Polígono de frecuencias

Es una alternativa al histograma*.*

<u>Construcción</u>

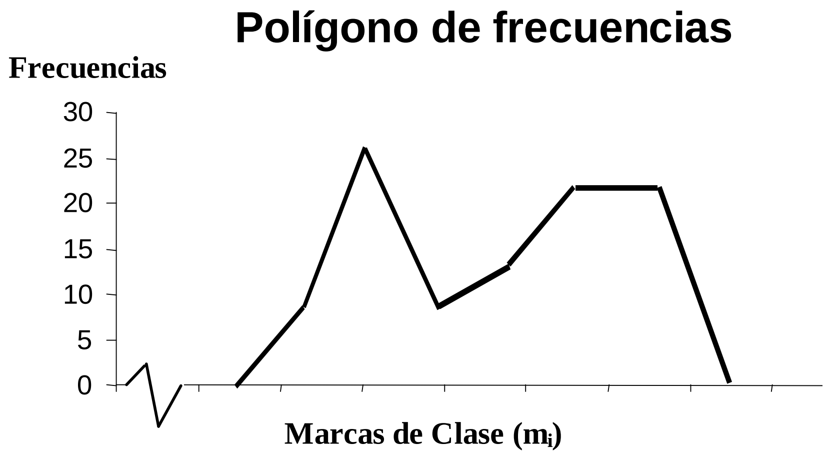
Paso 1. En el eje horizontal se escriben las *marcas de clase* de cada intervalo.

Paso 2. Para cada una de las m_i se colocan las alturas en el eje vertical, las cuales vienen dadas por las frecuencias respectivas (absolutas simple o relativas).

Paso 3. Luego, se marcan los puntos (m_i , fr_i *o* m_i , f_i) y se une con rectas en el plano cartesiano.

Paso 4. <u>Para cerrar la curva resultante con el eje de las abscisas</u>, se crean dos puntos medios ficticios, uno anterior al de la primera clase y otro posterior al de la última clase cada uno con frecuencia igual a cero. De esta manera se obtiene el polígono de frecuencias:

Representación gráfica



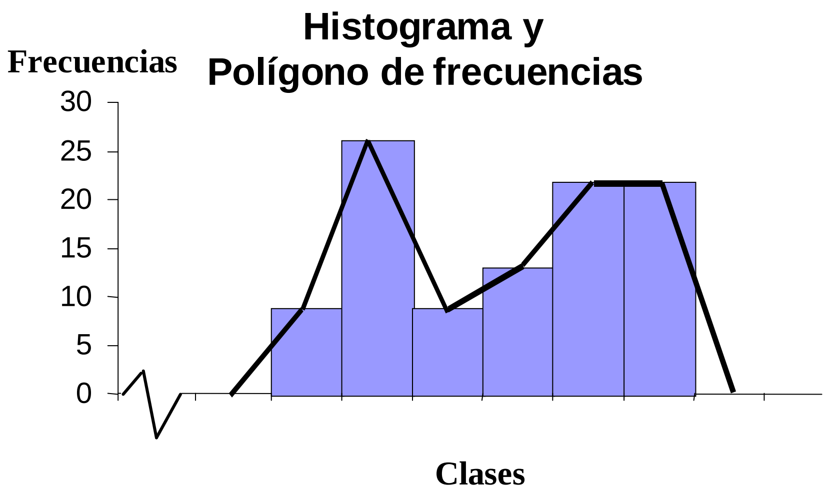


Diagrama de líneas de frecuencias

Es el equivalente al histograma en una distribución de frecuencias cuyas clases son valores individuales de la variable.

Construcción

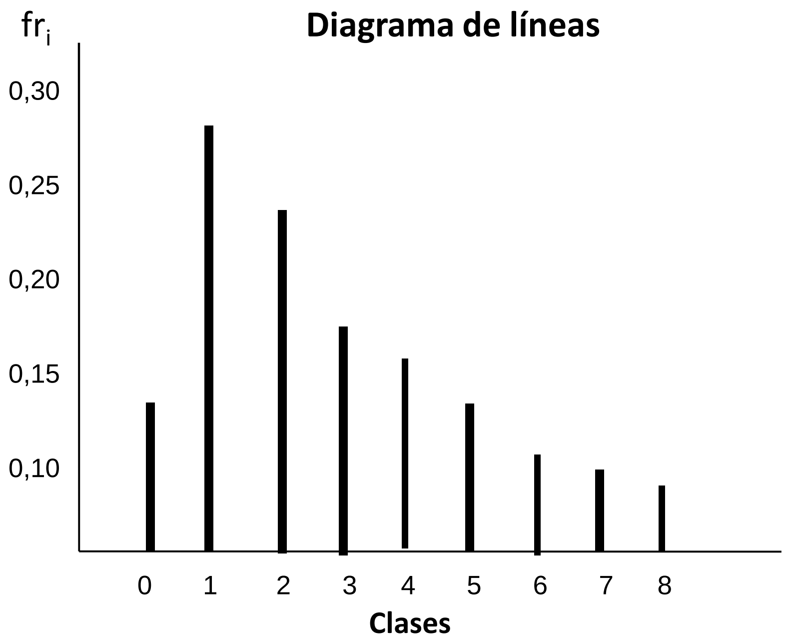
Paso 1: En el eje horizontal, marque sucesivamente las clases.

