CAPÍTULO 3

Medidas numéricas descriptivas

USO DE LA ESTADÍSTICA: Evaluación de los rendimientos de los fondos de inversión

OBJETIVOS DE APRENDIZAJE

En este capítulo, aprenderá:

- A describir las propiedades de tendencia central, variación y forma de los datos numéricos
- A calcular las medidas descriptivas de una población
- A construir e interpretar una gráfica de caja y bigote
- A describir la covarianza y el coeficiente de correlación

USO DE LA ESTADÍSTICA

Evaluación de los rendimientos de los fondos de inversión

Retomemos el estudio de los fondos de inversión presentado en el capítulo 2. Usted debe decidir en qué clases de fondos invertir. En el capítulo anterior se estudió cómo *presentar* datos en tablas y gráficas. Sin embargo, al ocuparse de datos numéricos como el rendimiento de las inversiones en los fondos de inversión durante 2003, también necesita resumir los datos y plantear preguntas estadísticas. ¿Cuál es la tendencia central del rendimiento de los diversos fondos? Por ejemplo, ¿cuál fue el rendimiento promedio de los fondos de inversión con riesgo bajo, medio y alto durante 2003? ¿Qué tanta variabilidad hay en los rendimientos? ¿El rendimiento de los fondos de alto riesgo varía más que el correspondiente a los de riesgo promedio o bajo? ¿Cómo puede utilizar esta información al decidir en cuáles fondos invertir?

Para las variables numéricas, usted necesita más que la simple imagen visual de una variable obtenida a partir de las gráficas analizadas en el capítulo 2. Por ejemplo, a usted le gustaría determinar no sólo si durante 2003 los fondos más riesgosos tuvieron un rendimiento superior, sino también si tuvieron más variación y cómo se distribuyeron en cada grupo de riesgo. También desea examinar si existe alguna relación entre el coeficiente de gastos y los rendimientos de 2003. La lectura de este capítulo le permitirá aprender sobre algunos métodos de medición:

- **Tendencia central**, es la medida que describe cómo todos los valores de los datos se agrupan en torno a un valor central.
- **Variación**, es la cantidad de disgregación o dispersión de los valores con respecto a un valor central.
- **Forma**, es el patrón de distribución de los valores desde el menor hasta el mayor.

También aprenderá sobre la covarianza y el coeficiente de correlación, que ayudan a medir la fuerza de asociación entre dos variables numéricas.

3.1 MEDIDAS DE TENDENCIA CENTRAL, VARIACIÓN Y FORMA

Es posible caracterizar cualquier conjunto de datos numéricos por la medición de su tendencia central, variación y forma. La mayoría de los conjuntos de datos presentan una tendencia central a agruparse en torno a un valor central. Cuando la gente habla de un "promedio", o "valor medio", o del valor más común o frecuente, se refiere de manera informal a la media, la mediana y la moda, tres medidas de tendencia central.

La variación mide la **distribución** o **dispersión** de valores que conforman el conjunto de datos. Una medida simple de la variación es el rango, que es la diferencia entre los valores máximo y mínimo. En la estadística, son de uso más común la desviación estándar y la varianza, dos medidas que se explican más adelante en esta sección. La forma de un conjunto de datos representa un patrón para todos los valores, desde el mínimo hasta el máximo. Como se observará más adelante en esta sección, muchos conjuntos de datos tienen un patrón semejante a una campana, cuya cima de valores está en alguna parte del centro.

La media

La **media aritmética** (por lo general llamada la **media**) es la medida más común de la tendencia central. La media es la medida más común en la que todos los valores desempeñan el mismo papel. La media sirve como "punto de equilibrio" del conjunto de datos (como el punto de apoyo de un balancín). La media se calcula sumando todos los valores del conjunto de datos y dividiendo el resultado por el *número* de valores considerados.

Para representar a la media de una muestra, utilice el símbolo $\overline{X}$, llamado *X testada*. Si se considera una muestra que contiene n valores, la ecuación de su media se escribe como:

$$\overline{X} = \frac{\text{suma de los valores}}{\text{número de valores}}$$

Al utilizar la serie $X_1, X_2, \ldots, X_n$ para representar al conjunto de n valores y n para representar al número de valores, la ecuación se convierte en:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Al utilizar la notación de sumatoria (que se explica en el apéndice B), reemplace el numerador $X_x + X_2 + \cdots + X_n$ por el término $\sum_{i=1}^{n} X_i$, que significa la suma de todos los valores X_i desde el primer valor de X, que es X_1, hasta el último valor de X, que es X_n, para formar la ecuación (3.1), una definición formal de la media de una muestra.

MEDIA DE UNA MUESTRA

La **media de una muestra** es la suma de los valores dividida por el número de valores.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \textbf{(3.1)}$$

donde

$\overline{X}$ = media de la muestra

n = número de valores o tamaño de la muestra

X_i = i-ésimo valor de la variable X

$\sum_{i=1}^{n} X_i$ = sumatoria de todos los valores X_i de la muestra

Como todos los valores desempeñan un papel semejante, una media se verá muy afectada por cualquier valor que difiera mucho de los demás en el conjunto de datos. Cuando tenga tales valores extremos, debe evitar el uso de la media.

La media sugiere cuál es un valor "típico" o central del conjunto de datos. Por ejemplo, si usted conoce el tiempo que le lleva arreglarse por las mañanas, podrá planear mejor su inicio del día y reducir al mínimo cualquier retraso (o adelanto) para llegar a su destino. Suponga que define en minutos (redondeando al minuto más cercano) el tiempo que le lleva arreglarse, desde que se levanta hasta que sale de casa. A lo largo de 10 días hábiles consecutivos, usted recaba los tiempos que se muestran a continuación: **TIMES**

Día:	1	2	3	4	5	6	7	8	9	10
Tiempo (minutos):	39	29	43	52	39	44	40	31	44	35

El tiempo medio es 39.6 minutos, que se calculó como sigue:

$$\bar{X} = \frac{\text{suma de los valores}}{\text{número de valores}}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{X} = \frac{39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10}$$

$$\bar{X} = \frac{396}{10} = 39.6$$

A pesar de que ni un solo día de la muestra tuvo en realidad el valor de 39.6 minutos, asignar 40 minutos a su arreglo personal sería un buen criterio para planear su inicio del día, pero sólo porque esos 10 días no contienen ningún valor extremo.

Compare lo anterior con el caso en que el valor del cuatro día fue de 102 minutos en lugar de 52. Este valor extremo provocaría que la media aumentara a 44.6 minutos, como se observa a continuación:

$$\bar{X} = \frac{\text{suma de los valores}}{\text{número de valores}}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$\bar{X} = \frac{446}{10} = 44.6$$

Un valor extremo elevó la media en más del 10%, de 39.6 a 44.6 minutos. En contraste con la media original, que estaba "en medio", mayor que cinco de los tiempos (y menor que los otros cinco), la nueva media es mayor que 9 de los 10 tiempos de arreglo. El valor extremo provocó que la media sea una mala medida de tendencia central.

EJEMPLO 3.1 EL RENDIMIENTO MEDIO EN 2003 DE LOS FONDOS DE INVERSIÓN PARA PEQUEÑOS CAPITALES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule el rendimiento medio en 2003 de los fondos de inversión de alto riesgo para capitales reducidos.

SOLUCIÓN

El rendimiento medio en 2003 de los fondos de inversión para capitales reducidos (**MUTUALFUNDS2004**) es 51.53 calculados de la siguiente manera:

$$\bar{X} = \frac{\text{suma de los valores}}{\text{número de valores}}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \frac{463.8}{9} = 51.53$$

El arreglo ordenado de los nueve fondos de inversión de alto riesgo para pequeños capitales es:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

Cuatro de estos rendimientos están por debajo de la media de 51.53, y cinco están por encima de ella.

La mediana

La **mediana** es el valor que divide en dos partes iguales a un conjunto de datos ya ordenado. La mediana no se ve afectada por los valores extremos, de manera que puede utilizarse cuando están presentes.

La mediana es el valor medio de un conjunto de datos ordenado de menor a mayor.

Para calcular la mediana del conjunto de datos, primero ordene los valores de menor a mayor. Utilice la ecuación (3.2) para calcular la clasificación del valor que corresponde a la mediana.

MEDIANA

El 50% de los valores son menores que la mediana y el otro 50% son mayores.

$$\text{Mediana} = \frac{n+1}{2} \text{ valor clasificado} \qquad \textbf{(3.2)}$$

Calcule el valor de la mediana siguiendo una de las dos reglas siguientes:

- ***Regla 1*** Si en el conjunto de datos hay un número *impar* de valores, la mediana es el valor colocado en medio.
- ***Regla 2*** Si en el conjunto de datos hay un número *par* de valores, entonces la mediana es el *promedio* de los dos valores colocados en medio.

Para calcular la mediana de la muestra de los 10 tiempos para arreglarse en las mañanas, los tiempos diarios se ordenan de la siguiente manera:

Valores ordenados:

29 31 35 39 39 40 43 44 44 52

Clasificación:

1 2 3 4 5 6 7 8 9 10

↑

Mediana = 39.5

Puesto que para esta muestra de 10 elementos el resultado de dividir $n + 1$ por 2 es $(10 + 1)/2 = 5.5$, debe utilizarse la regla 2 y promediar los valores clasificados quinto y sexto, 39 y 40. Por lo tanto, la mediana es 39.5. Una mediana de 39.5 significa que la mitad de los días, el tiempo necesario para arreglarse es menor o igual que 39.5 minutos, y la otra mitad de los días es mayor o igual que 39.5 minutos. Esta mediana de 39.5 minutos es muy cercana a la media del tiempo para arreglarse de 39.6 minutos.

EJEMPLO 3.2

CÁLCULO DE LA MEDIANA DE UNA MUESTRA CON UN NÚMERO IMPAR DE ELEMENTOS

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y con el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule la mediana del rendimiento en 2003 de los nueve fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN

Puesto que para esta muestra de nueve elementos el resultado de dividir $n + 1$ por 2 es $(9 + 1)/2 = 5$, al utilizar la regla 1, la mediana es el valor clasificado como quinto. Ordene el porcentaje del rendimiento en 2003 de los nueve fondos de inversión de alto riesgo para pequeños capitales de menor a mayor:

Valores ordenados:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

Clasificación:

1 2 3 4 5 6 7 8 9

↑
Mediana

La mediana del rendimiento es 53.8. La mitad de estos fondos de inversión tienen rendimientos iguales o menores que 53.8 y la otra mitad tiene rendimientos iguales o superiores.

La moda

La **moda** es el valor del conjunto de datos que aparece con mayor frecuencia. Al igual que en la mediana y a diferencia de la media, los valores extremos no afectan a la moda. Usted sólo debe utilizar la media con propósitos descriptivos, ya que varía más de una muestra a otra que la media o la mediana. Con frecuencia, en un conjunto de datos no existe moda, o bien, hay varias modas. Por ejemplo, considere los datos de tiempo para arreglarse que se muestran a continuación.

29 31 35 39 39 40 43 44 44 52

Existen dos modas, 39 y 44 minutos, ya que cada uno de estos valores aparece dos veces.

EJEMPLO 3.3 CÁLCULO DE LA MODA

El gerente de sistemas encargado de la red de una empresa lleva un registro del número de fallas del servidor que se presentan por día. Calcule la moda de los siguientes datos, que representan el número de fallas diarias del servidor durante las últimas dos semanas.

1 3 0 3 26 2 7 4 0 2 3 3 6 3

SOLUCIÓN
El arreglo ordenado de estos datos es:

0 0 1 2 2 3 3 3 3 3 4 6 7 26

Como el 3 aparece cinco veces, más que ningún otro valor, la moda es 3. De esta forma, el gerente de sistemas se dará cuenta de que la situación más común es la presencia de tres fallas del servidor al día. Para este conjunto de datos, la mediana también es igual a 3, mientras que la media es de 4.5. El valor extremo de 26 es atípico. Con estos datos, la mediana y la moda miden la tendencia central mejor que la moda.

Un conjunto de datos no tiene moda cuando ninguno de los valores es "más frecuente". En el ejemplo 3.4 aparece un conjunto de datos sin moda.

EJEMPLO 3.4 DATOS SIN MODA

Calcule la moda del rendimiento medio en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN
El arreglo ordenado para estos datos es:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

Estos datos no tienen moda. Ninguno de sus valores aparece con mayor frecuencia; cada uno aparece sólo una vez.

Cuartiles

Los **cuartiles** dividen a un conjunto de datos en cuatro partes iguales: el **primer cuartil Q_1** separa al 25.0%, que abarca a los valores más pequeños, del 75.0% restante, constituido por los que son mayores. El **segundo cuartil Q_2** es la mediana: 50.0% de sus valores son menores que la mediana y 50.0% son mayores. El **tercer cuartil Q_3** separa al 25.0%, que abarca a los valores más grandes, del 75.0% restante constituido por los que son menores. Las ecuaciones (3.3) y (3.4) definen a los cuartiles primero y tercero.[1]

*[1]El Q_1, la mediana y el Q_3 también son el 25, 50 y 75° percentil, respectivamente. Por lo general, las ecuaciones (3.2), (3.3) y (3.4) se expresan en términos de cálculo de percentiles: percentil (p * 100)° = valor clasificado p * (n + 1).*

PRIMER CUARTIL Q_1

El 25.0% de los valores son menores que el primer cuartil Q_1, y el 75.0% son mayores que el primer cuartil Q_1.

$$Q_1 = \frac{n+1}{4} \text{ valor clasificado} \qquad \textbf{(3.3)}$$

TERCER CUARTIL Q_3

El 75.0% de los valores son menores que el tercer cuartil Q_3, y el 25.0% son mayores que el tercer cuartil Q_3.

$$Q_3 = \frac{3(n+1)}{4} \text{ valor clasificado} \qquad \textbf{(3.4)}$$

Para calcular los cuartiles, se utilizan las siguientes reglas:

- ***Regla 1*** Si el resultado es un número entero, entonces el cuartil es igual al valor clasificado. Por ejemplo, si el tamaño de la muestra es $n = 7$, el primer cuartil Q_1 es igual a (7 + l)/4 = segundo valor clasificado.
- ***Regla 2*** Si el resultado es una fracción de mitad (2.5, 4.5, etcétera), entonces el cuartil es igual al promedio de los valores clasificados correspondientes. Por ejemplo, si el tamaño de la muestra es $n = 9$, el primer cuartil Q_1 es igual al valor clasificado como (9 + l)/4 = 2.5, la mitad entre los valores clasificados como segundo y tercero.
- ***Regla 3*** Si el resultado no es un número entero ni una fracción de mitad, se redondea al entero más cercano y se selecciona ese valor clasificado. Por ejemplo, si el tamaño de la muestra es $n = 10$, el primer cuartil Q_1 es igual a (10 + l)/4 = valor clasificado como 2.75. Se redondea el 2.75 a 3 y se utiliza en valor clasificado como tercero.

Con el fin de ilustrar el cálculo de los cuartiles para los datos referentes a los tiempos para arreglarse, se ordenan de menor a mayor.

Valores ordenados:

29	31	35	39	39	40	43	44	44	52

Clasificación:

1	2	3	4	5	6	7	8	9	10

El primer cuartil es el valor clasificado como $(n + l)/4 = (10 + 1)/4 = 2.75$. Al emplear la tercera regla de los cuartiles, redondeamos al tercer valor clasificado. Para los datos sobre el tiempo necesario para arreglarse, el valor clasificado como tercero es 35 minutos. Interprete el primer cuartil de 35 como que el 25% de los días el tiempo necesario para arreglarse es menor o igual a 35 minutos, y que el 75% de los días ese tiempo es mayor o igual a 35 minutos.

El tercer cuartil es el valor clasificado como $3(n + l)/4 = 3(10 + l)/4 = 8.25$. Empleando la tercera regla de los cuartiles, redondeamos al valor clasificado como octavo. El valor clasificado como octavo en los datos del tiempo necesario para arreglarse es de 44 minutos. Interprete esto como que el 75% de los días, el tiempo necesario para arreglarse es menor o igual que 44 minutos, y que el 25% de los días ese tiempo es mayor o igual que 44 minutos.

EJEMPLO 3.5 CÁLCULO DE LOS CUARTILES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño de capital invertido (pequeño, mediano y gran capital). Calcule el primer cuartil (Q_1) y el tercer cuartil (Q_3) del rendimiento en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN

Ordenados de menor a mayor, los porcentajes de rendimiento de los nueve fondos de inversión de alto riesgo para pequeños capitales durante 2003 son:

Valor clasificado:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

Clasificación:

1 2 3 4 5 6 7 8 9

Para estos datos:

$$Q_1 = \frac{(n+1)}{4} \text{ valor clasificado}$$

$$= \frac{9+1}{4} = 2.5 \text{ valor clasificado}$$

Por lo tanto, al utilizar la segunda regla, resulta que Q_1 es el valor clasificado como 2.5, que está justo a la mitad entre los valores clasificados como segundo y tercero. Como el valor clasificado como segundo es 39.2 y el tercero es 44.2, el primer cuartil Q_1 es el que está justo en medio de 39.2 y 44.2. De esta forma,

$$Q_1 = \frac{39.2 + 44.2}{2} = 41.7$$

Para encontrar el tercer cuartil Q_3:

$$Q_3 = \frac{3(n+1)}{4} \text{ valor clasificado}$$

$$= \frac{3(9+1)}{4} = 7.5 \text{ valor clasificado}$$

Así, al utilizar la segunda regla, Q_3 es el valor clasificado entre los valores séptimo y octavo. Como el valor clasificado como séptimo es 59.3 y el octavo es 62.4, el tercer cuartil Q_3 es el que está justo en medio de 59.3 y 62.4. De esta forma,

$$Q_3 = \frac{59.3 + 62.4}{2} = 60.85$$

Un primer cuartil de 41.7 señala que el 25% de los rendimientos obtenidos durante 2003 por los fondos de alto riesgo para pequeños capitales fueron menores o iguales que 41.7, mientras que el 75% de ellos fueron mayores o iguales que 41.7. El tercer cuartil de 60.85 indica que el 75% de los rendimientos obtenidos durante el mismo año por los fondos de alto riesgo para pequeños capitales fueron menores o iguales que 60.85 y que el 25% fueron mayores o iguales que 60.85.

La media geométrica

La media geométrica y la razón geométrica de rendimiento miden el estado de una inversión en el tiempo. La **media geométrica** mide la razón de cambio de una variable en el tiempo. La ecuación 3.5 define a la media geométrica.

MEDIA GEOMÉTRICA

La media geométrica es la raíz *n*-ésima del producto de *n* valores

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n} \qquad \textbf{(3.5)}$$

La ecuación 3.6 define a la media geométrica de la tasa de rendimiento.

MEDIA GEOMÉTRICA DE LA TASA DE RENDIMIENTO

$$\overline{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \qquad \textbf{(3.6)}$$

donde R_i es la tasa de rendimiento durante el periodo i

Para ilustrar el uso de estas medidas, considere una inversión de $100,000 que se reduce hasta tener un valor de $50,000 al final del año 1 y luego recupera su valor original de $100,000 al finalizar el año 2. La tasa de rendimiento de esta inversión en el periodo de dos años es 0, porque los valores inicial y final permanecen sin cambio. Sin embargo, la media aritmética de las tasas de rendimiento anuales de esta inversión es

$$\overline{X} = \frac{(-0.50) + (1.00)}{2} = 0.25 \text{ o } 25\%$$

ya que la tasa de rendimiento del año 1 es

$$R_1 = \left(\frac{50,000 - 100,000}{100,000}\right) = -0.50 \text{ o } -50\%$$

y la tasa de rendimiento del año 2 es

$$R_2 = \left(\frac{100,000 - 50,000}{50,000}\right) = 1.00 \text{ o } 100\%$$

Al utilizar la ecuación (3.6), se sabe que la media geométrica de la tasa de rendimiento para los dos años es

$$\begin{aligned}
\overline{R}_G &= [(1 + R_1) \times (1 + R_2)]^{1/n} - 1 \\
&= [(1 + (-0.50)) \times (1 + (1.0))]^{1/2} - 1 \\
&= [(0.50) \times (2.0)]^{1/2} - 1 \\
&= [1.0]^{1/2} - 1 \\
&= 1 - 1 = 0
\end{aligned}$$

Por lo tanto, la media geométrica de la tasa de rendimiento refleja con mayor exactitud el cambio (cero) del valor de la inversión durante el periodo de dos años de la media aritmética.

