

Unidad 3



Medidas de tendencia
central y de dispersión

Introducción

Los métodos tabulares y gráficos tienen algunas limitaciones para describir y analizar un conjunto de datos. Por ejemplo, si tenemos que realizar la descripción de un fenómeno ante un grupo de personas, estaríamos en seria desventaja si no contamos con el material y equipo necesario para elaborar tabulaciones o gráficos. Ante esta situación, acudimos al auxilio de otras herramientas proporcionadas por la estadística descriptiva: las medidas de tendencia central y de dispersión.

Las *medidas de tendencia central* son medidas descriptivas que señalan hacia dónde tienden a concentrarse los valores contenidos en un conjunto de datos. Su resultado debe ser un valor típico o representativo de la muestra o población, el cual es utilizado para describir o analizar un fenómeno. Al ser una idea abstracta y representativa del conjunto de datos, las medidas de tendencia central tienen la ventaja de poder ser transmitidas de manera verbal.

Por ejemplo, los medios de información dan a conocer el promedio semanal del índice de precios y cotizaciones de la bolsa de valores o el promedio mensual de las tasas de interés. Estos promedios son ejemplos de medidas de tendencia central, pues son datos típicos o representativos que nos describen la actividad bursátil en el piso de remates o el desempeño del mercado de dinero en un periodo determinado. Al ser una medida resumen puede ser transmitida con facilidad para dar una idea de la información contenida en un conjunto de datos.

Existen diversas medidas de tendencia central que son utilizadas según la naturaleza del fenómeno que se quiere investigar. Las medidas de tendencia central que se analizarán en esta unidad son:

- Media.
- Mediana.
- Moda.

Si bien, todas tienen como objetivo obtener un valor típico que describa hacia dónde se agrupan los valores de un conjunto de datos, cada una de ellas tiene ventajas y desventajas que hacen que las distingamos entre sí.

Sin embargo, en el análisis de muchos fenómenos también necesitamos conocer la manera en que los valores de una serie se dispersan entre sí. Para ello acudimos a otro tipo de medidas

descriptivas, las medidas de dispersión o de variabilidad, las cuales son tan importantes en el estudio de una serie de datos, como lo es localizar sus valores centrales.

Las medidas de dispersión proporcionan una idea mental con la cual se conoce cuánto varían o qué tanto se dispersan los valores de un conjunto de datos. Si la variación es muy pequeña, las medidas de dispersión también tendrían un valor muy pequeño e indicarían una gran uniformidad de los elementos de una serie. Por el contrario, si se obtiene un valor grande de las medidas de dispersión, señalaría gran variación entre los valores de los datos. La ausencia de dispersión es señal de uniformidad perfecta, lo cual quiere decir que todos los datos tienen el mismo valor.

En el estudio de algunos mercados las medidas de dispersión son utilizadas para medir la volatilidad, el nerviosismo o el riesgo que se presenta en una variable. Por ejemplo, cuando existe mucho nerviosismo entre los inversionistas en un mercado, se observará una enorme variación o volatilidad en sus precios.

Existen diversas medidas de dispersión que son utilizadas según la naturaleza del fenómeno que se quiere investigar. Las medidas de dispersión que se analizarán en esta unidad son:

- Rango.
- Varianza.
- Desviación estándar.
- Coeficiente de variación.
- Índice de asimetría.
- Índice de kurtosis.

3.1. Media, mediana y moda

También conocida como la media aritmética o el promedio, la media es la medida de tendencia central más utilizada en los negocios y en las ciencias sociales, pues se emplea con mucha frecuencia en trabajos empíricos. La media se utiliza únicamente para describir el comportamiento de variables cuantitativas.

Existen dos símbolos para representar a la media ($\bar{X}$ y μ). La $\bar{X}$ se refiere a un estadístico, es decir, es la media de una muestra; mientras que μ se refiere a un parámetro, es decir, es la media de una población. A la $\bar{X}$ se le conoce como la media muestral mientras que a la μ se le conoce como la media poblacional.

La manera de obtener la media muestral o poblacional depende de la forma como se encuentren organizados los datos, ya sea que estén no agrupados o agrupados. Se dice que trabajamos con *datos no agrupados* cuando se expone cada uno de los datos de la serie, mientras que los *datos agrupados* son aquellos que se encuentran organizados mediante tablas de frecuencias.

3.1.1. Media

a) Media para datos no agrupados

Cuando tenemos una serie con datos no agrupados $X_1, X_2, X_3, \dots, X_n$, la media se calcula sumando los valores de cada uno de los datos y su resultado se divide entre el número de datos que tiene la serie.

Para una población compuesta por los datos $X_1, X_2, X_3, \dots, X_N$, la fórmula de la *media poblacional* para datos no agrupados se describe de la siguiente manera:

$$\mu = \frac{(X_1 + X_2 + X_3 + \dots + X_n)}{N} = \frac{\sum X_i}{N}$$

Donde:

μ = Media aritmética de la población.

Σ = Suma.

N = Número de datos en la población.

X_i = El valor que toma cada uno de los datos.

Para una muestra que contenga $X_1, X_2, X_3, \dots, X_n$ datos, la *media muestral* para datos no agrupados se obtiene mediante la siguiente fórmula:

$$\bar{X} = \frac{(X_1 + X_2 + X_3 + \dots + X_n)}{N} = \frac{\Sigma X_i}{N}$$

Donde:

$\bar{X}$ = Media aritmética de la muestra.

Σ = Suma.

n = Número de datos incluidos en la muestra.

X_i = El valor que toma cada uno de los datos.

Ejemplo 1

En la tabla 3.1 se expone la cotización mensual del tipo de cambio entre el peso mexicano y el dólar estadounidense observada en algunas casas de cambio durante el año 2000.

- Si se realiza una inspección visual, ¿cuál sería tu opinión si alguien dijera que el tipo de cambio en el año 2000 estuvo alrededor de los 10.50 pesos por dólar?
- Encuentra la media para el tipo de cambio entre el peso y el dólar estadounidense en el año 2000.

Mes	Tipo de cambio en el 2000
Enero	9.47
Febrero	9.44
Marzo	9.29
Abril	9.37
Mayo	9.50
Junio	9.79
Julio	9.46
Agosto	9.28
Septiembre	9.33
Octubre	9.51
Noviembre	9.51
Diciembre	9.44

Fuente: Banco de México, www.banxico.org.mx

Tabla 3.1. Tipo de cambio mensual peso-dólar en el año 2000.

Contestando la pregunta del inciso a), desde luego que esta aseveración no es válida, pues en la tabla 3.1 los valores adquiridos por el tipo de cambio distan mucho de los 10.50 pesos por dólar. Si damos un vistazo a la tabla 3.1, podemos decir que los valores tienden a concentrarse alrededor de los 9.40 o 9.50 pesos por dólar.

Por lo tanto, es de esperarse que la media se encuentre muy cercana a los 9.40 o 9.50 pesos por dólar. Si nos preguntaran cuál sería un valor representativo o típico para describir el nivel del tipo de cambio durante el año 2000, llevamos a cabo la estimación de la media.

Debido a que el Banco de México únicamente seleccionó la paridad de algunas casas de cambio y no el total de las transacciones realizadas durante el año 2000, los datos de la tabla se refieren a una muestra. Adicionalmente, observamos que los datos no están agrupados, pues la tabla 3.1 no los organizó de acuerdo con su frecuencia, por lo que procedemos a estimar la media muestral para datos no agrupados de la siguiente manera:

$$\bar{X} = \frac{(9.47 + 9.44 + 9.29 + \dots + 9.44)}{12} = \frac{113.39}{12} = 9.44$$

El promedio del tipo de cambio durante el año 2000 fue 9.44 pesos por dólar, confirmando la apreciación hecha en el inciso a) de que el tipo de cambio estaría alrededor de los 9.40 o 9.50 pesos por dólar. El resultado 9.44 es utilizado como una medida típica o representativa que señala por dónde se concentraron las cotizaciones del dólar durante el año 2000. Si realizamos nuevamente una inspección visual a la tabla 3.1, se observa que en la mayoría de los meses existe un nivel cercano a los 9.44 pesos por dólar y únicamente durante el mes de julio la paridad se presionó ligeramente a los 9.79, como resultado del nerviosismo generado por las elecciones presidenciales del año 2000.

Ejemplo 2

En la tabla 3.2 se expone la participación mensual de la inversión extranjera en el mercado accionario de la Bolsa Mexicana de Valores, entre los meses de enero del año 2000 a octubre del 2001.

Encuentra el promedio de la participación extranjera en el mercado accionario para el periodo bajo estudio.

Mes	2000	2001
Enero	44.01	43.55
Febrero	46.58	40.17
Marzo	44.78	39.93
Abril	47.25	41.24
Mayo	45.07	41.21
Junio	46.69	40.95
Julio	44.07	39.87
Agosto	44.96	45.97
Septiembre	44.72	42.76
Octubre	44.62	43.85
Noviembre	43.03	
Diciembre	41.31	

Fuente: Bolsa Mexicana de Valores, www.bmv.com.mx

Tabla 3.2. Participación mensual de la inversión extranjera en la Bolsa Mexicana de Valores

En este ejemplo los datos tampoco se encuentran organizados mediante una tabla de frecuencias, por lo que se trata de un conjunto de datos no agrupados. Realizando una inspección visual, apreciamos que los valores se concentran alrededor de los números 43 o 44. Para confirmar lo

anterior, estimamos la media aritmética, pues en ocasiones resulta difícil determinar de manera visual hacia dónde se concentran los valores en un conjunto de datos

$$\mu = \frac{(44.01 + 46.58 + 44.78 + 47.25 + 45.07 + \dots + 43.85)}{22} = \frac{956.59}{22} = 43.48$$

Se puede decir que el promedio de la participación extranjera en el mercado accionario de la Bolsa Mexicana de Valores, entre enero del 2000 a octubre del 2001, fue de 43.48. Éste es un valor típico o representativo de la proporción de capitales extranjeros en la bolsa de valores, por lo que se puede decir que en este periodo 43.48% del capital negociado en el piso de remates fue de procedencia extranjera.

Ahora bien, ¿cómo podríamos mostrar de manera visual que la inversión extranjera representó un monto promedio de 43.48% respecto al total de las inversiones efectuadas en la bolsa de valores? Para ello construimos un gráfico de líneas en el que se muestren las participaciones mensuales de las inversiones extranjeras y su promedio en este periodo.

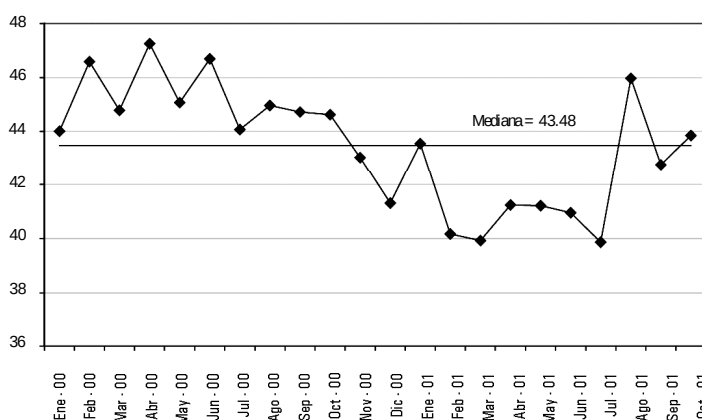


Gráfico 3.1. Participación mensual de la inversión extranjera en la Bolsa Mexicana de Valores

En el gráfico 3.1 observamos de manera visual el significado de la media de 43.48. Si bien es cierto que la participación extranjera en la Bolsa Mexicana de Valores tuvo un comportamiento irregular al presentarse una caída entre noviembre del 2000 a julio del 2001 como producto de la desaceleración económica mundial, la línea recta mostrada en la gráfica es una referencia que señala por dónde se concentró la participación extranjera en la bolsa de valores durante el periodo bajo estudio.

b) La media para datos agrupados

Cuando tenemos una serie con datos agrupados, es decir, que son presentados mediante una tabla de distribución de frecuencias, la media muestral $\bar{X}$ y la media poblacional μ se obtienen mediante las siguientes fórmulas:

$$\bar{X} = \frac{(m_1 \cdot f_1 + m_2 \cdot f_2 + \dots + m_n \cdot f_n)}{(f_1 + f_2 + \dots + f_n)} = \frac{\sum m_j f_j}{\sum f_j}$$

$$\mu = \frac{(m_1 \cdot f_1 + m_2 \cdot f_2 + \dots + m_n \cdot f_n)}{(f_1 + f_2 + \dots + f_n)} = \frac{\sum m_j f_j}{\sum f_j}$$

Donde:

$\bar{X}$ = Media aritmética de la muestra.

μ = Media aritmética de la población.

m_j = Punto medio para clase.

f_i = Frecuencia de cada clase.

$\sum f_i$ = Suma de las frecuencias de todas las clases

$\sum m_j f_i$ = Suma del producto de los puntos medios por las frecuencias de todas las clases

A diferencia de la fórmula para datos no agrupados, en este caso m_j representa el punto medio de cada clase, el cual se obtiene sumando el límite inferior y el límite superior de cada clase, y dividiendo este resultado entre 2.

Ejemplo 3

Una compañía aérea de transporte de paquetería desea conocer cuál es el peso promedio en kilogramos de los paquetes transportados, ya que de éste depende el costo y el número de paquetes que puede transportar sin violar los reglamentos de carga establecidos. Para ello, la compañía realizó un muestreo del peso en algunos paquetes cuyos resultados se presentan en la siguiente tabla de distribución de frecuencias

Peso en kg	f_i (frecuencia)
10.0 – 10.9	1
11.0 – 11.9	4
12.0 – 12.9	6
13.0 – 13.9	8
14.0 – 14.9	12
15.0 – 15.9	11
16.0 – 16.9	8
17.0 – 17.9	7
18.0 – 18.9	6
19.0 – 19.9	2

Tabla 3.3. Distribución de frecuencias de los paquetes transportados.

En este caso tenemos una serie con datos agrupados, pues sus valores son presentados mediante una tabla de distribución de frecuencias. Con los datos contenidos en la tabla 3.3 se puede obtener el punto medio de cada clase (véase la tabla 3.4), el cual sirve para el cálculo de la media aritmética.

Peso en kg	m_j (punto medio)	f_i	$m_j \cdot f_i$
10.0 – 10.9	10.45	1	10.45
11.0 – 11.9	11.45	4	45.8
12.0 – 12.9	12.45	6	74.7
13.0 – 13.9	13.45	8	107.6
14.0 – 14.9	14.45	12	173.4
15.0 – 15.9	15.45	11	169.95
16.0 – 16.9	16.45	8	131.6
17.0 – 17.9	17.45	7	122.15
18.0 – 18.9	18.45	6	110.7
19.0 – 19.9	19.45	2	38.9
		65	985.25

Tabla 3.4. Distribución de frecuencias del peso de los paquetes transportados, incluyendo el punto medio de cada clase.

Los resultados de la columna $m_j \cdot f_i$ se obtienen multiplicando cada uno de los puntos medios por la frecuencia de cada clase. Estos resultados se suman dando un monto de 985.25. Una vez realizadas estas operaciones procedemos a calcular la media muestral dividiendo 985.25 entre el monto obtenido por la suma de las frecuencias (65), tal como se señala en la siguiente fórmula:

$$\bar{x} = \frac{\sum m_j f_i}{\sum f_i} = \frac{985.25}{65} = 15.15$$

El peso promedio de los 65 paquetes transportados por esta compañía es de 15.15 kilogramos por paquete, lo que permitirá determinar el costo promedio de los paquetes que transporta esta compañía, además de conocer cuántos paquetes pueden ser transportados según el peso de carga permitido en cada vuelo que se realiza.

Ejemplo 4

De la información proporcionada por el XII Censo de Población y Vivienda, obtén la edad promedio de la población en México en el año 2000.

Eldades	Punto medio de clase m_j	Frecuencia f_i	$m_j \cdot f_i$
0 – 9 años	4.5	21 850 480	98 327 160.0
10 – 19 años	14.5	20 728 628	300 565 106.0
20 – 29 años	24.5	17 228 877	422 107 486.5
30 – 39 años	34.5	13 489 061	465 372 604.5
40 – 49 años	44.5	9 266 924	412 378 118.0
50 – 59 años	54.5	5 917 184	322 486 528.0
60 – 69 años	64.5	3 858 931	248 901 049.5
70 – 79 años	74.5	2 110 944	157 265 328.0
80 – 89 años	84.5	773 927	65 396 831.5
90 – más años	94.5	184 598	17 444 511.0
Total		95 409 554	2 510 244 723.0

Fuente XII Censo General de Población y Vivienda 2000, www.inegi.gob.mx

Tabla 3.5. Tabla de frecuencia de la población en México, incluyendo el punto medio de cada clase.

En este ejemplo se calcula la media poblacional μ para conocer la edad promedio en México, pues la información consultada fue obtenida de un censo de población. Cada uno de los puntos medios se multiplica por la frecuencia, que en este caso son los habitantes que corresponden a esa clase. Al obtener estos resultados, procedemos a calcular la media a través de la siguiente fórmula:

$$\mu = \frac{\sum m_j f_i}{\sum f_i} = \frac{2\,510\,244\,723}{95\,409\,554} = 26.31$$

La edad promedio de la población en México fue de 26.31 años, es decir, las edades de los habitantes en México tienden a concentrarse alrededor de los 26.31 años, lo que confirma la misma apreciación realizada en la unidad 2 de que la población en México está compuesta en su mayoría por gente joven. Incluso, se podría señalar que una persona con 26 años de edad es un habitante típico o representativo de la población en México.

Cabe señalar que en este cálculo fueron excluidas 2 073 858 personas que no especificaron su edad y suponemos que la marca de clase para las personas con 90 o más años es 94.5.

Ventajas y desventajas de la media

La media aritmética tiene diversas características que la hacen muy útil para los estudios realizados en los negocios y en las ciencias sociales.

1. Se puede calcular en cualquier conjunto de datos numéricos.
2. Un conjunto de datos numéricos tiene una y solo una media, de modo que siempre es única.
3. Toma en cuenta todos los datos de una muestra o población.

La media aritmética, en su carácter de ser un solo número que representa a todo conjunto de datos, tiene importantes ventajas.

- Primero, se trata de un término que es fácil de comprender y es intuitivamente claro.
- Segundo, todas las muestras o poblaciones tienen un valor promedio, evitando así confusiones en el análisis de datos.
- Por último, la media es útil para llevar a cabo procedimientos estadísticos como la comparación de medias entre diferentes conjuntos de datos.

El cálculo de la media se basa en todos los valores que toman los datos de una serie. Ninguna otra medida de tendencia central posee esta característica. Si bien es cierto que esta peculiaridad puede convertirse en una ventaja sobre otras medidas de tendencia central, la media aritmética resulta afectada por valores extremos o atípicos, es decir, por valores muy pequeños o valores demasiado grandes respecto al resto de los datos. En tales casos, la media aritmética representa una imagen distorsionada de la información que contienen los datos de un conjunto y no sería adecuado utilizarla para describir un fenómeno ni para ser empleada como una medida típica o representativa de una media o una población.

Ejemplo 5

Estima la media para la siguiente serie de datos 0, 1, 1, 3, 5 y 110.

Si se realiza una inspección visual se observa la presencia de un valor atípico, pues existe una gran diferencia entre los primeros cinco datos y el último dato de la serie, por lo que es de esperarse que la media aritmética no refleje un valor típico.

$$\mu = \frac{(0+1+1+3+5+110)}{6} = \frac{120}{6} = 20$$

Al obtener como resultado de la media aritmética un valor igual a 20, observamos que esta medida de tendencia central no cumple con su propósito de describir hacia dónde tienden a concentrarse los valores de una serie o de proporcionar un dato típico o representativo del conjunto de datos. De la serie de datos se puede observar que ningún valor se encuentra cercano al 20, por lo que este valor no puede ser representativo de la población. Esta distorsión es ocasionada por la presencia de un dato atípico en la serie de datos, que en este caso es 110.

Ante estas circunstancias necesitamos manejar otro tipo de medidas de tendencia central que no sean afectadas por valores atípicos. En el caso de la media aritmética su utilización únicamente es válida cuando los valores se encuentran muy cercanos entre sí, de lo contrario, no sería una medida de tendencia central confiable para analizar fenómenos.

Ejercicio 1

- De acuerdo con la información proporcionada por el Banco de México (www.banxico.org.mx) y el Instituto Nacional de Estadística, Geografía e Informática (www.inegi.gob.mx) en el año 2000 el Producto Interno Bruto a pesos corrientes fue de 5 432 354 825.00 miles de pesos y la población en el país era de 97 483 412 habitantes

Estima el Producto Interno Bruto per cápita o por habitante para el año 2000.

- Una alto ejecutivo se encuentra interesado en estudiar la maestría en negocios (Stanford Sloan Program) ofrecida por la Universidad de Stanford a personas con más de ocho años de experiencia en puestos de alta gerencia. Encuentra la edad promedio de los estudiantes de este programa de estudios, si se sabe que las edades de los estudiantes inscritos en este programa se encuentran distribuidas de la siguiente manera:

Edad	Número de estudiantes
30 – 34	18
35 – 39	18
40 – 44	10
45 – 50	2

Fuente www.gsb.stanford.edu/sloan

- El departamento de personal de una compañía ha tomado el tiempo que duran diferentes entrevistas de trabajo para que, de esa manera, se determine cuánto tiempo se debe destinar a cada entrevista. Para ello, se desea determinar la media. El tiempo de duración de cada entrevista observada (en minutos) es

37 30 23 46 42
 18 40 58 43 39
 55 64 42 28 21
 57 40 57 59 42
 35 26 13 42 38

- Una fábrica quiere conocer el tiempo que tardan 200 obreros en producir una pieza cada uno. Si la fábrica desea determinar el tiempo promedio que tarda cada obrero para establecer el tiempo de producción, con el fin de mejorar la eficiencia, calcula la media con la información de la siguiente tabla:

Tiempo de producción	f_i	m_i	$m_i f_i$	F_a
20.00 – 25.00	10	22.5	225	10
25.01 – 30.00	20	27.5	550	30
30.01 – 35.00	30	32.5	975	60
35.01 – 40.00	60	37.5	2250	120
40.01 – 45.00	50	42.5	2125	170
45.01 – 50.00	20	47.5	950	190
50.01 – 55.00	10	52.5	525	200
Σ	200		7 600	

Tiempo de producción de una pieza en minutos.

5. Una fábrica de ropa desea conocer cuántas chamarras terminadas y listas para ser entregadas produce en promedio, para de esta manera establecer un plan de ventas y mercadotecnia con la finalidad de lograr una mayor penetración en el mercado. Las chamarras terminadas y listas para ser entregadas por una fábrica de ropa por día contabilizadas durante un periodo de 20 días son:

142 163 108 157 124
 132 135 130 140 128
 136 133 146 137 149
 137 131 129 144 139

6. En la siguiente tabla se expone la distribución del tiempo que 75 clientes permanecieron en espera en la fila de un banco para pasar a cajas

Tiempo de espera	<i>fi</i>	<i>Fa</i>
0 – 14	7	7
15 – 29	19	26
30 – 44	27	53
45 – 59	13	66
60 – 74	6	72
75 – 89	3	75
Σ	75	

Tiempo de espera en un banco.

Si el banco quiere conocer el tiempo promedio que los clientes permanecen en espera en la fila para proporcionarles un mejor servicio, calcula la media.

3.1.2. La mediana (M_d)

Es una medida de tendencia central cuyo valor se encuentra exactamente a la mitad de una serie ordenada de datos. Por encima de la mediana se encuentra 50% de los datos con mayor valor de la serie y por debajo de ella 50% de los datos con menor valor de la serie. De esta forma, la mediana describe hacia dónde tienden a concentrarse los valores de una serie o de proporcionar un dato típico o representativo del conjunto de datos.

La mediana es representada por la expresión M_d y puede ser utilizada cuando la serie tiene valores extremos o atípicos, es decir, cuando existen diferencias significativas entre los valores que conforman la muestra o la población bajo estudio.

a) La mediana para datos no agrupados

Para encontrar la mediana muestral o poblacional de un conjunto de datos no agrupados se realizan los siguientes pasos:

1. Se ordenan los datos de la serie del valor más pequeño al valor más grande, es decir, se organiza la serie en orden creciente.
2. Observamos cuál es el tamaño de la muestra (n) o de la población (N) que se pretende analizar y procedemos a encontrar la mediana bajo uno de los siguientes criterios:
 - a) Si el total de datos analizados es un número impar, entonces la mediana es el valor que se encuentra exactamente en el centro de la serie ordenada. Es decir, es el valor del dato que ocupa la posición $\frac{(n+1)}{2}$ de la serie ordenada.
 - b) Si el total de datos analizados es un número par, entonces la mediana es el promedio de los dos valores que se encuentran en el centro de la serie ordenada. Es decir, es el promedio de los valores de los datos que ocupan las posiciones $\frac{n}{2}$ y $\frac{(n+2)}{2}$ de la serie ordenada.

Ejemplo 6

Estima la mediana para la serie de datos: 0, 1, 1, 3, 5, y 110.

Si se realiza una inspección visual se observa la presencia de un valor atípico, pues existe una gran diferencia entre los primeros cinco datos y el último dato de la serie, por lo que procedemos a calcular la mediana.

Al tener una serie con $n = 6$ (número par), promediamos los dos valores centrales de la serie ordenada y obtenemos la mediana:

$$M_d = \frac{1+3}{2} = \frac{4}{2} = 2$$

Como se puede apreciar, la mediana $M_d = 2$ no es afectada por la presencia de un dato atípico (110), por lo que puede ser utilizada como un dato típico o representativo del conjunto de datos.