Paso 2: En el eje vertical, marque, en la escala, los valores de las frecuencias absolutas o frecuencias relativas de las clases.

Paso 3: Para la primera clase, trace una línea vertical cuya altura es la frecuencia absoluta simple (o relativa) de esa clase;

Paso 4: Repita el procedimiento para las demás clases.

Representación gráfica



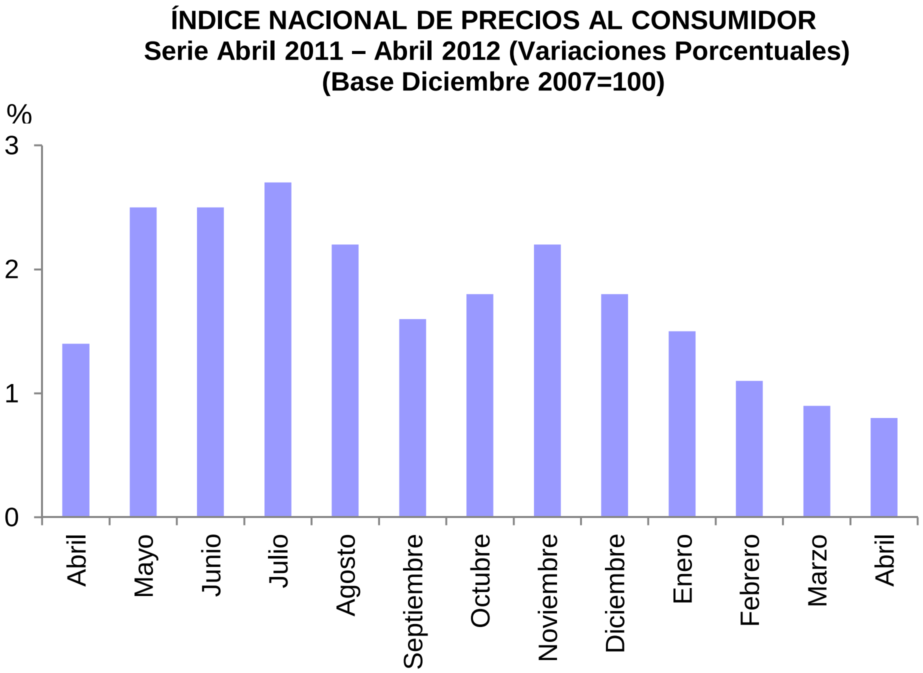
Ejercicio:

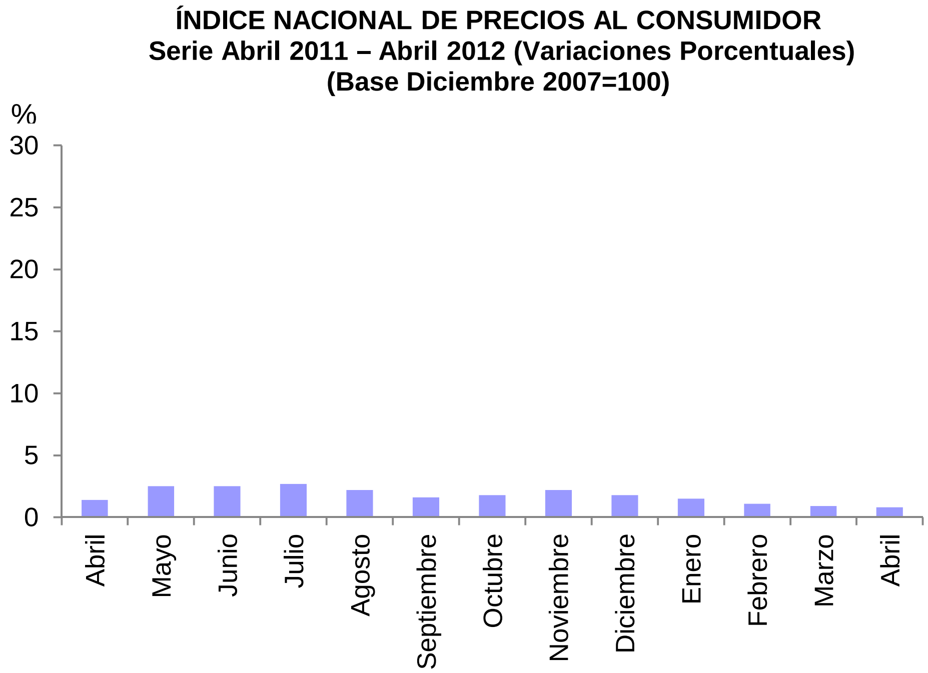
Construir un diagrama de líneas para la distribución de frecuencias de la variable número de hermanos.

