CAPÍTULO 3

Estadística descriptiva: medidas numéricas

CONTENIDO

ESTADÍSTICA *en* LA PRÁCTICA

SMALL FRY DESIGN*

SANTA ANA, CALIFORNIA

Small Fry Design, fundada en 1997, es una compañía de juguetes y accesorios que diseña e importa productos para niños. La línea de artículos de la empresa incluye osos de peluche, móviles, juguetes musicales, sonajas y cobertores de seguridad, y presenta diseños de juguetes de alta calidad para bebé con un énfasis en los colores, las texturas y los sonidos. Los productos se diseñan en Estados Unidos y se fabrican en China.

Small Fry Design emplea a representantes independientes para la venta de sus productos a minoristas de muebles infantiles, tiendas de accesorios y ropa para niños, negocios de regalos, tiendas departamentales exclusivas e importantes compañías de ventas por catálogo. En la actualidad, los productos de Small Fry Design se distribuyen en más de 1 000 puntos de venta minoristas en todo Estados Unidos.

La administración del flujo de efectivo es una de las actividades más importantes para la operación diaria de esta empresa. Garantizar que dicho flujo entrante sea suficiente para cumplir con las obligaciones de deudas tanto corrientes como a corto plazo puede significar la diferencia entre el éxito y el fracaso. Un factor crítico en la administración del flujo de efectivo es el análisis y control de las cuentas por cobrar. Al medir el tiempo promedio de cobro y el valor monetario de las facturas pendientes, la gerencia puede predecir la disponibilidad de efectivo y monitorear los cambios en el estado de las cuentas por cobrar. La empresa estableció las metas siguientes: la antigüedad promedio de las facturas pendientes no debe exceder los 45 días y el valor de las facturas con una antigüedad mayor a 60 días no debe exceder 5% del valor de todas las cuentas por cobrar.

En un resumen reciente del estado de las cuentas por cobrar se proporcionó la siguiente estadística descriptiva para la antigüedad de las facturas pendientes.

Media	40 días
Mediana	35 días
Moda	31 días

Móvil "Rey de la selva" de Small Fry Design.
© Joe-Higgins/South-Western.

La interpretación de estos datos estadísticos muestra que el tiempo promedio de cobro de una factura es de 40 días. La mediana señala que la mitad de estos documentos permanece pendiente 35 días o más. La moda de 31 días, el tiempo de cobro de una factura más frecuente, indica que el lapso más común en que ésta permanece pendiente es de 31 días. El resumen estadístico indica también que sólo 3% del valor de todas las cuentas por cobrar tiene un tiempo de cobro de más de 60 días. Con base en la información estadística, la gerencia quedó satisfecha, dado que las cuentas por cobrar y el flujo de efectivo entrante estaban bajo control.

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

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

En el capítulo 2 se estudiaron las presentaciones tabulares y gráficas utilizadas para resumir los datos. En este capítulo se presentan varias medidas numéricas que proporcionan otras opciones para la misma tarea.

Primero se verá el desarrollo de medidas numéricas para conjuntos de datos que constan de una sola variable. Cuando un conjunto de datos contiene más de una variable, las mismas medidas numéricas se calculan por separado para cada variable. Sin embargo, en el caso de dos variables, se desarrollarán también medidas de la relación entre éstas.

Se presentan las medidas numéricas de posición, dispersión, forma y asociación. Si las medidas se calculan para los datos de una muestra, se les llama **estadístico muestral**. Si se calculan para los datos de una población, se les llama **parámetros poblacionales**. En la inferencia estadística, un estadístico muestral se conoce como **estimador puntual** del parámetro poblacional correspondiente. En el capítulo 7 se verá con más detalle el proceso de la estimación puntual.

En los tres apéndices del capítulo se explica cómo se usan Minitab, Excel y StatTools para calcular las medidas numéricas descritas en el capítulo.

3.1 Medidas de posición o localización

Media

La **media**, o valor medio, es quizá la medida de ubicación más importante para una variable, pues proporciona una medida de la ubicación central de los datos. Si los datos son para una muestra, la media se denota por $\bar{x}$; si son para una población, se denota por la letra griega μ.

En las fórmulas estadísticas se acostumbra denotar el valor de la primera observación de la variable x mediante x_1, el valor de la segunda observación de la variable x por medio de x_2, y así sucesivamente. En general, el valor de la i-ésima observación de la variable x se representa por medio de x_i. Si se tiene una muestra con n observaciones, la fórmula para la media muestral es la siguiente.

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

MEDIA MUESTRAL

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

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

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

La letra griega Σ es el signo de sumatoria.

Para ilustrar el cálculo de una media muestral, considere los datos siguientes sobre el tamaño del grupo para una muestra de cinco grupos de estudiantes universitarios.

46 54 42 46 32

La notación x_1, x_2, x_3, x_4, x_5 se utiliza para representar el número de estudiantes en cada uno de los cinco grupos.

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

Por consiguiente, para calcular la media muestral se escribe

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

El tamaño de grupo de la media muestral es 44 estudiantes.

Otro ejemplo del cálculo de una media muestral se da en la situación siguiente. Suponga que una oficina de colocación de empleos a nivel universitario envió un cuestionario a una muestra de licenciados en administración de empresas recién egresados solicitando información sobre

TABLA 3.1 Sueldos mensuales iniciales para una muestra de 12 licenciados en administración de empresas recién egresados

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

WEB archivo
StartSalary

los sueldos mensuales iniciales. La tabla 3.1 exhibe los datos reunidos. El sueldo mensual inicial medio para la muestra de 12 licenciados en administración de empresas se calcula como sigue:

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12}$$

$$= \frac{3450 + 3550 + \cdots + 3480}{12}$$

$$= \frac{42480}{12} = 3540$$

La ecuación (3.1) ilustra cómo se calcula la media para una muestra con n observaciones. La fórmula para determinar la media de una población es la misma, pero se usa una notación diferente para indicar que se está trabajando con toda la población. El número de observaciones en una población se denota por N y el símbolo para la media poblacional es μ.

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

MEDIA POBLACIONAL

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

Mediana

La **mediana** es otra medida de ubicación central; es el valor de en medio cuando los datos están acomodados en orden ascendente (del valor menor al valor mayor). Con un número impar de observaciones, la mediana es el valor de en medio. Con un número par, no hay valor de en medio. En este caso se sigue la convención y la mediana se define como el promedio de los valores de las dos observaciones de en medio. Por conveniencia, la definición de la mediana se replantea como sigue.

MEDIANA

Ordene los datos de forma ascendente (del valor menor al valor mayor).

a) Para un número impar de observaciones, la mediana es el valor de en medio.
b) Para un número par de observaciones, la mediana es el promedio de los dos valores de en medio.

Esta definición se aplica para calcular la mediana de los tamaños de grupo para la muestra de cinco grupos de estudiantes universitarios. Al ordenar los datos de forma ascendente se obtiene la lista siguiente.

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

Dado que $n = 5$ es impar, la mediana es el valor de en medio. Por tanto, la mediana del tamaño de grupo es 46 estudiantes. Aun cuando este conjunto de datos contiene dos observaciones con valores de 46, cada una se trata de forma separada cuando los datos se acomodan en orden ascendente.

Suponga además que se calcula la mediana de los sueldos iniciales para los 12 licenciados en administración de empresas de la tabla 3.1. Primero se acomodan los datos en orden ascendente.

$$3\,310 \quad 3\,355 \quad 3\,450 \quad 3\,480 \quad 3\,480 \quad \underbrace{3\,490 \quad 3\,520}_{\text{Los dos valores de en medio}} \quad 3\,540 \quad 3\,550 \quad 3\,650 \quad 3\,730 \quad 3\,925$$

Como $n = 12$ es par, se identifican los dos valores de en medio: la mediana es el promedio de estos dos valores.

$$\text{Mediana} = \frac{3\,490 + 3\,520}{2} = 3\,505$$

La mediana es la medida de posición más empleada para los datos de los ingresos anuales y el valor de propiedad, debido a que algunos ingresos o valores de propiedad muy grandes pueden inflar la media. En tales casos, la mediana es la medida preferida de posición central.

Aunque la media es la medida de posición central de uso más común, en algunas situaciones se prefiere la mediana, ya que los valores de datos muy pequeños y muy grandes influyen en la media. Por ejemplo, suponga que uno de los licenciados recién graduados (tabla 3.1) tenía un sueldo inicial de $10 000 al mes (tal vez la empresa es propiedad de su familia). Si se cambia el sueldo mensual inicial más alto de la tabla 3.1 de $3 925 a $10 000 y se vuelve a calcular la media, la media muestral pasa de $3 540 a $4 046. Sin embargo, la mediana de $3 505 permanece igual, ya que $3 490 y $3 520 siguen siendo los dos valores de en medio. Si el sueldo inicial es sumamente alto, la mediana proporciona una mejor medida de posición central que la media. Al hacer una generalización, se afirma que siempre que un conjunto de datos contiene valores extremos, la mediana suele ser la medida preferida de posición central.

Moda

Una tercera medida de posición es la **moda**. Se define de la manera siguiente.

MODA

La moda es el valor que ocurre con mayor frecuencia.

Para ilustrar cómo identificar la moda, considere el tamaño de grupo de la muestra de cinco grupos de estudiantes universitarios. El único valor que ocurre más de una vez es el 46. Debido a que se presenta con una frecuencia de 2, que es la frecuencia más grande, se le considera la moda. Como otro ejemplo, considere la muestra de sueldos iniciales de los licenciados en administración de empresas. El único sueldo mensual inicial que ocurre más de una vez es $3 480. Dado que este valor tiene la frecuencia mayor, es la moda.

Hay situaciones en que la frecuencia mayor ocurre en dos o más valores diferentes; cuando esto sucede, existe más de una moda. Si los datos contienen exactamente dos modas, se dice que son *bimodales.* Si contienen más de dos, se dice que son *multimodales.* En estos casos, la moda casi nunca se presenta debido a que listar tres o más no resulta particularmente útil para describir la posición de los datos.

Percentiles

Un **percentil** proporciona información sobre cómo se distribuyen los datos en el intervalo del valor menor al valor mayor. Para datos que no contienen muchos valores repetidos, el percentil p-ésimo los divide en dos partes. Alrededor de p por ciento de las observaciones tiene valores menores que el percentil p-ésimo y cerca de $(100 - p)$ por ciento de las observaciones tiene valores mayores que el percentil p-ésimo. Éste se define formalmente del modo siguiente.

PERCENTIL

El percentil p-ésimo es un valor tal que *por lo menos* p por ciento de las observaciones es menor o igual que este valor, y *por lo menos* $(100 - p)$ por ciento de las observaciones es mayor o igual que este valor.

Los colegios y universidades suelen reportar los resultados de los exámenes de admisión en términos de percentiles. Por ejemplo, suponga que un solicitante obtiene una puntuación bruta de 54 en la parte verbal de un examen de admisión. Esta información no dice mucho acerca del desempeño que este estudiante tuvo en relación con otros que presentaron el mismo examen. Sin embargo, si la puntuación bruta de 54 corresponde al percentil 70, se sabe que aproximadamente 70% de los estudiantes obtuvo una puntuación menor a la de esta persona y alrededor de 30% alcanzó una puntuación mayor a la de esta persona.

El procedimiento siguiente se usa para calcular el p-ésimo percentil.

La ejecución de estos pasos facilita el cálculo de percentiles.

CÁLCULO DEL p-ÉSIMO PERCENTIL

Paso 1. Ordene los datos de modo ascendente (del valor menor al valor mayor).

Paso 2. Calcule un índice i

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

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

Paso 3. *a*) Si i *no es un entero, redondéelo.* El entero siguiente *mayor* que i denota la posición del p-ésimo percentil.

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

Como ejemplo de este procedimiento, se determinará el percentil 85 para los datos de los sueldos iniciales mensuales de la tabla 3.1.

Paso 1. Ordene los datos de modo ascendente.

3 310 3 355 3 450 3 480 3 480 3 490 3 520 3 540 3 550 3 650 3 730 3 925

Paso 2.

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

Paso 3. Como i no es un entero, se *redondea.* La posición del percentil 85 es el siguiente entero mayor que 10.2, es decir, la posición 11.

Observe de nuevo los datos: el percentil 85 es el valor de datos en la posición 11, o 3 730.

Como otro ejemplo de este procedimiento, considere el cálculo del percentil 50 para los datos de los sueldos iniciales. Al aplicar el paso 2 se obtiene

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

Dado que i es un entero, el paso 3*b*) establece que el percentil 50 es el promedio de los valores sexto y séptimo; por tanto, el percentil 50 es (3 490 + 3 520)/2 = 3 505. Observe que el *percentil 50 coincide con la mediana.*

Cuartiles

Los cuartiles son sencillamente percentiles específicos; por tanto, los pasos para calcular los percentiles se aplican directamente en el cálculo de cuartiles.

A menudo es recomendable dividir los datos en cuatro partes, cada una de las cuales contiene aproximadamente un cuarto, o 25% de las observaciones. La figura 3.1 muestra una distribución de datos dividida en cuatro partes. Los puntos de división se conocen como **cuartiles** y son definidos como:

Q_1 = primer cuartil, o percentil 25
Q_2 = segundo cuartil, o percentil 50 (también la mediana)
Q_3 = tercer cuartil, o percentil 75

Los datos sobre los sueldos iniciales mensuales se acomodan de nuevo en orden ascendente. Ya se identificó Q_2, el segundo cuartil (mediana), como 3 505.

3 310 3 355 3 450 3 480 3 480 3 490 3 520 3 540 3 550 3 650 3 730 3 925

El cálculo de los cuartiles Q_1 y Q_3 requiere el uso de la regla para obtener los percentiles 25 y 75. Estos cálculos son los siguientes.

Para obtener Q_1,

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

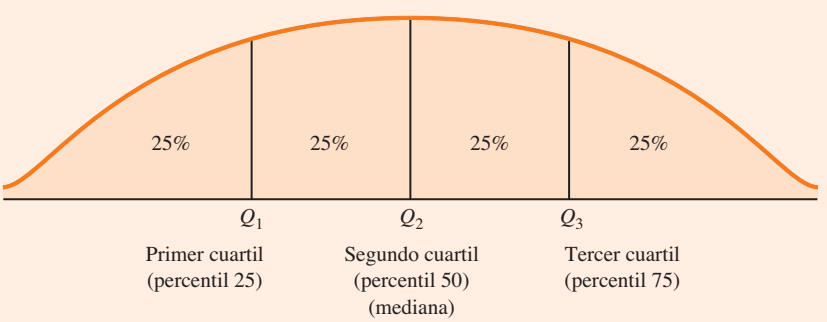
Como i es un entero, el paso 3 *b*) indica que el primer cuartil, o percentil 25, es el promedio del tercer y cuarto valores de los datos; por tanto, $Q_1 = (3\,450 + 3\,480)/2 = 3\,465$.

Para obtener Q_3,

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

Una vez más, dado que i es un entero, el paso 3*b*) indica que el tercer cuartil, o percentil 75, es el promedio del noveno y décimo valores de los datos; es decir, $Q_3 = (3\,550 + 3\,650)/2 = 3\,600$.

FIGURA 3.1 Posición de los cuartiles



Los cuartiles dividen los datos de los sueldos iniciales en cuatro partes, de las cuales cada una contiene 25% de las observaciones.

3 310 3 355 3 450 | 3 480 3 480 3 490 | 3 520 3 540 3 550 | 3 650 3 730 3 925

$Q_1 = 3\,465$ $Q_2 = 3\,505$ (mediana) $Q_3 = 3\,600$

Los cuartiles se definieron como los percentiles 25, 50 y 75; de ahí que se calculen de la misma manera que los percentiles. Sin embargo, a veces se usan otras convenciones para calcularlos, por lo que los valores reales reportados para los cuartiles pueden variar ligeramente, dependiendo de la convención utilizada. No obstante, el objetivo de todos los procedimientos es dividir los datos en cuatro partes iguales.

NOTAS Y COMENTARIOS

Cuando un conjunto de datos contiene valores extremos es preferible utilizar la mediana más que la media como medida de la ubicación central. Otra medida que se emplea a veces cuando hay valores extremos es la *media recortada.* Ésta se obtiene al eliminar un porcentaje de los valores menores y mayores de un conjunto de datos y luego calcular la media de los valores restantes. Por ejemplo, la media recortada al 5% se obtiene al eliminar 5% de los valores menores y 5% de los valores mayores de los datos y luego calcular la media de los valores restantes. Si se usa la muestra con $n = 12$ sueldos iniciales, $0.05(12) = 0.6$. El redondeo de este valor a 1 indica que la media recortada al 5% elimina el valor 1 menor y el valor 1 mayor. La media recortada al 5% utilizando las 10 observaciones restantes es 3 524.50.

Ejercicios

Métodos

1. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule la media y la mediana.
2. Asuma una muestra con los datos 10, 20, 21, 17, 16 y 12. Calcule la media y la mediana.
3. Considere una muestra con los datos 27, 25, 20, 15, 30, 34, 28 y 25. Calcule los percentiles 20, 25, 65 y 75.

AUTO evaluación

4. Considere una muestra con los datos 53, 55, 70, 58, 64, 57, 53, 69, 57, 68 y 53. Calcule la media, la mediana y la moda.

Aplicaciones

5. El índice Dow Jones de viajes informó cuánto pagan los viajeros de negocios por una noche en una habitación de hotel en las principales ciudades estadounidenses (*The Wall Street Journal,* 16 de enero de 2004). Las tarifas promedio de una habitación por noche para 20 ciudades son las siguientes:

WEB archivo

Hotels

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

a) ¿Cuál es la tarifa media de una habitación por noche?
b) ¿Cuál es la mediana de las tarifas de una habitación por noche?
c) ¿Cuál es la moda?
d) ¿Cuál es el primer cuartil?
e) ¿Cuál es el tercer cuartil?

6. Durante la temporada de basquetbol colegial de la NCAA 2007-2008 en Estados Unidos, los equipos de basquetbol varonil intentaron un número récord de tiros de 3 puntos, que promedió 19.07 tiros por partido (Associated Press Sports, 24 de enero de 2009). Al tratar de desalentar tantos tiros de 3 puntos y estimular a los estudiantes a hacer más jugadas, el comité de reglas de la NCAA movió la línea de tiro de 3 puntos de 19 pies, 9 pulgadas a 20 pies, 9 pulgadas al inicio de la temporada 2008-2009. En la tabla siguiente se aprecian los tiros de 3 puntos realizados y los encestes para una muestra de 19 partidos de basquetbol durante la temporada de referencia.

WEB archivo
3Points

3-Point Shots	Shots Made	3-Point Shots	Shots Made
23	4	17	7
20	6	19	10
17	5	22	7
18	8	25	11
13	4	15	6
16	4	10	5
8	5	11	3
19	8	25	8
28	5	23	7
21	7		

a) ¿Cuál es la media del número de tiros de 3 puntos realizados por partido?
b) ¿Cuál es la media del número de tiros de 3 puntos encestados por partido?
c) Al usar la línea de 3 puntos más cercana, los jugadores encestaban 35.2% de sus tiros. ¿Qué porcentaje de tiros encestan desde la nueva línea de 3 puntos?
d) ¿Cuál fue el impacto del cambio de reglas de la NCAA que retrocedió la línea de tiro a 20 pies, 9 pulgadas para la temporada 2008-2009? ¿Estaría usted de acuerdo con el artículo de Associated Press Sports que establece que "El retroceso de la línea de tiro de 3 puntos no ha cambiado drásticamente el juego"? Explique por qué.

7. El ingreso por donativos es una parte vital de los presupuestos anuales en los colegios y universidades. Un estudio realizado por los directivos administrativos de la Asociación Nacional de Colegios y Universidades informó que 435 instituciones encuestadas recibieron un total de $413 mil millones en donaciones. Las 10 universidades más ricas se listan a continuación (*The Wall Street Journal,* 27 de enero de 2009). Los montos se proporcionan en miles de millones de dólares.

Universidad	Donativo (miles de millones de dólares)	Universidad	Donativo (miles de millones de dólares)
Columbia	7.2	Princeton	16.4
Harvard	36.6	Stanford	17.2
MIT	10.1	Texas	16.1
Michigan	7.6	Texas A&M	6.7
Northwestern	7.2	Yale	22.9

a) ¿Cuál es la media de los donativos para estas universidades?
b) ¿Cuál es la mediana de los donativos?
c) ¿Cuál es la moda de estos apoyos?
d) Calcule el primer y el tercer cuartiles.

e) ¿Cuál es el donativo total para estas 10 universidades? Éstas representan 2.3% de los 435 colegios y universidades encuestados, ¿qué porcentaje del total de $413 mil millones en donativos recibieron?

f) *The Wall Street Journal* reportó que durante un periodo reciente de cinco meses, un declive económico ocasionó que los donativos disminuyeran 23%. ¿Cuál es la estimación en dólares de la reducción en los donativos totales que recibieron estas 10 universidades? Dada esta situación, ¿cuáles son algunos pasos que usted esperaría que los administradores universitarios tomaran en consideración?

AUTO evaluación

8. El costo de las compras que realizaron los consumidores, como vivienda unifamiliar, gasolina, servicios de Internet, declaración de impuestos y hospitalización fue difundido en un artículo de *The Wall Street Journal* (2 de enero de 2007). Los datos muestrales típicos sobre el costo de la declaración de impuestos por servicios tales como H&R Block se muestran en seguida.

WEB archivo
TaxCost

120	230	110	115	160
130	150	105	195	155
105	360	120	120	140
100	115	180	235	255

a) Calcule la media, la mediana y la moda.

b) Determine el primer y el tercer cuartiles.

c) Calcule e interprete el percentil 90.

9. Datos de la Asociación Nacional de Agentes Inmobiliarios de Estados Unidos muestran que las ventas de vivienda fueron las más bajas en 10 años (Associated Press, 24 de diciembre de 2008). A continuación se presentan los datos muestrales con el precio de venta representativo para las casas usadas y las nuevas. Los datos se expresan en miles de dólares.

Casas usadas	315.5	202.5	140.2	181.3	470.2	169.9	112.8	230.0	177.5
Casas nuevas	275.9	350.2	195.8	525.0	225.3	215.5	175.0	149.5	

a) ¿Cuál es la mediana de los precios de venta de las casas usadas?

b) ¿Cuál es la mediana de los precios de venta de las viviendas nuevas?

c) ¿Cuáles casas tienen la mediana de los precios de venta más alta: las usadas o las nuevas? ¿Cuál es la diferencia entre la mediana de los precios de venta?

d) Hace un año la mediana de los precios de venta de las casas usadas era de $208.4 mil y la de los precios de venta de las casas nuevas era de $249 mil. Calcule el cambio porcentual en la mediana de los precios de venta de unos y otros inmuebles durante un periodo de un año. ¿Cuáles viviendas tienen el cambio porcentual mayor en la mediana de los precios de venta: las usadas o las nuevas?

10. Un panel de economistas proporcionó pronósticos de la economía estadounidense para los primeros seis meses de 2007 (*The Wall Street Journal,* 2 de enero de 2007). Los cambios porcentuales en el producto interno bruto (PIB) pronosticados por 30 economistas son los siguientes.

