

Unidad 3

- Estadística descriptiva.

3 | Estadística descriptiva

3.1 | Introducción

Recuerde que el objetivo de la estadística moderna es **hacer inferencias acerca de un conjunto de datos, usualmente grande, llamado población, con base en la información contenida en una muestra**. Para formular las inferencias es necesario poder describir un conjunto de datos ya sea éste la población o una muestra.

Un buen número de textos tratan de los métodos de la estadística descriptiva, es decir, los métodos usados para describir conjuntos de datos numéricos. Estos métodos pueden ser clasificados en dos tipos: **métodos gráficos** y **métodos numéricos**. Aquí la discusión se restringe a unos cuantos métodos, tanto gráficos como numéricos, que son útiles no solamente para propósitos descriptivos sino también para hacer inferencias. Los métodos que han resultado obsoletos debido al advenimiento de las computadoras son omitidos y el lector que desee abundar más en el tema de la estadística descriptiva es referido a la bibliografía que se presenta al final del capítulo.

Los métodos gráficos que se presentan a continuación pueden ser aplicados a un conjunto de observaciones de la población o a un conjunto de observaciones de la muestra. Lo mismo es cierto para las medidas descriptivas numéricas pero se usan diferentes símbolos para indicar si la medida fue calculada a partir de la población o de la muestra.

3.2 | Distribuciones de frecuencias

Para introducir los métodos gráficos y numéricos presentados en este capítulo se hará uso de conjuntos de datos tomados de problemas reales. En el medio

financiero, los individuos y las organizaciones desean mantener una cartera de inversiones que les proporcione un beneficio máximo y que involucren niveles aceptables de riesgo. Una medida usual de la ganancia potencial y el riesgo inherente a una inversión es la razón de costo-beneficio. Es común que se prefieran inversiones con baja tasa de costo-beneficio a aquellos con tasa elevada. Los datos que se exhiben en la tabla 3.1 representan las razones de costo beneficio para 25 distintos tipos de acciones en el mercado de valores.

Tabla 3.1 Razón de costo-beneficio de 25 acciones en el mercado de valores.

20.5	19.5	15.6	24.1	9.9
15.4	12.7	5.4	17.0	28.6
16.9	7.8	23.3	11.8	18.4
13.4	14.3	19.2	9.2	16.8
8.8	22.1	20.8	12.6	15.9

De un vistazo a la tabla 3.1 se tiene que la tasa de costo-beneficio mayor es de 28.6 y la menor es 5.4. Surge la pregunta de cómo se distribuyen en este intervalo los otros 23 valores intermedios. Para responder a esta pregunta, dividimos el intervalo en cuestión en subintervalos de igual longitud. La elección del número de subintervalos debe basarse en la cantidad de datos que se consideran. Es común usar entre 5 y 20 subintervalos. A mayor el número de datos mayor el número de subintervalos. Para los datos de nuestro ejemplo se pueden usar los subintervalos de 5.00 a 8.99, de 9.00 a 12.99, de 13.00 a 16.99, Nótese que los puntos que definen los subintervalos fueron escogidos de forma tal que ninguno de los datos quede en el punto divisorio, para eliminar la ambigüedad al asignar las observaciones. En el lenguaje estadístico a estos subintervalos se les da el nombre de intervalos de clase o simplemente **clases**. Para determinar como se distribuyen las observaciones en las clases se pueden usar barritas de conteo como se muestra en la tabla 3.2.

Tabla 3.2 Frecuencias relativas para las 25 tasas de costo-beneficio

CLASE i	FRONTERAS DE CLASE	CONTEO	FRECUENCIA DE CLASE f_i	FRECUENCIA RELATIVA DE CLASE
1	5.00-8.99		3	3/25
2	9.00-12.99		5	5/25
3	13.00-16.99		7	7/25
4	17.00-20.99		6	6/25
5	21.00-24.99		3	3/25
6	25.00-28.99		1	1/25
Totales			25	1

Las observaciones caen en una de las seis clases, las cuales se numeran del 1 al 6 para facilitar su identificación. En la primera columna aparece el número de identificación de la clase, en la segunda las fronteras de clase y en la tercera

columna las barras de conteo, una por cada observación que cae en la clase. Por ejemplo, en la clase 1 caen 3 de las 25 observaciones, 5 en la clase 2 y así sucesivamente. El número de observaciones que caen en una clase dada, por ejemplo la clase i , es llamada la **frecuencia de clase** y se denota por el símbolo f_i . La frecuencia de clase aparece en la cuarta columna de la tabla 3.2. En la última columna de la tabla aparece la fracción del total de observaciones que caen en la clase o **frecuencia relativa**. Si se denota por n al número total de observaciones, $n = 25$ en nuestro ejemplo, entonces la frecuencia relativa para la i -ésima clase es f_i dividida entre el número de observaciones n .

$$\text{frecuencia relativa} = \frac{f_i}{n}$$

La tabulación final puede representarse gráficamente mediante un **histograma de frecuencias**, como en la figura 3.1. En un histograma de frecuencias se construyen rectángulos sobre cada una de las clases, con alturas proporcionales al número de elementos que caen en la clase (frecuencia de clase). Al observar el histograma se puede apreciar fácilmente como se distribuyen las tasas de costo-beneficio sobre el intervalo.

A veces conviene más graficar en el histograma las frecuencias relativas en lugar de las frecuencias de clase. La figura 3.2 muestra el histograma con las frecuencias relativas. Los estadísticos no hacen diferencia entre el histograma de frecuencias relativas y el de frecuencias y a ambos se refieren como histograma de frecuencia o simplemente como histograma. Si en el eje vertical se marcan los valores de las frecuencias y de las frecuencias relativas de manera correspondiente, los histogramas de frecuencias y de frecuencias relativas son idénticos (compare las figuras 3.1 y 3.2).

Aunque es importante describir el conjunto de $n = 25$ observaciones, es común que se tenga más interés en la población de la cual se tomó la muestra. Se puede pensar en las 25 razones de costo-beneficio como una muestra representativa tomada de la población de todas las tasas de costo-beneficio de todas

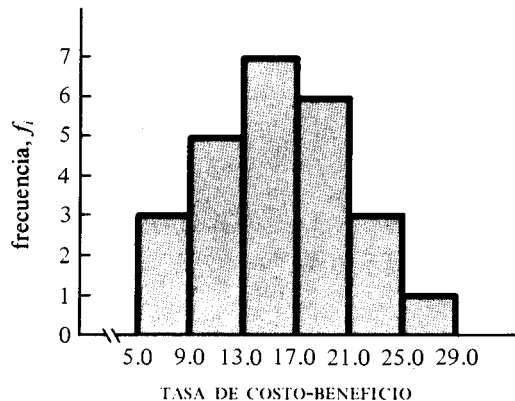


Figura 3.1 Histograma de frecuencias

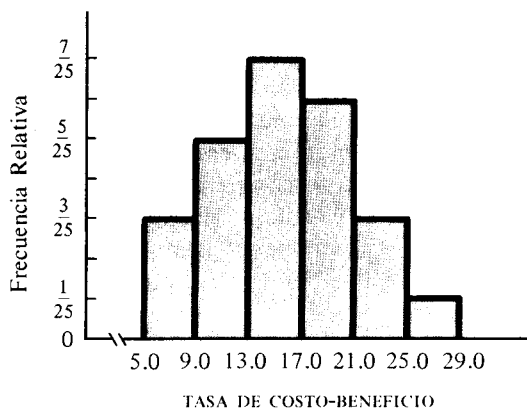


Figura 3.2 Histograma de frecuencias relativas

las acciones en el mercado de valores en 1976. Si se tuvieran las tasas de costo-beneficio de todas las acciones, se podría construir un histograma de frecuencias relativas de toda una población. Generalmente considerar los datos de toda una población puede resultar un proceso muy costoso y lento. Una muestra representativa puede en ocasiones ser la mejor y a veces la única información con que se cuenta para describir la población.

Considérese el histograma de frecuencias relativas de esta muestra con mayor detalle. ¿Qué fracción de las acciones tienen tasas de costo-beneficio de 17.0 o mayores? Observando el histograma es claro que esta fracción comprende las clases a la derecha de 17.0. De la tabla 3.2 se tiene que 10 acciones tienen tasas de costo-beneficio mayores o iguales a 17.0. Esto es, la fracción es de $\frac{10}{25}$, o 40%. Este es también el porcentaje del área total del histograma que está a la derecha de 17.0.

Suponga que cada uno de los 25 datos se anota en un pedazo de papel y cada papel se coloca en un sombrero. Se saca al azar un papel del sombrero. ¿Cuál es la probabilidad de que el papel tenga escrita una tasa de costo-beneficio mayor o igual a 17.0? Puesto que 10 de las 25 papeletas tienen tasas de costo-beneficio mayores o iguales a 17.0, se tienen 10 de 25 posibilidades. Se dice que la probabilidad en cuestión es de $\frac{10}{25}$. La palabra “probabilidad” se usa frecuentemente en el lenguaje cotidiano. La definición y significado formal de este término se describe en el capítulo 4.

Considere ahora la población de la cual fue tomada la muestra. ¿Qué proporción de todas las acciones en el mercado de valores tendrán tasas de costo-beneficio mayores o iguales a 17.0? Como no contamos con el histograma de frecuencias de la población entera, se podría inferir que el histograma de la población ha de ser similar al histograma de la muestra y entonces inferir que aproximadamente el 40% de las tasas de costo-beneficio de todas las acciones en el mercado de valores son mayores o iguales a 17.0. Es muy probable que esta estimación difiera del verdadero porcentaje en la población. La magnitud de este error de estimación será estudiada en el capítulo 8.

El histograma de frecuencias relativas es llamado a menudo **distribución de frecuencias** puesto que muestra cómo los datos se distribuyen en el eje

horizontal de la gráfica. Nótese que es posible dar a los rectángulos contruidos sobre cada intervalo de clase dos interpretaciones distintas pero relacionadas. Por una parte representan a la proporción de las observaciones que caen en esta clase. Por otra parte si se toma al azar una observación de los datos, la frecuencia relativa de una clase en particular representa también la probabilidad de que la observación caiga en esa clase. La característica más importante del histograma de frecuencias muestral es que proporciona información acerca del histograma de frecuencias de toda la población. Es sensato esperar que ambos histogramas, el de la muestra y el de la población, sean parecidos. En este caso el parecido será más notorio cuanto más observaciones de la población se consideren en la muestra. Si la muestra incluyese todas las observaciones en la población, obviamente ambos histogramas serían idénticos.

En la discusión anterior se vió como construir un histograma de frecuencias para las tasas de costo-beneficio de la tabla 3.1 y se explicó como interpretar esta distribución. Antes de finalizar este tema se hará un resumen de los principios que se emplean en la elaboración de una distribución de frecuencias de un conjunto de datos.

Para ilustrar el uso de estos principios suponga que se desea agrupar los 36 salarios de una pequeña empresa en cinco clases. Se tiene que el salario menor es de \$5,500.00 y el mayor de \$29,500.00. Para determinar la longitud de cada intervalo de clase, considérese la siguiente regla.

$$\text{longitud de intervalo} = \frac{\$29,500 - \$5,500}{5} = \$4,800$$

Puede resultar más conveniente usar \$5,000 como longitud de los intervalos. Las fronteras de clase serían entonces

\$5,000-\$9,999
 \$10,000-\$14,999
 \$15,000-\$19,999
 \$20,000-\$24,999
 \$25,000-\$29,999

Principios para la construcción de una distribución de frecuencias

- 1. Determinación del número de intervalos de clase.** Es conveniente tener entre 5 y 20 intervalos de clase. Entre más datos se tengan, más intervalos de clase deben considerarse. Si el número de intervalos es demasiado pequeño pueden estarse ocultando características importantes de los datos en esta agrupación. Si se tienen demasiados intervalos de clase se pueden tener muchos intervalos de clase vacíos que resten significado a la distribución. El número de intervalos de clase será determinado por el número de datos considerando también qué tan uniformes sean los datos. Una muestra pequeña requiere en general pocos intervalos de clase.
- 2. Determinación del tamaño de los intervalos.** Como regla general para encontrar la longitud de los intervalos divídase la diferencia de la mayor y la menor de las observaciones entre el número de intervalos de clase, modificando este cociente para obtener intervalos de clase de longitud

conveniente, cercana al cociente obtenido. Todas las clases deben tener la misma longitud, con posible excepción del primer y último intervalo. Esto es para permitir uniformidad en las comparaciones de frecuencias de clase.

3. **Determinación de las fronteras de clase.** Se debe comenzar por el primer intervalo, estando seguro de que incluye a la mínima de las observaciones. Se añaden los demás intervalos. Las fronteras deben seleccionarse de tal forma que sea imposible que una observación coincida con una frontera.

En nuestro ejemplo de los salarios las reglas para la construcción de distribuciones de frecuencia se siguen al pie de la letra. Cuando los datos se encuentran demasiado dispersos es posible dejar al primero o el último de los intervalos abierto por un extremo. Por ejemplo si el salario del presidente de la empresa fuera de \$100,000 y se incluye como el dato número 37 en la muestra, la última clase sería \$25,000 o más. Si no se hiciera así y se usara la regla para la longitud del intervalo se tendrían demasiados intervalos vacíos que harían confusa la interpretación de los resultados.