Gráficos engañosos

¡Cuidado! Cuando se observa un gráfico, particularmente como parte de un anuncio, sea cauteloso. Fíjese en las escalas utilizadas en los ejes vertical y horizontal. Se puede distorsionar la verdad con las técnicas estadísticas, tal como se muestra a continuación:

Gráficos engañosos



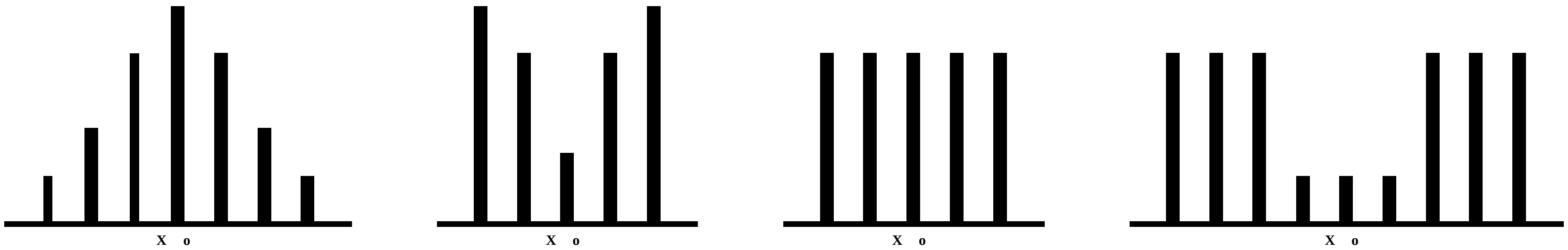


Distribución simétrica

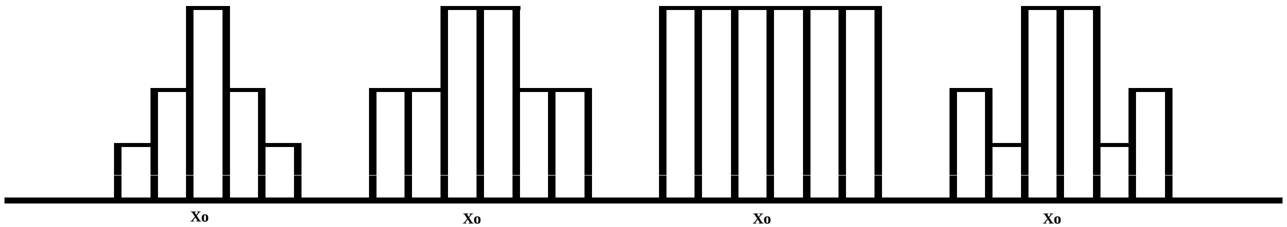
Una distribución de frecuencias es *simétrica* (o con sesgo cero) con respecto al valor central de la distribución, digamos x_0, cuando el gráfico a la izquierda de x_0 es el "espejo" de la derecha. En otras palabras, si a la izquierda y a la derecha de x_0 existe la misma cantidad de datos la distribución será simétrica.