WEB archivo
Economy

2.6	3.1	2.3	2.7	3.4	0.9	2.6	2.8	2.0	2.4
2.7	2.7	2.7	2.9	3.1	2.8	1.7	2.3	2.8	3.5
0.4	2.5	2.2	1.9	1.8	1.1	2.0	2.1	2.5	0.5

a) ¿Cuál es el pronóstico mínimo para el cambio porcentual en el PIB? ¿Cuál es el pronóstico máximo?

b) Calcule la media, la mediana y la moda.

c) Calcule el primer y el tercer cuartiles.

d) ¿Los economistas proporcionaron una perspectiva optimista o pesimita de la economía estadounidense? Comente.

11. En un experimento automotriz sobre millaje y consumo de gasolina se aplicó una prueba de circulación a 13 automóviles a lo largo de 300 millas tanto en ciudad como en autopista. Los datos siguientes se obtuvieron para el rendimiento en millas por galón.

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

Use la media, la mediana y la moda para señalar cuál es la diferencia en el rendimiento para la circulación en ciudad y en autopista.

12. Walt Disney Company compró Pixar Animation Studios, Inc. por 7 400 millones de dólares (sitio web de CNN Money, 24 de enero de 2006). Las películas animadas producidas por Disney y Pixar durante los 10 años previos a la compra se listan en la tabla siguiente. Los ingresos de taquilla (Revenue) se proporcionan en millones de dólares. Calcule el ingreso total, la media, la mediana y los cuartiles para comparar el éxito de taquilla de las películas producidas por ambas empresas. ¿Los estadísticos sugieren por lo menos una de las razones por las que Disney se interesó en comprar Pixar? Comente.

WEB archivo

Disney

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

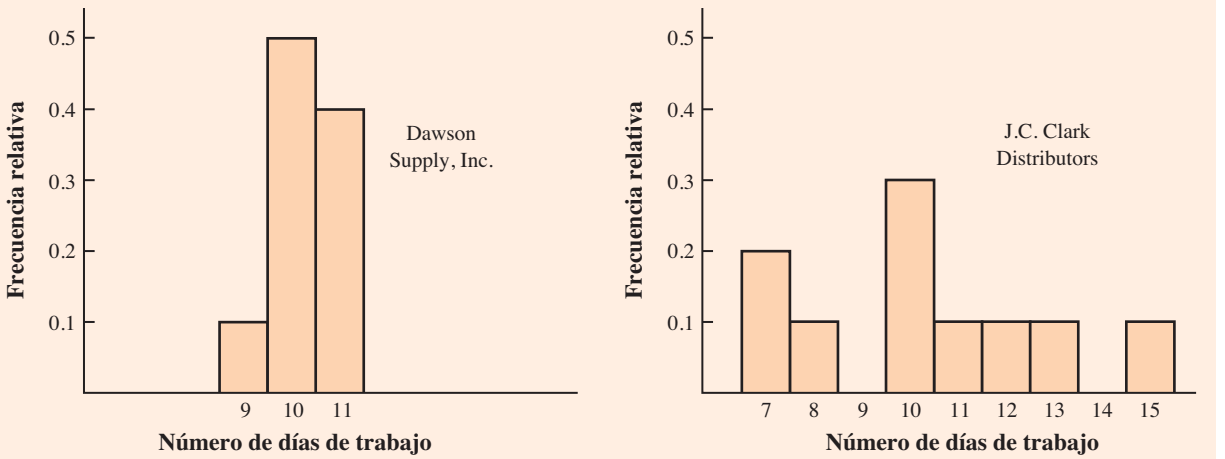
3.2 Medidas de variabilidad

La variabilidad en los plazos de entrega genera incertidumbre en la planeación de la producción. Los métodos presentados en esta sección ayudan a medir y entender la variabilidad.

Además de las medidas de posición, con frecuencia es conveniente considerar las medidas de variabilidad o dispersión. Por ejemplo, suponga que usted es un agente de compras de una empresa manufacturera grande y que coloca con regularidad pedidos con dos proveedores diferentes. Después de varios meses de operación, se da cuenta de que el número medio de días necesario para que ambos surtan los pedidos es de 10 días. Los histogramas que resumen el número de días de trabajo requeridos para que los proveedores suministren los pedidos se muestran en la figura 3.2. Aunque el número medio de días es 10 para los dos proveedores, ¿ambos muestran el mismo grado de confiabilidad en cuanto a efectuar las entregas a tiempo? Note la dispersión, o variabilidad, en los plazos de entrega indicados por los histogramas. ¿Qué proveedor prefiere usted?

Para la mayoría de las empresas es importante recibir a tiempo los materiales y suministros para sus procesos. Los plazos de entrega de 7 u 8 días mostrados para J.C. Clark Distributors podrían considerarse favorables, sin embargo, algunos plazos largos de 13 a 15 días podrían resultar desastrosos en términos de mantener ocupada a la fuerza de trabajo y la producción

FIGURE 3.2 Datos históricos que muestran el número de días requerido para surtir los pedidos



dentro de lo programado. Este ejemplo ilustra una situación en la que la variabilidad en los tiempos de entrega puede ser una consideración primordial al seleccionar a un proveedor. Para la mayoría de los agentes de compra, la menor variabilidad mostrada por Dawson Supply, Inc. lo haría el preferido.

Ahora se verá a la revisión de algunas medidas de variabilidad de uso común.

Rango

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

RANGO

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

Revise los datos sobre los sueldos iniciales para los licenciados en administración de empresas recién egresados que hemos venido trabajando de la tabla 3.1. El sueldo inicial mayor es de 3 925 y el menor es de 3 310. El rango es $3\,925 - 3\,310 = 615$.

Aun cuando el rango es la medida de variabilidad más fácil de calcular, pocas veces se usa como la única medida debido a que se basa sólo en dos de las observaciones y, por tanto, los valores extremos influyen mucho en él. Suponga que uno de los licenciados recién egresados recibe un sueldo inicial de $10 000 al mes. En este caso, el rango sería $10\,000 - 3\,310 = 6\,690$ en vez de 615. Este valor mayor para el rango no describe con claridad la variabilidad de los datos debido a que 11 de los 12 sueldos iniciales se agrupan estrechamente entre 3 310 y 3 730.

Rango intercuartílico

Una medida de la variabilidad que supera la dependencia sobre los valores extremos es el **rango intercuartílico (RIC)**. Esta medida de la variabilidad es la diferencia entre el tercer cuartil, Q_3, y el primer cuartil, Q_1. En otras palabras, el rango intercuartílico es el rango de la media de 50% de los datos.

RANGO INTERCUARTÍLICO

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

Para los datos sobre los sueldos mensuales iniciales, los cuartiles son $Q_3 = 3\,600$ y $Q_1 = 3\,465$. Por tanto, el rango intercuartílico es $3\,600 - 3\,465 = 135$.

Varianza

La **varianza** es una medida de la variabilidad que utiliza todos los datos. Se basa en la diferencia entre el valor de cada observación (x_i) y la media. La diferencia entre cada x_i y la media ($\bar{x}$ para una muestra; μ para una población) se llama *desviación respecto de la media.* Para una muestra, una desviación respecto de la media se escribe $(x_i - \bar{x})$; para una población, se escribe $(x_i - \mu)$. Si se desea calcular la varianza, las desviaciones respecto de la media *se elevan al cuadrado.*

Si los datos pertenecen a una población, el promedio de las desviaciones elevadas al cuadrado se llama *varianza poblacional*, la cual se denota por medio del símbolo griego σ^2. Para una población de N observaciones con una media poblacional μ, la definición de la varianza poblacional es la siguiente.

VARIANZA POBLACIONAL

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

En la mayoría de las aplicaciones estadísticas, los datos que se analizan provienen de una muestra. Cuando se calcula una varianza muestral, a menudo lo que interesa es usarla para estimar la varianza poblacional σ^2. Aunque una explicación detallada está más allá del alcance de este libro, puede mostrarse que si la suma de las desviaciones respecto de la media al cuadrado se divide entre $n - 1$, y no entre n, la varianza muestral resultante proporciona un estimador insesgado de la varianza poblacional. Por esta razón, la *varianza muestral,* denotada por s^2, se define como sigue.

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

VARIANZA MUESTRAL

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

Para ilustrar el cálculo de la varianza muestral se usarán los datos sobre los tamaños de grupo de la muestra de cinco grupos de estudiantes universitarios presentada en la sección 3.1. Un resumen de los datos, que incluye el cálculo de las desviaciones respecto de la media y los cuadrados de las desviaciones respecto de la media, se aprecia en la tabla 3.2. La suma de los cuadrados de estas desviaciones es $\Sigma(x_i - \bar{x})^2 = 256$. Por ende, si $n - 1 = 4$, la varianza muestral es

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

Antes de proseguir, observe que las unidades asociadas con la varianza muestral suelen causar confusión. Debido a que los valores que se suman para calcular la varianza, $(x_i - \bar{x})^2$, están elevados al cuadrado, las unidades asociadas con la varianza muestral también están *elevadas*

TABLA 3.2 Cálculo de desviaciones respecto de la media y desviaciones cuadradas respecto de la media de los datos de tamaños de grupo

Número de estudiantes en el grupo (x_i)	Tamaño de grupo medio ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Desviación cuadrada respecto de la media $(x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

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

al cuadrado. Por ejemplo, la varianza muestral para los datos del tamaño de grupo es $s^2 = 64$ (estudiantes)^2. Las unidades cuadradas asociadas con la varianza dificultan obtener una comprensión e interpretación intuitiva del valor numérico de ésta. Se recomienda considerarla como una medida útil en la comparación de la cantidad de variabilidad para dos o más variables. En una comparación de las variables, aquella con la varianza más grande muestra la mayor variabilidad. Una interpretación del valor de la varianza tal vez no sea necesaria.

Como otra ilustración del cálculo de una varianza muestral, considere los sueldos iniciales listados en la tabla 3.1 para los 12 licenciados en administración de empresas. En la sección 3.1 se observa que la media muestral de los sueldos es de 3 540. El cálculo de la varianza muestral ($s^2 = 27\,440.91$) se muestra en la tabla 3.3.

TABLA 3.3 Cálculo de la varianza muestral para los datos de los sueldos iniciales

Sueldo mensual (x_i)	Media muestral ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Desviación cuadrada respecto de la media $(x_i - \bar{x})^2$
3 450	3 540	−90	8 100
3 550	3 540	10	100
3 650	3 540	110	12 100
3 480	3 540	−60	3 600
3 355	3 540	−185	34 225
3 310	3 540	−230	52 900
3 490	3 540	−50	2 500
3 730	3 540	190	36 100
3 540	3 540	0	0
3 925	3 540	385	148 225
3 520	3 540	−20	400
3 480	3 540	−60	3 600
		0	301 850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

Usando la ecuación (3.5),

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

En las tablas 3.2 y 3.3 se aprecian la suma de las desviaciones sobre la media y la suma de las desviaciones cuadradas sobre la media. Para cualquier conjunto de datos, la suma de las desviaciones sobre la media *siempre será igual a cero.* Note que en esas tablas, $\Sigma(x_i - \bar{x}) = 0$. Las desviaciones positivas y negativas se cancelan entre sí, ocasionando que la suma de las desviaciones sobre la media sea igual a cero.

Desviación estándar

La **desviación estándar** se define como la raíz cuadrada positiva de la varianza. Siguiendo la notación que se adoptó para las varianzas muestral y poblacional, se usa s para denotar la desviación estándar muestral y σ para denotar la desviación estándar poblacional. La desviación estándar se deriva de la varianza de la manera siguiente.

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

DESVIACIÓN ESTÁNDAR

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

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

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

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

¿Qué se gana al convertir la varianza en la desviación estándar correspondiente? Recuerde que las unidades asociadas con la varianza están elevadas al cuadrado. Por ejemplo, la varianza muestral para los datos sobre los sueldos iniciales de los licenciados en administración de empresas recién egresados es $s^2 = 27\,440.91$ (dólares)^2. Debido a que la desviación estándar es la raíz cuadrada de la varianza, las unidades de esta última, los dólares al cuadrado, se convierten en dólares en la desviación estándar. Por consiguiente, la desviación estándar de los datos de los sueldos iniciales es $165.65. En otras palabras, ésta se mide en las mismas unidades que los datos originales; por esta razón la desviación estándar se compara más fácilmente con la media y con otros estadísticos que se miden en las mismas unidades que los datos originales.

Coeficiente de variación

En algunas situaciones nos interesa la estadística descriptiva que indique qué tan grande es la desviación estándar con respecto a la media. Esta medida se llama **coeficiente de variación**, y se expresa por lo general como un porcentaje.

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

COEFICIENTE DE VARIACIÓN

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

Para los datos de los tamaños de grupo, se encontró una media muestral de 44 y una desviación estándar muestral de 8. El coeficiente de variación es $[(8/44) \times 100]\% = 18.2\%$. Expresado con palabras, el coeficiente de variación indica que la desviación estándar muestral es 18.2% del valor de la media muestral. Para los datos de los sueldos iniciales con una media muestral de 3 540 y una desviación estándar muestral de 165.65, el coeficiente de variación, $[(165.65/3\,540) \times 100]\% = 4.7\%$, señala que la desviación estándar muestral es sólo 4.7% del valor de la media muestral. En general, el coeficiente de variación es un estadístico útil para comparar la variabilidad de las variables que tienen tanto desviaciones estándar como medias distintas.

NOTAS Y COMENTARIOS

1. El *software* y las hojas de cálculo para estadística se usan para obtener los estadísticos descriptivos presentados en este capítulo. Una vez que los datos se introducen en una hoja de cálculo, bastan unos comandos sencillos para generar el resultado deseado. En los tres apéndices del capítulo se explica cómo usar Minitab, Excel y StatTools para obtener estadísticos descriptivos.
2. La desviación estándar es una medida de uso común para el riesgo asociado con la inversión en acciones y fondos de acciones (*BusinessWeek,* 17 de enero de 2000). Proporciona una medida de cómo fluctúan los rendimientos mensuales en torno al rendimiento medio a largo plazo.
3. Cuando los valores de la media muestral $\bar{x}$ y los valores de los cuadrados de las desviaciones $(x_i - \bar{x})^2$ se redondean, se pueden introducir errores en la calculadora al obtener la varianza y la desviación estándar. Para reducir los errores de redondeo, se recomienda trabajar por lo menos con seis dígitos significativos durante los cálculos intermedios. La varianza o la desviación estándar resultantes pueden redondearse después a menos dígitos.
4. Una fórmula opcional para el cálculo de la varianza muestral es

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

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

Ejercicios

Métodos

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

AUTO evaluación

Aplicaciones

AUTO evaluación

16. Las puntuaciones que obtuvo un jugador de boliche en seis partidos fueron 182, 168, 184, 190, 170 y 174. Usando estos datos como una muestra, calcule los estadísticos descriptivos siguientes:
    a) Rango
    b) Varianza
    c) Desviación estándar
    d) Coeficiente de variación
17. Un sistema de teatro en casa (*home theater*) es la manera más fácil y económica de proporcionar sonido ambiental para un centro de entretenimiento en el hogar. Enseguida se presenta una muestra de precios (*Consumer Reports Buying Guide*, 2004) para modelos con y sin reproductor de DVD.

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

    a) Calcule el precio medio de los modelos con reproductor de DVD y el precio medio de los modelos sin reproductor de DVD. ¿Cuál es el precio adicional que se paga por tener un reproductor de DVD en el sistema de teatro en casa?
    b) Calcule el rango, la varianza y la desviación estándar de las dos muestras. ¿Qué le dice esta información sobre los precios de los modelos con y sin reproductor de DVD?

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

Ciudad	Tarifa diaria
Boston	$43
Atlanta	35
Miami	34
Nueva York	58
Orlando	30
Pittsburgh	30
Washington, D.C.	36

a) Calcule la media, la varianza y la desviación estándar de estas tarifas.
b) En una muestra similar de siete ciudades del oeste de Estados Unidos se obtuvo una media muestral de las tarifas de renta de automóviles de $38 por día. La varianza y la desviación estándar fueron 12.3 y 3.5, respectivamente. Comente la diferencia entre las tarifas de renta de las ciudades del este y del oeste de Estados Unidos.

19. *Los Ángeles Times* informa el índice de calidad del aire de varias zonas del sur de California. Una muestra de valores de este índice en Pomona proporcionó los datos siguientes: 28, 42, 58, 48, 45, 55, 60, 49 y 50.
a) Calcule el rango y el rango intercuartílico.
b) Calcule la varianza muestral y la desviación estándar muestral.
c) Una muestra de lecturas del índice de calidad del aire de Anaheim proporcionó una media muestral de 48.5, una varianza muestral de 136 y una desviación estándar muestral de 11.66. ¿Qué comparaciones puede hacer entre la calidad del aire en Pomona y en Anaheim sobre la base de estos estadísticos descriptivos?

20. Los datos siguientes se utilizaron para elaborar los histogramas del número de días requerido para que Dawson Supply, Inc. y J.C. Clark Distributors surtan pedidos (figura 3.2).

Días de entrega de Dawson Supply	11	10	9	10	11	11	10	11	10	10
Días de entrega de Clark Distributors	8	10	13	7	10	11	10	7	15	12

Use el rango y la desviación estándar para apoyar la observación anterior de que Dawson Supply proporciona los tiempos de entrega más consistentes y confiables.

21. ¿Cómo se comparan los costos de abarrotes en Estados Unidos? Usando una canasta básica que contiene 10 artículos que incluyen carne, leche, pan, huevos, café, papas, cereal y jugo de naranja, la revista *Where to Retire* calculó el costo de la canasta básica en seis ciudades y seis comunidades de jubilados en todo Estados Unidos (*Where to Retire,* noviembre/diciembre de 2003). Los datos con el costo de la canasta básica al dólar más cercano son los siguientes.

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

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

WEB archivo
BackToSchool

22. La Federación Nacional de Minoristas informó que los estudiantes universitarios de primer año gastan más en artículos de regreso a clases que cualquier otro grupo universitario (*USA Today,* 4 de agosto de 2006). El archivo BackToSchool contiene una base de datos muestrales que compara los gastos de regreso a clases de 25 estudiantes de primer año y 20 del último año.
   a) ¿Cuál es el gasto medio de regreso a clases de cada grupo? ¿Los datos son consistentes con el informe de la Federación Nacional de Minoristas?
   b) ¿Cuál es el rango de los gastos de cada grupo?
   c) ¿Cuál es el rango intercuartílico para cada grupo?
   d) ¿Cuál es la desviación estándar de los gastos de cada grupo?
   e) ¿Qué gastos de regreso a clases muestran más variación: los de los estudiantes de primer año o los de los universitarios de último año?

23. Las puntuaciones anotadas por un golfista amateur en el campo de golf de Bonita Fairways, en Bonita Springs, Florida, durante 2005 y 2006 son los siguientes.

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

   a) Use la media y la desviación estándar para evaluar el desempeño del golfista durante el periodo de dos años.
   b) ¿Cuál es la principal diferencia en su desempeño entre 2005 y 2006? ¿Qué mejora, si la hay, puede verse en las puntuaciones de 2006?

24. Los corredores de un equipo de atletismo universitario registraron los siguientes tiempos para los carreras de cuarto de milla y de milla (los tiempos están en minutos).

Tiempos de cuarto de milla	0.92	0.98	1.04	0.90	0.99
Tiempos de milla	4.52	4.35	4.60	4.70	4.50

   Después de ver esta muestra de tiempos, uno de los entrenadores comentó que los corredores de cuarto de milla registraron tiempos más consistentes. Utilice la desviación estándar y el coeficiente de variación para resumir la variabilidad de los datos. ¿El uso del coeficiente de variación indica que el comentario del entrenador es correcto?

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

Se han descrito varias medidas de ubicación y variabilidad para los datos. Además de éstas, es importante tener una medida de la forma de la distribución. En el capítulo 2 se vio que un histograma proporciona una representación gráfica de la forma de una distribución. Una medida numérica importante de la forma de una distribución es el **sesgo**.

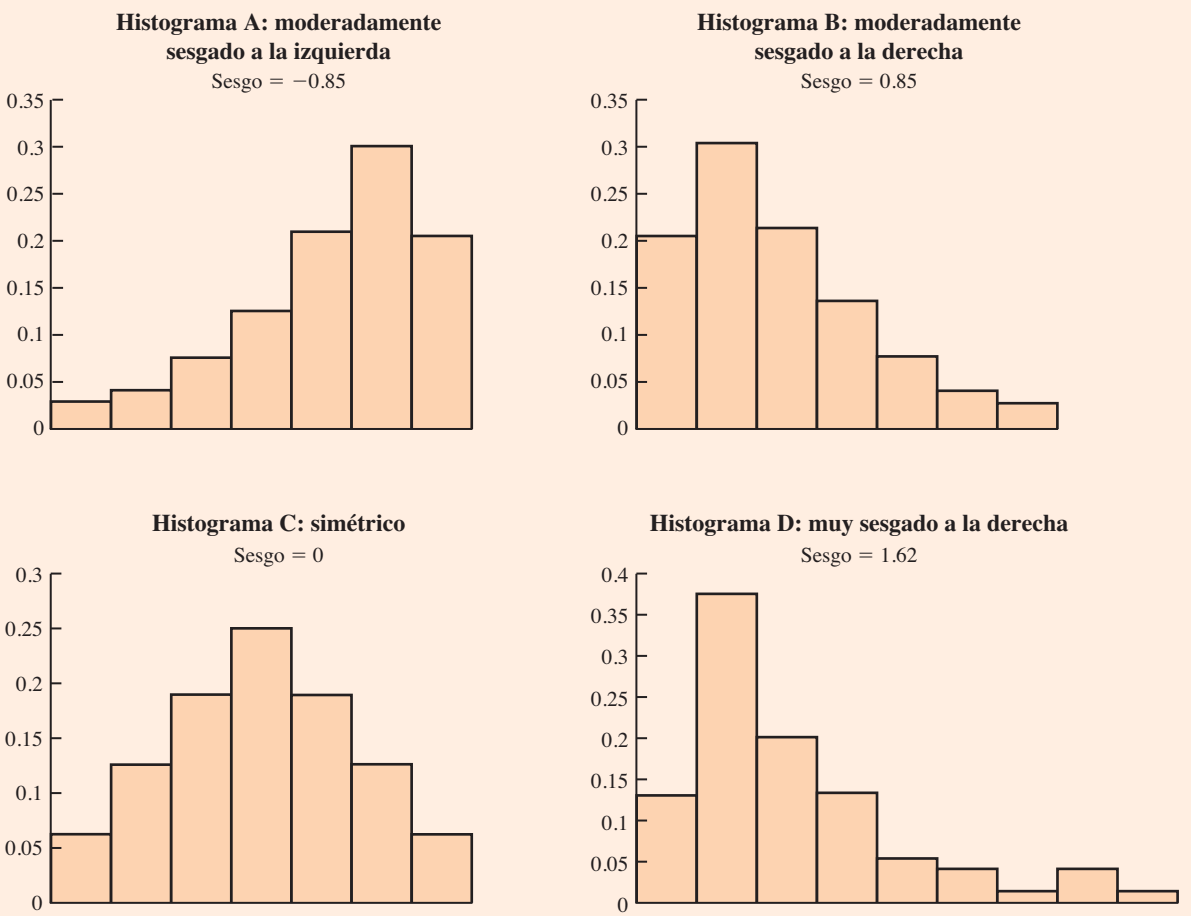
Forma de la distribución

En la figura 3.3 aparecen cuatro histogramas elaborados a partir de distribuciones de frecuencia relativa. Los histogramas A y B están moderadamente sesgados. El A está sesgado a la izquierda; su sesgo es −0.85. El B está sesgado a la derecha; su sesgo es +0.85. El histograma C es simétrico; su sesgo es cero. El D está muy sesgado a la derecha; su sesgo es 1.62. La fórmula empleada para calcular el sesgo es un tanto compleja.[1] Sin embargo, éste se obtiene fácilmente

[1] La fórmula para el sesgo de datos muestrales es:

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

FIGURA 3.3 Histogramas que muestran el sesgo de cuatro distribuciones



utilizando software para estadística. Para datos sesgados a la izquierda, el sesgo es negativo; para datos sesgados a la derecha, el sesgo es positivo. Si los datos son simétricos, el sesgo es cero.