Ejemplo 7

En la siguiente tabla se muestra el Índice de Precios y Cotizaciones (IPC) de la Bolsa Mexicana de Valores para cinco días del mes de noviembre del año 2001. Se desea conocer una medida de tendencia central del IPC para resumir el comportamiento bursátil durante esa semana.

Fecha	IPC
26/11/2001	5 759.49
27/11/2001	5 860.44
28/11/2001	5 848.21
29/11/2001	5 841.34
30/11/2001	5 832.83

Fuente: Bolsa Mexicana de Valores, www.bmv.com.mx

Tabla 3.6. IPC de la Bolsa Mexicana de Valores

Si se realiza una inspección visual a la tabla 3.6 se observa que el nivel del IPC del día 26 de noviembre representa un dato atípico (5 759.49 unidades), pues se encuentra muy por debajo del nivel registrado en el resto de la semana. En este caso la media no sería una medida de tendencia central apropiada para describir el nivel que el IPC mantuvo durante esta semana, por lo que conviene estimar la mediana.

1. Siguiendo los pasos para encontrar la mediana, ordenamos a la serie de datos del menor al mayor valor para quedar de la siguiente manera:

Posición	IPC	
1	5759.49	
2	5832.83	
3	5841.34	⇐ Mediana
4	5848.21	
5	5860.44	

Tabla 3.7. Serie en orden creciente del IPC.

2. Al tener un número de observaciones impar (son 5 observaciones) se procede a la aplicación de la siguiente fórmula:

$$N_d = \frac{(n+1)}{2} = \frac{(5+1)}{2} = \frac{6}{2} = 3$$

Donde N_d indica la posición del dato de la serie ordenada cuyo valor será la mediana.

El resultado anterior indica que se va a tomar el valor que se encuentre en la posición número tres de la serie ordenada, que en este caso viene representado por $M_d = 5841.34$. De esta manera se puede señalar que el nivel representativo del IPC de la Bolsa Mexicana de Valores observado durante la última semana del mes de noviembre de 2001 se ubicó en 5841.43 unidades. Alrededor de este número se ubicaron dos jornadas con valores superiores y dos jornadas con valores inferiores.

Ejemplo 8

En la siguiente tabla se muestra el tipo de cambio mensual observado por el Banco de México en algunas casas cambiarias del país durante el año 2000. Encuentra la mediana con la finalidad de que sea utilizada como medida representativa del tipo de cambio del año 2000.

Mes	Tipo de cambio en el 2000
Enero	9.47
Febrero	9.44
Marzo	9.29
Abril	9.37
Mayo	9.50
Junio	9.79
Julio	9.46
Agosto	9.28
Septiembre	9.33
Octubre	9.51
Noviembre	9.51
Diciembre	9.44

Fuente: Banco de México, www.banxico.org.mx

Tabla 3.8. Tipo de cambio mensual peso-dólar en el año 2000.

En esta información no se tiene la presencia de valores extremos o atípicos. No obstante se demostrará que cuando no se tiene la presencia de datos atípicos, el valor de la mediana es un muy cercano al valor de la media, es decir, ambas pueden ser utilizadas como medidas representativas de la serie de datos.

1. Siguiendo los pasos para encontrar la mediana, ordenamos a la serie de datos del menor al mayor valor para quedar de la siguiente manera:

Posición	Tipo de cambio en el 2000
1	9.28
2	9.29
3	9.33
4	9.37
5	9.44
6	9.44
7	9.46
8	9.47
9	9.50
10	9.51
11	9.51
12	9.79

← N_{d1}

← N_{d2}

Tabla 3.9. Tipo de cambio mensual peso-dólar en el año 2000.

2. Al tener un número de observaciones par (son 12 observaciones) se procede a la aplicación de la siguiente fórmula:

$$N_{d1} = \frac{n}{2} = \frac{12}{2} = 6 \quad N_{d2} = \frac{(n+2)}{2} = \frac{(12+2)}{2} = \frac{14}{2} = 7$$

Donde N_{d1} y N_{d2} indican la posición de los dos datos de la serie ordenada cuyos valores son utilizados para obtener la mediana. Ahora promediamos dichos valores y obtenemos la mediana.

$$M_d = \frac{(9.44 + 9.46)}{2} = \frac{18.9}{2} = 9.45$$

El resultado de la mediana es $M_d = 9.45$, que puede ser utilizado como un valor representativo del nivel que mantuvo el tipo de cambio entre el peso y el dólar durante el año 2000. También señala que 50% de los datos de la serie tiene un valor superior a 9.45 y el restante 50% tiene valores inferiores a 9.45. Observa que este valor difiere muy poco del valor obtenido en el ejemplo 1, donde la media muestral fue 9.44. Por esta razón, la media y la mediana son medidas de tendencia central que difieren muy poco cuando no se tiene la presencia de valores extremos o atípicos.

La mediana para datos agrupados

Cuando analizamos datos que se encuentran organizados mediante una tabla de frecuencias, la mediana para datos agrupados se obtiene utilizando la siguiente fórmula:

$$M_d = L_i + \left[\frac{\left(\frac{n}{2} - F_a \right)}{f_m} \right] I$$

Donde:

L_i = Límite inferior de la clase mediana.

n = Número de datos observados.

F_a = Frecuencia acumulada anterior a la clase mediana.

I = Amplitud del intervalo.

f_m = Frecuencia de la clase mediana.

Para localizar correctamente los componentes de esta fórmula debemos tomar en cuenta los siguientes puntos:

1. Las clases de la tabla de frecuencias deben estar organizadas en orden creciente y a la tabla se le debe adicionar una columna que contenga las frecuencias acumuladas de cada clase.
2. Identificamos la clase en donde se encuentra la mediana. Para ello se divide el total de datos que tiene la serie entre dos ($n/2$); posteriormente localizamos en la columna de las frecuencias acumuladas la clase en la que se encuentra el número ($n/2$).
3. Ésa es precisamente la clase donde se localiza la mediana, de la cual se toma su límite inferior (L_i), su frecuencia (f_m) y la amplitud del intervalo (I), el cual se obtiene de la diferencia entre el límite superior y el límite inferior de la clase.
4. El límite real inferior de la clase mediana (L_r) es un límite teórico que se obtiene sumando el límite inferior de la clase y el límite superior de la clase anterior y dividiendo esa suma entre dos.

$$\text{Límite real inferior} = \frac{\text{Límite inferior de clase} + \text{Límite superior de la clase anterior}}{2}$$

5. La amplitud del intervalo de la clase mediana (I) se obtiene de dos formas, ya sea con la diferencia de dos límites superiores de clase consecutivos o dos límites inferiores de clase consecutivos.
6. Se localiza la frecuencia acumulada inmediatamente inferior a la clase en donde se encuentra la mediana (F_a).

Cabe señalar que esta fórmula supone que los datos son continuos y que los valores observados dentro de cada clase forman una progresión aritmética

Ejemplo 9

Con el fin de conocer cuál es la situación del mercado laboral, una empresa recabó información de los salarios pagados en pesos por hora; esta información fue recolectada mediante una muestra de 100 obreros. Encuentra la mediana para determinar un salario representativo pagado por hora a los obreros. Los resultados de la muestra se observan en la tabla 3.10.

Salarios por hora	f_i	F_a
50 – 59.99	8	8
60 – 69.99	10	18
70 – 79.99	16	34
80 – 89.99	14	48
90 – 99.99	10	58
100 – 109.99	5	63
110 – 119.99	2	65
120 – 129.99	15	80
130 – 139.99	8	88
140 – 149.99	12	100
Σ	100	

← Clase mediana

Tabla 3.10. Distribución de frecuencias de los salarios pagados.

Con los datos presentados, el tamaño de muestra es $n = 100$. La clase mediana está definida por $n/2 = 100/2 = 50$, por lo que la clase que contiene la mediana es donde se encuentra la mitad de los obreros, siendo ésta la quinta clase en la cual los salarios fluctúan de 90 a 99.99 pesos por hora. El límite real inferior de la clase mediana se obtiene sumando el límite inferior de la clase mediana (90) al límite superior de la clase anterior a la mediana (89.99) y el resultado de esta suma se divide entre dos, dando $L_i = 89.995$. La frecuencia acumulada de la clase anterior a la clase mediana (80–89.99) es $F_a = 48$. La amplitud del intervalo de la clase mediana se define al hacer la diferencia de dos límites superiores de clases consecutivas, por ejemplo: $I = 99.99 - 89.99 = 10$ y la frecuencia de la clase mediana es $f_m = 10$.

$$M_d = L_i + \left[\frac{\left(\frac{n}{2} - F_a \right)}{f_m} \right] I = 89.995 + \left[\frac{\left(\frac{100}{2} - 48 \right)}{10} \right] 10 = 89.995 + \left[\frac{(50 - 48)}{10} \right] 10 = 89.995 + \left[\frac{2}{10} \right] 10$$

$$M_d = 89.995 + 2 = 91.995$$

El resultado obtenido por la empresa señala que 91.995 es el salario representativo de los obreros de esta empresa. Según la clase mediana del mercado laboral, 50% de los obreros perciben como máximo un salario de \$91.995 por hora y el 50% restante gana un salario mínimo de \$91.995.

Ejemplo 10

De acuerdo con la información proporcionada por el XII Censo de Población y Vivienda en México, encuentra la edad mediana para la población en México.

Eldades	Frecuencia f_i	Frecuencia acumulada
0 – 9 años	21 850 480	21 850 480
10 – 19 años	20 728 628	42 579 108
20 – 29 años	17,228,877	59 807 985
30 – 39 años	13 489 061	73 297 046
40 – 49 años	9 266 924	82 563 970
50 – 59 años	5 917 184	88 481 154
60 – 69 años	3 858 931	92 340 085
70 – 79 años	2 110 944	94 451 029
80 – 89 años	773 927	95 224 956
90 – más años	184 598	95 409 554
Total	95 409 554	

← Clase mediana

Fuente: XII Censo General de Población y Vivienda 2000, www.inegi.gob.mx

Tabla 3.11. Tabla de frecuencia de la población en México, incluyendo el punto medio de cada clase.

$$M_d = L_i + \left[\frac{\left(\frac{n}{2} - F_a \right)}{f_m} \right] I = 19.995 + \left[\frac{\left(\frac{95\,409\,554}{2} - 42\,579\,108 \right)}{17\,228\,877} \right] 10 = 19.995 + \left[\frac{5\,125\,669}{17\,228\,877} \right] 10$$

$$M_d = 19.995 + 2.975 = 22.48$$

La edad mediana en México es de 22.48, por lo que se puede decir que 50% de los habitantes en México tiene una edad mayor a los 22.48 años y el otro 50% tiene una edad menor a 22.48 años.

Ventajas y desventajas de la mediana

La mediana tiene diversas ventajas sobre otras medidas de tendencia central. Una de ellas es que nos señala el valor que se encuentra exactamente a la mitad de una serie ordenada de datos, por lo cual es considerada como el límite o el lindero que divide al 50% de los datos con mayor valor del 50% de los datos con menor valor.

La mediana también cuenta con algunas características de la media aritmética. Por ejemplo, también proporciona un solo número que representa a todo el conjunto de datos, por lo que es un término fácil de comprender y es intuitivamente claro; todas las muestras o poblaciones tienen una sola mediana; además, la mediana también es útil para la comparación de diferentes conjuntos de datos.

Sin embargo, la mediana no toma en cuenta todos los datos de una serie, sino únicamente el valor del dato que se encuentra exactamente a la mitad de la serie ordenada, en caso de que n sea impar, o los valores de los dos datos que se encuentran a la mitad de la serie ordenada, en caso de que n sea par. Esta peculiaridad puede considerarse como una ventaja o desventaja, dependiendo de la naturaleza del conjunto de datos.

Por ejemplo, a diferencia de la media, la mediana no se ve afectada cuando se tiene la presencia de datos extremos o atípicos, pues únicamente toma en cuenta uno o dos valores que se encuentran en el centro de la serie ordenada. Por esta razón, la mediana es la medida de tendencia central que más se utiliza cuando se tienen datos extremos.

Ejercicio 2

1. Una distribuidora de automóviles está interesada en conocer la eficiencia de diez de sus vendedores, según las ventas que realizan, con el fin de establecer cuántos autos es posible vender. El número de automóviles vendidos por cada vendedor es 2, 4, 7, 10, 10, 10, 12, 12, 14, 15. Calcula la mediana si ahora la distribuidora quiere conocer cuál es el número de autos vendidos más cerca del promedio.
2. Los pesos de una muestra de paquetes de una oficina de mensajería son: 21, 18, 30, 12, 14, 17, 28, 10, 16 y 25 kg. La oficina de paquetería quiere conocer el peso por paquete más cercano al peso promedio. Calcula la mediana.
3. Los salarios anuales (en pesos) de los ejecutivos de una corporación son 150 000, 100 000, 50 000, 40 000, 35 000, 35 000, 33 000, 30 000, 30 000, 30 000 y 28 000. Determina el salario que más se aproxima al promedio calculando la mediana.
4. El departamento de personal de una compañía ha tomado el tiempo que duran las entrevistas de trabajo para de esa manera determinar cuánto tiempo se debe destinar a cada entrevista. Para ello, se desea determinar la mediana. El tiempo de duración de cada entrevista (en minutos) es

37 30 23 46 42
 18 40 58 43 39
 55 64 42 28 21
 57 40 57 59 42

5. Una fábrica quiere conocer el tiempo que tardan 200 obreros en producir una pieza cada uno. Si la fábrica desea determinar el tiempo que más se acerca al tiempo promedio que tarda cada obrero para establecer el tiempo de producción con el fin de mejorar la eficiencia, calcula la mediana con la información de la siguiente tabla:

Tiempo de producción	f_i	m_i	$m_i f_i$	F_a
20.00 – 25.00	10	22.5	225	10
25.01 – 30.00	20	27.5	550	30
30.01 – 35.00	30	32.5	975	60
35.01 – 40.00	60	37.5	2250	120
40.01 – 45.00	50	42.5	2125	170
45.01 – 50.00	20	47.5	950	190
50.01 – 55.00	10	52.5	525	200
Σ	200		7 600	

Tiempo de producción de una pieza en minutos.

6. La siguiente tabla muestra la distribución de las cantidades de tiempo que un cliente permanece en espera en la fila de un banco para pasar a cajas de una muestra de 75 clientes.

Tiempo de espera	<i>fi</i>	<i>Fa</i>
0 – 14	7	7
15 – 29	19	26
30 – 44	27	53
45 – 59	13	66
60 – 74	6	72
75 – 89	3	75
Σ	75	

Tiempo de espera en un banco.

Si el banco quiere conocer el tiempo que más se acerca al tiempo promedio que permanecen los clientes en espera en la fila para proporcionarles un mejor servicio, calcula la mediana.

3.1.3. Moda

Es una medida de tendencia central cuyo valor es el más común en una serie de datos. La moda es representada por la expresión M_o y puede ser utilizada para describir series de datos con variables cuantitativas o variables cualitativas. En muchas ocasiones, esta medida es de gran utilidad en los negocios. Por ejemplo, algunas tiendas de autoservicio necesitan conocer cuál es el producto más demandado y en qué magnitud, con el propósito de tener al día sus inventarios.

a) La moda para datos no agrupados

La moda para datos no agrupados se define como el valor de la variable que se presenta con mayor frecuencia en una serie de datos.

Ejemplo 11

En la siguiente tabla se muestra el tipo de cambio mensual observado por el Banco de México en algunas casas cambiarias del país durante el año 2000. Encuentra la moda con la finalidad de que sea utilizada como medida representativa del tipo de cambio del año 2000.

Mes	Tipo de cambio
Enero	9.47
Febrero	9.44
Marzo	9.29
Abril	9.37
Mayo	9.50
Junio	9.79
Julio	9.46
Agosto	9.28
Septiembre	9.33
Octubre	9.51
Noviembre	9.51
Diciembre	9.44

Fuente: Banco de México, www.banxico.org.mx

Tabla 3.12. Tipo de cambio mensual en el 2000.

En este ejemplo se observa que los valores 9.44 y 9.51 aparecen en dos ocasiones cada uno, por lo que podemos señalar que en esta serie de datos existen dos modas $M_{o_1} = 9.44$ y $M_{o_2} = 9.51$, que son los datos más comunes o representativos del tipo de cambio durante el año 2000. Cuando existen dos modas en una serie de datos, como es el caso de este ejemplo, se dice que la serie es de tipo bimodal.

b) La moda para datos agrupados

Cuando se analizan *datos cualitativos* que están organizados mediante una tabla de frecuencias, la moda es la clase que tiene la mayor frecuencia.

Ejemplo 12

En el primer semestre del 2001, México colocó en las bolsas de Nueva York y Chicago 11 286 contratos de opciones “put” y “call” clasificados según el producto de la siguiente manera:

Producto	Contratos
Algodón	254
Café	1
Cártamo	7
Maíz	1,955
Sorgo	7,043
Soya	218
Trigo	1,808

Fuente: Claridades agropecuarias, ASERCA-SAGARPA, www.sagarpa.gob.mx

Tabla 3.13. Colocaciones de productos agrícolas

En este ejemplo se puede apreciar que el producto agrícola que más contratos de cobertura de precios celebró durante el primer semestre del año 2001 fue el sorgo con 7 043 contratos, convirtiéndose así en la moda de las colocaciones mexicanas en los mercados de futuros de las bolsas de Nueva York y Chicago.

Por otra parte, cuando se tiene la presencia de *datos cuantitativos* agrupados en una tabla de frecuencias, la moda se obtiene utilizando la siguiente fórmula:

$$M_o = L_i + \left[\frac{\Delta_1}{(\Delta_1 + \Delta_2)} \right] I$$

Donde:

M_o = Moda.

L_i = Límite real inferior de la clase modal (la que tiene la mayor frecuencia).

Δ_1 = Diferencia entre la mayor frecuencia y la frecuencia anterior.

Δ_2 = Diferencia entre la mayor frecuencia y la frecuencia que le sigue.

I = Amplitud del intervalo de la clase modal.

Ejemplo 13

Una casa de bolsa realizó un estudio comparativo de los rendimientos de ciertas acciones con el fin de conocer cuáles rendimientos fueron más atractivos para los compradores, según las acciones que fueron más vendidas. Mediante el cálculo de la moda determina el rendimiento de las acciones que fue más atractivo, considerando que la casa de bolsa elaboró la siguiente distribución sobre los rendimientos al vencimiento de una muestra de 65 acciones.

Rendimientos	f_i
50 – 59.99	8
60 – 69.99	10
70 – 79.99	16
80 – 89.99	14
90 – 99.99	10
100 – 109.99	5
110 – 119.99	2
Σ	65

← Clase modal

Tabla 3.14. Distribución de los rendimientos de acciones.

La clase que presenta una mayor frecuencia (16) es 70-79.99, por lo que el límite real inferior de la clase modal es $L_i = 69.995$. La diferencia entre la mayor frecuencia y la frecuencia anterior se define por: $\Delta_1 = 16 - 10 = 6$ y la diferencia entre la mayor frecuencia y la frecuencia posterior es $\Delta_2 = 16 - 14 = 2$. La amplitud del intervalo de clase donde se encuentra la mayor frecuencia es $I = 79.99 - 69.99 = 10$. En este caso, las clases muestran entre qué valores fluctúa el rendimiento más atractivo y la frecuencia representa el número de acciones que presentan tales rendimientos.

Al aplicar la fórmula de la moda con los datos anteriores se tiene:

$$M_o = L_i + \left[\frac{\Delta_1}{(\Delta_1 + \Delta_2)} \right] I = 69.995 + \left[\frac{6}{(6 + 2)} \right] 10 = 69.995 + \left[\frac{6}{8} \right] 10 = 69.995 + (0.75)(10)$$

$$M_o = 69.995 + 7.5 = 77.495$$

Debido a lo anterior el valor de la moda es igual a 77.495, por lo que la casa de bolsa puede concluir que el rendimiento que fue más atractivo para las 16 acciones que más se demandaron (frecuencia) es de 77.495.

Ejemplo 14

De acuerdo con la información proporcionada por el XII Censo de Población y Vivienda en México, encuentra la edad moda para la población en México.

Edades	Frecuencia f_i	Frecuencia acumulada
0 – 9 años	21 850 480	21 850 480
10 – 19 años	20 728 628	42 579 108
20 – 29 años	17 228 877	59 807 985
30 – 39 años	13 489 061	73 297 046
40 – 49 años	9 266 924	82 563 970
50 – 59 años	5 917 184	88 481 154
60 – 69 años	3 858 931	92 340 085
70 – 79 años	2 110 944	94 451 029
80 – 89 años	773 927	95 224 956
90 – más años	184 598	95 409 554
Total	95 409 554	

← Clase modal

Fuente: XII Censo General de Población y Vivienda 2000, www.inegi.gob.mx

Tabla 3.15. Tabla de frecuencia de la población en México, incluyendo el punto medio de cada clase.

La clase modal es (0 – 9), por lo que en este caso excepcional se toma el límite inferior $L_i = 0$, y no el límite real inferior. La razón radica en que la clase modal es la primera clase en la cual se encuentra contenido el número cero como límite inferior. En este caso no habría forma de tomar el límite real inferior para estimar la moda, pues al tratarse de un límite teórico, el límite real inferior resultaría un número negativo, el cual no tendría lógica alguna al estar manejando edades (no se puede hablar de edades negativas). Por otra parte, la diferencia entre la frecuencia mayor y su anterior es $\Delta_1 = 21\,850\,480 - 0 = 21\,850\,480$ y la diferencia con la posterior es: $\Delta_2 = 21\,850\,480 - 20\,728\,628 = 1\,121\,852$. El valor del intervalo de clase de la mayor frecuencia es: $I = 19 - 9 = 10$.

Al aplicar la fórmula de la moda con los datos anteriores se tiene:

$$M_o = L_i + \left[\frac{\Delta_1}{(\Delta_1 + \Delta_2)} \right] I = 0 + \left[\frac{21\,850\,480}{21\,850\,480 + 1\,121\,852} \right] 10 = 0 + (0.951)(10) = 9.51$$

$$M_o = 0 + 9.51 = 9.51$$

La moda de las edades en México es de 9.51 años.

Ventajas y desventajas de la moda

Al obtener la moda de un conjunto de datos pueden darse los siguientes casos:

1. Si no hay datos repetidos no existirá moda; por ejemplo, si se tienen los datos siguientes: 32, 45, 62, 35, 44.
2. Si hay datos repetidos que tengan valor cero, la moda es cero, pero no puede decirse que no hay moda; por ejemplo, si se tienen los siguientes datos de ventas de automóviles de lujo por día: 1, 0, 2, 0, 3, 0, 5.
3. Si hay más de un dato repetido igual número de veces existirá más de una moda, es decir, es una distribución multimodal, lo que representa una desventaja como medida de tendencia central; por ejemplo, si el siguiente conjunto de datos es el número de veces que aparece un comercial de tres productos (A, B, C) en la televisión en una hora: A, C, A, B, C, A, B, C, B. Con esos datos se tienen tres modas, ya que los comerciales de los productos A, B y C aparecen tres veces en una hora, por lo que la moda de los tres productos es tres.

La ventaja más sobresaliente de la moda es que puede ser utilizada para conocer una medida representativa de un conjunto de datos con valores cualitativos. Otra ventaja es que la moda no se ve afectada por datos extremos o atípicos. Sin embargo, la principal desventaja es que en algunas series de datos no existe la moda, lo que limita el propósito de conocer una medida representativa de un conjunto de datos.

Por último, se ha mencionado que en algunas series de datos puede presentarse el caso de que existen varias modas, lo que puede representar una ventaja o desventaja, dependiendo del problema que se estudie. La desventaja es que no tendríamos una medida representativa única de la serie de datos. Sin embargo, cuando la media y la mediana no son representativas, las modas pueden convertirse en las medidas más representativas para describir una serie de datos.

Ejercicio 3

1. Un distribuidor de automóviles está interesado en conocer la eficiencia de diez de sus vendedores, según las ventas que realizan con el fin de establecer cuántos autos es posible vender. El número de automóviles vendidos por cada vendedor es: 2, 4, 7, 10, 10, 10, 12, 12, 14, 15. Calcula la moda si la distribuidora de autos desea conocer el número de autos que más se vende.
2. Los pesos de una muestra de paquetes de una oficina de mensajería son: 21, 18, 30, 12, 14, 17, 28, 10, 16 y 25 kg. Calcula la moda si ahora la oficina de paquetería quiere conocer cuál es el peso por paquete que más se repite.
3. Los salarios anuales (en pesos) de los ejecutivos de una corporación son 150 000, 100 000, 50 000, 40 000, 35 000, 35 000, 33 000, 30 000, 30 000, 30 000 y 28 000. Calcula la moda para determinar cuál es el salario que predomina en la corporación.
4. El departamento de personal de una compañía ha tomado el tiempo que duran las entrevistas de trabajo para de esa manera determinar cuánto tiempo se debe destinar a cada entrevista. Calcula la moda para estimar el tiempo más usual que tarda una entrevista. El tiempo de duración de cada entrevista (en minutos) es

37 30 23 46 42
 18 40 58 43 39
 55 64 42 28 21
 57 40 57 59 42
 35 26 13 42 38

5. Una fábrica quiere conocer el tiempo que tardan 200 obreros en producir una pieza cada uno. Si la fábrica desea determinar el tiempo que más se repite, calcula la moda con la información de la siguiente tabla:

Tiempo de producción	f_i	F_a
20.00 – 25.00	10	10
25.01 – 30.00	20	30
30.01 – 35.00	30	60
35.01 – 40.00	60	120
40.01 – 45.00	50	170
45.01 – 50.00	20	190
50.01 – 55.00	10	200
Σ	200	

Tiempo de producción de una pieza en minutos

6. La siguiente tabla muestra la distribución de las cantidades de tiempo que los clientes permanecen en espera en la fila de un banco para pasar a cajas, la muestra es de 75 clientes

Tiempo de espera	f_i	F_a
0 – 14	7	7
15 – 29	19	26
30 – 44	27	53
45 – 59	13	66
60 – 74	6	72
75 – 89	3	75
Σ	75	

Tiempo de espera en un banco.