EJEMPLO 3.6 CALCULE LA MEDIA GEOMÉTRICA DE LA TASA DE RENDIMIENTO

El porcentaje de cambio del índice compuesto NASDAQ fue del −31.53% en 2002 y del +50.01% en 2003. Calcule la tasa geométrica de rendimiento.

SOLUCIÓN
Al utilizar la ecuación (3.6), se sabe que la media geométrica de la tasa de rendimiento del índice NASDAQ para los dos años es

$$\begin{aligned}\overline{R}_G &= [(1 + R_1) \times (1 + R_2)]^{1/n} - 1 \\ &= [(1 + (-0.3153)) \times (1 + (0.5001))]^{1/2} - 1 \\ &= [(0.6847) \times (1.5001)]^{1/2} - 1 \\ &= [1.0271]^{1/2} - 1 \\ &= 1.0135 - 1 = 0.0135\end{aligned}$$

La media geométrica de la tasa de rendimiento del índice NASDAQ para los dos años es del 1.35%.

Rango

El **rango** es la medida numérica descriptiva más sencilla de la variación en un conjunto de datos.

RANGO

El rango es igual al valor mayor menos el valor menor.

$$\text{Rango} = X_{\text{mayor}} - X_{\text{menor}} \qquad \textbf{(3.7)}$$

Para determinar el rango de los tiempos necesarios para arreglarse, los datos se ordenan de menor a mayor:

29 31 35 39 39 40 43 44 44 52

Al emplear la ecuación (3.7), se sabe que el rango es de 52 − 29 = 23 minutos. Un rango de 23 minutos señala que la mayor diferencia del tiempo necesario para arreglarse por la mañana entre dos días cualesquiera es de 23 minutos.

EJEMPLO 3.7 CALCULE EL RANGO DEL RENDIMIENTO EN 2003 DE LOS FONDOS DE INVERSIÓN DE ALTO RIESGO PARA PEQUEÑOS CAPITALES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule el rango del rendimiento en 2003 de los nueve fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN
Ordenados de menor a mayor, los rendimientos en 2003 de los nueve fondos de inversión de alto riesgo para pequeños capitales son:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

Por lo tanto, al utilizar la ecuación 3.7, se sabe que el rango = 66.5 − 37.3 = 29.2.

La mayor diferencia entre dos rendimientos cualesquiera de los fondos de inversión de alto riesgo para pequeños capitales es de 29.2.

El rango mide la *distribución total* del conjunto de datos. Aunque el rango es una medida simple de la variación total de los datos, no toma en cuenta *cómo* se distribuyen los datos entre los valores menor y mayor. En otras palabras, el rango no indica si los valores están distribuidos de manera uniforme a todo lo largo del conjunto de datos, agrupados cerca de la parte media, o agrupados cerca de uno o ambos extremos. De esta manera, resulta engañoso utilizar el rango como medida de la variación cuando al menos uno de los valores es extremo.

Rango intercuartil

El **rango intercuartil** (también llamado **dispersión media**) es la diferencia entre el *tercer* y *primer cuartil* de un conjunto de datos.

RANGO INTERCUARTIL

El rango intercuartil es la diferencia entre los cuartiles tercero y primero.

$$\text{Rango intercuartil} = Q_3 - Q_1 \qquad \textbf{(3.8)}$$

El rango intercuartil mide la dispersión en la mitad (parte central) de los datos, así que no se ve influido por los valores extremos. Para determinar el rango intercuartil de los tiempos necesarios para arreglarse

29 31 35 39 39 40 43 44 44 52

utilice la ecuación (3.8) y los resultados obtenidos en la página 77, $Q_1 = 35$ y $Q_3 = 44$.

$$\text{Rango intercuartil} = 44 - 35 = 9 \text{ minutos}$$

Por lo tanto, el rango intercuartil del tiempo necesario para arreglarse es de 9 minutos. Por lo general, al intervalo de 35 a 44 se le denomina *la mitad media*.

EJEMPLO 3.8

CALCULE EL RANGO INTERCUARTIL DEL RENDIMIENTO EN 2003 DE LOS FONDOS DE INVERSIÓN DE ALTO RIESGO PARA PEQUEÑOS CAPITALES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule el rango intercuartil del rendimiento en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN

Ordenados de menor a mayor, los rendimientos de los nueve fondos de inversión de alto riesgo para pequeños capitales durante 2003 son:

37.3 39.2 44.2 44.5 53.8 56.6 59.3 62.4 66.5

Utilice la ecuación 3.8 y los resultados obtenidos en la página 78, $Q_1 = 41.7$ y $Q_3 = 60.85$.

$$\text{Rango intercuartil} = 60.85 - 41.7 = 19.15$$

Así, el rango intercuartil de los rendimientos en 2003 es de 19.15.

Como el rango intercuartil no toma en cuenta ningún valor menor que Q_1 ni mayor que Q_3, no se ve afectado por los valores extremos. Las medidas de resumen como la mediana, Q_1, Q_3, y el rango intercuartil, que no reciben la influencia de valores extremos, se denominan medidas resistentes.

La varianza y la desviación estándar

A pesar de que el rango y el rango intercuartil son medidas de la variación, no contemplan *cómo* se distribuyen o se agrupan los valores que están entre los extremos. La **varianza** y la **desviación estándar** son dos medidas de la variación muy utilizadas para tomar en cuenta cómo se distribuyen los datos. Estos estadísticos miden la dispersión "promedio" alrededor de la media, es decir, qué tanto varían los valores más grandes que están por encima de ella y cómo se distribuyen los valores menores que están por debajo de ella.

Una medida simple de la variación alrededor de la media consideraría la diferencia entre cada uno de los valores y la media, y luego las sumaría. Sin embargo, si usted hiciera eso, podría descubrir que la media es el punto de equilibrio de un conjunto de datos y que tales diferencias sumarían cero en *todo* conjunto de datos. Una medida de la variación que sería distinta de un conjunto de datos a otro consistiría en *elevar al cuadrado* la diferencia entre cada uno de los valores y la media, y después sumarlas. En estadística, esta cantidad se denomina **suma de cuadrados** (o **SS**). Esta suma luego se divide entre el número de valores menos 1 (para datos de la muestra), con el fin de obtener una varianza de la muestra (S^2). La raíz cuadrada de la varianza de la muestra es la desviación estándar de la muestra (S).

Puesto que la suma de cuadrados es una suma de diferencias elevadas al cuadrado que, por las reglas aritméticas siempre será no negativa, *ni la varianza ni la desviación estándar podrán ser negativas*. En casi todos los conjuntos de datos, la varianza y la desviación estándar tendrán un valor positivo, aunque si no existe variación en todo el conjunto de datos y todos los valores de la muestra son los mismos, ambos estadísticos serán igual a cero.

En una muestra que contiene n valores, $X_1, X_2, X_3, \ldots, X_n$, la varianza de la muestra (representada por el símbolo S^2) es

$$S^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n - 1}$$

La ecuación 3.9 expresa esta ecuación utilizando la notación de sumatoria.

VARIANZA PARA UNA MUESTRA

La **varianza para una muestra** es la suma de las diferencias con respecto a la media elevada al cuadrado y dividida por el tamaño de la muestra menos uno.

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1} \qquad \textbf{(3.9)}$$

donde

$\overline{X}$ = media

n = tamaño de la muestra

X_i = i-ésimo valor de la variable X

$\sum_{i=1}^{n} (X_i - \overline{X})^2$ = sumatoria de los cuadrados de todas las diferencias entre los valores de X_i y $\overline{X}$.

DESVIACIÓN ESTÁNDAR DE LA MUESTRA

La **desviación estándar de una muestra** es la raíz cuadrada de la suma de los cuadrados de las diferencias con respecto a la media dividida por el tamaño de la muestra menos uno.

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}} \qquad \textbf{(3.10)}$$

Si el denominador fuese n en vez de $n - 1$, la ecuación (3.9) [y el término interno de la ecuación (3.10)] calcularía el promedio de las diferencias con respecto a la media elevadas al cuadrado. Sin embargo, se utiliza $n - 1$ porque ciertas propiedades matemáticas convenientes del estadístico S^2 lo hacen apropiado para la inferencia estadística (que analizaremos en el capítulo 7). Conforme aumenta el tamaño de la muestra, se hace cada vez más pequeña la diferencia entre dividir por n o por $n - 1$.

Es más probable que usted utilice la desviación estándar de la muestra como medida de la variación [definida en la ecuación (3.10)]. A diferencia de la varianza de la muestra, que es una cantidad elevada al cuadrado, la desviación estándar siempre es un número con las mismas unidades que los datos de muestra originales. La desviación estándar le ayuda a conocer de qué manera se agrupan o distribuyen un conjunto de datos con respecto a su media. En casi todos los conjuntos de datos, la mayoría de los valores observados quedan dentro de un intervalo de más menos una desviación estándar por encima y por debajo de la media. Por esa razón, conocer la media y la desviación estándar ayuda a definir por lo menos dónde se agrupa la mayoría de los valores de los datos.

Para calcular a mano la varianza S^2 y la desviación estándar S de una muestra:

Paso 1: Calcule la diferencia entre cada uno de los valores y la media.

Paso 2: Eleve al cuadrado cada una de esas diferencias.

Paso 3: Sume las diferencias elevadas al cuadrado.

Paso 4: Divida el total entre $n - 1$, para obtener la varianza de la muestra.

Paso 5: Extraiga la raíz cuadrada de la varianza de la muestra, para obtener la desviación estándar de la muestra.

La tabla 3.1 muestra los cuatro primeros pasos para calcular la varianza de los datos referentes al tiempo necesario para arreglarse, con una media ($\overline{X}$) = 39.6 (vea el cálculo de la media en la página 74). En la segunda columna se muestra el paso 1. En la tercera columna se muestra el paso 2. En la parte inferior se muestra la suma de la diferencias elevadas al cuadrado (paso 3). Luego, este total se divide entre 10 – 1 = 9, para calcular la varianza (paso 4).

TABLA 3.1
Cálculo de la varianza del tiempo necesario para arreglarse.

$\overline{X} = 39.6$

Tiempo (X)	***Paso 1:*** **($X_i - \overline{X}$)**	***Paso 2:*** **$(X_i - \overline{X})^2$**
39	–0.60	0.36
29	–10.60	112.36
43	3.40	11.56
52	12.40	153.76
39	–0.60	0.36
44	4.40	19.36
40	0.40	0.16
31	–8.60	73.96
44	4.40	19.36
35	–4.60	21.16
	Paso 3: **Suma:**	***Paso 4:*** **Dividido por (n – 1):**
	412.40	45.82

También es posible calcular la varianza si se sustituyen los valores de los términos en la ecuación 3.9:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$
$$= \frac{(39 - 39.6)^2 + (29 - 39.6)^2 + \cdots + (35 - 39.6)^2}{10 - 1}$$
$$= \frac{412.4}{9}$$
$$= 45.82$$

Puesto que la varianza está en unidades cuadradas (en minutos cuadrados en este caso), para calcular la desviación estándar se calcula la raíz cuadrada de la varianza. Al utilizar la ecuación (3.10) de la página 82, la desviación estándar S de la muestra es:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{45.82} = 6.77$$

Esto indica que los tiempos necesarios para arreglarse en esta muestra se agrupan dentro de los 6.77 minutos que circundan a la media de 39.6 minutos (es decir, se agrupan entre $\bar{X} - 1S = 32.83$ y $\bar{X} + 1S = 46.37$). De hecho, 7 de los 10 quedan dentro de este intervalo.

Al utilizar la segunda columna de la tabla 3.1, también es posible calcular que la suma de las diferencias entre cada uno de los valores y la media es cero. Para todo conjunto de datos, esta suma siempre será igual a cero:

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0 \text{ para todos los conjuntos de datos}$$

Esta propiedad es una de las razones por las que la media se utiliza como la medida más común de tendencia central.

EJEMPLO 3.9

CÁLCULO DE LA VARIANZA Y LA DESVIACIÓN ESTÁNDAR DEL RENDIMIENTO EN 2003 DE LOS FONDOS DE INVERSIÓN PARA PEQUEÑOS CAPITALES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule la varianza y la desviación estándar del rendimiento en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN

La tabla 3.2 ilustra el cálculo de la varianza y la desviación estándar del rendimiento en 2003 para los fondos de inversión de alto riesgo para pequeños capitales. Utilice la ecuación (3.9) de la página 82:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$
$$= \frac{(44.5 - 51.53)^2 + (39.2 - 51.53)^2 + \cdots + (66.5 - 51.53)^2}{9 - 1}$$
$$= \frac{891.16}{8}$$
$$= 111.395$$

TABLA 3.2
Cálculo de la varianza del rendimiento en 2003 para los fondos de inversión de alto riesgo para pequeños capitales.

$\overline{X} = 51.5333$

Rendimiento 2003	***Paso 1:*** $(X_i - \overline{X})$	***Paso 2:*** $(X_i - \overline{X})^2$
44.5	−7.0333	49.4678
39.2	−12.3333	152.1111
62.4	10.8667	118.0844
59.3	7.7667	60.3211
56.6	5.0667	25.6711
53.8	2.2667	5.1378
37.3	−14.2333	202.5878
44.2	−7.3333	53.7778
66.5	14.9667	224.0011
	Paso 3: **Suma:**	***Paso 4:*** **Dividido por (n − 1):**
	891.16	111.395

Al utilizar la ecuación (3.10) de la página 82, se sabe que la desviación estándar S de la muestra es:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} = \sqrt{111.395} = 10.55$$

La desviación estándar de 10.55 indica que los rendimientos en 2003 de los fondos de inversión de alto riesgo para pequeños capitales se agrupan dentro de los 10.55 que rodean a la media de 51.53 (es decir, se agrupan entre $\overline{X} - 1S = 40.98$ y $\overline{X} + 1S = 62.08$). De hecho, el 55.6% (5 de 9) de los rendimientos en 2003 quedan dentro de este intervalo.

A continuación se resumen las características del rango, del rango intercuartil, de la varianza y de la desviación estándar.

- Cuanto más esparcidos o dispersos están los datos, son mayores el rango, el rango intercuartil, la varianza y la desviación estándar.
- Cuanto más concentrados u homogéneos son los datos, son menores el rango, el rango intercuartil, la varianza y la desviación estándar.
- Si todos los valores son los mismos (de tal manera que no hay variación de los datos), el rango, el rango intercuartil, la varianza y la desviación estándar son iguales a cero.
- Ninguna de las medidas de la variación (rango, rango intercuartil, desviación estándar y varianza) puede ser negativa.

Coeficiente de variación

A diferencia de las medidas de la variación antes expuestas, el **coeficiente de variación** es una *medida relativa* de la variación que siempre se expresa como porcentaje, más que en términos de las unidades de los datos en particular. El coeficiente de variación, que se denota mediante el símbolo *CV*, mide de dispersión de los datos con respecto a la media.

COEFICIENTE DE VARIACIÓN

El coeficiente de variación es igual a la desviación estándar dividida por la media, multiplicada por 100%.

$$CV = \left(\frac{S}{\overline{X}}\right)100\% \qquad \textbf{(3.11)}$$

donde S = desviación estándar de la muestra

$\overline{X}$ = media de la muestra

Para la muestra de los 10 tiempos para arreglarse, como $\overline{X} = 39.6$ y $S = 6.77$, el coeficiente de variación es

$$CV = \left(\frac{S}{\overline{X}}\right)100\% = \left(\frac{6.77}{39.6}\right)100\% = 17.10\%$$

Para estos datos, la desviación estándar es el 17.1% del tamaño de la media.

El coeficiente de variación es muy útil al comparar dos o más conjuntos de datos medidos con unidades distintas, como ilustra el ejemplo 3.10.

EJEMPLO 3.10

COMPARACIÓN DE DOS COEFICIENTES DE VARIACIÓN CUANDO DOS VARIABLES TIENEN DISTINTAS UNIDADES DE MEDIDA

El gerente de operaciones de un servicio de entrega de paquetería está pensando si es conveniente adquirir una nueva flota de camiones. Al guardar los paquetes en los camiones para su entrega, se deben tomar en cuenta dos características principales: el peso (en libras) y el volumen (en pies cúbicos) de cada artículo.

El gerente de operaciones toma una muestra de 200 paquetes, y encuentra que la media del peso es 26.0 libras, con una desviación estándar de 3.9 libras, mientras que la media en volumen es de 8.8 pies cúbicos, con una desviación estándar de 2.2 pies cúbicos. ¿Cómo puede el gerente de operaciones comparar la variación de peso y volumen?

SOLUCIÓN

Como las unidades difieren para el peso y volumen, el gerente de operaciones debe comparar la variabilidad relativa en ambos tipos de medidas.

Para el peso, el coeficiente de variación es

$$CV_W = \left(\frac{3.9}{26.0}\right)100\% = 15.0\%$$

para el volumen, el coeficiente de variación es

$$CV_V = \left(\frac{2.2}{8.8}\right)100\% = 25.0\%$$

De esta forma, en relación con la media el volumen del paquete es mucho más variable que su peso.

Puntuaciones *Z*

Un **valor extremo** o **atípico** es un valor ubicado muy lejos de la media. Las puntuaciones Z son útiles para identificar atípicos. Cuanto mayor es la puntuación Z, mayor es la distancia entre tal valor y la media. La **puntuación *Z*** es igual a la diferencia entre ese valor y la media, dividida por la desviación estándar.

PUNTUACIONES Z

$$Z = \frac{X - \overline{X}}{S} \qquad \textbf{(3.12)}$$

Si se consideran los tiempos necesarios para arreglarse por la mañana, se observa que la media es de 39.6 minutos y la desviación estándar de 6.77 minutos. El tiempo necesario para arreglarse el primer día es de 39.0 minutos. La puntuación Z para el día 1 se calcula a partir de

$$\begin{aligned} Z &= \frac{X - \overline{X}}{S} \\ &= \frac{39.0 - 39.6}{6.77} \\ &= -0.09 \end{aligned}$$

La tabla 3.3 muestra las puntuaciones Z de los 10 días. La mayor es de 1.83 para el día 4, cuando el tiempo necesario para arreglarse fue de 52 minutos. La menor fue −1.57 para el día 2, cuando el tiempo necesario para arreglarse fue de 29 minutos. Como regla general, una puntuación Z se considera atípica si es menor que −3.0 o mayor que +3.0. Ninguno de los tiempos satisface este criterio.

TABLA 3.3

Puntuaciones Z para los 10 tiempos necesarios para arreglarse.

	Tiempo (X)	Puntuación Z
	39	−0.09
	29	−1.57
	43	0.50
	52	1.83
	39	−0.09
	44	0.65
	40	0.06
	31	−1.27
	44	0.65
	35	−0.68
Media	39.6	
Desviación estándar	6.77	

EJEMPLO 3.11

CÁLCULO DE LAS PUNTUACIONES Z DEL RENDIMIENTO EN 2003 DE LOS FONDOS DE INVERSIÓN DE ALTO RIESGO PARA PEQUEÑOS CAPITALES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule las puntuaciones Z del rendimiento en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN

La tabla 3.4 ilustra las puntuaciones Z de los rendimientos en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. La puntuación Z más grande es 1.42, correspondiente a un rendimiento porcentual de 66.5. La puntuación Z más baja es −1.35, correspondiente a un rendimiento porcentual de 37.3. Como regla general, se considera que una puntuación Z es atípica si es menor que −3.0 o mayor que +3.0. Ninguno de los rendimientos porcentuales satisface el criterio para considerarlo atípico.