En una distribución simétrica, la media y la mediana son iguales. Cuando los datos están sesgados positivamente, la media por lo general será mayor que la mediana; cuando están sesgados negativamente, la media será menor que la mediana. Los datos usados para construir el histograma D corresponden a las compras de los clientes de una tienda de ropa femenina. El monto medio de compra es $77.60 y la mediana del monto de compra es $59.70. Los pocos montos de compra grandes tienden a incrementar la media, mientras que a la mediana no le afectan. Cuando los datos están muy sesgados, se prefiere la mediana como medida de ubicación.

Valor *z*

Además de las medidas de posición, variabilidad y forma, también interesa la posición relativa de los valores dentro de un conjunto de datos. Las medidas de posición relativa ayudan a determinar a qué distancia de la media está un valor determinado.

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

denotados por $x_1, x_2, \ldots, x_n$. Asimismo, suponga que la media muestral, $\bar{x}$, y la desviación estándar muestral, s, ya se calcularon. Asociado con cada valor, x_i, hay otro valor llamado **valor z**. La ecuación (3.9) muestra cómo se calcula la puntuación z para cada x_i.

VALOR z

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

Donde

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

El valor z se llama *valor estandarizado*. El valor z, z_i, puede interpretarse como el *número de desviaciones estándar que x_i se encuentra de la media $\bar{x}$*. Por ejemplo, $z_1 = 1.2$ indicaría que x_1 es 1.2 desviaciones estándar mayor que la media muestral. De modo parecido, $z_2 = -0.5$ indicaría que x_2 es 0.5, o 1/2 desviaciones estándar menor que la media muestral. Un valor z mayor que cero ocurre para observaciones con un valor mayor que la media, y un valor z menor que cero ocurre para observaciones con un valor menor que la media. Un valor z de cero indica que el valor de la observación es igual a la media.

El valor z para cualquier observación puede interpretarse como una medida de la posición relativa de la observación en un conjunto de datos. Por tanto, se dice que las observaciones de dos conjuntos de datos diferentes con el mismo valor z tienen la misma posición relativa en términos de que presentan igual número de desviaciones estándar de la media.

Los valores z para los datos de los tamaños de grupo se calculan en la tabla 3.4. Recuerde la media muestral previamente calculada, $\bar{x} = 44$, y la desviación estándar muestral, $s = 8$. El valor z de -1.50 de la quinta observación indica que ésta es la más alejada de la media: está 1.50 desviaciones estándar por debajo de la media.

Teorema de Chebyshev

El **teorema de Chebyshev** permite hacer afirmaciones acerca de la proporción de los valores de datos que deben estar dentro de un número específico de desviaciones estándar de la media.

TABLA 3.4 Valores z de los datos de tamaños de grupo

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

TEOREMA DE CHEBYSHEV

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

A continuación se mencionan algunas implicaciones de este teorema cuando $z = 2$, 3 y 4 desviaciones estándar.

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

Como ejemplo del uso del teorema de Chebyshev, suponga que las calificaciones obtenidas en los exámenes parciales por 100 estudiantes universitarios en un curso de estadística para negocios tenían una media de 70 y una desviación estándar de 5. ¿Cuántos alumnos obtuvieron una calificación de entre 60 y 80 en los exámenes? ¿Cuántos obtuvieron calificaciones de entre 58 y 82?

Para calificaciones entre 60 y 80, observe que 60 está dos desviaciones estándar por abajo de la media, y 80 está dos desviaciones estándar por encima de la media. Usando el teorema de Chebyshev se ve que como mínimo 0.75, o por lo menos 75% de las observaciones debe tener valores dentro de dos desviaciones estándar de la media. Por tanto, 75% de los estudiantes como mínimo debió obtener una calificación de entre 60 y 80.

En el teorema de Chebyshev se requiere $z > 1$; pero no es necesario que z sea un número entero.

Si las calificaciones de los exámenes están entre 58 y 82, observe que $(58 - 70)/5 = -2.4$ indica que 58 está a 2.4 desviaciones estándar por debajo de la media y que $(82 - 70)/5 = +2.4$ indica que 82 está a 2.4 desviaciones estándar por encima de la media. Al aplicar el teorema de Chebyshev con $z = 2.4$, tenemos

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

Al menos 82.6% de los estudiantes debe obtener calificaciones de entre 58 y 82 en los exámenes.

Regla empírica

La regla empírica se basa en la distribución de probabilidad normal, la cual se estudia en el capítulo 6. La distribución normal se utiliza ampliamente en todo el libro.

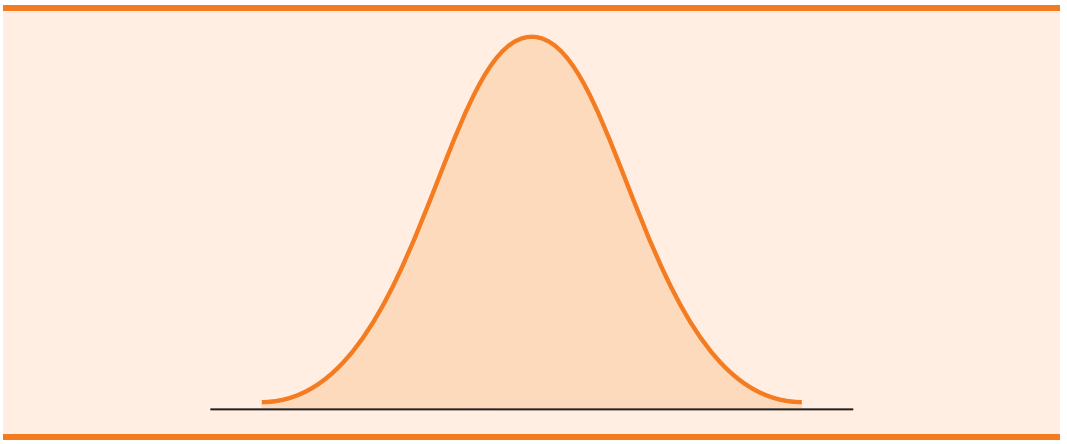
Una de las ventajas del teorema de Chebyshev estriba en que se aplica a cualquier conjunto de datos sin importar su forma de distribución. De hecho, podría usarse con cualquiera de las distribuciones de la figura 3.3. Sin embargo, en muchas aplicaciones prácticas los conjuntos de datos exhiben una distribución simétrica con forma de pila o de campana, como se aprecia en la figura 3.4. Cuando se piensa que los datos se aproximan a esta distribución, la **regla empírica** se usa para determinar el porcentaje de valores de datos que deben estar dentro de un número específico de desviaciones estándar de la media.

REGLA EMPÍRICA

Cuando los datos tienen una distribución en forma de campana:

- Aproximadamente 68% de los valores de datos estará dentro de una desviación estándar de la media.
- Aproximadamente 95% de los valores de datos estará dentro de dos desviaciones estándar de la media.
- Casi todos los valores de datos deben estar dentro de tres desviaciones estándar de la media.

FIGURA 3.4 Distribución simétrica con forma de pila o de campana



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

- Aproximadamente 68% de los envases llenos pesará entre 15.75 y 16.25 onzas (dentro de una desviación estándar de la media).
- Aproximadamente 95% de los envases llenos pesará entre 15.50 y 16.50 onzas (dentro de dos desviaciones estándar de la media).
- Casi todos los envases llenos pesarán entre 15.25 y 16.75 onzas (dentro de tres desviaciones estándar de la media).

Detección de observaciones atípicas

Un conjunto de datos a veces tiene una o más observaciones con valores inusualmente grandes o sumamente pequeños. Estos valores extremos se llaman **observaciones atípicas**. Los expertos en estadística experimentados emprenden acciones para identificar observaciones atípicas y luego revisan cada una con detalle. Una observación atípica suele ser un valor de datos que se registró incorrectamente; si esto ocurre, el error se corrige antes de un análisis posterior. También puede ser una observación que se introdujo de forma incorrecta en el conjunto de datos; si este es el caso, se elimina. Por último, puede consistir en un valor de datos inusual que se registró correctamente y pertenece al conjunto de datos. En tal caso, debe conservarse.

Es una buena idea buscar observaciones atípicas antes de tomar decisiones basadas en el análisis de datos. Suelen cometerse errores en el registro y la introducción de los datos en la computadora. Las observaciones atípicas no necesariamente tienen que eliminarse, pero debe verificarse qué tan exactas y apropiadas son.

Los valores estandarizados (puntuaciones z), se utilizan para identificar observaciones atípicas. Recuerde que la regla empírica permite concluir que cuando los datos tienen una distribución en forma de campana, casi todos los valores de datos están dentro de tres desviaciones estándar de la media. Por tanto, al usar puntuaciones z para identificar observaciones extremas, se recomienda tomar en cuenta como una observación atípica cualquier valor de datos con una puntuación z menor que -3 o mayor que $+3$. La exactitud de estos valores debe verificarse y determinar si pertenecen al conjunto de datos.

Vuelva a observar las puntuaciones z de los datos sobre los tamaños de grupo de la tabla 3.4. La puntuación z de -1.50 muestra que el tamaño del quinto grupo está más alejado de la media. Sin embargo, este valor estandarizado está dentro de los límites de -3 y $+3$ para las observaciones atípicas. Por esta razón, la puntuación z no indica que las observaciones atípicas estén presentes en los datos de los tamaños de clase.

NOTAS Y COMENTARIOS

1. El teorema de Chebyshev es aplicable a cualquier conjunto de datos y se utiliza para establecer el número mínimo de valores de datos que estará dentro de cierto número de desviaciones estándar de la media. Si se sabe que los datos tienen una forma aproximada de campana, se puede decir más.

Por ejemplo, la regla empírica permite afirmar que *aproximadamente* 95% de los valores de datos estará dentro de dos desviaciones estándar de la media; el teorema de Chebyshev sólo permite concluir que por lo menos 75% de estos valores estará dentro de ese intervalo.

2. Antes de analizar un conjunto de datos, los expertos en estadística efectúan varias revisiones para confirmar su validez. En un estudio grande no es raro que se cometan errores en el registro de los valores de datos o al introducirlos en una computadora. La identificación de las observaciones es una herramienta empleada para verificar la validez de los datos.

Ejercicios

Métodos

25. Considere una muestra con los datos 10, 20, 12, 17 y 16. Calcule el valor z de cada una de estas cinco observaciones.

26. Suponga una muestra con una media de 500 y una desviación estándar de 100. ¿Cuáles son los valores z de los datos siguientes: 520, 650, 500, 450 y 280?

AUTO evaluación

27. Considere una muestra con una media de 30 y una desviación estándar de 5. Utilice el teorema de Chebyshev para determinar el porcentaje de datos que se encuentra dentro de cada uno de los rangos siguientes.
    a) 20 a 40
    b) 15 a 45
    c) 22 a 38
    d) 18 a 42
    e) 12 a 48

28. Suponga que los datos tienen una distribución con forma de campana, una media de 30 y una desviación estándar de 5. Use la regla empírica para determinar el porcentaje de los datos que está dentro de cada uno de los rangos siguientes.
    a) 20 a 40
    b) 15 a 45
    c) 25 a 35

Aplicaciones

AUTO evaluación

29. Los resultados de una encuesta nacional revelaron que, en promedio, los adultos duermen 6.9 horas por noche. Imagine que la desviación estándar es de 1.2 horas.
    a) Use el teorema de Chebyshev para calcular el porcentaje de personas que duermen entre 4.5 y 9.3 horas.
    b) Con el teorema de Chebyshev calcule ahora el porcentaje que duerme entre 3.9 y 9.9 horas.
    c) Suponga que el número de horas de sueño sigue una distribución con forma de campana. Utilice la regla empírica para calcular el porcentaje de personas que duerme entre 4.5 y 9.3 horas por día. ¿Cómo se compara este resultado con el valor obtenido con el teorema de Chebyshev en el inciso *a*)?

30. La Oficina de Información Energética reportó que el precio medio por galón de gasolina de grado regular es de $2.05 (Energy Information Administration, mayo de 2009). Suponga que la desviación estándar es $0.10 y que el precio al detalle (o al menudeo) por galón tiene una distribución con forma de campana.
    a) ¿Qué porcentaje de gasolina de grado regular se vendió entre $1.95 y $2.15 por galón?
    b) ¿Qué porcentaje se vendió entre $1.95 y $2.25 por galón?
    c) ¿Qué porcentaje de gasolina de grado regular se vendió por más de $2.25 por galón?

31. El promedio nacional para la sección de matemáticas del examen de aptitudes escolares (College Board's Scholastic Aptitude Test, SAT) es 515 (*The World Almanac,* 2009). El Consejo Universitario vuelve a escalar en forma periódica las calificaciones del examen de tal manera que la desviación estándar sea aproximadamente 100. Responda las preguntas siguientes usando una distribución con forma de campana y la regla empírica para las calificaciones del examen verbal.

a) ¿Qué porcentaje de estudiantes obtuvo una calificación en el SAT verbal mayor que 615?
b) ¿Qué porcentaje obtuvo una calificación en el SAT verbal mayor que 715?
c) ¿Qué porcentaje de alumnos logró una calificación entre 415 y 515?
d) ¿Qué porcentaje obtuvo una calificación entre 315 y 615?

32. Los altos costos del mercado de bienes raíces en California han ocasionado que las familias que no pueden darse el lujo de comprar casas más grandes consideren los cobertizos de los patios traseros como una opción de ampliación. Muchos están usando las estructuras de sus patios para construir sus estudios, salas de arte y áreas de pasatiempos, así como para almacenamiento adicional. El precio medio de una estructura de tablillas de madera para patio trasero hecha a la medida es de $3 100 (*Newsweek,* 29 de septiembre de 2003). Suponga que la desviación estándar es $1 200.
a) ¿Cuál es el valor *z* para una estructura de patio trasero que cuesta $2 300?
b) ¿Cuál es el valor *z* para una estructura que cuesta $4 900?
c) Interprete los valores *z* en los incisos *a*) y *b*). Comente si alguna debe considerarse una observación atípica.
d) El artículo de *Newsweek* describió una combinación de oficina en el cobertizo del patio trasero construida con $13 000 en Albany, California. ¿Esta estructura debe considerarse una observación atípica? Explique por qué.

33. Florida Power & Light (FP&L) Company ha gozado de la reputación de reparar rápidamente un sistema eléctrico después de las tormentas. Sin embargo, durante las temporadas de huracanes de 2004 y 2005 la realidad fue otra: el método comprobado de la empresa para las reparaciones de emergencia ya no fue lo suficientemente bueno (*The Wall Street Journal,* 16 de enero de 2006). Los datos siguientes muestran los días requeridos para restablecer el servicio eléctrico después de siete huracanes durante los años de referencia.

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

Con base en esta muestra de siete huracanes, calcule los estadísticos descriptivos siguientes.
a) Media, mediana y moda.
b) Rango y desviación estándar.
c) ¿*Vilma* debe considerarse una observación atípica en términos de los días requeridos para restablecer el servicio eléctrico?
d) Los siete huracanes ocasionaron 10 millones de interrupciones en el servicio a los clientes. ¿Los estadísticos indican que FP&L debe considerar la necesidad de mejorar su método de reparaciones del sistema eléctrico? Comente.

34. Una muestra de puntuaciones de 10 partidos de basquetbol colegial de la NCAA proporcionó los datos siguientes (*USA Today,* 26 de enero de 2004).

WEB archivo
NCAA

Winning Team	Points	Losing Team	Points	Winning Margin
Arizona	90	Oregon	66	24
Duke	85	Georgetown	66	19
Florida State	75	Wake Forest	70	5
Kansas	78	Colorado	57	21
Kentucky	71	Notre Dame	63	8
Louisville	65	Tennessee	62	3
Oklahoma State	72	Texas	66	6

Winning Team	Points	Losing Team	Points	Winning Margin
Purdue	76	Michigan State	70	6
Stanford	77	Southern Cal	67	10
Wisconsin	76	Illinois	56	20

a) Calcule la media y la desviación estándar de los puntos anotados por el equipo ganador.
b) Suponga que los puntos anotados por los equipos triunfadores en todos los partidos de la NCAA siguen una distribución con forma de campana. Utilizando la media y la desviación estándar obtenidas en el inciso *a*), estime el porcentaje de los partidos de la NCAA en los cuales el equipo ganador anota 84 puntos o más. Calcule el porcentaje de los partidos de la NCAA en los cuales el equipo triunfador anota más de 90 puntos.
c) Calcule la media y la desviación estándar del margen de victoria. ¿Los datos contienen observaciones atípicas? Explique por qué.

35. *Consumer Reports* publica reseñas y calificaciones de una variedad de productos en su sitio web. A continuación se presenta una muestra de 20 sistemas de bocinas y sus calificaciones, las cuales varían en una escala de 1 a 5, en la que 5 es la mejor.

WEB archivo
Speakers

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

a) Calcule la media y la mediana.
b) Estime el primer y el tercer cuartiles.
c) Calcule la desviación estándar.
d) El sesgo de estos datos es −1.67. Comente la forma de la distribución.
e) ¿Cuáles son las puntuaciones z asociadas con Allison One y Omni Audio?
f) ¿Los datos contienen observaciones atípicas? Explique.

3.4 Análisis exploratorio de datos

En el capítulo 2 se introdujo el diagrama de tallo y hoja como una técnica de análisis exploratorio de datos. Recuerde que dicho análisis permite usar operaciones aritméticas simples y representaciones gráficas fáciles de dibujar para resumir los datos. En esta sección continúa el análisis exploratorio de datos considerando resúmenes de cinco números y diagramas de caja.

Resumen de cinco números

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

1. Valor menor
2. Primer cuartil (Q_1)
3. Mediana (Q_2)
4. Tercer cuartil (Q_3)
5. Valor mayor

La manera más fácil de elaborar un resumen de cinco números es colocar primero los datos en orden ascendente. Una vez hecho esto es fácil identificar el valor menor, los tres cuartiles y el valor mayor. Los sueldos mensuales de inicio mostrados en la tabla 3.1 para la muestra de 12 licenciados en administración de empresas recién egresados se repiten aquí en orden ascendente.

3 310 3 355 3 450 | 3 480 3 480 3 490 | 3 520 3 540 3 550 | 3 650 3 730 3 925

$Q_1 = 3\,465$ $Q_2 = 3\,505$ (mediana) $Q_3 = 3\,600$

La mediana de 3 505 y los cuartiles $Q_1 = 3\,465$ y $Q_3 = 3\,600$ se calcularon en la sección 3.1. Al revisar los datos se observa un valor menor de 3 310 y un valor mayor de 3 925. Por tanto, el resumen de cinco números para los datos de los sueldos iniciales es 3 310, 3 465, 3 505, 3 600 y 3 925. Entre los números adyacentes de un resumen de cinco números se encuentra aproximadamente un cuarto, o 25%, de las observaciones.

Diagrama de caja

Un **diagrama de caja** es un resumen gráfico de los datos basado en un resumen de cinco números. La clave para elaborar de un diagrama de caja es el cálculo de la mediana y los cuartiles Q_1 y Q_3. El rango intercuartílico, RIC $= Q_3 - Q_1$, también se utiliza. En la figura 3.5 se aprecia el diagrama de cuadro de los datos de los sueldos mensuales iniciales. Los pasos que se siguen para elaborarlo se presentan a continuación.

1. Se traza una caja con sus extremos ubicados en el primer y tercer cuartiles. Para los datos de los sueldos iniciales, $Q_1 = 3\,465$ y $Q_3 = 3\,600$. Este cuadro contiene la mitad, 50%, de los datos.
2. Se traza una línea vertical en el cuadro donde se ubica la mediana (3 505 para los datos de los sueldos iniciales).
3. Al usar el rango intercuartílico, RIC $= Q_3 - Q_1$, se localizan los *límites*. Para el diagrama de caja los límites son 1.5(RIC) por debajo de Q_1 y 1.5(RIC) por encima de Q_3. Para los datos de los sueldos, RIC $= Q_3 - Q_1 = 3\,600 - 3\,465 = 135$. Por tanto, los límites son $3\,465 - 1.5(135) = 3\,262.5$ y $3\,600 + 1.5(135) = 3\,802.5$. Los datos fuera de estos límites se consideran *observaciones atípicas.*
4. Las líneas punteadas de la figura 3.5 se llaman *bigotes.* Éstos se trazan desde los extremos de la caja hasta los valores menor y mayor *dentro de los límites* calculados en el paso 3. Por tanto, los bigotes terminan en los valores de los sueldos de 3 310 y 3 730.
5. Por último, la ubicación de cada observación atípica se señala con un asterisco (símbolo *). En la figura 3.5 se aprecia una observación, 3 925.

Los diagramas de caja proporcionan otra manera de identificar observaciones atípicas. Sin embargo, no necesariamente identifican los mismos valores que aquellos con una puntuación z menor que −3 o mayor que +3. Cualquiera de los dos procedimientos o ambos pueden usarse.

En la figura 3.5 se trazaron líneas que ilustran la posición de los límites superior e inferior, cómo se calculan los límites y dónde se ubican. Aunque los límites siempre se calculan, no se trazan

FIGURA 3.5 Diagrama de caja de los datos de los sueldos iniciales con líneas que muestran los límites superior e inferior

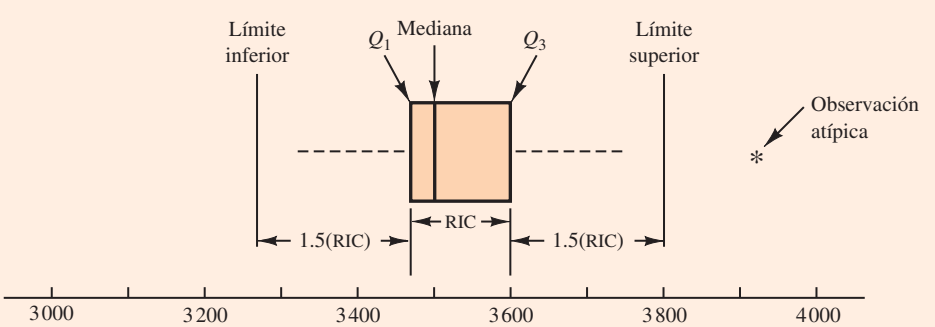
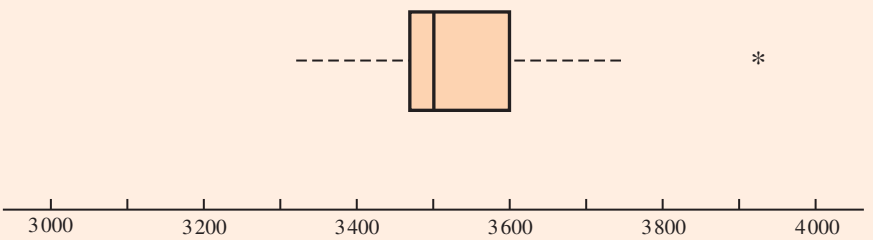


FIGURA 3.6 Diagrama de caja de los datos de los sueldos mensuales iniciales



por lo general en los diagramas de caja. La figura 3.6 muestra la apariencia usual de este tipo de diagrama para los datos de los sueldos.