Calcula la moda para conocer el tiempo que más tardan los clientes del banco en espera.

3.2. Relación entre la media, la mediana y la moda

Cuando se tiene que decidir cuál medida de tendencia central es la mejor para describir la forma en que tienden a concentrarse los datos, la respuesta dependerá de la figura que adquiera la distribución de frecuencias de los datos, pues ésta hace posible comparar la media, la mediana y la moda de manera simultánea.

La distribución de frecuencias se encuentra muy relacionada con el histograma visto en la unidad pasada. El eje vertical representa las frecuencias que adquieren los valores de la serie de datos y el eje horizontal incluye los valores que toma la variable a lo largo de la serie. Si la serie está compuesta de muchos datos, se observa que la gráfica se encuentra más suavizada que lo observado en los histogramas de la unidad pasada. Las distribuciones de frecuencias pueden adquirir las siguientes figuras

- Simétrica con una sola moda.
- Simétrica con dos o más modas.
- Asimétrica con sesgo positivo o derecho.
- Asimetría con sesgo negativo o izquierdo.

Una distribución simétrica es muy fácil de identificar. Su gráfica tiene la característica de que una mitad de la distribución es idéntica a la otra mitad, con la salvedad de que sus posiciones son distintas. Es decir, si la gráfica de una distribución es dividida exactamente a la mitad, y la figura de la primera mitad es muy similar con la otra, se dice que tenemos una distribución simétrica.

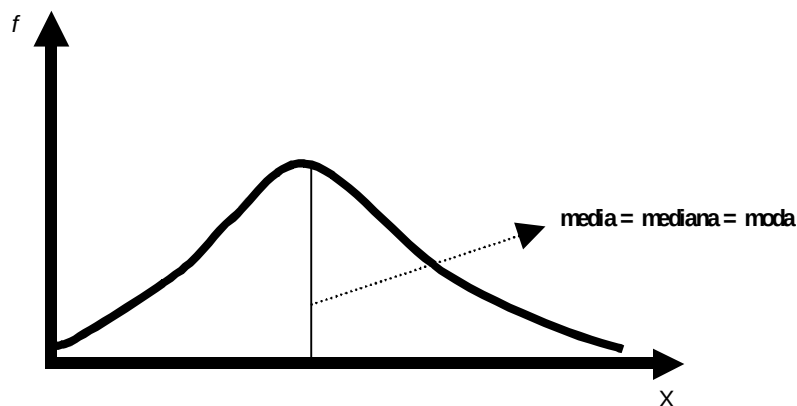


Figura 3.1. Distribución simétrica con una moda.

Por ejemplo, si trazamos una gráfica de distribución de frecuencias y la cortamos exactamente a la mitad, tal como se muestra en la figura 3.1, se puede observar que una mitad es idéntica a la otra, con la diferencia de que ocupan posiciones distintas. También se puede observar la existencia de una sola moda, pues únicamente existe una cima o "joroba" en la distribución de frecuencias (recuerda que la moda ocupa el valor donde se encuentra la mayor frecuencia).

Cuando se tiene una distribución perfectamente simétrica, media, mediana y moda coinciden en el mismo valor. En este caso daría lo mismo utilizar cualquiera de las tres medidas de tendencia central. Sin embargo, cuando la distribución de frecuencias no es exactamente simétrica y tiene una sola moda, es recomendable utilizar la mediana como la mejor medida de tendencia central.

En el caso de una distribución simétrica con dos o más modas es recomendable utilizar las modas como las mejores medidas de tendencia central, pues describe hacia dónde tienden a concentrarse los valores de la serie de datos.

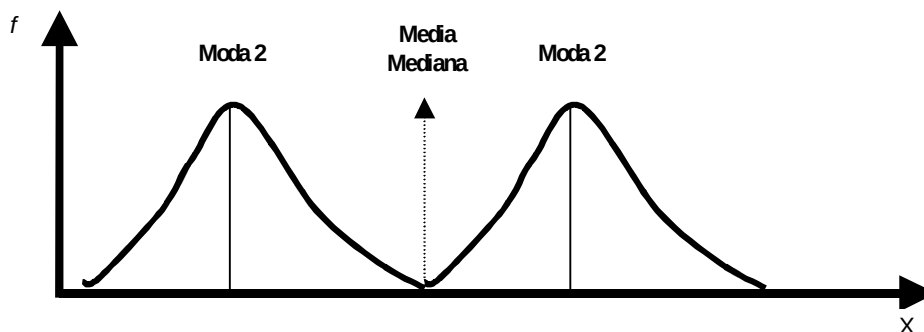


Figura 3.2. Distribución simétrica con dos modas.

En la figura 3.2. puede observarse una distribución simétrica con dos modas, las cuales nos señalan hacia dónde tienden a concentrarse los valores de los datos: hacia los valores de la moda 1 y de la moda 2. En este caso no sería recomendable tomar la media o la mediana como medidas de tendencia central, pues se aprecia que ningún dato tiende a agruparse alrededor de los valores de estas medidas descriptivas.

Si se divide una gráfica de distribución de frecuencias exactamente a la mitad, y una de ellas es muy distinta a la otra, se dice que es una distribución asimétrica. En estos casos se observará que la parte más alta o la cima de la figura queda cargada hacia uno de los lados, mientras que en el otro se observará que la figura tiende a alargarse dando el aspecto similar a una “cola”. A las distribuciones asimétricas también se le conoce como distribuciones sesgadas o distribuciones con algún tipo de sesgo.

Existen dos tipos de distribuciones asimétricas: las distribuciones con sesgo positivo o derecho y las distribuciones con sesgo negativo o izquierdo. En las distribuciones asimétricas con sesgo positivo o derecho se observará que la cola de la figura se encuentra a la derecha de la distribución, mientras que en su parte izquierda se ubicará la cima o el valor más alto de la distribución. En este caso, el valor de la media es superior al valor de la mediana; también se observará que el valor de la mediana es superior a la moda, tal como se señala en la figura 3.3.

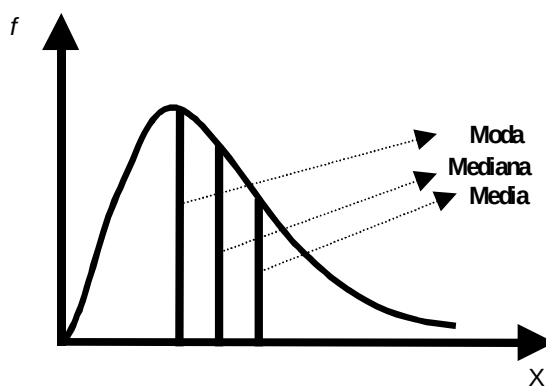


Figura 3.3. Distribución asimétrica positiva.

En las distribuciones asimétricas con sesgo negativo o izquierdo se observará que la cola de la figura se encuentra a la izquierda de la distribución, mientras que en su parte derecha se ubicará la cima o el valor más alto de la distribución. En este caso, el valor de la media es inferior al valor de la mediana; también se observará que el valor de la mediana es inferior a la moda, tal como se señala en la figura 3.4.

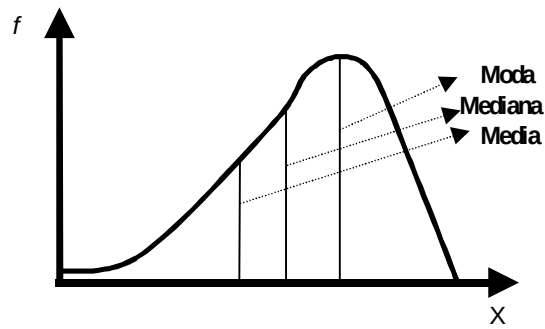


Figura 3.4. Distribución asimétrica negativa.

Cuando se tienen distribuciones asimétricas se señala que existe la presencia de valores extremos o atípicos en la serie de datos. Los valores atípicos se encuentran cargados hacia el lado de la cola. Por esa razón, el lado de la cola es el mismo hacia donde apunta el sesgo de la distribución, pues es en ese lugar donde se encuentran los valores extremos o atípicos.

Cuando se tiene la presencia de una distribución asimétrica no es recomendable utilizar la media como medida de tendencia central, pues al tener valores atípicos, obtendríamos una medida distorsionada. En el caso de distribuciones asimétricas es recomendable utilizar la mediana como la mejor medida de tendencia central, pues no se toman en cuenta los valores extremos de la serie de datos.

Ejercicio 4

1. En una distribución simétrica:
 - a) Media, mediana y moda son diferentes
 - b) Media, mediana y moda coinciden en el mismo valor.
 - c) La media es mayor que la mediana y la moda.
 - d) La moda es mayor que la media y la mediana.
2. En una distribución asimétrica sesgada hacia la derecha:
 - a) La mediana es mayor que la media y la moda.
 - b) Media, mediana y moda coinciden en el mismo valor.
 - c) La media es mayor que la mediana y la moda.
 - d) La moda es mayor que la mediana y la moda.
3. En una distribución asimétrica sesgada hacia la izquierda:
 - a) La mediana es mayor que la media y la moda.
 - b) Media, mediana y moda coinciden en el mismo valor.
 - c) La media es mayor que la mediana y la moda.
 - d) La moda es mayor que la mediana y la media.
4. De los ejemplos 4, 10 y 14 se sabe que la edad media de la población en México es $\mu = 26.31$, la edad mediana es $M_o = 22.48$ y la edad modal es $M_u = 9.51$.
 - a) Elabora la gráfica de distribución de frecuencias para la población en México, utilizando la información contenida en los ejemplos 4, 10 y 14.
 - b) Señala qué tipo de sesgo se observa en la gráfica de distribución de frecuencias para la población en México.

3.3. Cuartiles, deciles y percentiles

Una vez localizado el centro de la distribución de un conjunto de datos, el siguiente paso es analizar más detalladamente la manera en que se distribuye el resto de los valores. Por ejemplo, en algunas ocasiones resulta importante conocer la manera en que quedan distribuidos los datos de acuerdo con ciertos porcentajes que se observan en la serie de datos. Lo anterior también proporciona una imagen mental de la distribución de frecuencias.

En adición a las medidas de tendencia central, hay algunas medidas útiles de posición “no central” que suelen utilizarse al resumir o descubrir propiedades de grandes conjuntos de datos. A estas medidas se les denomina cuantiles. Algunos de los cuantiles más empleados son los cuartiles, los deciles y los percentiles, medidas que hacen posible un análisis más detallado de una distribución, representando qué porcentaje de los datos es más pequeño (si están a su izquierda) y qué porcentaje de los datos es más alto en valor (si están a su derecha).

En tanto que la mediana divide una distribución en dos partes iguales, donde 50% de los datos son menores y el otro 50% de los datos son mayores, los cuartiles son medidas descriptivas que dividen la distribución en cuatro partes, los deciles la dividen en diez partes y los percentiles la dividen en cien partes.

Cuartiles (Q)

Los cuartiles son aquellos valores que dividen una distribución de datos en cuatro partes y se representan por Q_1 , Q_2 y Q_3 , denominados primero, segundo y tercer cuartil, respectivamente.

Existen tres cuartiles, el primer cuartil (Q_1) es un punto tal que deja a la izquierda 25% de los datos que son menores que él y es menor que 75% de los datos restantes. El segundo cuartil (Q_2) tiene un valor igual a la mediana. El tercer cuartil Q_3 tiene un valor tal que sobrepasa en valor a 75% de los datos y es menor que el 25% restante.

Lo anterior se puede apreciar en la figura siguiente:

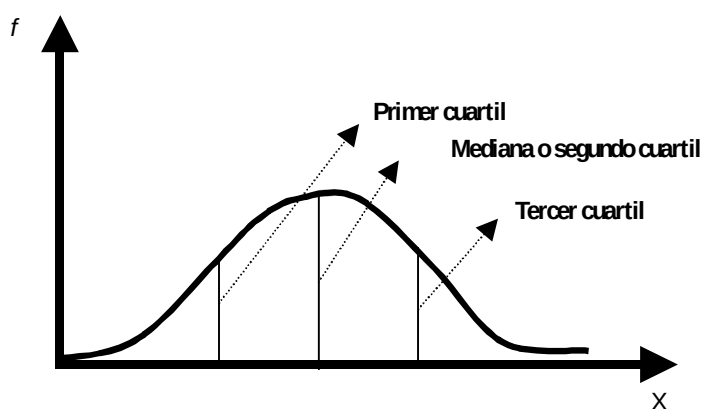


Figura 3.5. Cuartiles

En la figura anterior 25% del área queda a la izquierda del primer cuartil, mostrando que un cuarto del conjunto de datos tiene un valor menor y 75% a la derecha indica que tres cuartas partes de los datos son superiores en valor. El tercer cuartil muestra que 75% del área queda a la izquierda, con lo que tres cuartas partes de los datos son de menor valor y 25% a la derecha mostrando que una cuarta parte de los datos tiene un valor superior.

Los cuartiles para datos no agrupados en una serie se localizan de la siguiente manera: primero se ordenan los valores observados de acuerdo con su magnitud y, posteriormente, se determina el lugar que cada cuartil debe ocupar en la serie.

El lugar que debe tomar el primer cuartil se obtiene dividiendo el número de datos (n) entre cuatro. Esto se debe a que el valor de este cuartil deja a la izquierda 25% de los datos que son más pequeños y a la derecha 75% de los datos con valores mayores. La posición del primer cuartil se define por:

$$N_{OQ_1} = \frac{n}{4}$$

El lugar que debe ocupar el segundo cuartil se define dividiendo el número de datos (n) entre dos, ya que al ser igual que la mediana deja a la izquierda 50% de los datos menores y a la derecha 50% de los datos con mayores valores. Por ello, la fórmula para determinar la posición del segundo cuartil es:

$$N_{OQ_2} = \frac{(2 \cdot n)}{4} = \frac{n}{2}$$

El lugar que le corresponde al tercer cuartil se obtiene multiplicando el número de datos (n) por tres y dividiendo entre cuatro, debido a que considera que a su izquierda se encuentra 75% de los datos más pequeños y a la derecha 25% de los datos con valores mayores, siendo la fórmula para definir al tercer cuartil:

$$N_{OQ_3} = \frac{(3 \cdot n)}{4}$$

Ejemplo 15

El departamento de recursos humanos de una empresa desea dividir en cuatro partes iguales las solicitudes de empleo que recibe constantemente, con el fin de determinar los días en que la carga de trabajo aumenta. Para ello tomó una muestra de 18 días hábiles donde la cantidad de solicitudes de empleo, ordenadas de manera ascendente, fueron: 22, 26, 28, 31, 33, 34, 37, 39, 49, 50, 52, 59, 60, 62, 67, 69, 74 y 76. Para esto, se quiere hacer suposiciones mediante el cálculo de los cuartiles.

Los números de orden para cada uno de los cuartiles son:

Para el primer cuartil $N_{OQ_1} = \frac{18}{4} = 4.5$

Para el segundo cuartil $N_{OQ_2} = \frac{18}{2} = 9$

Para el tercer cuartil $N_{OQ_3} = \frac{(3 \cdot 18)}{4} = \frac{54}{4} = \frac{27}{2} = 13.5$

Los números de orden 4.5, 9 y 13.5 indican los lugares que ocupan en la serie ordenada cada uno de los cuartiles.

Para obtener los valores de los cuartiles de esta serie de datos se procede de la manera siguiente:

El primer cuartil está situado entre el cuarto y el quinto término, se suma el valor de estos términos y la suma se divide entre dos, lo cual da:

$$\frac{(31 + 33)}{2} = \frac{64}{2} = 32 \text{ que es el valor de } Q_1.$$

Esto quiere decir que 25% de los días se recibe menos de 32 solicitudes, mientras que 75% se recibe más de 32. Mostrando que el mínimo de solicitudes que se recibió en un día fue de 22 y el máximo fue 76, de lo que podemos concluir que el departamento tuvo una mayor carga de trabajo 75% de las veces.

El segundo cuartil tiene el número de orden 9, por lo tanto tiene como valor 49 que es el localizado en el noveno lugar, indicando que 50% de los días se recibe menos de 49 solicitudes y el otro 50% más de 49 solicitudes.

El tercer cuartil está entre el término 13 y 14, lo cual da $Q_3 = \frac{60+62}{2} = \frac{122}{2} = 61$. Por lo tanto, 75% de los días recibe menos de 61 solicitudes, mientras que sólo 25% de los días recibe más de 61 solicitudes.

El número de orden que ocupan los cuartiles para una serie de datos agrupados en una serie de frecuencias se obtiene mediante las relaciones $n/4$, $n/2$ y $3n/4$. Al tener estos números de orden se procede a buscar la frecuencia acumulada que los contenga. Una vez localizada esa frecuencia se elige la clase que contiene los distintos valores de la variable y el valor que corresponde a ese renglón es el valor del cuartil.

Este método exige que los datos sean continuos y que los valores observados en cada clase se distribuyan regularmente (en forma de progresión aritmética). Para situar cada uno de los cuartiles, primero hay que encontrar los números de orden que dividen a la serie en cuatro partes iguales, mediante las relaciones $n/4$, $n/2$ y $3n/4$. Posteriormente, se aplica la fórmula:

$$Q_i = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I$$

Donde:

L_i = Límite real inferior de la clase donde se encuentra el cuartil.

N_o = Lugar o posición que le corresponde al cuartil.

F_a = Frecuencia acumulada anterior a la clase donde se encuentra el cuartil.

I = Amplitud del intervalo donde se ubica el cuartil.

f_c = Frecuencia de la clase donde está el cuartil.

La cual es semejante a la utilizada en el cálculo de la mediana.

Ejemplo 16

Se desea conocer a partir de los cuartiles Q_1 , Q_2 y Q_3 la variación existente entre los salarios pagados por hora a 65 obreros. Los datos se presentan a continuación y se retoman de la tabla siguiente.

Salarios	f_i	F_a
50 – 59.99	8	8
60 – 69.99	10	18
70 – 79.99	16	34
80 – 89.99	14	48
90 – 99.99	10	58
100 – 109.99	5	63
110 – 119.99	2	65
Σ	65	

Tabla 3.16. Distribución de salarios pagados por hora.

$$N_{Q_1} = n/4 = 65/4 = 16.25$$

El número de orden 16.25 queda dentro de la segunda frecuencia acumulada, que es 18, que corresponde a la segunda clase de 60.00 a 69.99.

Si se aplica la fórmula, el resultado es

$$Q_i = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I$$

$$L_i = 59.995$$

$$N_o = 16.25$$

$$F_a = 8$$

$$f_c = 10$$

$$I = 69.99 - 59.99 = 10$$

Los datos se obtienen de la manera siguiente: la posición del primer cuartil es $N_o = 16.25$, por lo que la frecuencia de la clase donde se encuentra el primer cuartil es $f_c = 10$ y la frecuencia acumulada es $= 18$, correspondientes a la segunda clase $60 - 69.99$. Como el cuartil se encuentra en la segunda clase, la frecuencia acumulada de la clase anterior es $F_a = 8$; el límite real inferior de la segunda clase es $L_i = 59.995$ y la amplitud del intervalo de esa clase se obtiene restando al límite superior, de esa clase, el límite superior de la clase anterior.

Por lo tanto:

$$Q_1 = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I = 59.995 + \left[\frac{(16.25 - 8)}{10} \right] 10 = 59.995 + \left[\frac{8.25}{10} \right] 10 = 59.995 + \left[\frac{82.5}{10} \right]$$

$$Q_1 = 59.995 + 8.25 = 68.245$$

El dato muestra que 25% de los obreros recibe un salario por hora menor que 68.245 pesos, mientras que 75% recibe un salario mayor.

Para el cuartil 2:

$N_o Q_2 = n/2 = 65/2 = 32.5$ que se localiza en la tercera frecuencia acumulada que es 34, correspondiente a la tercera clase, por lo tanto:

$$L_i = 69.995$$

$$N_o = 32.5$$

$$F_a = 18$$

$$f_c = 16$$

$$I = 79.99 - 69.99 = 10$$

Sustituyendo en la fórmula.

$$Q_2 = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I = 69.995 + \left[\frac{(32.5 - 18)}{16} \right] 10 = 69.995 + \left[\frac{14.5}{16} \right] 10 = 69.995 + \left[\frac{145}{16} \right]$$

$$Q_2 = 69.995 + 9.0625 = 79.0575$$

Con esto se concluye que 50% de los obreros recibe un salario por hora menor que 79.0575 pesos, mientras que el otro 50% recibe un salario por hora mayor.

Para el cuartil 3:

$N_o Q_3 = 3n/4 = 3(65)/4 = 195/4 = 48.75$ que se localiza en la quinta frecuencia acumulada que es 58, correspondiente a la quinta clase, por lo tanto, el valor del tercer cuartil es

$$L_i = 89.995$$

$$N_o = 48.75$$

$$F_a = 48$$

$$f_c = 10$$

$$l = 99.99 - 89.99 = 10$$

Sustituyendo en la fórmula:

$$Q_3 = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] l = 89.995 + \left[\frac{(48.75 - 48)}{10} \right] 10 = 89.995 + \left[\frac{0.75}{10} \right] 10 = 89.995 + \left[\frac{7.5}{10} \right]$$

$$Q_3 = 89.995 + 0.75 = 90.745$$

El 75% de los obreros recibe un salario por hora menor que 90.745 pesos y 25% recibe un salario mayor.

Deciles

Los deciles son aquellos valores que dividen en diez partes una serie de datos y se representan por $D_1, D_2, \dots, D_9$, denominados primer decil, segundo decil, ..., noveno decil.

Si se desea dividir la serie ordenada de observaciones en diez partes iguales, resultan los deciles, desde el primero hasta el noveno, que dejan desde 10% hasta 90% de observaciones con categorías menores, respectivamente.

Para datos no agrupados, el primero, segundo, tercero, ..., noveno decil son los valores que se obtienen para los números de orden $n/10, 2 \cdot n/10, \dots, 9 \cdot n/10$ de los casos observados comenzando por la primera clase.

Ejemplo 17

Considerando el ejemplo 9 de las solicitudes de empleo recibidas por el departamento de recursos humanos de una empresa se pide calcular del decil D_1 al D_5 , con el fin de conocer las variaciones que presenta la distribución. Los datos son: 22, 26, 28, 31, 33, 34, 37, 39, 49, 50, 52, 59, 60, 62, 67, 69, 74 y 76.

Los números de orden para cada decil son:

$$N_{oD_1} = \frac{n}{10} = \frac{18}{10} = 1.8$$

$$N_{oD_2} = \frac{(2 \cdot n)}{10} = \frac{(2 \cdot 18)}{10} = \frac{36}{10} = 3.6$$

$$N_{oD_3} = \frac{(3 \cdot n)}{10} = \frac{(3 \cdot 18)}{10} = \frac{54}{10} = 5.4$$

$$N_{oD_4} = \frac{(4 \cdot n)}{10} = \frac{(4 \cdot 18)}{10} = \frac{72}{10} = 7.2$$

$$N_{oD_5} = \frac{(5 \cdot n)}{10} = \frac{(5 \cdot 18)}{10} = \frac{90}{10} = 9$$

El primer decil muestra que le corresponde la posición 1.8 que está situada entre el primero y el segundo dato (22 y 26), por lo que su valor es:

$$D_1 = \frac{(22+26)}{2} = \frac{48}{2} = 24$$

El decil muestra que 10% de los días se recibió 24 solicitudes o menos y 90% se recibió 24 solicitudes o más.

El segundo decil muestra que le corresponde la posición 3.6 por lo que su valor se encuentra entre el 28 y el 31, por lo que podemos tomar 30 como una aproximación del segundo decil. De esto se desprende que 20% de los días se recibió 30 solicitudes o menos y 80% se recibió 30 solicitudes o más.

Al trabajar deciles para datos agrupados es necesario seguir con una metodología similar a la de la mediana y de los cuartiles. Por ello, la fórmula para obtener el valor de los deciles es

$$D_1 = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I$$

Donde:

L_i = Límite real inferior de la clase donde se encuentra el decil.

N_o = Lugar o posición que le corresponde al decil.

F_a = Frecuencia acumulada anterior a la clase donde se encuentra el decil.

I = Amplitud del intervalo donde se ubica el decil.

f_d = Frecuencia de la clase donde está el decil.

Ejemplo 18

Retomando los datos del ejemplo 16 y aplicando la fórmula para interpolar (datos agrupados), que es la misma que la que se aplicó en el caso de los cuartiles, calcular los valores de los deciles 1, 2 y 5.

Salarios	f_i	F_a
50 – 59.99	8	8
60 – 69.99	10	18
70 – 79.99	16	34
80 – 89.99	14	48
90 – 99.99	10	58
100 – 109.99	5	63
110 – 119.99	2	65
Σ	65	

Tabla 3.17. Distribución de salarios pagados por hora.

