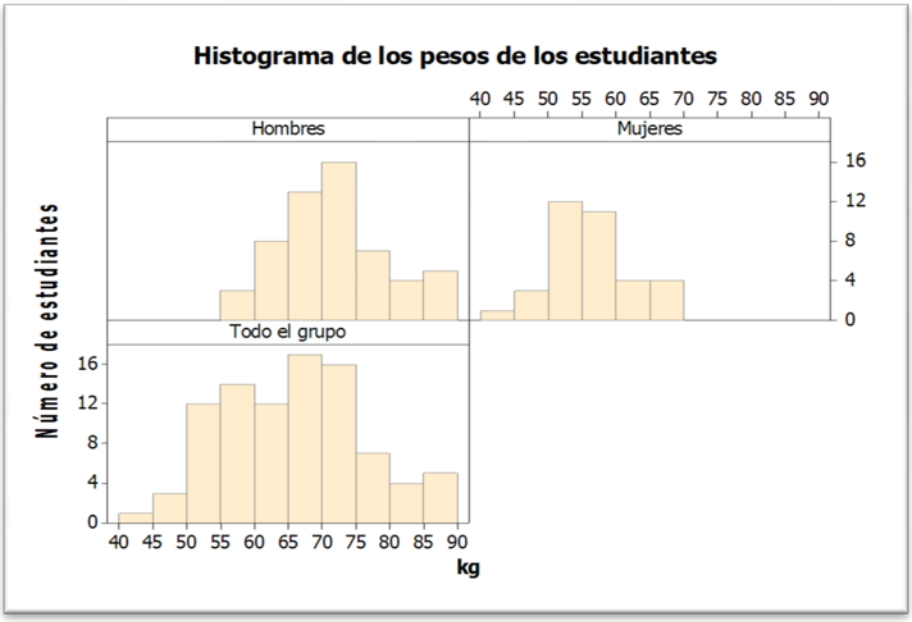
MEDIDAS DE TENDENCIA CENTRAL Y DISPERSIÓN

Suponga que le pedimos a un grupo de estudiantes de la asignatura de estadística que registren su peso en kilogramos. Con los datos del peso de los estudiantes obtenemos el histograma de los pesos para el grupo de estudiantes y un histograma para el peso de las mujeres y uno para el de los hombres.



¿Qué nos revelan los histogramas?

Un histograma es una gráfica muy utilizada en estadística. Se utiliza para datos cuantitativos y nos muestra la acumulación o tendencia de los datos, su variabilidad y la forma de la distribución. Entonces a partir de los histogramas elaborados, observamos que:

- La acumulación o tendencia del peso de los hombres se encuentra entre los 70 y 74 kg, mientras que la tendencia del peso de las mujeres es menor y se encuentra entre los 50 a 58 Kg.

- La variabilidad de todo el grupo está en un rango comprendido entre los 42 y los 90 kg. Sí se estudia únicamente el peso de los hombres se observa que se reduce la variabilidad y los pesos se encuentran ahora entre 54 y 90 kg. Para el grupo de las mujeres la variabilidad se reduce aún más y sus pesos se encuentran entre 42 y 70 Kg.

La variabilidad de todo el grupo es lógico que sea la mayor debido a que el grupo es muy heterogéneo ya que incluye los pesos de los hombres y de las mujeres. El que el grupo de las mujeres tenga menor variabilidad que el de los hombres nos indica que el grupo de las mujeres en cuanto a peso es más homogéneo que el grupo de los hombres.

Con los histogramas tenemos una medida burda de la tendencia y de la variabilidad

¿Se puede medir de una forma más precisa la acumulación o tendencia y la variabilidad?

La respuesta es afirmativa. Las medidas de tendencia o acumulación se conocen como medidas de Tendencia Central o de localización y las de variabilidad como medidas de dispersión o de variabilidad.

¿Cuáles son las medidas de tendencia central?

Las medidas de tendencia central más utilizadas, son la media aritmética, la mediana y la moda.

¿Qué es la media aritmética?

La media aritmética es la medida de tendencia central más utilizada y es igual a lo que conocemos como promedio. Entonces la media es la suma de los valores de todas las observaciones, dividida entre el número de observaciones realizadas.

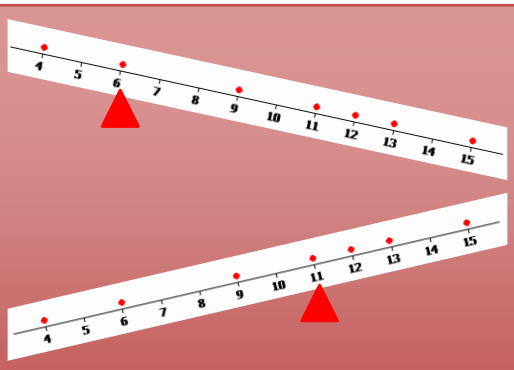
Sea n el tamaño de una muestra que contiene a las observaciones x_1, x_2, x_3, . . . , x_n, entonces la media aritmética, $\bar{x}$ es:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

En donde el subíndice i, indica un número de conteo para identificar cada observación.

La media de los números $x_1 = 13$, $x_2 = 15$, $x_3 = 9$, $x_4 = 6$, $x_5 = 4$, $x_6 = 12$, $x_7 = 11$ es:

$$\bar{x} = \frac{13+15+9+6+4+12+11}{7} = 10$$



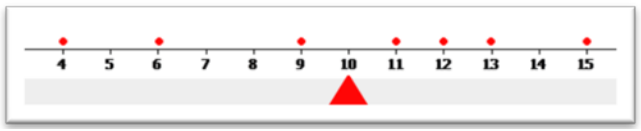
No hay equilibrio

El equilibrio se encuentra en 10. La suma de las distancias de los puntos a la derecha del punto de equilibrio es igual a la suma de las distancias de los puntos a la izquierda del punto de equilibrio

|4 – 10| + |6 – 10| + |9 – 10| = |11 – 10| + |12 – 10| + |13 – 10| + |15 – 10|

¿Por qué la media aritmética es una medida de tendencia central?

Una media 10, no significa que todos los datos sean igual a 10, es más para nuestros datos ninguno es igual a 10. Hay valores mayores y menores de 10. Veamos la gráfica de puntos siguiente, donde en una escala apropiada en el eje de las *X*, se representa cada dato mediante un punto. Si obtenemos las distancias de cada punto con respecto a 10, observamos que la suma de las distancias de los puntos a la derecha de 10, es igual a la suma de las distancias de los puntos a la izquierda de 10. Entonces, en 10 se equilibra la distribución de los datos, es decir es el punto de equilibrio o centro de gravedad de la distribución de los datos.



¿Se puede calcular la media aritmética a partir de los datos agrupados en una tabla de frecuencias?

La respuesta es afirmativa. Si lo único que tenemos es un resumen de los datos, en forma de tabla de frecuencias y no contamos con la información original, sí es posible calcular la media aritmética.

Con el fin de evitar cálculos aritméticos tediosos, no hace muchos años cuando se tenían numerosos datos, los datos originales se resumían en una tabla de frecuencias, y después se calculaban sus medidas de tendencia central y de variabilidad.

Hoy en día con el uso de software adecuado se pueden procesar fácilmente los datos originales, y ya no se justifica por éste motivo construir la tabla de frecuencias. Sin

Tiempo invertido en atender al cliente	No de clientes
141 – 157	2
157 – 173	13
173 – 189	17
189 – 205	14
205 – 221	3
221 - 237	1

"-" Indica a menos de:

Suponga que la información que tenemos es la siguiente Tabla que muestra el tiempo que tardaron 50 clientes en una caja bancaria y deseamos conocer cuál es el tiempo promedio que tardaron.

