3.1 Medidas de localización

Media

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

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

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

MEDIA MUESTRAL

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

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

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

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

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

46 54 42 46 32

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

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

Por tanto, para calcular la media muestral, escriba

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

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

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

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

archivo en CD
StartSalary

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

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

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

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

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

MEDIA POBLACIONAL

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

Mediana

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

MEDIANA

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

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

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

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

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

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

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

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

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

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

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

Moda

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

MODA

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

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

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

Percentiles

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

PERCENTIL

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

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

Para calcular el percentil p se emplea el procedimiento siguiente.

CÁLCULO DEL PERCENTIL p

Seguir estos pasos facilita el cálculo de los percentiles.

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

Paso 2. Calcular el índice i

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

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

Paso 3. (a) Si *i no es un número entero, debe redondearlo.* El primer entero mayor que i denota la posición del percentil p.
(b) Si *i es un número entero,* el percentil p es el promedio de los valores en las posiciones i e $i + 1$.

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

Paso 1. Ordenar los datos de menor a mayor

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

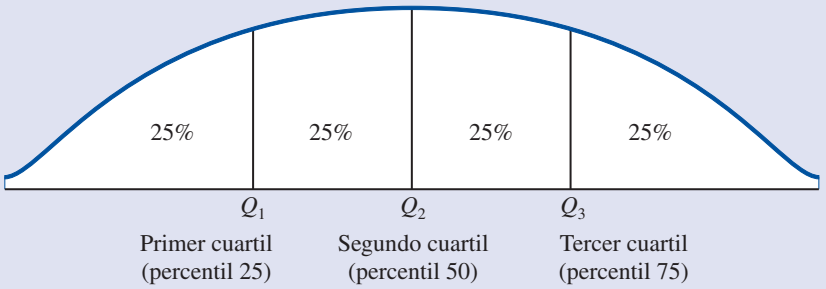
Paso 2.

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

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

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

FIGURA 3.1 LOCALIZACIÓN DE LOS CUARTILES



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

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

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

Cuartiles

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

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

Q_1 = primer cuartil, o percentil 25
Q_2 = segundo cuartil, o percentil 50
Q_3 = tercer cuartil, o percentil 75

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

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

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

Para hallar Q_1,

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

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

Para hallar Q_3,

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

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

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

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

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

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

NOTAS Y COMENTARIOS

Cuando el conjunto de datos contiene valores extremos, es preferible usar la mediana que la media como unidad de localización central. Otra medida que suele ser usada cuando hay valores extremos es la *media recortada*. La media recortada se obtiene eliminando del conjunto de datos un determinado porcentaje de los valores menores y mayores y calculando después la media de los valores restantes. Por ejemplo, la media recortada a 5% se obtiene eliminando el 5% menor y el 5% mayor de los valores y calculando después la media de los valores restantes. Con la muestra de los 12 sueldos iniciales, $0.05(12) = 0.6$. Redondear este valor a 1, indica que en la media recortada a 5% se elimina el valor (1) menor y el valor (1) mayor. La media recortada a 5% usando las 10 observaciones restantes es 3524.50.

