pero dos de ellas deben ser aclaradas.

Artículos	Número de artículos comprados
Ventas netas	Cantidad cargada a la tarjeta de crédito

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

### Informe para los directivos

Use los métodos de la estadística descriptiva presentados en este capítulo para resumir los datos y comente sus hallazgos. Su informe debe contener, por lo menos, lo siguiente:

1. Estadísticos descriptivos sobre las ventas netas y sobre las ventas a los distintos tipos de clientes.
2. Estadísticos descriptivos respecto de la relación entre edad y ventas netas.

**TABLA 3.14 MUESTRA DE 100 COMPRAS CON TARJETA DE CRÉDITO REALIZADAS EN LAS TIENDAS PELICAN**

Cliente	Tipo de cliente	Ar-tículos	Ventas netas	Modo de pago	Género	Estado civil	Edad
1	Regular	1	39.50	Discover	Masculino	Casado	32
2	Promocional	1	102.40	Proprietary Card	Femenino	Casada	36
3	Regular	1	22.50	Proprietary Card	Femenino	Casada	32
4	Promocional	5	100.40	Proprietary Card	Femenino	Casada	28
5	Regular	2	54.00	MasterCard	Femenino	Casada	34
6	Regular	1	44.50	MasterCard	Femenino	Casada	44
7	Promocional	2	78.00	Proprietary Card	Femenino	Casada	30
8	Regular	1	22.50	Visa	Femenino	Casada	40
9	Promocional	2	56.52	Proprietary Card	Femenino	Casada	46
10	Regular	1	44.50	Proprietary Card	Femenino	Casada	36
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Femenino	Casada	44
97	Promocional	9	253.00	Proprietary Card	Femenino	Casada	30
98	Promocional	10	287.59	Proprietary Card	Femenino	Casada	52
99	Promocional	2	47.60	Proprietary Card	Femenino	Casada	30
100	Promocional	1	28.44	Proprietary Card	Femenino	Casada	44

## Caso problema 2 Industria cinematográfica

La industria cinematográfica es un negocio muy competitivo. En más de 50 estudios se producen 300 a 400 películas por año y el éxito financiero de estas películas varía en forma considerable. Las variables usuales para medir el éxito de una película son ventas brutas (en millones de dólares) en el fin de semana del estreno, ventas brutas totales (en millones de dólares), número de salas donde se presenta la película, semanas en las que la película se encuentra entre las 60 mejores en ventas brutas. Los datos de una muestra de 100 películas producidas en 2005 se encuentran en el archivo titulado Movies. La tabla 3.15 muestra los datos de las 10 primeras películas que se encuentran en este archivo.

### Informe para los directivos

Use los métodos numéricos de la estadística descriptiva presentados en este capítulo para averiguar cómo contribuyen estas variables al éxito de una película. Su informe debe contener lo siguiente.

1. Estadísticos descriptivos para cada una de las cuatro variables con un análisis sobre la información que la estadística descriptiva proporciona acerca de la industria del cine.
2. ¿Hay alguna película que deba ser considerada como una observación atípica de alto desempeño?
3. Los estadísticos descriptivos muestran la relación entre ventas brutas y cada una de las otras variables. Argumente.

**TABLA 3.15** DATOS DEL ÉXITO DE 10 PELÍCULAS

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



## Caso problema 3 Las escuelas de negocios de Asia-Pacífico

En la actualidad se ha vuelto mundial el interés por tener un grado superior en estudios de negocios. En una investigación se encontró que en Asia cada vez más personas eligen una maestría en administración de negocios como camino hacia el éxito corporativo. De esta manera, en las escuelas de Asia-Pacífico, el número de solicitudes a cursos de maestría en administración de negocios sigue aumentando.

En esa región miles de personas suspenden sus carreras y pasan dos años en estudios para obtener una formación teórica en negocios. Los cursos en estas escuelas son bastante pesados y comprenden economía, banca, marketing, ciencias de la conducta, relaciones laborales, toma de decisiones, pensamiento estratégico, derecho internacional en negocios y otras áreas. En los datos que se presentan en la tabla 3.16 aparecen algunas de las características de las principales escuelas de negocios de Asia-Pacífico.