Sabemos, por ejemplo, que en la primera clase 2 clientes tardaron en la caja entre 141 y casi 157 segundos. No sabemos con exactitud cuánto tardó cada uno de ellos, sólo sabemos que tardaron un tiempo comprendido entre éstos dos límites. Para efectuar el cálculo de la media aritmética, supondremos que un valor representativo de la clase es su marca de clase ó punto medio, x_i

$$x_i = \frac{(\lim\ \sup + \lim\ \inf)_i}{2}$$

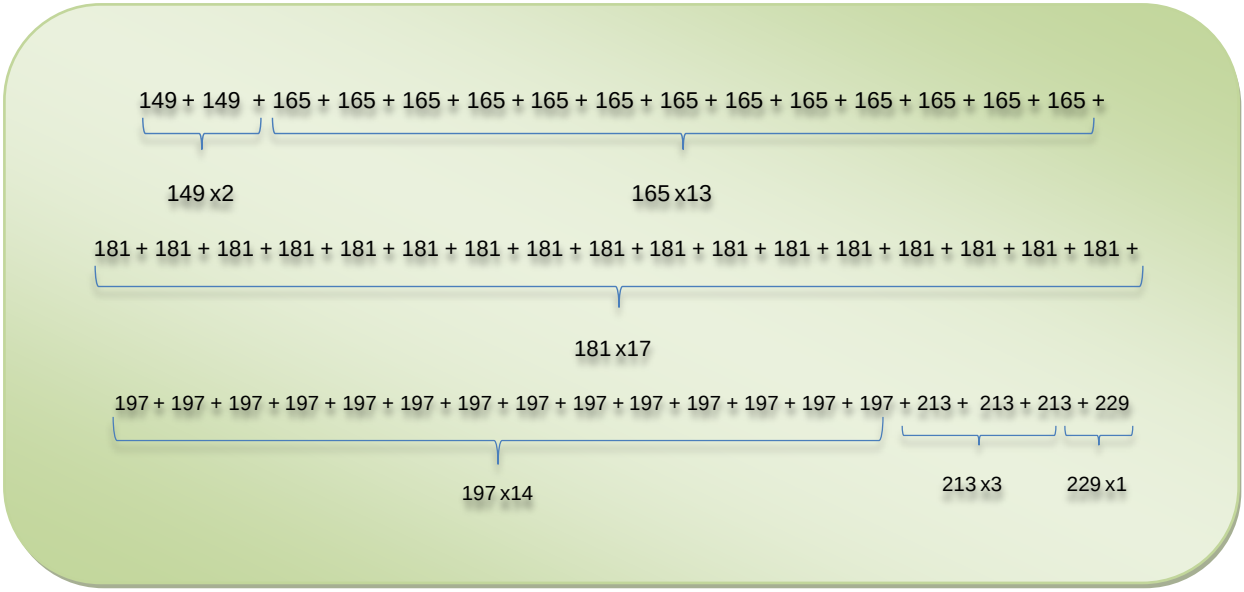
Entonces tenemos:

Tiempo invertido en atender al cliente	No de clientes f_i	Marca de clase x_i
141 – 157	2	149
157 – 173	13	165
173 – 189	17	181
189 – 205	14	197
205 – 221	3	213
221 - 237	1	229

"-" Indica a menos de:

¿Es un cálculo exacto? No, sólo es un valor aproximado.

Es decir, suponemos que tenemos 2 clientes que tardaron 149 segundos en la caja, 13 que tardaron 165, 17 que tardaron 181 segundos, etc. Entonces la suma de todos los datos sería igual a sumar 2 veces 149 más 13 veces 165 más 17 veces 181 más 14 veces 197 más 3 veces 213 más 1 vez 229.



La suma total es igual a:

$149 \times 2 + 165 \times 13 + 181 \times 17 + 197 \times 14 + 213 \times 3 + 229 \times 1 = 91$

Observe que la suma total es la suma de los productos marca de clase por frecuencia para cada clase.

La media será igual a la suma obtenida dividida entre el número de datos. Observe que se sumaron 50 datos, y que 50 es la suma de la columna de frecuencias, entonces:

$$\bar{x} = \frac{9146}{50} = 182.92\ seg$$

El cálculo anterior lo podemos sistematizar obteniendo una columna adicional en la Tabla de distribución de frecuencias. La columna expresará los productos $x_i f_i$ para cada clase.

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	$x_i f_i$
141 – 157	2	149	298
157 – 173	13	165	2145
173 – 189	18	181	3077
189 – 205	14	197	2758
205 – 221	3	213	639
221 - 237	1	229	229
Totales	50		9146

"-" Indica a menos de:

La suma de esta columna, 9146, entre el número de datos, nos da el valor de la media.

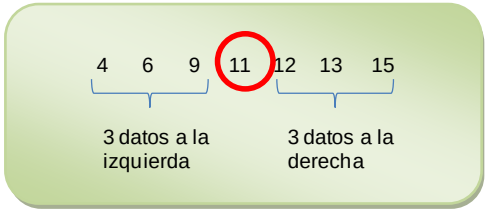
A partir de los cálculos realizados podemos escribir la expresión para la media calculada a partir de los datos agrupados en la Tabla de distribución de frecuencias.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$$

$$\bar{x} = \frac{9146}{50} = 182.92$$

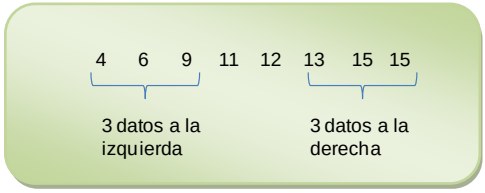
¿Qué es la mediana?

La mediana es el valor central que se localiza en una serie ordenada de datos. Para obtener la mediana de los números x_1 = 13, x_2 = 15, x_3 = 9, x_4 = 6, x_5 = 4, x_6 = 12, x_7 = 11, primero tenemos que ordenarlos:



Entonces la mediana es 11.

Si el número de datos fuera par, tendríamos dos valores centrales y la mediana sería la media de estos dos valores. Por ejemplo:

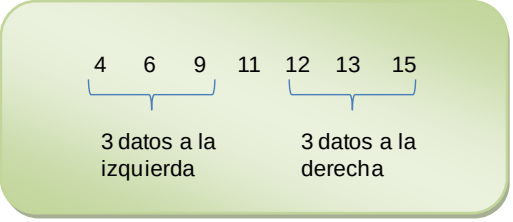


Tenemos dos valores centrales, 11 y 12, entonces la mediana es:

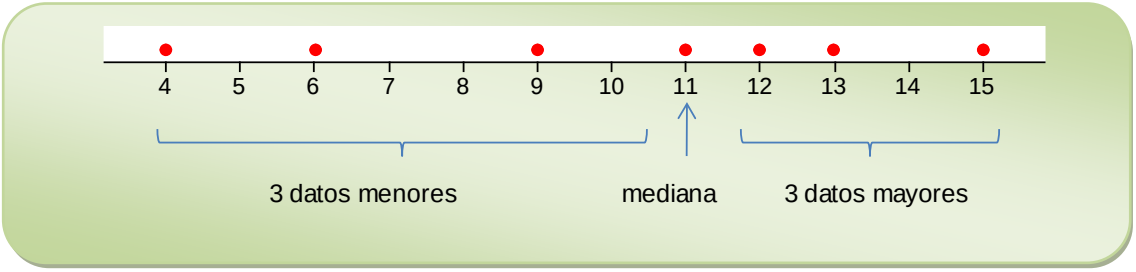
$$Mediana = \frac{11 + 12}{2} = 11.5$$

¿Qué representa la mediana?

Observe que la mediana divide la serie de datos en dos mitades y cada mitad tiene el mismo número de datos que la otra.



Representemos los datos mediante una gráfica de puntos. Por arriba de la mediana, 11, hay tres datos y por abajo también, sin importar el valor de los datos. Sólo toma en cuenta el número de datos y no le da importancia al hecho de que los valores por arriba de 11, estén más cerca de ella que los que están por debajo.



La mediana es el centro geométrico de la distribución de los datos.

¿Se puede calcular la mediana a partir de los datos agrupados en una tabla de frecuencias?

La respuesta es afirmativa. Al igual que la media, sí es posible calcular la mediana si sólo se cuenta con un resumen de los datos en forma de tabla de distribución de frecuencias.

A partir de nuestro ejemplo del tiempo que tardan unos clientes en una caja bancaria, calculemos la mediana.

Tiempo invertido en atender al cliente	No de clientes f_i
141 – 157	2
157 – 173	13
173 – 189	17
189 – 205	14
205 – 221	3
221 - 237	1

"-"Indica a menos de:

Debido a que la mediana es el valor por abajo del cual se encuentran el 50% de los datos y por arriba de él se encuentra también el 50% de los datos, entonces la mediana se debe de encontrar en la clase en la que la frecuencia relativa acumulada en una clase anterior sea menor de 0.5 (50%) y en ella la frecuencia relativa acumulada sea 0.5 o más. A esta clase se le llama clase mediana.

Para nuestro ejemplo, la clase mediana es la tercera. Una clase anterior, es decir la segunda clase, tiene una frecuencia relativa acumulada de 0.3 (menor a 0.5) y la tercera clase tiene una frecuencia relativa acumulada igual a 0.64 (mayor de 0.5).