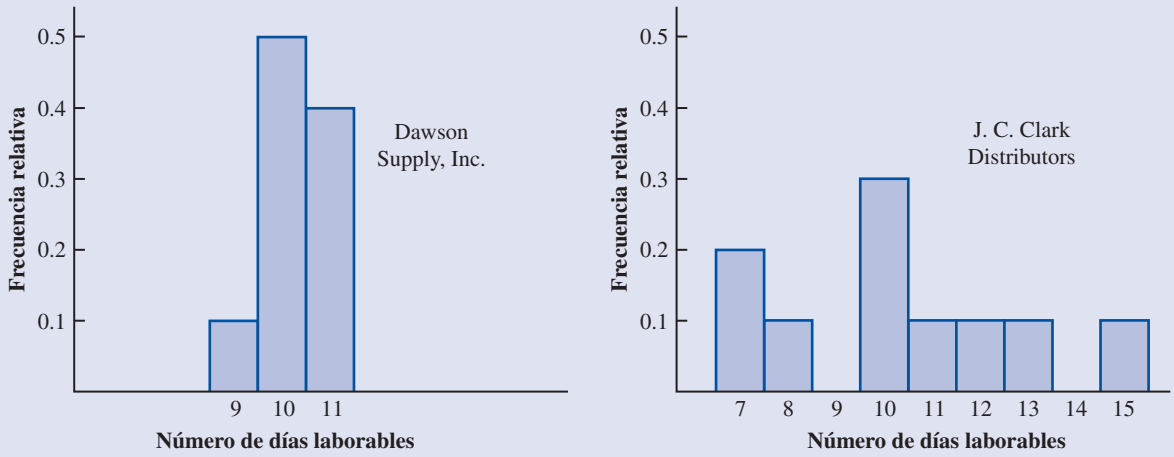
3.2 Medidas de variabilidad

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

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

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

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



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

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

Rango

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

RANGO

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

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

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

Rango intercuartílico

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

RANGO INTERCUARTÍLICO

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

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

Varianza

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

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

VARIANZA POBLACIONAL

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

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

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

VARIANZA MUESTRAL

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

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

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

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

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

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

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

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

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

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

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

Empleando la ecuación (3.5),

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

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

Desviación estándar

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

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

DESVIACIÓN ESTÁNDAR

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

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

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

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

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

Coeficiente de variación

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

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

COEFICIENTE DE VARIACIÓN

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

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

NOTAS Y COMENTARIOS

1. Los paquetes de software para estadística y las hojas de cálculo sirven para buscar los estadísticos descriptivos presentados en este capítulo. Una vez que los datos se han ingresado en una hoja de cálculo, basta emplear unos cuantos comandos sencillos para obtener los estadísticos deseados. En los apéndices 3.1 y 3.2 se muestra cómo usar Minitab y Excel para lograrlo.
2. La desviación estándar suele usarse como medida del riesgo relacionado con una inversión en acciones o en fondos de acciones (*BussinesWeek,* 7 de enero de 2000). Proporciona una medida de cómo fluctúa la rentabilidad mensual respecto de la rentabilidad promedio a largo plazo.
3. Redondear los valores de la media muestral $\bar{x}$ y de los cuadrados de las desviaciones $(x_i - \bar{x})^2$ puede introducir errores cuando se emplea una calculadora para el cálculo de la varianza y de la desviación estándar. Para reducir los errores de redondeo se recomienda conservar por lo menos seis dígitos significativos en los cálculos intermedios. La varianza o la desviación estándar obtenidos se redondean entonces a menos dígitos significativos.
4. Otra fórmula alterna para el cálculo de la varianza muestral es

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

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

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

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

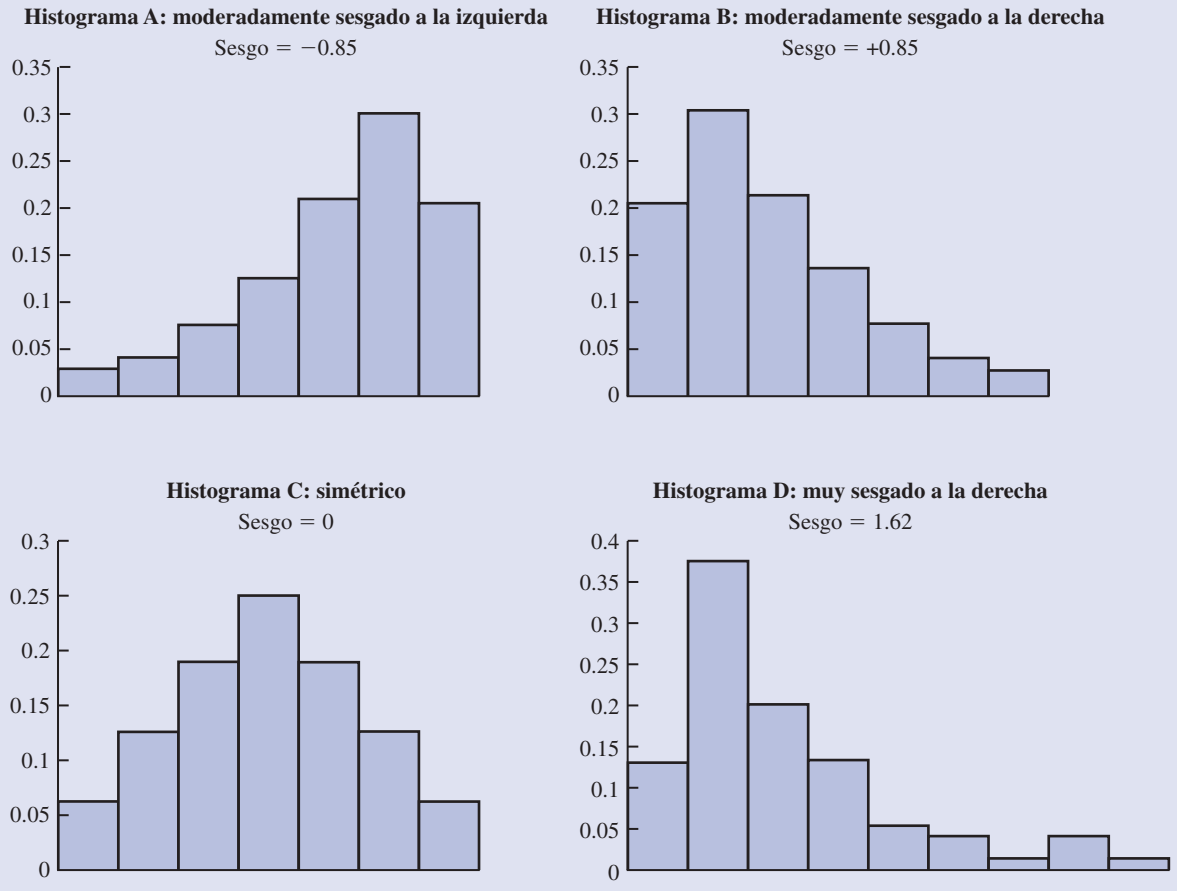
Forma de la distribución

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

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

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

FIGURA 3.3 HISTOGRAMAS QUE MUESTRAN EL SESGO DE CUATRO DISTRIBUCIONES



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

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

Puntos *z*

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

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

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

PUNTO z

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

donde

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

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

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

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

Teorema de Chebyshev

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

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

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

TEOREMA DE CHEBYSHEV

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

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

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

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

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

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

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

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

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