Ejercicios

3.1. Una agencia gubernamental desea investigar cómo se distribuyen los costos de arrendamiento de viviendas en una ciudad determinada. Se quieren encontrar intervalos para clasificar las $n = 100$ rentas mensuales de las viviendas seleccionadas. Las rentas mensuales varían de \$120 a \$550. Discuta usted el efecto de usar cada una de las siguientes longitudes para los intervalos de clase:

\$0.50, \$1.00, \$10.00, \$25.00, \$50.00, \$100.00

3.2 Se proponen 3 clasificaciones para los salarios por hora de un artesano. (Ver tabla E-1.) Critique usted cada una de las clasificaciones.

3.3. Como control de la ética publicitaria se requiere que el rendimiento, en millas por galón de gasolina, que las fabricantes de automóviles usan con fines publicitarios, esté basado en un buen número de pruebas efectuadas en diversas condiciones. Al tomar

una muestra de $n = 50$ automóviles se registran las siguientes observaciones en millas por galón:

27.9	29.3	31.8	22.5	34.2
34.2	32.7	26.5	26.4	31.6
35.6	31.0	28.0	33.7	32.0
28.5	27.5	29.8	31.2	28.7
30.0	28.7	33.2	30.5	27.9
31.2	29.5	28.7	23.0	30.1
30.5	31.3	24.9	26.8	29.9
28.7	30.4	31.3	32.7	30.3
33.5	30.5	30.6	35.1	28.6
30.1	30.3	29.6	31.4	32.4

a. Construya un histograma de frecuencias relativas usando cinco intervalos de clase de la misma longitud.

b. Los fabricantes afirman que su automóvil está diseñado para rendir al menos 30 millas por galón. ¿Qué porcentaje de autos en la muestra produce este rendimiento?

Tabla E-1

CLASIFICACIÓN I	CLASIFICACIÓN II	CLASIFICACIÓN III
\$ 0-3.00	\$ 0-3.50	\$ 0-menos de 3.00
3.00-6.00	3.51-7.00	4.00-menos de 7.00
6.00-9.00	7.01-10.00	8.00-menos de 11.00
9.00-12.00	10.01-15.00	12.00-menos de 15.00
12.00-15.00	más de 15.00	más de 15.00

3.4. Un cobrador de una empresa ha registrado el número de días que tarda en cobrar cada una de sus cuentas de crédito. Se han obtenido los siguientes 30 registros.

a. Construya un histograma de frecuencias relativas para este conjunto de datos.

b. Las cuentas con más de seis semanas (46 días) de retraso se consideran incobrables y se envían al de-

partamento legal. ¿Qué proporción de las cuentas tendrán que ser enviadas al departamento legal?

17	57	10	35	26	3
21	11	7	72	5	86
6	20	105	40	14	42
12	32	28	13	19	28
45	8	19	21	38	20

3.3 Otros métodos gráficos

Cuando se tienen datos que fueron colectados en distintos períodos de tiempo o distintas áreas geográficas es conveniente usar tablas o diagramas de diversos tipos para su presentación. En esta sección se presentan los principios para la elaboración y uso de este tipo de auxiliares gráficos.

En una tabla estadística se presenta una distribución de frecuencias clasificada o subdividida; en ella se comparan las frecuencias o las frecuencias relativas de muestras tomadas de dos o más poblaciones distintas. Las poblaciones pueden representar distintos períodos de tiempo, distintas áreas geográficas, distintas empresas de un ramo, distintos departamentos de una firma, etc. Dentro de cada muestra, las clasificaciones deben ser las mismas para permitir un análisis cruzado de los datos en forma adecuada.

Se cuenta también con auxiliares como los **diagramas de barras**, **las gráficas de líneas** o **los diagramas circulares** que sirven como un resumen visual de los datos. Hay muchos otros métodos gráficos que usa el estadístico en los negocios y en la economía pero la discusión de este capítulo se limita a los ya mencionados.

Tabla 3.3 *Análisis de los empleados en un despacho de asesoría*

	1974	1975	1976	1977
Total de Empleados	100	115	110	150
Profesionales	40	44	42	58
Hombres	37	40	39	48
Mujeres	3	4	3	10
No Profesionales	60	71	68	92
Hombres	50	58	56	69
Mujeres	10	13	12	23

La tabla 3.3 muestra, como ejemplo de una tabla estadística, el desglose por nivel académico y sexo de los empleados en un despacho de asesores para los años de 1974 a 1977. Las clasificaciones para cada uno de los cuatro años son las mismas, lo cual permite hacer comparaciones entre los distintos años. Las entradas de la tabla dan el número total de empleados, desglosado en profesionales y no profesionales para cada año. Note que la suma de profesionales y no profesionales da el número total de empleados para un año dado. La tabla también desglosa a profesionales y no profesionales en número de hombres y mujeres.

Algunas veces se usan porcentajes como entradas de las tablas estadísticas. Cuando sea éste el caso, las tablas deben ser examinadas cuidadosamente puesto que las comparaciones entre entradas correspondientes a distintas muestras pueden verse distorsionadas si el número de observaciones varía de muestra a muestra. Por ejemplo, dos entradas en un mismo renglón de la tabla pueden tener 50% ambas, pero la primera entrada puede representar 100 de 200 y la segunda puede representar 1 de 2. El 50% de la primera entrada basado en un total de 200 observaciones es, en este caso, más representativo que el segundo, basado en un total de 2. Por lo tanto, ambos porcentajes deben compararse con cuidado.

En la figura 3.3 se ejemplifica como se puede usar un **diagrama de barras** para exhibir gráficamente algunos de los datos de los empleados en la tabla 3.3. Por lo general los diagramas de barras no se subdividen tan finamente como las tablas puesto que particiones adicionales hacen confusa la apariencia de la gráfica, lo que desvirtúa el objetivo de producir una gráfica que sea de lectura sencilla y permita un análisis rápido de los datos.

Es posible construir distintos diagramas de barras que exhiban la misma información. En el ejemplo de la figura 3.3 es posible usar 3 barras para cada año, una barra exhibiendo el número de hombres, otra el número de mujeres y la tercera el número total de empleados. El tipo de diagrama usado no es importante, en tanto describa la información en forma clara y fácil de interpretar.

La figura 3.4 ilustra el uso conjunto de una **gráfica de líneas** y un **diagrama de barras**. Las barras muestran el número de viviendas en edificios multifamiliares construidos durante los años de 1965 a 1974, y la línea indica el número de viviendas unifamiliares construidas durante los mismos años. Note que el incremento en construcción de ambos tipos de habitación empieza en los años de 1965 a 1972 y declina en los años de 1973 y 1974. Instituciones tales como bancos, instituciones de crédito y agencias de bienes raíces pueden estar

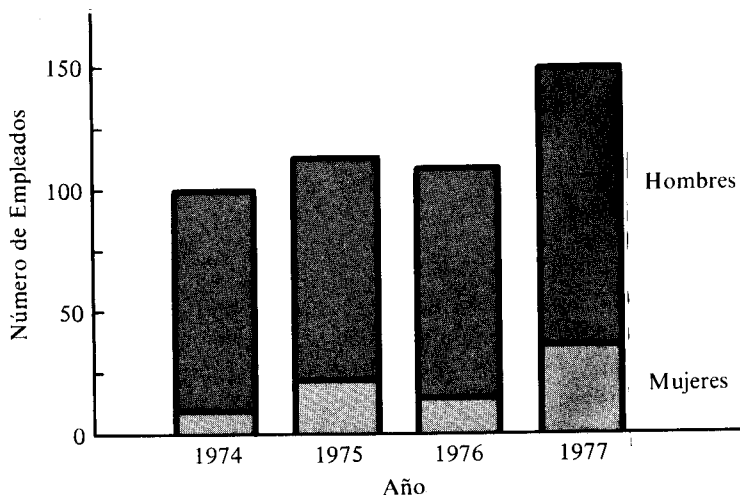


Figura 3.3 Diagrama de barras

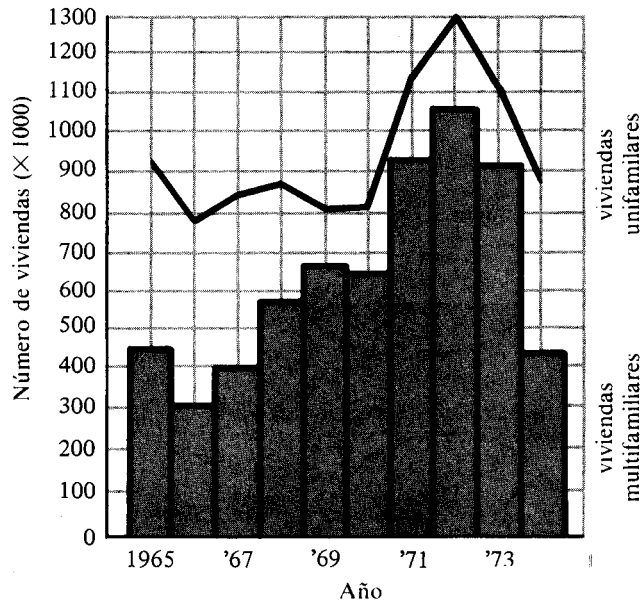


Figura 3.4 *Viviendas unifamiliares y multifamiliares construidas de 1965 a 1974*

interesadas, más que en el número exacto de unidades construido de cada tipo, en la relación entre uno y otro. Esta relación es mostrada claramente en la figura 3.4.

Los diagramas de barras sirven principalmente para representar el total de una cierta cantidad para cada año o para cada categoría presentada. En contraste, los **diagramas circulares** se usan para mostrar como una cantidad total se reparte en un grupo de categorías. Por ejemplo en la figura 3.5 se muestra el desglose por producto de los materiales transportados por vías acuáticas en los Estados Unidos en 1972. El diagrama en la parte superior corresponde al total del comercio. Este diagrama exhibe claramente el hecho de que el petróleo fue el producto que más se transportó por vías acuáticas (42.1% del total) mientras que las fibras representan sólo una pequeña parte (5.1%) del total. Es posible dar interpretaciones similares a los otros diagramas de la figura, pero lo principal es notar cómo el diagrama circular permite apreciar rápidamente qué proporción del total corresponde a cada subdivisión.

La construcción de un diagrama circular se facilita si se recuerda que el círculo completo tiene 360 grados y que este ángulo debe corresponder a un 100% del total representado. Por ejemplo considere el diagrama correspondiente al comercio total en la figura 3.5. Calcule la porción angular del círculo que se le debe asignar al petróleo y sus derivados. Puesto que esta porción representa el 42.1% del comercio total, se le debe asignar el 42.1% de los 360 grados totales del círculo. El ángulo correspondiente se calcula como sigue:

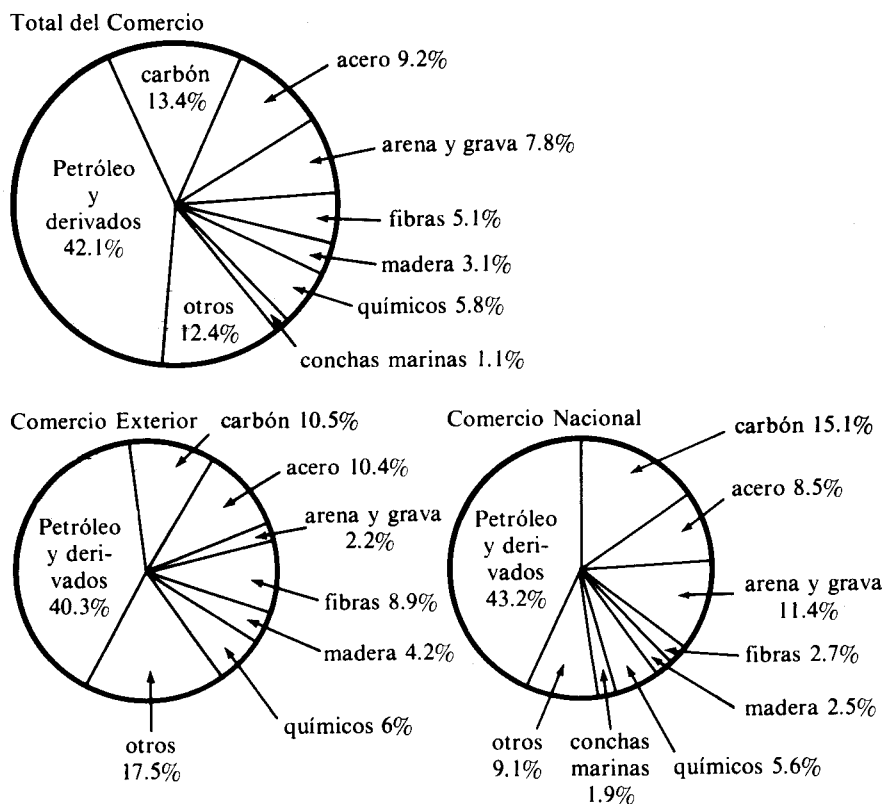


Figura 3.5 Materiales transportados por vías acuáticas en los Estados Unidos en 1972

$$\frac{42.1\%}{100\%} (360 \text{ grados}) = 151.56 \text{ grados}$$

Las porciones del círculo asignadas a otros productos se calculan de manera similar. El total de las porciones circulares para los diversos productos cubre los 360 grados del círculo.