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	Frecuencia relativa f_r	Frecuencia acumulada F_i	Frecuencia relativa acumulada Fr_i
141 – 157	2	149	0.04	2	0.04
157 – 173	13	165	0.26	15	0.30
173 – 189	18	181	0.34	32	0.64
189 – 205	14	197	0.28	46	0.92
205 – 221	3	213	0.06	49	0.98
221 - 237	1	229	0.02	50	1.00
Totales	50				

"-" Indica a menos de:

$$Mediana = Lim\ inf_{mediana} + \frac{(Lim\ sup - Lim\ inf)_{mediana}(0.5 - F_{r_{anterior}})}{f_{r_{mediana}}}$$

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	Frecuencia relativa f_r	Frecuencia acumulada F_i	Frecuencia relativa acumulada Fr_i
141 – 157	2	149	0.04	2	0.04
157 – 173	13	165	0.26	15	0.30
173 – 189	**18**	**181**	**0.34**	**32**	**0.64**
189 – 205	14	197	0.28	46	0.92
205 – 221	3	213	0.06	49	0.98
221 - 237	1	229	0.02	50	1.00
Totales	50				

"-" Indica a menos de:

$$Mediana = 173 + \frac{(189 - 173)\ \ (0.5 - 0.3)}{0.34} = 182.4$$

¿Qué es la Moda?

La moda es el valor más frecuente en una serie de datos. Por ejemplo, para los siguientes datos, la moda es 15, porque es el valor que se repite más.

4 6 9 11 12 13 15 15

¿En una serie de datos puede haber más de una moda?

Si. Sí se tiene dos o más valores con la misma frecuencia máxima, la distribución puede ser multimodal.

La siguiente serie de datos tiene dos modas, ya que el 11 y el 15, se repiten 2 veces, entonces se dice que la distribución de los datos es bimodal.

4 6 9 11 11 12 13 15 15

La siguiente serie de datos es trimodal, ya que el 4, el 11 y el 15 se repiten 3 veces.

4 4 4 6 9 11 11 11 12 13 15 15 15

¿En una serie de datos puede no existir la moda?

Sí. Si no hay un valor que se repita más veces que los otros, no existe la moda.

La siguiente serie de datos no tienen moda, porque no hay ningún dato que se repita más que otro. Todos tienen frecuencia 1

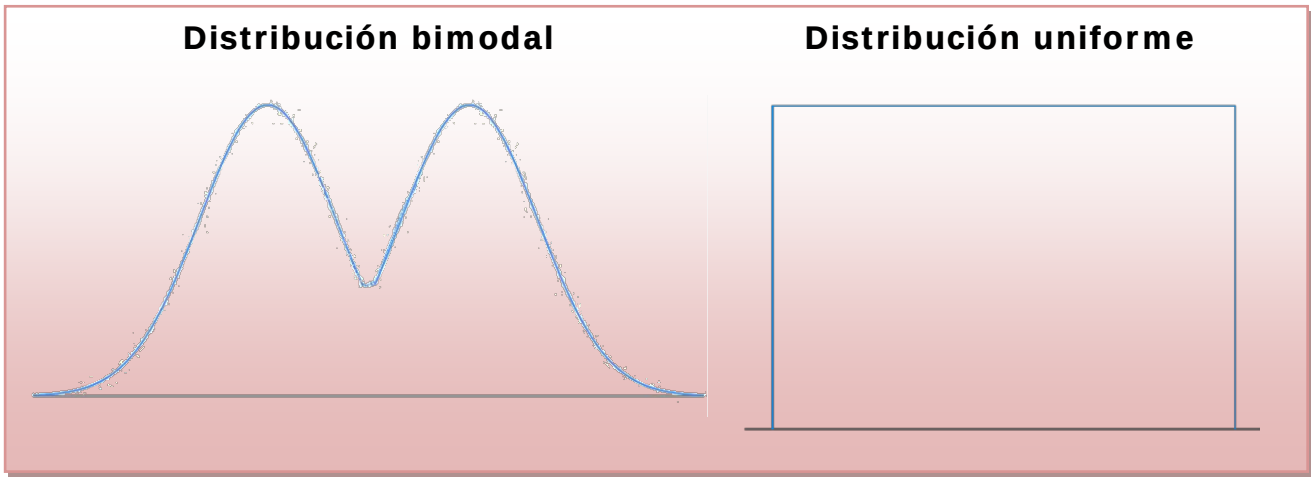
4 7 12 15 10 6 8

La siguiente serie de datos no tiene moda porque no hay ningún dato que se repita más que otro, todos tienen frecuencia 3.

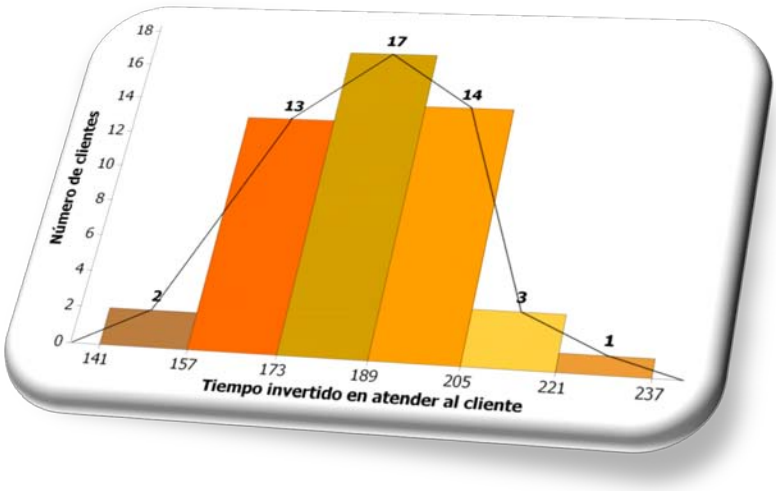
5 5 5 6 6 6 10 10 10

¿En una gráfica como se distingue la moda?

Cómo es el valor que se repite con mayor frecuencia la moda será el valor más alto o el pico de la distribución.



¿Se puede calcular la moda a partir de los datos agrupados en una tabla de frecuencias?

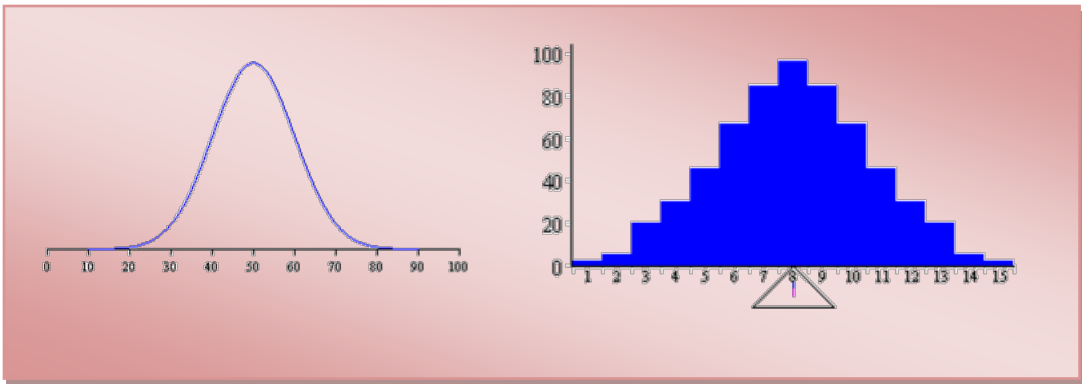


Resulta suficiente definir la clase modal, es decir la clase de mayor frecuencia (el pico de la distribución). Si se quiere establecer un valor, la moda será igual a la marca de clase de la clase modal.

Para nuestro ejemplo, la clase modal es la tercera. Entonces reportamos que la clase modal es de 173 a menos de 189 segundos y la moda es igual a 181.

¿En una serie de datos pueden ser iguales la media, la mediana y la moda?

Si, cuando la distribución es en forma de campana, lo que en estadística se conoce como distribución normal, coinciden los valores de la media, mediana y la moda. En la distribución que se muestra enseguida, la media, la mediana y la moda son iguales y tienen un valor de 50.