Los datos para obtener el valor del primer decil son los siguientes:

$$N_o = 6.5$$

$$L_i = 49.995$$

$$F_a = 0$$

$$f_d = 8$$

$$I = 69.99 - 59.99 = 10$$

Estos datos se obtienen de la manera siguiente: la posición del primer decil es $N_o = 6.5$, por lo que la frecuencia de la clase donde se encuentra el primer decil $f_d = 8$ y la frecuencia acumulada es $F_a = 8$, correspondiente a la primera clase que es 50 – 59.99. Como el decil se encuentra en la primera

dase, la frecuencia acumulada de la dase anterior es $F_a = 0$, el límite real inferior de la primera dase es $L_i = 49.995$ y la longitud del intervalo de esa dase se obtiene restando al límite superior de la siguiente dase, el límite superior de esta dase.

$$D_1 = 49.995 + \left[\frac{(6.5 - 0)}{8} \right] 10 = 49.995 + \left[\frac{6.5}{8} \right] 10 = 49.995 + \left[\frac{65}{8} \right] = 49.995 + 8.125 = 58.125$$

El primer decil muestra que 10% de los obreros recibe 58.12 pesos o menos por hora y 90% recibe 58.12 pesos por hora o más.

$$D_2 = 59.995 + \left[\frac{(13 - 8)}{10} \right] 10 = 59.995 + \left[\frac{5}{10} \right] 10 = 59.995 + \left[\frac{50}{10} \right] = 59.995 + 5 = 64.995$$

Del decil dos se tiene que 20% de los obreros recibe 64.995 pesos por hora o menos, mientras que 80% recibe 64.995 pesos o más.

$$D_5 = 69.995 + \left[\frac{(32.5 - 18)}{16} \right] 10 = 69.995 + \left[\frac{14.5}{16} \right] 10 = 69.995 + \left[\frac{145}{16} \right] = 69.995 + 9.0625 = 79.0625$$

El quinto decil muestra que 50% de los obreros recibe por hora 79.06 pesos o menos y el otro 50% recibe por hora 79.06 pesos o más.

Percentiles

El percentil p es un valor tal que a lo más p por ciento de los datos es menor que él y a lo más $(100 - p)$ por ciento de los datos es mayor.

Por ejemplo, el percentil 90 para un conjunto de datos es un valor que excede 90% de los datos y es menor que 10% de los datos.

En ocasiones se acostumbra también dividir una serie ordenada de observaciones en 100 partes iguales, dando lugar a los percentiles, desde el 1º hasta 99º, que dejan desde 1% hasta 99% de observaciones con categorías menores. El primero, segundo, tercero, ..., nonagésimo noveno percentil, son los valores que corresponden a los números de orden $n/100$, $2n/100$, $3n/100$, ..., $99n/100$ de los casos observados, comenzando por la primera dase.

La fórmula que define el cálculo de los percentiles es

$$P_i = L_i + \left[\frac{(N_o - F_a)}{f_p} \right] I$$

Donde:

L_i = Límite real inferior de la dase donde se encuentra el percentil.

N_o = Lugar o posición que le corresponde al percentil.

F_a = Frecuencia acumulada anterior a la dase donde se encuentra el percentil.

I = Amplitud del intervalo donde se ubica el percentil.

f_p = Frecuencia de la dase donde está el percentil.

Ejemplo 19

Considerando los datos de la tabla 3.17, el percentil 35, representado por P_{35} , es el valor que se obtiene para el número de orden $35n/100$, en este caso $35(65)/100 = 22.75$, que se considera está contenido en la tercera frecuencia acumulada 34 correspondiente a la tercera clase, aplicando la fórmula se obtiene:

$$P_{35} = 69.995 + \left[\frac{(22.75 - 18)}{16} \right] 10 = 69.995 + \left[\frac{4.75}{16} \right] 10 = 69.995 + \left[\frac{47.5}{16} \right] = 69.995 + 2.975 = 72.975$$

$$P_{75} = 89.995 + \left[\frac{(48.75 - 48)}{10} \right] 10 = 89.995 + \left[\frac{0.75}{10} \right] 10 = 89.995 + \left[\frac{7.5}{10} \right] = 89.995 + 0.75 = 90.74$$

De aquí que 35% de los trabajadores gana \$72.955 o menos, 75% gana \$90.74 o menos, y así sucesivamente.

Ejercicio 5

- Una fábrica quiere conocer el tiempo que tardan 200 obreros en producir una pieza cada uno. Si la fábrica desea determinar la variación que existe en el tiempo de producción al respecto tiempo promedio que tarda cada obrero, con el fin de mejorar la eficiencia, con los datos siguientes calcula:
 - El cuartil 1.
 - El decil 4.
 - El percentil 63.

Tiempo de producción	f_i	F_a
20.00 – 25.00	10	10
25.01 – 30.00	20	30
30.01 – 35.00	30	60
35.01 – 40.00	60	120
40.01 – 45.00	50	170
45.01 – 50.00	20	190
50.01 – 55.00	10	200
Σ	200	

Tabla 3.18. Tiempo de producción de una pieza en minutos.

- La siguiente es la distribución de las cantidades de tiempo que un cliente permanece en espera en la fila de un banco para pasar a cajas de una muestra de 75 clientes.

Tiempo de espera	f_i	F_a
0 – 14	7	7
15 – 29	19	26
30 – 44	27	53
45 – 59	13	66
60 – 74	6	72
75 – 89	3	75
Σ	75	

Tabla 3.19. Tiempo de espera en un banco.

El banco desea conocer cuál es la variación en el tiempo de espera en la fila. Calcula:

- El cuartil 3.
- El decil 5.
- El percentil 36.

3.4. Rango, varianza y desviación estándar

3.4.1. Rango

También conocido con el nombre de amplitud o recorrido, el *rango* se define como la diferencia que existe entre el valor máximo y el valor mínimo de un conjunto de datos. Es la medida de dispersión más fácil de calcular, y es especialmente útil en aquellas situaciones en que el objetivo de la investigación sólo consiste en averiguar el alcance de las variaciones extremas.

Por ejemplo, el desempeño del precio de las acciones en el mercado bursátil se suele reconocer por los rangos, al citar los precios máximos y mínimos de cada sesión. Es decir, la variación en el precio de una acción puede medirse obteniendo el rango existente entre los dos valores más extremos y así interpretar qué tanta volatilidad manifestó la acción en una jornada o periodo. Si se comparan dos acciones, se puede interpretar que la acción que tiene mayor variación es aquella que tiene mayor rango.

Ejemplo 20

Una compañía de seguros desea conocer la variación que existe en las ventas de sus ocho vendedores y de esa manera determinar la productividad de cada uno de ellos. Calcula el rango empleando la siguiente información de seguros vendidos durante un mes: 8, 11, 5, 14, 11, 8, 11, 16.

Si se desea hallar el *rango* de tales observaciones sólo hay que identificar el valor máximo (16) y el valor mínimo (5) y obtener la diferencia entre ellos:

$$\text{Rango} = \text{Valor máximo} - \text{Valor mínimo} = 16 - 5 = 11$$

El *rango* es 11, lo cual quiere decir que la diferencia entre el número de seguros vendidos por dos vendedores distintos, el mejor vendedor y el peor vendedor, es de 11, indicando una gran dispersión o variabilidad, ya que sería ilógico que si un vendedor logra vender 16 seguros, el otro sólo venda 5 si se trata de los mismos seguros. Lo anterior puede atribuirse a la experiencia, a la capacitación o a la cartera de clientes que cada vendedor tiene.

Ejemplo 21

Un analista desea comparar el desempeño de la Bolsa Mexicana de Valores de dos meses: septiembre y octubre de 2001. Para esto toma su principal indicador, el Índice de Precios y Cotizaciones (IPC), y obtiene las siguientes gráficas.

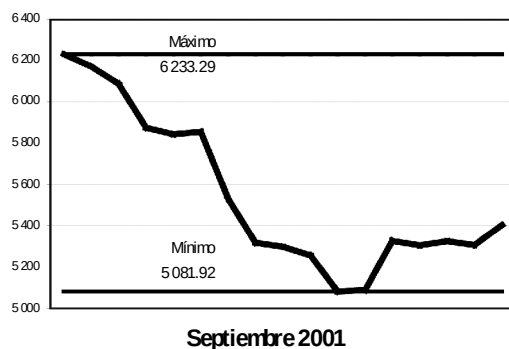


Figura 3.6. Bolsa Mexicana de Valores en septiembre de 2001.

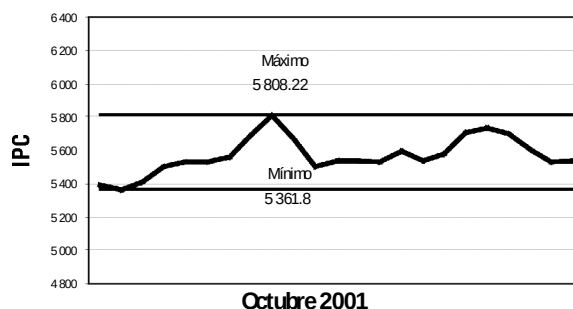


Figura 3.7. Bolsa Mexicana de Valores en octubre de 2001.

Si se desea conocer en cuál de los dos meses se presentó mayor volatilidad en el mercado de valores encontramos los rangos del IPC en cada uno de ellos

Rango en septiembre 2001 = $6\,233.29 - 5\,081.92 = 1\,151.37$

Rango en octubre 2001 = $5\,808.22 - 5\,361.8 = 446.42$

Se puede decir que en el mes de septiembre de 2001, la Bolsa Mexicana de Valores registró mayor volatilidad que en el mes de octubre, pues su rango de 1 151.37 fue superior al observado durante el mes de octubre de 446.42.

Este resultado también puede apreciarse de manera visual en las figuras 3.6. y 3.7., donde los rangos se representan por el diferencial existente entre el nivel máximo y el nivel mínimo del IPC. En el mes de septiembre se observa un rango mucho más ancho que el del mes de octubre, el cual se atribuyó al nerviosismo generado por los ataques terroristas del día 11 de septiembre en el Pentágono y en el World Trade Center de Nueva York.

Ventajas y desventajas del rango

La principal ventaja del rango radica en que es la medida de dispersión más fácil de obtener, pues únicamente se toman los dos valores extremos y se diferencian entre sí. Además, al medir la amplitud entre los dos valores más extremos en una serie de datos, esta medida de dispersión suele ser muy útil cuando se desea conocer qué tan extremos son los límites máximos y mínimos de una variable; por ejemplo, las temperaturas de ciertas ciudades del país o la ganancia de las casas de cambio que se obtienen diferenciando los precios de compra y los precios de venta para cada divisa.

Sin embargo, el hecho de que se tomen en cuenta únicamente los dos valores más extremos de un conjunto de datos, el rango puede ser una medida de dispersión que resulta afectada ante la presencia de datos atípicos.

Ejercicio 6

1. El rango se define como:
 - a) La amplitud entre el valor más grande y el valor más pequeño de la serie de datos.
 - b) La suma del valor más grande y el valor más pequeño de la serie de datos.
 - c) La diferencia entre los valores extremos y el valor central de la serie de datos.
 - d) La diferencia entre los valores centrales de la serie de datos.
2. El rango presenta fallas como medida de dispersión cuando:
 - a) Se tiene la presencia de medias desproporcionadas.
 - b) Se realiza un muestreo aleatorio.
 - c) Los datos emanan de una muestra y no de una población.
 - d) Se tiene la presencia de datos atípicos.
3. Es una de las ventajas de utilizar el rango:
 - a) Es una medida que señala hacia dónde se concentran los datos.
 - b) Es la medida de dispersión más fácil de calcular.
 - c) Es la medida de dispersión más exacta que existe en una serie.
 - d) Señala cómo se dispersan los datos de la media.
4. Si tenemos los siguientes datos 0, 1, 1, 3, y 5, entonces el rango es:
 - a) 5
 - b) 4
 - c) 2
 - d) 6
5. El departamento de crédito y cobranza de una empresa quiere conocer la variación que existe en una muestra de 15 datos, correspondientes a los próximos cobros (en pesos) que debe hacer. Calcula el rango para los datos siguientes

10 000	12 000	15 000	16 000	15 000
9 000	13 500	12 700	9 700	18 000
13 200	12 600	14 000	18 700	16 500

3.4.2. Varianza

Es una medida de variabilidad que toma en cuenta la dispersión que los valores de los datos tienen respecto a su media. Es decir, aquellos conjuntos de datos que tengan valores más alejados de la media, sea muestral o poblacional, tendrán una mayor varianza. Su resultado se expresa en unidades al cuadrado.

Existen dos símbolos para representar la varianza (σ^2 y S^2). La S^2 se refiere a un estadístico, es decir, a la varianza de una muestra; mientras que σ^2 se refiere a un parámetro, es decir, a la varianza de una población. A la S^2 se le conoce como la varianza muestral mientras que a σ^2 se le conoce como la varianza poblacional.

La manera de obtener la varianza de un conjunto de datos depende de la forma como se encuentren organizados los datos, ya sea que estén agrupados o no agrupados, así como del tipo de información con la que se trabaje, ya sea que provenga de una muestra o de una población.

a) La varianza para datos no agrupados

Cuando tenemos una variable cuya serie de datos no se encuentra agrupada, $X_1, X_2, X_3, \dots, X_n$, la *varianza poblacional* se calcula mediante la siguiente fórmula:

$$V(X) = \sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Donde:

$\sum (X_i - \mu)^2$ = Suma de los cuadrados de las desviaciones del valor de cada dato de la serie respecto a la media poblacional.

X_i = El valor de cada dato de la serie.

μ = La media poblacional.

N = Tamaño de la población.

Es decir, la varianza de una población para datos no agrupados es el promedio del cuadrado de las desviaciones respecto a su media μ .

Cuando tenemos una variable cuya serie de datos no se encuentra agrupada, $X_1, X_2, X_3, \dots, X_n$, la *varianza muestral* se calcula mediante la siguiente fórmula:

$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

Donde:

$\sum (X_i - \bar{X})^2$ = Suma de los cuadrados de las desviaciones del valor de cada dato de la serie respecto a la media muestral.

X_i = El valor de cada dato de la serie.

$\bar{X}$ = La media muestral.

N = Tamaño de la muestra.

A diferencia de lo que ocurre con otras fórmulas, la varianza de una muestra no equivale exactamente, en términos de cálculo, a la varianza de una población. El denominador de la fórmula de la varianza poblacional es el total de la población N , mientras que en la varianza muestral se incluye un factor de corrección $n - 1$.

Los pasos para obtener la varianza muestral o poblacional para datos no agrupados son los siguientes

1. Encuentra la media muestral o poblacional, según sea el caso.
2. Obtén cada una de las desviaciones respecto a la media, es decir, a cada uno de los datos $X_1, X_2, \dots, X_n$ se le resta la media obtenida en el paso anterior para quedar los siguientes valores

$$(X_1 - \mu), (X_2 - \mu), \dots, (X_n - \mu) \quad \text{en caso de una población.}$$

$$(X_1 - \bar{X}), (X_2 - \bar{X}), \dots, (X_n - \bar{X}) \quad \text{en caso de una muestra.}$$

3. Eleva al cuadrado cada una de las desviaciones obtenidas en el paso anterior y súmalas entre sí, para obtener la suma del cuadrado de las desviaciones

$$\sum (X - \mu)^2 = (X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_n - \mu)^2 \quad \text{en caso de una población.}$$

$$\sum (X - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \quad \text{en caso de una muestra.}$$

4. La suma del cuadrado de las desviaciones respecto a su media se divide entre N , en caso de una población; o entre $n - 1$, en caso de una muestra.

Tanto para una población como para una muestra, la fórmula de la varianza puede ser transformada en las siguientes expresiones, las cuales son conocidas como el método corto de la varianza:

$$\text{Varianza poblacional} \quad V(X) = \sigma^2 = \frac{\sum X_i^2}{N} - \mu^2$$

$$\text{Varianza muestral} \quad S^2 = \frac{\sum X_i^2 - n\bar{X}^2}{n-1}$$

Estas fórmulas tienen la ventaja de simplificar las operaciones que se deben realizar cuando se calcula la varianza, sea poblacional o muestral. Cabe señalar que las fórmulas establecidas por el método corto nos conducen al mismo resultado que si se hubieran empleado las fórmulas anteriores, siempre y cuando no se hayan omitido algunos dígitos en las distintas operaciones. La conveniencia de utilizar una u otra fórmula queda sujeta a la libre elección del lector, según la comodidad que le produzca cada una de ellas para realizar las operaciones.

Ejemplo 22

Emplea los datos de las ventas de seguros del ejemplo 20 y calcula la varianza, suponiendo que los datos constituyen la población total de los agentes de seguro de la compañía.

Se tiene que la media es:

$$\mu = \frac{\sum X}{N} = \frac{(8+11+5+14+11+8+11+16)}{8} = \frac{84}{8} = 10.5$$

Para calcular la varianza se requiere obtener cada una de las diferencias o desviaciones de los datos respecto a la media ($X - \mu$), elevarlas al cuadrado ($(X - \mu)^2$) y sumar estos resultados

X	(X - μ)	(X - μ) ²
8	-2.5	6.25
11	0.5	0.25
5	-5.5	30.25
14	3.5	12.25
11	0.5	0.25
8	-2.5	6.25
11	0.5	0.25
16	5.5	30.25
Σ	0	86

Tabla 3.20. Desviaciones de la venta de seguros

Ahora aplicamos la fórmula de varianza poblacional para datos no agrupados y obtenemos

$$V(X) = \sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N} = \frac{86}{8} = 10.75$$

Puede apreciarse que la varianza es de 10.75. Sin embargo, esta medida de variación no tiene un significado práctico debido a que el resultado obtenido está expresado en términos cuadrados, es decir, la variabilidad de seguros vendidos es de 10.75 seguros cuadrados.

Por esa razón, la varianza sólo tiene sentido lógico cuando comparamos diferentes conjuntos de datos con la misma unidad de medida, es decir, su interpretación es una medida relativa en el sentido de que aquel conjunto que tenga la mayor varianza será el de mayor grado de dispersión.

Por otra parte, si el lector hubiera optado por el método corto para estimar la varianza poblacional, el resultado hubiera sido el mismo. Para ello debemos estimar ΣX_i^2 y μ^2 :

$$\begin{aligned}\Sigma X_i^2 &= 8^2 + 11^2 + 5^2 + 14^2 + 11^2 + 8^2 + 11^2 + 16^2 \\ &= 64 + 121 + 25 + 196 + 121 + 64 + 121 + 256 = 968\end{aligned}$$

$$\mu^2 = 10.5^2 = 110.25$$

$$V(X) = \sigma^2 = \frac{\Sigma X_i^2}{N} - \mu^2 = \frac{968}{8} - 110.25 = 121 - 110.25 = 10.75$$

Si se compara este resultado mediante el método corto con el primer método, se puede apreciar que los resultados no fueron distintos.

Ejemplo 23

En las tablas 3.21 y 3.22 se exponen las cotizaciones mensuales del tipo de cambio entre el peso mexicano y el dólar estadounidense para los años de 1995 y 2000. Observa cuidadosamente la información contenida en cada tabla.

- Realizando una inspección visual, ¿en cuál de los dos años se observa una mayor estabilidad en el tipo de cambio?

- b) Encuentra la varianza para el tipo de cambio entre el peso y el dólar estadounidense en cada uno de los dos años

Mes	Tipo de cambio en 1995
Enero	5.69
Febrero	5.83
Marzo	6.81
Abril	5.78
Mayo	6.17
Junio	6.30
Julio	6.08
Agosto	6.31
Septiembre	6.41
Octubre	7.17
Noviembre	7.65
Diciembre	7.64

Fuente: Banco de México: www.banxico.org.mx

Tabla 3.21. Tipo de cambio mensual peso-dólar en el año 1995.

Mes	Tipo de cambio en el 2000
Enero	9.47
Febrero	9.44
Marzo	9.29
Abril	9.37
Mayo	9.50
Junio	9.79
Julio	9.46
Agosto	9.28
Septiembre	9.33
Octubre	9.51
Noviembre	9.51
Diciembre	9.44

Fuente: Banco de México: www.banxico.org.mx

Tabla 3.22. Tipo de cambio mensual peso-dólar en el año 2000

Se observa que los valores del tipo de cambio en el año de 1995 se encuentran muy dispersos entre sí, lo que indica una gran variabilidad o inestabilidad en el mercado cambiario. En contraste, en el año 2000 se puede observar que los valores de la divisa estadounidense se encuentran poco dispersos por lo que se esperaría que la varianza en este año sea menor a la de 1995.

Como los datos no se encuentran organizados mediante tablas de frecuencias, procedemos a encontrar la varianza muestral para datos no agrupados, obteniendo en primer lugar sus medias respectivas

$$\text{La media de 1995 es } \bar{X} = \frac{\sum X}{N} = \frac{(5.69 + 5.83 + 6.81 + \dots + 7.67)}{12} = \frac{77.84}{12} = 6.48$$

$$\text{La media de 2000 es } \bar{X} = \frac{\sum X}{N} = \frac{(9.47 + 9.44 + 9.29 + \dots + 9.44)}{12} = \frac{113.39}{12} = 9.44$$

Procedemos a encontrar la suma del cuadrado de las desviaciones del tipo de cambio respecto a la media, de acuerdo con las siguientes tablas

Mes	$(X - \bar{X})$	$(X - \bar{X})^2$
Enero	-0.79	0.6241
Febrero	-0.65	0.4225
Marzo	0.33	0.1089
Abril	-0.70	0.49
Mayo	-0.31	0.0961
Junio	-0.18	0.0324
Julio	-0.40	0.16
Agosto	-0.17	0.0289
Septiembre	-0.07	0.0049
Octubre	0.69	0.4761
Noviembre	1.17	1.3689
Diciembre	1.16	1.3456
Suma		5.1584

Tabla 3.23. Desviaciones del tipo de cambio en el año 1995.

Mes	$(X - \bar{X})$	$(X - \bar{X})^2$
Enero	0.03	0.0009
Febrero	0	0
Marzo	-0.15	0.0225
Abril	-0.07	0.0049
Mayo	0.06	0.0036
Junio	0.35	0.1225
Julio	0.02	0.0004
Agosto	-0.16	0.0256
Septiembre	-0.11	0.0121
Octubre	0.07	0.0049
Noviembre	0.07	0.0049
Diciembre	0	0
Suma		0.2023

Tabla 3.24. Desviaciones del tipo de cambio en el en el año 2000.

De los resultados obtenidos en las tablas 3.23. y 3.24., se divide la suma del cuadrado de las desviaciones entre $n - 1$ y así se obtiene la varianza muestral del tipo de cambio para los años de 1995 y 2000.

$$\text{Para el año de 1995} \quad S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{5.1584}{11} = 0.4689 \quad \text{pesos al cuadrado}$$

$$\text{Para el año 2000} \quad S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{0.2023}{11} = 0.0183 \quad \text{pesos al cuadrado}$$

Si bien los pesos al cuadrado continúan siendo una idea abstracta, ambas varianzas tienen sentido lógico cuando son comparadas entre sí, pues se encuentran expresadas en la misma unidad de medida. En este caso, el tipo de cambio en el año de 1995 tiene una mayor dispersión que el observado en el año 2000, tal como lo señalan ambas varianzas y tal como lo apreciamos de manera visual en el indiso anterior.

Este contraste se debe a la diferencia en los escenarios macroeconómicos que se vivieron durante esos años. Al ser mayor la varianza del año 1995, se refleja una gran volatilidad y nerviosismo en el mercado cambiario producido por una fuerte crisis económica que se vivía en ese año. En el año 2000 podemos observar que el peso mexicano gozó de una gran fortaleza, pues su cotización se mantuvo muy estable en el transcurso de los 12 meses, incluso en el mes de junio, cuando se presentaba la recta final de un proceso electoral en el país.

b) La varianza para datos agrupados

En el caso de datos agrupados, para encontrar la varianza es necesario conocer el punto medio de cada 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 un intervalo. El punto medio de la clase j se denota por m_j .

i) La *varianza poblacional* para datos agrupados se define como:

$$\sigma^2 = \frac{\sum [(m_j - \mu)^2 f_j]}{N}$$

Donde:

σ^2 = Varianza de la población.

m_j = Punto medio de clase.

μ = Media de la población.

N = Tamaño de la población.

f = Frecuencia de la clase.

ii) La fórmula para calcular la *varianza muestral* es:

$$S^2 = \frac{\sum [(m_j - \bar{X})^2 f_j]}{n-1}$$

Donde:

S^2 = Varianza de la muestra.

m_j = Punto medio de clase.

$\bar{X}$ = Media de la muestra.
 n = Tamaño de la muestra.
 f = Frecuencia de la clase.

Para obtener la varianza para datos agrupados, sea muestral o poblacional, se tienen que realizar los siguientes pasos:

1. Se obtiene la media muestral o poblacional para datos agrupados, según corresponda. Por ejemplo, si se pretende obtener la varianza muestral, entonces procedemos a encontrar la media a través de la siguiente fórmula:

$$\bar{X} = \frac{\sum m_j f}{\sum f}$$

2. Se encuentran los puntos medios para cada una de las clases $m_1, m_2, \dots, m_n$ y a cada uno se resta la media muestral o poblacional según corresponda. Por ejemplo, para el caso de la varianza muestral se encontrarían las siguientes desviaciones:

$$(m_1 - \bar{X}), (m_2 - \bar{X}), \dots, (m_n - \bar{X})$$

3. Se eleva al cuadrado cada una de las desviaciones de los puntos medios de clases respecto a la media. Por ejemplo, en caso de una población:

$$(X_1 - \mu)^2, (X_2 - \mu)^2, \dots, (X_n - \mu)^2$$

4. Cada uno de los cuadrados se multiplica por su respectiva frecuencia de clase. Por ejemplo, en el caso de una población:

$$(X_1 - \mu)^2 f_1, (X_2 - \mu)^2 f_2, \dots, (X_n - \mu)^2 f_n$$

5. Se suma cada uno de estos resultados y se divide, en el caso de la varianza poblacional, entre el número total de datos de la población (N), y en el caso de una muestra entre el $n-1$.