TABLA 3.4
Puntuación *Z* del rendimiento en 2003 de los fondos de inversión de alto riesgo para pequeños capitales.

	Rendimiento 2003	Puntuaciones *Z*
	44.5	–0.67
	39.2	–1.17
	62.4	1.03
	59.3	0.74
	56.6	0.48
	53.8	0.21
	37.3	–1.35
	44.2	–0.69
	66.5	1.42
Media	51.53	
Desviación estándar	10.55	

Forma

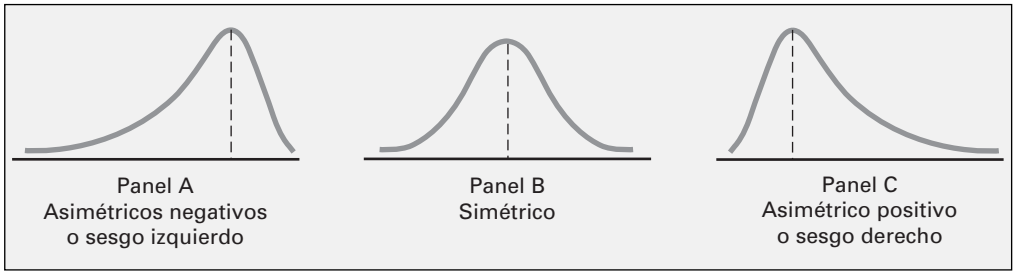
Una tercera e importante propiedad que describe a un conjunto de datos numéricos es la forma. Forma es el patrón de distribución de los valores de los datos a través del rango de todos los valores. La distribución puede ser **simétrica** cuando los valores pequeños y grandes se equilibran entre sí, o **asimétrica**, cuando muestra desequilibrio de los valores pequeños o grandes.

La forma influye en la relación de la media con la mediana de las siguientes maneras:

- Media < mediana; asimétrica negativa o sesgo izquierdo.
- Media = mediana; simétrica o asimetría cero.
- Media > mediana; asimétrica positiva o sesgo derecho.

La figura 3.1 describe tres conjuntos de datos, cada uno con distinta forma.

FIGURA 3.1
Comparación de tres conjuntos de datos con distinta forma.



Los datos del panel A son negativos, o **sesgados a la izquierda**. En este panel, la mayoría de los valores están en la parte superior de la distribución. Existe una cola larga y la distorsión hacia la izquierda es provocada por algunos valores muy pequeños. Estos valores extremadamente pequeños empujan la media hacia abajo, de manera que la media es menor que la mediana.

Los datos del panel B son simétricos. Cada mitad de la curva es una imagen al espejo del otro. Los valores bajos y altos de la escala se equilibran, y la media es igual a la mediana.

Los datos del panel C son **asimétricos positivos o sesgados a la derecha**. En este panel, la mayoría de los valores están en la parte inferior de la distribución. Existe una larga cola a la derecha de la distribución y cierta distorsión hacia la derecha provocada por algunos valores muy grandes. Estos valores sumamente grandes empujan a la media hacia arriba, de manera que la media resulta mayor que la mediana.

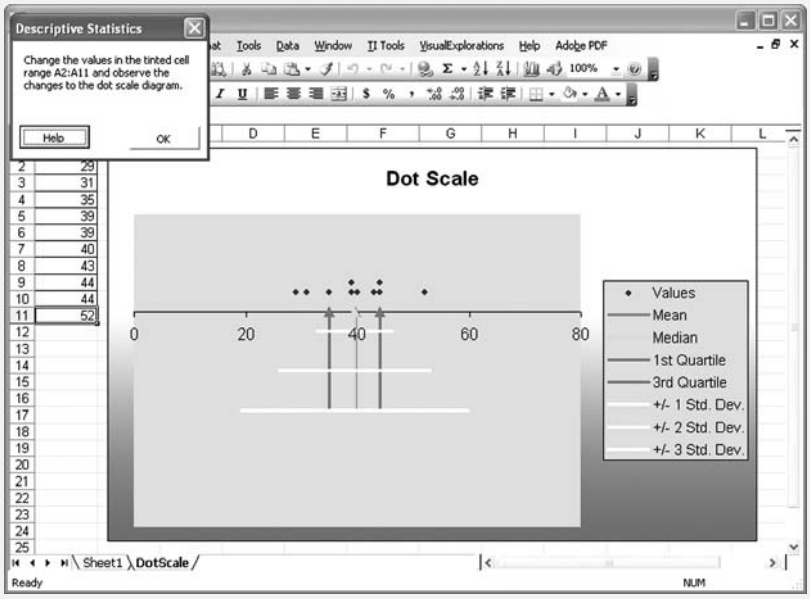
Resultados de la estadística descriptiva en Excel

El juego de herramientas de análisis de datos de Excel genera la media, mediana, moda, desviación estándar, varianza, rango, mínimo, máximo y cuenta (tamaño de la muestra) en una sola hoja de trabajo, todos ellos analizados en esta sección. Además, Excel calcula el error estándar, lo mismo que estadísticos para la curtosis y la asimetría. El *error estándar* es igual a la desviación estándar dividida por la raíz cuadrada del tamaño de la muestra, y se estudiará en el capítulo 7. La *asimetría* mide la falta de simetría en los datos, y se basa en un estadístico que está en función de las diferencias con

EXPLORACIONES VISUALES Exploración de la estadística descriptiva

Utilice el procedimiento Exploraciones Visuales de la Estadística Descriptiva para observar el efecto que tiene el cambio de valores en los datos sobre las medidas de tendencia central, variación y forma. Abra la macro de trabajo **Visual Explorations.xla** y seleccione **Visual Explorations ➜ Descriptive Statistics** en la barra de herramientas de Excel. Lea las instrucciones del cuadro que aparece (vea la ilustración que se muestra a continuación) y dé clic en **OK** para examinar el diagrama de puntos correspondiente a la muestra de 10 tiempos necesarios para arreglarse que utilizará a lo largo de este capítulo.

Experimente introduciendo un valor extremo como 10 minutos en una de las celdas de la columna A. ¿Qué medidas se ven afectadas por este cambio? ¿Cuáles no? Puede alternar entre los diagramas "previo" y "posterior" presionando repetidamente **Ctrl+Z** (deshacer) seguido de **Ctrl+Y** (rehacer) como ayuda para observar los cambios provocados por un valor extremo en el diagrama.



respecto a la media *elevadas al cubo*. Un valor de asimetría de cero indica una distribución simétrica. La *curtosis* mide la concentración relativa de valores en el centro de la distribución al compararlos con las colas y se basa en las diferencias con respecto a la media elevadas a la cuarta potencia. Esta medida no se analiza en el presente texto (vea la referencia 2).

A partir de la figura 3.2 de la página 90, los resultados de estadística descriptiva en Excel para el rendimiento de los fondos en 2003, con base en su nivel de riesgo, parecen mostrar ligeras diferencias para los tres niveles de riesgo en su rendimiento porcentual de 2003. Los fondos de alto riesgo tienen una media y una mediana ligeramente mayores que los de riesgo bajo y medio. Existe muy poca diferencia entre las desviaciones estándar de los tres grupos.

Resultados de la estadística descriptiva en Minitab

Para la estadística descriptiva, Minitab calcula el tamaño de la muestra (etiquetado como N), media, mediana, desviación estándar (etiquetada StDev), mínimo, máximo, coeficiente de variación (etiquetado CoefVar), primer y tercer cuartiles, rango y rango intercuartil (etiquetado IQR), todos analizados en esta sección.

A partir de la figura 3.3 de la página 90, los resultados de estadística descriptiva en Minitab para el rendimiento de los fondos en 2003, con base en su riesgo, parecen registrar ligeras diferencias del rendimiento porcentual en 2003 para los tres niveles de riesgo. Los fondos de alto riesgo tienen media, mediana y cuartiles ligeramente superiores a los de riesgo bajo y medio. Existe muy poca diferencia en las desviaciones estándar o el rango intercuartil de los tres grupos.

FIGURA 3.2
Estadística descriptiva en Excel para el rendimiento de los fondos en 2003 con base en su nivel de riesgo.

	A	B	C	D
1	Descriptive Statistics of 2003 Return by Risk			
2		Low	Average	High
3	Mean	41.36207	42.96304	45.99412
4	Standard Error	1.631596	2.045318	3.117886
5	Median	40.25	41	44.5
6	Mode	35.8	37.5	#N/A
7	Standard Deviation	12.42586	13.87202	12.85537
8	Sample Variance	154.402	192.433	165.2606
9	Kurtosis	-0.09275	-0.33478	0.711316
10	Skewness	0.358427	0.586116	-0.55254
11	Range	55.4	58.3	51.6
12	Minimum	15.8	19.7	14.9
13	Maximum	71.2	78	66.5
14	Sum	2399	1976.3	781.9
15	Count	58	46	17

FIGURA 3.3
Estadística descriptiva en Minitab para el rendimiento de los fondos en 2003 con base en su nivel de riesgo.

Results for: MUTUALFUNDS2004.MTW

Descriptive Statistics: Return 2003

Variable	Risk	N	Mean	StDev	CoefVar	Minimum	Q1	Median	Q3
Return 2003	average	46	42.96	13.87	32.29	19.70	32.63	41.00	49.83
	high	17	45.99	12.86	27.95	14.90	38.25	44.50	57.05
	low	58	41.36	12.43	30.04	15.80	31.23	40.25	48.95

Variable	Risk	Maximum	Range	IQR
Return 2003	average	78.00	58.30	17.20
	high	66.50	51.60	18.80
	low	71.20	55.40	17.73

PROBLEMAS PARA LA SECCIÓN 3.1

Aprendizaje básico

ASISTENCIA de PH Grade **3.1** A continuación se encuentra un conjunto de datos procedente de una muestra de $n = 5$:

7 4 9 8 2

a. Calcule la media, la mediana y la moda.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Calcule las puntuaciones *Z*. ¿Existe algún valor extremo?
d. Describa la forma del conjunto de datos.

ASISTENCIA de PH Grade **3.2** A continuación aparece un conjunto de datos procedente de una muestra de $n = 6$:

7 4 9 7 3 12

a. Calcule la media, la mediana y la moda.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Calcule las puntuaciones *Z*. ¿Existe algún valor extremo?
d. Describa la forma del conjunto de datos.

ASISTENCIA de PH Grade **3.3** A continuación aparece un conjunto de datos procedente de una muestra de $n = 7$:

12 7 4 9 0 7 3

a. Calcule la media, la mediana y la moda.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Describa la forma del conjunto de datos.

ASISTENCIA de PH Grade **3.4** A continuación aparece un conjunto de datos procedente de una muestra de $n = 5$:

7 −5 −8 7 9

a. Calcule la media, la mediana y la moda.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Describa la forma del conjunto de datos.

ASISTENCIA de PH Grade **3.5** Suponga que la tasa de rendimiento de una acción en particular durante los dos últimos años fue del 10 y del 30%. Calcule la media geométrica de la tasa de rendimiento (*Nota:* Una tasa de rendimiento del 10% se registra como 0.10 y una del 30% como 0.30).

Aplicación de conceptos

Puede resolver los problemas 3.6 a 3.20 manualmente o en Excel, Minitab o SPSS.

ASISTENCIA de PH Grade **3.6** El gerente de operaciones de una fábrica de llantas quiere comparar el diámetro interno real de dos tipos de neumáticos, que se espera sean de 575 milímetros en ambos casos. Se seleccionó una muestra de cinco llantas de cada tipo y se ordenaron de menor a mayor, como se aprecia a continuación:

Tipo *X*					Tipo *Y*				
568	570	575	578	584	573	574	575	577	578

a. Calcule la media, la mediana y la desviación estándar de ambos tipos de llantas.
b. ¿Cuál tipo de llanta es de mejor calidad? Explique por qué.
c. ¿Qué efecto tendría en sus respuestas a los incisos *a*) y *b*) si el último valor del tipo *Y* fuese 588 en lugar de 578? Explique su respuesta.

ASISTENCIA de PH Grade **3.7** Los siguientes datos representan el total de grasas en las hamburguesas y productos de pollo de una muestra tomada de cadenas de comida rápida. **FASTFOOD**

Hamburguesas

19 31 34 35 39 39 43

Pollo

7 9 15 16 16 18 22 25 27 33 39

Fuente: "Quick bites", Derechos reservados © 2000 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adoptado con autorización de Consumer Reports*, marzo de 2001, 46.*

Para las hamburguesas y los productos de pollo realice lo siguiente por separado:
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil y el coeficiente de variación.
c. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. Con base en los resultados de los incisos *a*) a *c*), ¿qué conclusiones se obtienen en relación con las diferencias en la grasa total de las hamburguesas y los productos de pollo?

3.8 La mediana del precio de una casa en diciembre de 2003 alcanza $173,200, un incremento del 6.7% respecto a diciembre de 2002. En todo el año, las ventas alcanzaron un récord de 6.1 millones de casas (James R. Hagerty, "Housing Prices Continue to Rise", *The Wall Street Journal*, 27 de enero, 2004, D1).
a. Describa la forma de la distribución correspondiente al precio de las casas vendidas.
b. ¿Por qué cree usted que el artículo informa sobre la mediana de los precios y no sobre la media?

3.9 En el ciclo escolar 2002-2003, muchas universidades públicas de Estados Unidos elevaron sus cuotas y tarifas de manutención, como resultado de la reducción de los subsidios estatales (Mary Beth Marklein, "Public Universities Raise Tuition, Fees-and Ire", *USA Today*, 8 de agosto, 2002, 1A-2A). A continuación se representa el cambio del costo de inscripción, un dormitorio compartido y el plan de alimentación más solicitado entre los ciclos escolares 2001-2002 y 2002-2003 en una muestra de 10 universidades públicas. **COLLEGECOST**

Universidad	Cambio en el costo ($)
University of California, Berkeley	1,589
University of Georgia, Athens	593
University of Illinois, Urbana-Champaign	1,223
Kansas State University, Manhattan	869
University of Maine, Orono	423
University of Mississippi, Oxford	1,720
University of New Hampshire, Durham	708
Ohio State University, Columbus	1,425
University of South Carolina, Columbia	922
Utah State University, Logan	308

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil, el coeficiente de variación y las puntuaciones *Z*.
c. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. Con base en los resultados de los incisos *a*) a *c*), ¿qué conclusiones se obtienen en relación con el cambio de los costos entre los ciclos escolares 2001-2002 y 2002-2003?

3.10 Los siguientes datos **COFFEDRINK** representan las calorías y la grasa (en gramos), que contienen las raciones con 16 onzas de bebidas a base de café servidas en Dunkin' Donuts y Starbucks.

Producto	Calorías	Grasa
Batido de moka helado de Dunkin' Donuts (pura leche)	240	8.0
Capuchino frapé de Starbucks	260	3.5
Raspado de café "Coolata" (crema) de Dunkin' Donuts	350	22.0
Café moka exprés helado de Starbucks (pura leche y crema batida)	350	20.0
Café moka batido helado de Starbucks (con crema batida)	420	16.0
Capuchino helado de Brownie de chocolate, de Starbucks (con crema batida)	510	22.0
Crema de chocolate batido helado de Starbucks (con crema batida)	530	19.0

Fuente: "Coffee as Candy at Dunkin' Donuts and Starbucks", Derechos Reservados © 2004 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057, organización sin fines de lucro. Adaptado con autorización de Consumer Reports*, junio de 2004, 9, sólo con propósitos educativos. No se autoriza su reproducción o uso comercial.* **www.ConsumerReports.org**

Para cada una de las variables (calorías y grasa):
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil, el coeficiente de variación y las puntuaciones *Z*. ¿Existe un valor atípico? Explique su respuesta.
c. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. A partir de los resultados de los incisos *a*) a *c*), ¿qué conclusiones se obtienen en relación con las calorías y la grasa de las bebidas heladas a base de café servidas en Dunkin' Donuts y en Starbucks?

3.11 Los siguientes datos representan el costo diario de una habitación de hotel y la renta de un automóvil en 20 ciudades estadounidenses durante una semana en octubre de 2003. **HOTEL-CAR**

Ciudad	Hotel	Automóviles
San Francisco	205	47
Los Ángeles	179	41
Seattle	185	49
Phoenix	210	38
Denver	128	32
Dallas	145	48
Houston	177	49
Minneapolis	117	41
Chicago	221	56
St. Louis	159	41
Nueva Orleáns	205	50
Detroit	128	32
Cleveland	165	34
Atlanta	180	46
Orlando	198	41
Miami	158	40
Pittsburg	132	39
Boston	283	67
Nueva York	269	69
Washington, D.C.	204	40

Fuente: The Wall Street Journal, *10 de octubre, 2003, W4.*

Para cada una de las variables (costo de hotel y costo del auto):
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil, el coeficiente de variación y las puntuaciones *Z*. ¿Existe un valor extremo? Explique su respuesta.
c. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. Con base en los resultados de los incisos *a*) a *c*), ¿qué conclusiones se obtienen en relación con el costo diario de una habitación de hotel y la renta de un automóvil?

3.12 A continuación se indica el costo de 14 modelos de cámara digital de 3 megapixeles en una tienda especializada. **CAMERA**

340 450 450 280 220 340 290
370 400 310 340 430 270 380

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil, el coeficiente de variación y las puntuaciones *Z*. ¿Existe un valor atípico? Explique su respuesta.
c. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. Con base en los resultados de los incisos *a*) a *c*), ¿qué conclusiones se obtienen en relación con el precio de las cámaras digitales de 3 megapixeles en una tienda especializada durante 2003?

3.13 Una empresa dedicada a la consultoría y desarrollo de software, ubicada en el área metropolitana de Phoenix, desarrolla programas para sistemas administrativos de cadenas de suministro, con base en la reutilización sistemática de software. En lugar de comenzar desde cero al elaborar y desarrollar nuevos sistemas de software personalizados, utiliza una base de datos que contiene componentes reutilizables que suman más de 2,000,000 de líneas de código, recopilados a lo largo de 10 años de labores continuas. Se pide a 8 analistas de la empresa que calculen la tasa de reutilización cuando se desarrolla un nuevo sistema de software. Los siguientes datos corresponden al porcentaje total de código que procede de la base de datos de reutilización y forma parte del sistema de software. **REUSE**

50.0 62.5 37.5 75.0 45.0 47.5 15.0 25.0

Fuente: M. A. Rothenberger y K. J. Dooley, "A Performance Measure for Software Reuse Projects", Decision Sciences, *30 (otoño de 1999), 1131-1153.*

a. Calcule la media, la mediana y la moda.
b. Calcule el rango, la varianza y la desviación estándar.
c. Interprete las medidas sintetizadas que se calculan en los incisos *a*) y *b*).

3.14 Un fabricante de baterías para flashes toma una muestra de 13 baterías de la producción del día y las utiliza de manera continua hasta que se agotan. El número de horas que se utilizaron hasta el momento de fallar fue: **BATTERIES**

342 426 317 545 264 451
1,049 631 512 266 492 562 298

a. Calcule la media, la mediana y la moda. Al observar la distribución de los tiempos transcurridos hasta la falla, ¿cuáles medidas de ubicación le parecen más apropiadas y cuáles menos adecuadas para utilizarlas con estos datos? ¿Por qué?
b. Calcule el rango, la varianza y la desviación estándar.
c. ¿Qué le recomendaría a un fabricante si quisiera anunciar que sus baterías "duran 400 horas"? (*Nota:* No existe una respuesta exacta para esta pregunta; se trata de decir cómo hacer precisa tal afirmación.)
d. Suponga que, en lugar de 342, el primer valor fue de 1,342. Repita los incisos *a*) a *c*) utilizando este valor. Elabore un comentario sobre la diferencia de los resultados.