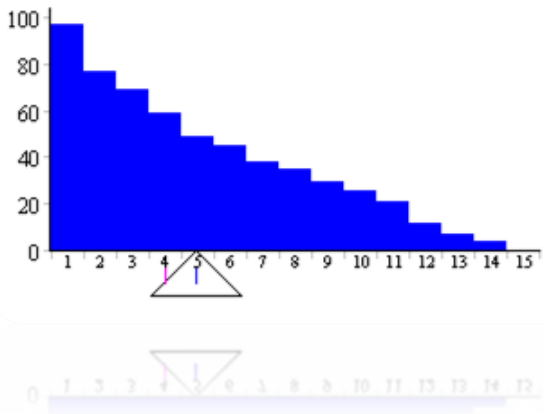


Sí la distribución no es simétrica ¿Cuál es la posición de la media, mediana y moda?

Si la distribución es simétrica coinciden los valores de la media y de la mediana. La moda puede o no existir.

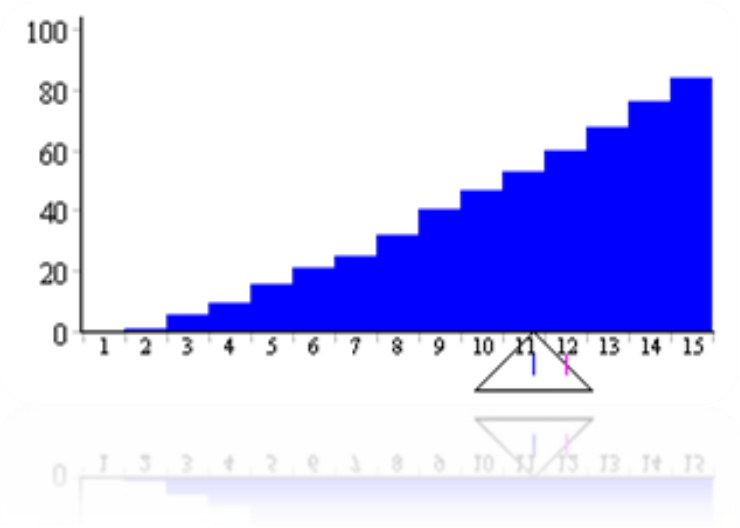
Para las distribuciones con sesgo a la derecha

$$\bar{x} > Mediana > Moda$$



Para las distribuciones con sesgo a la izquierda

$$Moda > Mediana > \bar{x}$$



¿Cuáles son las ventajas y las desventajas de cada una de las medidas de tendencia central revisadas?

La media es la más usada de las medidas de tendencia central, sus principales ventajas es que es muy fácil de calcular, para determinar su valor se toman en cuenta todos los datos y es muy importante en inferencia estadística por las propiedades de su distribución muestral. Su principal desventaja es que debido a que es el punto de equilibrio de la distribución su valor se ve muy afectado por datos extremos, por lo que si la distribución es muy sesgada no es conveniente utilizarla.

Suponga que tenemos los siguientes datos:

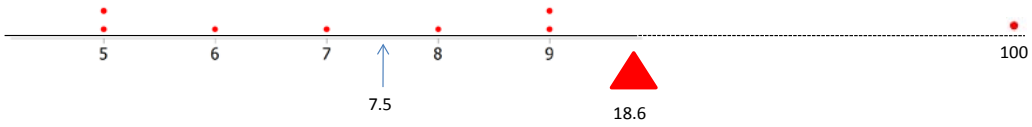
5 5 6 7 8 9 9

La media y la mediana tienen el mismo valor y éste es 7.

Suponga que en la serie introducimos un nuevo dato, extremo, 100. Ahora la serie es:

5 5 6 7 8 9 9 100

Entonces la media es igual a 18.6, mientras que la mediana es igual a 7.5



La mediana representa mejor a los datos ya que está muy cerca de siete de las observaciones, mientras que la media se ve muy afectada por el valor extremo.

La principal ventaja de la mediana es que no se ve afectada por valores extremos y por lo tanto si la distribución es muy asimétrica o sesgada es una medida que representa mejor a los datos. Su desventaja más importante es que su valor se determina con un solo dato, el dato central de la serie ordenada.

La moda por lo general no se usa debido a que no tiene un valor único ó puede ser que no exista. Para datos agrupados en tabla de frecuencia, la moda tiene importancia porque en éste caso si hay un valor único.

¿Cuáles son las medidas de variabilidad?

Las medidas de variabilidad son el rango o amplitud, la varianza, la desviación estándar y el coeficiente de variación.

¿Cómo se calculan estas medidas?

Para ejemplificar el cálculo y reafirmar el concepto de variabilidad, supongamos que tenemos dos muestras de tres datos cada una:

Muestra 1			**Muestra 2**		
17	**18**	**19**	**15**	**16**	**23**

El resumen de los datos de cada muestra seria:

Muestra 1	Muestra 2
$n = 3; \quad \bar{x} = 18$	$n = 3; \quad \bar{x} = 18$

De tal forma, que si nos referimos a una muestra de tamaño 3 y media 18, no sabemos si hablamos de la muestra 1 o de la muestra 2. Es decir, la media no es una medida suficiente para poder distinguir una muestra de la otra. Es necesario, entonces construir otra medida que permita diferenciarlas.

Si inspeccionamos las muestras vemos que la primera varía de 17 a 19, mientras que la segunda de 15 a 23. Esta amplitud o rango es la primera medida de variabilidad.

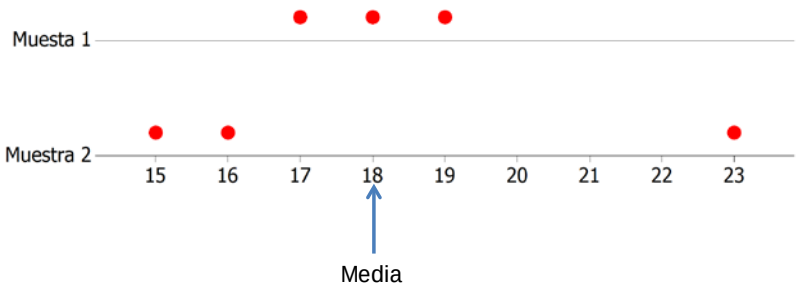
Muestra 1	**Muestra 2**
$\mathrm{R} = \mathrm{y}_{\max} - \mathrm{y}_{\min}$	$\mathrm{R} = \mathrm{y}_{\max} - \mathrm{y}_{\min}$
$R = 19 - 17 = 2$	$R = 23 - 15 = 8$

Sin embargo, es una medida que solo toma en cuenta dos datos, el valor máximo y el mínimo y por tanto se ve afectada por los valores extremos. Es una medida que proporciona la variabilidad en forma muy burda.

Si observamos las muestras vemos que la muestra 1, tiene sus valores más agrupados alrededor de la media, 18, mientras que los valores de la muestra 2, están más alejados de ella. Entonces, se hace necesaria una medida que valore la variabilidad o distancia promedio de los datos con respecto a su media.

La idea sería obtener las distancias de cada dato con respecto a su media, y a partir de estas obtener la distancia promedio. Note que una medida construida de esta manera, toma en cuenta todos los datos.

Muestra 1		Muestra 2	
x_i	distancia $(x_i - \bar{x})$	x_i	distancia $(x_i - \bar{x})$
17	17 – 18 = -1	15	15 – 18 = -3
18	18 – 18 = 0	16	16 – 18 = -2
19	19 – 18 = 1	23	23 – 18 = 5
	$\sum d = 0$		$\sum d = 0$
	$d_{prom} = 0$		$d_{prom} = 0$



¿Por qué en ambos casos la suma de las distancias resulta igual a cero? Sabemos que si obtenemos las distancias de cada dato con respecto a su media, la suma de las distancias de los datos mayores a la media, es igual a la suma de las distancias de los datos menores a ella, y si a las distancias de los datos menores a la media les asignamos signos negativos y a las mayores signos positivos, la suma siempre resultara cero y por esta vía resulta imposible obtener la distancia o variabilidad promedio.

Una manera de resolver este problema es elevar al cuadrado las distancias, con lo cual se resolvería el problema de los signos. Obtendríamos la distancia cuadrática promedio, lo que nos daría la medida que buscamos elevada al cuadrado, la cual se conoce como varianza y se representa con s^2. Una vez calculada la varianza, obtenemos su raíz cuadrada y con esto la medida buscada, que se conoce como desviación estándar.