El propósito principal de cualquier ayuda gráfica es dar una representación pictórica de los datos que resulte de fácil y rápida interpretación. El tipo de gráfica o cuadro que se presente o el formato usado es secundario. Lo importante es que la gráfica o tabla presentada comunique el mensaje que contienen los datos en un lenguaje que sea claro y rápidamente inteligible para casi cualquier persona.

Ejercicios

3.5. Los datos en la tabla E-2 representan los gastos de consumo personal por tipo de bienes en los Estados Unidos, para algunos años entre 1929 y 1969 (en miles de millones de dólares):

- Elabore una gráfica de barras que represente estos datos.
- Use la gráfica para discutir los cambios en el tiempo de los gastos en los 3 tipos de bienes en el período de 1929 a 1969.

Tabla E-2

TIPO DE BIENES	AÑO			
	1929	1949	1959	1969
bienes perdurables	9.2	24.6	44.3	90.0
bienes perecederos	37.7	94.6	146.6	245.8
servicios	30.3	54.6	120.3	241.6

3.6. El costo estimado, en centavos por milla, para la operación de un automóvil compacto modelo 1974 durante 10 años o 100,000 km se muestran en la tabla siguiente. Construya una gráfica circular para

Depreciación	2.36
Mantenimiento	2.12
Gasolina	1.82
Seguros	1.47
Estacionamiento	1.96
Impuestos	0.92

representar estos costos de operación.

3.7. La distribución en la siguiente tabla muestra el número de accionistas de una empresa clasificado por el número de acciones que posee:

ACCIONES	ACCIONISTAS
menos de 1000	486
1000-1999	372
2000-2999	210
3000-3999	117
4000-4999	43
5000-5999	27
6000 o más	13

- Construya un histograma de frecuencias.
- Construya una gráfica circular que describa estos datos.

3.4 Gráficas que engañan

Aún cuando las técnicas gráficas para describir datos son muy útiles, deben ser examinadas con extremo cuidado. Es posible en algunos casos construir gráficas que induzcan al lector inexperto a conclusiones equivocadas. El método más fácil para confundir al lector es modificar la escala de los ejes de la gráfica.

Por ejemplo, suponga que el número de conatos de accidente por mes en un aeropuerto es registrado en el período de Enero a Mayo, obteniéndose respectivamente 13, 14, 14, 15 y 15.

Si se desea que el crecimiento observado parezca pequeño (trabajando para las autoridades del aeropuerto) se pueden exhibir los datos en una gráfica como la figura 3.6. En esta gráfica el crecimiento es aparente pero no parece muy grande. Si se quiere exagerar esta crecimiento (si se desea protestar por este hecho) se puede usar una gráfica como en la figura 3.7, en donde se ha ampliado el eje vertical y excluye los puntos del 1 al 12. Nótese el efecto visual que produce al aumentar la pendiente a una más pronunciada.

Otro modo de lograr el mismo efecto, aumentar o disminuir la pendiente, consiste en encoger o alargar el eje horizontal. Hay, por supuesto, limitaciones para estos estiramientos y encogimientos de los ejes, si se quiere lograr una gráfica que parezca razonable a la vista del lector. Por ejemplo en las figuras 3.6 y 3.7 no se pueden encoger o alargar demasiado los ejes debido al número limitado de datos ($n = 5$).

El alargar o encoger los ejes para aumentar las pendientes en las gráficas de barras, histogramas, polígonos de frecuencia y demás figuras de este tipo, comúnmente toma por sorpresa al lector descuidado. Es siempre necesario un análisis cuidadoso de las escalas usadas en ambos ejes. Sin embargo, el hecho

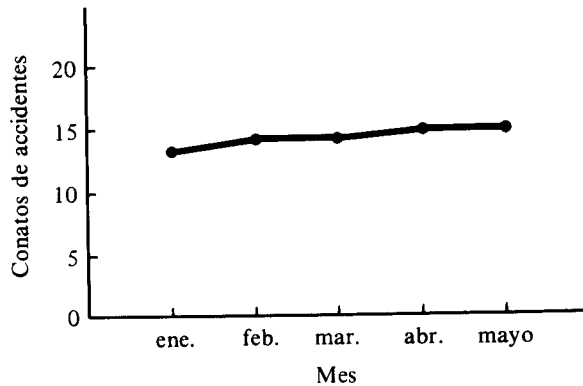


Figura 3.6 *Número de conatos de accidentes por mes.*

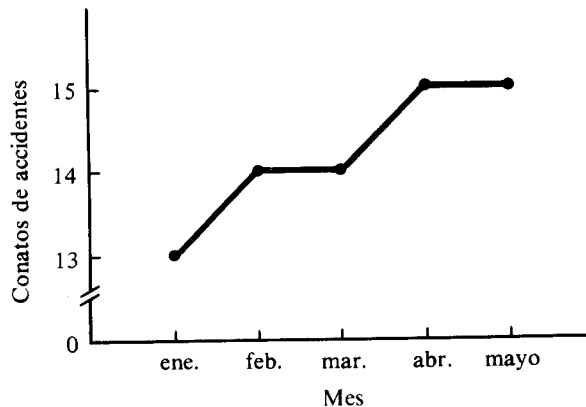


Figura 3.7 *Número de conatos de accidentes por mes.*

de que los incrementos o decrementos en la respuesta resulten grandes o pequeños depende de la importancia que el lector dé a estos cambios y no de la pendiente que se muestra en la representación gráfica.

Los ejemplos anteriores nos dan una sencilla ilustración de como la realidad puede ser deformada, accidentalmente o intencionalmente, usando métodos gráficos descriptivos. Es necesario al contemplar una representación gráfica examinarla detenidamente prestando especial atención a las escalas usadas. Debe observarse también si los ejes aparecen interrumpidos como en el caso de la figura 3.6 y, lo más importante, reflexione si los cambios observados en la variable descrita son o no importantes desde un punto de vista práctico. Las conclusiones acerca de los datos en una gráfica deben sacarse con extremo cuidado.

3.5 Medidas descriptivas numéricas

Los métodos gráficos son extremadamente útiles para lograr una descripción de los datos que sea captada rápidamente por el lector. Como se dice comúnmente una imagen dice más que cien palabras. Sin embargo las técnicas gráficas presentan limitaciones para la descripción y análisis de conjuntos de datos. Por ejemplo supóngase que se desea discutir los datos con un grupo de personas y es necesario describirlos verbalmente. Puesto que no es posible utilizar un histograma, es necesario usar otras medidas descriptivas que puedan dar al auditorio la idea mental de un histograma.

Una segunda limitación del histograma y otras técnicas gráficas, que no resulta tan obvia, es que son difíciles de usar para hacer inferencia estadística. Supuestamente el histograma de la muestra se usa para hacer inferencias acerca de la forma y la posición del histograma que describe a la población entera y que es desconocido. Esta inferencia se basa en la acertada suposición de que debe existir cierta similitud entre los dos histogramas. Se sabe cuando los histogramas son idénticos, pero esta situación no es común en la práctica. Si el histograma de la muestra y el de la población no coinciden, es necesario poder medir hasta qué grado difieren, o bien el grado de similitud. Por ejemplo puede preguntarse acerca del grado de similitud entre el histograma en la figura 3.2 y el histograma de frecuencias de la población entera de tasas de costo-beneficio de la cual la muestra fue tomada. Aunque si bien estas dificultades no son insuperables, es preferible, en ocasiones, utilizar otras medidas descriptivas que puedan usarse como predictores de la forma de la distribución de frecuencias de la población.

Las limitaciones de los métodos gráficos pueden salvarse con el uso de medidas descriptivas numéricas. Con base en los datos muestrales es posible calcular un conjunto de números que proporcionen al estadístico una buena imagen mental de la distribución de frecuencias de la población y que resulte útil para hacer inferencias acerca de las características de la población.

Definición

Las medidas descriptivas numéricas calculadas a partir del total de observaciones de la población se denominan **parámetros**; aquellas calculadas de las observaciones de una muestra se denominan **estadísticas**.

3.6 Medidas de tendencia central

Para construir una imagen mental de la distribución de frecuencias de un conjunto de datos, puede hacerse uso de distintas medidas descriptivas. Una de las primeras medidas de interés es una **medida de tendencia central**, esto es, una medida del centro de la distribución. Nótese en el ejemplo de las tasas de costo-beneficio que los datos van de 5.4 a 28.6, y que el centro del histograma

está localizado en la vecindad de 16.0. Aquí se consideran reglas para localizar el centro de la distribución de un conjunto de datos.

Una de las medidas de tendencia central más comunes y de las más útiles es el promedio aritmético del conjunto de observaciones. A esta medida se le conoce también como **media aritmética** o simplemente la **media** de un conjunto de observaciones.

Definición

La **media aritmética** de un conjunto de n observaciones $y_1, y_2, \dots, y_n$ es igual a la suma de las observaciones dividida entre n .

Es siempre conveniente tener presente tanto la muestra como la población. Cada una de éstas tiene su media. Para distinguir entre las dos se usará el símbolo $\bar{y}$ (y -barra) para la media muestral y la letra griega μ (mu) para denotar a la media de la población. Denotando a las observaciones muestrales por $y_1, y_2, \dots, y_n$, la fórmula para la media muestral se da en el siguiente cuadro.

Media muestral

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Es fácil verificar que $\bar{y}$ cae cerca del centro de las observaciones a partir de las cuales fue calculada.

Ejemplo 3.1

Encuentre la media de las observaciones 2, 9, 11, 5, 6.

Solución Substituyendo las observaciones en la fórmula se tiene

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{2 + 9 + 11 + 5 + 6}{5} = 6.6$$

Como se ha visto, $\bar{y}$ se usa para localizar el centro de una muestra de observaciones. Otro uso de mayor importancia de $\bar{y}$ es el de servir de estimador (predictor) del valor desconocido de la media μ de la población. Por ejemplo, el promedio de la muestra dada en la tabla 3.1 es

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{400}{25} = 16.0$$

Observe que este valor cae aproximadamente en el centro del conjunto de medidas de esta muestra. La media de la población es desconocida, pero si se

necesitara hacer una estimación de este valor, nuestra estimación de μ sería 16.0.

Una segunda medida de tendencia central es la **mediana**.

Definición

La **mediana** de un conjunto de n observaciones $y_1, y_2, y_3, \dots, y_n$ se define como la observación que cae en el centro cuando las observaciones se ordenan en orden creciente. Si el número de observaciones es par, se escoge como mediana al valor medio entre las dos observaciones de en medio.

Ejemplo 3.2 Considere las observaciones muestrales

9, 2, 7, 11, 14

Ordenándolas de acuerdo a su magnitud, 2, 7, 9, 11, 14, se tiene que 9 es la mediana.

Ejemplo 3.3 Considere las observaciones muestrales

9, 2, 7, 11, 14, 6

Arregladas en orden creciente se tiene 2, 6, 7, 9, 11, 14. La mediana en este caso es el valor medio entre 7 y 9, el cual es 8.

La regla para localizar la mediana cuando el número de observaciones es par parece un tanto arbitraria. Recuerde que la mediana se calcula ya sea con fines únicamente descriptivos o bien como un estimador de la mediana de la población. Con fines descriptivos se puede ser tan arbitrario como se desee. Si se pasa a estimar la mediana de la población, la bondad de la regla para calcularla depende de si tiende a producir una estimación buena o mala de la mediana de la población.

Una tercera medida que se usa en ocasiones como medida de tendencia central es la **moda**.

Definición

La **moda** de un conjunto de n observaciones $y_1, y_2, y_3, \dots, y_n$ se define como el valor que ocurre con mayor frecuencia.

Ejemplo 3.4 Considere las observaciones muestrales

9, 2, 7, 11, 14, 7, 2, 7.

El valor 7 ocurre 3 veces, el 2 aparece 2 veces y el resto ocurre una vez cada uno. Luego 7 es la moda de estas observaciones muestrales.

La moda no es una medida de tendencia central muy usual, pero sirve en los negocios, por ejemplo para determinar qué tamaño del producto es el de

mayor demanda. Por ejemplo un fabricante de camisetas está interesado en cuál es la talla de mayor demanda. Similarmente para programar la producción de un medicamento el fabricante estará interesado en cual es la dosis más comúnmente recetada por los médicos. Todas estas medidas son descritas adecuadamente por la moda. La figura 3.8 muestra una distribución de frecuencia típica en la cual se indica la moda.

