

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:

1. El valor central o típico de los datos
2. La dispersión de los datos
3. 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 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. La media aritmética se calcula al sumar los datos y al dividir este resultado entre el número de valores.

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.

Dada una colección de datos representada por $x_1, x_2, \dots, x_n$, la media aritmética de una muestra se denotará por el símbolo $\bar{x}$ (que se lee "equis barra"), y su calculo se puede expresar matemáticamente como:

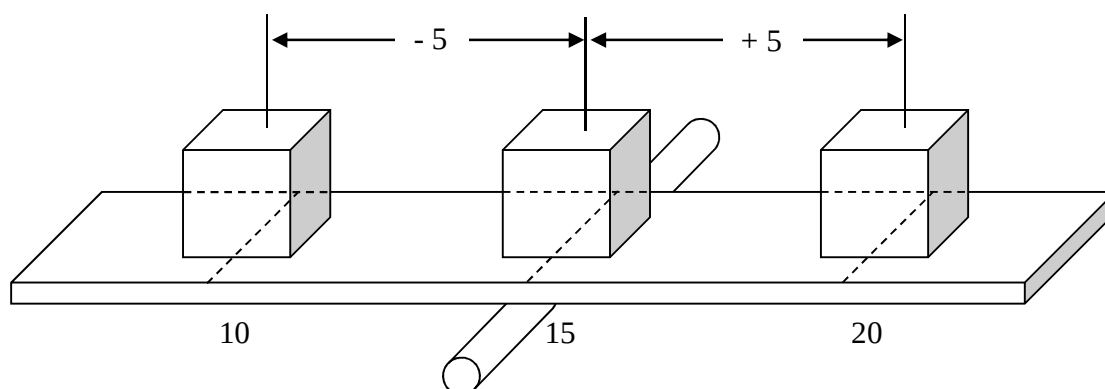
$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

El procedimiento para calcular la media aritmética es el mismo, independientemente si un conjunto de datos se refiere a las observaciones de la muestra o a todos los valores de la población. Sin embargo, se utiliza el símbolo μ para la media de una población y N para el número de elementos en la misma:

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

Nota

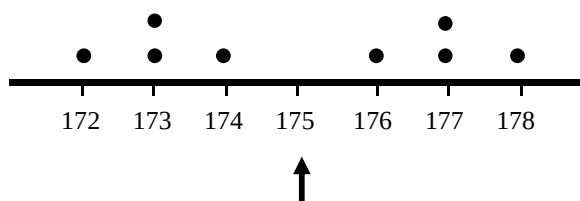
- i. La media aritmética viene expresada en las mismas unidades que los datos originales.
- ii. La media aritmética no tiene que coincidir con alguno de los datos de la colección. Como se observa en el último ejemplo el valor $\bar{x} = 175$ Kg. no aparece en los pesos del grupo de cerdos.
- iii. Quizás 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:



Nótese como la media es el *punto de equilibrio* de la tabla; las diferencias positivas y negativas se contrabalancean entre sí.

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.

**Nota**

No debe interpretarse la media como punto medio de los datos. La media representa el punto de equilibrio de las observaciones, el cual no tiene que ser igual al punto medio. En el gráfico anterior el punto de equilibrio coincide con el punto medio debido a que esos datos se distribuyen *simétricamente*.

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. Habitualmente, suele suceder así, sin embargo, existen algunas excepciones. 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 $x_1, x_2, \dots, x_n$, cuyas respectivas ponderaciones son $w_1, w_2, \dots, w_n$ 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

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

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 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

Así el promedio ponderado del estudiante fue de:

$$\bar{x}_p = \frac{6(10) + 3(16) + 5(13) + 3(20)}{6 + 3 + 5 + 3} = 13,71 \text{ puntos}$$

y su promedio aritmético simple:

$$\bar{x} = \frac{10 + 16 + 13 + 20}{4} = 14,75 \text{ puntos}$$

¿A qué se debe que los dos promedios anteriores sean distintos?

c. *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:

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

Ejercicio:

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.

Nota:

En el caso de una *distribución de frecuencias para valores individuales de la variable*, mediante la fórmula se obtendrá la misma respuesta como si se trabajara con datos originales. Si las clases de la distribución de frecuencias son intervalos, el agrupamiento hace que se pierda información y por tanto la media resultante es una aproximación. El uso de los puntos medios de clase (marcas de clase) los considera como promedios de clase, que representan a la clase respectiva, lo cual no siempre se cumple. Sin embargo, si no se dispone de datos originales, no existe otra alternativa razonable. Además la aproximación de esta fórmula a la verdadera media es generalmente buena.

Propiedades de la media aritmética

La media aritmética presenta ciertas propiedades útiles e interesantes, que explican por qué es la medida de tendencia central que se utiliza más ampliamente.