Descripción de la forma en que se distribuyen los datos

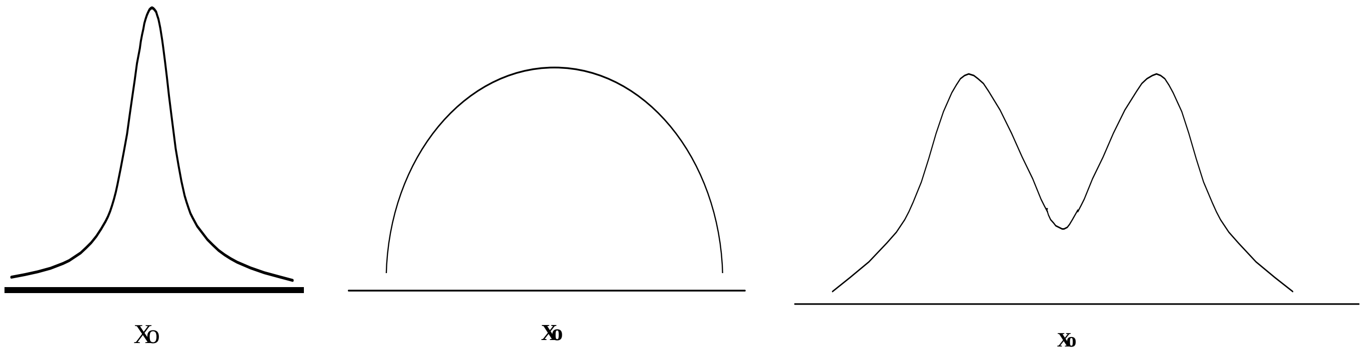
DIAGRAMAS DE LÍNEAS



HISTOGRAMAS



POLÍGONOS

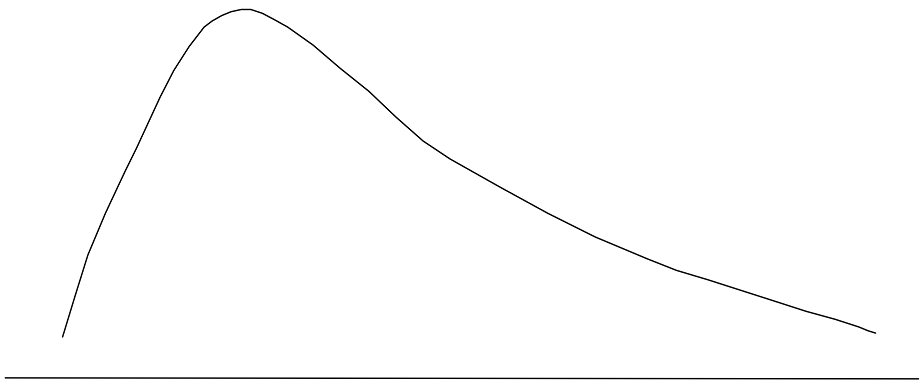


Distribución asimétrica

Si una distribución no es simétrica, se dice que es asimétrica (o sesgada). Existen dos casos de asimetría:

Asimetría Positiva o por la derecha

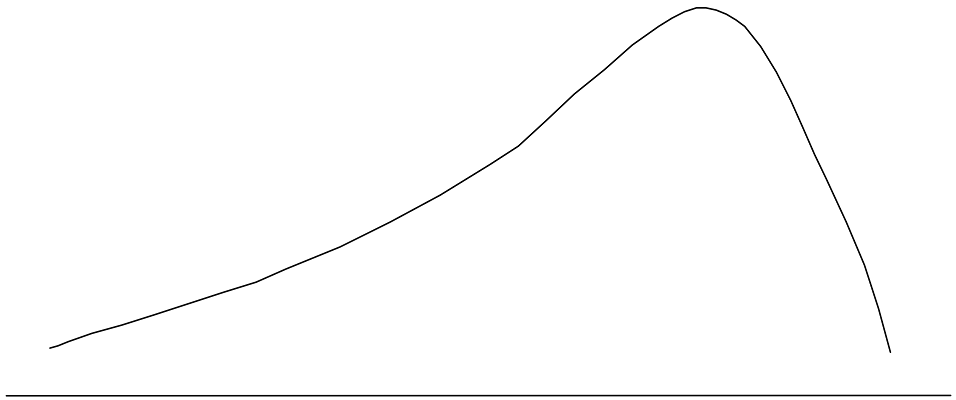
Este tipo de asimetría se presenta cuando existe una mayor concentración de datos en las primeras clases en comparación con las últimas. Se puede visualizar fácilmente cuando el extremo o "cola" de la derecha del gráfico se prolonga más que el de la izquierda.



Asimetría positiva

Asimetría Negativa o por la izquierda

Cuando hay mayor concentración de datos en las últimas clases en comparación con las primeras, es decir, cuando la "cola" izquierda de la curva se prolonga más que la derecha se dice que, la distribución de frecuencias es *asimétrica negativa o por la izquierda*.



Asimetría Negativa

Diagrama de tallo y hojas

(Basado en Walpole, R. y Myers, R. (1993) *Probabilidad y Estadística*. Págs. 59-61.)

Un diagrama de *tallo y hojas* es una representación visual de los datos que es a la vez una tabla y un gráfico. Es como un histograma horizontal con el cual se puede visualizar rápidamente la distribución de los datos. El diagrama de *tallo y hojas* provee más detalles que un histograma, ya que cada punto del gráfico representa un valor individual de los datos.

Construcción

Para ejemplificar la elaboración de un diagrama de tallo y hojas, considérese los datos de la siguiente tabla, que representan la duración de 40 baterías de carro similares. Las baterías estaban garantizadas para durar 3 años.

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Diagrama de tallo y hojas

Primero, se divide cada observación en dos partes que consisten en un tallo y una hoja de tal forma que el primero represente el dígito que es el entero y la hoja corresponda a la parte decimal del número.