WEB archivo
MajorSalary

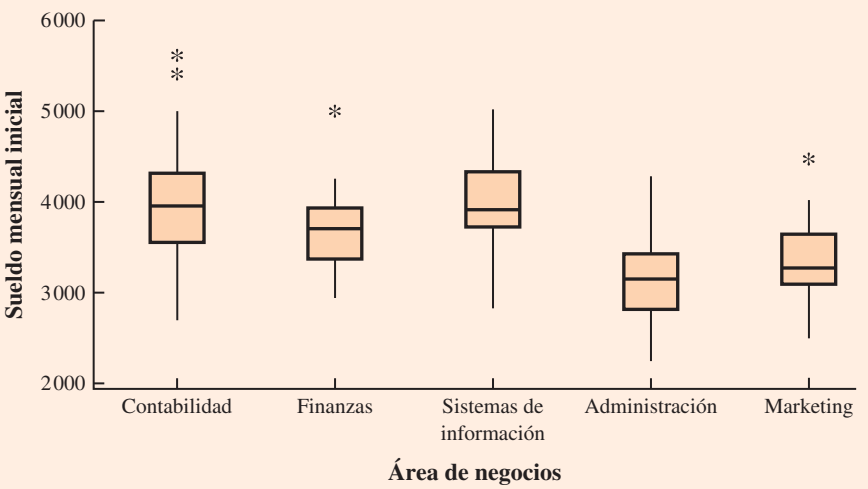
Con la finalidad de comparar los sueldos mensuales iniciales de los licenciados en administración de empresas por área de especialización, se seleccionó una muestra de 111 licenciados recién graduados. Se registraron el campo de especialización y el sueldo mensual inicial de cada profesional. La figura 3.7 muestra los diagramas de caja de Minitab para contabilidad, finanzas, sistemas de información, administración y marketing. Observe que el área de especialización aparece en el eje horizontal, y cada diagrama de caja en el eje vertical por encima del área correspondiente. Mostrar los diagramas de caja de esta manera es una técnica gráfica excelente para hacer comparaciones entre dos o más grupos.

¿Qué observaciones puede hacer acerca de los sueldos iniciales por área de especialización usando los diagramas de caja de la figura 3.7? En específico se observa lo siguiente.

- Los sueldos más altos corresponden a contabilidad; los sueldos más bajos corresponden a administración y marketing.
- Con base en las medianas, la de los sueldos de contabilidad y sistemas de información es similar y mayor. Le sigue finanzas, y administración y contabilidad muestran sueldos con una mediana inferior.
- Existen observaciones atípicas de sueldos altos para las áreas de contabilidad, finanzas y marketing.
- Los sueldos en el área de finanzas parecen tener menos variación, mientras que en contabilidad parecen tener la mayor variación.

Tal vez pueda ver otras interpretaciones basadas en estos diagramas de caja.

FIGURA 3.7 Diagramas de cuadro de Minitab de los sueldos mensuales iniciales por área de especialización



NOTAS Y COMENTARIOS

1. Una ventaja de los procedimientos del análisis exploratorio de datos estriba en que son fáciles de usar, ya que requieren pocos cálculos numéricos. Sencillamente los valores de datos se clasifican en orden ascendente y se identifica el resumen de cinco números. Entonces puede trazarse el diagrama de caja. No es necesario calcular la media y la desviación estándar de los datos.
2. En el apéndice 3.1 se explica cómo elaborar un diagrama de caja de los datos de los sueldos iniciales usando Minitab. El diagrama obtenido se parece al de la figura 3.6, pero girado hacia un lado.

Ejercicios

Métodos

36. Considere una muestra con los datos 27, 25, 20, 15, 30, 34, 28 y 25. Proporcione el resumen de cinco números de los datos.
37. Elabore el diagrama de caja de los datos del ejercicio 36.

AUTO evaluación

38. Muestre el resumen de cinco números y el diagrama de caja de los datos siguientes: 5, 15, 18, 10, 12, 16, 10, 6.
39. Un conjunto de datos tiene un primer cuartil de 42 y un tercer cuartil de 50. Calcule los límites inferior y superior del diagrama de caja correspondiente. ¿Un valor de datos de 65 debe considerarse una observación atípica?

Aplicaciones

40. Naples, Florida, celebra un medio maratón (carrera de 13.1 millas) en enero de cada año. El evento atrae a corredores de todo Estados Unidos y de otras partes del mundo. En enero de 2009 entraron 22 hombres (Men) y 31 mujeres (Women) en la clase de edades de 19 a 24 años. Los tiempos de llegada a la meta en minutos se listan enseguida (*Naples Daily News,* 19 de enero de 2009). Los tiempos se muestran en orden de llegada (Finish).

WEB archivo

Runners

Finish	Men	Women	Finish	Men	Women	Finish	Men	Women
1	65.30	109.03	11	109.05	123.88	21	143.83	136.75
2	66.27	111.22	12	110.23	125.78	22	148.70	138.20
3	66.52	111.65	13	112.90	129.52	23		139.00
4	66.85	111.93	14	113.52	129.87	24		147.18
5	70.87	114.38	15	120.95	130.72	25		147.35
6	87.18	118.33	16	127.98	131.67	26		147.50
7	96.45	121.25	17	128.40	132.03	27		147.75
8	98.52	122.08	18	130.90	133.20	28		153.88
9	100.52	122.48	19	131.80	133.50	29		154.83
10	108.18	122.62	20	138.63	136.57	30		189.27
						31		189.28

a) George Towett, de Marietta, Georgia, llegó en primer lugar de los hombres y Lauren Wald, de Gainesville, Florida, llegó en primer lugar de las mujeres. Compare los tiempos de llegada de los primeros lugares para ambos grupos. Si los 53 corredores hombres y mujeres hubieran competido como un grupo, ¿en qué lugar habría terminado Lauren?

b) ¿Cuál es el tiempo medio para los corredores hombres y mujeres? Compare a los corredores y a las corredoras con base en la mediana de sus tiempos.

c) Proporcione un resumen de cinco números tanto de los hombres como de las mujeres.

d) ¿Hay observaciones atípicas en alguno de los dos grupos?

e) Muestre los diagramas de caja para los dos grupos. ¿Quiénes tienen la mayor variación en los tiempos de llegada: los hombres o las mujeres? Explique.

AUTOevaluación

41. A continuación se proporcionan las ventas anuales, en millones de dólares, de 21 compañías farmacéuticas.

8 408	1 374	1 872	8 879	2 459	11 413
608	14 138	6 452	1 850	2 818	1 356
10 498	7 478	4 019	4 341	739	2 127
3 653	5 794	8 305			

a) Proporcione un resumen de cinco números.
b) Calcule los límites inferior y superior.
c) ¿Los datos contienen observaciones atípicas?
d) Las ventas de $14 138 millones de Johnson & Johnson son las más altas de la lista. Suponga que cometió un error al introducir los datos (una transposición) y que las ventas se introdujeron como $41 138 millones. ¿El método de detección de observaciones del inciso *c*) identifica este problema y permite corregir errores en la introducción de datos?
e) Muestre un diagrama de caja.

42. *Consumer Reports* proporcionó calificaciones de satisfacción del cliente en general para los servicios de telefonía celular AT&T, Sprint, T-Mobile y Verizon en zonas metropolitanas importantes de todo Estados Unidos. La calificación de cada servicio refleja la satisfacción del cliente considerando una variedad de factores como el costo, los problemas de conectividad, las llamadas suspendidas, la interferencia estática y el soporte técnico. Se utilizó una escala de satisfacción de 0 a 100, en la cual 0 indica una insatisfacción total y 100 una satisfacción total. Las calificaciones para los cuatro servicios de telefonía celular en 20 zonas metropolitanas se muestran en seguida (*Consumer Reports,* enero de 2009).

WEB archivo
CellService

Metropolitan Area	AT&T	Sprint	T-Mobile	Verizon
Atlanta	70	66	71	79
Boston	69	64	74	76
Chicago	71	65	70	77
Dallas	75	65	74	78
Denver	71	67	73	77
Detroit	73	65	77	79
Jacksonville	73	64	75	81
Las Vegas	72	68	74	81
Los Ángeles	66	65	68	78
Miami	68	69	73	80
Minneapolis	68	66	75	77
Philadelphia	72	66	71	78
Phoenix	68	66	76	81
San Antonio	75	65	75	80
San Diego	69	68	72	79
San Francisco	66	69	73	75
Seattle	68	67	74	77
St. Louis	74	66	74	79
Tampa	73	63	73	79
Washington	72	68	71	76

a) Considere T-Mobile primero. ¿Cuál es la mediana de la calificación?
b) Elabore un resumen de cinco números para el servicio de esta empresa.
c) ¿Hay observaciones atípicas para T-Mobile? Explique por qué.
d) Repita los incisos *b*) y *c*) para los otros tres servicios de telefonía celular.

e) Presente los diagramas de caja para los cuatro servicios de telefonía celular en una gráfica. Comente qué indica la comparación de diagramas acerca de los cuatro servicios. ¿Cuál recomendó *Consumer Reports* como el mejor en cuanto a la satisfacción del cliente en general?

43. Los Phillies de Filadelfia triunfaron en la Serie Mundial de beisbol de las grandes ligas de 2008 al derrotar a Mantarrayas de Tampa Bay 4 a 3 (*The Philadelphia Inquirer,* 29 de octubre de 2008). Antes, en la clasificatoria de las grandes ligas de beisbol, los Phillies de Filadelfia ganaron el Campeonato de la Liga Nacional al vencer a Los Dodgers de Los Ángeles, mientras que Mantarrayas de Tampa Bay se llevó el Campeonato de la Liga Americana al derrotar a los Medias Rojas de Boston Red Sox. El archivo MLBSalaries contiene los sueldos de los 28 jugadores de cada uno de estos cuatro equipos (base de datos de sueldos de *USA Today*, octubre de 2008). Los datos, mostrados en miles de dólares, se han ordenado del sueldo mayor al menor para cada equipo.

WEB archivo
MLBSalaries

a) Analice los sueldos para el campeón mundial Phillies de Filadelfia. ¿Cuál es la nómina total del equipo? ¿Cuál es la mediana del sueldo? Proporcione el resumen de cinco números.

b) ¿Hay observaciones atípicas para los Phillies de Filadelfia? De ser así, ¿cuántos y de cuánto son los montos de los sueldos?

c) ¿Cuál es la nómina total de cada uno de los otros tres equipos? Elabore el resumen de cinco números para cada equipo e identifique cualesquiera observaciones atípicas.

d) Muestre los diagramas de caja de los sueldos para los cuatro equipos. ¿Cuáles son sus interpretaciones? De estos cuatro equipos, ¿parece que el equipo con sueldos más altos ganó los campeonatos de la liga y la Serie Mundial?

WEB archivo
Mutual

44. Un listado de 46 fondos de inversión y su rendimiento porcentual total de 12 meses se muestra en la tabla 3.5 (*Smart Money,* febrero de 2004).

a) ¿Cuáles son la media y la mediana de los porcentajes de rendimiento para estos fondos de inversión?

b) ¿Cuáles son el primer y el tercer cuartiles?

c) Proporcione un resumen de cinco números.

d) ¿Los datos contienen alguna observación atípica? Muestre un diagrama de caja.

TABLA 3.5 Rendimiento de 12 meses para fondos de inversión

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

3.5 Medidas de asociación entre dos variables

Hasta ahora hemos examinado los métodos numéricos que resumen los datos de *una variable a la vez*. Un gerente o quien toma decisiones se interesa con frecuencia en la *relación entre dos variables*. En esta sección se presentan la covarianza y la correlación como medidas descriptivas de la relación entre dos variables.

Para empezar, reconsidere la aplicación referente a una tienda de estéreos y equipos de sonido en San Francisco que se presentó en la sección 2.4. El gerente del establecimiento quiere determinar la relación entre el número de comerciales de televisión transmitidos el fin de semana y las ventas en la tienda durante la semana siguiente. Los datos muestrales con las ventas expresadas en cientos de dólares se proporcionan en la tabla 3.6. Ésta registra 10 observaciones ($n = 10$), una para cada semana. El diagrama de dispersión de la figura 3.8 indica una relación positiva, con las ventas más altas (y) asociadas con un número mayor de comerciales (x). De hecho, el diagrama de dispersión sugiere que se podría usar una línea recta como una aproximación de la relación. En el análisis siguiente se introduce la **covarianza** como una medida descriptiva de la asociación lineal entre dos variables.

Covarianza

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

COVARIANZA MUESTRAL

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

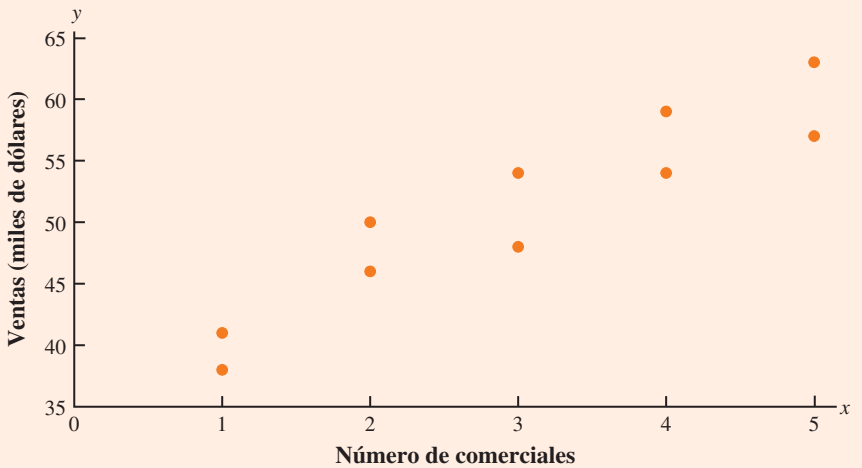
Esta fórmula empareja cada x_i con una y_i. Luego se suman los productos obtenidos al multiplicar la desviación de cada x_i de su media muestral $\bar{x}$ por la desviación de la y_i correspondiente de su media muestral $\bar{y}$; esta suma se divide entonces por $n - 1$.

TABLA 3.6 Datos muestrales para la tienda de estéreos y equipos de sonido

WEB archivo
Stereo

Week	Number of Commercials x	Sales Volume ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

FIGURA 3.8 Diagrama para la tienda de estéreos y equipos de sonido



Para medir la solidez de una relación lineal entre el número de comerciales (Number of Commercials) x y el volumen de ventas (Sales Volume) y en el problema de la tienda de estéreos y equipos de sonido, use la ecuación (3.10) a efecto de calcular la covarianza muestral. La tabla 3.7 presenta el cálculo de $\Sigma(x_i - \bar{x})(y_i - \bar{y})$. Observe que $\bar{x} = 30/10 = 3$, y $\bar{y} = 510/10 = 51$. Usando la ecuación (3.10) se obtiene una covarianza muestral de

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

TABLA 3.7 Cálculos de la covarianza muestral

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

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

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

COVARIANZA POBLACIONAL

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

En la ecuación (3.11) la notación μ_x denota la media poblacional de la variable x, y μ_y denota la media poblacional de la variable y. La covarianza poblacional σ_{xy} se define para una población de tamaño N.

Interpretación de la covarianza

Para ayudar en la interpretación de la covarianza muestral, considere la figura 3.9; es igual al diagrama de dispersión de la figura 3.7, con una línea punteada vertical en $\bar{x} = 3$ y una línea punteada horizontal en $\bar{y} = 51$. Las líneas dividen la gráfica en cuatro cuadrantes. Los puntos del cuadrante I corresponden a x_i mayor que $\bar{x}$ y y_i mayor que $\bar{y}$; los puntos del cuadrante II corresponden a x_i menor que $\bar{x}$ y y_i menor que $\bar{y}$, etc. Por tanto, el valor de $(x_i - \bar{x})(y_i - \bar{y})$ debe ser positivo para los puntos del cuadrante I, negativo para los del cuadrante II, positivo para los del cuadrante III, y negativo para los puntos del cuadrante IV.

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

Si el valor de s_{xy} es positivo, los puntos con la mayor influencia en s_{xy} deben estar en los cuadrantes I y III. Por ende, un valor positivo para s_{xy} indica una asociación lineal positiva entre x y y; es decir, a medida que el valor de x aumenta, el valor de y también. Si el valor de s_{xy} es negativo, no obstante, los puntos con la mayor influencia en s_{xy} están en los cuadrantes II y IV. Por ende, un valor negativo para s_{xy} indica una asociación lineal negativa entre x y y; es decir, a medida que el valor de x aumenta, el valor de y disminuye. Por último, si los puntos están distribuidos de manera uniforme en los cuatro cuadrantes, el valor de s_{xy} será cercano a cero, lo que indica que no existe una asociación lineal entre x y y. En la figura 3.10 se aprecian los valores de s_{xy} que se expresan con tres tipos distintos de diagramas de dispersión.

FIGURA 3.9 Diagrama de dispersión particionado para la tienda de estéreos y equipos de sonido

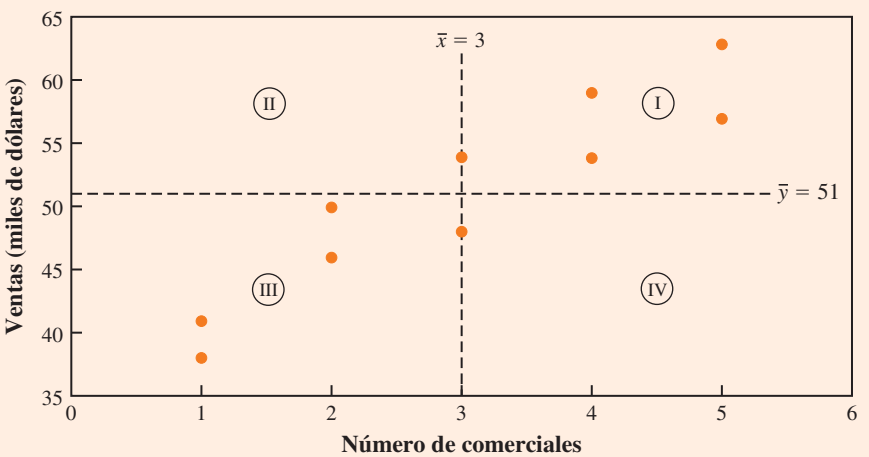
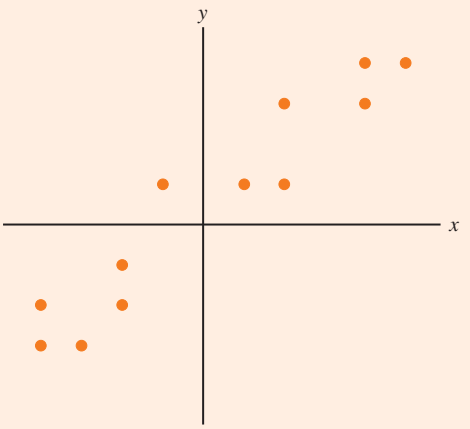
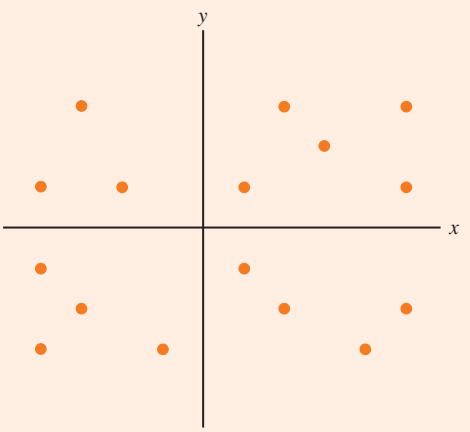


FIGURA 3.10 Interpretación de la covarianza muestral

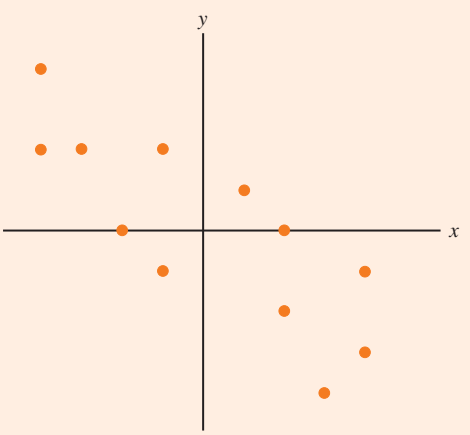
s_{xy} positiva:
(x y y se relacionan linealmente de manera positiva)



s_{xy} aproximadamente 0:
(x y y no se relacionan de manera lineal)



s_{xy} negativa:
(x y y se relacionan linealmente de manera negativa)



Observe de nuevo la figura 3.9. El diagrama de dispersión para la tienda de estéreos y equipos de sonido sigue el patrón del panel superior de la figura 3.10. Como es de esperarse, el valor de la covarianza muestral indica una relación lineal positiva en la que $s_{xy} = 11$.

A partir del análisis anterior, podría parecer que un valor positivo grande para la covarianza indica una relación lineal positiva sólida, y un valor negativo grande indica una relación lineal negativa sólida. Sin embargo, un problema con la covarianza como medida de la solidez de una relación lineal estriba en que su valor depende de las unidades de medida para x y y. Por ejemplo, suponga que estamos interesados en la relación entre la estatura x y el peso y de las personas. Desde luego, la solidez de la relación debe ser la misma, ya sea que la estatura se mida en pies o pulgadas. Sin embargo, la medición en pulgadas no da valores numéricos mucho mayores para $(x_i - \bar{x})$ que cuando la estatura se mide en pies. Por tanto, con la altura medida en pulgadas se obtendría un valor mayor para el numerador $\Sigma(x_i - \bar{x})(y_i - \bar{y})$ en la ecuación (3.10) —y por consiguiente una covarianza mayor—, cuando de hecho la relación no cambia. Una medida de la relación entre dos variables que no se ve afectada por las unidades de medición para x y y es el **coeficiente de correlación**.

Coeficiente de correlación

Para los datos muestrales, el coeficiente de correlación del producto-momento de Pearson se define como se indica a continuación.

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

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

donde

r_{xy} = coeficiente de correlación muestral
s_{xy} = covarianza muestral
s_x = desviación estándar muestral de x
s_y = desviación estándar muestral de y

La ecuación (3.12) indica que el coeficiente de correlación del producto-momento de Pearson para los datos muestrales (conocido comúnmente de manera más simple como *coeficiente de correlación muestral*) se calcula al dividir la covarianza muestral entre el producto de la desviación estándar muestral de x y la desviación estándar muestral de y.

A continuación se calcula el coeficiente de correlación muestral para la tienda de estéreos y equipos de sonido. Usando los datos de la tabla 3.7 se pueden estimar las desviaciones estándar muestrales para las dos variables:

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

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

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

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

La fórmula para calcular el coeficiente de correlación de una población, denotado por la letra griega ρ_{xy} (*ro*), se presenta a continuación.

El coeficiente de correlación muestral r_{xy} es el estimador del coeficiente de correlación poblacional ρ_{xy}.

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

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

donde

ρ_{xy} = coeficiente de correlación poblacional
σ_{xy} = covarianza poblacional
σ_x = desviación estándar poblacional de x
σ_y = desviación estándar poblacional de y

El coeficiente de correlación muestral r_{xy} proporciona una estimación del coeficiente de correlación poblacional ρ_{xy}.