3.15 Una sucursal bancaria ubicada en una zona comercial de la ciudad, desarrolló un proceso mejorado para atender a sus clientes desde la hora del almuerzo al mediodía, hasta la 1:00 PM. Se registra el tiempo de espera en minutos (definido como el tiempo transcurrido desde que el cliente se forma en la fila hasta que llega a la ventanilla del cajero) de todos los clientes

durante ese horario por una semana. Se selecciona una muestra aleatoria de 15 clientes y se tienen los siguientes resultados: **BANK1**

4.21 5.55 3.02 5.13 4.77 2.34 3.54
3.20 4.50 6.10 0.38 5.12 6.46 6.19 3.79

a. Calcule la media, la mediana, primero y tercer cuartiles.

b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil, el coeficiente de variación y las puntuaciones *Z*. ¿Existe algún valor atípico? Explique su respuesta.

c. ¿Los datos son asimétricos? De ser así, ¿cómo?

d. Un cliente llega a la sucursal durante la hora del almuerzo y pregunta al gerente cuánto tendrá que esperar, éste le responde "Menos de cinco minutos, con toda seguridad". Con base en sus resultados de los incisos *a*) y *b*), evalúe la exactitud de tal afirmación.

3.16 Suponga que otra sucursal, ubicada en una zona residencial, también se preocupa por el tiempo de espera desde de la hora del almuerzo hasta la 1:00 PM. Se registra el tiempo de espera en minutos (definido como el tiempo transcurrido desde que el cliente se forma en la fila hasta que llega a la ventanilla del cajero) de todos los clientes durante ese horario por una semana. Se selecciona una muestra aleatoria de 15 clientes y se tienen los siguientes resultados: **BANK2**

9.66 5.90 8.02 5.79 8.73 3.82 8.01
8.35 10.49 6.68 5.64 4.08 6.17 9.91 5.47

a. Calcule la media, la mediana, primero y tercer cuartiles.

b. Calcule la varianza, la desviación estándar, el rango, el rango intercuartil y el coeficiente de variación. ¿Existe algún valor atípico? Explique su respuesta.

c. ¿Los datos son asimétricos? De ser así, ¿cómo?

d. Un cliente llega a la sucursal durante la hora del almuerzo y pregunta al gerente cuánto tendrá que esperar, éste le responde: "Menos de cinco minutos, con toda seguridad". Con base en sus resultados de los incisos *a*) y *b*), evalúe la exactitud de tal afirmación.

AUTO Examen **3.17** China tiene el mercado con crecimiento más rápido en ventas de automóviles de pasajeros y es el cuarto mercado más grande, detrás de Estados Unidos, Japón y Alemania. Las ventas aumentaron un 61% en 2002 y un 55% en 2003 (Peter Wonacott, "A Fear Amid China's Car Boom", *The Wall Street Journal,* 2 de febrero, 2004, A17). Calcule la media geométrica de la tasa de incremento. (*Sugerencia:* Denote el crecimiento del 61% como $R_1 = 0.61$.)

3.18 Durante el periodo transcurrido desde 2000 hasta 2003, se observó una gran volatilidad en el valor de las acciones. Los datos que se presentan en la siguiente tabla **STOCKRETURN** representan las tasas de rendimiento total del índice industrial Dow Jones, del índice Standard & Poor's 500, del índice Russell 2000, y del índice Wilshire 5000 de 2000 a 2003.

Año	DJIA	SP500	Russell2000	Wilshire5000
2003	25.30	26.40	45.40	29.40
2002	−15.01	−22.10	−21.58	−20.90
2001	−5.44	−11.90	−1.03	−10.97
2000	−6.20	−9.10	−3.02	−10.89

Fuente: The Wall Street Journal, *2 de enero, 2004.*

a. Calcule la tasa de rendimiento geométrica de los índices Dow Jones, Standard & Poor's 500, Russell 2000 y Wilshire 5000.

b. ¿Qué conclusiones se obtienen en relación con las tasas de rendimiento geométricas de los cuatro índices bursátiles?

c. Compare los resultados del inciso *b*) con los de los problemas 3.19*b*) y 3.20*b*).

3.19 Durante el periodo de 2000 a 2003, se observó una gran volatilidad en el valor de las inversiones. Los datos que se presentan en la siguiente tabla **BANKRETURN** representan la tasa de rendimiento total de un certificado de depósito a un año, de un certificado de depósito a 30 meses y de un depósito en el mercado de dinero de 2000 a 2003.

Año	A 1 año	A 30 meses	Mercado de dinero
2003	1.20	1.76	0.61
2002	1.98	2.74	1.02
2001	3.60	3.97	1.73
2000	5.46	5.64	2.09

Fuente: The Wall Street Journal**,** *2 de enero, 2004.*

a. Calcule la tasa de rendimiento geométrica de los certificados de depósito a un año, 30 meses y en el mercado de dinero.

b. ¿Qué conclusiones se obtienen en relación con las tasas de rendimiento geométricas de los tres depósitos?

c. Compare los resultados del inciso *b*) con los de los problemas 3.18*b*) y 3.20*b*).

3.20 Durante el periodo de 2000 a 2003, se observó una gran volatilidad en el valor de los metales. Los datos que se presentan en la siguiente tabla **METALRETURN** representan la tasa de rendimiento total de platino, oro y plata de 2000 a 2003.

Año	Platino	Oro	Plata
2003	34.2	19.5	24.0
2002	24.5	24.5	5.5
2001	−21.3	1.2	−3.0
2000	−23.3	1.8	−5.9

Fuente: The Wall Street Journal, *2 de enero, 2004.*

a. Calcule la tasa de rendimiento geométrica de platino, oro y plata.

b. ¿Qué conclusiones se obtienen en relación con las tasas de rendimiento geométricas de los tres metales?

c. Compare los resultados del inciso *b*) con los de los problemas 3.18*b*) y 3.19*b*).

3.2 MEDIDAS NUMÉRICAS DESCRIPTIVAS DE UNA POBLACIÓN

En la sección 3.1 se expusieron varios *estadísticos* que describen las propiedades de la tendencia central, la variación y la forma de una *muestra*. Si su conjunto de datos representa medidas numéricas de toda una *población*, necesita calcular e interpretar los *parámetros*, medidas sintetizadas para una población. En esta sección, aprenderá sobre tres parámetros descriptivos de la población, la media poblacional, la varianza poblacional y la desviación estándar poblacional.

Como ayuda para ilustrar estos parámetros, vea primero la tabla 3.5, que contiene los cinco mayores bonos de capital (en términos de activos totales) para el primero de marzo de 2004. También se indica el rendimiento a 52 semanas de cada uno de ellos. **LARGEST BONDS**

TABLA 3.5
Rendimiento en 2003 de la población compuesta por los cinco mayores bonos de capital.

Fondo de capital	Rendimiento a 52 semanas (en porcentaje)
Vanguard GNMA	3.8
Vanguard Total Bond Index	6.5
Pimco Total Return Admin	7.0
Pimco Total Return Instl	7.3
America Bond Fund	12.9

Fuente: The Wall Street Journal, *25 de marzo, 2004, C2.*

La media poblacional

La **media poblacional** se representa por medio del símbolo μ, la letra griega *mu* minúscula. La ecuación (3.13) define a la media poblacional.

MEDIA POBLACIONAL

La media poblacional es la suma de los valores de la población dividida por el tamaño de la población N.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \qquad \textbf{(3.13)}$$

donde

μ = media poblacional

X_i = i-ésimo valor de la variable X

$\sum_{i=1}^{N} X_i$ = sumatoria de todos los valores X_i de la población

Para calcular el rendimiento medio de la población de bonos de capital listados en la tabla 3.5, se utiliza la ecuación (3.13),

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{3.8 + 6.5 + 7.0 + 7.3 + 12.9}{5} = \frac{37.5}{5} = 7.5$$

De esta manera, el rendimiento medio en 2003 de tales bonos de capital es del 7.5%.

Varianza y desviación estándar poblacionales

La **varianza poblacional** y la **desviación estándar poblacional** miden la variación en una población. Al igual que los estadísticos muestrales relacionados, la desviación estándar poblacional es igual a la raíz cuadrada de la varianza poblacional. El símbolo σ^2, que es la letra griega *sigma* minúscula elevada al cuadrado, representa la varianza poblacional y el símbolo σ, la misma letra griega minúscula pero sin elevar al cuadrado, representa la desviación estándar poblacional. Las ecuaciones (3.14) y (3.15) definen esos parámetros. Los denominadores de los términos de la derecha de estas ecuaciones utilizan N y no el término $(n - 1)$ que se emplea para la varianza y la desviación estándar de las muestras [vea las ecuaciones (3.9) y (3.10) de la página 82].

VARIANZA POBLACIONAL

La varianza poblacional es la suma de las diferencias con respecto a la media de la población elevada al cuadrado y dividida por el tamaño de la población N.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad \textbf{(3.14)}$$

donde μ = media poblacional

X_i = i-ésimo valor de la variable X

$\sum_{i=1}^{N}(X_i - \mu)^2$ = sumatoria de todas las diferencias entre los valores X_i y μ, elevadas al cuadrado

DESVIACIÓN ESTÁNDAR POBLACIONAL

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} \qquad \textbf{(3.15)}$$

Para calcular la varianza poblacional correspondiente a los datos de la tabla 3.5 de la página 94, se utiliza la ecuación (3.14),

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \\ &= \frac{(3.8 - 7.5)^2 + (6.5 - 7.5)^2 + (7.0 - 7.5)^2 + (7.3 - 7.5)^2 + (12.9 - 7.5)^2}{5} \\ &= \frac{13.69 + 1.00 + 0.25 + 0.04 + 29.16}{5} \\ &= \frac{44.14}{5} = 8.828\end{aligned}$$

De esta forma, la varianza de los rendimientos es de 8.828 unidades porcentuales de rendimiento al cuadrado. Las unidades cuadradas hacen que la varianza sea difícil de interpretar. Debe utilizarse la desviación estándar, que emplea las unidades originales de los datos (rendimiento porcentual). A partir de la ecuación (3.15),

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} = \sqrt{8.828} = 2.97$$

Por lo tanto, el rendimiento típico en 2003 difiere de la media de 7.5 en aproximadamente 2.97. Esta enorme variación sugiere que los grandes bonos de capital tienen resultados muy distintos.

La regla empírica

En la mayoría de los conjuntos de datos, una gran parte de los valores tienden a agruparse en algún lugar cercano a la mediana. En los conjuntos de datos asimétricos a la derecha, el agrupamiento se presenta a la izquierda de la media, es decir en un valor menor que la media. En los conjuntos de datos asimétricos a la izquierda, el agrupamiento se presenta a la derecha de la media, es decir en un valor mayor que la media. En los conjuntos de datos simétricos, donde la mediana y la media son iguales, con frecuencia los valores tienden a agruparse alrededor de la media y la mediana, generando una distribución con forma de campana. En las distribuciones de esta clase, utilizar la **regla empírica** permite examinar la variabilidad:

- Aproximadamente el 68% de los valores se encuentran a una distancia de ±1 desviación estándar de la media.
- Aproximadamente el 95% de los valores se encuentran a una distancia de ±2 desviaciones estándar de la media.
- Aproximadamente el 99.7% se encuentran a una distancia de ±3 desviaciones estándar de la media.

La regla empírica ayuda a medir cómo se distribuyen los valores por encima y debajo de la media. Esto permite identificar los valores atípicos cuando se analiza un conjunto de datos numéricos. La regla empírica implica que, en las distribuciones con forma de campana, aproximadamente sólo uno de cada 20 valores estará alejado de la media más allá de dos desviaciones estándar en cualquier dirección. Por regla general, los valores que no se encuentran en el intervalo $\mu \pm 2\sigma$ se consideran como posibles atípicos. Esta regla también implica que sólo alrededor de tres de cada 1,000 estarán alejados de la media más allá de tres desviaciones estándar. Por lo tanto, casi siempre se consideran como extremos los valores que no se encuentran en el intervalo $\mu \pm 3\sigma$. En los conjuntos de datos con mucha asimetría, o en los que por alguna otra razón no tienen forma de campana, en lugar de la regla empírica se debe aplicar la regla de Chebyshev, que se explica en la página 97.

EJEMPLO 3.12 USO DE LA REGLA EMPÍRICA

La cantidad media de llenado de una población integrada por 12 latas de gaseosa es de 12.06 onzas, con una desviación estándar de 0.02. También se sabe que esta población tiene forma de campana. Describa la distribución de la cantidad de llenado de las latas. ¿Existe una gran probabilidad de que una lata tenga menos de 12 onzas de gaseosa?

SOLUCIÓN

$$\mu \pm \sigma = 12.06 \pm 0.02 = (12.04, 12.08)$$
$$\mu \pm 2\sigma = 12.06 \pm 2(0.02) = (12.02, 12.10)$$
$$\mu \pm 3\sigma = 12.06 \pm 3(0.02) = (12.00, 12.12)$$

Utilizando la regla empírica, aproximadamente el 68% de las latas tendrán entre 12.04 y 12.08 onzas, aproximadamente el 95% tendrá entre 12.02 y 12.10 onzas, y aproximadamente el 99.7% tendrá entre 12.00 y 12.12 onzas. Así que es muy poco probable que una lata tenga menos de 12 onzas.

La regla de Chebyshev

La **regla de Chebyshev** (referencia 1) establece que para todo conjunto de datos, independientemente de su forma, el porcentaje de valores que se encuentran a una distancia de k desviaciones estándar o menos de la media, debe ser por lo menos igual a

$$(1 - 1/k^2) \times 100\%$$

Puede usar esta regla para todo valor de k mayor que 1. Considere una $k = 2$. La regla de Chebyshev establece que al menos $[1 - (1/2)^2] \times 100\% = 75\%$ de los valores deben estar dentro de ±2 desviaciones estándar de la media.

La regla de Chebyshev es muy general y se aplica a cualquier tipo de distribución. La regla señala *por lo menos* el porcentaje de valores que quedan dentro de una distancia dada de la media. Sin embargo, si el conjunto de datos tiene una forma que se aproxima a la de campana, la regla empírica reflejará con mayor precisión la mayor concentración de datos cerca de la media. En la tabla 3.6 se comparan la regla empírica y la de Chebyshev.

TABLA 3.6
Variación de los datos con respecto a la media.

	Porcentaje de valores encontrados en intervalos alrededor de la media	
Intervalo	**Chebyshev (para toda distribución)**	**Regla empírica (distribución con forma de campana)**
$(\mu - \sigma, \mu + \sigma)$	Al menos 0%	Aproximadamente 68%
$(\mu - 2\sigma, \mu + 2\sigma)$	Al menos 75%	Aproximadamente 95%
$(\mu - 3\sigma, \mu + 3\sigma)$	Al menos 88.89%	Aproximadamente 99.7%

EJEMPLO 3.13 USO DE LA REGLA DE CHEBYSHEV

Como en el ejemplo 3.12, la media de la cantidad de llenado de una población integrada por 12 latas de gaseosa es de 12.06 onzas y una desviación estándar de 0.02. Sin embargo, no se conoce la forma de la población y no es posible suponer que tiene forma de campana. Describa la distribución de la cantidad de llenado de las latas. ¿Existe una gran probabilidad de que una lata tenga menos de 12 onzas de gaseosa?

SOLUCIÓN

$$\mu \pm \sigma = 12.06 \pm 0.02 = (12.04, 12.08)$$
$$\mu \pm 2\sigma = 12.06 \pm 2(0.02) = (12.02, 12.10)$$
$$\mu \pm 3\sigma = 12.06 \pm 3(0.02) = (12.00, 12.12)$$

Como la distribución posiblemente sea asimétrica, no es pertinente utilizar la regla empírica. Usando la regla de Chebyshev no se puede decir algo sobre el porcentaje de latas que tienen entre 12.04 y 12.08 onzas. Es posible determinar que al menos el 75% de las latas tendrán entre 12.02 y 12.10 onzas, y que por lo menos el 88.89% tendrán entre 12.00 y 12.12 onzas. Por lo tanto, entre 0 y 11.11% de las latas tienen menos de 12 onzas.

Cuando se tienen datos muestrales, estas dos reglas permiten entender cómo se distribuyen los datos alrededor de la media. En todo caso, use el valor de $\bar{X}$ que calculó, en lugar de μ y el que calculó para S en lugar de σ. Los resultados calculados empleando los estadísticos muestrales son aproximaciones, ya que utilizó estadísticos muestrales ($\bar{X}$, S) y no parámetros poblacionales (μ, σ).

PROBLEMAS PARA LA SECCIÓN 3.2

Aprendizaje básico

ASISTENCIA de PH Grade **3.21** A continuación se presenta un conjunto de datos para una población con $N = 10$:

7 5 11 8 3 6 2 1 9 8

a. Calcule la media poblacional.
b. Calcule la desviación estándar poblacional.

ASISTENCIA de PH Grade **3.22** A continuación se presenta un conjunto de datos para una población con $N = 10$:

7 5 6 6 6 4 8 6 9 3

a. Calcule la media poblacional.
b. Calcule la desviación estándar poblacional.

Aplicación de conceptos

AUTO Examen **3.23** Los siguientes datos representan las declaraciones trimestrales de impuestos por ventas (en miles de dólares), correspondientes al periodo que finalizó en marzo de 2004, enviados al contralor del poblado Fair Lake por los 50 negocios establecidos en dicha localidad: **TAX**

10.3	11.1	9.6	9.0	14.5
13.0	6.7	11.0	8.4	10.3
13.0	11.2	7.3	5.3	12.5
8.0	11.8	8.7	10.6	9.5
11.1	10.2	11.1	9.9	9.8
11.6	15.1	12.5	6.5	7.5
10.0	12.9	9.2	10.0	12.8
12.5	9.3	10.4	12.7	10.5
9.3	11.5	10.7	11.6	7.8
10.5	7.6	10.1	8.9	8.6

a. Calcule la media, la varianza y la desviación estándar de esta población.
b. ¿Qué proporción de estos negocios tienen declaraciones trimestrales de impuestos sobre ventas dentro de ±1, ±2 o ±3 desviaciones estándar de la media?
c. Compare y encuentre las diferencias entre sus hallazgos con lo que cabría esperar de acuerdo con la regla empírica. ¿Le sorprenden los resultados obtenidos en *b*)?

ASISTENCIA de PH Grade **3.24** Considere una población de 1,024 fondos de inversión que invierten principalmente en empresas grandes. Usted determinó que μ, la media del porcentaje total anual de rendimientos obtenidos por todos los fondos es 8.20 y que σ, la desviación estándar, es 2.75. Suponga además que determinó que el rango del porcentaje total anual va de −2.0 a 17.1 y que los cuartiles son 5.5 (Q_1) y 10.5(Q_3), respectivamente. De acuerdo con la regla empírica, ¿qué porcentaje de estos fondos se espera que estén
a. dentro de ±1 desviaciones estándar de la media?
b. dentro de ±2 desviaciones estándar de la media?
c. De acuerdo con la regla de Chebyshev, ¿qué porcentaje de estos fondos se espera que estén dentro de ±1, ±2 o ±3 desviaciones estándar de la media?
d. De acuerdo con la regla de Chebyshev, se espera que al menos el 93.75% de estos fondos tengan rendimientos totales anuales entre ¿cuáles dos cantidades?

3.25 En la siguiente tabla **ASSETS** se representan los activos de cinco grandes fondos de capital, en miles de millones de dólares.

Fondo de capital	Activos (miles de millones de dólares)
Vanguard GNMA	19.5
Vanguard Total Bond Mkt. Index	16.8
Bond Fund of America A	13.7
Franklin Calif. Tax-Free Inc. A	12.8
Vanguard Short-Term Corp.	10.9

a. Calcule la media de esta población constituida por los cinco bonos de capital más grandes. Interprete este parámetro.
b. Calcule la varianza y la desviación estándar de esta población. Interprete estos parámetros.
c. ¿Existe mucha variabilidad en los activos de los fondos de capital?