Tallos	Hojas	Frecuencia
1	69	2
2	25696	5
3	4318514723628297130097145	25
4	71354172	8

El *tallo* de 1.9

La *hoja* de 1.9

Diagrama de tallo y hojas

Tallos	Hojas	Frecuencia
1 S	69	2
2 I	2	1
2 S	5696	4
3 I	431142322130014	15
3 S	8576897975	10
4 I	13412	5
4 S	757	3

En este diagrama modificado de doble tallo y hojas, donde los tallos que corresponden a las hojas 0, 1, 2, 3 y 4 han sido identificados con el símbolo I (inferior), y los tallos correspondientes a las hojas 5 a 9 por el símbolo S (superior).

Diagrama de tallo y hojas

Puede lograrse un incremento adicional en el número de tallos al escribir cada valor del tallo cinco veces en el lado izquierdo de la línea vertical donde se podría, ahora, identificar con la letra *a* al tallo para las hojas 0 y 1, con la *b* para las hojas 2 y 3, la *c* para las hojas 4 y 5, la *d* para las 6 y 7 y la *e* para las hojas 8 y 9. Para los datos de la Tabla 1 se usarían entonces, los tallos 1*d*, 1*e*, 2*a*, 2*b*, 2*c*, 2*d* y 2*e* para dibujar una tabla de cinco tallos y hojas.

Si los datos consisten en cantidades, desde $8.800 a $9.600, las cuales representan los mejores tratos posibles de 100 automóviles nuevos de un cierto distribuidor y se construye un diagrama de tallo y hojas, los tallos serían 88, 89, 90, . . , 96 y ahora cada hoja tendría dos dígitos. Un vehículo que se vende a $9,385 tendría un valor de tallo de 93 y la hoja de dos dígitos 85.

Diagrama de tallo y hojas

Las hojas de múltiples dígitos que pertenecen a un mismo tallo generalmente se separan con comas en el diagrama de tallo y hojas. Los puntos decimales en los datos casi siempre se ignoran cuando todos los dígitos a la derecha del punto decimal representan la hoja. Tal es el caso de la Tabla 2 y la Tabla 3. Sin embargo, si los datos consisten en los números en el rango de 21,8 a 74,9, se podrían seleccionar como tallos los dígitos 2, 3, 4, 5, 6, 7, de tal manera que un número, por ejemplo el 48,3, tendría un valor de tallo de 4 y de hoja 8,3.

También se acostumbra a redondear el valor de la variable a la hoja más cercana para presentarla sólo con un dígito. Por ejemplo, si una variable toma el valor 6,25 tendrá un tallo de 6 y una hoja de 3.

Diagrama de tallo y hojas

<u>Ejemplo</u>

A continuación se ilustra el diagrama de tallo y hoja para 25 observaciones del rendimiento por lote de un proceso químico. (Tomado de Newbold, P. (1998). *Estadística para los Negocios y la Economía.* Pág. 7.)

Tallo	Hoja
6I	134
6S	556
7I	0113
7S	57889
8I	1344
8S	788
9I	23
9S	5

Las *hojas* están ordenadas

Medidas Descriptivas Numéricas

Frecuentemente una colección de datos se puede reducir a una o unas cuantas medidas numéricas sencillas que resumen al conjunto total. Tales medidas son más fáciles de comprender que el conjunto de datos originales o ya agrupados. Tres características importantes de los datos que las medidas numéricas ponen de manifiesto son:

El valor central o típico de los datos

La dispersión de los datos

La forma de la distribución de los datos

Medidas de Posición o Localización (tendencia central)

Las medidas de posición se utilizan para indicar un valor que tiende a tipificar o a ser el más representativo de un conjunto de datos. Las tres medidas que más comúnmente se emplean son la ***media***, la ***mediana*** y la **moda**.

1. **Media**

   a. ***Media Aritmética***

   La media aritmética se calcula al sumar los datos y al dividir este resultado entre el número de valores.

   La media aritmética es lo que viene a la mente de la mayoría de las personas cuando se menciona la palabra "promedio".

   Como este término tiene ciertas propiedades matemáticas deseables, es la más importante de las medidas de tendencia central.

Dada una colección de datos representada por $x_1, x_2, \ldots, x_n$, la media aritmética de una **muestra** se denotará por el símbolo $\bar{x}$, y su cálculo se puede expresar matemáticamente como:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

La media aritmética de una **población** se denotará por μ :

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Ejemplo:

Si un granjero quiere conocer el peso promedio de sus ocho cerdos cuyos pesos en kilogramos son: 172, 177, 178, 173, 177, 174, 176, 173; realizará el siguiente cálculo:

$$\frac{172+177+178+173+177+174+176+173}{8}=\frac{1400}{8}=175$$

Es decir, el peso promedio de esos cerdos es 175 Kg.