Interpretación del coeficiente de correlación

Primero se considerará un ejemplo sencillo que ilustra el concepto de una relación lineal positiva perfecta. El diagrama de dispersión de la figura 3.11 representa la relación entre x y y con base en los datos muestrales siguientes.

x_i	y_i
5	10
10	30
15	50

FIGURA 3.11 Diagrama de dispersión que representa una relación lineal positiva perfecta

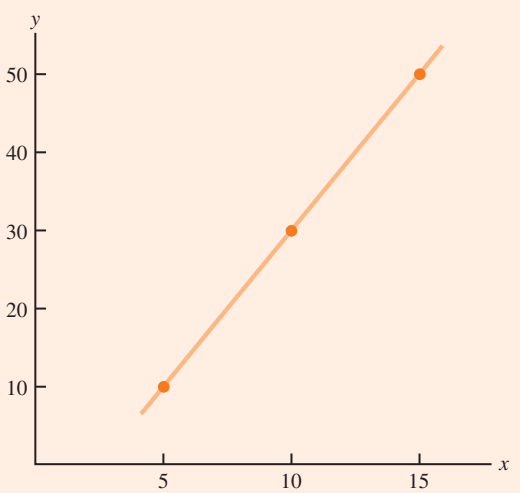


TABLA 3.8 Cálculos utilizados para obtener el coeficiente de correlación muestral

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

La línea recta trazada a través de cada uno de los tres puntos muestra una relación lineal perfecta entre x y y. Con el fin de aplicar la ecuación (3.12) para calcular la correlación muestral, primero se calculan s_{xy}, s_x y s_y. Algunos cálculos se presentan en la tabla 3.8. Con los resultados de esta tabla encontramos

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

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

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

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

El coeficiente de correlación varía de −1 a +1. Los valores cercanos a −1 o a + 1 indican una relación lineal sólida. Entre más se acerque la correlación a cero, más débil es la relación.

Por tanto, se aprecia que el valor del coeficiente de correlación muestral es 1.

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

Suponga ahora que cierto conjunto de datos indica una relación lineal positiva entre x y y pero la relación no es perfecta. El valor de r_{xy} será menor que 1, lo que indica que los puntos en el diagrama de dispersión no estarán todos sobre una línea recta. A medida que los puntos se desvían más y más de una relación lineal positiva perfecta, el valor de r_{xy} se vuelve cada vez más y más pequeño. Cuando éste es igual a cero, indica que no existe una relación lineal entre x y y, y los valores de r_{xy} cercanos a cero indican una relación lineal débil.

Para los datos de la tienda de estéreos y equipos de sonido, $r_{xy} = 0.93$. Por consiguiente, se concluye que existe una relación lineal positiva sólida entre el número de comerciales y las ventas. De manera más específica, un aumento en el número de comerciales se asocia con un incremento en las ventas.

En resumen, se observa que la correlación proporciona una medida de asociación lineal y no necesariamente de causalidad. Una correlación alta entre dos variables no significa que los cambios en una variable ocasionarán cambios en la otra. Por ejemplo, podemos encontrar que la calificación de calidad y el precio típico de la comida en los restaurantes se correlacionan de manera positiva. Sin embargo, un simple incremento en el precio de la comida no causará que la calificación de la calidad aumente.

Ejercicios

Métodos

AUTOevaluación

45. A continuación se presentan cinco observaciones tomadas para dos variables.

x_i	4	6	11	3	16
y_i	50	50	40	60	30

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

46. A continuación se presentan cinco observaciones tomadas para dos variables.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a) Elabore un diagrama de dispersión con estos datos.
b) ¿Qué indica el diagrama de dispersión acerca de la relación entre x y y?
c) Calcule e interprete la covarianza muestral.
d) Determine e interprete el coeficiente de correlación muestral.

Aplicaciones

47. Nielsen Media Research proporciona dos medidas de la audiencia televisiva: el rating, que es el porcentaje de hogares que cuenta con un aparato y está viendo un programa, y el share, que es el porcentaje de hogares que tiene el equipo encendido cuyos miembros están viendo un programa determinado. Las cifras siguientes muestran los datos de las calificaciones y las cuotas de Nielsen de la Serie Mundial de Beisbol de las Grandes Ligas durante un periodo de nueve años (Associated Press, 27 de octubre de 2003).

Rating	19	17	17	14	16	12	15	12	13
Share	32	28	29	24	26	20	24	20	22

a) Elabore un diagrama de dispersión con el rating en el eje horizontal.
b) ¿Cuál es la relación entre rating y share? Explique por qué.
c) Calcule e interprete la covarianza muestral.
d) Calcule el coeficiente de correlación muestral. ¿Qué indica este valor acerca de la relación entre rating y share?

48. Un estudio de un departamento de transporte sobre la velocidad de manejo y las millas por galón para automóviles de tamaño mediano dio como resultado los datos siguientes.

Velocidad (millas por hora)	30	50	40	55	30	25	60	25	50	55
Millas por galón	28	25	25	23	30	32	21	35	26	25

Calcule e interprete el coeficiente de correlación muestral.

49. A principios de 2009 el declive económico ocasionó la pérdida de empleos y un incremento en los préstamos morosos para vivienda. La tasa nacional de desempleo fue de 6.5% y el porcentaje de préstamos morosos de 6.12% (*The Wall Street Journal,* 27 de enero de 2009). En la proyección de hacia dónde se dirigía el mercado de bienes raíces el siguiente año, los economistas estudiaron la relación entre la tasa de desempleo y el porcentaje de préstamos morosos. La expectativa era que si la primera seguía en aumento, habría también un incremento en el porcentaje de préstamos con deudores morosos. Los datos siguientes muestran la tasa de

desempleo y el porcentaje de préstamos morosos para 27 de los principales mercados de bienes raíces.

WEB archivo
Housing

Metro Area	Jobless Rate (%)	Delinquent Loan (%)	Metro Area	Jobless Rate (%)	Delinquent Loan (%)
Atlanta	7.1	7.02	Nueva York	6.2	5.78
Boston	5.2	5.31	Orange County	6.3	6.08
Charlotte	7.8	5.38	Orlando	7.0	10.05
Chicago	7.8	5.40	Philadelphia	6.2	4.75
Dallas	5.8	5.00	Phoenix	5.5	7.22
Denver	5.8	4.07	Portland	6.5	3.79
Detroit	9.3	6.53	Raleigh	6.0	3.62
Houston	5.7	5.57	Sacramento	8.3	9.24
Jacksonville	7.3	6.99	St. Louis	7.5	4.40
Las Vegas	7.6	11.12	San Diego	7.1	6.91
Los Ángeles	8.2	7.56	San Francisco	6.8	5.57
Miami	7.1	12.11	Seattle	5.5	3.87
Minneapolis	6.3	4.39	Tampa	7.5	8.42
Nashville	6.6	4.78			

a) Calcule el coeficiente de correlación. ¿Existe una correlación positiva entre la tasa de desempleo (Jobless Rate) y el porcentaje de préstamos de vivienda morosos (Delinquent Loan)? ¿Cuál es su interpretación?

b) Muestre un diagrama de dispersión de la relación entre la tasa de desempleo y el porcentaje de préstamos de vivienda morosos.

50. El promedio industrial Dow Jones (DJIA) y el índice 500 de Standard & Poor's (S&P 500) miden el desempeño del mercado de valores. El DJIA se basa en el precio de las acciones de 30 empresas grandes; el S&P 500, en el precio de las acciones de 500 empresas. Si tanto el DJIA como el S&P 500 miden el desempeño del mercado de valores, ¿cómo se correlacionan? Los datos siguientes ilustran el incremento o el decremento porcentual diario en el DJIA y el S&P 500 para una muestra de nueve días durante un periodo de tres meses (*The Wall Street Journal,* 15 de enero a 10 de marzo de 2006).

WEB archivo
StockMarket

DJIA	0.20	0.82	−0.99	0.04	−0.24	1.01	0.30	0.55	−0.25
S&P 500	0.24	0.19	−0.91	0.08	−0.33	0.87	0.36	0.83	−0.16

a) Elabore un diagrama de dispersión.

b) Calcule el coeficiente de correlación muestral para estos datos.

c) Comente la asociación entre el DJIA y el S&P 500. ¿Necesita revisarlos antes de darse una idea general sobre el desempeño diario del mercado de valores?

51. Las temperaturas diarias altas (High) y bajas (Low) para 14 ciudades de todo el mundo se muestran en el siguiente cuadro (The Weather Channel, 22 de abril de 2009).

WEB archivo
WorldTemp

City	High	Low	City	High	Low
Athens	68	50	London	67	45
Beijing	70	49	Moscow	44	29
Berlin	65	44	Paris	69	44
Cairo	96	64	Rio de Janeiro	76	69
Dublin	57	46	Rome	69	51
Geneva	70	45	Tokyo	70	58
Hong Kong	80	73	Toronto	44	39

a) ¿Cuál es la media muestral de la temperatura alta?
b) ¿Cuál es la media muestral de la temperatura baja?
c) ¿Cuál es la correlación entre las temperaturas alta y baja? Comente.

3.6 Media ponderada y trabajo con datos agrupados

En la sección 3.1 se presentó la media como una de las medidas más importantes de ubicación central. La fórmula para la media de una muestra con n observaciones se vuelve a establecer como sigue.

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

En esta fórmula, cada x_i recibe igual importancia o peso. Aunque esta práctica es la más común, en algunos casos la media se calcula confiriendo a cada observación un peso que refleje su importancia. Una media calculada de esta manera se conoce como **media ponderada**.

Media ponderada

La media ponderada se calcula como sigue.

MEDIA PONDERADA

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

donde

x_i = valor de observación i
w_i = peso de la observación i

Cuando los datos provienen de una muestra, la ecuación (3.15) proporciona la media muestral ponderada. Cuando son de una población, μ reemplaza a $\bar{x}$ y la misma ecuación proporciona la media poblacional ponderada.

Como ejemplo de la necesidad de una media ponderada, considere la muestra siguiente de cinco compras de una materia prima durante los tres meses pasados.

Compra	Costo por libra ($)	Número de libras
1	3.00	1 200
2	3.40	500
3	2.80	2 750
4	2.90	1 000
5	3.25	800

Observe que el costo por libra varía de $2.80 a $3.40, y la cantidad comprada varía de 500 a 2 750 libras. Suponga que un gerente solicitó información sobre el costo medio por libra de la materia prima. Debido a que las cantidades ordenadas varían, se debe usar la fórmula para una media ponderada. Los cinco valores de datos del costo por libra son $x_1 = 3.00$; $x_2 = 3.40$; $x_3 = 2.80$; $x_4 = 2.90$, y $x_5 = 3.25$. El costo medio ponderado por libra se obtuvo al ponderar

cada costo por su cantidad correspondiente. Para este ejemplo, los pesos son $w_1 = 1\,200$; $w_2 = 500$; $w_3 = 2\,750$; $w_4 = 1\,000$, y $w_5 = 800$. Con base en la ecuación (3.15), la media ponderada se calculó como sigue.

$$\bar{x} = \frac{1\,200(3.00) + 500(3.40) + 2\,750(2.80) + 1\,000(2.90) + 800(3.25)}{1\,200 + 500 + 2\,750 + 1\,000 + 800}$$

$$= \frac{18\,500}{6\,250} = 2.96$$

Por tanto, el cálculo de la media ponderada indica que el costo medio por libra para la materia prima es \$2.96. Observe que utilizando la ecuación (3.14) en vez de la fórmula de la media ponderada se habrían obtenido resultados erróneos. En este caso, la media de los cinco valores del costo por libra es (3.00 + 3.40 + 2.80 + 2.90 + 3.25)/5 = 15.35/5 = \$3.07, el cual exagera el costo medio real por libra adquirida.

El cálculo de un promedio escolar es un buen ejemplo del uso de la media ponderada.

La opción de los pesos para el cálculo de una media ponderada en particular depende de la aplicación. Un ejemplo muy conocido para los estudiantes universitarios es el cálculo de un promedio escolar. En éste, los valores de datos manejados son por lo general 4 para una calificación A; 3 para una calificación B; 2 para una calificación C; 1 para una calificación D, y 0 para una calificación F. Los pesos son el número de horas de los créditos ganados por cada calificación. El ejercicio 54 al final de esta sección proporciona un ejemplo de este cálculo de la media ponderada. En otros cálculos, las cantidades como las libras, los dólares o el volumen suelen usarse como pesos. Sea como fuere, cuando las observaciones varían en importancia, el analista debe elegir el peso que mejor refleje la importancia de cada observación en la determinación de la media.

Datos agrupados

En la mayoría de los casos, las medidas de posición y variabilidad se calculan con valores de datos individuales. No obstante, los datos en ocasiones están disponibles sólo en forma agrupada o en forma de distribución de frecuencia. En el análisis siguiente se explica cómo usar la fórmula de la media ponderada para obtener aproximaciones de la media, la varianza y la desviación estándar para **datos agrupados**.

En la sección 2.2 se proporcionó una distribución de frecuencia del tiempo en días requerido para completar las auditorías de fin de año de la firma de contabilidad pública Sanderson and Clifford. La distribución de frecuencia de la duración de las auditorías se ilustra en la tabla 3.9. Con base en esta distribución, ¿cuál es la media muestral de la duración de las auditorías?

Para calcular la media usando sólo los datos agrupados, el punto medio de cada clase se trata como si fuera representativo de los elementos de la clase. Sea M_i el punto medio para la clase i, y f_i la frecuencia de la clase i. La fórmula de la media ponderada (3.15) se utiliza entonces con los valores de datos denotada como M_i y los pesos dados por las frecuencias f_i. En este caso,

TABLA 3.9 Distribución de frecuencia de la duración de la auditoría

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

el denominador de la ecuación es la suma de las frecuencias, la cual es el tamaño muestral n. Es decir, $\Sigma f_i = n$. Por tanto, la ecuación para la media muestral de los datos agrupados es la siguiente.

MEDIA MUESTRAL PARA DATOS AGRUPADOS

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

donde

M_i = punto medio para la clase i
f_i = frecuencia para la clase i
n = tamaño muestral

Con los puntos medios de clase, M_i, a medio camino entre los límites de clase, la primera de 10–14 en la tabla 3.9 tiene un punto medio en $(10 + 14)/2 = 12$. Los cinco puntos medios de clase y el cálculo de la media ponderada para los datos de duración de la auditoría se resumen en la tabla 3.10. Como puede verse, la media muestral de la duración de la auditoría es de 19 días.

Para calcular la varianza de datos agrupados se usa una versión ligeramente alterada de la fórmula para la varianza proporcionada en la ecuación (3.5). En esta ecuación las desviaciones cuadradas de los datos con respecto a la media muestral $\bar{x}$ se escribieron como $(x_i - \bar{x})^2$. Sin embargo, con los datos agrupados, los valores no se conocen. En este caso, el punto medio de la clase, M_i, se trata como si fuera representativo de los x_i valores en la clase correspondiente. Por tanto, las desviaciones cuadradas respecto de la media muestral, $(x_i - \bar{x})^2$, se remplazan por $(M_i - \bar{x})^2$. Así, del mismo modo que con los cálculos de la media muestral para los datos agrupados, se pesa cada valor por la frecuencia de la clase, f_i. La suma de las desviaciones cuadradas con respecto a la media para todos los datos se aproxima por medio de $\Sigma f_i(M_i - \bar{x})^2$. El término $n - 1$ en vez de n aparece en el denominador con el fin de hacer de la varianza muestral la estimación de la varianza poblacional. De ahí que la fórmula siguiente se use con objeto de obtener la varianza muestral para los datos agrupados.

VARIANZA MUESTRAL PARA DATOS AGRUPADOS

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

TABLA 3.10 Cálculo de la media muestral de la duración de la auditoría para los datos agrupados

Duración de la auditoría (días)	Punto medio de la clase (M_i)	Frecuencia (f_i)	$f_i M_i$
10–14	12	4	48
15–19	17	8	136
20–24	22	5	110
25–29	27	2	54
30–34	32	1	32
		20	380

Media muestral $\bar{x} = \frac{\Sigma f_i M_i}{n} = \frac{380}{20} = 19$ días

TABLA 3.11 Cálculo de la varianza muestral de la duración de la auditoría para los datos agrupados (media muestral $\bar{x} = 19$)

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

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

El cálculo de la varianza muestral para la duración de la auditoría con base en los datos agrupados se ilustra en la tabla 3.11. La varianza muestral es 30.

La desviación estándar para los datos agrupados es sencillamente la raíz cuadrada de la varianza para tales datos. Para los datos de duración de la auditoría, la desviación estándar muestral es $s = \sqrt{30} = 5.48$.

Antes de concluir con esta sección sobre el cálculo de las medidas de posición y dispersión para los datos agrupados, observe que las fórmulas (3.16) y (3.17) son para una muestra. Las medidas para la población se calculan de modo parecido. Las fórmulas de los datos agrupados para una media y varianza poblacionales se presentan a continuación.

MEDIA POBLACIONAL PARA DATOS AGRUPADOS

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

VARIANZA POBLACIONAL PARA DATOS AGRUPADOS

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

NOTAS Y COMENTARIOS

En el cálculo de la estadística descriptiva para los datos agrupados, los puntos medios de las clases se utilizan para aproximar los valores de datos de cada clase. Como resultado, la estadística descriptiva para los datos agrupados se aproxima a la estadística descriptiva que resultaría directamente del uso de los datos originales. Por consiguiente, siempre que sea posible es recomendable calcular los estadísticos descriptivos a partir de los datos originales en vez de hacerlo a partir de los datos agrupados.

Ejercicios

Métodos

52. Considere los datos siguientes y sus pesos correspondientes.

x_i	Peso (w_i)
3.2	6
2.0	3
2.5	2
5.0	8

a) Calcule la media ponderada.
b) Calcule la media muestral de los cuatro valores de datos sin ponderar. Observe la diferencia en los resultados proporcionados por los dos cálculos.

AUTO evaluación

53. Considere los datos muestrales en la frecuencia de distribución siguiente.

Clase	Punto medio	Frecuencia
3–7	5	4
8–12	10	7
13–17	15	9
18–22	20	5

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

Aplicaciones

AUTO evaluación

54. El promedio de calificaciones para los estudiantes universitarios se basa en el cálculo de una media ponderada. Para la mayoría de los estudiantes, las calificaciones se proporcionan con los valores de datos siguientes: A (4), B (3), C (2), D (1) y F (0). Después de 60 horas de clase de estudios superiores, un alumno de la universidad estatal obtuvo 9 horas de clase de A, 15 de clase de B, 33 de clase de C y 3 horas de clase de D.
a) Calcule el promedio de calificaciones del estudiante.
b) Los alumnos de la universidad estatal deben mantener un promedio de calificaciones de 2.5 para sus primeras 60 horas de clases de estudios superiores con el fin de ser admitidos en el colegio de administración. ¿Este estudiante será admitido?

55. Morningstar da seguimiento al rendimiento total de un número grande de fondos de inversión. La tabla siguiente registra el rendimiento total y el número de fondos para cuatro categorías de fondos de inversión (*Morningstar Funds500*, 2008).

Tipo de fondo	Número de fondos	Rendimiento total (%)
Capital nacional	9 191	4.65
Capital internacional	2 621	18.15
Capital especializado	1 419	11.36
Híbridos	2 900	6.75

a) Usando el número de fondos como pesos, calcule el rendimiento total promedio ponderado para los fondos de inversión cubiertos por Morningstar.
b) ¿Hay alguna dificultad asociada con el uso del "número de fondos" como pesos en el cálculo del rendimiento total promedio ponderado para Morningstar en el inciso *a*)? Comente. ¿Qué más podría usarse para los pesos?
c) Suponga que invirtió $10 000 en fondos de inversión a principios de 2007 y que diversificó la inversión al colocar $2 000 en fondos de capital nacional, $4 000 en fondos de capital

internacional, $3 000 en fondos de capital especializado y $1 000 en fondos híbridos. ¿Cuál es el rendimiento esperado sobre el portafolio?

56. Con base en una encuesta de 425 programas de la maestría en administración de empresas, el informe de *U.S. News & World Report* calificó el programa de la Escuela de Negocios de la Universidad Kelley de Indiana como el 20o. mejor del país (*America's Best Graduate Schools,* 2009). La calificación se basó en parte en encuestas a decanos de la escuela de negocios y a reclutadores corporativos. Se solicitó a todos los consultados que evaluaran la calidad académica general del programa de maestría en una escala de 1 "marginal" a 5 "sobresaliente". Use la muestra de respuestas listada abajo para calcular la calificación media ponderada de los decanos de la escuela de negocios y los reclutadores corporativos. Comente.

Evaluación de la calidad	Decanos de la escuela de negocios	Reclutadores corporativos
5	44	31
4	66	34
3	60	43
2	10	12
1	0	0

57. La distribución de frecuencia siguiente muestra el precio por acción de las 30 empresas del promedio industrial Dow Jones (*Barron's,* 2 de febrero de 2009).

Precio por acción	Número de empresas
$ 0–9	4
$10–19	5
$20–29	7
$30–39	3
$40–49	4
$50–59	4
$60–69	0
$70–79	2
$80–89	0
$90–99	1

a) Calcule el precio medio por acción y la desviación estándar del precio por acción para las empresas del promedio industrial Dow Jones.

b) El 16 de enero de 2006, el precio medio por acción era de $45.83 y la desviación estándar de $18.14. Comente los cambios ocurridos en el precio por acción durante el periodo de tres años.

Resumen

En este capítulo se introdujeron varios estadísticos descriptivos que se utilizan para resumir la posición, la variabilidad y la forma de una distribución de datos. A diferencia de los procedimientos tabulares y gráficos del capítulo 2, las medidas en este capítulo resumen los datos en términos de valores numéricos. Cuando los valores numéricos se obtienen de una muestra, se les llama *estadísticos muestrales*; cuando se obtienen de una población se llaman *parámetros poblacionales*. En seguida se presenta parte de la notación utilizada para ambos conceptos.

En la inferencia estadística, la estadística muestral se conoce como estimador puntual del parámetro poblacional.

	Estadístico muestral	Parámetro poblacional
Media	$\bar{x}$	μ
Varianza	s^2	σ^2
Desviación estándar	s	σ
Covarianza	s_{xy}	σ_{xy}
Correlación	r_{xy}	ρ_{xy}

Se definieron la media, la mediana y la moda como medidas de la posición central. Luego se utilizó el concepto de percentiles para describir otras posiciones en el conjunto de datos. A continuación se presentaron el rango, el rango intercuartílico, la varianza, la desviación estándar y el coeficiente de variación como medidas de variabilidad o dispersión. Nuestra medida principal de la forma de una distribución de datos fue el sesgo. Los valores negativos indican una distribución de datos sesgada a la izquierda; los valores positivos indican una distribución de datos sesgada a la derecha. Luego se describió cómo se usan la media y la desviación estándar al aplicar el teorema de Chebyshev y la regla empírica para proporcionar más información sobre la distribución de los datos e identificar observaciones atípicas.

En la sección 3.4 se muestra cómo elaborar un resumen de cinco números y un diagrama de caja para proporcionar información simultánea sobre la ubicación, la variabilidad y la forma de la distribución. En la sección 3.5 se introdujeron la covarianza y el coeficiente de correlación como medidas de asociación entre dos variables. En la sección final se explicó cómo calcular una media ponderada, así como la media, la varianza y la desviación estándar para datos agrupados.