3.26 Los datos del archivo **ENERGY** contienen el consumo de energía per cápita en kilowatts-hora de cada uno de los 50 estados y el distrito de Columbia, que constituyen a Estados Unidos, durante 1999.
a. Calcule la media, la varianza y desviación estándar de la población.
b. ¿Qué proporción de estos estados tienen un consumo de energía promedio per cápita dentro de ±1 desviación estándar de la media, dentro de ±2 desviaciones estándar de la media, y dentro de ±3 desviaciones estándar de la media?
c. Compare y encuentre las diferencias entre sus hallazgos contra lo que cabría esperar de acuerdo con la regla empírica. ¿Le sorprenden los resultados obtenidos en *b*)?
d. Eliminando los datos correspondientes al distrito de Columbia en los incisos *a*) a *c*), ¿cómo cambian los resultados?

3.27 Los datos en el archivo **DOWRETURN** muestran el rendimiento anualizado de 10 años (1994-2003) correspondiente a 30 empresas incluidas en el Dow Jones Industrials.
a. Calcule la media de esta población. Interprete este número.
b. Calcule la varianza y la desviación estándar de esta población. Interprete la desviación estándar.
c. Utilice la regla empírica o la de Chebyshev, la que resulte apropiada, para explicar aún más la variación de este conjunto de datos.
d. Utilizando los resultados de *c*), ¿existen algunos datos atípicos? Explique su respuesta.

3.3 ANÁLISIS EXPLORATORIO DE DATOS

En la sección 3.1 se analizaron estadísticos muestrales para datos numéricos como son las medidas de tendencia central, variación y forma. Otra manera de describir datos numéricos es mediante el análisis exploratorio de datos, que incluye el resumen de cinco números y la gráfica de caja y bigote (referencias 5 y 6).

Resumen de cinco números

Un **resumen de cinco números** compuesto por:

$$X_{\text{menor}} \quad Q_1 \quad \text{Mediana} \quad Q_3 \quad X_{\text{mayor}}$$

permite determinar la forma de la distribución. En la tabla 3.7 se explica cómo las relaciones entre los "cinco números" le permiten reconocer la forma del conjunto de datos.

TABLA 3.7
Relaciones entre el resumen de cinco números y el tipo de distribución

	Tipo de distribución		
Comparación	**Asimétrico a la izquierda**	**Simétrico**	**Asimétrico a la derecha**
La distancia de X_{menor} a la mediana contra la distancia de la mediana a X_{mayor}.	La distancia de X_{menor} a la mediana es mayor que la distancia de la mediana a X_{mayor}.	Ambas distancias son iguales.	La distancia de X_{menor} a la mediana es menor que la distancia de la mediana a X_{mayor}.
La distancia de X_{menor} a Q_1 contra la distancia de Q_3 a X_{mayor}.	La distancia de X_{menor} a Q_1 es mayor que la distancia de Q_3 a X_{mayor}.	Ambas distancias son iguales.	La distancia de X_{menor} a Q_1 es menor que la distancia de Q_3 a X_{mayor}.
La distancia de Q_1 a la mediana contra la distancia de la mediana a Q_3	La distancia de Q_1 a la mediana es mayor que la distancia de la mediana a Q_3	Ambas distancias son iguales.	La distancia de Q_1 a la mediana es menor que la distancia de la mediana a Q_3.

Para la muestra de 10 tiempos necesarios para arreglarse, el menor valor es 29 minutos y el mayor es 52 minutos (vea las páginas 75 y 77). Los cálculos ya realizados en la sección 3.1 indican que la mediana = 39.5, el primer cuartil = 35, y el tercer cuartil = 44. Por lo tanto, el resumen de cinco puntos es:

$$29 \quad 35 \quad 39.5 \quad 44 \quad 52$$

La distancia de X_{menor} a la mediana (39.5 − 29 = 10.5) es ligeramente menor que la distancia de la mediana a X_{mayor} (52 − 39.5 = 12.5). La distancia de X_{menor} a Q_1 (35 − 29 = 6) es ligeramente menor que la distancia de Q_3 a X_{mayor} (52 − 44 = 8). De esta forma, los tiempos para arreglarse son ligeramente asimétricos a la derecha.

EJEMPLO 3.14

CÁLCULO DEL RESUMEN DE LOS CINCO NÚMEROS DEL PORCENTAJE DE RENDIMIENTO EN 2003 DE LOS FONDOS DE INVERSIÓN DE ALTO RIESGO PARA PEQUEÑOS CAPITALES

Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72), se clasifican de acuerdo con el nivel de riesgo (bajo, medio y alto) y el tamaño del capital invertido (pequeño, mediano y gran capital). Calcule el resumen de cinco puntos del rendimiento en 2003 de los nueve fondos de inversión de alto riesgo para pequeños capitales. **MUTUALFUNDS2004**

SOLUCIÓN

De los cálculos previos realizados a los rendimientos en 2003 de los fondos de alto riesgo para pequeños capitales (vea las páginas 76 y 78), la mediana = 53.8, el primer cuartil = 41.7, y el tercer cuartil = 60.85. Además, el menor valor del conjunto de datos es 37.3 y el mayor es 66.5. Por lo tanto, el resumen de cinco puntos es:

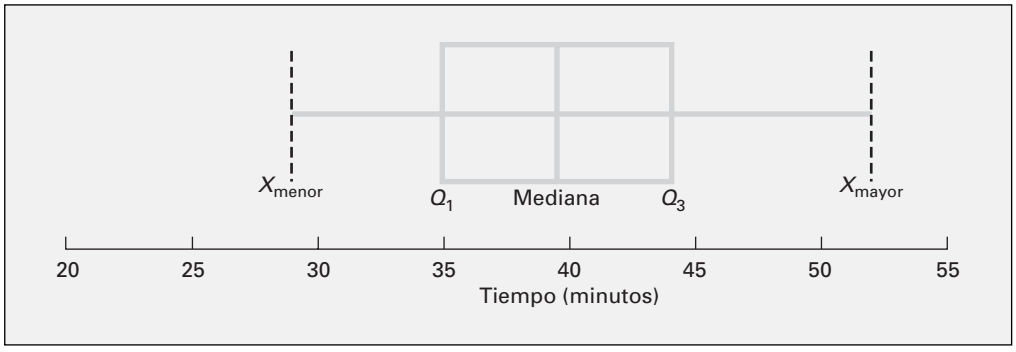
37.3 41.7 53.8 60.85 66.5

La distancia de X_{menor} a la mediana (53.8 − 37.3 = 16.5) es mayor que la distancia de la mediana a X_{mayor} (66.5 − 53.8 = 12.7). Esto indica asimetría a la izquierda. La distancia de X_{menor} a Q_1 (41.7 − 37.3 = 4.4) es ligeramente menor que la distancia de Q_3 a X_{mayor} (66.5 − 60.85 = 5.65). Esto indica una ligera asimetría a la derecha. Por lo tanto, los resultados son incongruentes.

Gráfica de caja y bigote

La **gráfica de caja y bigote** ofrece una representación visual de los datos basada en el resumen de cinco números. En la figura 3.4 se ilustra la gráfica de caja y bigote de los tiempos necesarios para arreglarse.

FIGURA 3.4
Gráfica de caja y bigote del tiempo necesario para arreglarse.



La línea vertical dibujada dentro de la caja representa a la mediana. La línea vertical a la izquierda de la caja representa la ubicación de Q_1 y la línea vertical a la derecha de la caja representa la ubicación de Q_3. De esta forma, la caja contiene al 50% de los valores de la distribución. El 25% inferior de los datos se representa mediante una línea (es decir, un *bigote*) que une el lado izquierdo de la caja con la ubicación del menor valor, X_{menor}. De la misma manera, el 25% superior de los datos se representa mediante un bigote que une el lado derecho de la caja con la ubicación del valor mayor, X_{mayor}.

La gráfica de caja y bigote de los tiempos necesarios para arreglarse que aparece en la figura 3.4 muestra una muy ligera asimetría a la derecha, ya que la distancia entre la mediana y el valor mayor es levemente mayor que la distancia entre el menor valor y la mediana. El bigote derecho es un poco más largo que el izquierdo.

EJEMPLO 3.15 GRÁFICA DE CAJA Y BIGOTE DEL RENDIMIENTO PORCENTUAL EN 2003 DE LOS FONDOS DE INVERSIÓN DE RIESGO BAJO, PROMEDIO Y ALTO

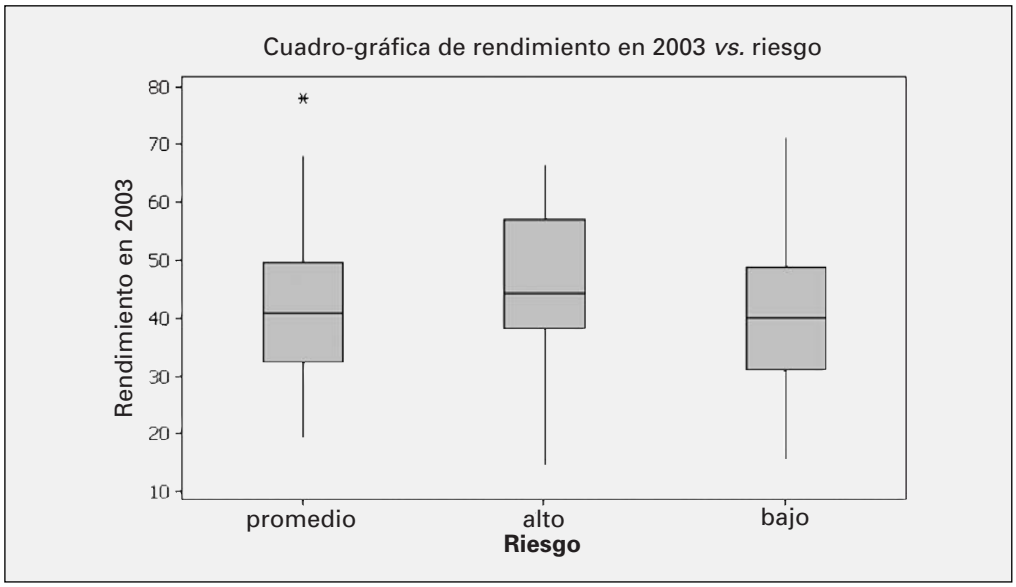
Los 121 fondos de inversión que forman parte del escenario "Uso de la estadística" (vea la página 72) se clasifican de acuerdo con su nivel de riesgo (bajo, medio y alto) y tamaño del capital invertido (pequeño, mediano y gran capital). Construya la gráfica de caja y bigote para los rendimientos en 2003 para los fondos de inversión de riesgo bajo, promedio y alto. **MUTUALFUNDS2004**

SOLUCIÓN

En la figura 3.5 aparece la gráfica de caja y bigote de los rendimientos en 2003 para los fondos de inversión de riesgo bajo, promedio, y alto, elaborada en Minitab. Este programa muestra la gráfica de manera vertical, de inferior (bajo) a superior (alto). El asterisco (*) de los fondos de riesgo promedio representa la presencia de valores atípicos.[2] La media del porcentaje de rendimiento y los cuartiles de los fondos de alto riesgo son mayores que los correspondientes a los fondos de riesgo bajo o promedio. Los fondos de riesgo promedio son asimétricos a la derecha, a causa del rendimiento extremadamente alto de uno de ellos (78). Los fondos de alto riesgo aparecen asimétricos a la izquierda por el largo bigote inferior, pero la mediana del rendimiento está más cerca del primer cuartil que del tercero. Los fondos de bajo riesgo aparecen ligeramente asimétricos a la derecha porque el bigote superior es más largo que el interior.

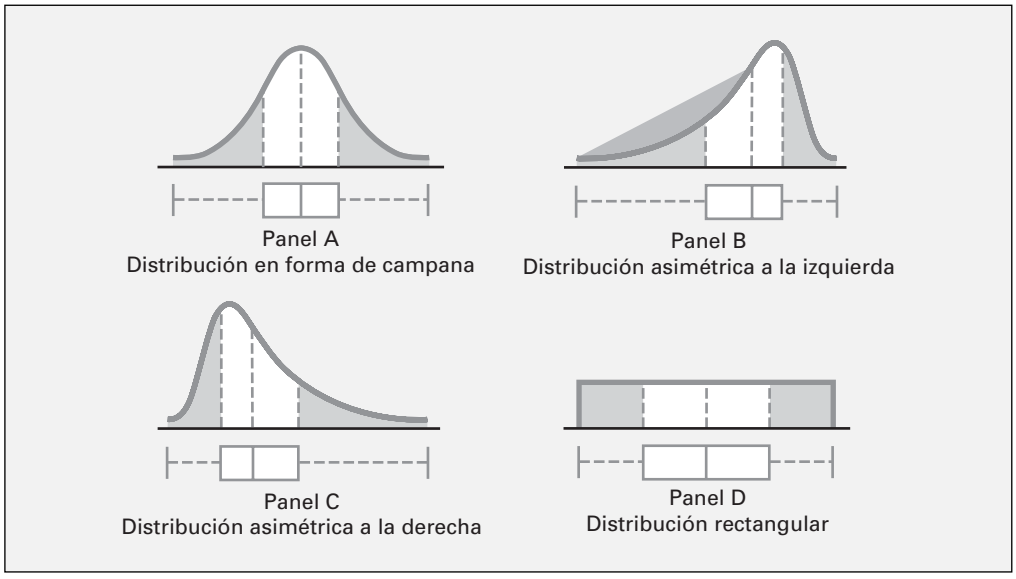
[2]Si existen valores atípicos, los bigotes de la gráfica de caja y bigote de Minitab se extienden 1.5 veces el rango intercuartil más allá de los cuartiles o hasta el valor más alto.

FIGURA 3.5
Gráfica de caja y bigote de los rendimientos en 2003, en Minitab, para los fondos de inversión de riesgo bajo, promedio y alto.



En la figura 3.6 se comprueba la relación que existe entre la gráfica de caja y bigote y el polígono de cuatro tipos distintos de distribución. (*Nota:* El área bajo cada polígono se divide en cuartiles que corresponden al resumen de cinco números de la gráfica de caja y bigote.)

FIGURA 3.6
Gráficas de caja y bigote, y sus polígonos correspondientes, de cuatro distribuciones.



Los paneles A y D de la figura 3.6 son simétricos. En estas distribuciones, la media y la mediana son iguales. Además, la longitud del bigote izquierdo es igual a la del derecho, y la línea que representa a la mediana divide la caja por la mitad.

El panel B de la figura 3.6 es asimétrico a la izquierda. Los pocos valores pequeños inclinan a la media hacia la punta izquierda. Para esta distribución asimétrica a la izquierda, la asimetría indica que existe un marcado agrupamiento de los valores en el extremo superior de la escala (es decir, el lado derecho); el 75% de todos los valores se encuentran entre el extremo izquierdo de la caja (Q_1) y el extremo del bigote derecho (X_{mayor}). Por lo tanto, el largo bigote izquierdo contiene al 25% más pequeño de los valores, lo que demuestra distorsión de la simetría de este conjunto de datos.

El panel C de la figura 3.6 es asimétrico a la derecha. La concentración de valores está en el extremo inferior de la escala (es decir, en el lado izquierdo de la gráfica de caja y bigote). Aquí, el 75% de todos los valores se encuentran entre el principio del bigote izquierdo (X_{menor}) y el extremo derecho de la caja (Q_3), y el 25% restante de los valores se encuentran dispersos a lo largo del bigote derecho, en el extremo superior de la escala.

PROBLEMAS PARA LA SECCIÓN 3.3

Aprendizaje básico

ASISTENCIA de PH Grade **3.28** A continuación se presenta un conjunto de datos para una muestra con $n = 6$:

7 4 9 7 3 12

a. Elabore el resumen de cinco números.
b. Construya su gráfica de caja y bigote, y describa la forma.
c. Compare su respuesta del inciso *b*) con la del problema 3.2*d*) de la página 90. Analícela.

ASISTENCIA de PH Grade **3.29** A continuación se presenta un conjunto de datos para una muestra con $n = 7$:

12 7 4 9 0 7 3

a. Elabore el resumen de cinco números.
b. Realice su gráfica de caja y bigote, y describa la forma.
c. Compare su respuesta del inciso *b*) con la del problema 3.3*c*) de la página 90. Analícela.

3.30 A continuación se presenta un conjunto de datos para una muestra con $n = 5$:

7 –5 –8 7 9

a. Elabore el resumen de cinco números.
b. Construya su gráfica de caja y bigote, y describa la forma.
c. Compare su respuesta del inciso *b*) con la del problema 3.4*c*) de la página 90. Analícela.

Aplicación de conceptos

Puede resolver los problemas 3.31 a 3.36 manualmente o en Excel, Minitab o SPSS.

ASISTENCIA de PH Grade AUTO Examen **3.31** Un fabricante de baterías para flash fotográfico tomó una muestra de 13 baterías de la producción diaria y las utilizó de manera continua hasta agotarlas. El número de horas que funcionaron está en el archivo. **BATTERIES**

342 426 317 545 264 451
1,049 631 512 266 492 562 298

a. Elabore el resumen de cinco números.
b. Construya su gráfica de caja y bigote, y describa la forma.

3.32 Durante el ciclo escolar 2002-2003, muchas universidades estadounidenses elevaron sus cuotas y tarifas de manutención, como consecuencia de la reducción de los subsidios estatales (Mary Beth Marklein, "Public Universities Raise Tuition, Fees -and Ire", *USA Today*, 8 de agosto, 2002, 1A-2A). A continuación se representa el cambio del costo de inscripción, un dormitorio compartido y el plan de alimentación más solicitado entre los ciclos escolares 2001-2002 y 2002-2003, para una muestra de 10 universidades públicas. **COLLEGECOST**

Universidad	Cambio en el costo ($)
University of California, Berkeley	1,589
University of Georgia, Athens	593
University of Illinois, Urbana-Champaign	1,223
Kansas State University, Manhattan	869
University of Maine, Orono	423
University of Mississippi, Oxford	1,720
University of New Hampshire, Durham	708
Ohio State University, Columbus	1,425
University of South Carolina, Columbia	922
Utah State University, Logan	308

a. Elabore el resumen de cinco números.
b. Construya su gráfica de caja y bigote, y describa la forma.

3.33 Una empresa dedicada a la consultoría y al desarrollo de software, ubicada en el área metropolitana de Phoenix, desarrolla software para sistemas administrativos de cadenas de suministro y se vale de la reutilización sistemática de software. En lugar de comenzar desde cero para elaborar y desarrollar nuevos sistemas personalizados de software, utiliza una base de datos que contiene componentes reutilizables que suman más de 2,000,000 de líneas de código, recopilados a lo largo de 10 años de actividades continuas. Se pide a ocho analistas de la empresa que calculen la tasa de reutilización cuando se desarrolla un nuevo sistema de software. Los siguientes datos corresponden al porcentaje total de código que procede de la base de datos de reutilización y forma parte del sistema de software. **REUSE**

50.0 62.5 37.5 75.0 45.0 47.5 15.0 25.0

Fuente: M. A. Rothenberger y K. J. Dooley, "A Performance Measure for Software Reuse Projects", Decision Sciences, *30 (Otoño de 1999), 1131-1153.*

a. Elabore el resumen de cinco números.
b. Realice su gráfica de caja y bigote, y describa la forma de los datos.

3.34 Los siguientes datos representan la tarifa (en dólares) por cheque devuelto de una muestra de 23 bancos, para los clientes de depósito directo que conservan un saldo de $100 y la cuota (en dólares) mensual por manejo de cuenta, si sus cuentas no conservan el saldo mínimo requerido de $1,500, de una muestra de 26 bancos. **BANKCOST1 BANKCOST2**

Tarifa por cheque devuelto

26 28 20 20 21 22 25 25 18 25 15 20 18 20 25 25 22 30 30 30 15 20 29

Cuota mensual por manejo de cuenta

12 8 5 5 6 6 10 10 9 7 10 7 7 5 0 10 6 9 12 0 5 10 8 5 5 9

Fuente: "The New Face of Banking", Copyright © 2000 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, *junio de 2000.*

a. Elabore el resumen de cinco números de la tarifa por cheque devuelto y de la cuota mensual por manejo de cuenta.
b. Realice la gráfica de caja y bigote de la tarifa por cheque devuelto y de la cuota mensual por manejo de cuenta.
c. ¿Qué similitudes y diferencias existen en la distribución de la tarifa por cheque devuelto y de la cuota mensual por manejo de cuenta?