Muestra 1

x_i	distancia $(x_i - \bar{x})$	$(x_i - \bar{x})^2$
17	17 – 18 = -1	1
18	18 – 18 = 0	0
19	19 – 18 = 1	1
		2

$$s^2 = \frac{2}{2} = 1\ unidad^2$$

$$s = \sqrt{1} = 1\ unidad$$

Muestra 2

x_i	distancia $(x_i - \bar{x})$	$(x_i - \bar{x})^2$
15	15 – 18 = -3	9
16	16 – 18 = -2	4
23	23 – 18 = 5	25
		38

$$s^2 = \frac{38}{2} = 19\ unidades^2$$

$$s = \sqrt{19} = 4.36\ unidades$$

El resumen de las muestras es el siguiente:

Muestra 1

$n = 3; \quad \bar{x} = 18; \quad s = 1$

Muestra 2

$n = 3; \quad \bar{x} = 18; \quad s = 4.36$

Ahora si es posible distinguir con este resumen una muestra de la otra. Es necesario, entonces una medida de tendencia central y una de variabilidad.

¿Qué significado tiene la varianza y la desviación estándar?

La varianza no tiene significado. Se expresa en las unidades de los datos elevadas al cuadrado. Sí estas estudiando el número de clientes que llegan a un autolavado, la varianza tiene como unidades clientes^2, lo cual no tiene ningún significado. La desviación estándar tiene las mismas unidades que los datos y nos proporciona la variabilidad promedio de los datos con respecto a su media.

La muestra 1, tiene tres datos, su promedio es 18. No significa que todos los datos sean 18, unos serán mayores y otros serán menores. ¿Qué tanto se alejan los datos individuales con respecto a 18? Unos se alejan más, otros se alejan menos, pero en promedio se alejan 1 unidad, que es su desviación estándar.

La muestra 2, tiene tres datos, su promedio es 18. No significa que todos los datos sean 18, unos serán mayores y otros serán menores. ¿Qué tanto se alejan los datos individuales con respecto a 18? Unos se alejan más, otros se alejan menos, pero en promedio se alejan 4.36 unidades, que es su desviación estándar.

Tal vez te preguntes por qué el denominador de s^2 es $n-1$, en lugar de n, si estamos buscando una variabilidad promedio. Desde el punto de vista de la Estadística Descriptiva es irrelevante dividir entre uno u otro. Desde el punto de vista de la Inferencia Estadística si es importante la selección del divisor y se divide entre N si se trata de la variabilidad de la población y entre $n-1$ si es la variabilidad de una muestra.

Entonces, la varianza y la desviación estándar tienen las siguientes expresiones:

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

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

¿Qué es el coeficiente de variación?

El coeficiente de variación es una medida de variabilidad relativa de una serie de datos y se obtiene dividiendo la desviación estándar de los datos entre su media.

$$c.v. = \frac{s}{\bar{x}}$$

Debido a que la desviación estándar y la media tienen las mismas unidades, el coeficiente de variación se expresa por lo general en proporción o en porcentaje y por lo tanto, se utiliza para comparar la variabilidad de dos o más series de datos.

Muestra 1

$$c.v. = \frac{1}{18} * 100 = 5.56\%$$

Muestra 2

$$c.v. = \frac{4.36}{18} * 100 = 24.22$$

¿Se pueden calcular la varianza y la desviación estándar a partir de los datos agrupados en una tabla de frecuencias?

Si es posible calcular éstas medidas a partir de una Tabla de distribución de frecuencias. Utilizando la Tabla de distribución de frecuencias que nos indica el tiempo que tardaron algunos clientes en una caja bancaria, determinemos la varianza y la desviación estándar.

Tiempo invertido en atender al cliente	No de clientes f_i
141 – 157	2
157 – 173	13
173 – 189	17
189 – 205	14
205 – 221	3
221 - 237	1

"-"Indica a menos de:

Lo primero que tenemos que calcular es la media como lo hicimos anteriormente.

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	$x_i f_i$
141 – 157	2	149	298
157 – 173	13	165	2145
173 – 189	18	181	3077
189 – 205	14	197	2758
205 – 221	3	213	639
221 - 237	1	229	229
Totales	50		9146

"-" Indica a menos de:

Calculemos la distancia de cada dato con respecto a la media. Recuerda que suponemos que lo valores de los datos son las marcas de clase. Incluiremos una columna donde se registren estas distancias

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	$x_i f_i$	$(x_i - \bar{x})$
141 – 157	2	149	298	-33.92
157 – 173	13	165	2145	-17.92
173 – 189	18	181	3077	-1.92
189 – 205	14	197	2758	14.08
205 – 221	3	213	639	30.08
221 - 237	1	229	229	46.08
Totales	50		9146	

"-" Indica a menos de:

Añadimos una columna donde se anoten las distancias al cuadrado

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	$x_i f_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
141 – 157	2	149	298	-33.92	1150.57
157 – 173	13	165	2145	-17.92	321.13
173 – 189	18	181	3077	-1.92	3.69
189 – 205	14	197	2758	14.08	198.25
205 – 221	3	213	639	30.08	904.81
221 - 237	1	229	229	46.08	2123.37
Totales	50		9146		

"-" Indica a menos de:

Finalmente estas distancias cuadráticas corresponden a la distancia al cuadrado de cada dato con respecto a su media, pero recuerda que suponemos que cada dato o marca de clase se repite un número igual a su frecuencia, por lo que tenemos que obtener en una columna los productos.

Tiempo invertido en atender al cliente	No. de clientes f_i	Marca de clase x_i	$x_i f_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
141 – 157	2	149	298	-33.92	1150.57	2301.13
157 – 173	13	165	2145	-17.92	321.13	4174.64
173 – 189	18	181	3077	-1.92	3.69	66.36
189 – 205	14	197	2758	14.08	198.25	2775.45
205 – 221	3	213	639	30.08	904.81	2714.42
221 - 237	1	229	229	46.08	2123.37	2123.37
Totales	50		9146			14155.37

"-" Indica a menos de:

Entonces la varianza es:

$$s^2 = \frac{14155.37}{49} = 288.89$$

La desviación estándar es:

$$s = \sqrt{288.89} = 16.99$$

El coeficiente de variación es:

$$c.v. = \left(\frac{288.89}{16.99}\right) 100 = 9.3\%$$

La expresión para la varianza y la desviación estándar a partir de datos agrupados en tablas de frecuencias viene dada por

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{n-1}$$

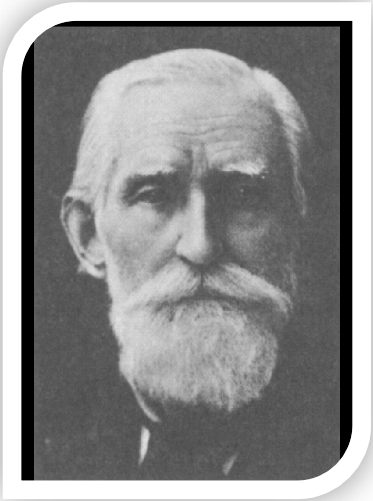
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i}{n-1}}$$

Ya que la desviación estándar es una medida de variabilidad ¿Se puede usar para determinar cuantos datos se encuentran en diferentes intervalos alrededor de la media?

Si, la desviación estándar nos permite determinar, con un buen grado de precisión, donde están localizados los valores de una distribución de frecuencias con relación a la media.

Podemos hacer esto de acuerdo con un teorema establecido por el matemático ruso, P.L. Chebyshev, el cual establece que independientemente de la forma de la distribución, la proporción de valores que se encuentran a *k* desviaciones estándares de la media es al menos $1-\frac{1}{k^2}$, siendo *k* cualquier número mayor que 1.

Suponga que las calificaciones del primer examen parcial de Estadística descriptiva del grupo 1304 compuesto por 50 alumnos, de la carrera de Lic. en Administración, obtuvieron un promedio de 70 con una desviación estándar de 5. ¿Cuántos alumnos tuvieron calificaciones entre 60 y 80?



Al aplicar el teorema de Chebyshev, observamos que cuando menos 75% de los 50 alumnos, es decir 38, deben haber obtenido calificaciones entre 60 y 80.

Si la distribución de los datos es simétrica con forma de campana, lo que conocemos en estadística como distribución normal, se puede aplicar una regla empírica, para determinar con más precisión el porcentaje de observaciones que caen dentro de determinada cantidad de desviaciones estándar respecto a la media aritmética. En este caso podemos decir que:

Para datos con distribución normal:

- Aproximadamente 68% de los valores caen dentro de ± 1 desviación estándar a partir de la media.
- Aproximadamente 95% de los valores caen dentro de ± 2 desviaciones estándar a partir de la media.
- Aproximadamente 99% de los valores caen dentro de ± 3 desviaciones estándar a partir de la media.