Los estadísticos descriptivos estudiados pueden obtenerse por medio de software para estadística y hojas de cálculo. En los apéndices del capítulo se explica cómo se usan Minitab, Excel y StatTools para elaborar los estadísticos descriptivos que se trabajaron en este capítulo.

Glosario

Coeficiente de correlación Medida de la asociación lineal entre dos variables que toma los valores entre -1 y $+1$. Los valores cercanos a $+1$ indican una relación lineal positiva sólida; los valores cercanos a -1 indican una relación lineal negativa sólida, y los valores cercanos a cero, la falta de una relación lineal.

Coeficiente de variación Medida de variabilidad relativa calculada al dividir la desviación estándar entre la media y multiplicar por 100.

Covarianza Medida de la asociación lineal entre dos variables. Los valores positivos indican una relación positiva; los valores negativos indican una relación negativa.

Cuartiles Los percentiles 25, 50 y 75, conocidos como primer cuartil, segundo cuartil (mediana) y tercer cuartil, respectivamente. Los cuartiles se usan para dividir un conjunto de datos en cuatro partes, con cada parte conteniendo aproximadamente 25% de los datos.

Datos agrupados Datos disponibles en intervalos de clase según se resumen por una distribución de frecuencia. Los valores individuales de los datos originales no están disponibles.

Desviación estándar Medida de variabilidad calculada al tomar la raíz cuadrada positiva de la varianza.

Diagrama de caja Resumen gráfico de los datos basado en un resumen de cinco números.

Estadístico muestral Valor numérico usado como medida de resumen para una muestra (por ejemplo, la media muestral, $\bar{x}$, la varianza muestral, s^2, y la desviación estándar de la muestra, s).

Estimador puntual Los estadísticos muestrales, como $\bar{x}$, s^2 y s, cuando se utilizan para estimar el parámetro poblacional correspondiente.

Media Medida de la ubicación central calculada al resumir los valores de datos y dividir entre el número de observaciones.

Media ponderada La media obtenida al asignar a cada observación un peso que refleje su importancia.

Mediana Medida de la posición central proporcionada por el valor de en medio cuando los datos se acomodan en orden ascendente.

Moda Medida de la posición, definida como el valor que ocurre con mayor frecuencia.

Observación atípica Valor de datos inusualmente pequeño o inusualmente grande.

Parámetro poblacional Valor numérico utilizado como una medida de resumen para una población (por ejemplo, la media poblacional, μ, la varianza poblaciónal, σ^2, y la desviación estándar de la población, σ).

Percentil Valor tal que por lo menos p por ciento de las observaciones es menor o igual que este valor, y como mínimo $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor. El percentil 50 es la mediana.

Rango Medida de la variabilidad definida para ser el valor mayor menos el valor menor.

Rango intercuartílico (RIC) Medida de variabilidad definida como la diferencia entre el tercer y el primer cuartiles.

Regla empírica Se usa para calcular el porcentaje de valores de datos que deben estar dentro de una, dos y tres desviaciones estándar de la media para los datos que exhiben una distribución con forma de campana.

Resumen de cinco números Técnica de análisis exploratorio de datos que usa cinco números para resumir los datos: valor menor, primer cuartil, mediana, tercer cuartil y valor más grande.

Sesgo Medida de la forma de una distribución de datos. Los datos sesgados a la izquierda dan como resultado un sesgo negativo; una distribución de datos simétrica genera un sesgo de cero, y los datos sesgados a la derecha producen un sesgo positivo.

Teoremoa de Chebyshev Se utiliza para hacer enunciados sobre la proporción de los valores de datos que deben estar dentro de un número especificado de desviaciones estándar de la media.

valor z Valor calculado al dividir la desviación con respecto a la media $(x_i - \bar{x})$ entre la desviación estándar s. Una puntuación z se conoce como un valor estandarizado y denota el número de desviaciones estándar x_i a partir de la media.

Varianza Medida de variabilidad basada en las desviaciones cuadradas de los valores de datos con respecto a la media.

Fórmulas clave

Media muestral

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

Media poblacional

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

Rango intercuartílico

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

Varianza poblacional

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

Varianza muestral

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

Desviación estándar

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

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

Coeficiente de variación

$$\left(\frac{\text{desviación estándar}}{\text{media}} \times = 100\right)\% \tag{3.8}$$

Valor z

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

Covarianza muestral

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

Covarianza poblacional

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

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

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

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

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

Media ponderada

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

Media muestral para datos agrupados

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

Varianza muestral para datos agrupados

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

Media poblacional para datos agrupados

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

Varianza poblacional para datos agrupados

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

Ejercicios complementarios

58. Según la encuesta del gasto anual de los consumidores, el promedio mensual del cargo a la tarjeta de crédito Visa del Bank of America fue de $1 838 (*U.S. Airways Attaché Magazine,* diciembre de 2003). Una muestra de cargos mensuales a tarjetas de crédito proporciona los datos siguientes.

WEB archivo
Visa

236	1 710	1 351	825	7 450
316	4 135	1 333	1 584	387
991	3 396	170	1 428	1 688

a) Calcule la media y la mediana.
b) Estime el primer y tercer cuartiles.
c) Calcule el rango y el rango intercuartílico.
d) Determine la varianza y la desviación estándar.
e) La medida del sesgo para estos datos es 2.12. Comente la forma de esta distribución. ¿Es la que usted esperaría? ¿Por qué?
f) ¿Los datos contienen observaciones atípicas?

59. La Oficina del Censo de Estados Unidos (U.S. Census Bureau) proporciona estadísticas sobre la vida familiar en este país, incluyendo la edad en el primer matrimonio, el estado marital actual y el tamaño de la vivienda (sitio web U.S. Census Bureau, 20 de marzo de 2006). Los datos siguientes muestran la edad en el primer matrimonio para una muestra de hombres (Men) y una de mujeres (Women).

WEB archivo
Ages

Hombres	26	23	28	25	27	30	26	35	28
	21	24	27	29	30	27	32	27	25
Mujeres	20	28	23	30	24	29	26	25	
	22	22	25	23	27	26	19		

a) Determine la edad media en la época del primer matrimonio para hombres y mujeres.
b) Calcule el primer y tercer cuartiles para ambos grupos.
c) Hace 25 años la edad media en la época del primer matrimonio era de 25 para los hombres y 22 para las mujeres. ¿Qué elementos proporciona esta información para comprender la decisión de cuándo casarse entre la gente joven en la actualidad?

60. El rendimiento del dividendo es el dividendo anual por acción que una empresa paga, dividido entre el precio por acción actual de mercado expresado como porcentaje. Una muestra de 10 empresas grandes proporciona los siguientes datos de rendimiento del dividendo (*The Wall Street Journal,* 16 de enero de 2004).

Empresa	Rendimiento %	Empresa	Rendimiento %
Altria Group	5.0	General Motors	3.7
American Express	0.8	JPMorgan Chase	3.5
Caterpillar	1.8	McDonald's	1.6
Eastman Kodak	1.9	United Technology	1.5
ExxonMobil	2.5	Wal-Mart Stores	0.7

a) ¿Cuáles son la media y la mediana de los rendimientos?
b) ¿Cuáles son la varianza y la desviación estándar?
c) ¿Qué empresa proporciona el rendimiento del dividendo más alto?
d) ¿Cuál es el valor z para McDonald's? Interprete este valor z.
e) ¿Cuál es el valor z para General Motors? Interprételo.
f) Con base en la puntuación z, ¿los datos contienen alguna observación atípica?

61. El Departamento de Educación de Estados Unidos informa que alrededor de 50% de todos los estudiantes universitarios usa un préstamo estudiantil para ayudarse a cubrir los gastos escolares (National Center for Educational Studies, enero de 2006). En la siguiente lista se observa una muestra de alumnos que se graduaron con una deuda de préstamos estudiantil. Los datos, en miles de dólares, registran montos típicos de deuda después de la graduación.

10.1 14.8 5.0 10.2 12.4 12.2 2.0 11.5 17.8 4.0

a) Para aquellos alumnos que usan un préstamo estudiantil, ¿cuál es la deuda media después de la graduación?
b) ¿Cuál es la varianza? ¿La desviación estándar?

62. Los propietarios de pequeñas empresas con frecuencia acuden a compañías de servicios externos para manejar la nómina de sus empleados. Esto se debe a que las pequeñas empresas se enfrentan a regulaciones fiscales complicadas, y las multas por errores en las declaraciones fiscales son costosas. Según el Internal Revenue Service, 26% de todas las devoluciones de impuestos de empleo de las pequeñas empresas contenía errores que dieron como resultado una multa fiscal al propietario (*The Wall Street Journal,* 30 de enero de 2006). La multa fiscal para una muestra de 20 pequeñas empresas se presenta a continuación.

WEB archivo
Penalty

820	270	450	1 010	890	700	1 350	350	300	1 200
390	730	2 040	230	640	350	420	270	370	620

a) ¿Cuál es la multa fiscal media para las devoluciones de impuestos sobre nómina llenadas de manera inapropiada?
b) ¿Cuál es la desviación estándar?
c) ¿La multa más alta de $2 040 es una observación atípica?
d) ¿Cuáles son algunas ventajas de contratar una empresa de servicios de nómina para el propietario de una pequeña empresa a efecto de que maneje los servicios de nómina de los empleados, incluidas las devoluciones de impuestos de empleo?

63. El transporte público y el automóvil son dos medios que un empleado puede usar para ir al trabajo cada día. Las muestras de los tiempos registrados para cada método se listan enseguida; los tiempos se proporcionan en minutos.

Transporte público	28	29	32	37	33	25	29	32	41	34
Automóvil	29	31	33	32	34	30	31	32	35	33

a) Calcule el tiempo de la media muestral para ir al trabajo en cada medio.
b) Calcule la desviación estándar muestral para cada método.
c) Con base en sus resultados de los incisos *a*) y *b*), ¿cuál método de transporte debe preferirse? Explique por qué.
d) Elabore un diagrama de caja para cada método. ¿Una comparación de los diagramas apoya su conclusión del inciso *c*)?

64. La Asociación Nacional de Agentes Inmobiliarios (National Association of Realtors) informó el precio medio de la vivienda en Estados Unidos y el incremento en éste durante un periodo de cinco años (*The Wall Street Journal,* 16 de enero de 2006). Utilice los precios de la muestra de viviendas listados aquí para responder las preguntas siguientes.

WEB archivo
Homes

995.9	48.8	175.0	263.5	298.0	218.9	209.0
628.3	111.0	212.9	92.6	2 325.0	958.0	212.5

a) ¿Cuál es el la mediana del precio de la muestra de vivienda?
b) En enero de 2001, la Asociación Nacional de Agentes Inmobiliarios informó que en Estados Unidos la mediana del precio de la vivienda fue de $139 300. ¿Cuál fue el incremento del porcentaje en la mediana del precio durante el periodo de cinco años?
c) ¿Cuál es el primer y el tercer cuartil para los datos de la muestra?
d) Proporcione un resumen de cinco números para los precios de la vivienda.
e) ¿Los datos contienen algunas observaciones atípicas.
f) ¿Cuál es el precio medio de la vivienda para la muestra? ¿Por qué la Asociación Nacional de Agentes Inmobiliarios prefiere usar la mediana del precio de las casas en su informe?

65. La Encuesta de la Comunidad Estadounidense de la Oficina del Censo de Estados Unidos dio a conocer el porcentaje de niños menores de 18 años que había vivido por debajo del nivel de pobreza durante los 12 meses anteriores (sitio web del U.S. Census Bureau, agosto de 2008). Las regiones de Estados Unidos noreste (NE), sureste (SE), oeste medio (MW), suroeste (SW) oeste (W) y el porcentaje de niños menores de 18 años que había vivido por debajo del nivel de pobreza se listan para cada estado.

WEB archivo
PovertyLevel

State	Region	Poverty %	State	Region	Poverty %
Alabama	SE	23.0	Montana	W	17.3
Alaska	W	15.1	Nebraska	MW	14.4
Arizona	SW	19.5	Nevada	W	13.9
Arkansas	SE	24.3	New Hampshire	NE	9.6
California	W	18.1	New Jersey	NE	11.8
Colorado	W	15.7	New Mexico	SW	25.6
Connecticut	NE	11.0	New York	NE	20.0
Delaware	NE	15.8	North Carolina	SE	20.2
Florida	SE	17.5	North Dakota	MW	13.0
Georgia	SE	20.2	Ohio	MW	18.7
Hawaii	W	11.4	Oklahoma	SW	24.3
Idaho	W	15.1	Oregon	W	16.8
Illinois	MW	17.1	Pennsylvania	NE	16.9
Indiana	MW	17.9	Rhode Island	NE	15.1
Iowa	MW	13.7	South Carolina	SE	22.1
Kansas	MW	15.6	South Dakota	MW	16.8
Kentucky	SE	22.8	Tennessee	SE	22.7
Louisiana	SE	27.8	Texas	SW	23.9
Maine	NE	17.6	Utah	W	11.9
Maryland	NE	9.7	Vermont	NE	13.2
Massachusetts	NE	12.4	Virginia	SE	12.2
Michigan	MW	18.3	Washington	W	15.4
Minnesota	MW	12.2	West Virginia	SE	25.2
Mississippi	SE	29.5	Wisconsin	MW	14.9
Missouri	MW	18.6	Wyoming	W	12.0

a) ¿Cuál es la mediana del porcentaje del nivel de pobreza (Poverty) para los 50 estados?
b) ¿Cuáles son el primer y el tercer cuartiles? ¿Cuál es su interpretación de los cuartiles?
c) Muestre un diagrama de caja para los datos. Interprete el diagrama respecto de lo que indica acerca del nivel de pobreza para los niños de Estados Unidos. ¿Algún estado (State) se considera una observación atípica? Comente.
d) Identifique los estados en el cuartil inferior. ¿Cuál es su interpretación de este grupo y qué región o regiones se representan en este cuartil?

66. La revista *Travel + Leisure* presentó su lista anual de los 500 mejores hoteles del mundo (*Travel + Leisure,* enero de 2009). La revista proporciona una calificación para cada hotel junto con una breve descripción que incluye su tamaño, servicios y costo por noche en habitación doble. Una muestra de 12 de los hoteles de más alta calificación en Estados Unidos se presenta a continuación.

WEB archivo
Travel

Hotel	Location	Rooms	Cost/Night
Boulders Resort & Spa	Phoenix, AZ	220	499
Disney's Wilderness Lodge	Orlando, FL	727	340
Four Seasons Hotel Beverly Hills	Los Ángeles, CA	285	585
Four Seasons Hotel	Boston, MA	273	495
Hay-Adams	Washington, DC	145	495
Inn on Biltmore Estate	Asheville, NC	213	279
Loews Ventana Canyon Resort	Phoenix, AZ	398	279
Mauna Lani Bay Hotel	Isla de Hawaii	343	455
Montage Laguna Beach	Laguna Beach, CA	250	595
Sofitel Water Tower	Chicago, IL	414	367
St. Regis Monarch Beach	Dana Point, CA	400	675
The Broadmoor	Colorado Springs, CO	700	420

a) ¿Cuál es el número medio de habitaciones (Rooms)?
b) ¿Cuál es el costo medio por noche (Cost/Night) para una habitación doble?

c) Elabore un diagrama de dispersión con el número de habitaciones en el eje horizontal y el costo por noche en el eje vertical. ¿Parece haber una relación entre el número de habitaciones y el costo por noche? Comente.

d) ¿Cuál es el coeficiente de correlación muestral? ¿Qué le dice sobre la relación entre el número de habitaciones y el costo por noche para una habitación doble? ¿Esto le parece razonable? Comente.

67. Morningstar da seguimiento al rendimiento de un gran número de empresas y publica una evaluación de cada una. Junto con una variedad de datos financieros, Morningstar incluye una estimación del valor justo (Fair Value) para el precio que debe pagarse por una acción de las acciones comunes de la empresa. Los datos para 30 empresas se encuentran en el archivo llamado *FairValue*. Los datos incluyen la estimación del precio justo por acción de las acciones comunes, el precio por acción más reciente y la utilidad por acción para la empresa (*Morningstar Stocks500,* 2008).

WEB archivo
FairValue

a) Elabore un diagrama de dispersión para los datos del precio justo y del precio por acción, con este último sobre el eje horizontal. ¿Cuál es el coeficiente de correlación muestral y qué puede decir acerca de la relación entre las variables?

b) Desarrolle un diagrama de dispersión para los datos del precio justo y del precio por acción con este último sobre el eje horizontal. ¿Cuál es el coeficiente de correlación muestral y qué puede decir acerca de la relación entre las variables?

68. ¿El registro de un equipo de béisbol de ligas mayores durante el entrenamiento de primavera indica cómo jugará durante la temporada regular? En los últimos seis años el coeficiente de correlación entre el porcentaje de victorias de un equipo en el entrenamiento de primavera y su porcentaje de triunfos en la temporada regular es de 0.18 (*The Wall Street Journal,* 30 de marzo de 2009). Enseguida se listan los porcentajes de victorias para los 14 equipos de la Liga Americana durante la temporada 2008.

WEB archivo
SpringTraining

Team	Spring Training	Regular Season	Team	Spring Training	Regular Season
Baltimore Orioles	0.407	0.422	Minnesota Twins	0.500	0.540
Boston Red Sox	0.429	0.586	New York Yankees	0.577	0.549
Chicago White Sox	0.417	0.546	Oakland A's	0.692	0.466
Cleveland Indians	0.569	0.500	Seattle Mariners	0.500	0.377
Detroit Tigers	0.569	0.457	Tampa Bay Rays	0.731	0.599
Kansas City Royals	0.533	0.463	Texas Rangers	0.643	0.488
Los Ángeles Angels	0.724	0.617	Toronto Blue Jays	0.448	0.531

a) ¿Cuál es el coeficiente de correlación entre los porcentajes de victoria del entrenamiento de primavera (Spring Training) y de la temporada regular (Regular Season)?

b) ¿Qué indica su conclusión acerca del registro de un equipo durante el entrenamiento de primavera sobre cómo jugará durante la temporada regular? ¿Cuáles son algunas razones para que esto ocurra? Comente.

69. Los días para el vencimiento de una muestra de cinco fondos del mercado de dinero se listan enseguida junto con los montos en dólares de las cantidades invertidas en los fondos. Utilice la media ponderada para determinar el número medio de días para el vencimiento de los dólares invertidos en estos cinco fondos del mercado de dinero.

Días para el vencimiento	Valor monetario (millones)
20	20
12	30
7	10
5	15
6	10

70. La velocidad de los automóviles que viajan por una autopista con un límite de velocidad establecido de 55 millas por hora se comprueba mediante un sistema de radar de la policía estatal. A continuación se presenta una distribución de frecuencia de las velocidades.

Velocidad (millas por hora)	Frecuencia
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10
Total	475

a) ¿Cuál es la velocidad media de los automóviles que viajan en esta autopista?
b) Calcule la varianza y la desviación estándar.

Caso a resolver 1 Pelican Stores

Pelican Stores, una división de National Clothing, es una cadena de tiendas de ropa para dama que opera en todo Estados Unidos. La cadena lanzó recientemente una promoción en la que se enviaron cupones de descuento a los clientes de otras tiendas de National Clothing. Los datos recabados de una muestra de 100 transacciones de tarjetas de crédito en Pelican Stores durante un día, mientras la promoción estuvo vigente, se encuentran en el archivo llamado *PelicanStores*. La tabla 3.12 presenta una parte del conjunto de datos. El método de pago proprietary card se refiere a los cargos realizados usando una tarjeta de National Clothing. A los clientes (Customer)

WEB archivo
PelicanStores

TABLA 3.12 Muestra de 100 compras con tarjeta de crédito en Pelican Stores

Customer	Type of Customer	Items	Net Sales	Method of Payment	Gender	Marital Status	Age
1	Regular	1	39.50	Discover	Male	Married	32
2	Promotional	1	102.40	Proprietary card	Female	Married	36
3	Regular	1	22.50	Proprietary card	Female	Married	32
4	Promotional	5	100.40	Proprietary card	Female	Married	28
5	Regular	2	54.00	MasterCard	Female	Married	34
6	Regular	1	44.50	MasterCard	Female	Married	44
7	Promotional	2	78.00	Proprietary card	Female	Married	30
8	Regular	1	22.50	Visa	Female	Married	40
9	Promotional	2	56.52	Proprietary card	Female	Married	46
10	Regular	1	44.50	Proprietary card	Female	Married	36
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
96	Regular	1	39.50	MasterCard	Female	Married	44
97	Promotional	9	253.00	Proprietary card	Female	Married	30
98	Promotional	10	287.59	Proprietary card	Female	Married	52
99	Promotional	2	47.60	Proprietary card	Female	Married	30
100	Promotional	1	28.44	Proprietary card	Female	Married	44

que efectuaron una compra utilizando un cupón de descuento se les llama *clientes de promoción* y a los que compraron, pero no usaron un cupón de descuento se les denomina *clientes regulares*. Dado que los cupones promocionales no se enviaron a los compradores regulares de Pelican Stores, la gerencia considera las ventas realizadas a personas que presentaron los cupones de promoción como ventas que de lo contrario no se hubieran hecho. Por supuesto, Pelican también espera que los clientes de promoción sigan comprando en sus tiendas.

La mayoría de las variables mostradas en la tabla 3.12 se explican por sí mismas, pero dos requieren una aclaración.

Artículos (Items)	Número total de artículos adquiridos.
Ventas netas (Net Sales)	Monto total ($) cargado a la tarjeta de crédito.

A la gerencia de Pelican le gustaría usar estos datos muestrales para enterarse de su base de clientes y evaluar la promoción de los cupones de descuento.

Informe gerencial

Utilice los métodos tabular y gráfico de la estadística descriptiva para resumir los datos y comente sus hallazgos. Como mínimo, su informe debe incluir lo siguiente:

1. Estadísticos descriptivos sobre las ventas netas y sobre las ventas netas por varias clasificaciones de clientes.
2. Estadísticos descriptivos concernientes a la relación entre la edad (Age) y las ventas netas.

Caso a resolver 2 Industria del cine

La industria estadounidense del cine es un negocio competitivo. Más de 50 estudios producen un total de 300 a 400 películas nuevas cada año (Motion Pictures), y el éxito financiero de cada una varía considerablemente. Las ventas brutas del fin de semana de estreno (Opening Gross Sales), las ventas brutas totales (Total Gross Sales), el número de cines (Number of Theaters) donde la película se exhibe y el número de semanas que ésta estuvo entre las primeras 60 (Weeks in Top 60) en ventas brutas son variables comunes utilizadas para medir el éxito de un título. Los datos recabados de una muestra de 100 filmes producidos en 2005 se incluyen en el archivo llamado *Movies*. La tabla 3.13 muestra los datos de las primeras 10 películas de este archivo.

WEB archivo
Movies

TABLA 3.13 Datos del desempeño de 10 películas

Motion Picture	Opening Gross Sales ($millions)	Total Gross Sales ($millions)	Number of Theaters	Weeks in Top 60
Coach Carter	29.17	67.25	2574	16
Ladies in Lavender	0.15	6.65	119	22
Batman Begins	48.75	205.28	3858	18
Unleashed	10.90	24.47	1962	8
Pretty Persuasion	0.06	0.23	24	4
Fever Pitch	12.40	42.01	3275	14
Harry Potter and the Goblet of Fire	102.69	287.18	3858	13
Monster-in-Law	23.11	82.89	3424	16
White Noise	24.11	55.85	2279	7
Mr. and Mrs. Smith	50.34	186.22	3451	21

Informe gerencial

Utilice los métodos numéricos de la estadística descriptiva presentados en este capítulo para saber cómo estas variables contribuyen al éxito de una película. Incluya lo siguiente en su informe.