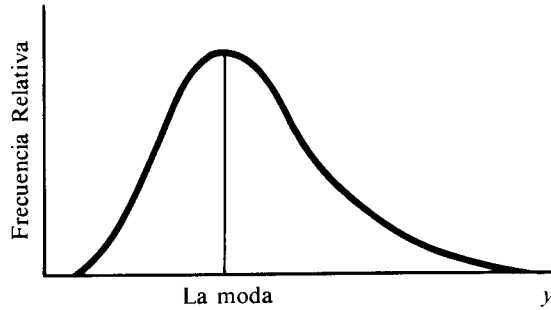
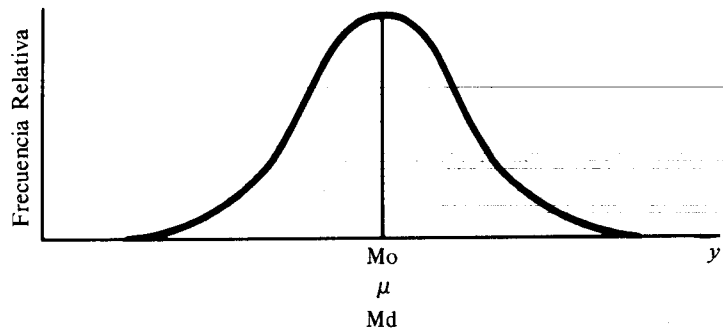


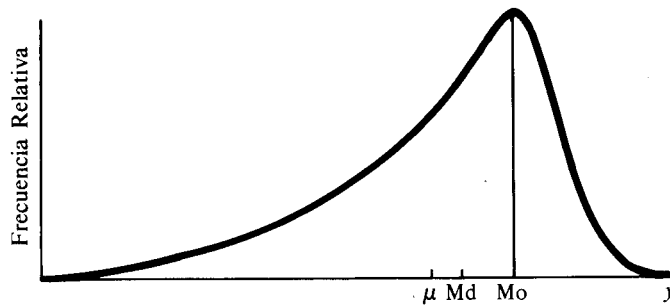
Figura 3.8 Colocando la moda para una distribución de frecuencias

La relación entre la media (μ), la mediana (Md), y la moda (Mo) se ilustra en la figura 3.9. Para una distribución de frecuencias simétrica, esto es, una para la cual los valores equidistantes a la media ocurren con igual frecuencia, como se muestra en la figura 3.9(a), los valores de la media, la mediana y la moda son idénticos. Si la distribución está desplazada a la izquierda la media, la mediana y la moda están en este orden, como se muestra en la figura 3.9(b). Si la distribución está desplazada a la derecha entonces la moda, la mediana y la media quedan en ese orden, como se muestra en la figura 3.9(c).

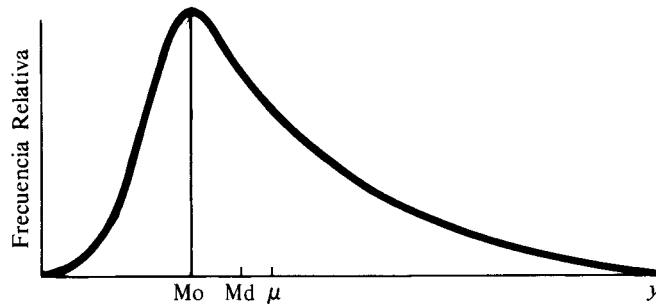


(a) Distribución Simétrica

Figura 3.9 Relaciones entre la media (μ), mediana (Md) y moda (Mo)



(b) Distribución desplazada a la izquierda



(c) Distribución desplazada a la derecha

Figura 3.9 Relaciones entre la media (μ), mediana (Md) y moda (Mo) (continuación)

La media mide el “centro de gravedad” de un conjunto de datos y es consecuentemente afectada por los valores extremos. Esta propiedad de la media se puede observar en las figuras 3.9(b) y 3.9(c). Si la distribución está desplazada a la derecha, la media se desplaza a la derecha de la moda; similarmente si la distribución se desplaza a la izquierda. La mediana ignora valores extremos, éstos son tomados en cuenta sólo para localizar el valor central en los datos ordenados.

La inferencia estadística resulta en general más simple cuando se usa la media. Puesto que en los siguientes capítulos es la inferencia el tema principal y debido también a que la media es más usada para este propósito, de aquí en adelante concentraremos la atención en la media como medida de tendencia central.

Ejercicios

3.8. Los siguientes datos representan el número de interrupciones por día de trabajo debidas a fallas mecánicas en una planta procesadora de alimentos:

2, 3, 0, 5, 4, 3, 1, 3, 5, 2

Calcule la media, la mediana y encuentre el número modal de interrupciones diarias.

3.9. Diga usted qué medidas de tendencia central serían más útiles en cada uno de los siguientes casos.

a. El gerente de producción de una fábrica de envases de vidrio quiere saber cuál es el tamaño de envase que debe fabricar en mayor cantidad. El tiene a la mano un buen número de datos de los tamaños de envase ordenados por los clientes.

b. El gerente de ventas de una compañía que produce mobiliario de lujo desea seleccionar regiones para establecer salas de exhibición. ¿En qué medida del ingreso familiar por región estará más interesado, en la media o en la mediana?

c. Un analista de la bolsa de valores está interesado en describir el cambio diario en el precio en el mercado de una acción de cierta compañía. Rara vez el precio cambia más de un punto, pero hay ocasiones en que el precio cambia hasta cuatro puntos. ¿Qué medida debe usar el analista para describir el cambio de precio de la acción en cuestión, la media, la mediana o la moda de los cambios de precio en el mercado?

3.7 | Medidas de variabilidad

Una vez localizado el centro de la distribución de un conjunto de datos, el siguiente paso es buscar una medida de la **variabilidad** o **dispersión** de los datos. Considere las dos distribuciones que se exhiben en la figura 3.10. Ambas distribuciones están centradas en $y = 4$, pero existe una gran diferencia en la variabilidad de las observaciones alrededor de la media para estas distribuciones. Las observaciones en la figura 3.10(a) van de 3 a 5; en la figura 3.10(b) van de 0 a 8.

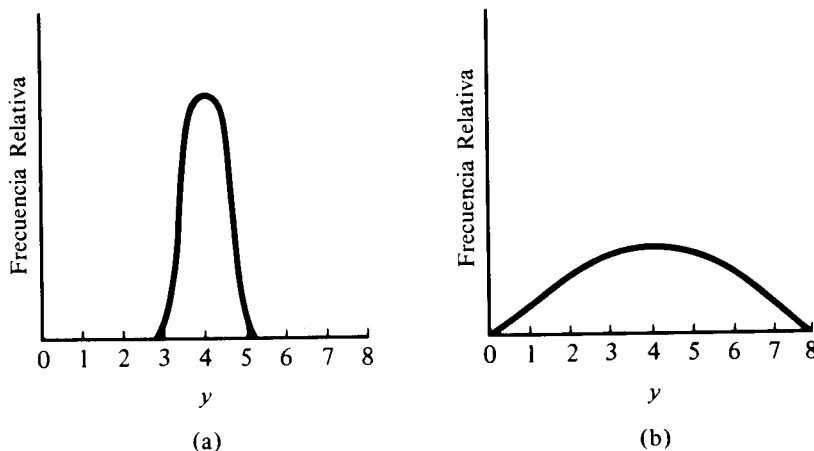


Figura 3.10 Variabilidad o dispersión de los datos

La variabilidad es una característica muy importante de un conjunto de datos. Por ejemplo si se está fabricando tornillos, la variación excesiva en el diámetro de los tornillos implica un alto porcentaje de productos defectuosos.

Además de la importancia práctica de la variación en los datos, es obvio que se necesita una medida de esta característica para lograr una imagen mental de la distribución de frecuencias. Existen diversas medidas de variabilidad, pero aquí se discutirán solamente algunas de las más importantes.

La medida de variación más simple es el recorrido.

Definición

El **recorrido** de un conjunto de n observaciones $y_1, y_2, \dots, y_n$ se define como la diferencia entre la mayor y la menor de las observaciones.

En el ejemplo de las tasas de costo-beneficio de la tabla 3.1 las observaciones varían de 5.4 a 28.6, entonces el recorrido es $28.6 - 5.4 = 23.2$.

Desafortunadamente el recorrido no resulta totalmente satisfactorio como medida de variación. Considere las dos distribuciones de la figura 3.11. Ambas distribuciones tienen el mismo recorrido pero los datos en la figura 3.11(b) son más variables que los datos de la figura 3.11(a).

Otras medidas de variación que ayudan a salvar la dificultad anterior son los **cuartiles** y los **porcentiles**. Recuérdese que si se especifica un intervalo en el eje horizontal del histograma, el porcentaje del área bajo el histograma, que queda arriba del intervalo, es igual al porcentaje del total de las observaciones que caen en ese intervalo. Puesto que la mediana está a la mitad de las observaciones cuando éstas se ordenan en sentido ascendente la mediana será el valor y tal que la mitad del área del histograma queda a la izquierda y la otra mitad a la derecha. Similarmente se definen los **cuartiles** como aquellos valores que dividen el área del histograma en cuartos.

Definición

Sean $y_1, y_2, y_3, \dots, y_n$ un conjunto de n observaciones ordenadas por orden de magnitud. El **primer cuartil** es aquel que deja a la izquierda $1/4$ de las observaciones y es menor que $3/4$ de las observaciones. El **segundo cuartil** es la mediana. El **tercer cuartil** sobrepasa $3/4$ de las observaciones y es menor que $1/4$ de ellas.

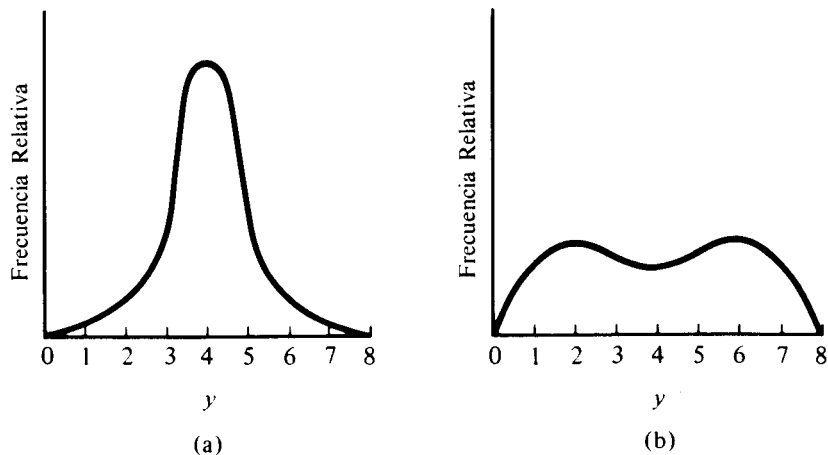


Figura 3.11 Distribuciones con igual recorrido y distinta variabilidad

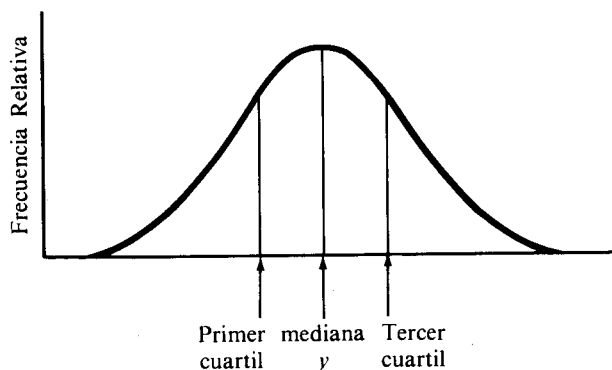


Figura 3.12 Los cuartiles

En la figura 3.12 se muestra como $1/4$ del área queda a la izquierda del primer cuartil, $3/4$ a la derecha. El tercer cuartil es el valor de y tal que $3/4$ del área queda a la izquierda y $1/4$ a la derecha.

Cuando se tienen pocos datos, es posible que los cuartiles queden entre dos observaciones y entonces es posible tomar una infinidad de números que satisfagan la definición. Existen reglas arbitrarias para, en estos casos, localizar un valor específico para los cuartiles. Pero son omitidas en este texto dado que es poco común calcular cuartiles cuando se tienen pocas observaciones. Para grandes cantidades de datos los cuartiles o se localizan en uno de los datos, o bien, los intervalos entre dos observaciones son tan pequeños que el punto exacto donde se les localice carece de importancia práctica. Entonces la definición anterior satisface por lo general las necesidades prácticas.

En algunas aplicaciones, principalmente cuando se tiene una gran cantidad de datos, es preferible usar los *porcentiles*.

Definición

Sean $y_1, y_2, y_3, \dots, y_n$ un conjunto de n observaciones arregladas por orden de magnitud. El **porcentil p** es un valor tal que a lo más p por ciento de las observaciones es menor que él y a lo más $(100 - p)$ por ciento de las observaciones son mayores.

Por ejemplo el porcentil 90 para un conjunto de datos es un valor que excede al 90% de las observaciones y es excedido por 10% de las observaciones. Como en el caso de los cuartiles, 90% del área del histograma queda a la izquierda del porcentil 90.

Una ventaja del recorrido es que se expresa como un solo número. Los cuartiles y los porcentiles dan mayor información sobre la localización y la variación de los datos pero se necesitan varios números para dar una descripción adecuada. La pregunta ahora es ¿es posible encontrar una medida de variabilidad que se exprese como un solo número y que sea más significativo que el recorrido?