Sea $x_1, x_2, \dots, x_n$, una colección de datos cuya media aritmética es $\bar{x}$, entonces se cumple que:

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

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

Ejemplo:

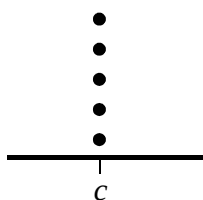
En el ejemplo de los pesos de los cerdos se obtuvo que la media aritmética es 175 Kg. Ahora calculando las desviaciones con respecto a $\bar{x}=175$ se tiene que:

$$\begin{array}{rcl}
 172 - 175 & = & -3 \\
 177 - 175 & = & +2 \\
 178 - 175 & = & +3 \\
 173 - 175 & = & -2 \\
 177 - 175 & = & +2 \\
 174 - 175 & = & -1 \\
 176 - 175 & = & +1 \\
 173 - 175 & = & -2 \\
 \hline
 & & 0
 \end{array}$$

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

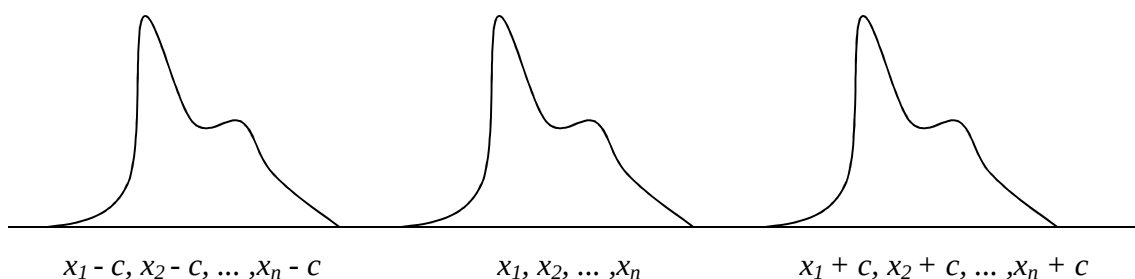
Si se calcula la expresión anterior sustituyendo $\bar{x}$ por cualquier otro valor arbitrario que se nos ocurra, se obtiene un valor mayor al que se consigue utilizando $\bar{x}$.

- iii. 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 :



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

Esta situación se puede visualizar gráficamente de la siguiente manera:

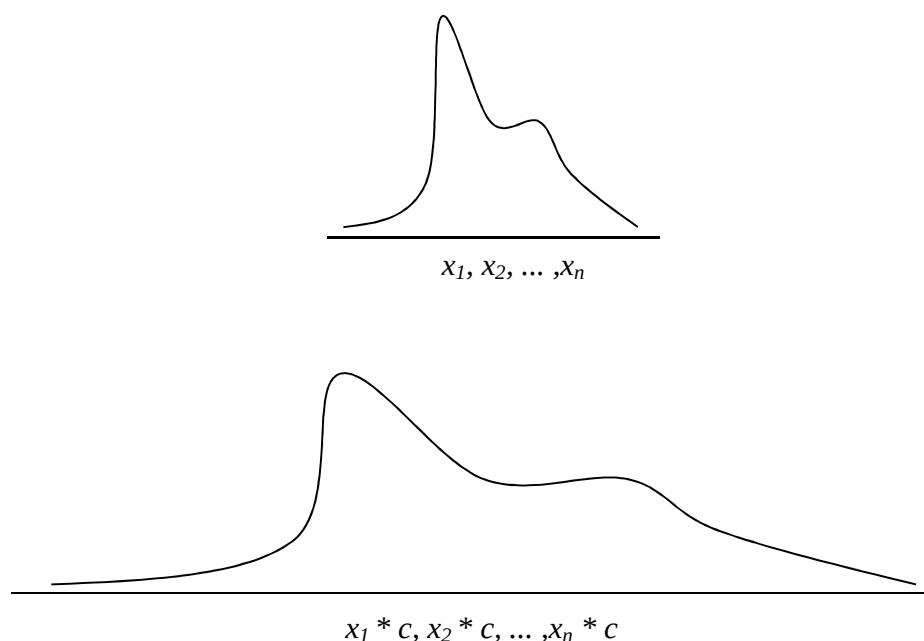


Al sumar la misma constante a cada uno de los datos, realmente lo que estamos haciendo es desplazar sobre el eje horizontal los datos hacia la derecha si la constante

es positiva o hacia la izquierda si la constante es negativa. Entonces la media aritmética se "corre" con los datos.

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

En la siguiente ilustración se puede observar como se ensancha la distribución de los datos originales cuando estos han sido modificados al multiplicar cada uno por una constante, con lo cual la media se ve afectada.