Ejemplo 24

Una gran empresa de ventas por teléfono quiere conocer la variación existente en las ventas realizadas (en miles de pesos) por sus operadores. Para esto realiza una muestra de 25 operadores telefónicos, obteniendo los resultados de la siguiente tabla. Calcula la varianza muestral.

Ventas (miles \$)	f
5.00 – 8.99	3
9.00 – 12.99	5
13.00 – 16.99	7
17.00 – 20.99	6
21.00 – 24.99	3
25.00 – 28.99	1
Σ	25

Tabla 3.25 Distribución de las ventas por teléfono.

Las clases denotan las ventas realizadas en miles de pesos y la frecuencia del número de operadores telefónicos.

Ventas (miles de \$)	F	m_j	$(m_j \cdot f)$	$(m_j - \bar{X})$	$(m_j - \bar{X})^2$	$[(m_j - \bar{X})^2] f$
5.00 – 8.99	3	6.995	20.985	-8.64	74.6496	223.9488
9.00 – 12.99	5	10.995	54.975	-4.64	21.5296	107.648
13.00 – 16.99	7	14.995	104.965	-0.64	0.4096	2.8672
17.00 – 20.99	6	18.995	113.97	3.36	11.2896	67.7376
21.00 – 24.99	3	22.995	68.985	7.36	54.1696	162.5088
25.00 – 28.99	1	26.995	26.995	11.36	129.0496	129.0496
Σ	25		390.875			693.76

Tabla 3.26 Distribución de las ventas por teléfono.

Para obtener la varianza, en primer lugar se debe calcular la media muestral para datos agrupados, encontrando el punto medio de clase, multiplicarlo por su frecuencia de la clase correspondiente, y sus resultados se suman para obtener la media, tal y como se muestra a continuación:

$$\bar{X} = \frac{\Sigma m_j \cdot f}{n} = \frac{390.875}{25} = 15.635$$

Se obtiene la varianza restandole a cada punto medio de clase la media muestral, elevando cada una de estas diferencias al cuadrado y multiplicando cada diferencia cuadrática por la frecuencia respectiva de clase de la manera siguiente:

$$s^2 = \frac{\Sigma [(m_j - \bar{X})^2 f_j]}{n - 1} = \frac{693.76}{(25 - 1)} = \frac{693.76}{24} = 28.90666667 \text{ pesos al cuadrado}$$

La varianza obtenida señala que la dispersión existente entre las ventas entre

$(n - 1)$ es de 28.90666667 miles de pesos al cuadrado.

Ventajas y desventajas de la varianza

La varianza mide la variabilidad tomando en cuenta la dispersión que los valores de los datos tienen respecto a su media. Es decir, aquellos conjuntos que tengan valores más alejados de la media, sea muestral o poblacional, tendrán una mayor varianza, mientras que aquellos conjuntos con valores más cercanos a la media mostrarán una mayor uniformidad al contar con una varianza menor.

La varianza únicamente adquiere valores mayores o iguales a cero, nunca valores negativos, y se utiliza para comparar la dispersión de dos o más conjuntos de datos que se encuentren expresados en la misma unidad de medida; por ejemplo, para observar la variación existente entre dos líneas de producción, la tasa de interés de dos instrumentos financieros, las ventas de productos expresados en la misma moneda, etcétera.

La principal desventaja de la varianza es que su resultado se expresa en unidades al cuadrado, resultando darle una interpretación lógica. Además, la varianza no puede comparar la dispersión de dos conjuntos de datos expresados en diferentes unidades de medida; por ejemplo, chamarras con coches, diferentes divisas, el IPC de la Bolsa Mexicana de Valores con el índice Dow Jones de la Bolsa de Nueva York, etcétera.

Ejercicio 7

- Grandes varianzas implican:
 - Que los datos no varían.
 - Que hay gran variación en los datos.
 - Que hay poca variación en los datos.
 - Que las medias son desproporcionadas.
- Uno de los inconvenientes de utilizar la varianza como medida de dispersión es que:
 - La varianza muestral es sesgada y la poblacional no.
 - La varianza se ve afectada por el tipo de dato que estamos utilizando.
 - Las varianzas poblacionales y muestrales son distintas.
 - Los resultados se expresan en unidades al cuadrado.
- Si tenemos cinco datos cuyos valores son las constantes 2, 2, 2, 2 y 2; entonces la varianza es
 - Cualquier valor.
 - Un valor mayor o igual a cero.
 - Un valor igual a cero.
 - Tanto valores positivos como negativos, excepto el cero.
- Una serie compuesta con los siguientes datos 0, 1, 1, 3 y 5, su varianza será:
 - 2
 - 4
 - 0
 - 1
- Con los siguientes datos de crédito y cobranza, calcula la varianza para datos no agrupados, con el fin de determinar la variabilidad de los datos de los próximos cobros (en pesos).

10 000	12 000	15 000	16 000	15 000
9 000	13 500	12 700	9 700	18 000
13 200	12 600	14 000	18 700	16 500

- Un despacho de consultoría en cuestiones de mercado hace una encuesta de los ingresos anuales (en miles de pesos) de 300 familias para clasificarlas por nivel de ingreso y con esto establecer qué artículos son susceptibles de promocionarse y posicionar en el mercado, considerando las variaciones existentes. Con la información siguiente calcula la varianza:

Ingreso (miles de \$)	f
1.50 – 2.999	25
3.00 – 4.999	31
5.00 – 6.999	42
7.00 – 8.999	45
9.00 – 10.999	52
11.00 – 12.999	42
13.00 – 14.999	35
15.00 – 16.999	28
Σ	300

Distribución de salarios

3.4.3. Desviación estándar

Al igual que la varianza, la desviación estándar es una medida de variabilidad que también toma en cuenta la dispersión de los valores de los datos respecto a su media. Sin embargo, su significado es más valioso que el de la varianza, pues su resultado se encuentra expresado en las mismas unidades de la variable que se examina y no en valores elevados al cuadrado como lo hace la varianza.

La desviación estándar se representa mediante la letra griega σ para el caso de una población, o por S en el caso de una muestra. Se obtiene sacando la raíz cuadrada al resultado de la varianza, no importa si ésta se trata de una varianza para datos no agrupados o para datos agrupados, o provenientes de una muestra o de una población. Al proporcionar sus resultados en unidades no cuadradas, la desviación estándar es muy fácil de interpretar y su resultado tiene mayor significado en el análisis de un fenómeno.

Las fórmulas para la *desviación estándar para datos no agrupados* son:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} \quad \text{o} \quad S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

Cuando se trabaja con *datos agrupados*, la desviación estándar también se calcula sacando la raíz cuadrada, pero empleando las fórmulas respectivas de la varianza para datos agrupados

$$\sigma = \sqrt{\frac{\sum [(m - \mu)^2 f_j]}{N}} \quad \text{o} \quad S = \sqrt{\frac{\sum [(m - \bar{X})^2 f_j]}{n-1}}$$

Tanto en datos no agrupados como en datos agrupados, σ indica la desviación estándar para una población, mientras que la S representa la desviación estándar para una muestra.

Ejemplo 25

Una casa de bolsa desea realizar un comparativo entre los rendimientos anuales y los riesgos de dos instrumentos financieros que han estado operando durante los últimos siete años. Sus rendimientos anuales, expresados en porcentajes, son los siguientes

Instrumento A:	4.0%	14.3%	19.5%	14.7%	26.5%	37.2%	23.8%
Instrumento B:	6.5%	4.4%	4.8%	6.9%	8.5%	5.8%	5.1%

Obtener la media y la desviación estándar de los rendimientos observados por los dos instrumentos financieros

En primer lugar se obtiene el rendimiento promedio por instrumento:

$$\mu_A = \frac{\sum X}{N} = \frac{(4 + 14.3 + 19.5 + 14.7 + 26.5 + 37.2 + 23.8)}{7} = \frac{140}{7} = 20\%$$

$$\mu_B = \frac{\sum X}{N} = \frac{(6.5 + 4.4 + 4.8 + 6.9 + 8.5 + 5.8 + 5.1)}{7} = \frac{42}{7} = 6\%$$

Como puede observarse, el instrumento que presenta el mayor rendimiento promedio es A con 20%, mientras que el instrumento B tiene un rendimiento promedio de 6%. En ese sentido, resultaría más atractivo invertir en el fondo A que en el fondo B.

Para medir el riesgo de cada uno de los fondos encontramos sus desviaciones estándar; para esto, primero se deben obtener las varianzas poblacionales y posteriormente se les saca la raíz cuadrada:

Acción A

$$\Sigma(X - \mu)_A^2 = (4 - 20)^2 + (14.3 - 20)^2 + \dots + (23.8 - 20)^2 = 669.36$$

$$V(X)_A = \sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{669.36}{7} = 95.62285714$$

$$\sigma_A = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{95.62285714} = 9.778694041$$

Acción B

$$\Sigma(X - \mu)_B^2 = (6.5 - 6)^2 + (4.4 - 6)^2 + \dots + (5.1 - 6)^2 = 12.16$$

$$V(X)_B = \sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{12.16}{7} = 1.737142857$$

$$\sigma_B = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{1.737142857} = 1.318007154$$

Puede observarse que el instrumento A tiene una variabilidad de 9.778694041%, mientras que el instrumento B tuvo una variabilidad de 1.318007154%. Esto indica que los rendimientos del instrumento A tienen una mayor dispersión que los rendimientos del instrumento B.

En el contexto de este ejemplo puede pensarse en la desviación estándar como una medida de la incertidumbre o riesgo de la rentabilidad de una inversión. Es decir, la rentabilidad promedio fue mayor para el instrumento A, pero su riesgo en términos de la desviación estándar de la rentabilidad también fue mayor.

Por otra parte, para obtener la desviación estándar cuando se trabaja con datos agrupados se utiliza la misma metodología que en el caso de los datos no agrupados. En primer lugar se encuentra la varianza a través de su respectiva fórmula y posteriormente se le saca la raíz cuadrada.

Ejemplo 26

Con los datos del ejemplo 5 calcula la desviación estándar.

$$S = \sqrt{\frac{\Sigma(m_j - \bar{X})^2 f_j}{n-1}} = \sqrt{\frac{693.76}{24}} = \sqrt{28.90666667} = 5.376492041$$

Con este resultado se deduce que la variación promedio que existe en las ventas realizadas por teléfono es de 5.38 miles de pesos. Esto puede ayudar a la empresa a analizar las ventas que realizan los operadores de una manera más sencilla que utilizando ventas al cuadrado.

Ventajas y desventajas de la desviación estándar

La principal ventaja de la desviación estándar es que indica la manera en que se dispersan los datos respecto a la media en las mismas unidades de la variable que se examina y no en valores elevados al cuadrado. Al igual que la varianza, la desviación estándar únicamente adquiere valores mayores o iguales a cero, nunca valores negativos.

Es utilizada para comparar la dispersión entre distintos conjuntos de datos. Aquellos conjuntos que tengan valores más alejados de la media tendrán una mayor desviación estándar, mientras que aquellos conjuntos con valores más cercanos a la media mostrarán una menor desviación estándar.

Al igual que la varianza, una desventaja de la desviación estándar es que tampoco puede comparar la dispersión de dos conjuntos de datos que se expresan en diferentes unidades de medida.

Ejercicio 8

1. Con los datos de crédito y cobranza que se presentan a continuación, calcula la desviación estándar de los próximos cobros.

10 000	12 000	15 000	16 000	15 000
9 000	13 500	12 700	9 700	18 000
13 200	12 600	14 000	18 700	16 500

2. Con los siguientes datos de los ingresos anuales (en miles de pesos) de 300 familias, calcula la desviación estándar.

Ingreso (miles de \$)	<i>f</i>
1.50 – 2.999	25
3.00 – 4.999	31
5.00 – 6.999	42
7.00 – 8.999	45
9.00 – 10.999	52
11.00 – 12.999	42
13.00 – 14.999	35
15.00 – 16.999	28
Σ	300

Distribución de salarios

3.5. Interpretación de la desviación estándar y su aplicación en los negocios

La desviación estándar es utilizada como unidad de medida para conocer a qué distancia se encuentra alejado el valor de un dato respecto a la media, es decir, a cuántas veces la desviación estándar se ubica del valor X de la media. Para esto requerimos de la siguiente fórmula:

$$Z = \frac{X_i - \bar{X}}{S} \quad \text{En el caso de una muestra}$$

$$Z = \frac{X_i - \mu}{\sigma} \quad \text{En el caso de una población}$$

Esta fórmula es conocida con el nombre de estandarización, donde Z indica a qué distancia se encuentra un valor alejado de la media en términos de la desviación estándar como unidad de medida. Si el resultado de Z es negativo, se dice que el valor X es inferior al valor de la media y se encuentra Z veces la desviación estándar por debajo de la media. Si el resultado de Z es positivo, se dice que el valor X es superior al valor de la media y se encuentra Z veces la desviación estándar por encima de la media.

Ejemplo 27

Si tenemos una muestra cuya media es $\bar{X} = 50.5$ y su desviación estándar es $S = 10$, y se desea conocer a qué distancia de la media se encuentra un dato en específico de la muestra, por ejemplo $X = 18.5$, aplicamos la siguiente fórmula:

$$Z = \frac{X_i - \bar{X}}{S} = \frac{18.5 - 50.5}{10} = -3.2$$

Lo anterior quiere decir que el número 18.5 se encuentra a 3.2 veces la desviación estándar por debajo de la media 50.5. Observa que el 18.5 es inferior al valor 50.5 de la media.

Ejemplo 28

Si tenemos una población cuya media es $\mu = 300$ y su desviación estándar es $\sigma = 100$, y se desea conocer a qué distancia de la media se encuentra un dato en específico de la población, por ejemplo $X = 450$, aplicamos la siguiente fórmula:

$$Z = \frac{X_i - \mu}{\sigma} = \frac{450 - 300}{100} = 1.5$$

Es decir, el número 450 se encuentra a 1.5 veces la desviación estándar por encima de la media 300. Observa que el 450 es superior al valor 300 de la media.

Teorema de Tchebysheff

El teorema de Tchebysheff señala cuál es el porcentaje mínimo de datos que se acumulan alrededor de la media dentro de una distancia equivalente a Z veces la desviación estándar de la media. Para esto se utiliza la siguiente fórmula:

$$\text{Porcentaje m\u00ednimo de los datos de un conjunto} = \left(1 - \frac{1}{Z^2}\right) 100\%$$

Donde:

Z = N\u00famero de desviaciones est\u00e1ndar

Por lo tanto, el teorema de Tchebysheff se\u00f1ala que para cualquier tipo de distribuci\u00f3n de datos se cumple lo siguiente:

1. Al considerar una distancia de dos desviaciones est\u00e1ndar ($Z=2$), al menos 75% de los datos $\left(1 - \frac{1}{(2^2)}\right) 100\% = 75\%$ debe estar contenido dentro del rango que se encuentra a 2 desviaciones est\u00e1ndar por encima de la media ($\mu + 2\sigma$) y a 2 desviaciones est\u00e1ndar por debajo de la media ($\mu - 2\sigma$).
2. Al tomarse en cuenta una distancia de tres desviaciones est\u00e1ndar ($Z=3$), al menos 88.89% de los datos $\left(1 - \frac{1}{(3^2)}\right) 100\% = 88.89\%$ debe estar contenido dentro del rango que se encuentra a 3 desviaciones est\u00e1ndar por encima de la media ($\mu + 3\sigma$) y a 3 desviaciones est\u00e1ndar por debajo de la media ($\mu - 3\sigma$).
3. Si la distancia es de cuatro desviaciones est\u00e1ndar ($Z=4$), al menos 93.75% de los datos $\left(1 - \frac{1}{(4^2)}\right) 100\% = 93.75\%$ debe estar incluido dentro del rango que se encuentra a 4 desviaciones est\u00e1ndar por encima de la media ($\mu + 4\sigma$) y a 4 desviaciones est\u00e1ndar por debajo de la media ($\mu - 4\sigma$).

Por ejemplo, si tenemos un conjunto de datos cuya distribuci\u00f3n de frecuencias se representa mediante la siguiente figura, entonces la relaci\u00f3n que existe entre el m\u00ednimo de datos acumulados alrededor de la media y la desviaci\u00f3n est\u00e1ndar es:

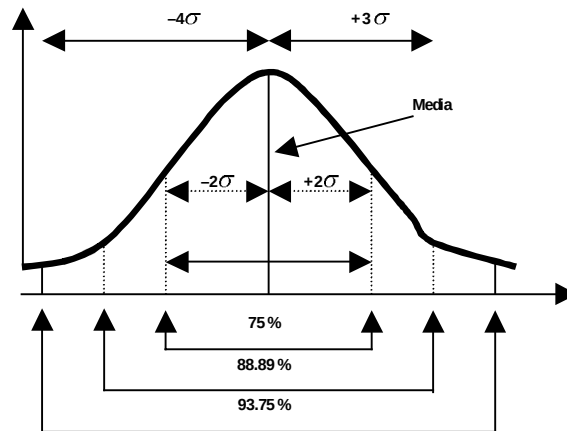


Figura 3.8. Dispersi\u00f3n de datos.

La figura anterior indica que al medir la dispersi\u00f3n de los datos, si \u00e9stos se alejan de la media de la distribuci\u00f3n a una distancia de 2 desviaciones est\u00e1ndar, en el intervalo comprendido por los extremos derecho ($+2\sigma$) e izquierdo (-2σ) estar\u00e1n agrupados al menos 75% de los datos, concentr\u00e1ndose 37.5% ($75\% / 2$) a la derecha de la media y el restante 37.5% a la izquierda de la media. Cabe destacar que mientras m\u00e1s se alejan los datos de la media, mayor ser\u00e1 la desviaci\u00f3n est\u00e1ndar y, por lo tanto, mayor ser\u00e1 la dispersi\u00f3n.

Ejemplo 29

Con los datos del ejemplo 22, calcula el porcentaje mínimo de los datos que se encuentran dentro del rango de 2.5 desviaciones estándar por encima y por debajo de la media, así como los valores que delimitan este rango.

La media de los datos es 10.5, la varianza 10.75 y la desviación estándar 3.278719262. A una distancia de 2.5 desviaciones estándar ($Z = 2.5$):

$$\left(1 - \frac{1}{(2.5)^2}\right) 100\% = \left(1 - \frac{1}{6.25}\right) 100\% = [1 - (0.16)] 100 = (0.84)(100) = 84\%$$

El resultado implica que al menos 84% de las ventas debe estar a una distancia de 2.5 veces la desviación estándar por encima y por debajo de la media.

Para calcular los valores exactos que delimitan el rango de 2.5 veces la desviación estándar, por encima y por debajo de la media, se realizan las siguientes operaciones:

$$\text{Valor inferior: } \mu - Z\sigma = 10.5 - (2.5)(3.278719262) = 2.3032$$

$$\text{Valor superior: } \mu + Z\sigma = 10.5 + (2.5)(3.278719262) = 18.6967$$

Es decir, dentro del intervalo denotados por los valores 2.3032 y 18.6967 se encontrarán concentrados como mínimo 84% de los datos alrededor de la media, $\mu = 10.5$. Si verificamos en el ejemplo 22, los datos de la serie 8, 11, 5, 14, 11, 8, 11 y 16 observamos que en este ejemplo se cumple con facilidad el teorema de Tchebysheff, pues dentro del intervalo arriba expuesto se encuentran depositados todos los datos de la serie (100% de los datos), superando así el porcentaje mínimo señalado de 84% del total de los datos. Estos resultados se pueden apreciar en la figura siguiente.

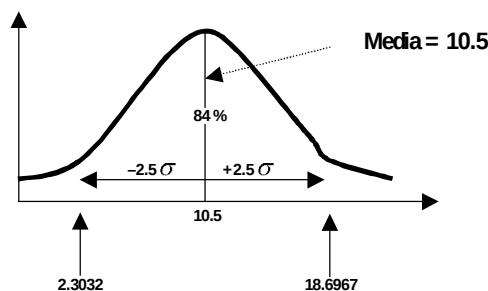


Figura 3.9. Dispersión de datos.

Ejemplo 30

En la tabla siguiente se expone la participación mensual de la inversión extranjera en el mercado accionario de la Bolsa Mexicana de Valores, durante el año 2000.

- Calcula el porcentaje mínimo de datos que se encuentra dentro del rango de 1.5 veces la desviación estándar por arriba y por debajo de la media.
- Encuentra los valores superior e inferior que determinan este rango.

Mes	2000
Enero	44.01
Febrero	46.58
Marzo	44.78
Abril	47.25
Mayo	45.07
Junio	46.69
Julio	44.07
Agosto	44.96
Septiembre	44.72
Octubre	44.62
Noviembre	43.03
Diciembre	41.31

Fuente: Bolsa Mexicana de Valores. www.bmv.com.mx

Tabla 3.27. Participación mensual de la inversión extranjera en la Bolsa Mexicana de Valores

La media de los datos es 44.7575 y la desviación estándar 1.6328. A una distancia de 1.5 desviaciones estándar de la media ($Z = 1.5$):

$$\left(1 - \frac{1}{(1.5)^2}\right) 100\% = \left(1 - \frac{1}{2.25}\right) 100\% = [1 - (0.4444)] 100 = (0.5556) (100) = 55.56\%$$

Al menos 55.56% de los datos debe estar a una distancia de 1.5 veces la desviación estándar por encima y por debajo de la media. Los valores exactos que delimitan el rango de 1.5 veces la desviación estándar, por encima y por debajo de la media, son los siguientes:

$$\text{Valor inferior: } \mu - Z\sigma = 44.7575 - (1.5)(1.6328) = 42.3083$$

$$\text{Valor superior: } \mu + Z\sigma = 44.7575 + (1.5)(1.6328) = 47.2067$$

Como mínimo, 55.56% de los 12 datos registrados en la tabla anterior debe estar a una distancia de 1.5 desviaciones estándar respecto a su media (44.7575), es decir, entre los valores 42.3083 y 47.2067. Para constatar que se cumple el teorema de Tchebysheff se expone esta información en el siguiente diagrama:

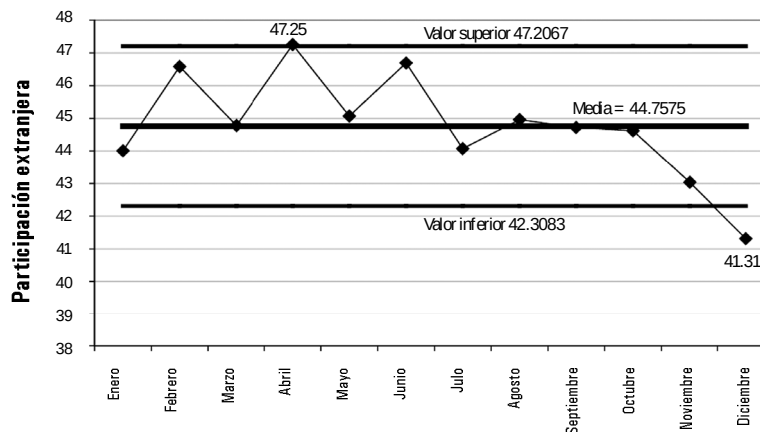


Figura 3.10 Participación de la inversión extranjera en la Bolsa Mexicana de Valores.

En la figura anterior se observa que únicamente dos valores, 47.25 y 41.31, quedaron excluidos del rango determinado por los valores 47.2067 y 42.3083. De esta manera observamos que 10 datos de los 12 que tiene la serie están incluidos dentro del rango señalado por los valores $\mu - 1.5\sigma$ y $\mu + 1.5\sigma$, lo que representa un porcentaje de 83.33% $[(10/12)*100]$, cumpliendo con facilidad el mínimo requerido por el teorema de Tchebysheff, de 55.55%, para una desviación estándar de 1.5.

La regla empírica

Un caso particular de los conceptos señalados por el teorema de Tchebysheff es cuando tenemos un conjunto de datos cuya distribución tiene la figura acampanada y simétrica. En este caso, la relación que existe entre el porcentaje de datos que se encuentran contenidos dentro de un intervalo y la desviación estándar respecto a la media que determina este intervalo es la siguiente:

1. Aproximadamente más de dos terceras partes centrales del conjunto de datos (68%) están comprendidas entre dos valores que se encuentran a una distancia de la media equivalente a la desviación estándar, tanto por la parte superior como por la parte inferior de la media.
2. Aproximadamente 95% de los datos centrales de un conjunto de datos están contenidos a una distancia de la media equivalente dos veces la desviación estándar, tanto por la parte superior como por la parte inferior de la media.
3. Aproximadamente 99% de los datos centrales de un conjunto de datos están contenidos a una distancia de la media equivalente a tres veces la desviación estándar, tanto por la parte superior como por la parte inferior de la media.