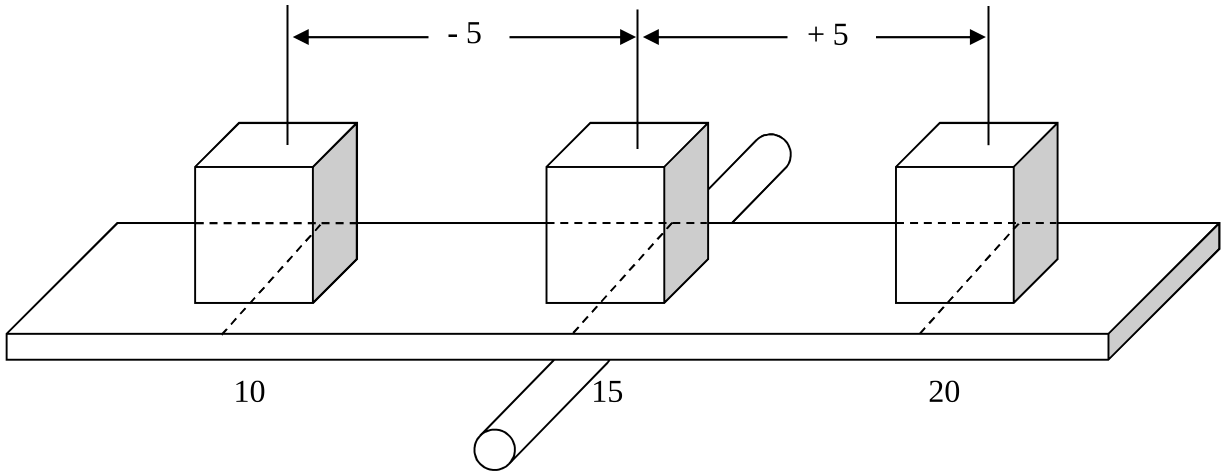
Nota

i. La media aritmética viene expresada en las mismas unidades que los datos originales.

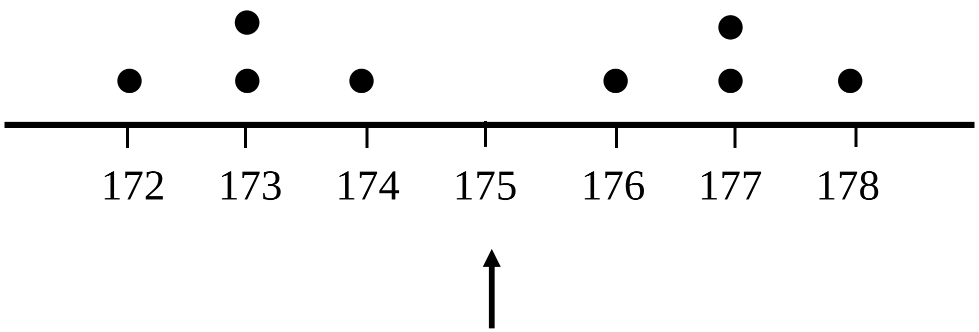
ii. La media aritmética no tiene que coincidir con alguno de los datos de la colección. Como se observa en el último ejemplo el valor $\overline{X}$=175 Kg. no aparece en los pesos del grupo de cerdos.

iii. Quizás la manera más adecuada de interpretar la media aritmética sea la que se hace desde el punto de vista de la física, en el sentido de que la media de una serie de datos representa el *centro de gravedad* o *punto de equilibrio* de esos datos.

Una representación física de la media es imaginar una barra con un punto de apoyo central que sostiene pesos iguales en sitios correspondientes a los valores de un conjunto. La media de los números 10, 15 y 20 se puede ilustrar como se observa en la siguiente figura:



En el último ejemplo también podemos observar visualmente el punto de equilibrio o centro de gravedad de esos datos:



$\bar{x}$ =175 Kg. constituye el punto en donde se logra el equilibrio.

Ejercicio:

Para los datos no agrupados, de estudio en clase, calcule la media aritmética para las variables peso, número de hermanos, visitas a la discoteca, visitas al cine, estatura e ingreso mensual del hogar.

b. ***Media Ponderada***

La fórmula de la *media aritmética* supone que cada observación es de igual importancia.

Por ejemplo, un profesor informa a su clase que efectuará cuatro parciales. Estos, con respecto a la calificación final del curso equivalen a:

Parcial 1: 20%, Parcial 2: 30%, Parcial 3: 20% y Parcial 4: 30%

El cálculo de la media deberá considerar las diferentes ponderaciones de los exámenes.

Se conoce como ***peso*** o ***ponderación*** a los factores cuantitativos que modifican a cada uno de los datos.