- vi. Si se tienen m diferentes grupos de datos de distintos tamaños $n_1, n_2, \dots, n_m$ respectivamente, entonces la media de todos esos datos juntos viene dada por:

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

Nota:

Obsérvese que: $\bar{\bar{x}} \neq \frac{\bar{x}_1 + \bar{x}_2 + \dots + \bar{x}_m}{m}$

Ejemplo:

Si en un semestre un estudiante aprobó sus cuatro materias con 15 puntos ¿Cuál fue su calificación promedio?

De acuerdo a la propiedad iii, la media aritmética de sus calificaciones fue de 15 puntos.

Ejemplo:

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?

Nótese que todos los cerdos aumentan 20 Kg., así que a cada uno de los pesos originales se le debe sumar la constante $c = 20$. En consecuencia, de acuerdo a la propiedad iv. dentro de quince días el peso promedio de los cerdos debe ser $175+20 = 195$ Kg.

Ejemplo:

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?

Ejemplo:

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: $\bar{x}_1 = 15$ y $\bar{x}_2 = 12$. Entonces la nota promedio del primer parcial para las dos secciones juntas es:

$$\bar{\bar{x}} = \frac{66*15 + 73*12}{139} = 13,42 \text{ puntos}$$

¡OJO es falso que! $\bar{\bar{x}} = \frac{12+15}{2} = 13,5$

Ejemplo:

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

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

Este valor 210 Kg. transmite una idea equivocada de la realidad en cuanto al peso de la mayoría en ese grupo de cerdos. ¿Qué es lo que provoca que $\bar{x}$ no sea representativa de los pesos de los cerdos?

En estos casos no debe utilizarse la media para calcular el peso promedio, sino que se recomiendan otras medidas de tendencia central.

Desventajas de la media aritmética

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

Ventajas de la media aritmética

- Es un promedio que toma en cuenta todos los valores de una colección de datos.
- 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.
- 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.

2. Mediana

La mediana de una colección de datos, que previamente han sido *ordenados*, es aquél valor más central o que está más en medio en el conjunto de datos. En otras palabras, la mediana es *mayor* que aproximadamente la mitad de los datos y *menor* que (aproximadamente) la otra mitad. Así se tiene que aproximadamente 50% de las observaciones se encuentran por arriba y 50% (aproximadamente) por debajo de ella. La mediana se denota Md (también algunos autores la denotan como $\tilde{X}$).

Ejemplo:

Los tiempos de los miembros de un equipo de atletismo en una carrera de 1,6 Km están dados en la siguiente tabla, calcule la mediana.

Miembro	1	2	3	4	5	6	7
Tiempo (en minutos)	4.2	9.0	4.7	5.0	4.3	5.1	4.8

En primer lugar se deben ordenar los datos: 4.2 4.3 4.7 4.8 5.0 5.1 9.0.


Mediana

$Md = 4.8$ minutos, es el valor que está en el centro de los datos.

Ejemplo:

Calcule la mediana para el número de pacientes tratados en la sala de emergencias de un hospital durante ocho días consecutivos:

Día	1	2	3	4	5	6	7	8
No. de pacientes	49	52	86	30	35	31	43	11

Los datos ordenados son: 86 52 49 43 35 31 30 11


Centro de los datos

La mediana en este caso puede ser 43 ó 35, o también cualquier valor entre 43 y 35. Para evitar esta imprecisión, se acepta tomar como mediana la suma de los dos valores centrales y se dividen entre dos:

$$Md = \frac{43 + 35}{2} = 39.$$

Nota:

Si se tienen n observaciones ordenadas, la mediana es la observación que ocupa la posición $\frac{n+1}{2}$ cuando n es impar y la media de las observaciones que ocupan las posiciones $\frac{n}{2}$ y $\frac{n+2}{2}$ cuando n es par.

Ejemplo:

Regresando al ejemplo de los tiempos del equipo de atletismo, se pide calcular la media y comparar este resultado con el de la mediana ya obtenida.

Entonces, se obtiene que $\bar{x} = 5.3$ minutos y antes se obtuvo que $Md = 4.8$ minutos. Nótese que en esos datos existe un *valor atípico*: 9.0 minutos. Por tanto, la media aritmética $\bar{x}$ se distorsiona. La mediana, en cambio, *no se ve distorsionada* por la presencia del valor 9.0. Este valor pudo haber sido 15.0 o incluso 45.0 y la mediana ¡seguirá siendo la misma!