1. Los estadísticos descriptivos de cada una de las cuatro variables junto con un análisis de lo que cada estadístico descriptivo indica sobre la industria del cine.
2. ¿Qué películas, si las hay, deben considerarse observaciones atípicas de alto desempeño? Explique por qué.
3. La estadística descriptiva muestra la relación entre las ventas brutas totales y cada una de las otras variables. Comente.

Caso a resolver 3 Escuelas de negocios de Asia-Pacífico

WEB archivo
Asian

La consecución de un título de posgrado en los negocios es ahora internacional. Una encuesta muestra que cada vez más asiáticos eligen la ruta de la maestría en administración de empresas (MBA) para lograr el éxito corporativo. Como resultado, el número de solicitantes para los cursos de MBA en escuelas de Asia-Pacífico sigue aumentando.

En toda la región, miles de asiáticos muestran una creciente voluntad de dejar de lado temporalmente su carrera y pasar dos años en la búsqueda de un título de negocios teórico. Los cursos en estas escuelas son notoriamente difíciles e incluyen economía, banca, marketing, ciencias del comportamiento, relaciones laborales, toma de decisiones, pensamiento estratégico, derecho de los negocios, y mucho más. El conjunto de datos de la tabla 3.14 muestra algunas características de las principales escuelas de negocios de Asia-Pacífico.

Informe gerencial

Use los métodos de la estadística descriptiva para resumir los datos de la tabla 3.14. Comente sus hallazgos.

1. Incluya un resumen para cada variable del conjunto de datos. Comente e interprete con base en los máximos y los mínimos, así como los medios y las proporciones apropiados. ¿Qué elementos de comprensión nuevos proporcionan estos estadísticos descriptivos respecto de las escuelas de negocios de Asia-Pacífico?
2. Resuma los datos para comparar lo siguiente:
   a) Cualquier diferencia entre los costos de clases locales y en el extranjero.
   b) Alguna diferencia entre los sueldos iniciales medios para las escuelas que requieren y no requieren experiencia laboral.
   c) Cualquier diferencia entre los sueldos iniciales para escuelas que requieren y no requieren pruebas de inglés.
3. ¿Los sueldos iniciales parecen estar relacionados con las clases?
4. Presente resúmenes gráficos y numéricos adicionales que sean benéficos para comunicar los datos de la tabla 3.14 a otras personas.

Caso a resolver 4 Transacciones del sitio web de Heavenly Chocolates

Heavenly Chocolates fabrica y vende productos de chocolate de calidad en su planta y tienda minorista ubicada en Saratoga Springs, Nueva York. Hace dos años la empresa desarrolló un sitio web y comenzó a vender sus productos por Internet. Las ventas electrónicas han excedido las expectativas de la empresa y la gerencia ahora está considerando estrategias para incrementarlas aún más. Para saber más sobre los clientes del sitio web, se seleccionó una muestra de 50 transacciones de Heavenly Chocolate de las ventas del mes anterior. Datos que ilustran

TABLA 3.14 Datos de 25 escuelas de negocios de Asia-Pacífico

Escuela de negocios	Inscripción de tiempo completo	Estudiantes por facultad	Clases locales ($)	Clases en el extranjero ($)	Edad	Extranjero%	GMAT	Examen de inglés	Experiencia de trabajo	Sueldo inicial ($)
Melbourne Business School	200	5	24 420	29 600	28	47	Sí	No	Sí	71 400
University of New South Wales (Sydney)	228	4	19 993	32 582	29	28	Sí	No	Sí	65 200
Indian Institute of Management (Ahmedabad)	392	5	4 300	4 300	22	0	No	No	No	7 100
Chinese University of Hong Kong	90	5	11 140	11 140	29	10	Sí	No	No	31 000
International University of Japan (Niigata)	126	4	33 060	33 060	28	60	Sí	Sí	No	87 000
Asian Institute of Management (Manila)	389	5	7 562	9 000	25	50	Sí	No	Sí	22 800
Indian Institute of Management (Bangalore)	380	5	3 935	16 000	23	1	Sí	No	No	7 500
National University of Singapore	147	6	6 146	7 170	29	51	Sí	Sí	Sí	43 300
Indian Institute of Management (Calcutta)	463	8	2 880	16 000	23	0	No	No	No	7 400
Australian National University (Canberra)	42	2	20 300	20 300	30	80	Sí	Sí	Sí	46 600
Nanyang Technological University (Singapore)	50	5	8 500	8 500	32	20	Sí	No	Sí	49 300
University of Queensland (Brisbane)	138	17	16 000	22 800	32	26	No	No	Sí	49 600
Hong Kong University of Science and Technology	60	2	11 513	11 513	26	37	Sí	No	Sí	34 000
Macquarie Graduate School of Management (Sydney)	12	8	17 172	19 778	34	27	No	No	Sí	60 100
Chulalongkorn University (Bangkok)	200	7	17 355	17 355	25	6	Sí	No	Sí	17 600
Monash Mt. Eliza Business School (Melbourne)	350	13	16 200	22 500	30	30	Sí	Sí	Sí	52 500
Asian Institute of Management (Bangkok)	300	10	18 200	18 200	29	90	No	Sí	Sí	25 000
University of Adelaide	20	19	16 426	23 100	30	10	No	No	Sí	66 000
Massey University (Palmerston North, New Zealand)	30	15	13 106	21 625	37	35	No	Sí	Sí	41 400
Royal Melbourne Institute of Technology Business Graduate School	30	7	13 880	17 765	32	30	No	Sí	Sí	48 900
Jamnalal Bajaj Institute of Management Studies (Mumbai)	240	9	1 000	1 000	24	0	No	No	Sí	7 000
Curtin Institute of Technology (Perth)	98	15	9 475	19 097	29	43	Sí	No	Sí	55 000
Lahore University of Management Sciences	70	14	11 250	26 300	23	2.5	No	No	No	7 500
University Sains Malaysia (Penang)	30	5	2 260	2 260	32	15	No	Sí	Sí	16 000
De La Salle University (Manila)	44	17	3 300	3 600	28	3.5	Sí	No	Sí	13 100

TABLA 3.15 Muestra de 50 transacciones del sitio web de Heavenly Chocolates

WEB archivo
Shoppers

Customer	Day	Browser	Time (min)	Pages Viewed	Amount Spent ($)
1	Mon	Internet Explorer	12.0	4	54.52
2	Wed	Other	19.5	6	94.90
3	Mon	Internet Explorer	8.5	4	26.68
4	Tue	Firefox	11.4	2	44.73
5	Wed	Internet Explorer	11.3	4	66.27
6	Sat	Firefox	10.5	6	67.80
7	Sun	Internet Explorer	11.4	2	36.04
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48	Fri	Internet Explorer	9.7	5	103.15
49	Mon	Other	7.3	6	52.15
50	Fri	Internet Explorer	13.4	3	98.75

el día de la semana (Day) en que se realizó cada transacción, el tipo de explorador (Browser) usado por el cliente, el tiempo invertido en el sitio web (Time), el número de páginas visitadas (Pages Viewed,) y la cantidad gastada (Amount Spent) por cada uno de los 50 clientes están contenidos en el archivo llamado *Shoppers*. Una porción de los datos se muestra en la tabla 3.15.

A Heavenly Chocolates le gustaría usar los datos de la muestra para determinar si los compradores en línea que pasaron más tiempo y vieron más páginas también gastaron más dinero durante su visita al sitio web. A la empresa también le gustaría investigar el efecto que el día de la semana y el tipo de explorador tienen sobre las ventas.

Informe gerencial

Use los métodos de la estadística descriptiva para saber más acerca de los clientes que visitan el sitio web de Heavenly Chocolates. Incluya lo siguiente en su informe.

1. Resúmenes gráficos y numéricos para el tiempo que el comprador pasa en el sitio web, el número de páginas visitadas y la cantidad media gastada por transacción. Comente los datos que obtuvo acerca de los compradores en línea de Heavenly Chocolates a partir de estos resúmenes numéricos.
2. Resuma la frecuencia, los dólares totales y la cantidad media gastados por transacción para cada día de la semana. ¿Qué observaciones puede usted hacer sobre el negocio de Heavenly Chocolates con base en el día de la semana? Comente.
3. Resuma la frecuencia, los dólares totales y la cantidad media gastados por transacción para cada tipo de navegador. ¿Qué observaciones puede hacer acerca del negocio con base en el tipo de explorador? Comente.
4. Elabore un diagrama de dispersión y calcule el coeficiente de correlación muestral para explorar la relación entre el tiempo invertido en el sitio web y la cantidad gastada. Utilice el eje horizontal para el tiempo invertido. Comente.
5. Prepare un diagrama de dispersión y calcule el coeficiente de correlación muestral para explorar la relación entre el número de páginas visitadas y la cantidad gastada. Utilice el eje horizontal para el número de páginas web consultadas. Comente.
6. Elabore un diagrama de dispersión y calcule el coeficiente de correlación muestral para explorar la relación entre el tiempo pasado en el sitio web y el número de páginas visitadas. Use el eje horizontal para representar el número de páginas visitadas. Comente.

Apéndice 3.1 Estadística descriptiva usando Minitab

En este apéndice se describe cómo se usa Minitab para calcular una variedad de estadísticos descriptivos y desplegar diagramas de caja. Luego se explica su uso para obtener las medidas de covarianza y de correlación para dos variables.

Estadística descriptiva

La tabla 3.1 proporcionó los sueldos iniciales de 12 licenciados en administración de empresas recién graduados de la escuela de negocios. Estos datos están disponibles en el archivo Start-Salary. La figura 3.12 muestra la estadística descriptiva de los datos de los sueldos iniciales obtenidos con Minitab. Las definiciones de los encabezados se muestran en seguida.

N	Número de valores de datos
N*	Número de valores de datos faltantes
Mean	Media
SE Mean	Error estándar de la media
StDev	Desviación estándar
Minimum	Valor de datos mínimo
Q1	Primer cuartil
Median	Mediana
Q3	Tercer cuartil
Maximum	Valor de datos máximo

La etiqueta SE Mean se refiere al *error estándar de la media.* Se calcula dividiendo la desviación estándar entre la raíz cuadrada de *N*. La interpretación y el uso de esta medida se estudian en el capítulo 7, cuando se presentan los temas de muestreo y distribuciones del muestreo.

Aunque las medidas numéricas del rango, el rango intercuartílico, la varianza y el coeficiente de variación no aparecen en el resultado de Minitab, estos valores se calculan fácilmente a partir de los resultados de la figura 3.12 como sigue.

$$\text{Rango} = \text{máximo} - \text{mínimo}$$
$$\text{RIC} = Q_3 - Q_1$$
$$\text{Varianza} = (\text{StDev})^2$$
$$\text{Coeficiente de variación} = (\text{StDev/Mean}) \times 100$$

Por último, observe que los cuartiles de Minitab $Q_1 = 3\,457.5$ y $Q_3 = 3\,625$ son ligeramente diferentes de los cuartiles $Q_1 = 3\,465$ y $Q_3 = 3\,600$ calculados en la sección 3.1. Las distintas convenciones* que se usaron para identificar los cuartiles explican esta variación. Por consiguiente, los valores Q_1 y Q_3 proporcionados por una convención tal vez no sean idénticos a los derivados de otra convención. No obstante, cualesquiera diferencias tienden a ser insignificantes

FIGURA 3.12 Estadísticos descriptivos proporcionados por Minitab

N	N*	Mean	SE Mean	StDev
12	0	3540.0	47.8	165.7
Minimum	**Q1**	**Median**	**Q3**	**Maximum**
3310.0	3457.5	3505.0	3625.0	3925.0

* Con las n observaciones arregladas en orden ascendente (del valor menor al valor mayor), Minitab usa las posiciones dadas por $(n + 1)/4$ y $3(n + 1)/4$ para ubicar a Q_1 y Q_3, respectivamente. Cuando una posición es fraccional, Minitab hace una interpolación entre los dos valores de datos ordenados adyacentes para determinar el cuartil correspondiente.

y los resultados proporcionados no deben inducir al usuario a errores al hacer las interpretaciones usuales asociadas con los cuartiles.

Enseguida se explicará cómo se generan los estadísticos de la figura 3.12. Los datos de los sueldos iniciales están en la columna C2 de la hoja de trabajo de StartSalary. Los pasos siguientes guían para generar los estadísticos descriptivos.

WEB archivo
StartSalary

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **Display Descriptive Statistics**.
Paso 4. Cuando el cuadro de diálogo Display Descriptive Statistics aparezca:
Introduzca C2 en el cuadro **Variables**.
Haga clic en **OK**.

Diagrama de caja

Los pasos siguientes usan el archivo StartSalary para generar el diagrama de caja sobre los datos de los sueldos iniciales.

Paso 1. Seleccione el menú **Graph**.
Paso 2. Elija **Boxplot**.
Paso 3. Seleccione **Simple** y haga clic en **OK**.
Paso 4. Cuando aparezca el cuadro de diálogo Boxplot-One Y, Simple:
Introduzca C2 en el cuadro **Graph variables**.
Haga clic en **OK**.

Covarianza y correlación

WEB archivo
Stereo

La tabla 3.6 proporciona el número de comerciales y el volumen de ventas de una tienda de estéreos y equipos de sonido. Estos datos están disponibles en el archivo *Stereo*; el número de comerciales se encuentra en la columna C2 y el volumen de ventas en la columna C3. Los pasos siguientes muestran cómo se usa Minitab para calcular la covarianza de las dos variables.

Paso 1. Seleccione el menú **Stat**.
Paso 2. Elija **Basic Statistics**.
Paso 3. Elija **Covariance**.
Paso 4. Cuando el cuadro de diálogo Covariance aparezca:
Introduzca C2 C3 en el cuadro **Variables**.
Haga clic en **OK**.

Para obtener el coeficiente de correlación del número de comerciales y el volumen de ventas sólo es necesario realizar un cambio en el procedimiento anterior. En el paso 3 elija la opción **Correlation**.

Apéndice 3.2 Estadística descriptiva usando Excel

Excel se puede utilizar para generar los estadísticos descriptivos de este capítulo. En este apéndice se explica cómo se usa para obtener varias medidas de posición y variabilidad para una sola variable, así como la covarianza y el coeficiente de correlación como medidas de asociación entre dos variables.

Uso de las funciones de Excel

Excel proporciona funciones para calcular la media, la mediana, la moda, la varianza muestral y la desviación estándar de la muestra. El uso de estas funciones se explica mediante el cálculo

FIGURA 3.13 Uso de las funciones de Excel para calcular la media, mediana, moda y desviación estándar

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

	A	B	C	D	E	F
1	**Graduate**	**Starting Salary**		**Mean**	3540	
2	1	3450		**Median**	3505	
3	2	3550		**Mode**	3480	
4	3	3650		**Variance**	27440.91	
5	4	3480		**Standard Deviation**	165.65	
6	5	3355				
7	6	3310				
8	7	3490				
9	8	3730				
10	9	3540				
11	10	3925				
12	11	3520				
13	12	3480				
14						

WEB archivo
StartSalary

de la media, la mediana, la varianza muestral y la desviación estándar muestral de los datos de los sueldos iniciales de la tabla 3.1. Vuelva a observar la figura 3.13 mientras se describen los pasos involucrados. Los datos se introducen en la columna B.

La función AVERAGE de Excel se usa para calcular la media al introducir la fórmula siguiente en la celda E1.

=AVERAGE(B2:B13)

De modo parecido, las fórmulas =MEDIAN(B2:B13), =MODE(B2:B13), =VAR(B2:B13) y =STDEV(B2:B13) se introducen en las celdas E2:E5, respectivamente, para calcular la mediana, la moda, la varianza y la desviación estándar. La hoja de trabajo en segundo plano muestra que los valores estimados con las funciones de Excel son los mismos que aquellos calculados antes en el capítulo.

Excel proporciona también funciones para calcular la covarianza y el coeficiente de correlación. Debe tener cuidado cuando las use debido a que la función de covarianza trata los datos como una población y la función de correlación los trata como una muestra. Por tanto, el resultado obtenido usando la función de covarianza de Excel debe ajustarse para proporcionar la covarianza muestral. Enseguida se explica cómo usar estas funciones para calcular la covarianza muestral y el coeficiente de correlación muestral para los datos de la tienda de estéreos y equipos de sonido de la tabla 3.7. Vuelva a observar la figura 3.14 mientras se presentan los pasos involucrados.

WEB archivo
Stereo

La función de covarianza de Excel, COVAR, sirve para calcular la covarianza poblacional al introducir la fórmula siguiente en la celda F1.

=COVAR(B2:B11,C2:C11)

De manera similar, la fórmula =CORREL(B2:B11,C2:C11) se introduce en la celda F2 para calcular el coeficiente de correlación muestral. La hoja de trabajo en segundo plano muestra los

FIGURA 3.14 Uso de las funciones de Excel para calcular la covarianza y la correlación

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	=COVAR(B2:B11,C2:C11)	
2	1	2	50		Sample Correlation	=CORREL(B2:B11,C2:C11)	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

	A	B	C	D	E	F	G
1	Week	Commercials	Sales		Population Covariance	9.90	
2	1	2	50		Sample Correlation	0.93	
3	2	5	57				
4	3	1	41				
5	4	3	54				
6	5	4	54				
7	6	1	38				
8	7	5	63				
9	8	3	48				
10	9	4	59				
11	10	2	46				
12							

valores estimados usando las funciones de Excel. Observe que el valor del coeficiente de correlación muestral (0.93) es el mismo que se calculó usando la ecuación (3.12). Sin embargo, el resultado proporcionado por la función COVAR de Excel, 9.9, se obtuvo al tratar los datos como una población. Por tanto, este resultado debe ajustarse para obtener la covarianza muestral. El ajuste es muy sencillo. Primero note que la fórmula de la covarianza poblacional, la ecuación (3.11), requiere que se divida entre el número total de observaciones en el conjunto de datos, pero la fórmula para la covarianza muestral, la ecuación (3.10), requiere que se divida entre el número total de observaciones menos 1. Por tanto, para usar el resultado de Excel de 9.9 a efecto de calcular la covarianza muestral, sencillamente se multiplica 9.9 por $n/(n-1)$. Como $n = 10$, se obtiene

$$s_{xy} = \left(\frac{10}{9}\right)9.9 = 11$$

Por tanto, la covarianza muestral de los datos de la tienda de estéreos y equipos de sonido es 11.

Uso de la herramienta Descriptive Statistics de Excel

WEB archivo
StartSalary

Como ya se demostró, Excel proporciona funciones estadísticas para calcular los estadísticos descriptivos de un conjunto de datos. Estas funciones se usan para determinar un estadístico a la vez (por ejemplo, la media, la varianza, etc.). Excel también cuenta con una variedad de herramientas para análisis de datos. Una de estas herramientas, llamada Descriptive Statistics, permite al usuario calcular una variedad de estadísticos descriptivos en una sola operación. Enseguida se explica cómo usar esta herramienta para calcular los estadísticos descriptivos de los datos de los sueldos iniciales de la tabla 3.1.

Paso 1. Haga clic en la ficha **Data** de la cinta de opciones.
Paso 2. En el grupo **Analysis** haga clic en **Data Analysis**.
Paso 3. Cuando el cuadro de diálogo Data Analysis aparezca:
Elija **Descriptive Statistics**.
Haga clic en **OK**.

FIGURA 3.15 Resultado de la herramienta Descriptive Statistics de Excel

	A	B	C	D	E	F
1	**Graduate**	**Starting Salary**		***Starting Salary***		
2	1	3450				
3	2	3550		**Mean**	3540	
4	3	3650		Standard Error	47.82	
5	4	3480		**Median**	3505	
6	5	3355		**Mode**	3480	
7	6	3310		**Standard Deviation**	165.65	
8	7	3490		**Sample Variance**	27440.91	
9	8	3730		Kurtosis	1.7189	
10	9	3540		**Skewness**	1.0911	
11	10	3925		**Range**	615	
12	11	3520		**Minimum**	3310	
13	12	3480		**Maximum**	3925	
14				**Sum**	42480	
15				**Count**	12	
16						

Paso 4. Cuando el cuadro de diálogo Descriptive Statistics aparezca:
Introduzca B1:B13 en el cuadro **Input Range**.
Seleccione **Grouped By Columns**.
Elija **Labels in First Row**.
Seleccione **Output Range**.
Introduzca D1 en el cuadro **Output Range** (para identificar la esquina superior izquierda de la sección de la hoja de trabajo donde aparecerá el estadístico descriptivo).
Seleccione **Summary statistics**.
Haga clic en **OK**.

Las celdas D1:E15 de la figura 3.15 muestran los estadísticos descriptivos proporcionados por Excel. Las entradas en negritas son los estadísticos descriptivos que se cubren en este capítulo. Los que no están en negritas se cubren después en el libro o se estudian en libros más avanzados.

Apéndice 3.3 Estadística descriptiva usando StatTools

En este apéndice se describe cómo se utiliza StatTools para calcular una variedad de estadísticos descriptivos y desplegar diagramas de caja. Luego se muestra cómo se usa StatTools para obtener las medidas de covarianza y correlación para dos variables.

Estadística descriptiva

WEB archivo
StartSalary

Los datos de los sueldos iniciales de la tabla 3.1 son útiles para ilustrar. Primero se verá el uso de Data Set Manager con el fin de crear un conjunto de datos StatTools para estos datos usando el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes generarán una variedad de estadísticos descriptivos.

Paso 1. Haga clic en la ficha **StatTools** de la cinta de opciones.
Paso 2. En **Analyses Group,** haga clic en **Summary Statistics**.
Paso 3. Elija la opción **One-Variable Summary**.

Paso 4. Cuando el cuadro de diálogo One-Variable Summary Statistics se abra:
En la sección **Variables** seleccione **Starting Salary**.
Haga clic en **OK**.

Aparecerá una variedad de estadísticos descriptivos.

Diagramas de caja

Los datos de los sueldos iniciales de la tabla 3.1 se usan para ilustrar. Primero se utiliza Data Set Manager con el fin crear un conjunto de datos StatTools para estos datos mediante el procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes crearán un diagrama de caja para estos datos.

WEB archivo
StartSalary

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En **Analyses Group**, haga clic en **Summary Graphs**.
Paso 3. Elija la opción **Box-Whisker Plot**.
Paso 4. Cuando el cuadro de diálogo StatTools–Box–Whisker Plot aparezca:
En la sección **Variables** seleccione **Starting Salary**.
Haga clic en **OK**.

El símbolo □ se usa para identificar una observación atípica, y x para identificar la media.

Covarianza y correlación

Utilizamos los datos de la tienda de estéreos y equipos de sonido de la tabla 3.7 para demostrar el cálculo de la covarianza muestral y el coeficiente de correlación muestral. Primero se usa Data Set Manager con el fin crear un conjunto de datos StatTools para estos datos por medio del procedimiento descrito en el apéndice del capítulo 1. Los pasos siguientes proporcionarán la covarianza muestral y el coeficiente de correlación muestral.