Suponga que en una línea de producción, se llenan automáticamente bolsas de plástico con detergente en polvo. Con frecuencia, los pesos de llenado tienen una distribución en forma de campana. Si el peso promedio de llenado es de 1 kilogramo y la desviación estándar es de 5 gramos, se puede aplicar la regla empírica para hacer las siguientes conclusiones:

- Aproximadamente el 68% de las bolsas llenas tienen entre 995 y 1005 gramos de detergente en polvo.

- Aproximadamente el 95% de las bolsas llenas tienen entre 990 y 1010 gramos de detergente en polvo.

- Aproximadamente el 99% de las bolsas llenas tienen entre 985 y 1015 gramos de detergente en polvo.

¿Cuáles son las ventajas y desventajas de cada una de las medidas de variabilidad?

El rango es muy fácil de calcular pero su desventaja es que solo toma en cuenta dos valores, el valor menor y el valor mayor. La desviación estándar es fácil de calcular, toma en cuenta todos los datos y es una medida importante en el estudio de la inferencia estadística. Su principal desventaja es que es sensible a los valores extremos. El coeficiente de variación es muy útil cuando se quiere comparar la variabilidad de dos o más muestras o poblaciones, debido a que su valor es independiente de las unidades de medición.

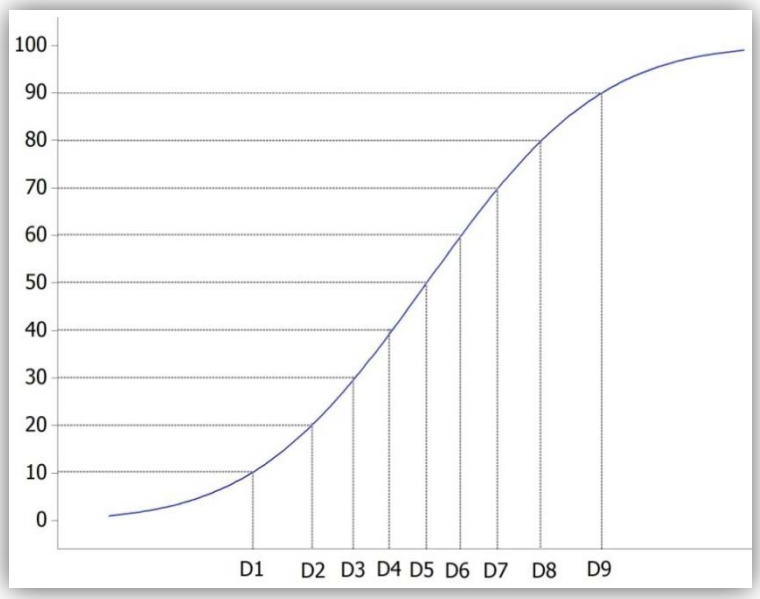
¿Existe otra forma de describir la variabilidad de un conjunto de datos?

Si. Una forma consiste en determinar la posición de los valores que dividen una serie de datos en partes iguales. A estas medidas se les conoce por lo general como medidas de posición.

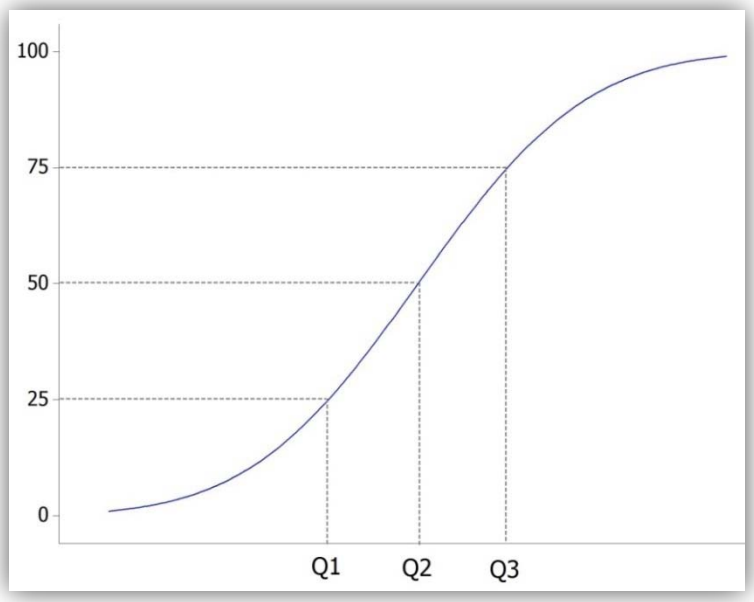
Si un conjunto de datos lo dividimos en 100 partes iguales, a cada parte se le conoce como percentil. Si lo dividimos en 10 partes iguales, a cada parte se le conoce como decil y sí lo dividimos en 4 partes iguales a cada parte se le conoce como cuartil.

¿Qué indica la posición?

Una medida de posición indica el porcentaje de datos que son menores a ella. Por ejemplo, el 10% de los datos son menores que el decil 1, el 30% de los datos son menores que el decil 3, etc.



Sí dividimos el conjunto de datos en 4 partes, por ejemplo, el 25% de los datos son menores que el cuartil 1, el 50% de los datos son menores que el cuartil 2 y el 75% de los datos son menores que el cuartil 3.



¿Cómo se determina el valor de una medida de posición?

Suponga que tenemos los siguientes datos

$x_1 = 13$, $x_2 = 15$, $x_3 = 9$, $x_4 = 6$, $x_5 = 4$, $x_6 = 12$, $x_7 = 11$. Lo primero que tenemos que hacer es ordenarlos,

4 6 9 11 12 13 15

Para determinar el valor de una medida de posición pensemos, en la mediana, que es el valor que divide el número de observaciones en dos partes iguales. ¿Qué posición le correspondería a la mediana?, su posición se encuentra en (n+1)/2; es decir en (7+1)/2 = 4. La mediana es el valor que le corresponde al cuarto dato, es decir 11.

Si el número de datos fuera par, por ejemplo

4 6 9 11 12 13 15 15

La expresión para localizar la mediana sigue siendo válida. La mediana se encuentra en (n+1)/2; es decir en (8+1)/2 = 4.5. La mediana es el valor que corresponde a la posición 4.5, es decir el promedio de los valores correspondientes a la posición 4 y a la posición 5. La mediana es (11+12)/2= 11.5.

Generalizando, obtenemos las expresiones para localizar los percentiles, deciles y cuartiles, respectivamente:

$$L_P = (n+1)\frac{P}{100}$$

$$L_Q = (n+1)\frac{Q}{4}$$

$$L_D = (n+1)\frac{D}{10}$$

30	55	38	34	30	24	45	28	51	51
22	47	42	3	39	65	26	37	44	33
62	21	33	49	57	47	19	43	27	51
21	14	25	36	61	46	48	35	40	36
67	56	45	35	54	49	36	34	27	54

Consideremos ahora los siguientes datos, que se refieren al tiempo de entrega de la comida a domicilio del Restaurante The Ramen, especialista en comida japonesa ubicado en Cuautitlán Izcalli.

Definamos los valores del cuartil 1, 2 y 3.

Lo primero que tenemos que hacer es ordenar los datos e indicar su posición.

Número	Tiempo	Número	Tiempo	Número	Tiempo	Número	Tiempo	Número	Tiempo
1	21	11	28	21	36	31	45	41	51
2	21	12	29	22	36	32	45	42	54
3	22	13	30	23	36	33	46	43	54
4	23	14	30	24	37	34	47	44	55
5	24	15	33	25	38	35	47	45	56
6	24	16	33	26	39	36	48	46	57
7	25	17	34	27	40	37	49	47	61
8	26	18	34	28	42	38	49	48	62
9	27	19	35	29	43	39	51	49	65
10	27	20	35	30	44	40	51	50	67

Entonces:

$$L_{Q1} = (50 + 1)\frac{1}{4} = 12.75$$

El cuartil 1, se encuentra en la posición 12.75. El dato que ocupa la posición 12 es 29 y el que ocupa la posición 13 es 30. Es decir, para la diferencia de una unidad de posición hay una diferencia de un minuto (30–29), entonces mediante la siguiente regla de tres simple, determinamos, los minutos que le corresponden a 0.75.