3.35 Los siguientes datos representan el total de grasas en hamburguesas y artículos de pollo tomados de una muestra de cadenas de comida rápida. **FASTFOOD**

Hamburguesas

19 31 34 35 39 39 43

Pollo

7 9 15 16 16 18 22 25 27 33 39

Fuente: "Quick Bites", Copyright © 2001 por Consumers Union of US., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, *marzo de 2001, 46.*

a. Elabore el resumen de cinco puntos para las hamburguesas y para los productos de pollo.
b. Construya la gráfica de caja y bigote para las hamburguesas y los productos de pollo, y describa la forma de la distribución de cada una.
c. ¿Qué similitudes y diferencias existen en la distribución de hamburguesas y de productos de pollo?

3.36 Una sucursal bancaria ubicada en una zona comercial de la ciudad desarrolló un proceso mejorado para atender a sus clientes durante la hora del almuerzo a mediodía, hasta la 1:00 PM. Durante una semana se registra el tiempo de espera en minutos (definido de manera operacional como el tiempo transcurrido desde que el cliente se forma en la fila hasta que llega a la ventanilla del cajero) de todos los clientes en ese horario. Se selecciona una muestra aleatoria de 15 clientes, y los resultados son los siguientes: **BANK1**

4.21 5.55 3.02 5.13 4.77 2.34 3.54
3.20 4.50 6.10 0.38 5.12 6.46 6.19 3.79

Otra sucursal, ubicada en una zona residencial, también está preocupada por el horario del almuerzo de mediodía hasta la 1:00 PM. Durante una semana, se registra el tiempo de espera en minutos (definido como el tiempo transcurrido desde que el cliente se forma en la fila hasta que llega a la ventanilla del cajero) de todos los clientes en ese horario. Se selecciona una muestra aleatoria de 15 clientes, y los resultados son los siguientes: **BANK2**

9.66 5.90 8.02 5.79 8.73 3.82 8.01
8.35 10.49 6.68 5.64 4.08 6.17 9.91 5.47

a. Elabore el resumen de cinco números para tiempo de espera en ambas sucursales bancarias.
b. Construya la gráfica de caja y bigote, y describa la forma de la distribución de las dos sucursales.
c. ¿Qué similitudes y diferencias existen en la distribución de los tiempos de espera en ambas sucursales bancarias?

3.4 LA COVARIANZA Y EL COEFICIENTE DE CORRELACIÓN

En la sección 2.5, usted utilizó los diagramas de dispersión para examinar de forma visual la relación que existe entre dos variables numéricas. En esta sección, se analizan la covarianza y el coeficiente de correlación, que miden la fortaleza de la relación entre dos variables numéricas.

La covarianza

La **covarianza** mide la fortaleza de la relación lineal entre dos variables numéricas (X y Y). La ecuación 3.16 define la **covarianza de una muestra** y el ejemplo 3.16 ilustra su uso.

LA COVARIANZA MUESTRAL

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \qquad \textbf{(3.16)}$$

EJEMPLO 3.16

CÁLCULO DE LA COVARIANZA DE UNA MUESTRA

Considere el coeficiente de gastos y los rendimientos en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. Calcule la covarianza de la muestra.

SOLUCIÓN

La tabla 3.8 presenta el coeficiente de gastos y los rendimientos de los fondos de inversión de alto riesgo para pequeños capitales, y en la figura 3.7 aparece una hoja de Excel que calcula la covarianza de esos datos. El área de cálculos de la figura 3.7 descompone la ecuación (3.16) en un conjunto de cálculos más pequeños. A partir de la celda C17, o directamente por la ecuación (3.16), se sabe que la covarianza es 1.19738.

$$\text{cov}(X,Y) = \frac{9.579}{9-1}$$
$$= 1.19738$$

TABLA 3.8

Coeficiente de gastos y rendimientos en 2003 de los fondos de inversión de alto riesgo para pequeños capitales.

Coeficiente de gastos	Rendimiento en 2003
1.25	37.3
0.72	39.2
1.57	44.2
1.40	44.5
1.33	53.8
1.61	56.6
1.68	59.3
1.42	62.4
1.20	66.5

FIGURA 3.7

Hoja de Excel que calcula la covarianza entre el coeficiente de gastos y los rendimientos en 2003 de los fondos de alto riesgo para pequeños capitales.

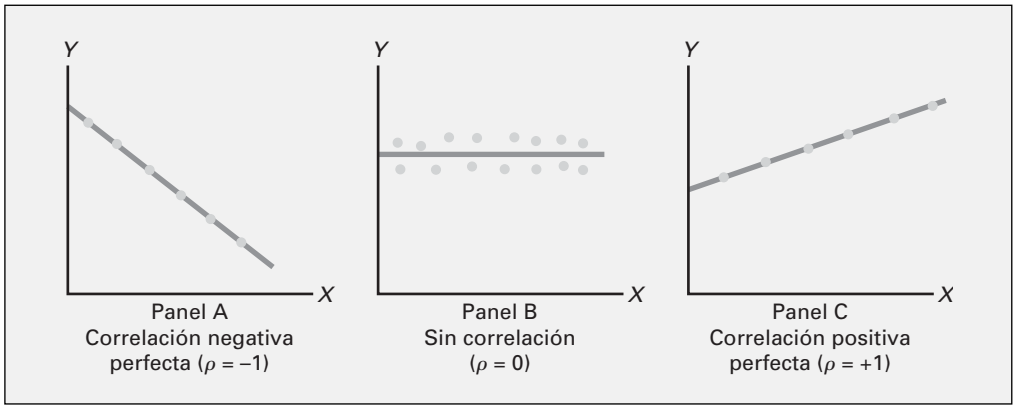
	A	B	C	
1	Expense ratio (X)	Return 2003 (Y)	(X-XBar)(Y-YBar)	
2	1.25	37.3	1.47078	=(A2 - C13) * (B2 - C14)
3	0.72	39.2	7.81111	=(A3 - C13) * (B3 - C14)
4	1.57	44.2	-1.58889	=(A4 - C13) * (B4 - C14)
5	1.4	44.5	-0.32822	=(A5 - C13)* (B5 - C14)
6	1.33	53.8	-0.05289	=(A6 - C13) * (B6 - C14)
7	1.61	56.6	1.30044	=(A7 - C13) * (B7 - C14)
8	1.68	59.3	2.53711	=(A8 - C13) * (B8 - C14)
9	1.42	62.4	0.72444	=(A9 - C13) * (B9 - C14)
10	1.2	66.5	-2.29489	=(A10 - C13) * (B10 - C14)
11				
12		Calculations		
13		XBar	1.353333333	=AVERAGE(A2:A10)
14		YBar	51.53333333	=AVERAGE(B2:B10)
15		*n*-1	8	=COUNT(A2:A10) - 1
16		Sum	9.57900	=SUM(C2:C10)
17		Covariance	1.19738	=C16/C15

La covarianza tiene un defecto importante como medida de la relación lineal entre dos variables numéricas. Como la covarianza puede tener cualquier valor, es imposible determinar la fortaleza relativa de la relación. Para ello, es necesario calcular el coeficiente de correlación.

Coeficiente de correlación

El **coeficiente de correlación** mide la fortaleza relativa de una relación lineal entre dos variables numéricas. Los valores del coeficiente de correlación varían desde −1 para una correlación negativa perfecta, hasta +1 para una correlación positiva perfecta. *Perfecta* quiere decir que si se trazaran los puntos en un diagrama de dispersión, todos ellos se podrían unir por medio de una línea recta. Al tratar con datos poblacionales para variables numéricas, se utiliza la letra griega ρ como símbolo del coeficiente de correlación. En la figura 3.8 se ilustran tres tipos diferentes de asociación entre dos variables.

FIGURA 3.8
Tipos de asociación entre variables.



En el panel A de la figura 3.8 hay una relación lineal negativa perfecta entre X y Y. De esta manera, el coeficiente de relación ρ es igual a −1, y al aumentar X, Y disminuye de una manera perfectamente predecible. El panel B ilustra una situación en la que no existe relación entre X y Y. En este caso, el coeficiente de correlación ρ es igual a 0, y al aumentar X no existe tendencia de Y a aumentar ni disminuir. El panel C ilustra una relación positiva perfecta en la que ρ es igual a +1. En este caso, Y aumenta de una manera perfectamente predecible cuando lo hace X.

Cuando se tienen datos muestrales, se calcula el coeficiente muestral de correlación r. Al utilizar los datos de una muestra, es difícil que se tenga un coeficiente muestral de exactamente +1 o −1. En la figura 3.9 de la página 106 se presentan diagramas de dispersión, con sus respectivos coeficientes muestrales de correlación r para seis conjuntos de datos, cada uno de los cuales contiene 100 valores de X y Y.

En el panel A, el coeficiente de correlación r es −0.9. Como se observa, donde los valores de X son más pequeños existe una fuerte tendencia a que los valores de Y sean grandes. De la misma forma, los valores pequeños de X tienden a hermanarse con valores pequeños en Y. No todos los datos quedan sobre una línea recta, por lo que la asociación entre X y Y no se describe como *perfecta*. Los datos del panel B tienen un coeficiente de correlación igual a −0.6, y los valores pequeños de X tienden a hermanarse con los valores grandes de Y. La relación lineal entre X y Y en el panel B no es tan fuerte como en el panel A. Así, el coeficiente de correlación en el panel B no es tan negativo como en el panel A. En el panel C, la relación lineal entre X y Y es muy débil, $r = -0.3$, y sólo existe una ligera tendencia de los valores pequeños de X a hermanarse con los más grandes de Y. En los paneles D a F se describen conjuntos de datos con coeficientes de correlación positivos, porque los valores pequeños de X tienden a hermanarse con los valores pequeños de Y, y los valores grandes de X tienden a asociarse con los valores grandes de Y.

En el análisis de la figura 3.9, las relaciones se describieron deliberadamente como *tendencias* y no como *causa-efecto*. Ese término se utilizó con un propósito. La sola correlación no prueba que

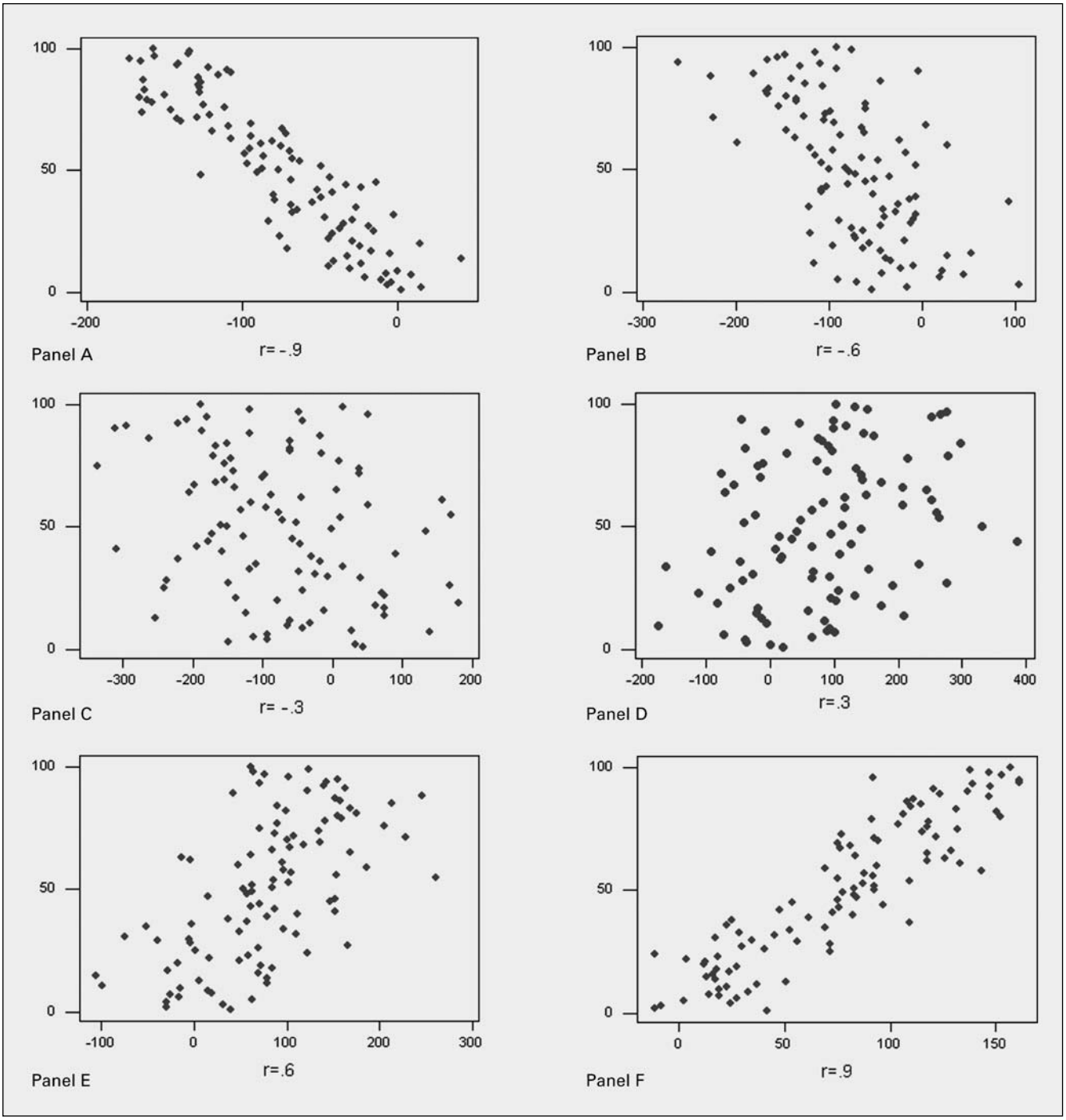


FIGURA 3.9 Seis diagramas de dispersión creados con Minitab y sus respectivos coeficientes de correlación *r*.

existe un efecto de causalidad, es decir, que el cambio en el valor de una variable *causó* el cambio en la otra variable. Una correlación fuerte puede producirse por simple coincidencia, por el efecto de una tercera variable que no se tomó en cuenta en el cálculo, o por una relación de causa-efecto. Sería necesario realizar un análisis adicional para determinar cuál de estas tres situaciones produce verdaderamente la correlación. Por tanto, se afirma que la causalidad implica correlación, pero la sola correlación no implica causalidad.

La ecuación (3.17) define el **coeficiente muestral de correlación** *r* y el ejemplo 3.17 ilustra su uso.

COEFICIENTE MUESTRAL DE CORRELACIÓN

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y} \qquad \textbf{(3.17)}$$

donde

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$$

El ejemplo 3.17 ilustra el cálculo del coeficiente muestral de correlación r mediante la ecuación (3.17).

EJEMPLO 3.17

CÁLCULO DEL COEFICIENTE MUESTRAL DE CORRELACIÓN

Considere el coeficiente de gastos y los rendimientos en 2003 de los fondos de inversión de alto riesgo para pequeños capitales. A partir de la figura 3.10 y de la ecuación (3.17), calcule el coeficiente muestral de correlación.

SOLUCIÓN

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

$$= \frac{1.19738}{(0.287663)(10.554383)}$$

$$= 0.3943786$$

FIGURA 3.10
Hoja de Excel que calcula el coeficiente de correlación entre los gastos y los rendimientos en 2003 de los fondos de alto riesgo para pequeños capitales.

	A	B	C	D	E
1	Expense ratio	Return 2003	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
2	1.25	37.3	0.0107	202.5878	1.4708
3	0.72	39.2	0.4011	152.1111	7.8111
4	1.57	44.2	0.0469	53.7778	-1.5889
5	1.4	44.5	0.0022	49.4678	-0.3282
6	1.33	53.8	0.0005	5.1378	-0.0529
7	1.61	56.6	0.0659	25.6711	1.3004
8	1.68	59.3	0.1067	60.3211	2.5371
9	1.42	62.4	0.0044	118.0844	0.7244
10	1.2	66.5	0.0235	224.0011	-2.2949
11		Sums:	0.662	891.16	9.5790
12					
13				Calculations	
14				XBar	1.353333333
15				YBar	51.53333333
16				*n*-1	8
17				Covariance	1.19738
18				S_X	0.287662997
19				S_Y	10.55438298
20				*r*	0.394378596

(formulas for range C2:E11 not shown)

=AVERAGE(A2:A10)
=AVERAGE(B2:B10)
=COUNT(A2:A10) - 1
=E11/E16
=SQRT(C11/E16)
=SQRT(D11/E16)
=CORREL(A2:A10,B2:B10)
or
=E17/(E18 * E19)

El coeficiente de gastos y los rendimientos en 2003 de los fondos de inversión de alto riesgo para pequeños capitales están correlacionados de forma positiva. Los fondos de inversión con menores coeficientes de gastos tienden a relacionarse con los menores rendimientos en 2003. Los fondos de inversión con mayores coeficientes de gastos tienden a relacionarse con los mayores rendimientos en 2003. Esta relación es muy débil, como lo indica el coeficiente de correlación, $r = 0.394$.

No es posible suponer que tener un bajo coeficiente de gastos provocó los bajos rendimientos en 2003. Sólo se puede decir que eso es lo que tiende a ocurrir en la muestra. Como con todas las inversiones, los resultados del pasado no avalan los del futuro.

En resumen, el coeficiente de correlación señala la relación, o asociación, lineal entre dos variables numéricas. Cuando el coeficiente de correlación se acerca a +1 o –1, es más fuerte la relación lineal entre las dos variables. Cuando el coeficiente de correlación se acerca a 0, existe poca o ninguna relación lineal. El signo del coeficiente de correlación señala si los datos se correlacionan de manera positiva (es decir, los valores más grandes de X se suelen hermanar con los valores más grandes de Y) o negativa (es decir, los valores más grandes de X se suelen hermanar con los valores más pequeños de Y). La existencia de una correlación fuerte no implica un efecto causal. Sólo señala las tendencias presentes en los datos.

PROBLEMAS PARA LA SECCIÓN 3.4

Aprendizaje básico

3.37 A continuación se presenta un conjunto de datos para una muestra con $n = 11$ elementos:

X	7	5	8	3	6	10	12	4	9	15	18
Y	21	15	24	9	18	30	36	12	27	45	54

a. Calcule la covarianza.
b. Calcule el coeficiente de correlación.
c. ¿Qué tan fuerte es la relación entre X y Y? Explique su respuesta.

Aplicación de conceptos

Puede resolver los problemas 3.38 a 3.43 manualmente o en Excel, Minitab o SPSS.

3.38 En un artículo publicado recientemente (J. Clements, "Why Investors Should Put up to 30% of Their Stock Portfolio in Foreign Funds", *The Wall Street Journal,* 26 de noviembre, 2003, D1) que analiza las inversiones en acciones extranjeras asegura que: el coeficiente de correlación entre el rendimiento de inversiones en acciones estadounidenses y acciones internacionales de gran capital fue de 0.80; entre acciones estadounidenses y acciones internacionales de pequeño capital fue de 0.53; entre acciones estadounidenses y bonos internacionales fue de 0.03; entre acciones estadounidenses y acciones de mercados emergentes fue de 0.71; y entre acciones estadounidenses y deuda de mercados emergentes fue de 0.58.

a. ¿Qué conclusiones se obtienen sobre la fortaleza de la relación entre el rendimiento de inversiones en acciones estadounidenses y los otros cinco tipos de inversiones?
b. Compare los resultados de *a*) con los del problema 3.39*a*).