A este conjunto de relaciones se le conoce como la regla empírica. Por ejemplo, en el caso de una distribución de frecuencias simétrica podemos observar el siguiente gráfico:

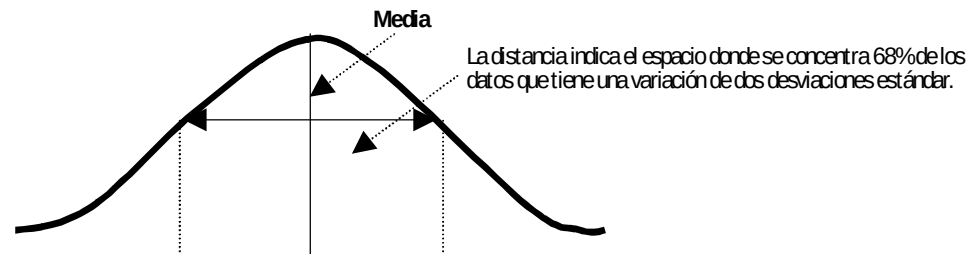


Figura 3.11. Agrupamiento de datos

Por esto, la desviación estándar como medida de dispersión promedio alrededor de la media ayuda a comprender cómo se distribuyen los datos por encima (o la derecha) y por debajo (o la izquierda) del valor de la media.

3.5.1. Interpretación de la desviación estándar

Ahor bien, la aplicación e interpretación en los negocios tiene muchos ejemplos y para ello podemos ejemplificar el caso de McDonald's Corporation.

La historia de McDonald's se inició en 1948, cuando los hermanos Richard y Maurice McDonald abrieron en San Bernardino (California) su primer restaurante McDonald's, establecimiento en el que se hacían los pedidos sin tener que bajarse del coche. Por aquella época, Ray Kroc, un pequeño empresario de máquinas de batidos, consiguió la cesión del derecho de la marca convencido de sus posibilidades de expansión. Un menú limitado y un alto volumen de ventas caracterizaron el éxito del nuevo restaurante. En 1954, Ray Kroc, por entonces proveedor de la máquina de batidos, sorprendido por la magnitud del pedido de equipos "multi-mixers" solicitado, visitó el local de los hermanos McDonald y les propuso abrir más locales. Un año más tarde, los hermanos McDonald le otorgaron a Kroc los derechos exclusivos para la comercialización y explotación del negocio de McDonald's.

De esta manera abrió su primer restaurante en Des Plaines, Illinois (1955), estableciendo un nuevo concepto de restaurante basado en ofrecer al cliente los más altos estándares de calidad, servicio y limpieza, los valores básicos sobre los que se ha constituido la compañía, al tiempo que ponía en marcha el sistema de franquicia. Posteriormente, McDonald's añadió a estas tres premisas de funcionamiento un cuarto principio: valor, es decir, la mejor relación calidad-precio.

A partir de ese momento, Ray Kroc hizo de los McDonald's la mayor organización de servicio rápido del mundo, basando su éxito en la filosofía operativa del *sistema McDonald's "calidad, servicio, limpieza y valor"*.

McDonald's es la compañía líder en el sector de restauración de servicio rápido en el mundo con más de 30,000 restaurantes en 119 países.

La década de los sesenta representa el periodo de expansión de McDonald's por Estados Unidos, una etapa que culminó con tres hitos destacados. A mediados de 1967, McDonald's Corporation abrió su primer restaurante en Canadá, iniciándose así el periodo de expansión de McDonald's por todo el mundo. Un año después, Jim Deligatti, franquiciado de Pittsburg, creaba la hamburguesa que se acabaría convirtiendo en el producto estrella de la compañía: el Big Mac. Y en 1969 se fundaba la Universidad de la Hamburguesa, en Illinois, uno de los centros de formación corporativa más avanzados del mundo, por el que pasan anualmente más de 3,000 estudiantes (entre directivos, franquiciados y empleados).

En 1967 la cadena abrió su primera sucursal fuera de los Estados Unidos y en 1990 se inauguró el primer McDonald's en Moscú, todo un símbolo de los nuevos tiempos.

McDonald's es una empresa que cambió la forma de hacer los negocios en el mundo. Es una marca basada en la filosofía de Ray Kroc, quien impulsó el negocio y tuvo la visión de construir una gran familia de hombres y mujeres que trabajan con todo el mundo para servir al cliente, ofreciéndole una comida de la mejor calidad en forma rápida, en un ambiente limpio y seguro y con una atención amistosa y amable.

McDonald's es la mayor red de locales de servicio rápido de comidas del mundo con un importante potencial de crecimiento: 45 millones de personas de la población mundial comen por día en uno de los 27 000 locales McDonald's distribuidos en los 120 países de los 5 continentes.

El 29 de octubre de 1985 McDonald's abrió su primer restaurante en México, en el sur de la Ciudad de México.

La imagen corporativa de McDonald's son los arcos dorados, los cuales son un símbolo que da la bienvenida a la persona no importando su edad o estilo de vida.

En México actualmente existen restaurantes McDonald's en 48 ciudades de la República Mexicana, donde se atiende cada mes a más de 4 millones de mexicanos.

Los principales proveedores de McDonald's en México son Coca Cola, Bimbo y Grupo Lala, su proveedor de carne 100% de ganado bovino es Trosi.

Asimismo, McDonald's está afiliado con organizaciones como Disney Co., Mattel, Copa Mundial de Fútbol Soccer (FIFA), NFL, Juegos Olímpicos y NBA.

Ahora bien, se realizó lo siguiente y se pide cierta información e interpretación.

Se tomó una muestra de ventas anuales en miles de pesos en 100 sucursales diferentes de McDonald's en el país, complementa el siguiente cuadro:

Resultados de las ventas una muestra de 100 sucursales de McDonald's en el país

Clase	Punto medio m_i	Frecuencia f_i	$f_i m_i$	Media $\bar{X}$	$m_i - \bar{X}$	$(m_i - \bar{X})^2$	$f_i (m_i - \bar{X})^2$
500 – 599	550	4					
600 – 699	650	7					
700 – 799	750	8					
800 – 899	850	10					
900 – 999	950	12					
1 000 – 1 099	1 050	17					
1 100 – 1 199	1 150	13					
1 200 – 1 299	1 250	9					
1 300 – 1 399	1 350	8					
1 400 – 1 499	1 450	7					
1 500 – 1 599	1 550	3					
1 600 – 1 699	1 650	2					

1. De la siguiente muestra de ventas en 100 locales de McDonald's en el país, determina la media.
2. De la muestra de ventas en 100 sucursales o franquicias de McDonald's en el país, determina la varianza.
3. De la siguiente muestra de ventas en 100 establecimientos o franquicias de McDonald's en el país, determina la desviación estándar.

Proceso de elaboración:
Caso McDonald's

Clase	Punto medio m_i	Frecuencia f_i	$f_i m_i$	Media $\bar{X}$	$m_i - \bar{X}$	$(m_i - \bar{X})^2$	$f_i (m_i - \bar{X})^2$
500–599	550	4	2 200	1 056	–506	256 036	1 024 144
600–699	650	7	4 550	1 056	–406	164 836	1 153 852
700–799	750	8	6 000	1 056	–306	93 636	729 088
800–899	850	10	8 500	1 056	–206	42 436	424 360
900–999	950	12	11 400	1 056	–106	11 236	134 832
1 000–1 099	1 050	17	17 850	1 056	–6	36	612
1 100–1 199	1 150	13	14 950	1 056	94	8 836	114 868
1 200–1 299	1 250	9	11 250	1 056	194	37 636	338 724
1 300–1 399	1 350	8	10 800	1 056	294	86 436	691 488
1 400–1 499	1 450	7	10 150	1 056	394	155 236	1 086 652
1 500–1 599	1 550	3	4 650	1 056	494	244 036	732 108
1 600–1 699	1 650	2	3 300	1 056	594	352 836	705 672
	1 100	100	105 600				7 156 400

$$1. \text{ Media } \bar{X} = \frac{\sum f_i M_i}{n} = \frac{105\,600}{100} = 1\,056$$

$$2. \text{ Varianza } S^2 = \frac{\sum f_i (M_i - \bar{X})^2}{n-1} = \frac{7\,156\,400}{99} = 72\,286.9$$

Este resultado nos dice que tenemos una dispersión de 72,286.9 [\$/]² respecto de la media; esta medida de dispersión por sí misma no proporciona información relevante; si nosotros comparásemos este resultado con la varianza resultante de las ventas anuales de 100 tiendas de McDonald's de otro país, podríamos comparar ambos resultados y la que tenga la mayor varianza será la que tenga un mayor grado de dispersión y por tanto podríamos decir que las ventas anuales son menos estables que en el país que tiene menor grado de dispersión.

$$3. \text{ Desviación estándar } S = \sqrt{S^2} = \sqrt{72\,286.9} = 268.7$$

Se obtiene una desviación estándar de 268 pesos; con este dato podemos determinar la distancia, en unidades de desviación estándar, a la que se encuentran distanciadas las ventas anuales de cada tienda de McDonald's respecto de la media. Por ejemplo, una sucursal que tiene ventas anuales de 520 mil pesos, ¿a cuántas desviaciones estándar estará alejada de la media de las ventas anuales de todas las tiendas?

Ejercicio 9

1. Si tenemos una distribución de datos cuya media es $\bar{X} = 20$ y su desviación estándar es $S = 6$, encuentra, en términos de la desviación estándar (Z), a qué distancia está de la media el valor $X = 30$
 - a) 5 desviaciones estándar.
 - b) 3.33 desviaciones estándar.
 - c) 1.66 desviaciones estándar.
 - d) 0.3 desviaciones estándar.
2. En una distribución de datos, al menos 75% de los datos está contenido dentro de:
 - a) 2 desviaciones estándar por encima y por debajo de la media.
 - b) 1 desviación estándar por encima y por debajo de la media.
 - c) 3 desviaciones estándar por encima y por debajo de la media.
 - d) 4 desviaciones estándar por encima y por debajo de la media.
3. En cualquier distribución, al menos 93.75 % de los datos se encuentran contenidos dentro del rango que se encuentra a:
 - a) 2 desviaciones estándar por encima y por debajo de la media.
 - b) 1 desviación estándar por encima y por debajo de la media.
 - c) 3 desviaciones estándar por encima y por debajo de la media.
 - d) 4 desviaciones estándar por encima y por debajo de la media.
4. Señala el porcentaje mínimo de los datos centrales que deben estar contenidos a una distancia de 3 veces la desviación estándar por encima y por debajo de la media ($k = 3$):
 - a) 32.5 % de los datos.
 - b) 50 % de los datos.
 - c) 75 % de los datos.
 - d) 88.89 % de los datos.
5. La regla empírica señala que aproximadamente 68% de los datos se encuentra entre los valores que encuentran a:
 - a) 1 desviación estándar por encima y por debajo de la media.
 - b) 2 desviaciones estándar por encima y por debajo de la media.
 - c) 3 desviaciones estándar por encima y por debajo de la media.
 - d) 4 desviaciones estándar por encima y por debajo de la media.
6. La regla empírica se cumple para distribuciones:
 - a) Asimétricas y acampanadas.
 - b) Acampanadas y simétricas.
 - c) De cualquier tipo.
 - d) Simétricas y no acampanadas.

7. La regla empírica señala que los valores que se encuentran a 3 veces la desviación estándar por encima y por debajo de la media, aproximadamente, se observa el:
 - a) 68% de los datos
 - b) 75% de los datos
 - c) 90% de los datos
 - d) 99% de los datos
8. Si en una distribución cuya media es 16.5 y desviación estándar de 4.3, calcula el porcentaje mínimo que se encuentra dentro del rango de tres veces la desviación estándar por encima y por debajo de la media, así como sus respectivos valores que delimitan este rango.
9. Si en una distribución cuya media es 2000 y desviación estándar de 300, calcula el porcentaje mínimo que se encuentra dentro del rango de 1.2 veces la desviación estándar por encima y por debajo de la media, así como sus respectivos valores que delimitan este rango.
10. Si en una distribución cuya media es 95 y desviación estándar de 25, calcula el porcentaje mínimo que se encuentra dentro del rango de 2.4 veces la desviación estándar por encima y por debajo de la media, así como sus respectivos valores que delimitan este rango.

3.6. Coeficiente de variación

Es una medida de dispersión que señala qué tan grande es la magnitud de la desviación estándar respecto a la media del conjunto de datos que se examina. A diferencia de otras medidas de variabilidad, el coeficiente de variación mide la dispersión en términos de porcentaje y no en unidades de medida. De esta manera, este coeficiente se utiliza para comparar la dispersión entre dos conjuntos de datos expresados en diferentes unidades de medidas.

Por ejemplo, si los analistas de un despacho de bienes raíces están interesados en determinar si el valor de un avalúo tiene mayor variabilidad que el tamaño del lote, resultaría imposible comparar en forma directa la dispersión mediante el rango, la varianza o la desviación estándar, pues el valor del avalúo se mide en unidades monetarias, por ejemplo en miles de pesos, mientras que el tamaño del lote se mide en metros cuadrados. En este caso, los analistas pueden utilizar el coeficiente de variación, expresado en porcentajes, y así comparar la dispersión de dos variables expresadas en distintas unidades de medida.

El coeficiente de variación se representa mediante la expresión CV y se obtiene dividiendo la desviación estándar entre la media, multiplicando este resultado por 100, no importando si se trata de datos no agrupados o de datos agrupados, o que provengan de una muestra o de una población.

El coeficiente de variación se puede calcular mediante la fórmula siguiente:

$$CV = \left(\frac{S}{\bar{X}} \right) 100\% \quad \text{En caso de una muestra}$$

$$CV = \left(\frac{\sigma}{\mu} \right) 100\% \quad \text{En caso de una población}$$

Donde:

CV = Coeficiente de variación.

S = Desviación estándar de la muestra.

$\bar{X}$ = Media de los datos.

σ = Desviación estándar de la población.

μ = Media poblacional.

Ejemplo 31

Con los datos del ejemplo 25, calcula el coeficiente de variación con el fin de hacer una comparación de los rendimientos de las acciones.

$$CV_A = \left(\frac{S_A}{\mu_A} \right) 100\% = \left(\frac{9.778694041}{20} \right) 100 = (0.488934702) (100) = 48.8934702\%$$

$$CV_B = \left(\frac{S_B}{\mu_B} \right) 100\% = \left(\frac{1.318007154}{6} \right) 100 = (0.219667859) (100) = 21.9667859\%$$

La acción que presenta la menor variabilidad es la B, que como ya se había mencionado es la que presenta un menor rendimiento promedio y una menor desviación estándar (menor riesgo), con lo que se concluye que la acción más conveniente para invertir sin incurrir en un gran riesgo es la acción B.

Ejemplo 32

Los analistas de un centro financiero desean comparar el desempeño del tipo de cambio y el porcentaje de la participación extranjera en el mercado accionario de la Bolsa Mexicana de Valores durante el año 2000. Para esto se calcula el coeficiente de variación para cada uno de los mercados.

Mes	Tipo de cambio en el 2000
Enero	9.47
Febrero	9.44
Marzo	9.29
Abril	9.37
Mayo	9.50
Junio	9.79
Julio	9.46
Agosto	9.28
Septiembre	9.33
Octubre	9.51
Noviembre	9.51
Diciembre	9.44

Fuente: Banco de México: www.banxico.org.mx

Tabla 3.28. Tipo de cambio mensual peso-dólar en el año 2000.

Mes	Inversión extranjera en el 2000
Enero	44.01
Febrero	46.58
Marzo	44.78
Abril	47.25
Mayo	45.07
Junio	46.69
Julio	44.07
Agosto	44.96
Septiembre	44.72
Octubre	44.62
Noviembre	43.03
Diciembre	41.31

Fuente: Banco de México: www.banxico.org.mx

Tabla 3.29. Participación extranjera en la bolsa en el año 2000.

Las variables que se desean comparar vienen expresadas en diferentes unidades de medida; el tipo de cambio se expresa en pesos mientras que la inversión extranjera se representa en proporciones. Por tal razón, se calculan los coeficientes de variación para cada una de las variables y así se compara la variabilidad de ambos mercados. Para ello tomamos las medias y las desviaciones estándar de los ejemplos 4 y 11.

$$CV_{\text{Tipodecambio}} = \left(\frac{S}{\bar{X}} \right) 100\% = \left(\frac{0.1352}{9.44} \right) 100 = (0.0143)(100) = 1.43\%$$

$$CV_{\text{Inv. extranjera}} = \left(\frac{S}{\bar{X}} \right) 100\% = \left(\frac{1.6328}{44.7575} \right) 100 = (0.0364)(100) = 3.64\%$$

Los analistas de este centro financiero pueden concluir que el mercado cambiario durante el año 2000 tuvo mayor estabilidad que la participación extranjera en el mercado accionario, pues el coeficiente de variación del primero fue de 1.43%, mientras que el del segundo fue de 3.64%. De esta forma, los analistas comparan la variación de dos mercados que tienen distintas unidades de medición.

Ventajas y desventajas del coeficiente de variación

El coeficiente de variación es útil cuando pretende comparar la variabilidad de dos o más conjuntos de datos expresados en diferentes unidades de medición, pues el resultado será señalado en porcentajes.

La única desventaja que adolece el coeficiente de variación es cuando se tienen que comparar dos conjuntos de datos donde uno tiene una media con valores negativos y el otro tiene una media positiva. Para el primer conjunto, el coeficiente de variación será negativo; mientras que para el segundo, el coeficiente de variación será positivo, haciendo difícil la comparación entre ambos. Esto puede solucionarse tomando los valores absolutos del resultado que se obtenga en ambos coeficientes.

Ejercicio 10

1. El coeficiente de variación es una medida de dispersión que expresa sus resultados como:
 - a) Unidades métricas.
 - b) Desviaciones estándar.
 - c) Porcentajes.
 - d) Desviaciones respecto a la media.
2. El coeficiente de variación tiene la ventaja de:
 - a) Comparar conjuntos de datos expresados en diferentes unidades de medición.
 - b) Comparar conjuntos de datos expresados en diferentes unidades cuadradas.
 - c) Comparar conjuntos de datos expresados en desviaciones.
 - d) Comparar conjuntos de datos expresados en porcentajes.
3. Si tenemos tres diferentes acciones A, B y C, y el coeficiente de variación de sus precios son $CV_A=13\%$, $CV_B=15\%$ y $CV_C=7\%$, entonces
 - a) La acción A es la de mayor variabilidad y la acción B es la de menor variabilidad.
 - b) La acción B es la de mayor variabilidad y la acción A es la de menor variabilidad.
 - c) La acción C es la de mayor variabilidad y la acción A es la de menor variabilidad.
 - d) La acción B es la de mayor variabilidad y la acción C es la de menor variabilidad.
4. Una casa de cambio desea conocer la variación existente entre el valor de dos monedas (pesos/ dólar y pesos/ libra) en las transacciones de 10 días para determinar qué moneda es la que representa una mayor estabilidad. Con los siguientes datos, calcula el coeficiente de variación.

Dólar	150	125	120	200	250	175	200	250	180	140
Libra	200	275	180	195	280	250	240	200	300	290

3.7. Índice de asimetría y kurtosis

Cuando se estudiaron las medidas de tendencia central se analizó la relación que existe entre la media, la mediana y la moda, señalando que esta relación depende de la forma en que se distribuyen los datos. Se dijo que el posicionamiento de las medidas de tendencia central estaba en función del tipo de sesgo que se observaba en la distribución de frecuencias.

Por ejemplo, cuando se tiene sesgo positivo o derecho, la media es mayor que la mediana y que la moda, es decir, la media se encuentra más a la derecha de la moda, dando así una distribución con una cola alargada que se extiende hacia el lado derecho. En el caso contrario, cuando la media es menor que la moda, la distribución tiene un sesgo negativo o izquierdo, dando así una distribución con una cola alargada que se extiende hacia el lado izquierdo. También se señaló que cuando las tres medidas de tendencia central coinciden con el mismo valor, la distribución de frecuencias era acampanada y simétrica.

En esta sección se analizará de manera más formal la asimetría de una distribución de frecuencias a través del índice de asimetría. Este aspecto es sumamente importante en el análisis de datos, pues dependiendo del tipo y de la magnitud del sesgo que se observe en una distribución de frecuencias se conocerá con más detalle la forma en que se dispersan los datos de una serie, detectando con mayor facilidad la presencia de datos atípicos.

El índice de asimetría es una medida de dispersión mediante la cual se conoce el tipo y la magnitud de sesgo en una distribución de frecuencias. Se representa mediante la expresión α_3 .

Para el caso de datos no agrupados, las fórmulas del índice de asimetría son:

$$\alpha_{3 \text{ Poblacional}} = \frac{\left(\frac{\sum (X_j - \mu)^3}{N} \right)}{(\sigma)^3} \qquad \alpha_{3 \text{ Muestral}} = \frac{\left(\frac{\sum (X_j - \bar{X})^3}{n-1} \right)}{(S)^3}$$

Para el caso de datos agrupados, las fórmulas del índice de asimetría son:

$$\alpha_{3 \text{ Poblacional}} = \frac{\left(\frac{\sum [(m_j - \mu)^3] f}{N} \right)}{(\sigma)^3} \qquad \alpha_{3 \text{ Muestral}} = \frac{\left(\frac{\sum [(m_j - \bar{X})^3] f}{n-1} \right)}{(S)^3}$$

Donde:

α_3 = Coeficiente de asimetría.	f = Frecuencia de clase.
m_j = Punto medio de clase.	σ = Desviación estándar de la población.
μ = Media poblacional.	S = Desviación estándar de la muestra.
$\bar{X}$ = Media muestral	N = Tamaño de la población.
n = Tamaño de la muestra.	

La interpretación del índice de asimetría se define según el caso que se trate:

1. Si el índice de asimetría es igual o cercano a cero ($\alpha_3 = 0$), la distribución es simétrica o in sesgada; es decir, si la distribución es dividida exactamente a la mitad, y la figura de la primera mitad es idéntica a la otra mitad.
2. Si el índice de asimetría es mayor que cero ($\alpha_3 > 0$), la distribución es asimétricamente positiva o sesgada hacia la derecha, es decir, si la distribución es dividida exactamente a la mitad, se observará que la cola de la figura se extiende hacia la derecha de la distribución, mientras que su cima o valor más alto de la distribución se ubicará en la parte izquierda.

3. Si el índice de asimetría es menor que cero ($\alpha_3 < 0$), la distribución es asimétrica negativa o sesgada hacia la izquierda. Es decir, si la distribución es dividida exactamente a la mitad, se observará que la cola de la figura se encuentra hacia la izquierda de la distribución, mientras que su cima o valor más alto de la distribución se ubicará en la parte derecha.

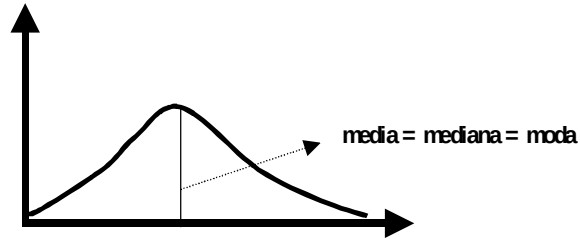


Figura 3.12. Distribución simétrica $\alpha_3 = 0$.

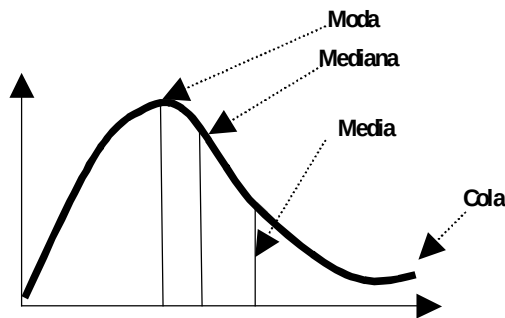


Figura 3.13. Distribución sesgada a la derecha $\alpha_3 > 0$.

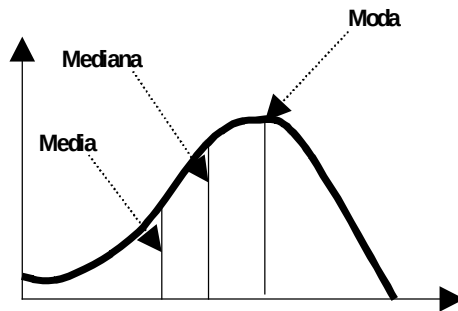


Figura 3.14. Distribución sesgada a la izquierda $\alpha_3 < 0$.

Ejemplo 33

Calcula el índice de asimetría para determinar qué tipo de sesgo tiene la siguiente serie de datos de una población: 1, 1, 2, 2, 2, 3, 3, 4, 5 y 6.

Para obtener el índice de asimetría, primero debemos encontrar cada uno de los elementos de su fórmula.

Se encuentra la media poblacional:

$$\mu = \frac{\sum X}{N} = \frac{(1+1+2+...+6)}{10} = \frac{29}{10} = 2.9$$

Se encuentra la varianza poblacional:

$$V(X) = \sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{24.9}{10} = 2.49$$

Se encuentra la desviación estándar:

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = 1.57$$

Se eleva al cubo la desviación estándar:

$$\sigma^3 = 3.86$$

Se obtiene la suma del cubo de las desviaciones con respecto a la media:

$$\Sigma(X - \mu)^3 = 24.9$$

Finalmente, se sustituyen estos resultados en la fórmula del índice de asimetría:

$$\alpha_3 = \frac{\left(\frac{\Sigma(X_j - \mu)^3}{n} \right)}{(\sigma)^3} = \frac{\left(\frac{24.48}{10} \right)}{3.86} = \frac{2.448}{3.86} = 0.6341$$

Se obtiene un índice de asimetría positivo, por lo que se puede decir que la distribución tiene un pequeño sesgo positivo o derecho. Si se observa la figura de la distribución de frecuencias, se notará que tiene una cola que se alarga hacia el lado derecho de la distribución:

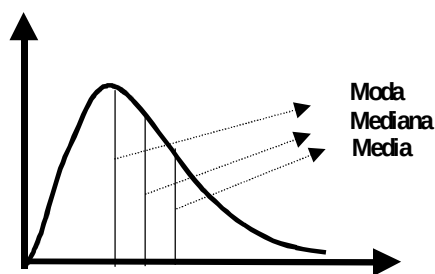


Figura 3.15. Distribución asimétrica positiva.