La ***media ponderada*** de una colección de datos $x_1, x_2, \ldots, x_n$, cuyas respectivas ponderaciones son $w_1, w_2, \ldots, w_n$ se define como:

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

Así, un alumno que logre las siguientes calificaciones:

Evaluación	**Calificación**	**Ponderación**
1	15	0,30
2	12	0,20
3	19	0,20
4	12	0,30

$$\bar{x} = \frac{0{,}30(15) + 0{,}20(12) + 0{,}20(19) + 0{,}30(12)}{0{,}30 + 0{,}20 + 0{,}20 + 0{,}30} = 14{,}3$$

Este alumno obtendrá un promedio de 14,3 puntos.

Si todas las evaluaciones poseen la misma importancia, entonces el promedio sería 14,5 puntos. ¿Por qué?

Ejemplo:

Supóngase que el semestre anterior un estudiante dela Carrera de Estadística cursó Matemática I, Inglés, Métodos Estadísticos I y Sociología, obteniendo las siguientes calificaciones:

Materia	Unidades Crédito	Calificación
Matemáticas I	6	10
Sociología	3	16
Métodos Estadísticos I	5	13
Inglés	3	20

$$\bar{x}_p = \frac{6(10)+3(16)+5(13)+3(20)}{6+3+5+3} = 13{,}71 \qquad \bar{x} = \frac{10+16+13+20}{4} = 14{,}75$$

a. ***Media aritmética para datos agrupados en distribuciones de frecuencias***

Es posible utilizar una variante de la fórmula para calcular la media ponderada, a fin de obtener la media de una distribución de frecuencias. Las ponderaciones son sustituidas por las frecuencias absolutas simples y la fórmula se convierte en:

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i m_i}{n}$$

<u>Ejercicio</u>:

Calcular la media aritmética para las distribuciones de frecuencias de las variables peso, visitas a la discoteca, estatura, número de hermanos, visitas al cine e ingreso mensual del hogar.

Medidas Descriptivas Numéricas

Propiedades de la media aritmética

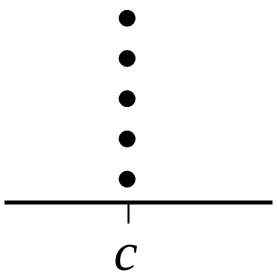
Sea $x_1, x_2, \ldots, x_n$, una colección de datos cuya media aritmética es $\overline{X}$, entonces se cumple que:

1. La suma de las desviaciones o diferencias de cada uno de los datos con respecto a su media, es cero:

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = 0$$

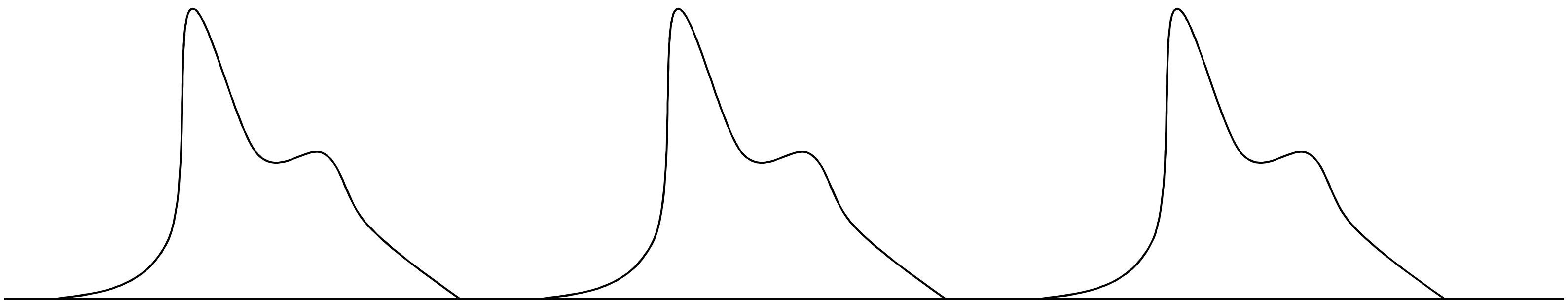
2. $\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ es un valor mínimo.