Asian

**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
Jannalal Bajaj Institute of Management Studies (Bombay)	240	9	1 000	1 000	24	0	No	No	Sí	7 000
Curtin Institute of Technology (Perth)	98	15	9 475	19 097	29	43	Sí	No	Sí	55 000
Lahore University of Management Sciences	70	14	11 250	26 300	23	2.5	No	No	No	7 500
Universiti Sains Malaysia (Penang)	30	5	2 260	2 260	32	15	No	Sí	Sí	16 000
De La Salle University (Manila)	44	17	3 300	3 600	28	3.5	Sí	No	Sí	13 100

## Informe para los directivos

Use los métodos de la estadística descriptiva para resumir los datos de la tabla 3.16. Argumente sobre sus hallazgos.

1. Para cada variable presente un resumen del conjunto de datos. Haga comentarios e interpretaciones con base en máximos y mínimos, así como en las medias y proporciones adecuadas. ¿Qué conclusiones nuevas proporcionan estos estadísticos descriptivos respecto de las escuelas de negocios de Asia–Pacífico?
2. Resuma los datos para hacer las comparaciones siguientes:
  - a. Diferencias entre las colegiaturas para alumnos locales y de fuera.
  - b. Diferencias entre los salarios promedio iniciales para egresados de escuelas que exigen experiencia laboral y de escuelas que no la exigen.
  - c. Discrepancias entre los salarios promedio iniciales de egresados de escuelas que exigen una prueba de inglés y de escuelas que no la exigen.
3. ¿Parece haber relación entre los salarios iniciales y las colegiaturas?
4. Presente cualquier gráfica y resumen numérico que pueda servir para comunicar a otras personas la información presentada en la tabla 3.16.

## Apéndice 3.1 Estadística descriptiva usando Minitab

En este apéndice se describe cómo usar Minitab para obtener estadísticos descriptivos. En la tabla 3.1 aparecen los sueldos iniciales de 12 recién egresados de la carrera de administración. En el panel A de la figura 3.11 están los estadísticos descriptivos obtenidos para resumir los datos usando Minitab. A continuación se dan las definiciones de los títulos que se observan en el panel A.

N	número de valores en los datos
N*	número de datos faltantes
Mean	media
SE Mean	error estándar de la media
StDev	desviación estándar
Minimum	valor mínimo (menor) en los datos
Q1	primer cuartil
Median	mediana
Q3	tercer cuartil
Maximum	valor máximo (mayor) en los datos

El título SE mean se refiere al *error estándar de la media*. Este valor se obtiene dividiendo la desviación estándar entre la raíz cuadrada de  $N$ . La interpretación y uso de esta medición se verá en el capítulo 7, cuando se introduzca el tema del muestreo y de la distribución muestral.

Aunque en los resultados de Minitab no aparecen el rango, el rango intercuartílico, la varianza y el coeficiente de variación, estas medidas son fáciles de calcular a partir de los resultados que aparecen en la figura 3.11; se calculan como sigue.