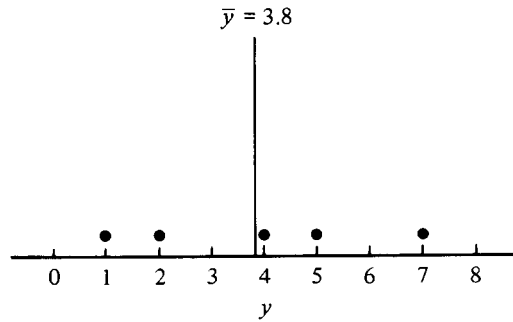


Figura 3.13 Diagrama de puntos

Para ilustrar con un ejemplo, considérense las observaciones muestrales 5, 7, 1, 2, 4. Estas observaciones se muestran en el **diagrama de puntos** que se presenta en la figura 3.13.

Si se calcula la media como medida de tendencia central se obtiene:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{19}{5} = 3.8$$

La localización de $\bar{y}$ también se muestra en el diagrama de puntos. Es posible ver la variabilidad en términos de las distancias entre cada punto (observación) y la media $\bar{y}$. Si las distancias son grandes podemos decir que hay más variabilidad que si las distancias son pequeñas. Más explícitamente, se define la **desviación** de una observación de su media por la medida $(y_i - \bar{y})$. Nótese que las observaciones a la derecha de la media producen desviaciones positivas y observaciones a la izquierda producen desviaciones negativas. Los valores de las observaciones y de las desviaciones para este ejemplo se muestran en la primera y segunda columnas de la tabla 3.4.

Tabla 3.4 Cálculo de $\sum_{i=1}^n (y_i - \bar{y})^2$

	y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	y_i^2
	5	1.2	1.44	25
	7	3.2	10.24	49
	1	-2.8	7.84	1
	2	-1.8	3.24	4
	4	.2	.04	16
Totales	19	0	22.80	95

Si se está de acuerdo en que las desviaciones contienen información sobre la variación de las observaciones, el siguiente paso es construir una fórmula basada en las desviaciones y que sea una buena medida de la variación. Como

una posibilidad se puede considerar el promedio de las desviaciones. Desafortunadamente esta idea no funcionará porque algunas desviaciones son positivas, otras negativas, y su suma es siempre 0. Esto último puede demostrarse usando los teoremas sobre sumatorias del capítulo 2. La demostración es como sigue:

Dadas n observaciones $y_1, y_2, y_3, \dots, y_n$,

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y}) &= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} = \sum_{i=1}^n y_i - n\bar{y} \\ &= \sum_{i=1}^n y_i - n \frac{\sum_{i=1}^n y_i}{n} = \sum_{i=1}^n y_i - \sum_{i=1}^n y_i = 0\end{aligned}$$

Note que las desviaciones en la segunda columna de la tabla 3.4 suman a cero.

Se puede preguntar ¿por qué no calcular el promedio de los valores absolutos, (las magnitudes ignorando el signo), de las desviaciones? Esta idea se ha usado de hecho para construir medidas de variabilidad pero tiende a resultar insatisfactoria en algunos aspectos de la inferencia estadística. Para evitar la dificultad causada por los signos, es preferible usar la suma de los cuadrados de las desviaciones

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

Para un número fijo de observaciones, cuando esta cantidad es grande, los datos son más variables que cuando la cantidad es pequeña.

Definición

La **varianza de una población** de N observaciones $y_1, y_2, y_3, \dots, y_N$ se define como el promedio del cuadrado de las desviaciones con respecto a su media μ . La varianza de la población se denota por σ^2 (σ es la letra griega sigma) y está dada por la fórmula

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$$

Nótese que se ha usado la letra mayúscula N para denotar el número de elementos en la población y la letra minúscula n para el número de elementos en la muestra.

Comúnmente no se cuenta con el total de observaciones de la población entera y sólo se dispone de una muestra de observaciones tomadas de la población. En este caso se debe usar la varianza de la muestra como se define en el siguiente cuadro.

Definición

La **varianza de una muestra** de n observaciones $y_1, y_2, y_3, \dots, y_n$ se define como la suma de los cuadrados de las desviaciones de las observaciones respecto de su media $\bar{y}$, dividida esta suma entre $(n - 1)$. La varianza de la muestra se denota por s^2 y está dada por la fórmula

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2$$

Por ejemplo, se puede calcular la varianza del conjunto de la muestra de $n = 5$ observaciones que se presentan en la tabla 3.4. El cuadrado de la desviación de cada observación aparece en la tercera columna de la tabla 3.4. La suma de los cuadrados de las desviaciones es

$$\sum_{i=1}^5 (y_i - \bar{y})^2 = 22.80$$

y la varianza muestral es

$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{22.80}{4} = 5.70$$

Existe una aparente inconsistencia entre las definiciones de la varianza de la población y de la varianza de una muestra. Recuerdese que la media muestral $\bar{y}$ se usa como un estimador de la media de la población μ . Aunque no se ha establecido, es conveniente mencionar que la media de la muestra es un buen estimador de μ . Siguiendo un razonamiento similar parecería razonable suponer que

$$s'^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

debería ser un buen estimador de la varianza de la población σ^2 . Sin embargo se puede demostrar que cuando el tamaño de la muestra (n) es pequeño, el estimador s'^2 tiende a subestimar a σ^2 mientras que la varianza muestral s^2 es un mejor estimador de σ^2 . Nótese que s'^2 y s^2 difieren sólo en el denominador y cuando n es grande s'^2 y s^2 serán aproximadamente iguales. En capítulos posteriores se hará uso de s^2 como estimador de σ^2 en numerosas ocasiones, **s'^2 no será mencionado más en este texto.**

Tal vez, hasta aquí, resulte decepcionante el significado práctico que se le ha dado a la varianza como medida de variabilidad. Grandes varianzas implican gran variación de los datos, pero una afirmación de este tipo permite solamente comparaciones entre distintos conjuntos de datos. Hasta aquí, la varianza no dice nada específico acerca de un solo conjunto de datos. Por ejemplo ¿qué se puede decir de la variabilidad de un conjunto de datos que

tenga varianza igual a 100? La respuesta no puede darse con los elementos que disponemos. Se intentará remediar esta situación por medio de una nueva definición y, en la siguiente sección, de un teorema y una regla práctica.

Definición

La **desviación estándar** de un conjunto de n observaciones $y_1, y_2, y_3, \dots, y_n$ es igual a la raíz cuadrada positiva de la varianza.

La varianza se mide en términos del cuadrado de las unidades originales. Si las observaciones están medidas en centímetros la varianza está dada en centímetros cuadrados. Al tomar la raíz cuadrada de la varianza, se obtiene la desviación estándar, con lo que se regresa a las unidades originales de las observaciones.

Desviación Estándar Muestral

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

La desviación estándar de la población es σ .

Ahora que se tiene la desviación estándar, puede preguntarse ¿para qué entonces se definió primero la varianza? El hecho es que ambas medidas desempeñan un papel importante en la estadística.

3.8 El significado práctico de la desviación estándar

En esta sección se presenta un interesante y útil teorema desarrollado por el matemático ruso Tchebysheff. La demostración de este teorema no es muy difícil pero se omite en esta discusión.

El Teorema de Tchebysheff

Dados un número k mayor o igual que uno, y un conjunto de observaciones $y_1, y_2, y_3, \dots, y_n$, al menos $(1 - 1/k^2)$ de las observaciones caen dentro de k desviaciones estándar de la media.

El teorema de Tchebysheff se refiere a cualquier conjunto de observaciones; por lo tanto se puede aplicar tanto a una muestra como a la población. Con fines de ilustración se usará la notación correspondiente a la población pero tenga presente que lo discutido en este caso se aplica de manera idéntica a la muestra.

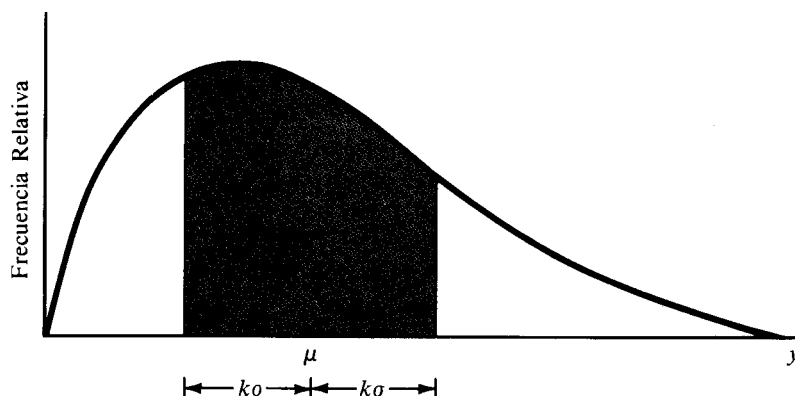


Figura 3.14 Ilustración del teorema de Tchebysheff.

La idea contenida en el teorema de Tchebysheff se ilustra en la figura 3.14. En esta figura se construye un intervalo midiendo una distancia de $k\sigma$ a ambos lados de la media. Note que el resultado del teorema es cierto para cualquier número k siempre y cuando éste sea mayor o igual a uno. Entonces dentro de este intervalo se tendrá una fracción del $(1 - 1/k^2)$ del número total de observaciones n .

Para entender mejor el resultado del teorema, se calculan algunos valores de la fracción $(1 - 1/k^2)$. Para $k = 1$ el teorema afirma que cuando menos $1 - 1/(1)^2 = 0$ de las observaciones caen en el intervalo de $(\mu - \sigma)$ a $(\mu + \sigma)$, resultado que no proporciona información en absoluto. Pero para $k = 2$, $1 - 1/(2)^2 = 3/4$ de las observaciones caen en el intervalo de $(\mu - 2\sigma)$ a $(\mu + 2\sigma)$. Cuando menos $8/9$ de las observaciones caen dentro de 3 desviaciones estándar de la media, es decir dentro del intervalo de $(\mu - 3\sigma)$ a $(\mu + 3\sigma)$. Aunque $k = 2$ y $k = 3$ se usan comúnmente al aplicar este teorema, el número k no tiene que ser entero. Nótese que a $k = 2.5$ desviaciones estándar de la media se tienen al menos $1 - 1/(2.5)^2 = .84$ observaciones.

Cuando se aplica el teorema de Tchebysheff a datos de una muestra es posible usar s' (definido en la sección 3.7) en lugar de s para construir los intervalos centrados en la media. s' es una cantidad un poco menor que s y da intervalos algo más cortos pero el asunto es de poca importancia práctica

Tabla 3.5 Algunos Valores de $(1 - 1/k^2)$

k	$(1 - 1/k^2)$
1	0
2	3/4
3	8/9

puesto que el teorema es también cierto cuando se usa s' en lugar de s . Además, para n grande, s y s' son casi iguales.

Nótese que aunque el objetivo principal es el de describir poblaciones, los ejemplos que se presentan describen conjuntos con pocas observaciones y son usados únicamente para ilustrar el uso del teorema de Tchebysheff.

Ejemplo 3.5

La media y la varianza de una muestra de $n = 25$ son 75 y 100 respectivamente. Use el teorema de Tchebysheff para describir la distribución de las observaciones.

Solución Se tiene $\bar{y} = 75$ y $s^2 = 100$. La desviación estándar es $s = \sqrt{100} = 10$. La distribución de las observaciones está centrada en $\bar{y} = 75$ y el teorema de Tchebysheff establece que:

1. Al menos $3/4$ de las 25 observaciones caen en el intervalo $(\bar{y} \pm 2s) = [75 \pm 2(10)]$, esto es, de 55 a 95.
2. Al menos $8/9$ de las 25 observaciones caen en el intervalo $(\bar{y} \pm 3s) = [75 \pm 3(10)]$, es decir, de 45 a 105.

El énfasis hecho en el término “al menos” se debe a que el resultado del teorema de Tchebysheff es un tanto conservador ya que se aplica a cualquier conjunto de observaciones. En la mayoría de los casos la proporción de observaciones que caen en el intervalo correspondiente excede a $(1 - 1/k^2)$.

Se presenta ahora una regla que describe adecuadamente la variabilidad de una distribución acampanada y razonablemente bien la variabilidad de otras distribuciones que se acercan a esta forma. Una buena cantidad de mediciones de características de seres vivos y otras variables que se observan en la naturaleza siguen una distribución en forma de campana u otra forma similar a ésta. De aquí la importancia práctica de la siguiente regla que se conoce como la **regla empírica**.

La regla empírica

Dada una distribución de las observaciones con forma aproximadamente acampanada (como en la figura 3.15), entonces, el intervalo

$(\mu \pm \sigma)$ contiene aproximadamente al 68% de las observaciones

$(\mu \pm 2\sigma)$ contiene aproximadamente al 95% de las observaciones

$(\mu \pm 3\sigma)$ contiene casi todas las observaciones

La distribución acampanada de la figura 3.15 se conoce como la **distribución normal**. Esta distribución se discutirá en detalle en el capítulo 7. La importancia de la regla empírica consiste en su utilidad para describir adecuadamente la variación de un gran número de tipos de datos.

Ejemplo 3.6

Se realiza un estudio de tiempos y movimientos para determinar el tiempo necesario para realizar una operación de una fábrica. Se mide el tiempo necesario para realizar la operación para $n = 40$ obreros. Se calculan la media y la desviación estándar obteniéndose 12.8 y 1.7 respectivamente. Use la regla empírica para describir los datos en esta muestra.