3.39 Un artículo publicado recientemente (J. Clements, "Why Investors Should Put up to 30% of Their Stock Portfolio in Foreign Funds", *The Wall Street Journal,* 26 de noviembre, 2003, D1) que analiza las inversiones en bonos extranjeros asegura que: el coeficiente de relación entre el rendimiento de la inversión en bonos estadounidenses y acciones internacionales de gran capital fue de –0.13; entre bonos estadounidenses y acciones internacionales de pequeño capital fue de –0.18; entre bonos estadounidenses y bonos internacionales fue de 0.48; entre bonos estadounidenses y acciones de mercados emergentes fue de –0.20; y entre bonos estadounidenses y deuda de mercados emergentes fue de 0.10.

a. ¿Qué conclusiones se obtienen sobre la fortaleza de la relación entre el rendimiento de las inversiones en bonos estadounidenses y los otros cinco tipos de inversiones?
b. Compare los resultados de *a*) con los del problema 3.38*a*).

3.40 Los siguientes datos **COFFEEDRINK** representan las calorías y la grasa (en gramos) que contienen las raciones con 16 onzas de bebidas a base de café servidas en Dunkin' Donuts y en Starbucks.

Producto	Calorías	Grasa
Batido de moka helado de Dunkin' Donuts (pura leche)	240	8.0
Capuchino frapé de Starbucks	260	3.5
Raspado de café "Coolata" (crema) de Dunkin' Donuts	350	22.0
Café moka exprés helado de Starbucks (pura leche y con crema batida)	350	20.0
Café moka batido helado de Starbucks (con crema batida)	420	16.0
Capuchino helado de Brownie de chocolate, de Starbucks (con crema batida)	510	22.0
77Crema de chocolate helado de Starbucks (con crema batida)	530	19.0

Fuente: "Coffee as Candy at Dunkin 'Donuts and Starbucks", Derechos Reservados © 2004 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057, organización sin fines de lucro. Adaptado de Consumer Reports, *junio de 2004, 9, sólo con propósitos educativos. No se autoriza su reproducción o uso comercial.* **www.ConsumerReports.org**

a. Calcule la covarianza de la muestra.
b. Calcule el coeficiente de correlación.
c. ¿Qué le parece más útil para expresar la relación que existe entre calorías y grasa: la covarianza o el coeficiente de correlación? Explique por qué.
d. ¿Qué conclusiones deduce acerca de la relación entre calorías y grasa?

3.41 Los siguientes datos representan el valor de exportaciones e importaciones de varios países en 2001: **EXPIMP**

País	Exportaciones	Importaciones
Unión Europea	874.1	912.8
Estados Unidos	730.8	1180.2
Japón	403.5	349.1
China	266.2	243.6
Canadá	259.9	227.2
Hong Kong	191.1	202.0
México	158.5	176.2
Corea del Sur	150.4	141.1
Taiwán	122.5	107.3
Singapur	121.8	116.0

Fuente: N. King y S. Miller, "Post-Iraq Influence of U.S. Faces Test at New Trade Talks", The Wall Street Journal, 9 de septiembre, 2003, A1.

a. Calcule la covarianza.
b. Calcule el coeficiente de correlación.
c. ¿Qué le parece más útil para expresar la relación que existe entre exportaciones e importaciones: la covarianza o el coeficiente de correlación? Explique por qué.
d. ¿Qué conclusiones puede deducir acerca de la relación entre exportaciones e importaciones?

AUTO Examen **3.42** Los siguientes datos **SECURITY** representan el porcentaje de traspaso durante 1998-1999 de los dispositivos de vigilancia utilizados antes de abordar en los aeropuertos, y las infracciones de seguridad detectadas por millón de pasajeros.

Ciudad	Traspaso	Infracciones
St. Louis	416	11.9
Atlanta	375	7.3
Houston	237	10.6
Boston	207	22.9
Chicago	200	6.5
Denver	193	15.2
Dallas	156	18.2
Baltimore	155	21.7
Seattle/Tacoma	140	31.5
San Francisco	110	20.7
Orlando	100	9.9
Washington-Dulles	90	14.8
Los Ángeles	88	25.1
Detroit	79	13.5
San Juan	70	10.3
Miami	64	13.1
Nueva York-JFK	53	30.1
Washington-Reagan	47	31.8
Honolulu	37	14.9

Fuente: Alan B. Krueger, "A Small Dose of Common Sense Would Help Congress Break the Gridlock over Airport Security", The New York Times, 15 de noviembre, 2001, C2.

a. Calcule la covarianza.
b. Calcule el coeficiente de correlación.
c. ¿Qué conclusiones obtiene sobre la relación que existe entre la tasa de traspaso de los dispositivos y las infracciones de seguridad detectadas?

3.43 Los siguientes datos **CELLPHONE** representan el tiempo en horas de uso de teléfonos móviles en modo digital y la capacidad de la batería en miliamperios.

Tiempo de uso	Capacidad de la batería	Tiempo de uso	Capacidad de la batería
4.50	800	1.50	450
4.00	1500	2.25	900
3.00	1300	2.25	900
2.00	1550	3.25	900
2.75	900	2.25	700
1.75	875	2.25	800
1.75	750	2.50	800
2.25	1100	2.25	900
1.75	850	2.00	900

Fuente: "Service Shortcomings", Copyright 2002 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, febrero de 2002, 25.

a. Calcule la covarianza.
b. Calcule el coeficiente de correlación.
c. ¿Qué conclusiones se obtienen sobre la relación entre la capacidad de la batería y el tiempo de uso en modo digital?
d. Usted espera que los teléfonos con batería de mayor capacidad tengan un tiempo de uso superior.¿Lo sustentan los datos?

3.5 ERRORES EN LAS MEDIDAS NUMÉRICAS DESCRIPTIVAS Y CONSIDERACIONES ÉTICAS

En este capítulo estudió cómo se definen las características de un conjunto de datos numéricos mediante varios estadísticos que miden las propiedades de su tendencia central, variación y forma. El siguiente paso es el análisis e interpretación de los estadísticos calculados. Su análisis es *objetivo*; su interpretación es *subjetiva*. Usted debe evitar los errores que surjan en la objetividad de su análisis o en la subjetividad de su interpretación.

El análisis de los fondos de inversión con base en el nivel de riesgo es *objetivo* y revela varios descubrimientos imparciales. Objetividad al analizar datos significa reportar las medidas numéricas descriptivas más apropiadas para un conjunto de datos determinado. Ahora que ha leído el capítulo y se ha familiarizado con varias medidas numéricas descriptivas y sus fortalezas y debilidades, ¿cómo continuará con el análisis objetivo? Como los datos se distribuyen de una manera ligeramente asimétrica, ¿no debería reportar la mediana además de la media? ¿La desviación estándar no ofrece más información sobre la propiedad de variación que el rango? ¿Debe describir al conjunto de datos como asimétrico a la derecha?

Por otra parte, la interpretación de datos es *subjetiva*. Al interpretar los descubrimientos analíticos, las personas elaboran conclusiones distintas. Todos vemos el mundo desde perspectivas diferentes. De esta manera, puesto que la interpretación de datos es subjetiva, usted debe hacerla de manera imparcial, neutral y clara.

Aspectos éticos

En todos los análisis de datos, los aspectos éticos son de vital importancia. Como consumidor cotidiano de información, usted debe cuestionar lo que lee en periódicos y revistas, lo que escucha en la radio y la televisión, así como lo que ve en Internet. A lo largo del tiempo, se ha manifestado mucho escepticismo sobre el propósito, el enfoque y la objetividad de los estudios que se publican. Quizá ningún comentario al respecto es más representativo que la frase atribuida al famoso estadista británico del siglo XIX, Benjamin Disraeli: "Existen tres clases de mentiras: las mentiras, las mentiras detestables y la estadística".

Las consideraciones éticas aparecen al decidir cuáles resultados incluir en un reporte. Usted debe documentar los resultados tanto buenos como malos. Además, al hacer exposiciones orales y presentar reportes escritos, debe comunicar los resultados de manera imparcial, objetiva y neutral. El comportamiento falto de ética se presenta al seleccionar de forma deliberada una medida resumida inapropiada (por ejemplo, la media de un conjunto de datos muy asimétrico), para distorsionar los hechos con el fin de respaldar una posición en particular. También es ético dejar de reportar de manera selectiva descubrimientos pertinentes, cuando éstos no respaldan una posición en particular.

RESUMEN

Este capítulo trató sobre las medidas descriptivas. En éste y el capítulo anterior, estudió la estadística descriptiva: cómo se presentan los datos en tablas y gráficas y luego su resumen, descripción, análisis e interpretación. Al manejar los datos relacionados con los fondos de inversión, usted tuvo la oportunidad de presentar información útil mediante el uso de diagramas circulares, histogramas y otros métodos gráficos. Exploró las características del desempeño en el pasado, como la tendencia central, variabilidad y forma, utilizando medidas descriptivas numéricas como la media, la mediana, los cuartiles, el rango, la desviación estándar y el coeficiente de correlación. En la tabla 3.9 se presenta una lista de las medidas descriptivas numéricas incluidas en este capítulo.

En el capítulo siguiente, se estudiarán los principios básicos de la probabilidad, con el fin de eliminar la brecha entre el tema de la estadística descriptiva y el de la estadística inferencial.

TABLA 3.9
Resumen de las medidas numéricas descriptivas.

Tipo de análisis	Datos numéricos
Describir la tendencia central, variación y forma de una variable numérica	Media, mediana, moda, cuartiles, media geométrica, rango, rango intercuartil, desviación estándar, varianza, coeficiente de variación, puntuaciones *Z*, gráfica de caja y bigote **(secciones 3.1-3.3)**
Describir la relación entre dos variables numéricas	Covarianza, coeficiente de correlación **(sección 3.4)**

FÓRMULAS IMPORTANTES

Media de una muestra

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \textbf{(3.1)}$$

Mediana

$$\text{Mediana} = \frac{n+1}{2} \text{ valor clasificado} \quad \textbf{(3.2)}$$

Primer cuartil Q_1

$$Q_1 = \frac{n+1}{4} \text{ valor clasificado} \quad \textbf{(3.3)}$$

Tercer cuartil Q_3

$$Q_3 = \frac{3(n+1)}{4} \text{ valor clasificado} \quad \textbf{(3.4)}$$

Media geométrica

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n} \quad \textbf{(3.5)}$$

Media geométrica de la tasa de rendimiento

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \quad \textbf{(3.6)}$$

Rango

$$\text{Rango} = X_{\text{mayor}} - X_{\text{menor}} \quad \textbf{(3.7)}$$

Rango intercuartil

$$\text{Rango intercuartil} = Q_3 - Q_1 \quad \textbf{(3.8)}$$

Varianza para una muestra

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \quad \textbf{(3.9)}$$

Desviación estándar de la muestra

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \quad \textbf{(3.10)}$$

Coeficiente de variación

$$CV = \left(\frac{S}{\bar{X}}\right)100\% \quad \textbf{(3.11)}$$

Puntuaciones *Z*

$$Z = \frac{X - \bar{X}}{S} \quad \textbf{(3.12)}$$

Media poblacional

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \quad \textbf{(3.13)}$$

Varianza poblacional

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} \quad \textbf{(3.14)}$$

Desviación estándar poblacional

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \quad \textbf{(3.15)}$$

La covarianza muestral

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \quad \textbf{(3.16)}$$

Coeficiente muestral de correlación

$$r = \frac{\text{cov}(X,Y)}{S_X S_Y} \quad \textbf{(3.17)}$$

CONCEPTOS CLAVE

PROBLEMAS DE REPASO

Revisión de su comprensión

3.44 ¿Cuáles son las propiedades de un conjunto de datos numéricos?

3.45 ¿Qué expresa la propiedad tendencia central?

3.46 ¿Cuáles son las diferencias entre media, mediana y moda, y cuáles son las ventajas y desventajas de cada una de ellas?

3.47 ¿Cómo interpreta el primer cuartil, la mediana y el tercer cuartil?

3.48 ¿Qué expresa la propiedad variación?

3.49 ¿Qué mide la puntuación Z?

3.50 ¿Cuáles son las diferencias entre las diversas medidas de la variación como rango, rango intercuartil, varianza, desviación estándar y coeficiente de variación, y cuáles son las ventajas y desventajas de cada una?

3.51 ¿Cómo nos ayuda la regla empírica a explicar de qué maneras se agrupan y distribuyen los valores de un conjunto de datos numéricos?

3.52 ¿En qué difieren la regla empírica y la regla de Chebyshev?

3.53 ¿Qué expresa la propiedad forma?

3.54 ¿En qué difieren la covarianza y el coeficiente de correlación?

Aplicación de conceptos

Puede resolver los problemas 3.55 a 3.61 manualmente o en Excel, Minitab, o SPSS. Le recomendamos resolver los problemas 3.62 a 3.80 con Excel, Minitab, o SPSS.

3.55 Una característica de calidad que resulta de interés en el proceso de llenado de bolsitas de té es el peso que contienen. Si las bolsas quedan semivacías, se presentan dos problemas. Primero, los clientes no podrían prepararse el té tan cargado como lo desean. Segundo, la empresa podría infringir las leyes de veracidad en lo descrito en la etiqueta. En este producto, el peso impreso en la etiqueta del paquete señala que, en promedio, hay 5.5 gramos de té en cada bolsa. Si la cantidad media de té en una bolsa supera ese peso, la empresa está regalando producto. Resulta complicado introducir la cantidad exacta de té en cada bolsa, puesto que la variación en las condiciones de temperatura y humedad dentro de la fábrica, las diferencias en la densidad del té y la rápida operación de llenado que realiza la máquina (aproximadamente 170 bolsas por minuto). La siguiente tabla muestra el peso, en gramos, de una muestra compuesta por 50 bolsas de té elaboradas en una hora por una sola máquina. **TEABAGS**

5.65	5.44	5.42	5.40	5.53	5.34	5.54	5.45	5.52	5.41
5.57	5.40	5.53	5.54	5.55	5.62	5.56	5.46	5.44	5.51
5.47	5.40	5.47	5.61	5.53	5.32	5.67	5.29	5.49	5.55
5.77	5.57	5.42	5.58	5.58	5.50	5.32	5.50	5.53	5.58
5.61	5.45	5.44	5.25	5.56	5.63	5.50	5.57	5.67	5.36

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Interprete las medidas de tendencia central y variación dentro del contexto de este problema. ¿Por qué debería preocuparse la compañía por la tendencia central y la variación?
d. Realice una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
e. ¿La empresa satisface el requisito dispuesto en la etiqueta de que, en promedio, hay 5.5 gramos de té por bolsa? Si usted estuviera a cargo de este proceso, ¿qué cambios, en caso necesario, trataría de hacer con respecto a la distribución de los pesos de las bolsas individuales?

3.56 En el estado de Nueva York las cajas de ahorro tienen permitido vender cierta clase de seguro de vida, llamado Seguro de Vida de Caja de Ahorro (SBLI, siglas en inglés para Savings Bank Life Insurance). El proceso de aprobación se compone de cada etapa de suscripción, la cual incluye una revisión de la solicitud, una consulta a la oficina de información médica, posibles peticiones de información médica adicional y exámenes médicos, así como la etapa de consolidación durante la cual se generan las pólizas y se envían al banco para su entrega. La capacidad de entregar a los clientes de manera oportuna las pólizas aprobadas resulta vital para que este servicio sea rentable para el banco. En el transcurso de un mes, se seleccionó una muestra aleatoria de 27 pólizas aprobadas, y se registró el siguiente tiempo de procesamiento total, en días: **INSURANCE**

73	19	16	64	28	28	31	90	60	56	31	56	22	18
45	48	17	17	17	91	92	63	50	51	69	16	17	

a. Calcule la media, la mediana, primero y tercer cuartiles.

b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.

c. Elabore una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?

d. ¿Qué le respondería usted a un cliente que entra al banco con el fin de comprar este tipo de póliza de seguros y le pregunta cuánto dura el proceso de aprobación?

3.57 Una de las principales medidas de la calidad del servicio que brinda cualquier organización es la velocidad con la que responde a las quejas del cliente. Una gran tienda departamental, propiedad de una familia que vende muebles y pisos, incluyendo alfombras, emprendió una importante expansión durante los últimos años. En particular el departamento de pisos se amplió de dos equipos de instalación a un supervisor de instalación, un medidor, y 15 equipos de instalación. Se seleccionó una muestra de 50 quejas relacionadas con la instalación de alfombras, recibidas durante uno de los últimos años. Los siguientes datos representan el número de días transcurridos desde que se recibió la queja hasta su solución. **FURNITURE**

54	5	35	137	31	27	152	2	123	81	74	27
11	19	126	110	110	29	61	35	94	31	26	5
12	4	165	32	29	28	29	26	25	1	14	13
13	10	5	27	4	52	30	22	36	26	20	23
33	68										

a. Calcule la media, la mediana, primero y tercer cuartiles.

b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.

c. Elabore una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?

d. Con base en los resultados de los incisos *a*) a *c*), si usted tuviera que informar al presidente de la empresa cuánto tendrá que esperar un cliente para ver su queja resuelta, ¿qué le diría? Explique su respuesta.

3.58 Una empresa de manufactura produce gabinetes de acero para equipo eléctrico. El principal componente del gabinete es una canaleta que se elabora con lámina de acero calibre 14. Se produce utilizando una troqueladora de deslizamiento progresivo de 250 toneladas, que genera dos formaciones de 90 grados en el acero plano, haciendo el canal. La distancia de un lado al otro de estas formaciones resulta de especial importancia, por la impermeabilización para aplicaciones a la intemperie. La empresa necesita que la canaleta tenga una anchura de entre 8.31 y 8.61 pulgadas. A continuación encuentran las anchuras, en pulgadas, de una muestra de $n = 49$ canaletas. **TROUGH**

8.312	8.343	8.317	8.383	8.348	8.410	8.351	8.373	8.481	8.422
8.476	8.382	8.484	8.403	8.414	8.419	8.385	8.465	8.498	8.447
8.436	8.413	8.489	8.414	8.481	8.415	8.479	8.429	8.458	8.462
8.460	8.444	8.429	8.460	8.412	8.420	8.410	8.405	8.323	8.420
8.396	8.447	8.405	8.439	8.411	8.427	8.420	8.498	8.409	

a. Calcule la media, la mediana, el rango y la desviación estándar de la anchura. Interprete estas medidas de tendencia central y variabilidad.

b. Elabore el resumen de cinco números.

c. Realice su gráfica de caja y bigote y describa la forma.

d. ¿Qué concluye sobre el número de canaletas que satisfacen las necesidades de la empresa, al medir entre 8.31 y 8.61 pulgadas de ancho?

3.59 La empresa del problema 3.58 también fabrica aislantes eléctricos. Si los aislantes se rompen al estar en uso, es probable que ocurra un cortocircuito. Para poner a prueba la fuerza de los aislantes, se efectúa una prueba de destrucción con la finalidad de determinar cuánta *fuerza* se necesita para romperlos. La fuerza se mide al observar cuántas libras se aplican al aislante antes de que se rompa. A continuación se presentan los datos de 30 aislantes en este experimento: **FORCE**

1,870	1,728	1,656	1,610	1,634	1,784	1,522	1,696	1,592	1,662
1,866	1,764	1,734	1,662	1,734	1,774	1,550	1,756	1,762	1,866
1,820	1,744	1,788	1,688	1,810	1,752	1,680	1,810	1,652	1,736

a. Calcule la media, la mediana, el rango y la desviación estándar de la variable fuerza.

b. Interprete las medidas de tendencia central y de variabilidad del inciso *a*).

c. Construya su gráfica de caja y bigote y describa la forma.

d. ¿Qué concluye sobre la resistencia de los aislantes, si la empresa necesita una medición de al menos 1,500 libras de fuerza?