Ejemplo 34

Con la información del ejemplo 5, calcula el coeficiente de asimetría para saber hacia qué lado se carga la cola de la curva de estos datos.

Tiempo de servicio	f	m_j	$(m_j \cdot f)$	$(m_j - \bar{X})$	$(m_j - \bar{X})^3$	$[(m_j - \bar{X})^3] f$
5.00 – 8.99	3	6.995	20.985	-8.64	-644.972544	-1 934.917632
9.00 – 12.99	5	10.995	54.975	-4.64	-99.897344	-499.48672
13.00 – 16.99	7	14.995	104.965	-0.64	-0.262144	-1.835008
17.00 – 20.99	6	18.995	113.97	3.36	37.933056	227.598336
21.00 – 24.99	3	22.995	68.985	7.36	398.688256	1 196.064768
25.00 – 28.99	1	26.995	26.995	11.36	1 466.003456	1 466.003456
Σ	25		390.875			453.4272

Tabla 3.30. Distribución de las ventas por teléfono.

Los datos obtenidos son:

$$\bar{X} = \frac{\sum(m_j \cdot f)}{n} = \frac{390.875}{25} = 15.635$$

$$S^2 = \frac{\sum(m_j - \bar{X})^2 f}{n-1} = \frac{693.76}{(25-1)} = \frac{693.76}{24} = 28.90666667$$

$$S = \sqrt{\frac{\sum(m_j - \bar{X})^2 f}{n-1}} = \sqrt{28.90666667} = 5.376492041$$

El numerador de la fórmula empleada para calcular el coeficiente se denota por:

$$\frac{\sum(m_j - \bar{X})^3 f}{n-1} = \frac{453.4272}{24} = 18.8928$$

Con los datos anteriores, el coeficiente de asimetría es

$$\alpha_3 = \frac{\left(\frac{\sum(m_j - \bar{X})^3 f}{n-1} \right)}{(S^3)} = \frac{18.8928}{(5.376492041)^3} = \frac{18.8928}{155.4164633} = 0.121562411$$

Con el resultado se puede observar que el coeficiente es cercano a cero, así la distribución se caracteriza por ser incesgada, es decir, que la curva tiene una forma simétrica tal que las colas tienden a ser iguales.

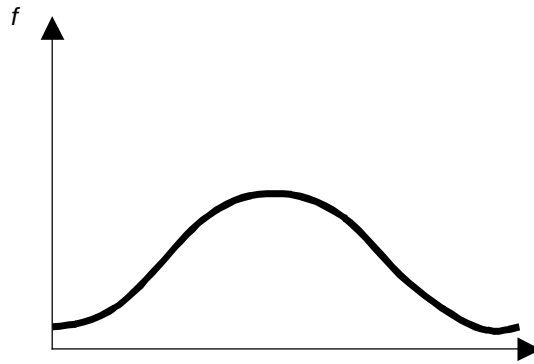


Figura 3.16. Distribución incesgada de las ventas por teléfono.

Ejercicio 11

1. En una distribución con $\alpha_3 = 0$:
 - a) Media, mediana y moda son diferentes
 - b) Media, mediana y moda coinciden en el mismo valor.
 - c) La media es mayor que la mediana y la moda.
 - d) La moda es mayor que la media y la mediana.
2. En una distribución con $\alpha_3 > 0$:
 - a) La mediana es mayor que la media y la moda.
 - b) Media, mediana y moda coinciden en el mismo valor.
 - c) La media es mayor que la mediana y la moda.
 - d) La moda es mayor que la mediana y la moda.
3. En una distribución $\alpha_3 < 0$:
 - a) La mediana es mayor que la media y la moda.
 - b) Media, mediana y moda coinciden en el mismo valor.
 - c) La media es mayor que la mediana y la moda.
 - d) La moda es mayor que la mediana y la moda.
4. Encuentra el índice de asimetría para una serie conformada por los siguientes datos provenientes de una muestra: 0, 1, 1, 3 y 5, y señala qué tipo de distribución es.
5. Con los datos de los ingresos anuales (en miles) de 300 familias que se presentan a continuación, calcula el coeficiente de asimetría para saber cómo es el sesgo de la distribución.

Ingreso (miles de \$)	<i>f</i>
1.50 – 2.999	25
3.00 – 4.999	31
5.00 – 6.999	42
7.00 – 8.999	45
9.00 – 10.999	52
11.00 – 12.999	42
13.00 – 14.999	35
15.00 – 16.999	28
Σ	300

Distribución de salarios

3.7.1. Kurtosis

El índice de kurtosis es una medida de dispersión mediante la cual se conoce qué tan concentrados o qué tan dispersos se encuentran los datos alrededor de la media. Su resultado representa el grado de apuntamiento de una distribución, es decir, qué tan puntiaguda o qué tan aplanada es la curva de una distribución. Cuando es muy puntiaguda se dice que los datos se encuentran muy concentrados alrededor de la media, mientras que si es muy chata o aplanada, se dice que existe una gran dispersión de los datos alrededor de la media.

Para encontrar el índice de kurtosis, las fórmulas dependen de la información con la que se trabaje y de la manera en que se encuentren organizados los datos, ya sea que se trate de una muestra o de una población, o que los datos se encuentren no agrupados o agrupados. Se representa mediante la expresión α_4 .

Para el caso de datos no agrupados, la kurtosis poblacional y muestral se expresan mediante las siguientes fórmulas:

$$\alpha_{4 \text{ Poblacional}} = \frac{\left(\frac{\sum (X_j - \mu)^4}{N} \right)}{(\sigma)^4} \quad \alpha_{4 \text{ Muestral}} = \frac{\left(\frac{\sum (X_j - \bar{X})^4}{n-1} \right)}{(S)^4}$$

Para el caso de datos agrupados, la kurtosis poblacional y muestral se obtienen utilizando las siguientes fórmulas:

$$\alpha_{4 \text{ Poblacional}} = \frac{\left(\frac{\sum (m_j - \mu)^4 f}{N} \right)}{(\sigma)^4} \quad \alpha_{4 \text{ Muestral}} = \frac{\left(\frac{\sum (m_j - \bar{X})^4 f}{n-1} \right)}{(S)^4}$$

Donde:

α_4 = Coeficiente de kurtosis

m_j = Punto medio de clase.

$\bar{X}$ = Media de la muestra.

f = frecuencia de la clase.

μ = Media poblacional.

n = Tamaño de la muestra.

N = Tamaño de la población.

σ = Desviación estándar poblacional.

S = Desviación estándar de la muestra.

La interpretación del índice de kurtosis se define según el caso que se trate:

1. Si el índice de kurtosis es igual a tres ($\alpha_4 = 3$), la distribución no es ni tan puntiaguda ni tan plana. A este tipo de distribución se le conoce como distribución mesocúrtica.
2. Si el índice de kurtosis es mayor a tres ($\alpha_4 > 3$), la distribución es muy puntiaguda, es decir, los datos se encuentran muy concentrados alrededor de la media. A este tipo de distribución se le conoce como distribución leptocúrtica.
3. Si el índice de kurtosis es menor a tres ($\alpha_4 < 3$), la distribución es muy plana, es decir, los datos se encuentran muy dispersos del valor de la media. A este tipo de distribución se le conoce como distribución platocúrtica.

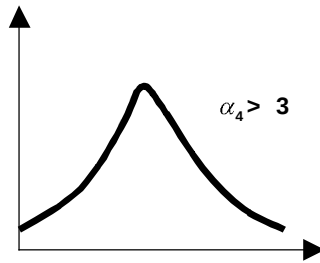


Figura 3.17. Distribución leptocúrtica.

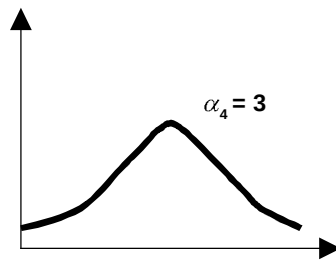


Figura 3.18. Distribución mesocúrtica.

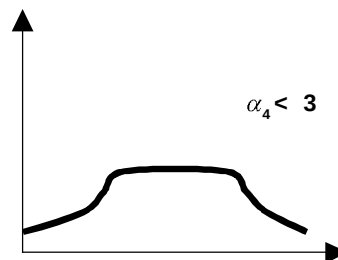


Figura 3.19. Distribución platocúrtica.

Ejemplo 35

Empleando los datos del ejemplo 5, calcula el coeficiente de kurtosis para saber cómo es la forma de la curva de estos datos.

Tiempo de servicio	f	m_j	$(m_j \cdot f)$	$(m_j - \bar{X})$	$(m_j - \bar{X})^4$	$[(m_j - \bar{X})^4] f$
5.00 – 8.99	3	6.995	20.985	-8.64	5 572.56278	1 6717.68834
9.00 – 12.99	5	10.995	54.975	-4.64	463.5236762	2 317.618381
13.00 – 16.99	7	14.995	104.965	-0.64	0.16777216	1.17440512
17.00 – 20.99	6	18.995	113.97	3.36	127.4550682	764.7304092
21.00 – 24.99	3	22.995	68.985	7.36	2 934.345564	8 803.036692
25.00 – 28.99	1	26.995	26.995	11.36	16 653.79926	16 653.79926
Σ	25		390.875			45 258.04749

Tabla 3.31. Distribución de las ventas por teléfono.

Obtenemos la información necesaria para encontrar la kurtosis

$$\bar{X} = \frac{\Sigma(m_j \cdot f)}{n} = \frac{390.875}{25} = 15.635$$

$$s^2 = \frac{\Sigma(m_j - \bar{X})^2 f}{n-1} = \frac{693.76}{(25-1)} = \frac{693.76}{24} = 28.90666667$$

$$S = \sqrt{\frac{\Sigma(m_j - \bar{X})^2 f}{n-1}} = \sqrt{28.90666667} = 5.376492041$$

El numerador de la fórmula empleada para calcular el coeficiente se denota por:

$$\frac{\sum(m_j - \bar{X})^4 f}{n-1} = \frac{45258.049}{24} = 1\,885.751979$$

Con los datos anteriores, el coeficiente de kurtosis es:

$$\alpha_4 = \frac{\left(\frac{\sum(m_j - \bar{X})^4 f}{n-1} \right)}{(S^4)} = \frac{1\,885.751979}{(5.376492041)^4} = \frac{1\,885.751979}{835.5953778} = 2.25677646$$

Con el resultado se puede observar que el coeficiente es menor a tres, por lo que la distribución se caracteriza por ser platicúrtica, es decir, que la curva tiene una forma tal que su apuntamiento es achatado, tal y como se muestra a continuación:

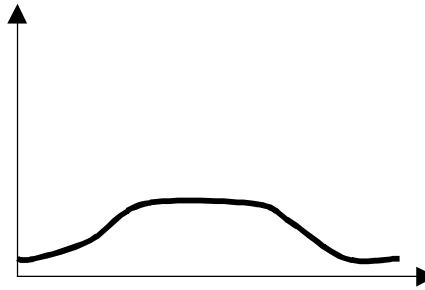


Figura 3.20. Distribución de las ventas por teléfono.

Ejercicio 12

1. El índice de kurtosis mide:
 - a) La simetría de una distribución.
 - b) Un valor típico o representativo de la distribución.
 - c) La dispersión existente entre el valor mayor y el menor.
 - d) El grado de apuntamiento de una distribución.
2. Si el índice de kurtosis α_4 es igual a tres, entonces
 - a) La distribución es asimétrica.
 - b) La distribución es mesocúrtica.
 - c) La distribución es leptocúrtica.
 - d) La distribución es platicúrtica.
3. Si el índice de kurtosis α_4 es menor a tres, entonces
 - a) La distribución es asimétrica.
 - b) La distribución es mesocúrtica.
 - c) La distribución es leptocúrtica.
 - d) La distribución es platicúrtica.
4. Si el índice de kurtosis α_4 es mayor a tres, entonces
 - a) La distribución es asimétrica.
 - b) La distribución es mesocúrtica.
 - c) La distribución es leptocúrtica.
 - d) La distribución es platicúrtica.
5. Con los siguientes datos de los ingresos anuales (en miles) de 300 familias, calcula el coeficiente de kurtosis para conocer cómo es la forma de la curva de distribución:

Ingreso (miles de \$)	f
1.50 – 2.999	25
3.00 – 4.999	31
5.00 – 6.999	42
7.00 – 8.999	45
9.00 – 10.999	52
11.00 – 12.999	42
13.00 – 14.999	35
15.00 – 16.999	28
Σ	300

Distribución de salarios

3.8. Agrupamiento de datos muestrales en el proceso decisorio de los negocios

El manejo de datos agrupados en los negocios proporciona mayor facilidad en la manipulación y procesamiento de la información que requerimos para la toma de decisiones al interior de la organización. Cuando iniciamos con mediciones del comportamiento de un determinado proceso podemos analizar todos los datos obtenidos sin la necesidad de agruparlos pero con el inconveniente de que si se decide ampliar el número de observaciones la complejidad puede aumentar y el tiempo de procesamiento también.

Otra opción es el uso de agrupaciones de datos que consiste en establecer clases (límites inferior y superior) y estimar la frecuencia de cada uno de ellos, con esto estaremos dando un peso ponderado, en otras palabras se generará una distribución de frecuencias.

Si se decide ampliar el número de observaciones no complicará la manipulación de la información dado que sólo incrementaremos la frecuencia que corresponda o generaremos más rangos.

En muchos casos se requiere dar un peso o valor o grado de importancia a los datos recabados durante las observaciones tal como peso, volumen, costo, etc., y se requiere conocer la media de estas observaciones, a este cálculo se le da el nombre de *media ponderada*.

Para calcular la media ponderada utilizaremos la fórmula siguiente:

$$\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$$

Donde:

X_i = valor del dato i .

w_i = frecuencia o peso del dato i .

Para el caso de una población aplica la misma fórmula pero sustituimos la $\bar{X}_w$ por μ_w .

Resumiendo, si el origen de los datos es una muestra, entonces estaremos hablando de la media ponderada de la muestra; pero si el origen es una población, entonces nos referiremos a la media ponderada poblacional.

Ejemplo 36

Un almacén general necesita actualizar sus tarifas de almacenaje por día para mejorar su competitividad. El administrador del almacén solicitó información sobre el costo promedio por metro cúbico de almacenaje.

Contrato	Costo por m³	m³
1	\$50.00	1
2	\$43.00	8
3	\$38.00	25
4	\$30.00	30
5	\$40.00	10
6	\$43.00	9
7	\$39.00	27
8	\$50.00	4
9	\$45.00	6
10	\$40.00	15

Para obtener la media ponderada aplicamos la fórmula descrita anteriormente, así que:

$$\bar{X}_w = \frac{50(1) + 43(8) + 38(25) + 30(30) + 40(10) + 43(9) + 39(27) + 50(4) + 45(6) + 40(15)}{1 + 8 + 25 + 30 + 10 + 9 + 27 + 4 + 6 + 15}$$

$$= \frac{5154}{135} = 38.18 \text{ Media ponderada.}$$

De acuerdo con el cálculo realizado, el costo promedio diario ponderado por m³ almacenado es de \$38.18.

Este resultado lo utilizaremos más adelante para tomar decisiones al respecto.

Continuando con el almacén general, el administrador necesita conocer cuál es el tiempo promedio que se almacenan las diferentes mercancías, para lo cual se desarrolló la siguiente distribución de frecuencias.

Días de almacenaje	Frecuencia (f_i)
1 a 10	20
11 a 20	28
21 a 30	19
31 a 40	30
41 a 50	12
51 a 60	14
61 a 70	25
71 a 80	8
81 a 90	15

Para determinar el punto medio de cada clase m_i , realizamos lo siguiente: (límite inferior + límite superior) / 2, así que el punto medio para la primera clase será $(1+10) / 2 = 5.5$, para la segunda clase será $(11+20) / 2 = 15.5$ y así sucesivamente.

Posteriormente realizamos el producto de la frecuencia (f_i) por el punto medio (m_i) de cada clase obteniendo la tabla siguiente:

Días de almacenaje	Frecuencia (f_i)	Punto medio de clase (m_i)	$f_i m_i$
1 a 10	20	5.5	110.0
11 a 20	28	15.5	434.0
21 a 30	19	25.5	484.5
31 a 40	30	35.5	1 065.0
41 a 50	12	45.5	546.0
51 a 60	14	55.5	777.0
61 a 70	25	65.5	1 637.5
71 a 80	8	75.5	604.0
81 a 90	15	85.5	1 282.5
	171		6 940.5

Aplicamos la fórmula para determinar el promedio muestral:

$$\bar{X} = \frac{\sum f_i m_i}{\sum f_i} = \frac{6\,940}{171} = 40.58 \text{ días de almacenaje.}$$

Continuando con el ejemplo procedemos a calcular la varianzay la desviación estándar:

Días de almacenaje	Punto medio de clase (m_i)	Frecuencia (f_i)	Desviación ($m_i - \bar{X}$)	$(m_i - \bar{X})^2$	$f_i(m_i - \bar{X})^2$
1 a 10	5.5	20	-35.08	1 230.61	24 612.13
11 a 20	15.5	28	-25.08	629.01	17 612.18
21 a 30	25.5	19	-15.08	227.41	4 320.72
31 a 40	35.5	30	-5.08	25.81	774.19
41 a 50	45.5	12	4.92	24.21	290.48
51 a 60	55.5	14	14.92	222.61	3 116.49
61 a 70	65.5	25	24.92	621.01	15 525.16
71 a 80	75.5	8	34.92	1 219.41	9 755.25
81 a 90	85.5	15	44.92	2 017.81	30 267.10
					106 273.69

Aplicando la fórmula de la varianzay para una muestra:

$$S^2 = \frac{\sum f_i(m_i - \bar{X})^2}{n-1} = \frac{106\,273.69}{170} = 625.14$$

La desviación estándar:

$$S = \sqrt{625.14} = 25.0027 = 25 \text{ días}$$

Conclusión sobre el problema.

Un costo promedio diario ponderado por m³ de almacenaje de \$38.18.

El promedio de días de almacenaje es de 40.58 días.

La varianzay es igual a 625.14 días².

La desviación estándar es de 25 días.

Con esta información el administrador puede fomentar mediante tarifas preferenciales a los contratos que tengan sus mercancías por menos de 40 días, con la finalidad de disponer de espacio en corto plazo, en otras palabras, rentar el espacio un mayor número de veces.

Autoevaluación

1. Una medida de tendencia central señala:
 - a) Hacia dónde se concentran los valores de una serie de datos.
 - b) El grado de dispersión de una muestra o una población.
 - c) Un valor que divide una serie de datos en cuatro partes iguales.
 - d) Una fórmula que presenta información de los percentiles.
2. Es el valor que ocupa el lugar central en una serie de datos, ubicándose 50% por encima de los datos con mayor valor y 50% por debajo de los datos con menor valor:
 - a) Media.
 - b) Percentil.
 - c) Mediana.
 - d) Moda.
3. La fórmula de la media se define como:
 - a) La variable que se presenta con mayor frecuencia en una distribución.
 - b) Una medida de dispersión.
 - c) Una medida que se divide en dos partes.
 - d) La suma total de las observaciones dividida entre el número de valores.
4. En la serie simple: 2, 4, 5, 1, 2, 5, 6, 8, el cuartil Q_1 está representado por:
 - a) 6
 - b) 3
 - c) 2
 - d) 4
5. En la serie simple: 2, 4, 5, 1, 2, 5, 6, 8, el decil D_5 está representado por:
 - a) 8
 - b) 3
 - c) 4
 - d) 1
6. En la serie 2, 4, 6, 5, 1, 7, 11, 6, 8, la moda está representada por:
 - a) 6
 - b) 5
 - c) 4
 - d) No hay moda.
7. Es una medida de tendencia central:
 - a) El rango.
 - b) La moda.
 - c) La varianza.
 - d) La desviación estándar.

8. La medida de tendencia central que toma en cuenta todos los datos de la serie es
- La moda
 - La mediana.
 - El cuartil.
 - La media.
9. Los datos agrupados se refieren:
- A los datos que se obtienen en forma aleatoria.
 - Los datos que se ordenan en forma ascendente.
 - Los datos organizados en una distribución de frecuencias.
 - Todos los valores observados de la variable que se enlistan.
10. Cuando la distribución de frecuencia es simétrica:
- La moda es diferente que la media.
 - La mediana es igual que la moda pero diferente que la media.
 - La moda y mediana son diferentes.
 - La media, la moda y la mediana son iguales.
11. Un banco toma una muestra de 20 analistas financieros y les pide que hagan una predicción sobre las ganancias por acción (dólares por acción) de una gran empresa para el próximo año. Los resultados que obtuvieron son:

Predicción (dólares por acción)	f	F_a
9.950 – 10.449	2	2
10.450 – 10.949	8	10
10.950 – 11.449	6	16
11.450 – 11.949	3	19
11.950 – 12.449	1	20
Σ	20	

Tabla 3.31. Distribución de las ganancias por acción.

Con los datos anteriores:

11.1. La media es

- 10.4720
- 11.0245
- 11.1034
- 12.0130

11.2. La mediana es

- 9.9895
- 10.3745
- 10.995
- 10.949

11.3. El valor de la moda es

- a) 10.8245
- b) 11.9450
- c) 12.3445
- d) 9.9745

11.4. El valor del cuartil dos (Q_2) es

- a) 10.0235
- b) 11.1035
- c) 10.9745
- d) 10.9495

11.5. El valor del decil cuatro (D_4) es

- a) 11.3425
- b) 10.9350
- c) 10.8245
- d) 10.4584

11.6. El valor del percentil sesenta (F_{60}) es

- a) 10.0130
- b) 11.1161
- c) 11.1040
- d) 12.3020

12. El rango se obtiene:

- a) Sumando el valor mínimo y el valor máximo de un conjunto de datos
- b) Restando al valor máximo, el valor mínimo de un conjunto de datos
- c) Restando al valor mínimo, el valor máximo de un conjunto de datos
- d) Promediando el valor máximo y el valor mínimo de un conjunto de datos

13. El rango tiene la siguiente característica:

- a) Es sensible a valores desproporcionados de un conjunto de datos
- b) No es sensible a valores desproporcionados de un conjunto de datos
- c) Es útil para calcular variaciones con datos desproporcionados
- d) No sirve para calcular variaciones con datos proporcionados

14. Una desventaja de la varianza radica en que:

- a) No señala la manera en que se concentran los datos
- b) No se puede encontrar la desviación estándar.
- c) Sus resultados son expresados en unidades al cuadrado.
- d) No considera las desviaciones respecto a la media.

15. Para calcular la varianza con datos agrupados
 - a) Se debe conocer la mediana.
 - b) Se debe conocer la moda.
 - c) Se debe conocer la frecuencia relativa.
 - d) Se debe conocer el punto medio de clase.
16. En los conjuntos cuya distribución es simétrica y acampanada:
 - a) Gran parte de los datos se encuentran cercanos de la media.
 - b) Gran parte de los datos se encuentran cercanos a la varianza.
 - c) Gran parte de los datos se encuentran cercanos al cero.
 - d) Pocos datos se agrupan alrededor de la moda.
17. El índice de kurtosis mide:
 - a) La simetría de una distribución.
 - b) Un valor típico o representativo de la distribución.
 - c) La dispersión existente entre el valor mayor y el menor.
 - d) El grado de apuntamiento de una distribución.
18. Una distribución que tiene un pico muy alto se denomina:
 - a) Distribución mesocúrtica.
 - b) Distribución de frecuencias.
 - c) Distribución platocúrtica.
 - d) Distribución leptocúrtica.
19. Si el índice de kurtosis α_4 es igual a tres, entonces
 - a) La distribución es asimétrica.
 - b) La distribución es mesocúrtica.
 - c) La distribución es leptocúrtica.
 - d) La distribución es platocúrtica.
20. Si el índice de kurtosis α_4 es menor a tres, entonces
 - a) La distribución es asimétrica.
 - b) La distribución es mesocúrtica.
 - c) La distribución es leptocúrtica.
 - d) La distribución es platocúrtica.
21. Si el índice de kurtosis α_4 es mayor a tres, entonces
 - a) La distribución es asimétrica.
 - b) La distribución es mesocúrtica.
 - c) La distribución es leptocúrtica.
 - d) La distribución es platocúrtica.

22. El coeficiente de asimetría mide:
- La altitud de la curva de distribución.
 - El sesgo de la curva de distribución.
 - La media de una distribución.
 - La varianza de una distribución.
23. Una distribución asimétrica negativa se caracteriza por:
- Tener un sesgo hacia la izquierda.
 - No tener sesgo.
 - Tener sesgo hacia la derecha.
 - Es simétrica.
24. Una empresa mayorista distribuidora de aparatos eléctricos desea estudiar sus cuentas por cobrar (en miles de pesos) para dos meses sucesivos (abril y mayo). Se seleccionan dos muestras independientes de cincuenta cuentas para cada uno de los meses. Los datos que se recolectaron son:

Monto (miles \$)	f (abril)	Monto (miles \$)	f (mayo)
0 – 1.999	6	0 – 1.999	10
2.000 – 3.999	13	2.000 – 3.999	14
4.000 – 5.999	17	4.000 – 5.999	13
6.000 – 7.999	10	6.000 – 7.999	10
8.000 – 9.999	4	8.000 – 9.999	0
10.000 – 11.999	0	10.000 – 11.999	3
Σ	50	Σ	50

Tabla 3.32. Distribución de las cuentas por cobrar.