(13 - 12) unidades de posición ---------- (30 – 29) minutos
0.75 unidades de posición ---------- x

$$x = \frac{0.75\ (30 - 29)}{(13 - 12)}$$

A lo posición 12, le corresponden 29 minutos y a la fracción de 0.75 unidades entre la posición 12 y 13, le corresponde 0.75 minutos; luego a la posición 12.75 le corresponderá 29 + 0.75 = 29.75 minutos que es la posición del cuartil 1, Q_1. Es decir, que el 25% de los pedidos se entrega en menos de 29.75 minutos y el 75% de los pedidos tarda más de 29.75 minutos en entregarse.

El cuartil 2, Q_2, se encuentra en la posición

$$L_{Q2} = (50 + 1)\frac{2}{4} = 25.5$$

A la posición 25 le corresponde el valor de 38 minutos y a la posición 26 le corresponde 39 minutos. Es decir, para una unidad de posición de diferencia, hay una diferencia de un minuto (39-38), y la fracción de 0.5 unidades de posición entre la posición 25 y 26, le corresponderá 0.5 minutos. Entonces, la posición 25.5 corresponde a 38 + 0.5 = 38.5, que es el valor del cuartil 2, Q_2. Es decir, el 50% de los tiempos de entrega son menores a 38.5 minutos, lo cual corresponde a la mediana.

El cuartil 3, Q_3, se encuentra en la posición

$$L_{Q3} = (50 + 1)\frac{3}{4} = 38.25$$

El dato que ocupa la posición 38, es 49 y el que ocupa la posición 39 es 51. Es decir, ahora a una unidad de posición de diferencia le corresponde una diferencia de dos minutos (51 - 49), y a la fracción de 0.25 unidades entre la posición 38 y la 39, le corresponde (0.25)(2) = 0.5 minutos. Entonces, a la posición 38.25, le corresponde 49 + 0.5 = 49.5, que es el valor del cuartil 3, Q_3. Es decir, el 75% de los pedidos se entregan en menos de 49.5 minutos y solo el 25% de los tiempos de entrega exceden este valor.

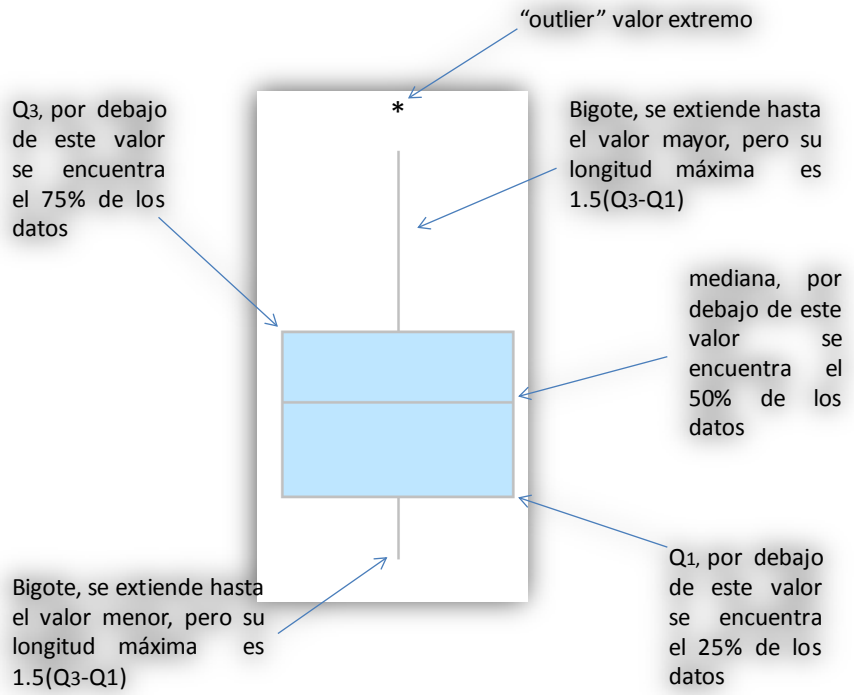
Como ejercicio calcule los percentiles 15, 40 y 85 e interprételos.

¿Existe algún tipo de gráfica que muestre las medidas de posición?

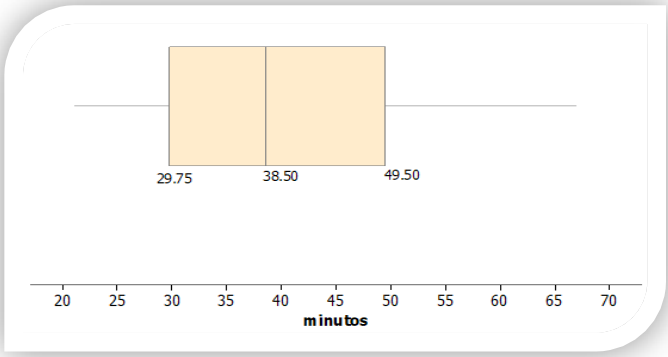
Si, únicamente para los cuartiles. Una gráfica de caja es la representación gráfica de la distribución de los datos basada en los cuartiles.

El primer paso consiste en elaborar una escala adecuada ya sea vertical u horizontal. Enseguida se dibuja una caja que inicie en el cuartil 1, Q_1 y termine en el cuartil 3, Q_3. Dentro de la caja con una línea se indica el cuartil 2, Q_2. El tamaño de la caja que es igual a (Q_3 - Q_1), se conoce como rango intercuartílico. A partir de las tapas de la caja se trazan líneas, conocidas como bigotes, de longitud máxima a 1.5 (Q_3 - Q_1). Sí se encuentra el valor máximo o mínimo antes de esta longitud, el bigote termina ahí. Esta es la razón por la que algunas veces los bigotes no son de igual tamaño. Si por el contrario, una vez trazada la línea con la longitud máxima no se incluyeran algunos datos, éstos se señalan con asteriscos, para indicar que son valores extremos.

La siguiente figura muestra los componentes de una gráfica de caja.



La gráfica de caja correspondiente a los tiempos de entrega de la comida japonesa, se muestra enseguida:



El hecho de que el bigote del lado derecho sea más largo que el izquierdo, indica que la distribución de los tiempos de entrega presenta sesgo a la derecha o positivo.

¿Y en cuanto a la forma de la distribución existen medidas para describirla?

La forma de la distribución se puede describir mediante el sesgo.

Una distribución puede tener una de cuatro formas, simétrica, sesgada a la derecha, sesgada a la izquierda o multimodal.

Si se va a realizar el cálculo manualmente la forma más sencilla de medir el sesgo es mediante la fórmula del coeficiente de sesgo de Pearson:

$$Sesgo = \frac{3(\bar{x} - Mediana)}{s}$$

El valor del coeficiente de sesgo basado en las desviaciones cúbicas de los datos con respecto a su media, viene dado por la siguiente fórmula:

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

El término $\left(\frac{x_i - \bar{x}}{s}\right)$ se refiere a la estandarización de los valores x_i. Los valores estandarizados son independientes de las unidades empleadas.

Insistimos que al igual que las otras medidas descriptivas el valor del sesgo se puede determinar utilizando software apropiado, sin embargo, realicemos los cálculos en forma manual para ejemplificar el uso de la fórmula

Suponga que se tiene el siguiente conjunto de datos 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4 y 5. Se desea calcular el sesgo

Primero obtenemos la media

$$\bar{x} = \frac{1 + 1 + \cdots + 5}{15} = 2.333$$

Ahora calculemos la desviación estándar

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	-1.33	1.78
1	-1.33	1.78
1	-1.33	1.78
1	-1.33	1.78
1	-1.33	1.78
2	-0.33	0.11
2	-0.33	0.11
2	-0.33	0.11
2	-0.33	0.11
3	0.67	0.44
3	0.67	0.44
3	0.67	0.44
4	1.67	2.78
4	1.67	2.78
5	2.67	7.11
		23.33

$$s = \sqrt{\frac{23.33}{14}} = 1.29$$

Agregamos otra columna para obtener el sesgo

	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$\left(\frac{x_i - \overline{x}}{s}\right)$	$\left(\frac{x_i - \overline{x}}{s}\right)^3$
1	-1.33	1.78	-1.03	-1.10
1	-1.33	1.78	-1.03	-1.10
1	-1.33	1.78	-1.03	-1.10
1	-1.33	1.78	-1.03	-1.10
1	-1.33	1.78	-1.03	-1.10
2	-0.33	0.11	-0.26	-0.02
2	-0.33	0.11	-0.26	-0.02
2	-0.33	0.11	-0.26	-0.02
2	-0.33	0.11	-0.26	-0.02
3	0.67	0.44	0.52	0.14
3	0.67	0.44	0.52	0.14
3	0.67	0.44	0.52	0.14
4	1.67	2.78	1.29	2.15
4	1.67	2.78	1.29	2.15
5	2.67	7.11	2.07	8.82
		23.33		7.96

$$sesgo = \frac{15}{(15-1)(15-2)}(7.96) = 0.66$$

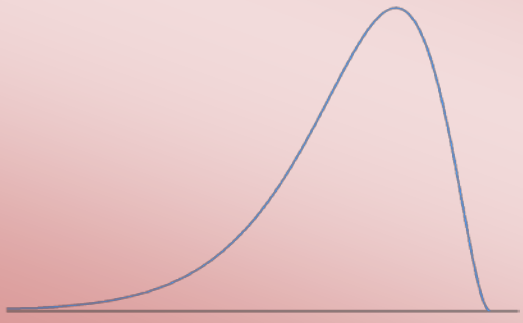
Obtengamos el sesgo utilizando la fórmula del coeficiente de sesgo de Pearson. Primero hay que calcular la mediana, ordenando los datos la mediana seria la posición 8.