3.60 Los problemas de una línea telefónica que impiden hacer o recibir llamadas desconciertan tanto al cliente como a la empresa telefónica. Los siguientes datos representan muestras de 20 problemas reportados a dos oficinas distintas de una empresa telefónica, y el tiempo transcurrido para resolverlos (en minutos) desde la línea del cliente: **PHONE**

Central telefónica I Tiempo para resolver problemas (minutos)

1.48	1.75	0.78	2.85	0.52	1.60	4.15	3.97	1.48	3.10
1.02	0.53	0.93	1.60	0.80	1.05	6.32	3.93	5.45	0.97

Central telefónica II Tiempo para resolver problemas (minutos)

7.55	3.75	0.10	1.10	0.60	0.52	3.30	2.10	0.58	4.02
3.75	0.65	1.92	0.60	1.53	4.23	0.08	1.48	1.65	0.72

Para ambas centrales telefónicas:

a. Calcule la media, la mediana, primero y tercer cuartiles.

b. Calcule el rango, rango intercuartil, varianza, desviación estándar y coeficiente de variación.

c. Elabore una gráfica de barras de lado a lado y una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?

d. Con base en los resultados de los incisos *a*) a *c*), ¿existen algunas diferencias entre ambas centrales? Explique su respuesta.

3.61 En muchos procesos de manufactura se utiliza el término "trabajo-en-proceso" (con frecuencia abreviado WIP, por las siglas en inglés para "work-in-process"). En una planta que produce libros, el WIP representa el tiempo que transcurre para que se doblen, junten, cosan, peguen por un extremo y encuadernen las hojas procedentes de la prensa. Los siguientes datos representan muestras de 20 libros en dos plantas de producción y el tiempo de procesamiento (definido de forma operacional como el tiempo, en días, transcurrido desde que las hojas salen de la prensa hasta que los libros se empacan en cajas) para estos trabajos. **WIP**

Planta A

5.62	5.29	16.25	10.92	11.46	21.62	8.45	8.58	5.41	11.42
11.62	7.29	7.50	7.96	4.42	10.50	7.58	9.29	7.54	8.92

Planta B

9.54	11.46	16.62	12.62	25.75	15.41	14.29	13.13	13.71	10.04
5.75	12.46	9.17	13.21	6.00	2.33	14.25	5.37	6.25	9.71

Para ambas plantas:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore las gráficas de barra de lado a lado y de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. Con base en los resultados de los incisos *a*) a *c*), ¿existen algunas diferencias entre ambas plantas? Explique su respuesta.

3.62 Los datos incluidos en el archivo **CEREALS** se componen del costo monetario por onza, calorías, fibra en gramos y azúcar en gramos, de 33 cereales para desayunar.

Fuente: Obtenido de Copyrigh 1999 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, octubre de 1999, 33-34.

Para cada una de las variables:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué concluye en relación con el costo por onza en centavos, calorías, fibra en gramos y azúcar en gramos, de los 33 cereales para desayunar?

3.63 Los recortes presupuestales estatales forzaron el aumento en los costos de manutención para las universidades públicas durante el ciclo escolar 2003-2004. Los datos que se encuentran en el archivo **TUITION** incluyen la diferencia en los costos de manutención entre los ciclos 2002-2003 y 2003-2004 para los alumnos procedentes del mismo estado donde se encuentra la institución y los procedentes de otros estados.

a. Calcule la media, la mediana, primero y tercer cuartiles de la diferencia en los costos de manutención entre los ciclos 2002-2003 y 2003-2004 para los alumnos procedentes del mismo estado donde se encuentra la institución y los procedentes de otros estados.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación de la diferencia en los costos de manutención entre los ciclos 2002-2003 y 2003-2004 para los alumnos procedentes del mismo estado donde se encuentra la institución y los procedentes de otros estados.
c. Elabore la gráfica de caja y bigote de la diferencia en los costos de manutención entre los ciclos 2002-2003 y 2003-2004 para los alumnos procedentes del mismo estado donde se encuentra la institución y los procedentes de otros estados. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtendría en relación con la diferencia en los costos de manutención entre los ciclos 2002-2003 y 2003-2004 para los alumnos procedentes del mismo estado donde se encuentra la institución y los procedentes de otros estados?

3.64 Las promociones de marketing, como la entrada gratis a las personas con gorra, ¿aumentan la asistencia a los juegos de la Liga Mayor de Béisbol? Un artículo publicado en *Sport Marketing Quarterly* informa sobre la efectividad de las promociones de marketing [T. C. Boyd y T. C. Krehbiel, "Promotion Timing in Major League Baseball and the Stacking Effects of Factors that Increase Game Attractiveness", *Sport Marketing Quarterly,* 12(2003), 173-183]. El archivo de datos **ROYALS** incluye las siguientes variables para los Reales de Kansas City durante la temporada 2002:

GAME = juegos como local en el orden en que se jugaron.
ATTENDANCE = espectadores con boleto pagado en ese juego.
PROMOTION-Y = hubo promoción; N = no hubo promoción.

a. Calcule la media y la desviación estándar de los espectadores con boleto pagado para los 43 juegos en los que hubo promoción y para los 37 juegos sin promoción.
b. Elabore un resumen de cinco números para los 43 juegos en los que hubo promoción y para los 37 juegos sin promoción.
c. Realice una representación que contenga dos gráficas de caja y bigote; una de los 43 juegos en los que hubo promoción y otra de los 37 juegos sin promoción.
d. Analice los resultados de los incisos a) a c) y comente sobre la eficacia de las promociones en los juegos de los Reales durante la temporada 2002.

3.65 Los datos incluidos en el archivo **PETFOOD2** se componen del costo por ración, tasas por lata, proteína en gramos y grasa en gramos de 97 variedades de comida seca y enlatada para perro y para gato.

Fuente: Obtenido de Copyright 1998 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, febrero de 1998, 18-19.

Realice lo siguiente para los cuatro tipos de comida (comida seca para perro, comida enlatada para perro, comida seca para gato y comida enlatada para gato), y para las variables costo por servicio, proteína en gramos y grasa en gramos:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.

c. Elabore las gráficas de barras de lado a lado y la de caja y bigote, de los cuatro tipos (comida seca para perro, comida enlatada para perro, comida seca para gato y comida enlatada para gato). ¿Son asimétricos los datos de alguno de los tipos de comida? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en relación con las diferencias entre los cuatro tipos (comida seca para perro, comida enlatada para perro, comida seca para gato y comida enlatada para gato)?

3.66 Un fabricante de tejas de asfalto de Boston y Vermont ofrece a sus clientes una garantía de 20 años en la mayoría de sus productos. Para determinar si una teja dura tanto como el periodo de garantía, se realiza una prueba de vida acelerada en la planta. En la prueba, realizada en un laboratorio, la teja se expone a las tensiones que recibiría en toda su vida útil de uso normal, mediante un experimento que lleva tan sólo unos minutos. En esta prueba, se cepilla repetidamente una teja durante un breve lapso, y se pesa la cantidad de gránulos (en gramos) desprendidos por el cepillado. Se espera que las tejas con menor desprendimiento duren más en uso normal que las que experimentan gran cantidad de desprendimiento. Ante esta situación, si se espera que dure tanto como el periodo de garantía, una teja no debe tener un desprendimiento superior a 0.8 gramos. El archivo **GRANULE** contiene los datos de una muestra compuesta por 170 medidas realizadas en las tejas de la empresa en Boston y 140 medidas realizadas en las tejas de Vermont.
a. Elabore el resumen de cinco puntos para las tejas de Boston y las tejas de Vermont.
b. Realice las gráficas barras de lado a lado y de caja y bigote para ambos tipos de teja, y describa la forma de las distribuciones.
c. Comente sobre la capacidad de las tejas para conseguir un desprendimiento de 0.8 gramos o menos.

3.67 Los datos del archivo **STATES** representan los resultados de la Encuesta de la Comunidad Estadounidense (American Community Survey), con una muestra de 700,000 hogares emprendida en todos los estados durante el censo de EUA del año 2000. Realice lo siguiente para las variables tiempo promedio de traslado al trabajo en minutos, porcentaje de hogares con ocho o más habitaciones, ingreso medio y porcentaje de propietarios con hipoteca, cuyos costos de vivienda superan el 30% de sus ingresos:
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Realice una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en relación con el tiempo promedio de traslado al trabajo en minutos, porcentaje de hogares con ocho o más habitaciones, ingreso medio y porcentaje de propietarios con hipoteca cuyos costos de vivienda superan el 30% de sus ingresos?

3.68 Las finanzas del béisbol han provocado mucha controversia, pues los propietarios aseguran que pierden dinero, los jugadores afirman que los propietarios ganan dinero, y los aficionados se quejan por lo costoso que resulta asistir a los juegos o verlos por televisión de paga. Además de los datos relacionados con las estadísticas del equipo durante la temporada 2001, el archivo **BB2001** contiene las estadísticas de todos los equipos sobre precios de las entradas, índice de costo por aficionado, ingresos por entradas en temporada regular, ingresos por televisión local, radio y cable; todos los demás ingresos de operación, compensación y beneficios del jugador; datos locales y nacionales e ingresos por operaciones de béisbol. Para cada una de estas variables:
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. Calcule la correlación que existe entre el número de victorias y las compensaciones y beneficios del jugador. ¿Qué tan fuerte es la relación entre estas dos variables?
e. ¿Qué conclusiones obtiene en relación con los ingresos por entradas en temporada regular, ingresos por televisión local, radio y cable; todos los demás ingresos de operación, compensación y beneficios del jugador; datos locales y nacionales e ingresos por operaciones de béisbol?

3.69 Los datos incluidos en el archivo **AIRCLEANERS** representan el precio, el costo anual de energía y el costo anual del filtro de unos limpiadores de aire.
a. Calcule el coeficiente de correlación entre el precio y el costo de energía.
b. Calcule el coeficiente de correlación entre el precio y el costo del filtro.
c. ¿Qué conclusiones obtiene sobre la relación del costo de energía y del costo del filtro con el precio de los limpiadores de aire?

Fuente: "Portable Room Air Cleaners", Copyright © 2002 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, febrero de 2002, 47.

3.70 Los datos incluidos en el archivo **PRINTERS** representan el precio, la velocidad de texto, el costo de texto, el tiempo de fotografía a color y el costo de la fotografía color de unas impresoras de computadora.
a. Calcule el coeficiente de correlación entre el precio y cada una de las siguientes características: velocidad de texto, costo de texto, tiempo de fotografía a color y costo de fotografía a color.
b. Con base en los resultados del inciso *a*), ¿cree usted que alguna de las demás variables podría ser útil para pronosticar el precio de la impresora? Explique su respuesta.

Fuente: "Printers", Copyright © 2002 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, marzo de 2002, 51.

3.71 Usted quiere estudiar las características de los automóviles modelo 2002, en términos de las siguientes variables: millas por galón, longitud, anchura, necesidades de circunferencia de viraje, peso y capacidad del compartimiento de equipaje. **AUTO2002**

Fuente: "The 2002 Cars", Copyright © 2002 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, abril de 2002.

Para cada una de esas variables:
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore una gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en relación con los automóviles 2002?

3.72 Consulte los datos del problema 3.71. Usted quiere comparar los vehículos utilitarios (o SUV, siglas en inglés para sports utility vehicles) con los que no son de ese tipo, en términos de millas por galón, longitud, anchura, necesidades de circunferencia de viraje, peso y capacidad del compartimiento de carga. Para cada una de esas variables, y considerando dos tipos de vehículos:
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore las gráficas de barras de lado a lado y de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en relación con las diferencias entre los SUV y los vehículos de otra clase?

3.73 Zagat's publica las calificaciones de restaurantes en varias ciudades de Estados Unidos. El archivo **RESTRATE** contiene los datos de la calificación para la comida, decorado, servicio y precio por persona de una muestra compuesta por 50 restaurantes localizados en la ciudad de Nueva York, y 50 localizados en Long Island.

Fuente: Zagat Survey 2002 New York City Restaurants and Zagat Survey 2002 Long Island Restaurants.

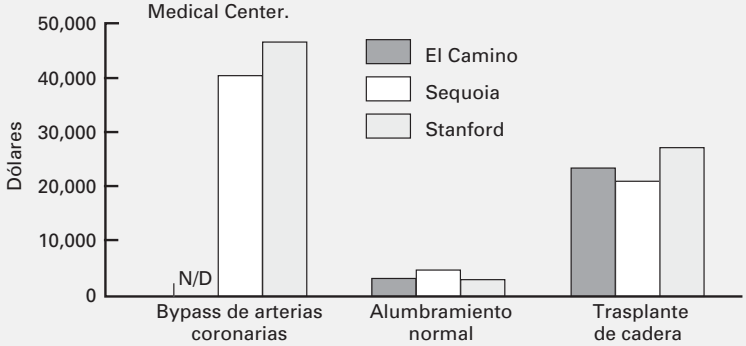
Para los restaurantes de Nueva York y Long Island, las variables calificación de la comida, calificación del decorado, calificación del servicio y calificación del precio por persona:
a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore una gráfica de caja y bigote y una de barras de lado a lado de Nueva York y Long Island. ¿Son asimétricos los datos de alguna de las variables? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en relación con las diferencias que existen entre los restaurantes de Nueva York y Long Island?

3.74 Como un ejemplo del mal uso de la estadística, un artículo de Glenn Kramon ("Coaxing the Stanford Elephant to Dance", *The New York Times* Sunday Business Section, 11 de noviembre, 1990) describe que los costos del Stanford Medical Center se habían elevado más que los de la competencia ya que era más probable que brindara atención a personas indigentes, más enfermas, beneficiarios de Medicare y Medicaid, y pacientes con problemas más complejos. Se utilizó la gráfica que aparece más adelante para comparar los precios promedio en 1989 y 1990 de tres procedimientos médicos (bypass de arterias coronarias, alumbramiento normal y trasplante de cadera) en tres instituciones competidoras (El Camino, Sequoia y Stanford).

Suponga que trabaja en un centro de salud. La directora general sabe que usted está tomando un curso de estadística y le llama para analizar esto. Le dice que anoche se presentó ese artículo en el marco de una discusión de grupo, como parte de una reunión de directores generales de los centros de salud de la zona, y que uno de ellos mencionó que la gráfica era totalmente irrelevante y le pidió su opinión. Ahora ella le pide que prepare la respuesta. Usted sonríe, respira profundo y responde...

3.75 Usted planea estudiar para su examen de estadística con un grupo de compañeros, uno de los cuales está especialmente interesado en impresionarlo. Este individuo se

Cuánto cuestan los servicios de salud
Comparación de los precios hospitalarios promedio de varias operaciones durante 1989-90 en California. Los hospitales Sequoia y El Camino son los principales competidores locales del Stanford Medical Center.



Para el alumbramiento normal con una estancia de dos días y para el trasplante de cadera con una estancia de nueve días, los costos de El Camino son el promedio de los precios alto y bajo.

En todas las operaciones, los costos del Sequoia son los promedios del 50% medio de todos los precios.

Los datos de Stanford son el costo promedio de todas las operaciones.

Fuente: Stanford Medical Center, Sequoia Hospital y Hospital El Camino.

ofreció a trabajar voluntariamente con Excel, Minitab o SPSS para obtener información resumida, tablas y gráficas necesarias para el conjunto de datos que contiene diversas variables numéricas y categóricas estipulado por el maestro como objeto de estudio. Se le acerca con los resultados impresos y exclama: "Lo tengo todo: —las medias, las medianas, las desviaciones estándar, las gráficas de caja y bigote, y los diagramas de pastel— de todas nuestras variables. El problema es que algunos de los resultados parecen extraños, como las gráficas de caja y bigote para género y mayores de edad, y los diagramas de pastel del índice de nivel de estudios y de la estatura. Tampoco entiendo por qué el profesor Krehbiel dice que no podemos obtener la estadística descriptiva de algunas de las variables; ¡las tengo para todo! Mira, la media de la estatura es 68.23, la media del índice de nivel de estudios es 2.76, la media del género es 1.50, la media para los mayores de edad es 4.33". ¿Cuál sería su respuesta?

Ejercicios de reporte por escrito

3.76 Los datos que aparecen en el archivo **BEER** representan el precio de un paquete de cerveza con 6 botellas de 12 onzas cada una, las calorías en 12 onzas líquidas, el porcentaje de contenido alcohólico en 12 onzas líquidas, el tipo de cerveza (artesanales de baja fermentación, artesanales de alta fermentación, importadas de baja fermentación, regulares y frías, y cervezas *light* y sin alcohol), y el país de origen (estadounidenses y del resto del mundo) de cada una de las 69 cervezas incluidas en la muestra.

Su tarea consiste en escribir un reporte con base en una evaluación descriptiva completa de las variables numéricas (precio, calorías y contenido alcohólico) independientemente del tipo u origen del producto. Luego realice una evaluación similar, comparando cada una de esas variables numéricas con base en el tipo de producto (artesanales de baja fermentación, artesanales de alta fermentación, importadas de baja fermentación, regulares y frías, y cervezas *light* y sin alcohol). Efectúe también una evaluación similar, para comparar y establecer las diferencias de cada una de esas variables numéricas, con base en el origen de las cervezas: las preparadas en Estados Unidos contra las del resto del mundo. Junto con su reporte debe anexar todas las tablas, los diagramas y las medidas numéricas descriptivas apropiadas.

Fuente: "Beers", Copyright © 1996 por Consumers Union of U.S., Inc., Yonkers, NY 10703-1057. Adaptado con autorización de Consumer Reports, junio de 1996.

PROYECTO EN EQUIPO

El archivo **MUTUALFUNDS2004** contiene información relacionada con 12 variables a partir de una muestra de 121 fondos de inversión. Las variables son:

Fund —Nombre del fondo de inversión.
Category —Tipo de acciones que abarca el fondo de inversión: pequeño, mediano o gran capital.
Objective —Objetivo de las acciones que abarca el fondo de inversión: crecimiento o valor.
Assets —Activos en millones de dólares.
Fees —Cargos por venta (no o sí).
Expense ratio —Relación entre gastos y activos netos, en porcentaje.
2003 Return —Rendimiento en los 12 meses de 2003.
Three-year return —Rendimiento anualizado 2001 a 2003.
Five-year return —Rendimiento anualizado 1999 a 2003.
Risk —Factor de riesgo de pérdida del fondo de inversión, clasificado como bajo, medio o alto.
Best quarter —Mejor resultado trimestral 1999 a 2003.
Worst quarter —Peor resultado trimestral 1999 a 2003.

3.77 Para la relación de gastos en porcentaje, el rendimiento en 2003, el rendimiento trianual y el rendimiento quinquenal:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore la gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en relación con estas variables?

3.78 Usted quiere comparar los fondos de inversión que tienen cuotas o cargos con los que no los tienen. Realice lo siguiente con cada uno de los dos grupos, para las variables relación de gastos en porcentaje, rendimiento en 2003, rendimiento trianual y rendimiento quinquenal:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore la gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en cuanto a las diferencias que existen entre los fondos de inversión con y sin cuotas?

3.79 Usted quiere comparar los fondos de inversión que tienen un objetivo de crecimiento con los que tienen un objetivo de valor. Realice lo siguiente con cada uno de los dos grupos, para las variables coeficiente de gastos en porcentaje, rendimiento en 2003, rendimiento trianual y rendimiento quinquenal:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore la gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene en cuanto a las diferencias que existen entre los fondos con objetivo de crecimiento y los fondos con objetivo de valor?

3.80 Usted quiere comparar los fondos de inversión para pequeño, mediano y gran capital. Realice lo siguiente con cada uno de los tres grupos, para las variables coeficiente de gastos en porcentaje, rendimiento en 2003, rendimiento trianual y rendimiento quinquenal:

a. Calcule la media, la mediana, primero y tercer cuartiles.
b. Calcule el rango, el rango intercuartil, la varianza, la desviación estándar y el coeficiente de variación.
c. Elabore la gráfica de caja y bigote. ¿Los datos son asimétricos? De ser así, ¿cómo?
d. ¿Qué conclusiones obtiene con respecto a las diferencias que existen entre los fondos de inversión para pequeño, mediano y gran capital?