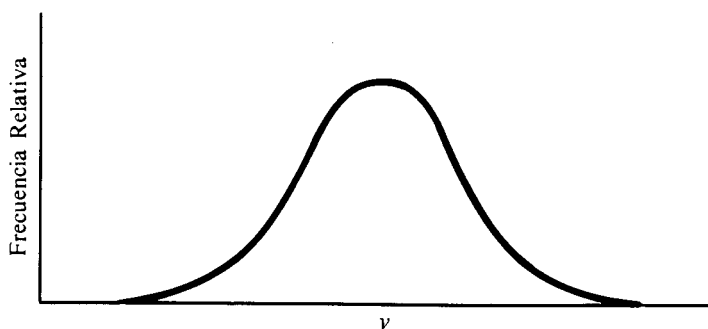


Figura 3.15 *La distribución normal*

Solución Para describir los datos se calculan los intervalos

$$(\bar{y} \pm s) = (12.8 \pm 1.7) \quad \text{o} \quad 11.1 \text{ a } 14.5$$

$$(\bar{y} \pm 2s) = [12.8 \pm 2(1.7)] \quad \text{o} \quad 9.4 \text{ a } 16.2$$

$$(\bar{y} \pm 3s) = [12.8 \pm 3(1.7)] \quad \text{o} \quad 7.7 \text{ a } 17.9$$

De acuerdo con la regla empírica se espera que aproximadamente el 68% de las observaciones estarán en el intervalo de 11.1 a 14.5, 95% de las observaciones estarán en el intervalo de 9.4 a 16.2, y casi todas ellas en el intervalo de 7.7 a 17.9.

Si se duda que la distribución de las observaciones tiene forma parecida a la acampanada, o por alguna otra razón se prefiere proceder conservadoramente, se puede entonces aplicar el teorema de Tchebysheff para estar absolutamente seguro de tales afirmaciones. El teorema de Tchebysheff asegura en este caso que al menos $3/4$ de las observaciones se encuentran en el intervalo de 9.4 a 16.2 y al menos $8/9$ de las observaciones en el intervalo de 7.7 a 17.9.

Es interesante observar como se comporta la regla empírica con el conjunto de datos de razón de costo-beneficio de la tabla 3.1. En la sección 3.9 se descubrirá que la media y la desviación estándar para las $n = 25$ observaciones son $\bar{y} = 16.0$ y $s = 5.6$. En la tabla 3.6 se exhiben los intervalos calculados y el número de observaciones que se observan en cada intervalo. En la primera columna de esta tabla aparece el índice k , en la segunda el intervalo $(\bar{y} \pm ks)$, en la tercera columna se tiene el número de observaciones en cada intervalo y la frecuencia relativa aparece en la cuarta columna. Note que el histograma de frecuencias de la figura 3.2 sólo tiene un ligero parecido a la forma de una campana. Sin embargo, los porcentajes que se observan en los correspondientes intervalos concuerdan razonablemente con los resultados de la regla empírica.

Tabla 3.6 *Frecuencias de observaciones que se encuentran a k desviaciones estándar de la media para los datos de la tabla 3.1*

k	INTERVALO ($\bar{y} \pm ks$)	FRECUENCIA EN INTERVALO	FRECUENCIA RELATIVA
1	10.4-21.6	16	.64
2	4.8-27.2	24	.96
3	-.8-32.8	25	1.00

Ejercicios

3.10. La mayoría de las escuelas de administración norteamericanas requieren que las personas que solicitan admisión presenten el resultado de una prueba diseñada para tal efecto. El promedio de este examen se ha calculado de 480 con una desviación estándar de 100. La experiencia anterior muestra evidencia de que la distribución de las calificaciones de este examen sigue una distribución de forma parecida a la acampanada.

- ¿Qué porcentaje de las calificaciones se espera encontrar en el intervalo de 380 a 580?
- ¿Qué porcentaje de las calificaciones se espera encontrar en el intervalo de 280 a 680?
- Una de las escuelas más prestigiosas sólo considera una solicitud si es acompañada por una calificación mayor o igual a 680. ¿Qué porcentaje de las personas que toman el examen cumplen con este requisito?

3.11. Las técnicas de control de calidad que se usan en el proceso de manufactura de un cierto producto,

sirven para garantizar la consistencia y uniformidad del producto terminado. Un ingeniero de control de calidad de una fábrica de botellas desea establecer normas de control. Con el proceso de fabricación bajo control se selecciona aleatoriamente una muestra de $n = 30$ botellas y se registran sus pesos. Se encuentra un peso promedio de $\bar{y} = 82$ gramos y una desviación estándar de 9 gramos. Para describir la distribución de pesos de las botellas cuando el proceso se encuentra bajo control

- Use el teorema de Tchebysheff.
 - Use la regla empírica. ¿Considera usted apropiado usar la regla empírica en este caso. ¿Por qué?
- 3.12. En el ejercicio 3.11, si la muestra seleccionada hubiera sido de sólo $n = 4$ botellas en lugar de $n = 30$, ¿consideraría usted adecuado usar la regla empírica para describir la variación de los pesos de las botellas? ¿Por qué?

3.9

Un método que facilita el cálculo de la varianza

Independiente del procedimiento que se use, el cálculo de la varianza y de la desviación estándar resulta un tanto tedioso. Particularmente si se emplea para su cálculo la regla que sugiere la definición, es decir, calculando cada una de las desviaciones respecto a la media como se ilustra en la tabla 3.4. Se hará uso de los datos de la tabla 3.4 para ilustrar un procedimiento de cálculo más corto. Los cálculos necesarios se muestran en la tabla 3.7. La primera columna en esta tabla contiene las observaciones mismas y la segunda columna las observaciones elevadas al cuadrado.

Tabla 3.7 Tabla para el cálculo simplificado de $\sum_{i=1}^n (y_i - \bar{y})^2$

y_i	y_i^2
5	25
7	49
1	1
2	4
4	16
Totales	19 95

Ahora se calcula

$$\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = 95 - \frac{(19)^2}{5} = 95 - \frac{361}{5} = 95 - 72.2 = 22.8$$

Nótese que este resultando es exactamente igual a la suma de los cuadrados de las desviaciones

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

que aparece en la tercera columna de la tabla 3.4. Esto no ocurre por casualidad. La suma de los cuadrados de las desviaciones es siempre igual a la fórmula en el siguiente cuadro.

Fórmula abreviada para calcular la suma de cuadrados de las desviaciones

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$

Esta fórmula puede demostrarse haciendo uso de los teoremas para sumatorias del capítulo 2. (Ver ejercicio 2.42.)

La suma de los cuadrados de las desviaciones, cuyo cálculo se facilita mediante la fórmula anterior, es necesaria en las fórmulas para la varianza y la desviación estándar. Esta fórmula facilita el cálculo puesto que elimina el cálculo de las desviaciones individuales. Otra ventaja de la fórmula es que, en general, proporciona una mejor precisión que cuando se calculan individualmente las desviaciones. Es común que el estudiante encuentre que la varianza que él ha calculado sea distinta de la respuesta del texto. Esto es debido generalmente al redondeo que se usa en los cálculos intermedios. Es conveniente que este tipo de redondeo de cifras sea evitado lo más posible, puesto que puede afectar seriamente el resultado del cálculo de la varianza. Otra ventaja es que el método corto es especialmente apropiado para el uso de calculadora, puesto que algunas de estas acumulan simultáneamente las cantidades

$$\sum_{i=1}^n y_i \quad \text{y} \quad \sum_{i=1}^n y_i^2$$

El procedimiento se ilustra ahora con los datos de las $n = 25$ tasas de costo-beneficio de la tabla 3.1. Tomando los datos de la tabla 3.1 se pueden verificar los siguientes cálculos

$$\sum_{i=1}^n y_i = 400 \quad \text{y} \quad \sum_{i=1}^n y_i^2 = 7154.02$$

Ahora usando la fórmula del procedimiento corto se tiene

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = 7154.02 - \frac{(400)^2}{25} \\ &= 7154.02 - 6400 = 754.02\end{aligned}$$

Se sigue que la desviación estándar (correcta hasta una cifra decimal) está dada por

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{754.02}{24}} = \sqrt{31.4175} = 5.6$$

Ejemplo 3.7 Calcule $\bar{y}$ y s para las observaciones 85, 70, 60, 90 y 81.

Solución

$$\bar{y} = \frac{386}{5} = 77.2$$

y_i	y_i^2
85	7,225
70	4,900
60	3,600
90	8,100
81	6,561
386	30,386

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \\ &= 30,386 - \frac{(386)^2}{5} \\ &= 30,386 - 29,799.2 \\ &= 586.8\end{aligned}$$

Por lo tanto

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{586.8}{4}} = \sqrt{146.7} = 12.1$$

Ejercicios

3.13 Calcule $\bar{y}$, s^2 , y s a partir de los siguientes datos:

2, 1, 3, 0, 4, 2, 1, 3

Usese la fórmula abreviada para calcular la suma de los cuadrados de las desviaciones.

3.14 Usando la fórmula abreviada para el cálculo de la suma de cuadrados de las desviaciones, calcule $\bar{y}$, s^2 , s para los siguientes datos:

4, 1, 0, 3, 6, 4

3.15 Calcule $\bar{y}$, s^2 , y s a partir de los siguientes datos

0, -2, -5, -3, 2, -4, 0, 3, 0, -1

3.16 Para los datos de las interrupciones de trabajo en el ejercicio 3.8 calcule la varianza del número de interrupciones diarias. ¿Cuál es la desviación estándar?

3.17. Los préstamos hipotecarios para compra de vivienda tienen distintas tasas de amortización dependiendo de la ubicación, del tipo de vivienda, del sueldo promedio del comprador y otros factores. Una muestra de $n = 25$ tasas de amortización para préstamos hipotecarios se presentan a continuación

5.2	6.0	7.5	8.0	5.0
7.9	6.6	9.2	7.4	6.5
8.0	9.0	7.3	7.5	7.0
4.6	8.5	5.5	9.3	9.5
6.5	7.5	6.5	8.1	8.2

a. Construya un histograma de frecuencias relativas para estos datos.

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

c. Calcule s^2 y s .

3.18 Con los datos del ejercicio 3.17 encuentre el percentil 50 para la distribución de estas tasas de interés. Encuentre el percentil 90 e interprete estos

valores que ha calculado. Encuentre también el primer y el tercer cuartil para estas tasas de interés.

3.19 Para los datos de las tasas de interés del ejercicio 3.17,

a. Use el teorema de Tchebysheff.

b. Use la regla empírica. ¿Considera usted que la regla empírica resulta apropiada para describir este conjunto de datos? ¿Por qué?

c. Calcule el porcentaje de observaciones que queden a una, dos y tres desviaciones estándar de la media. Compárelos con los porcentajes obtenidos mediante la regla empírica y con los que se esperan de acuerdo al teorema de Tchebysheff.

3.20 De acuerdo a la ética publicitaria, una planta refinadora de gasolina ha realizado $n = 10$ pruebas experimentales para comparar un aditivo nuevo con la marca regular ya en el mercado. Los incrementos en millas por galón en las 10 pruebas son los siguientes

1.5, 3.4, 0.2, 1.6, 2.0, 1.8, 2.9, 3.7, 0.4, 2.6

a. Calcule $\bar{y}$, el aumento promedio en millas por galón debido al nuevo aditivo.

b. Use el procedimiento corto de cálculo para calcular s^2 y s para estos datos.

3.10 Estimación de la media y la varianza a partir de datos agrupados (optativo)

Es común encontrarse con que los únicos datos disponibles para el análisis se nos proporcionen en la forma de un histograma de frecuencias. Los reportes de algunas compañías contienen, a veces, los datos listados en términos de frecuencias de clases, los datos proporcionados por el gobierno o por las agencias noticiosas son presentados, muchas veces, a manera de diagrama de barras u otro tipo de agregación de los datos. En tales casos no se conoce el valor exacto de las observaciones que caen en los intervalos de clase. Cuando esto ocurre no es posible calcular exactamente la media y la varianza de la muestra.

Se tiene, sin embargo, un método para calcular una aproximación de la media o de la varianza cuando la información disponible se tiene en forma de frecuencias de clase. **El método se basa en la suposición de que el punto medio de cada clase es aproximadamente igual a la media aritmética de las medidas contenidas en el intervalo.** El punto medio de la clase i se denota por m_i . Suponiendo que el punto medio de cada clase fuese realmente la media de las observaciones de la clase, entonces, para una clase particular, por ejemplo la clase i , si se multiplica m_i por f_i , la frecuencia de la clase i , se obtiene la suma de las f_i observaciones en esa clase i . Si se suma el total en cada

clase se obtiene la suma del total de observaciones que forman la distribución de frecuencias y $\bar{y}$ puede calcularse al dividir esta suma total de las observaciones entre el número total de observaciones, n , en la forma acostumbrada. Por supuesto, la bondad de la aproximación de la media obtenida por este método depende directamente del grado en el que los puntos medios de las clases representan adecuadamente al promedio de observaciones en la clase. Usualmente, las aproximaciones que se obtienen son adecuadas, principalmente cuando las frecuencias de clase son suficientemente grandes en relación con la longitud del intervalo. Para aproximar la varianza s^2 cuando se cuenta con datos agrupados se usa una fórmula que generaliza el procedimiento de cálculo para s^2 , presentado en la sección 3.9.