$$\text{Rango} = \text{Máximo} - \text{Mínimo}$$

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

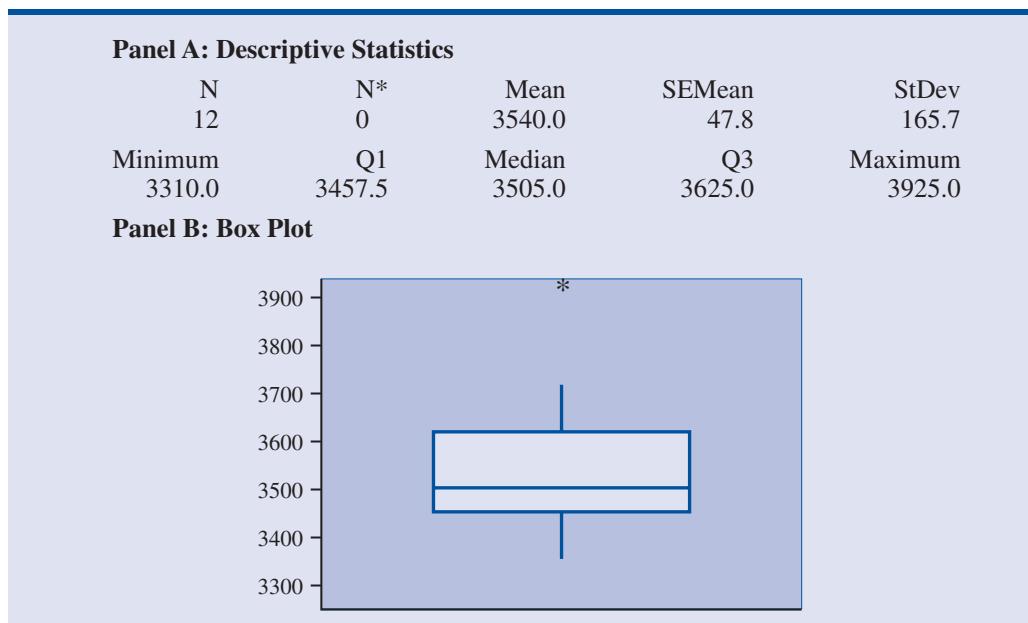
$$\text{Varianza} = (\text{StDev})^2$$

$$\text{Coeficiente de variación} = (\text{StDev}/\text{Media}) \times 100$$

Por último, observe que los cuartiles que da Minitab,  $Q_1 = 3457.5$  y  $Q_3 = 3625$ , son ligeramente diferentes a los calculados en la sección 3.1. Esto se debe al empleo de convenciones\* dirigidas

\*Cuando se tienen  $n$  observaciones ordenadas de menor a mayor (en orden ascendente), para localizar los cuartiles  $Q_1$  y  $Q_3$  Minitab usa las posiciones dadas por  $(n + 1)/4$  y  $3(n + 1)/4$ , respectivamente. Si se obtiene un número fraccionario, Minitab interpola entre los valores de los datos adyacentes ordenados para determinar el cuartil correspondiente.

**FIGURA 3.11** ESTADÍSTICOS DESCRIPTIVOS Y DIAGRAMA DE CAJA PROPORCIONADOS POR MINITAB



ferentes para identificar los cuartiles. De manera que los valores  $Q_1$  y  $Q_3$  obtenidos con una convención quizás no sean idénticos a los valores  $Q_1$  y  $Q_3$  obtenidos con otra. Sin embargo, estas diferencias tienden a ser despreciables y los resultados no afectan al hacer las interpretaciones relacionadas con los cuartiles.

Ahora verá cómo se generan los estadísticos que aparecen en la figura 3.11. Los datos de los sueldos iniciales se encuentran en la columna C2 de la hoja de cálculo de Minitab. Para generar los estadísticos descriptivos realice los pasos siguientes:



- Paso 1.** Seleccionar el menú **Stat**
- Paso 2.** Elegir **Basic Statistics**
- Paso 3.** Elegir **Display Descriptive Statistics**
- Paso 4.** Cuando aparece el cuadro de diálogo **Display Descriptive Statistics**:
  - Ingresar C2 en el cuadro **Variables**
  - Dar clic en **OK**

El panel B de la figura 3.11 es un diagrama de caja obtenido con Minitab y contiene entre el primer y tercer cuartil 50% de los datos. La línea dentro de la caja corresponde a la mediana. El asterisco indica que hay una observación atípica en 3925.

Con los pasos siguientes se genera el diagrama de caja que aparece en la figura 3.11.

- Paso 1.** Seleccionar el menú **Graph**
- Paso 2.** Elegir **Boxplot**
- Paso 3.** Elegir **Simple** y hacer clic en **OK**
- Paso 4.** Cuando aparezca el cuadro de diálogo **Boxplot-One Y, Simple**:
  - Ingresar C2 en el cuadro **Graph variables**
  - Dar clic en **OK**

La medida del sesgo tampoco aparece como parte de los resultados estándar de estadística descriptiva que proporciona Minitab. Sin embargo, puede incluirse mediante los pasos siguientes.