Con los datos anteriores, para los meses de abril y mayo respectivamente:

24.1. Las medias son:

- 4.7195 y 4.3995
- 4.0078 y 4.4475
- 3.9945 y 4.5785
- 4.1975 y 4.9935

24.2. Las varianzas son:

- 5.5065 y 6.9958
- 5.0628 y 7.0612
- 5.2145 y 7.1628
- 5.0325 y 7.0022

24.3. Las desviaciones estándar son:

- 2.3465 y 2.6449
- 2.2835 y 2.6763
- 2.2500 y 2.6572
- 2.2433 y 2.6461

24.4. Los coeficientes de variación son:

- a) 47.67% y 60.39%
- b) 58.54% y 59.46%
- c) 57.16% y 58.45%
- d) 53.44% y 52.99%

24.5. Los coeficientes de kurtosis son:

- a) 2.3456 y 3.2658
- b) 2.5546 y 3.2245
- c) 2.1033 y 3.1555
- d) 2.2893 y 3.0866

24.6. Los coeficientes de asimetría son:

- a) 0.3022 y 0.8854
- b) 0.2547 y 0.4458
- c) 0.1029 y 0.6630
- d) 0.0547 y 0.7078

24.7. Según el apuntamiento del mes de abril, la distribución es

- a) Leptocúrtica.
- b) Platicúrtica.
- c) Mesocúrtica.
- d) Asimétrica positiva.

Respuestas a los ejercicios _____

Ejercicio 1

1. $\bar{X} = \frac{\sum X}{n} = \frac{5\,432\,354\,825.00}{97\,483\,412} = 55.72594$ miles de pesos por habitante.
2. $\bar{X} = \frac{\sum m_j f_j}{\sum f_j} = \frac{1\,757}{48} = 36.60$ años
3. $\bar{X} = \frac{\sum X}{n} = \frac{(37+30+23+\dots+26+13+42+38)}{25} = \frac{995}{25} = 39.8$ minutos.
5. $\bar{X} = \frac{\sum X}{n} = \frac{(142+163+108+\dots+139)}{20} = \frac{2\,740}{20} = 137$ chamarras
4. $\bar{X} = \frac{\sum m_j f_j}{\sum f_j} = \frac{1\,757}{48} = 36.60$ minutos
6. $\bar{X} = \frac{\sum m_j f_j}{\sum f_j} = \frac{2\,790}{75} = 37.2$

Ejercicio 2

1. $M_d = \frac{(10+10)}{2} = \frac{20}{2} = 10$. La mediana indica que se vendieron 10 automóviles
2. $M_d = \frac{(17+18)}{2} = \frac{35}{2} = 17.5$. El peso por paquete que se encuentra en la mediana es de 17.5 kg
3. $N_d = \frac{(n+1)}{2} = \frac{(11+1)}{2} = \frac{12}{2} = 6$. El valor que se encuentra colocado en el número 6 es \$35 000, que es la mediana de los salarios de los ejecutivos
4. Como es número par se toma el promedio de los valores que se encuentran en $N_{d1} = \frac{n}{2}$ y $N_{d2} = \frac{n+2}{2}$
 $N_{d1} = \frac{20}{2} = 10$ $N_{d2} = \frac{n+2}{2} = 11$
 El valor que ocupa el lugar 10 es 42 y el lugar 11 es 42, por tanto la mediana es 42, siendo la mediana del tiempo en minutos que duran las entrevistas
5. $M_d = L_j + \left[\frac{\left(\frac{n}{2} - F_a \right)}{f_m} \right] I = 35.005 + \left[\frac{\left(\frac{200}{2} - 60 \right)}{60} \right] 5 = 35.005 + \left[\frac{(100 - 60)}{60} \right] 5 = 35.005 + \left[\frac{40}{60} \right] 5$
 $M_d = 29.5 + \left[\frac{172.5}{27} \right] = 29.5 + 6.39 = 35.89$

$$6. \quad M_d = L_i + \left[\frac{\left(\frac{n}{2} - F_a \right)}{f_m} \right] I = 29.5 + \left[\frac{\left(\frac{75}{2} - 26 \right)}{27} \right] 15 = 29.5 + \left[\frac{37.5 - 26}{27} \right] 15 = 29.5 + \left[\frac{11.5}{27} \right] 15$$

$$M_d = 29.5 + \left[\frac{172.5}{27} \right] = 29.5 + 6.39 = 35.89$$

Ejercicio 3

1. La observación 10 representa la moda, ya que es la observación que se repite más veces.
2. No hay moda, ya que ninguna observación se repite.
3. El \$30 000 representa la moda, ya que es la observación que más se repite.
4. El 42 representa la moda, ya que es la observación que se repite más veces.

$$5. \quad M_o = L_i + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right] * I = 35.005 + \left[\frac{30}{(30+10)} \right] 5 = 35.005 + \left[\frac{30}{40} \right] 5 = 35.005 + \left[\frac{150}{40} \right]$$

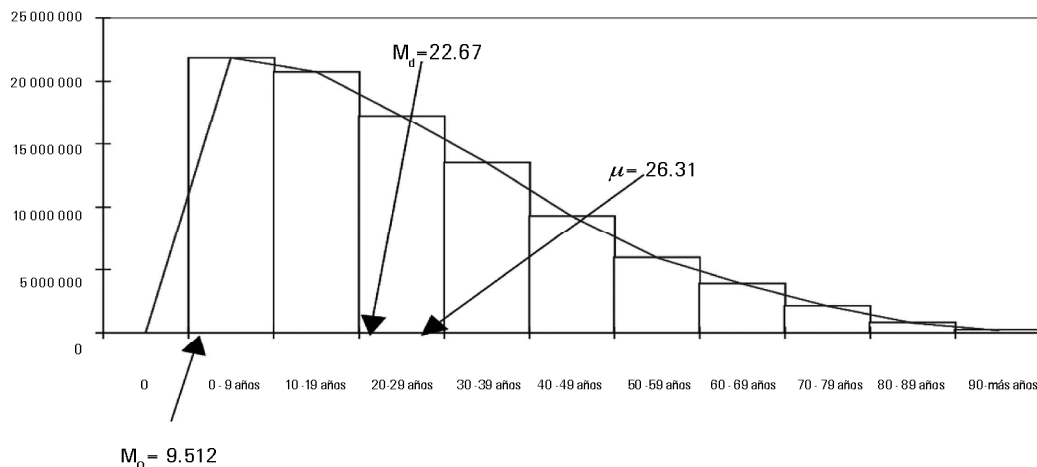
$$M_o = 29.5 + 5.45 = 34.95$$

$$6. \quad M_o = L_i + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right] * I = 29.5 + \left[\frac{8}{8+14} \right] 15 = 29.5 + \left[\frac{8}{22} \right] 15 = 29.5 + \left[\frac{120}{22} \right]$$

$$M_o = 29.5 + 5.45 = 34.95$$

Ejercicio 4

1. b)
2. c)
3. d)
4. Inciso a) Diagrama de frecuencias para la población en México.



4. Inciso b) La gráfica tiene la cima cargada hacia el lado derecho y su cola se encuentra en la parte izquierda, por lo que es una distribución de frecuencias asimétrica con sesgo derecho o positivo. Además, se observa que las medidas de tendencia central se encuentran de la siguiente manera:

Moda es (9.512) < Mediana (22.97) < Media (26.31)

Lo que confirma el resultado de que la distribución tiene sesgo derecho o positivo.

Ejercicio 5

1.

a) $N_o = 11 / 4 = 50$

$$Q_1 = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I = 30.005 + \left[\frac{(50 - 30)}{30} \right] 5 = 30.005 + \left[\frac{20}{30} \right] 5 = 30.005 + \left[\frac{100}{30} \right]$$

$$Q_1 = 30.005 + 3.33 = 33.335$$

b) $N_o = 41 / 10 = 80$

$$D_4 = 35.005 + \left[\frac{(80 - 60)}{60} \right] 5 = 35.005 + \left[\frac{20}{60} \right] 5 = 35.005 + \left[\frac{100}{60} \right] = 35.005 + 1.66 = 36.665$$

c) $N_o = 63 / 100 = 126$

$$P_{63} = 40.005 + \left[\frac{(126 - 120)}{50} \right] 5 = 40.005 + \left[\frac{6}{50} \right] 5 = 40.005 + \left[\frac{30}{50} \right] = 40.005 + 0.6 = 40.605$$

2.

a) $N_o = 31 / 4 = 56.25$

$$Q_3 = L_i + \left[\frac{(N_o - F_a)}{f_c} \right] I = 44.5 + \left[\frac{(56.25 - 53)}{13} \right] 15 = 44.5 + \left[\frac{3.25}{13} \right] 15 = 44.5 + \left[\frac{48.75}{13} \right]$$

$$Q_3 = 44.5 + 3.75 = 48.25$$

b) $N_o = 51 / 10 = 37.5$

$$D_5 = 29.5 + \left[\frac{(37.5 - 26)}{27} \right] 15 = 29.5 + \left[\frac{11.5}{27} \right] 15 = 29.5 + \left[\frac{172.5}{27} \right] = 29.5 + 6.38 = 35.88$$

c) $N_o = 36 / 100 = 27$

$$P_{36} = 29.5 + \left[\frac{(27 - 26)}{27} \right] 15 = 29.5 + \left[\frac{1}{27} \right] 15 = 29.5 + \left[\frac{14}{27} \right] = 29.5 + 0.55 = 30.51$$

Ejercicio 6

1. a)
2. d)
3. b)
4. a)

5. Para hallar el rango se debe identificar el valor más alto y el valor más bajo. El valor máximo es 18 700 y el valor mínimo es 9 000. Por esto, el rango es

$$\text{Rango} = \text{Valor máximo} - \text{Valor mínimo} = 18\,700 - 9\,000 = 9\,700$$

El rango es 9 700, por lo que la diferencia existente entre los cobros es de 9 700 pesos, mostrando que la variabilidad es considerable por la diferencia existente.

Ejercicio 7

- b)
- d)
- c)
- b)
- En primer lugar, hay que calcular la media de los datos para posteriormente obtener las desviaciones y las desviaciones cuadradas.
La media se define por:

$$\bar{X} = \frac{\sum X}{N} = \frac{(10\,000 + 12\,000 + 15\,000 + \dots + 14\,000 + 18\,700 + 16\,500)}{15} = \frac{205\,900}{15}$$

$$\bar{X} = 13\,726.66667$$

Como el cálculo de la varianza requiere obtener las diferencias de los datos con respecto a la media $(X - \bar{X})$ y las diferencias cuadradas $(X - \bar{X})^2$, se tiene:

X	$(X - \bar{X})$	$(X - \bar{X})^2$	X	$(X - \bar{X})$	$(X - \bar{X})^2$
10 000	-3 726.666667	13 888 044.45	9 700	-4 026.666667	16 214 044.45
12 000	-1 726.666667	2 981 377.779	18 000	4 273.333333	18 261 377.77
15 000	1 273.333333	1 621 377.777	13 200	-526.6666666	277 377.7777
16 000	2 273.333333	5 168 044.443	12 600	-1 126.666667	1 269 377.779
15 000	1 273.333333	1 621 377.777	14 000	273.3333334	74 711.11115
9 000	-4 726.666667	18 289 877.78	18 700	4 973.333333	24 734 044.44
13 500	-226.6666666	51 377.77775	16 500	2 773.333333	7 691 377.776
12 700	-1 026.666667	1 054 044.445	Σ		113 197 833.3

Desviaciones de los cobros.

Cada diferencia se obtuvo restando a cada valor de X la media, y la diferencia cuadrada elevando al cuadrado cada diferencia. Empleando la fórmula para determinar la varianza, considerando que los datos de cobros constituyen la muestra:

$$V(X) = S^2 = \frac{\sum (X - \bar{X})^2}{n-1} = \frac{113\,197\,833.3}{14} = 8\,085\,559.521$$

Puede apreciarse que la varianza es de 8 085 559.521, pero no tiene significado práctico esta medida de variabilidad.

- Las dases denotan el ingreso familiar anual y la frecuencia el número de familias encuestadas

Ingreso (miles de \$)	f	m_i	$(m_i - f)$	$(m_i - \bar{X})$	$(m_i - \bar{X})^2$	$[(m_i - \bar{X})]^2 f$
1.50 - 2.999	25	2.2495	56.2375	-6.93083	48.03640449	1 200.910112
3.00 - 4.999	31	3.9995	123.9845	-8.78083	77.10297549	2 390.192240
5.00 - 6.999	42	5.9995	251.979	-3.18083	10.11767949	424.942539
7.00 - 8.999	45	7.9995	359.9775	-1.18083	1.39435949	62.746177
9.00 - 10.999	52	9.9995	519.974	0.81917	0.67103949	34.894053
11.00 - 12.999	42	11.9995	503.979	2.81917	7.94771949	333.804219
13.00 - 14.999	35	13.9995	489.9825	4.81917	23.22439949	812.853982
15.00 - 16.999	28	15.9995	447.986	6.81917	46.50107949	1 302.030226
Σ	300		2 754.100			6 562.373548

Distribución de los ingresos familiares.

Para obtener la varianza, en primer lugar, se debe calcular el punto medio de clase sumando los límites inferior y superior de cada clase y, posteriormente, dividir la suma entre dos. Al haber calculado el punto medio de cada clase, debe multiplicarse éste por la frecuencia y sumar los productos para obtener la media, tal y como se muestra a continuación:

$$\bar{X} = \frac{\sum(m_j \cdot f)}{n} = \frac{2\,754}{300} = 9.18033 \longrightarrow \text{La clase que presenta la media es la quinta}$$

Después de calcular la media se debe restar a cada punto medio de clase la media, elevar la diferencia al cuadrado y multiplicar esta diferencia cuadrática por la frecuencia respectiva para obtener la varianza de la manera siguiente:

$$s^2 = \frac{\sum[(m_j - \bar{X})^2 f]}{n-1} = \frac{6\,562.373548}{(300-1)} = \frac{6\,562.3735}{299} = 21.94773762$$

La varianza obtenida no tiene un sentido lógico, sólo puede deducirse que la clase que presenta la mayor variación es la quinta, donde el rango promedio de ingresos familiares está entre 9 y 10.999 miles de pesos, para 52 familias

Ejercicio 8

1. Partiendo de los datos del punto 1 del ejercicio se tiene:

$$s^2 = 8\,085\,559.521$$

La desviación estándar es la raíz cuadrada de la varianza:

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} = \sqrt{8\,085\,559.521} = 2\,843.511829$$

El resultado muestra que los cobros de la empresa tienen una variación de 2 843.511829 pesos, es decir, que puede cobrar 2 843.511829 pesos por abajo o por encima de la media.

2. Con los datos del punto 2 ejercicio, la desviación estándar será:

$$s^2 = 21.94773762$$

$$S = \sqrt{\frac{\sum(m_j - \bar{X})^2 f}{n-1}} = \sqrt{21.94773762} = 4.684841259$$

Esto muestra que la variación que existe entre los ingresos recibidos por las familias es en promedio de 4.684841259 miles de pesos.

Ejercicio 9

1. c)
2. a)
3. d)
4. d)
5. a)
6. b)
7. d)

8.

$$\left(1 - \frac{1}{(3)^2}\right) 100\% = \left(1 - \frac{1}{9}\right) 100\% = [1 - (0.1111)] 100 = (0.8889) (100) = 88.89\%$$

Valor inferior: $16.5 - (3) (4.3) = 3.6$

Valor superior: $16.5 + (3) (4.3) = 29.4$

9.

$$\left(1 - \frac{1}{(1.2)^2}\right) 100\% = \left(1 - \frac{1}{1.44}\right) 100\% = [1 - (0.6944)] 100 = (0.3056) (100) = 30.56\%$$

Valor inferior: $2000 - (1.2) (300) = 1\ 640$

Valor superior: $2000 + (1.2) (300) = 2\ 360$

10.

$$\left(1 - \frac{1}{(2.4)^2}\right) 100\% = \left(1 - \frac{1}{5.76}\right) 100\% = [1 - (0.1736)] 100 = (0.8263) (100) = 82.63\%$$

Valor inferior: $95 - (2.4) (25) = 35$

Valor superior: $95 + (2.4) (25) = 155$

Ejercicio 10

1. c)

2. a)

3. d)

4. En primer lugar se debe obtener la media de los rendimientos, es decir, el rendimiento promedio por acción para calcular las desviaciones cuadradas de los datos con respecto a la media:

$$X_D = \frac{\sum X}{n} = \frac{(150 + 125 + 120 + 200 + 250 + 175 + 200 + 250 + 180 + 140)}{10} = \frac{1\ 790}{10} = 179$$

$$X_L = \frac{\sum X}{n} = \frac{(200 + 275 + 180 + 195 + 280 + 250 + 240 + 200 + 300 + 290)}{10} = \frac{2\ 410}{10} = 241$$

Como puede observarse, la moneda que presenta la mayor venta promedio es la libra con una venta promedio de 247 transacciones diarias, mientras que el dólar tiene una venta promedio de 179 transacciones diarias.

El siguiente paso es obtener las diferencias de cada valor con respecto a su media, elevarlas al cuadrado y sumarlas para obtener la varianza y, posteriormente, sacarles la raíz cuadrada para obtener la desviación estándar:

Dólar:

$$\sum (X - \bar{X})_D^2 = (150 - 179)^2 + (125 - 179)^2 + \dots + (140 - 179)^2 = 19\ 740$$

Libra:

$$\sum (X - \bar{X})_L^2 = (200 - 247)^2 + (275 - 247)^2 + \dots + (290 - 247)^2 = 18\ 200$$

Una vez calculadas las sumas de las desviaciones cuadradas se procede a definir las varianzas:

$$V(X)_D = S^2 = \frac{\sum (X - \bar{X})^2}{n-1} = \frac{19\ 740}{9} = 2\ 193.333333$$

$$V(X)_L = S^2 = \frac{\sum (X - \bar{X})^2}{n-1} = \frac{18\ 200}{9} = 2\ 022.222222$$

Puede observarse que la varianza de las transacciones de la libra es menor que la varianza de las transacciones del dólar. Esto indica que la venta de dólares tiene una mayor dispersión, pero no se puede tener una interpretación coherente del resultado debido a que se emplean cuadrados, por lo cual se sacan las raíces para obtener las desviaciones estándar:

$$S_D = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} = \sqrt{2\,193.333333} = 46.83303677$$

$$S_L = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} = \sqrt{2\,022.222222} = 44.96912521$$

Puede observarse que la libra tiene una mayor venta promedio con una variabilidad de 44.96922521, pero la venta del dólar tiene una variabilidad mayor (46.83303677).

Con estos datos es posible calcular el coeficiente de variación y comparar la variabilidad de las transacciones de ambas monedas en términos porcentuales

$$CV_D = \left(\frac{S_D}{\bar{X}_D} \right) 100\% = \left(\frac{46.83303677}{179} \right) 100 = (0.261637076) (100) = 26.1637076\%$$

$$CV_L = \left(\frac{S_L}{\bar{X}_L} \right) 100\% = \left(\frac{44.96912521}{247} \right) 100 = (0.182061235) (100) = 18.2061235\%$$

Con los resultados se aprecia que las transacciones de libras presentan la menor variación, por lo que se puede concluir que la libra resulta ser una moneda más estable en cuanto a su venta.

Ejercicio 11

1. b)
2. c)
3. d)
- 4.

Se obtiene la desviación estándar de la serie: $S = 2$.

Se obtiene la suma del cubo de las desviaciones.

$$\sum (X - \bar{X})^3 = (-2)^3 + (-1)^3 + (-1)^3 + (3)^3 + (5)^3 = -8 -1 -1 +1 +27 = 18$$

Se obtiene el índice muestral de asimetría para datos no agrupados

$$\alpha_3 = \frac{\left(\frac{\sum(X_j - \bar{X})^3}{n-1} \right)}{(S)^3} = \frac{\left(\frac{18}{4} \right)}{(2)^3} = 0.56. \text{ Tiene sesgo derecho o positivo.}$$

5. Empleando los datos se tiene lo siguiente:

Ingreso (miles de \$)	f	n_i	$(u_i \cdot f)$	$(u_i - \bar{X})$	$(u_i - \bar{X})f$	$(u_i - \bar{X})^4$	$[(u_i - \bar{X})^4] f$
1.50 – 2.999	25	2.2495	56.2375	-6.93083	1200.91	2 307.496156	57 687.4039
3.00 – 4.999	31	3.9995	123.9845	-8.78083	2390.1899	5 944.868829	184 290.9337
5.00 – 6.999	42	5.9995	251.979	-3.18083	424.9392	102.3674382	4 299.432404
7.00 – 8.999	45	7.9995	359.9775	-1.18083	62.7435	1.944238384	87.49072728
9.00 – 10.999	52	9.9995	519.974	0.81917	34.892	0.450293995	23.41528774
11.00 – 12.999	42	11.9995	503.979	2.81917	333.8034	63.16624507	2 652.982293
13.00 – 14.999	35	13.9995	489.9825	4.81917	812.8505	539.3727316	18 878.04561
15.00 – 16.999	28	15.9995	447.986	6.81917	1302.028	2 162.350394	60 545.81103
Σ	300		2 754.100		6568.3565		328 465.5149

Distribución de los ingresos familiares

$$\bar{X} = \frac{\Sigma(m_j \cdot f)}{n} = \frac{2\,754.1}{300} = 9.18$$

$$S^2 = \frac{\Sigma(m_j - \bar{X})^2 f}{n-1} = \frac{6\,562.3565}{299} = 21.9476$$

$$S = \sqrt{\frac{\Sigma(m_j - \bar{X})^2 f}{n-1}} = \sqrt{21.9476} = 4.6848$$

El denominador de la fórmula empleada para calcular el coeficiente se denota por:

$$\frac{\Sigma(m_j - \bar{X})^4 f}{n-1} = \frac{328\,465.5149}{299} = 1\,098.546872$$

Con los datos anteriores, el coeficiente de kurtosis es:

$$\alpha_4 = \frac{\left(\frac{\Sigma(m_j - \bar{X})^4 f}{n-1} \right)}{(S^4)} = \frac{1\,098.546872}{(4.685)^2} = \frac{1\,098.546872}{481.768478} = 28.407967$$

Con el resultado se puede observar que el coeficiente es mayor a tres, por lo que la distribución se caracteriza por ser leptocúrtica, es decir, que la curva tiene una forma tal que su apuntamiento es muy alto.

Ejercicio 12

1. d)
2. b)
3. d)
4. c)
5. Los datos a emplear son:

Ingreso (miles de \$)	f	n_j	$(m_j \cdot f)$	$(m_j - \bar{X})$	$(m_j - \bar{X})^3$	$[(m_j - \bar{X})^3] f$
1.50 – 2.999	25	2.2495	56.2375	-6.93083	-332.9321533	-8 323.303833
3.00 – 4.999	31	3.9995	123.9845	-8.78083	-677.0281203	-20 987.87173
5.00 – 6.999	42	5.9995	251.979	-3.18083	-32.18261845	-1 351.669975
7.00 – 8.999	45	7.9995	359.9775	-1.18083	-1.646501515	-74.09256818
9.00 – 10.999	52	9.9995	519.974	0.81917	0.549695418	28.58416174
11.00 – 12.999	42	11.9995	503.979	2.81917	22.40597235	941.0508387
13.00 – 14.999	35	13.9995	489.9825	4.81917	111.9223293	3 917.281526
15.00 – 16.999	28	15.9995	447.986	6.81917	317.0987662	8 878.765454
Σ	300		2 754.100			-16 971.25623

Distribución de los ingresos familiares

Los datos obtenidos son:

$$\bar{X} = \frac{\Sigma(m_j \cdot f)}{n} = \frac{2\,754.1}{300} = 9.18$$

$$s^2 = \frac{\sum(m_j - \bar{X})^2 f}{n-1} = \frac{6\,562.3565}{299} = 21.9476$$

$$S = \sqrt{\frac{\sum(m_j - \bar{X})^2 f}{n-1}} = \sqrt{21.9476} = 4.6848$$

El denominador de la fórmula empleada para calcular el coeficiente se denota por:

$$\frac{\sum(m_j - \bar{X})^3 f}{n-1} = \frac{-16\,971.25623}{299} = -54.160054$$

Con los datos anteriores, el coeficiente de asimetría es

$$\alpha_3 = \frac{\left(\frac{\sum(m_j - \bar{X})^3 f}{n-1} \right)}{(S)^3} = \frac{-54.160054}{(4.685)^3} = -0.5519676$$

Con el resultado se puede observar que el coeficiente es negativo, por lo que la distribución se caracteriza por ser sesgada negativamente, es decir, que la curva tiene una forma tal que la cola izquierda es más larga.

Respuestas a la autoevaluación

-
1. a)
 2. c)
 3. d)
 4. c)
 5. c)
 6. a)
 7. b)
 8. d)
 9. c)
 10. d)
 11.
 - 11.1. b)
 - 11.2. d)
 - 11.3. a)
 - 11.4. d)
 - 11.5. c)
 12. b)
 13. a)
 14. c)
 15. d)
 16. a)
 17. d)
 18. d)
 19. b)
 20. d)
 21. c)
 22. b)
 23. a)
 24.
 - 24.1. a)
 - 24.2. b)
 - 24.3. c)
 - 24.4. a)
 - 24.5. d)
 - 24.6. c)
 - 24.7. b)