La Media y la Varianza de Datos Agrupados

Si los datos se encuentran agrupados de acuerdo a la frecuencia con que ocurren en cada una de k clases excluyentes, sin puntos comunes, la media $\bar{y}$ y la varianza s^2 de las observaciones que hayan sido agrupadas, pueden ser aproximadas por

$$\bar{y} \approx \frac{\sum_{i=1}^k f_i m_i}{n}$$

$$s^2 \approx \frac{\sum_{i=1}^k f_i m_i^2 - \left(\sum_{i=1}^k f_i m_i \right)^2 / n}{n - 1}$$

en donde m_i denota al punto medio de la clase i y f_i la frecuencia de observaciones dentro de la clase i . (Nota: El símbolo $\approx$ significa "es aproximadamente igual a".)

En la tabla 3.8 se resumen los cálculos necesarios para calcular $\bar{y}$ y s^2 a partir de la distribución de frecuencias de las $n = 25$ tasas de costo-beneficio de 25 acciones en el mercado de valores. Los cálculos necesarios para los datos agrupados se presentan en el cuadro anterior. El uso de estas fórmulas se simplifica cuando los datos se organizan como se muestra en la tabla 3.8.

Tabla 3.8 Frecuencias de clase y puntos medios de la clase para las 25 tasas de costo-beneficio de la tabla 3.2

CLASE i	FRONTERAS DE CLASE	f_i	m_i	$f_i m_i$	$f_i m_i^2$
1	5.00-8.99	3	7	21	147
2	9.00-12.99	5	11	55	605
3	13.00-16.99	7	15	105	1575
4	17.00-20.99	6	19	114	2166
5	21.00-24.99	3	23	69	1587
6	25.00-28.99	1	27	27	729
Totales		25		391	6809

De acuerdo a la fórmula la media puede ser aproximada a partir de estos datos agrupados de la siguiente manera

$$\bar{y} \approx \frac{\sum_{i=1}^6 f_i m_i}{25} = \frac{391}{25} = 15.64$$

y la varianza de estas observaciones se aproxima como sigue

$$\begin{aligned} s^2 &\approx \frac{\sum_{i=1}^6 f_i m_i^2 - \left[\left(\sum_{i=1}^6 f_i m_i \right)^2 / 25 \right]}{24} = \frac{6809 - [(391)^2 / 25]}{24} \\ &= \frac{693.76}{24} = 28.91 \end{aligned}$$

La desviación estándar aproximada es

$$s \approx \sqrt{28.91} = 5.377$$

En las secciones 3.6 y 3.9 se calcularon la media y la desviación estándar de las $n = 25$ observaciones sin agrupar, encontrándose

$$\bar{y} = 16 \quad y \quad s = 5.6$$

En este caso estos valores son aproximados satisfactoriamente por los valores de $\bar{y}$ y s calculados a partir de datos agrupados.

Aunque en la sección 3.2 se sugiere que las clases tengan todas la misma longitud, esto no es necesario para aplicar las fórmulas para datos agrupados. Lo que sí se supone es que los puntos medios de las clases son aproximadamente iguales a las medias de las observaciones dentro de las clases. Los procedimientos para datos agrupados no son aplicables en el caso en que una o más de las clases sea abierta (una frontera de clase en $-\infty$ o $+\infty$), puesto que, en este caso, resulta imposible encontrar los puntos medios para estas clases.

Ejercicios

3.21 Con los datos del ejercicio 3.3 y usando las fórmulas para datos agrupados calcule la media y la varianza aproximadas de los rendimientos por galón a partir del histograma de frecuencias relativas encontrado en el ejercicio 3.3(a). (Sugerencia: para encontrar la frecuencia en cada clase multiplique cada frecuencia relativa por 50.)

3.22 Use las fórmulas para datos agrupados para aproximar la media y la varianza de las tasas de amortización del ejercicio 3.17, usando la distribución de frecuencias obtenida en el ejercicio 3.17(a). Compare la aproximación con los valores para $\bar{y}$ y s^2 obtenidos en los ejercicios 3.17(b) y 3.17(c).

3.23 El contador de un almacén desea estimar el balance promedio, en dólares, de las 10,000 cuentas

de crédito que maneja el almacén. La distribución de frecuencias se presenta en la tabla y fue construida a partir de una muestra de $n = 100$ cuentas seleccionadas al azar de los archivos de crédito del almacén.

BALANCE DE LA CUENTA	FRECUENCIA
de \$0 a menos de 20	10
de 20 a menos de 40	15
de 40 a menos de 60	40
de 60 a menos de 80	22
de 80 a menos de 100	13
	100

- a. Use las fórmulas para datos agrupados para aproximar la media y la desviación estándar de los balances de estas cuentas.
- b. Explique por qué estos valores no son los verdaderos valores de $\bar{y}$ y s de los datos representados en el análisis.

- c. ¿Aproximadamente qué porcentaje de las cuentas de crédito tienen un saldo menor que 30 dólares?
- d. ¿Aproximadamente qué porcentaje de las 10,000 cuentas de crédito tienen un saldo mayor que 98 dólares?

3.11 Transformaciones lineales de los datos (optativo)

En ocasiones se desean hacer comparaciones entre dos o más conjuntos de datos, pero sucede que algunas veces los datos se encuentran representados en distintos sistemas de medidas. Un sindicato desea comparar los salarios de los obreros de dos plantas armadoras de automóviles una en Detroit y la otra en París. O un comprador desea comparar el rendimiento de un auto inglés expresado en kilómetros por galón imperial al rendimiento de un auto americano expresado en millas por galón. Para tales comparaciones es necesario transformar los conjuntos de datos a una escala de medida común. Lo que se debe hacer es encontrar una relación matemática entre uno de los conjuntos de datos, usado como referencia, y los otros conjuntos de datos. Por medio de esta relación, se transforman todos los datos a la escala de medida del conjunto seleccionado como referencia.

Si la comparación que se desea hacer entre los conjuntos de datos involucra comparaciones de medias, varianzas o desviaciones estándar, basándose en una escala de medida común, no es necesario transformar primero los datos a la escala común. Por ejemplo si el sindicato mencionado en el párrafo anterior toma una muestra de 50 obreros parisinos no es necesario convertir cada uno de los 50 salarios de francos a dólares antes de calcular las medidas descriptivas numéricas en dólares. Conociendo la media, la varianza o la desviación estándar en francos es posible convertirlos a sus equivalencias en dólares por medio del siguiente teorema, que puede llamarse el *teorema del cambio de escala*. La prueba se omite pero puede construirse a partir de las definiciones de media y varianza.

Teorema 3.1 Teorema del cambio de escala

Suponga que $\bar{y}$ y s_y^2 son la media y la varianza de un conjunto de n observaciones $y_1, y_2, y_3, \dots, y_n$. Si cada observación y_i es transformada mediante la transformación lineal $x_i = a + by_i$, donde a y b son números fijos, entonces, la media de los datos transformados es

$$\bar{x} = a + b\bar{y}$$

La varianza de los datos transformados es

$$s_x^2 = b^2 s_y^2$$

y la desviación estándar de los datos transformados es

$$s_x = |b|s_y$$

Aunque el teorema de cambio de escala se ha establecido para la media y la varianza muestrales, el teorema se aplica también para media y varianza de la población cuando las observaciones de ésta se transforman bajo una relación lineal.

Ejemplo 3.8 Los empleados de la planta armadora de la Volkswagen, en Munich, ganaron en 1973 un salario promedio de 15,200 marcos con una desviación estándar de 1,000 marcos. Para comparar estos salarios con los de los empleados de plantas similares en los Estados Unidos es necesario transformar la media y la desviación estándar de marcos a dólares. La tasa de cambio es de 1 marco = 0.39 dólares. Entonces en la fórmula para $\bar{x}$ se tiene $a = 0$ y $b = .39$ y entonces

$$\begin{aligned}\bar{x} \text{ (dólares)} &= .39\bar{y} \text{ (marcos)} = .39(15,200) = 5.928 \\ s_x \text{ (dólares)} &= .39s_y \text{ (marcos)} = .39(1,000) = 390\end{aligned}$$

Puesto que la relación entre dólares y marcos está dada por la multiplicación de una constante, se supone que la constante a de nuestro teorema es igual a cero. Si los datos se obtuvieron mediante la adición de una constante, la constante b del teorema se toma igual a uno y en tal caso la varianza permanece igual.

Ejemplo 3.9 La temperatura máxima diaria para Estocolmo, Suecia, para el mes de abril se ha estimado que tiene un promedio de 10 grados Celsius y una varianza de 50 grados cuadrados, Celsius. Si se desea comparar con el equivalente en grados Fahrenheit es necesario aplicar la transformación grados Fahrenheit = $32 + (9/5)$ grados Celsius. Usando el teorema del cambio de escala se tiene

$$\begin{aligned}\bar{x} \text{ (Fahrenheit)} &= 32 + \left(\frac{9}{5}\right) \cdot 10 = 50 \text{ grados Fahrenheit} \\ s_x^2 \text{ (Fahrenheit)}^2 &= \left(\frac{9}{5}\right)^2 \cdot 50 = 162 \text{ (Fahrenheit)}^2 \text{ grados}\end{aligned}$$

Ejercicios

3.24. Substraiga una constante apropiada a las observaciones: 27.995, 27.998, 28.005, 28.003, 28.004, calcule la media y la varianza de las observaciones transformadas y use luego el teorema del cambio de escala para obtener la media y la varianza de las observaciones originales.

3.25. Multiplique por una constante apropiada las siguientes observaciones muestrales: .029, .030, .028, .031, .032. Calcule media y varianza de las observaciones transformadas y use el teorema del cambio de escala para obtener la media y la varianza de las observaciones originales.

3.26. Un estudio económico muestra que el gasto en alimentos de una familia de cuatro personas es en promedio \$218 mensuales con una desviación estándar de \$32. Use el teorema del cambio de escala para

encontrar la media y la varianza del gasto diario en alimentos.

3.27. Las conversiones del sistema inglés al sistema métrico decimal son muy comunes. Use el teorema del cambio de escala para convertir las siguientes medidas descriptivas a su equivalente métrico.

- a. En 25 pruebas una nueva marca de auto americano se obtiene un rendimiento promedio de 27.8 millas por galón (mpg), con desviación estándar de 4.4 mpg. Expresé estos resultados en términos de kilómetros por galón (1 kilómetro = .622 millas).
- b. Expresé las estadísticas descriptivas de la parte a arriba, en kilómetros por litro (1 litro = 1.056 cuartos de galón).
- c. Una empacadora de mariscos reporta que el sal-

món pescado por pesqueros comerciales tiene un peso promedio de 9.8 libras por pez, con varianza de 17.3. Encuentre la media y la desviación estándar del salmón pescado por pesqueros comerciales expresadas en kilogramos (1 kilogramo = 2.205 libras).

d. Una compañía aserradora extranjera exporta troncos con una longitud promedio de 31 pies 6 pulgadas, con desviación estándar de 19 pies 2 pulgadas. Convierta estas estadísticas a metros (1 metro = 39.37 pulgadas, 1 pie = 12 pulgadas).

3.12

Resumen

El objetivo de un estudio estadístico es hacer inferencias acerca de una característica de una población, usando la información contenida en una muestra. En vista de que en este contexto las poblaciones son conjuntos de datos, se hace necesario contar con métodos que permitan describir conjuntos de observaciones.

Los métodos para describir conjuntos de observaciones pueden clasificarse en dos categorías, métodos gráficos y métodos numéricos. El histograma de frecuencias relativas es un método gráfico que resulta de gran utilidad para caracterizar conjuntos de observaciones. Hay otros métodos gráficos que resultan útiles en tanto proporcionen un resumen de los datos que sea fácil de leer e interpretar. Las medidas descriptivas numéricas son números que tienen por objeto proveer una imagen mental de la distribución de frecuencias. La atención se ha enfocado en las medidas de tendencia central y en las medidas de variación. Las más usuales entre éstas son la media y la desviación estándar. Aunque la moda no es precisamente una medida de tendencia central, se ha incluido debido a su importancia para caracterizar niveles de demanda de los productos de una empresa. Se hizo ver que la media tiene significado intuitivo propio, mientras que la desviación estándar adquiere una interpretación intuitiva al introducirse el teorema de Tchebysheff y la regla empírica. El objetivo central es el describir a la población de la cual se ha tomado la muestra. Para este fin la media y la varianza muestrales $\bar{y}$ y s^2 se usan como estimadores de la media y la varianza de la población μ y σ^2 .

Aunque el número de métodos descriptivos que aquí se introducen es considerable, éstos son sólo un pequeño porcentaje de todos los métodos que pudieran haber sido discutidos. También se han omitido muchas técnicas de cálculo que se acostumbra introducir en textos elementales. Estas omisiones se hacen necesarias debido a la restricción de tiempo que se tiene en los cursos elementales y debido también a que el uso de las computadoras ha minimizado la importancia de fórmulas especiales para facilitar los cálculos. Pero sobre todo la razón principal para la omisión de muchos de las técnicas descriptivas se debe a que a menudo distraen al estudiante y obscurecen el objetivo principal de la estadística moderna y de este texto—la inferencia estadística.