**FIGURA 3.12** COVARIANZA Y CORRELACIÓN OBTENIDAS USANDO MINITAB CON LOS DATOS DEL NÚMERO DE COMERCIALES Y VENTAS

**Covariances: No. of Commercials, Sales Volume**

	No. of Comme	Sales Volume
No. of Comme	2.22222	
Sales Volume	11.00000	62.88889

**Correlations: No. of Commercials, Sales Volume**

Pearson correlation of No. of Commercials and Sales Volume = 0.930  
P-Value = 0.000

#### **Paso 1.** Seleccionar el menú Stat

## Paso 2. Elegir Basic Statistics

### **Paso 3. Elegir Display Descriptive Statistics**

**Paso 4.** Cuando aparezca el cuadro de diálogo Display Descriptive Statistics:

## Clic en Statistics

## Elegir Skewness

Clic en OK

Clic en **OK**

La medida del sesgo, 1.09, aparecerá en su hoja de cálculo.

La figura 3.12 muestra los resultados que da Minitab para la covarianza y la correlación con los datos de la tienda de equipos de sonido presentados en la tabla 3.7. En la parte de la figura que corresponde a la covarianza, *No. of Comme* denota el número de semanas que se televisaron los comerciales y *Sales Volume* las ventas durante la semana siguiente. El valor que aparece en la columna *No. of Comme* y en el renglón *Sales Volume*, 11, es la covarianza muestral que se calculó en la sección 3.5. El valor de la columna *No. of Comme* y en el renglón *No. of Comme*, 2.22222, es la varianza muestral del número de comerciales, y el valor que se encuentra en la columna *Sales Volume* y en el renglón *Sales Volume*, 62.88889, es la varianza muestral de las ventas. El coeficiente de correlación muestral, 0.930, aparece en los resultados, en la parte correspondiente a la correlación. Nota: la interpretación del valor-*p* = 0.000 se verá en el capítulo 9.

Ahora se describe cómo obtener la información que se muestra en la figura 3.12. En la columna C2 de la hoja de cálculo de Minitab ingrese los datos del número de comerciales y en la columna C3 los datos de las ventas. Los pasos necesarios para obtener los resultados que se muestran en los tres primeros renglones de la figura 3.12 son los siguientes.

#### **Paso 1.** Seleccionar el menú Stat

## Paso 2. Elegir Basic Statistics

### Paso 3. Elegir Covariance

**Paso 4.** Cuando aparezca el cuadro de diálogo Covariance:

Ingresar C2 C3 en el cuadro **Variable**

Clic en OK

Para obtener el resultado correspondiente a la correlación, que se observa en la tabla 3.12, sólo hay que hacer una modificación a estos pasos para la covarianza. En el paso 3 seleccione la opción **Correlation**.

Apéndice 3.2 Estadísticos descriptivos usando Excel

Emplee Excel para generar los estadísticos descriptivos vistos en este capítulo. Ahora aprenderá a usar Excel para generar diversas medidas de localización y de variabilidad para una variable, así como la covarianza y el coeficiente de correlación para medir la asociación entre dos variables.

**FIGURA 3.13** USO DE LAS FUNCIONES DE EXCEL PARA CALCULAR LA MEDIA, MEDIANA, MODA, VARIANZA Y DESVIACIÓN ESTÁNDAR

	A	B	C	D		E	F
1	Graduate	Starting Salary		Mean	=AVERAGE(B2:B13)		
2	1	3450		Median	=MEDIAN(B2:B13)		
3	2	3550		Mode	=MODE(B2:B13)		
4	3	3650		Variance	=VAR(B2:B13)		
5	4	3480		Standard Deviation	=STDEV(B2:B13)		
6	5	3355					
7	6	3310		A	B	C	D
8	7	3490		1 Graduate	Starting Salary		Mean 3540
9	8	3730		2	1	3450	Median 3505
10	9	3540		3	2	3550	Mode 3480
11	10	3925		4	3	3650	Variance 27440.91
12	11	3520		5	4	3480	Standard Deviation 165.65
13	12	3480		6	5	3355	
14				7	6	3310	
				8	7	3490	
				9	8	3730	
				10	9	3540	
				11	10	3925	
				12	11	3520	
				13	12	3480	
				14			

## Uso de las funciones de Excel



Excel tiene funciones para calcular media, mediana, moda, varianza muestral y desviación estándar muestral. Con los datos de los sueldos iniciales de la tabla 3.1 ilustrará el uso de las funciones de Excel para calcular la media, mediana, moda, varianza muestral y desviación estándar muestral. Al ir siguiendo los pasos necesarios, consulte la figura 3.13. Ingrese los datos en la columna B.