$$Mediana = 2$$

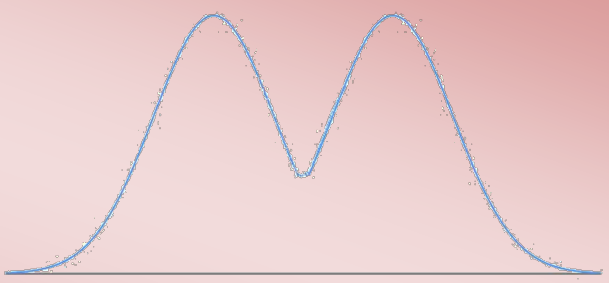
Número	
1	1
2	1
3	1
4	1
5	1
6	2
7	2
8	2
9	2
10	3
11	3
12	3
13	4
14	4
15	5

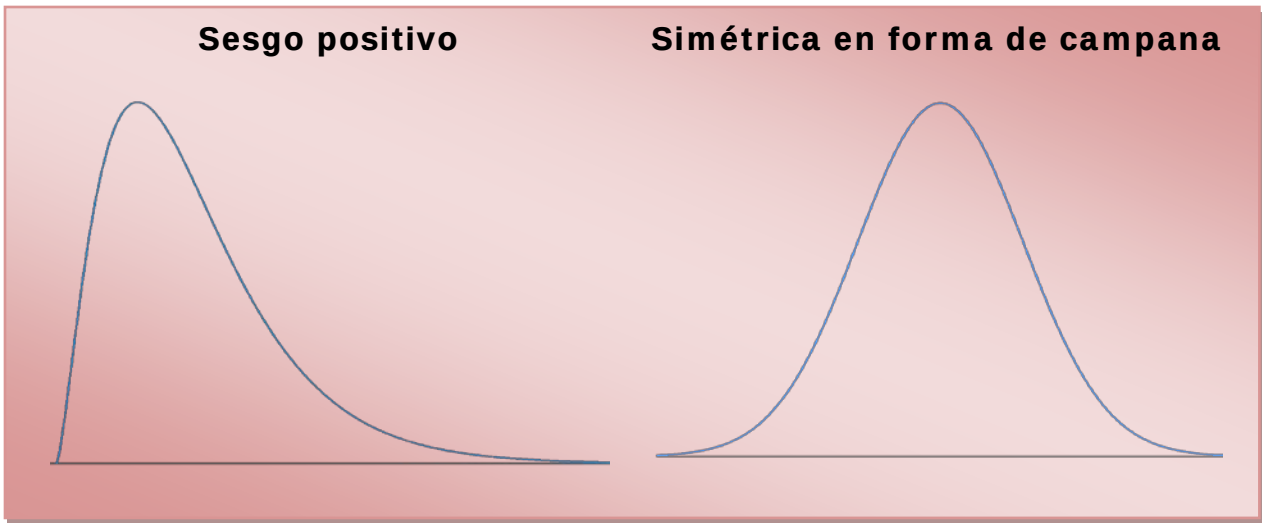
$$sesgo = \frac{3(2.33 - 2)}{1.29} = 0.77$$

Sesgo negativo.



Bimodal





¿Hay otras medidas de Tendencia Central?

Sí, entre ellas podemos mencionar a la media geométrica, media armónica, la media truncada y la media de Windsor.

¿Qué es la media geométrica?

La media geométrica de una serie de datos se define como la raíz n-ésima del producto de los datos. Los datos tienen que ser positivos

$$media\ geométrica = \sqrt[n]{x_1 x_2 x_3 \ldots . x_n}$$

Esta medida se utiliza cuando se quiere determinar el cambio promedio de tasas, razones, porcentajes o velocidades.

Suponga que una empresa ha aumentado su producción en un 25% en 2008 y en un 40% en 2009. ¿Cuál sería el aumento promedio de la producción en esto dos años? Un aumento del 25%, lo representamos como 1.25 y un aumento de 40%, como 1.4.

Tenemos:

$$media\ geométrica = \sqrt[2]{(1.25)(1.4)} = 1.323$$

Entonces el % de crecimiento promedio es de 32.3%

¿Qué es la media armónica?

Se define como el reciproco de la media aritmética de los recíprocos de las observaciones

$$media\ armónica = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

Se utiliza para procesar datos de razones que tienen dimensiones físicas, como por ejemplo, rendimiento del combustible en un automóvil medido en kilómetros por litro, velocidad promedio medida en kilómetros por hora, etc.

¿Qué es la media truncada?

En el caso de las distribuciones sesgadas, se ha mencionado que la media es muy sensible a los valores extremos y que una mejor medida descriptiva de la tendencia central de la distribución es la mediana. Sin embargo, el uso de la mediana tiene el inconveniente que solo toma en cuenta un valor, el dato central. Una medida que se propone que tome en cuenta en su cálculo un mayor número de datos, es la media truncada o podada. Para calcular esta medida se eliminan las colas de la serie de datos, es decir, se elimina un porcentaje de los datos extremos, menores y mayores. El porcentaje de datos a eliminar, puede ser hasta del 25% en cada extremo. En todo caso, la idea es eliminar los datos que afecten a la media.

Considere la siguiente serie de datos a fin de calcular y entender la media truncada.

Observamos que existen dos valores extremos 1150 y 1155, que van a afectar a la media. La media truncada elimina los valores extremos, con el fin de eliminar su influencia sobre la media y poder calcular una medida de tendencia central con la mayor cantidad de información posible.

Calculamos la mediana, la media y la media truncada eliminando el 5% en cada uno de los extremos. Se seleccionó un 5%, porque este es suficiente para eliminar los valores extremos.

20	28	21	27	18
23	18	22	23	23
17	21	26	20	20
17	16	19	21	16
17	15	22	17	20
10	14	13	22	20
1150	1155			

La media es igual a 90.3, que como se puede observar no es una buena representación de los datos.

Calculemos la media truncada. Se ordenan los datos y se elimina el 5% del total de ellos en cada extremo. En nuestro caso, se eliminan (32)(0.05)=1.6, eliminamos 2 datos en cada extremo y se calcula la media con los 28 datos restantes.

10	13	14	15	16
16	17	17	17	17
18	18	19	20	20
20	20	20	21	21
21	22	22	22	23
23	23	26	27	28
1150	1155			

Entonces la media truncada es igual 20.1, valor que representa bien a los datos.

Una media truncada aproximadamente igual a la media aritmética, indica poco sesgo en la distribución

¿Qué es la media de Windsor?

La media de Windsor es una variante de la media truncada. En esta medida se sustituyen los datos que se eliminan en el extremo inferior por el dato menor no eliminado y los datos que se eliminan en el extremo superior por el dato mayor no eliminado.

Para nuestro ejemplo, después de eliminar los dos datos menores, el primer dato es 14, entonces los datos eliminados los sustituimos por 14. Lo mismo hacemos en el extremo superior, después de eliminar los dos datos mayores, el dato mayor será 28, entonces sustituimos los dos eliminados por 28, obteniendo la serie de datos siguiente:

La media de Windsor se calcula con estos datos y resulta ser igual a 20.21

14	**14**	14	15	16
16	17	17	17	17
18	18	19	20	20
20	20	20	21	21
21	22	22	22	23
23	23	26	27	28
28	**28**			