Ejercicios complementarios

3.28. Realice el siguiente experimento: lance una moneda 10 veces y registre el número y de caras obtenidas. Repita este proceso $n = 50$ veces y con estos

50 valores de y , construya un histograma de frecuencias relativas para estos datos.

3.29. La gran variedad de factores a considerar en

la compra de una vivienda, lugar, precio, tasa de amortización, tipo de construcción y otros, hacen que el tiempo que un comprador tarda en llegar a su decisión final sea muy variable. Los datos que se presentan representan la duración de la búsqueda (en semanas) de 25 compradores de vivienda en una cierta población.

15	17	7	15	20
5	3	19	10	3
11	10	4	8	13
9	15	6	2	8
12	1	2	13	4

- a. Construya un histograma de frecuencias para estos datos.
b. ¿Qué le dice a usted esta descripción gráfica acerca del tiempo de búsqueda que invierten los compradores de vivienda?

3.30. Los datos de la tabla presentan las emisiones de unos bonos de una empresa gubernamental y se indica cuantas emisiones fueron ofrecidas al público y cuantas colocadas en forma privada (los datos se dan en millones \$).

AÑO	TOTAL	CUENTAS OFRE- CIDAS AL PÚBLICO	CUENTAS COLO- CADAS EN FORMA PRIVADA
1965	\$12,456	\$ 4,389	\$8,067
1966	13,695	6,262	7,434
1967	17,470	10,883	6,597
1968	14,103	8,068	6,032
1969	14,310	9,636	4,671
1970	27,658	23,121	4,540
1971	26,490	19,777	6,714
1972	23,502	14,940	8,562
1973	20,286	12,404	7,882
1974	31,075	24,886	6,188

Construya un diagrama de barras como el de la figura 3.3 para describir estos datos.

3.31. Los $n = 6$ datos que se presentan representan el número de días laborales perdidos el último año por 6 empleados de una compañía:

3, 8, 4, 10, 5, 6

Calcule $\bar{y}$, s^2 y s para estos datos.

3.32. Las $n = 7$ observaciones representan la pérdida o ganancia de una acción en el mercado de valores al precio de cierre durante 7 días laborales consecutivos:

-1, 3, 4, 1, 0, -4, -3

Calcule $\bar{y}$, s^2 y s para estos datos.

3.33. Calcule $\bar{y}$, s^2 y s para los datos del ejercicio 3.28.

3.34. Calcule $\bar{y}$, s^2 y s para los datos del ejercicio 3.29.

3.35. En el histograma construido en el ejercicio 3.28 encuentre qué fracción de las observaciones caen en el intervalo $(\bar{y} \pm 2s)$. (Use los resultados del ejercicio 3.33.) Diga si este resultado es consistente con el teorema de Tchebysheff. Diga usted si considera que el histograma del ejercicio 3.28 tiene forma acampanada. Diga si considera que la regla empírica describe adecuadamente estos datos.

3.36. Repita las instrucciones del ejercicio 3.35 pero use el intervalo $(\bar{y} \pm s)$.

3.37. Para los datos del ejercicio 3.29, encuentre la proporción de mediciones que caen en los intervalos $(\bar{y} \pm s)$ y $(\bar{y} \pm 2s)$. Diga qué opina acerca de si estos resultados son o no consistentes con el teorema de Tchebysheff y con la regla empírica.

3.38. La distribución en la tabla muestra, en millas-tonelada los fletes movidos por los cinco medios primarios de transporte en los Estados Unidos en 1968.

MODO DE TRANSPORTE	MILLAS-TONELADA (BILLONES)
ferrocarril	756.8
carretera	396.3
agua	287.0
tubería	332.3
aérea	2.9

Construya un diagrama circular que indique los porcentajes del total de millas-tonelada para cada medio de transporte.

3.39. La distribución en la tabla muestra el total de los gastos en publicidad, por medio de difusión en Estados Unidos en 1970.

MEDIO DE DIFUSIÓN	CANTIDAD VENDIDA
Periódicos	\$4936
Revistas	1061
Televisión	2853
Radio	1128
Correo	2548
Aparadores	839
Otros	941

Construya un diagrama circular que represente estos datos.

3.40. Sugiera usted las fronteras de clase para clasificar cada uno de los siguientes conjuntos en 10 clases:

- El número de semanas antes de que una máquina lavadora falle por primera vez. Se tienen datos de 964 compradores, datos que van de 0 a 113 semanas.
- El gasto mensual en primas para seguro de vida de 583 empleados de una empresa comercial. Las cantidades varían de \$0 a \$467.
- Los salarios mensuales de los empleados de una importante compañía con 1280 empleados, que incluyen al presidente de la compañía y a otros ejecutivos. Los salarios varían entre \$6,850 y \$85,000.
- El número de clientes que compraron en un supermercado en los pasados 100 días, los datos varían de 92 a 471 clientes.

3.41. Calcule el recorrido para los datos del ejercicio 3.29. Calcule el cociente del recorrido entre s . Si se tuviera una gran cantidad de datos cuya distribución se asemeja en forma a una campana, ¿a cuántas desviaciones estándar se espera que el recorrido sea aproximadamente igual? (Nótese que esta respuesta proporciona una aproximación burda del valor de s .)

3.42. La regla práctica para estimar s con base en el recorrido, presentada en el ejercicio 3.41, puede extenderse para usarse con cantidades pequeñas de datos obtenidos al tomar una muestra de una distribución de forma acampanada. En estos casos el valor calculado de s deberá estar cercano al cociente del rango entre la constante apropiada que se encuentra en la siguiente tabla.

NÚMERO DE MEDICIONES	5	10	25
----------------------	---	----	----

Cociente esperado del recorrido a s	2.5	3	4
---------------------------------------	-----	---	---

- Use esta regla para estimar s con los datos del ejercicio 3.17. Compare esta estimación con el valor calculado de s .
- Hago lo mismo de la parte (a) para los datos del ejercicio 3.29.

3.43. Una muestra de 10 compañías farmacéuticas revela los siguientes porcentajes del gasto de promoción con respecto al total de ventas del producto:

21, 18, 25, 26, 25, 20, 24, 19, 28, 22

- Observe los datos y adivine el valor de s usando la aproximación del recorrido.
- Calcule $\bar{y}$ y s y compare este último con la aproximación del recorrido de la parte a.

3.44. A partir de los siguientes datos, un estudiante calculó s y obtuvo $s = 0.263$. ¿En qué se basaría usted para dudar de la exactitud de su cálculo? ¿Cuál es el valor correcto (aproximado a centésimos)?

17.2	17.1	17.0	17.1	16.9
17.0	17.1	17.0	17.3	17.2
17.1	17.0	17.1	16.9	17.0
17.1	17.3	17.2	17.4	17.1

3.45. Una máquina fabrica tornillos con diámetro promedio de 0.51 centímetros y desviación estándar de 0.01. Si la distribución de los diámetros es acampanada, ¿qué porcentaje de la producción total tendrá diámetros en el intervalo 0.49 a 0.53?

3.46. Con referencia al ejercicio 3.45. Suponga que las especificaciones para los tornillos requieren un diámetro para los tornillos de $(.5 \pm .02)$ centímetros. Los tornillos que no satisfacen este requerimiento se consideran defectuosos. Si la máquina funciona como en el ejercicio 3.45, ¿qué porcentaje de la producción se espera resulte defectuosa?

3.47. Los anuncios comerciales de una estación de televisión promedian 35 segundos en el aire con una desviación estándar de 5 segundos. Si la distribución del tiempo en el aire de los anuncios comerciales es de forma acampanada, ¿qué fracción del total de anuncios se espera que dure entre 30 y 40 segundos?

3.48. Se ha demostrado que los anuncios comerciales que duran más de 40 segundos son largos, costosos y poco efectivos para transmitir el mensaje central. ¿Qué proporción de los anuncios descritos en el ejercicio 3.47 pueden ser considerados demasiado largos?

3.49. La tasa de interés de diversas acciones industriales da un promedio de 8.55% con una desviación estándar de .70. Suponga que la distribución de estas tasas de interés sigue una distribución acampanada.

- Describa la distribución de las tasas de interés.
- ¿Qué porcentaje de los bonos se espera que tengan una tasa mayor de 9.95?

3.50. Un estudio de una compañía de seguros de daños revela que la reclamación promedio por daños a automóviles es de \$495 con una desviación estándar de \$75. Suponga que las reclamaciones siguen una distribución normal.

- Describa la distribución de las reclamaciones.
- ¿Qué porcentaje de las reclamaciones excede \$570?

3.51. La distribución de frecuencias de la tabla muestra la cantidad de impuestos pagados por concepto de posesión de bienes raíces en una colonia de 442 habitantes.

CANTIDAD PAGADA (DÓLARES)	FRECUENCIA
---------------------------	------------

Menos de 100	27
100-199	85
200-299	217
300-399	81
400-499	32

- a. Use las fórmulas para datos agrupados para obtener una aproximación de la media y la desviación estándar de los impuestos pagados por los residentes.
- b. Haga uso de la regla empírica para estimar la proporción de propietarios que pagan entre \$156 y \$347 de impuestos.

3.52. Use la regla empírica para estimar el percentil 84 con los resultados del ejercicio 3.51.

3.53. Use las fórmulas para datos agrupados para aproximar la media y la varianza de los datos de búsqueda de vivienda del ejercicio 3.29. Compare estos resultados con los $\bar{y}$ y s^2 calculados en el ejercicio 3.34.

3.54. Substraiga 10 unidades a cada una de las observaciones del ejercicio 3.29, calcule la media y la varianza de los datos transformados y use el teorema del cambio de escala para obtener media y varianza para los datos originales.

3.55. A cada observación del ejercicio 3.17 réstele primero 4.6 y al resultado multiplíquelo por 10. Calcule la media y la varianza de los datos transformados y use el teorema del cambio de escala para obtener la

media y la varianza de los datos originales.

3.56. Los Estados Unidos desean exportar una alfombra que vale 8.50 dólares por yarda cuadrada. ¿Cuál es el precio correspondiente por metro cuadrado? (1 metro cuadrado = 10.76 pies cuadrados, 1 yarda = 3 pies.)

3.57. Con referencia al ejercicio 3.56, suponga que las alfombras han de exportarse a la Gran Bretaña. Calcule el precio en libras esterlinas por metro cuadrado (1 libra esterlina = 1.70 dólares).

3.58. Si un automóvil norteamericano anuncia que su nueva máquina tiene 290 pulgadas cúbicas de desplazamiento en sus pistones, diga el equivalente en centímetros cúbicos (1 pulgada = 2.54 centímetros).

3.59. La temperatura mínima promedio en el mes de mayo, en un estado de la Unión Americana es de 50 grados Fahrenheit con desviación estándar de 18 grados Fahrenheit. ¿Cuáles son los correspondientes valores en grados Celsius?

3.60. Con referencia al problema 3.59, use la regla empírica con los datos transformados para determinar qué porcentaje de los días de mayo tuvieron temperaturas mínimas bajo 0° Celsius.

Práctica con Datos Reales

1. Seleccione usted una revista especializada en su área de trabajo o bien un documento gubernamental o el reporte de alguna empresa. Seleccione una muestra de $n = 25$ observaciones en una de las variables en el conjunto de datos.

- a. Defina la población de la cual la muestra fue tomada.
- b. Construya un histograma de frecuencias relativas para estos datos.
- c. Calcule $\bar{y}$ y s para estos datos.
- d. Diga usted si considera que estos datos tienen una distribución acampanada y por lo tanto son adecuadamente descritos por la regla empírica.
- e. Diga qué proporción de las observaciones cae a dos desviaciones estándar de $\bar{y}$, ¿a tres desviaciones? Diga si estos resultados se ajustan al teorema de Tchebysheff se ajustan a la regla empírica.

2. Realice el siguiente experimento para estimar el volumen diario de ventas de algún establecimiento comercial. Observe la cantidad de cada venta en 5

periodos distintos de 30 minutos. Construya un histograma de frecuencias relativas para representar sus datos.

- a. Use las fórmulas para los datos agrupados para aproximar la media y la desviación estándar del importe por venta, basándose en las frecuencias de clase.
- b. Encuentre la fracción de las observaciones que queda en el intervalo $(\bar{y} \pm s)$ y en el intervalo $(\bar{y} \pm 2s)$. Opine si la regla empírica proporciona una descripción satisfactoria de los datos.
- c. Conociendo el número de horas que el establecimiento seleccionado está abierto en la día de la investigación, se puede estimar fácilmente el ingreso total diario. Si se tienen n observaciones entonces $n\bar{y}$ estima el total de ventas en los 5 segmentos de 30 minutos observados. Si el establecimiento está abierto 10 horas, el total de ventas del día es estimado por $4n\bar{y}$, o en general si el establecimiento está abierto T horas el total de ventas es estimado por $(T/2.5)(n\bar{y})$. El qué tan confiable es esta estimación será discutido posteriormente en el texto.