3. Si todos los datos son iguales a un mismo valor fijo o constante *c*, entonces la media de esos datos también es igual a *c*:



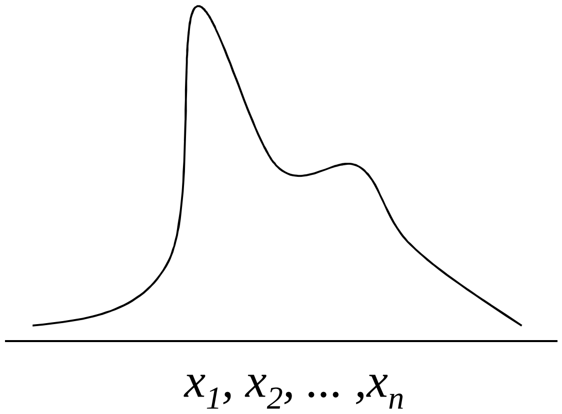
4. Si a cada uno de los datos originales se le suma un mismo número real *c*, entonces se tiene una nueva colección de datos $x_1 + c,\ x_2 + c,\ \ldots,\ x_n + c$, cuya media viene dada por $\bar{X} + c$.

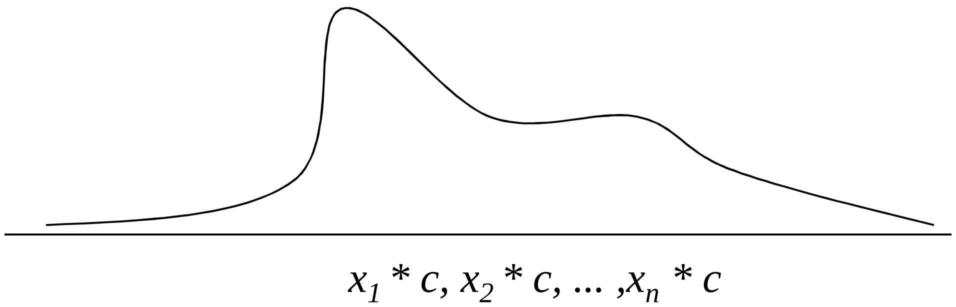
Esta situación se puede visualizar gráficamente de la siguiente manera:



$x_1 - c,\ x_2 - c,\ \ldots,\ x_n - c$ $\qquad$ $x_1,\ x_2,\ \ldots,\ x_n$ $\qquad$ $x_1 + c,\ x_2 + c,\ \ldots,\ x_n + c$

5. Si cada uno de los datos originales se multiplica por un mismo número real c, entonces se genera una nueva colección de datos $x_1 * c$, $x_2 * c$, ... , $x_n * c$, cuya media viene dada por $\bar{X} * c$.





6. Si se tienen *m* diferentes grupos de datos de distintos tamaños n_1, n_2, ... , n_m respectivamente, entonces la media de todos esos datos juntos viene dada por:

$$\overline{\overline{x}} = \frac{\sum_{i=1}^{m} n_i \overline{x}_i}{\sum_{i=1}^{m} n_i}$$

***Nota*:**

Obsérvese que: $\overline{\overline{x}} \neq \frac{\overline{x}_1 + \overline{x}_2 + \cdots + \overline{x}_m}{m}$

Ejercicios:

1. Si en un semestre un estudiante aprobó sus cuatro materias con 15 puntos ¿Cuál fue su calificación promedio?
2. Haciendo referencia al ejemplo de los pesos de los cerdos, suponga que al granjero le han recomendado un nuevo alimento para cerdos que según parece los engorda 20 Kg. en quince días. ¿Cuál será el peso promedio de los cerdos dentro de quince días, luego de utilizar el nuevo alimento?
3. Suponga ahora que todos los cerdos del granjero se enferman a causa de un virus y se detecta cinco días después que todos estos animales han disminuido exactamente 10 Kg. ¿cuál es ahora el peso promedio de los cerdos?
4. Las secciones 03 y 05 de la asignatura Estadística I, tienen 66 y 73 alumnos respectivamente. Se realiza la primera evaluación y se obtienen las siguientes notas promedio por sección: $\overline{x}_1 = 15$ y $\overline{x}_2 = 12$. ¿ Cuál es la nota promedio del primer parcial para las dos secciones juntas?

Si en el ejemplo de los cerdos, se incluye otro de esos animales cuyo peso es de 490 Kg.

a. Calcule la media aritmética.

b. ¿Esta media aritmética representa adecuadamente a los pesos de los cerdos?

$$\overline{x} = \frac{172+177+178+173+177+174+176+173+490}{9} = \frac{1890}{9} = 210$$

Desventajas de la media aritmética

1. No puede calcularse cuando los datos están agrupados en distribuciones de frecuencias que tienen un intervalo de clase abierto.
2. La principal desventaja es que se ve afectada por la presencia de *valores extremos o atípicos* en los datos.

Ventajas de la media aritmética

1. Es un promedio que toma en cuenta todos los valores de una colección de datos.
2. Es fácil de calcular y se presta a operaciones algebraicas, lo que la convierte en la medida de tendencia central más utilizada tanto en estudios descriptivos como para realizar inferencias.
3. En general, para una serie dada de datos existe una buena aproximación entre el valor de la media para los datos no agrupados y la media de los datos agrupados.