WEB archivo
Stereo

Paso 1. Haga clic en la ficha **StatTools** en la cinta de opciones.
Paso 2. En **Analyses Group,** haga clic en **Summary Statistics**.
Paso 3. Elija la opción **Correlation and Covariance**.
Paso 4. Cuando el cuadro de diálogo StatTools–Correlation and Covariance aparezca:
En la sección **Variables**:
Elija **No. of Commercials**.
Seleccione **Sales Volume**.
En la sección **Tables to Create**:
Seleccione **Table of Correlations**.
Elija **Table of Covariances**.
En la sección **Table Structure**, seleccione **Symmetric**.
Haga clic en **OK**.

Una tabla muestra el coeficiente de correlación y la covarianza aparecerá.

CAPÍTULO 4

Introducción a la probabilidad

CONTENIDO

ESTADÍSTICA *en* LA PRÁCTICA

OCEANWIDE SEAFOOD*

SPRINGBORO, OHIO

Oceanwide Seafood es el principal proveedor de pescado y mariscos de calidad del suroeste de Ohio. La empresa vende más de 90 variedades de mariscos frescos y congelados de todo el mundo y prepara cortes especiales según las especificaciones de sus clientes, que incluyen los principales restaurantes y minoristas de alimentos en Ohio, Kentucky e Indiana. La empresa, fundada en 2005, ha logrado tener éxito al proporcionar un excelente servicio al cliente y mariscos de calidad excepcional.

La probabilidad y la información estadística se utilizan para la toma de decisiones operativas y de marketing. Por ejemplo, para seguir la pista del crecimiento de la empresa y establecer los futuros niveles meta de ventas, se utiliza una serie de tiempo que muestra las ventas mensuales. Estadísticos como el tamaño medio de los pedidos del cliente y el número medio de días que tarda en hacer los pagos ayudan a identificar a los mejores clientes de la empresa, así como a proporcionar puntos de referencia para el manejo de los problemas de las cuentas por cobrar. Además, los datos sobre los niveles mensuales de inventario se usan en el análisis de la utilidad de operación y las tendencias en las ventas de productos.

El análisis de probabilidad ha ayudado a Oceanwide a determinar precios razonables y rentables para sus productos. Por ejemplo, cuando recibe un pescado entero fresco de uno de sus proveedores, éste se procesa y corta para cumplir con los pedidos de cada cliente. Un atún entero fresco de 100 libras conservado en hielo podría costarle a Oceanwide \$500. A simple vista, el costo para la empresa parece ser \$500/100 = \$5 por libra. Sin embargo, debido a la pérdida en la operación de procesamiento y corte, un atún entero de 100 libras no proporcionará 100 libras de producto terminado. Si la operación de procesamiento y corte produce 75% del atún entero, el número de libras de producto terminado disponible para vender a los clientes sería 0.75(100) = 75 libras, no 100 libras. En este caso, el costo real del atún para la empresa sería \$500/75 = \$6.67 por libra. Por tanto, Oceanwide necesitaría determinar un costo de \$6.67 por libra para que el precio que fija a sus clientes sea rentable.

El atún de aleta azul se envía a Oceanwide Seafood casi todos los días. © Gregor Kervina, 2009/Fotografía usada con autorización de Shutterstock.com.

Para ayudar a determinar el porcentaje del rendimiento probable del procesamiento y corte de atún entero, se recabaron datos sobre el rendimiento de una muestra del producto entero. La variable y denota el porcentaje de rendimiento del producto. Utilizando los datos, Oceanwide pudo determinar que 5% de las veces dicho rendimiento fue por lo menos de 90%. En la notación de probabilidad condicional, ésta se escribe $P(Y \geq 90\% \mid \text{atún}) = 0.05$; es decir, la probabilidad de que el rendimiento sea por lo menos de 90%, teniendo en cuenta que el pescado es un atún, es 0.05. Si Oceanwide estableció el precio de venta del producto sobre la base de un rendimiento de 90%, la empresa obtendrá un rendimiento menor al esperado 95% de las veces. Como resultado, estaría subestimando su costo por libra y también el precio para sus clientes. Otra información de probabilidad condicional para otros porcentajes de rendimiento ayudaron a la gerencia a seleccionar un rendimiento de 70% como base para determinar el costo del atún y el precio que fija para sus clientes. Probabilidades condicionales parecidas sobre otros productos del mar permitieron establecer porcentajes para fijar precios por rendimiento para cada tipo de producto del mar. En este capítulo usted aprenderá a calcular e interpretar las probabilidades condicionales y otras más que son útiles en el proceso de toma de decisiones.

* Los autores agradecen a Dale Hartlage, presidente de Oceanwide Seafood Company, por proporcionar este artículo para la sección Estadística en la práctica.

Los gerentes o administradores suelen basar sus decisiones en un análisis de incertidumbre como los siguientes:

1. ¿Qué posibilidades hay de que las ventas disminuyan si los precios aumentan?
2. ¿Cuál es la probabilidad de que un nuevo método de ensamble mejore la productividad?
3. ¿Qué tan probable es que este proyecto se complete a tiempo?
4. ¿Qué posibilidad hay de que una nueva inversión sea rentable?

Algunos de los primeros trabajos sobre probabilidad tuvieron su origen en una serie de cartas entre Pierre de Fermat y Blaise Pascal en la década de 1650.

La **probabilidad** es una medida numérica de la posibilidad de que un evento ocurra. Por tanto, se utiliza como una medida del grado de incertidumbre asociado con cada uno de los cuatro eventos previamente listados. Si las probabilidades están disponibles, se puede determinar la posibilidad de ocurrencia de cada evento.

Los valores de probabilidad siempre se asignan en una escala de 0 a 1. Una probabilidad cercana a 0 indica que es poco probable que un evento ocurra, una probabilidad cercana a 1 indica que es casi seguro que un evento se produzca. Otras probabilidades entre 0 y 1 representan grados de posibilidad de que un evento ocurra. Por ejemplo, si se considera el evento "lluvia para mañana", se entiende que cuando el informe del clima indica "una probabilidad de lluvia casi nula", significa que la posibilidad de lluvia es muy baja. Sin embargo, si se informa una probabilidad de 0.90 de que llueva, es probable que llueva. Una medida de 0.50 indica que la probabilidad de que llueva es igual a la de que no llueva. La figura 4.1 representa el punto de vista de la probabilidad como una medida numérica de la posibilidad de que un evento ocurra.

4.1 Experimentos, reglas de conteo y asignación de probabilidades

En el estudio de la probabilidad, un **experimento** se define como un proceso que genera resultados bien definidos. En cada repetición ocurre uno y sólo uno de los resultados posibles del experimento. En seguida se listan varios ejemplos de experimentos y sus resultados correspondientes.

Experimento	Resultados del experimento
Lanzar una moneda	Cara, cruz
Seleccionar una parte para inspeccionarla	Defectuosa, sin defectos
Hacer una llamada de ventas	Comprar, no comprar
Arrojar un dado	1, 2, 3, 4, 5, 6
Jugar un partido de futbol americano	Ganar, perder, empatar

Cuando se especifican todos los resultados posibles del experimento, el **espacio muestral** de éste queda definido.

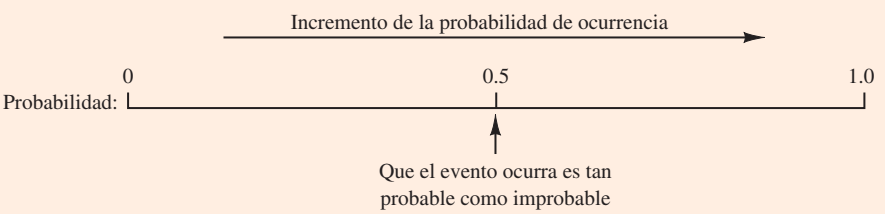
ESPACIO MUESTRAL

El espacio muestral es el conjunto de todos los resultados del experimento.

A los resultados del experimento también se les llama puntos de la muestra.

Un resultado experimental también se conoce como **punto de la muestra** para identificarlo como un elemento del espacio muestral.

FIGURA 4.1 La probabilidad como una medida numérica de la posibilidad de que un evento ocurra



Considere el primer experimento de la tabla anterior, es decir, el lanzamiento de una moneda. La cara que cae hacia arriba, ya sea cara o cruz, determina los resultados del experimento (puntos de la muestra). Si S denota el espacio muestral, se utiliza la siguiente notación para describirlo.

$$S = \{\text{cara, cruz}\}$$

El espacio muestral para el segundo experimento de la tabla, en el que se selecciona una parte para inspeccionarla, se describe como sigue:

$$S = \{\text{defectuosa, sin defectos}\}$$

Los dos ejemplos que se acaban de describir tienen dos resultados del experimento (puntos de la muestra). Sin embargo, suponga que se considera el cuarto caso listado en la tabla: el tiro de un dado. Los resultados del experimento posibles, que se definen como el número de puntos que tiene la cara superior del dado, son los seis puntos del espacio muestral de este experimento.

$$S = \{1, 2, 3, 4, 5, 6\}$$

Reglas de conteo, combinaciones y permutaciones

La identificación y el conteo de los resultados del experimento es un paso necesario en la asignación de probabilidades. Ahora se estudiarán tres reglas de conteo útiles.

Experimentos de pasos múltiples La primera regla de conteo se aplica a los experimentos de pasos múltiples. Considere un experimento que consiste en lanzar dos monedas. Los resultados se definen en función del patrón de caras y cruces que muestra la cara superior de las dos monedas. ¿Cuántos resultados son posibles para este experimento? El ejemplo de lanzar dos monedas se considera un experimento de dos pasos en el cual el paso 1 es el lanzamiento de la primera moneda y el paso 2 el lanzamiento de la segunda. Si se utiliza H para denotar una cara y T para una cruz, (H, H) indica el resultado experimental en el que hay una cara en la primera moneda y una cara en la segunda. Siguiendo esta notación, el espacio muestral (S) para este experimento se describe como sigue:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Por tanto, hay cuatro resultados experimentales posibles. En este caso, es fácil listarlos todos.

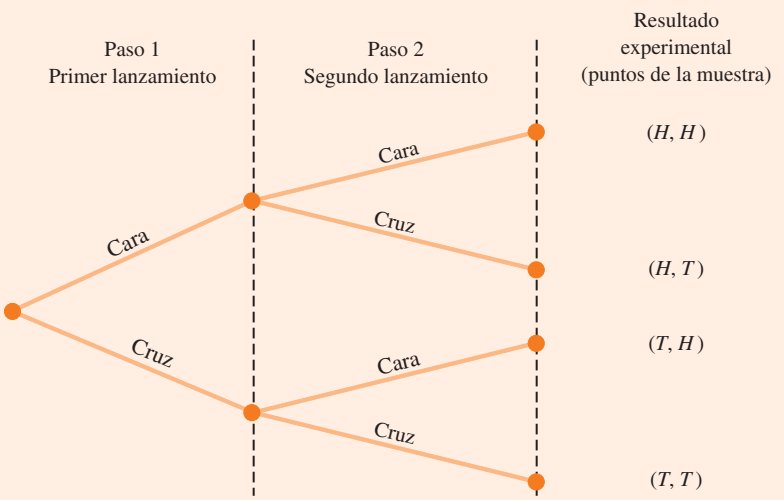
La regla de conteo para experimentos de pasos múltiples permite determinar el número de resultados del experimento sin listarlos.

REGLA DE CONTEO PARA EXPERIMENTOS DE PASOS MÚLTIPLES

Si un experimento se describe como una secuencia de k pasos con n_1 resultados posibles en el primer paso, n_2 resultados posibles en el segundo paso, y así sucesivamente, el número total de resultados del experimento está dado por $(n_1)\,(n_2)\,\ldots\,(n_k)$.

Si se considera el experimento del lanzamiento de dos monedas como la secuencia de lanzar primero una moneda $(n_1 = 2)$ y luego la otra $(n_2 = 2)$, al aplicar la regla de conteo puede verse que $(2)(2) = 4$, por lo que hay cuatro resultados experimentales distintos posibles. Como se mostró, estos resultados son $S = \{(H, H), (H, T), (T, H), (T, T)\}$. El número de resultados en un experimento que consiste en lanzar seis monedas es $(2)(2)(2)(2)(2)(2) = 64$.

FIGURA 4.2 Diagrama de árbol para el experimento del lanzamiento de dos monedas



Sin el diagrama de árbol, podría pensarse que hay sólo tres resultados experimentales posibles para dos lanzamientos de una moneda: 0 caras, 1 cara y 2 caras.

Un **diagrama de árbol** es una representación gráfica que ayuda a visualizar un experimento de pasos múltiples. La figura 4.2 muestra un diagrama de árbol para el experimento del lanzamiento de dos monedas. La secuencia de pasos va de izquierda a derecha a través del árbol. El paso 1 corresponde al lanzamiento de la primera moneda y el paso 2, al lanzamiento de la segunda. En cada paso, los dos resultados posibles son cara o cruz. Observe que a cada resultado posible del paso 1 le corresponden las dos ramas de los dos resultados posibles del paso 2. Cada uno de los puntos en el extremo derecho del árbol representa un resultado experimental. Cada trayectoria que recorre por el árbol desde el nodo que está en el extremo izquierdo hasta uno de los nodos en el extremo derecho es una secuencia única de resultados.

Ahora se explicará cómo se utilizan la reglas de conteo para experimentos de pasos múltiples mediante el análisis de un proyecto de expansión de Kentucky Power & Light Company (KP&L), el cual tiene la finalidad de incrementar la capacidad de generación de una de sus plantas en el norte de Kentucky. El proyecto está dividido en dos etapas o pasos secuenciales: etapa 1 (diseño) y etapa 2 (construcción). Aun cuando cada una se programará y controlará lo más detalladamente posible, la gerencia no puede predecir el tiempo exacto requerido para completar cada etapa. Un análisis de proyectos de construcción similares reveló que la duración posible de la etapa de diseño sería de 2, 3 o 4 meses y la duración probable de la fase de construcción sería de 6, 7 u 8 meses. Además, debido a la necesidad apremiante de tener más electricidad, la gerencia fijó una meta de 10 meses para completar todo el proyecto.

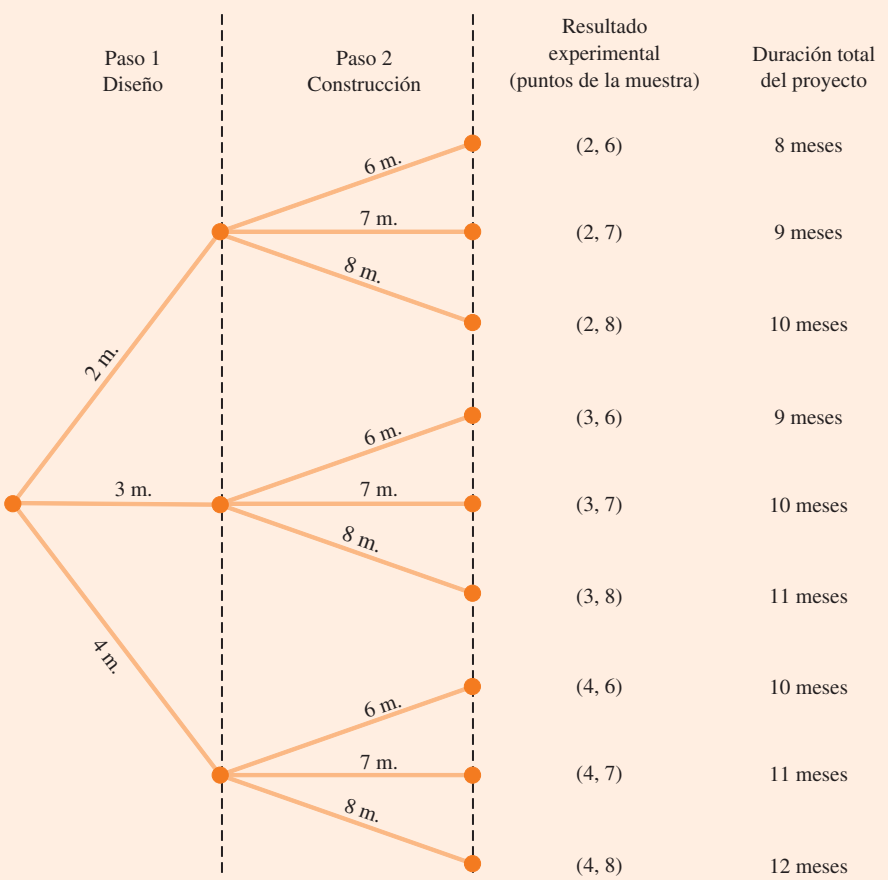
Como este proyecto tiene tres tiempos de terminación posibles para la etapa de diseño (paso 1) y tres tiempos de terminación posibles para la de construcción (paso 2), se aplica la regla de conteo para los experimentos de pasos múltiples para determinar un total de $(3)(3) = 9$ resultados del experimento. Para describir dichos resultados se utiliza una notación de dos números; por ejemplo, (2, 6) indica que la etapa de diseño se completará en 2 meses y la de construcción en 6. Este resultado experimental implica un total de $2 + 6 = 8$ meses para completar todo el plan. La tabla 4.1 resume los nueve resultados del experimento del problema de KP&L. El diagrama de árbol de la figura 4.3 muestra cómo ocurren los nueve resultados (puntos de la muestra).

La regla de conteo y el diagrama de árbol ayudan al gerente de proyectos a identificar los resultados del experimento y a determinar la duración posible del proyecto. A partir de la

TABLA 4.1 Resultados del experimento (puntos de la muestra) del proyecto de KP&L

Duración (meses)			
Etapa 1 Diseño	**Etapa 2 Construcción**	**Notación para resultados del experimento**	**Duración total del proyecto (meses)**
2	6	(2, 6)	8
2	7	(2, 7)	9
2	8	(2, 8)	10
3	6	(3, 6)	9
3	7	(3, 7)	10
3	8	(3, 8)	11
4	6	(4, 6)	10
4	7	(4, 7)	11
4	8	(4, 8)	12

FIGURA 4.3 Diagrama de árbol del proyecto de KP&L



información de la figura 4.3 se ve que éste durará de 8 a 12 meses, y que seis de los nueve resultados del experimento tienen la duración deseada de 10 meses o menos. Aun cuando la identificación de los resultados del experimento puede parecer útil, es necesario considerar cómo se asignan los valores de probabilidad a dichos resultados antes de evaluar la probabilidad de que el proyecto se complete dentro de los 10 meses deseados.

Combinaciones Una segunda regla de conteo útil permite contar el número de resultados cuando el experimento consiste en la selección de n objetos de un conjunto (generalmente mayor) de N objetos. Ésta se conoce como *regla de conteo para combinaciones*.

REGLA DE CONTEO PARA COMBINACIONES

El número de combinaciones de N objetos tomados n a la vez es

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!} \qquad (4.1)$$

donde

$$N! = N(N-1)(N-2)\cdots(2)(1)$$
$$n! = n(n-1)(n-2)\cdots(2)(1)$$

y, por definición,

$$0! = 1$$

En el muestreo de una población finita de tamaño N, la regla de conteo para combinaciones ayuda a determinar el número de muestras diferentes de tamaño n que pueden seleccionarse.

La notación ! significa *factorial;* por ejemplo, 5 factorial es 5! = (5)(4)(3)(2)(1) = 120.

Como ejemplo del uso de la regla de conteo para combinaciones, considere un procedimiento de control de calidad en el cual un inspector selecciona al azar de dos a cinco partes para buscar defectos. En un grupo de cinco partes, ¿cuántas combinaciones de dos partes pueden seleccionarse? La regla de conteo de la ecuación (4.1) muestra que con $N = 5$ y $n = 2$; tenemos

$$C_2^5 = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{(5)(4)(3)(2)(1)}{(2)(1)(3)(2)(1)} = \frac{120}{12} = 10$$

Por tanto, 10 resultados son posibles para el experimento de selección de dos partes al azar de un grupo de cinco. Si las cinco partes se etiquetan como A, B, C, D y E, las 10 combinaciones o resultados del experimento son AB, AC, AD, AE, BC, BD, BE, CD, CE y DE.

Como otro ejemplo, considere el sistema de lotería de Florida que utiliza la selección al azar de seis enteros de un grupo de 53 para determinar al ganador de la semana. La regla de conteo para combinaciones, la ecuación (4.1), se utiliza para determinar el número de maneras en que seis enteros diferentes pueden seleccionarse de un grupo de 53.

$$\binom{53}{6} = \frac{53!}{6!(53-6)!} = \frac{53!}{6!47!} = \frac{(53)(52)(51)(50)(49)(48)}{(6)(5)(4)(3)(2)(1)} = 22\,957\,480$$

La regla de conteo para combinaciones muestra que el evento de ganar la lotería es muy poco probable.

La regla de conteo para combinaciones establece que casi 23 millones de resultados experimentales son posibles en el sorteo de la lotería. Una persona que compra un billete tiene 1 oportunidad en 22 957 480 de ganar.

Permutaciones Una tercera regla de conteo que en ocasiones es útil es la regla de conteo para permutaciones. Ésta permite que una persona calcule el número de resultados experimentales cuando se seleccionan n objetos de un conjunto de N objetos y el orden de selección es

importante. Los mismos n objetos seleccionados en un orden distinto se consideran un resultado experimental diferente.

REGLA DE CONTEO PARA PERMUTACIONES

El número de permutaciones de N objetos tomados n a la vez está dado por

$$P_n^N = n!\binom{N}{n} = \frac{N!}{(N-n)!} \tag{4.2}$$

La regla de conteo para permutaciones se relaciona estrechamente con la regla de conteo para combinaciones; sin embargo, un experimento produce más permutaciones que combinaciones para el mismo número de objetos debido a que cada selección de n objetos se ordena de $n!$ maneras distintas.

Como ejemplo, considere de nuevo el proceso de control de calidad en el que un inspector selecciona dos de cinco partes distintas para inspeccionarlas en busca de defectos. ¿Cuántas permutaciones pueden seleccionarse? La regla de conteo de la ecuación (4.2) muestra que con $N = 5$ y $n = 2$ se tiene

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = \frac{(5)(4)(3)(2)(1)}{(3)(2)(1)} = \frac{120}{6} = 20$$

Por tanto, hay 20 resultados posibles para el experimento de seleccionar dos partes al azar de un grupo de cinco cuando se toma en cuenta el orden de selección. Si las partes se etiquetan como A, B, C, D y E, las 20 permutaciones son AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE y ED.

Asignación de probabilidades

Ahora se explicará cómo asignar las probabilidades a los resultados del experimento. Los enfoques de tres pasos más usuales son el método clásico, el de frecuencia relativa y el subjetivo. Sea cual fuere el método empleado, se deben cumplir dos **requisitos básicos para la asignación de probabilidades**.

REQUISITOS BÁSICOS PARA LA ASIGNACIÓN DE PROBABILIDADES

1. La probabilidad asignada a cada resultado experimental debe estar entre 0 y 1, inclusive. Si E_i denota el i-ésimo resultado del experimento y $P(E_i)$ su probabilidad, entonces este requisito se escribe como

$$0 \leq P(E_i) \leq 1 \text{ para toda } i \tag{4.3}$$

2. La suma de las probabilidades para todos los resultados del experimento debe ser igual a 1. Para n resultados, este requisito se escribe como

$$P(E_1) + \mathrm{P(E_2)} + \cdots + P(E_n) = 1 \tag{4.4}$$

El **método clásico** de asignación de probabilidades es apropiado cuando todos los resultados del experimento son igualmente probables. Si n resultados son posibles, una probabilidad de $1/n$ se asigna a cada resultado experimental. Cuando se utiliza este método, los dos requisitos básicos para la asignación de probabilidades se cumplen de manera automática.