Para calcular la media emplee la función AVERAGE (PROMEDIO) de Excel ingresando la fórmula siguiente en la celda E1:

=AVERAGE(B2:B13)

De manera similar ingrese en las celdas E2:E5 las fórmulas =MEDIAN(A2:B13), =MODA(B2:B13), =VAR(B2:B13) y =DESVEST(B2:B13) para calcular, respectivamente, la mediana, moda, varianza y desviación estándar. La hoja de cálculo que aparece en primer plano muestra que los valores calculados usando las funciones de Excel son iguales a los ya calculados en este capítulo.

Excel tiene también funciones para calcular la covarianza y el coeficiente de correlación. Al usar estas funciones debe tener cuidado, dado que la función covarianza trata a los datos como población y la función correlación como muestra. Por tanto, los resultados obtenidos con la función covarianza de Excel deben ajustarse para obtener la covarianza muestral. Se le muestra cómo usar estas funciones de Excel para el cálculo de la covarianza muestral y del coeficiente de correlación muestral empleando los datos de la tienda que vende equipos de sonido y que se presentaron en la figura 3.14.



**FIGURA 3.14** USO DE LAS FUNCIONES DE EXCEL PARA CALCULAR LA COVARIANZA Y LA CORRELACIÓN

	A	B	C	D	E		F		G		
1	Week	Commercials	Sales		Population Covariance		=COVAR(B2:B11:C2:C11)				
2	1	2	50		Sample Correlation		=CORREL(B2:B11,C2:C11)				
3	2	5	57								
4	3	1	41		A	B	C	D	E	F	G
5	4	3	54		1	Week	Commercials	Sales	Population Covariance	9.90	
6	5	4	54		2	1	2	50	Sample Correlation	0.93	
7	6	1	38		3	2	5	57			
8	7	5	63		4	3	1	41			
9	8	3	48		5	4	3	54			
10	9	4	59		6	5	4	54			
11	10	2	46		7	6	1	38			
12					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	<b>Graduate</b>	<b>Starting Salary</b>		<i>Starting Salary</i>		
2	1	3450				
3	2	3550		<b>Mean</b>	3540	
4	3	3650		Standard Error	47.82	
5	4	3480		<b>Median</b>	3505	
6	5	3355		<b>Mode</b>	3480	
7	6	3310		<b>Standard Deviation</b>	165.65	
8	7	3490		<b>Sample Variance</b>	27440.91	
9	8	3730		Kurtosis	1.7189	
10	9	3540		<b>Skewness</b>	1.0911	
11	10	3925		<b>Range</b>	615	
12	11	3520		<b>Minimum</b>	3310	
13	12	3480		<b>Maximum</b>	3925	
14				<b>Sum</b>	42480	
15				<b>Count</b>	12	
16						



esta herramienta para calcular los estadísticos descriptivos del conjunto de datos referidos a los sueldos iniciales presentados en la tabla 3.1. Consulte la figura 3.15 a medida que se le describen los pasos necesarios.

**Paso 1.** Seleccionar el menú **Herramientas**

**Paso 2.** Elegir **Análisis de datos**

**Paso 3.** Cuando aparezca el cuadro de diálogo Análisis de datos:

Elegir **Estadística descriptiva**

Clic en **OK**

**Paso 4.** Cuando aparezca el cuadro de diálogo Estadística descriptiva:

Ingresar B1:B13 en el cuadro **Rango de entrada**

Seleccionar **Agrupados por Columnas**

Seleccionar **Rótulos en la primera fila**

Seleccionar **Rango de salida**

Ingresar D1 en la caja para el rango de salida (para identificar la esquina superior izquierda de la hoja de cálculo en la que aparecerá la estadística descriptiva)

Seleccionar **Resumen de estadísticas**

Clic en **OK**.

Las celdas D1:D15 de la figura 3.15 muestran la estadística descriptiva obtenida con Excel. Las entradas en negritas son los estadísticos descriptivos que se estudiaron en este capítulo. Los estadísticos descriptivos que no están en negritas se estudiarán en capítulos subsiguientes o en textos más avanzados.