Cálculo de la mediana para datos agrupados en distribuciones de frecuencias**i. Cuando las clases son intervalos**

- Se ubica la *clase mediana*, la cual viene dada por aquella clase que contiene a la frecuencia acumulada $\frac{n}{2}$ o equivalentemente a la frecuencia relativa acumulada 0,5.
- Luego de ubicada la clase mediana, el cálculo de la mediana se hace mediante un proceso de interpolación el cual conduce a la siguiente fórmula:

$$Md = LI_m + \left(\frac{\frac{n}{2} - F_{am}}{f_m} \right) * C_m$$

en donde,

LI_m : Límite inferior de la clase mediana

n : No. total de observaciones o datos

F_{am} : Frecuencia acumulada anterior a la clase mediana

f_m : Frecuencia absoluta de la clase mediana

c_m : Amplitud de la clase mediana

Ejemplo: Calcular la mediana para la distribución de frecuencias de la variable peso.

En primer lugar se debe ubicar la clase mediana, para esto se debe calcular:

$$\frac{n}{2} = \frac{43}{2} = 21,5$$

Ahora se ubica la frecuencia acumulada que contiene a 21,5:

	Clases	mi	fi	fri	Fi	Fri
	[40-49)	44,5	4	0,0930	4	0,0930
	[49-58)	53,5	10	0,2326	14	0,3256
Clase mediana	[58-67)	62,5	15	0,3488	29	0,6744
	[67-76)	71,5	7	0,1628	36	0,8372
	[76-85)	80,5	5	0,1163	41	0,9535
	[85-94)	89,5	1	0,0233	42	0,9767
	[94-103)	98,5	1	0,0233	43	1
	Totales		43		1	

Frecuencia acumulada
que contiene a 21,5

También se puede ubicar la clase mediana encontrando la frecuencia relativa acumulada que contiene a 0,5000.

Entonces, se tiene que:

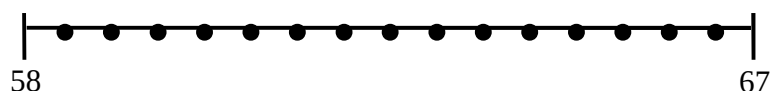
$$Md = 58 + \left(\frac{21,5 - 14}{15} \right) * 9$$

$$Md = 62,5$$

De esta manera, $Md = 62,5$ Kg. representa el valor central de los pesos. Es decir, aproximadamente la mitad de los estudiantes de Métodos Estadísticos I tienen un peso inferior a 62,5 Kg. y aproximadamente la otra mitad pesa más de 62,5 Kg.

Nota:

En la fórmula de la mediana se está suponiendo que los valores en el intervalo de clase que contiene la mediana están *uniformemente espaciados (o equidistantes)*. Entonces, en el ejemplo anterior se está suponiendo que los 15 valores que contiene la clase mediana están uniformemente espaciados en [58 -67):



Ejercicio:

Calcule la mediana para las distribuciones de frecuencias correspondientes a las variables estatura, índice académico e ingreso mensual del hogar.

ii. Cuando las clases son valores individuales

- ♦ Se calcula $n/2$ (o se considera el valor 50% de las observaciones)
- ♦ Si el valor $n/2$ **NO APARECE** en la columna de la F_i , entonces se ubica aquella frecuencia acumulada que lo contiene y la mediana será el valor de la variable correspondiente a esa frecuencia acumulada.

- ♦ Una forma equivalente de hacer lo anterior es la siguiente, si el valor 50% no aparece en la columna de las $Fr_i * 100$ entonces se ubica aquella frecuencia que lo contenga y la mediana será el valor de la variable correspondiente a esa clase.
- ♦ Si el valor $n/2$ **APARECE** en la columna de las F_i , es decir que coincide con la frecuencia acumulada de alguna clase, entonces la mediana viene dada por la media aritmética de ese valor de la variable y el siguiente valor.
- ♦ También, si el valor 50% coincide con alguna de las $Fr_i * 100$, entonces la mediana viene dada por el promedio de los valores de la variable correspondiente a esa clase y a la siguiente.

Ejemplo:

La siguiente distribución de frecuencias corresponde al número de materias que cursan 112 estudiantes de la carrera de Contaduría Pública. Calcule la mediana.

Inicialmente se debe calcular $\frac{n}{2} = 56$. Entonces 56 no aparece en la columna de las F_i .

Por tanto, $Md = 4$ materias.

Número de materias				
	f_i	f_{ri}	F_i	Fr_i
	1	0,0089	1	0,0089
	2	0,0089	2	0,0179
	3	12	0,1071	0,1250
Mediana →	4	56	0,5000	0,6250
	5	40	0,3571	0,9821
	6	2	0,0179	1,0000
Totales 112 1,0000				

Frecuencia acumulada que contiene a $\frac{n}{2} = 56$

Ejemplo:

Calcule la mediana para la siguiente distribución de frecuencias, en donde $\frac{n}{2} = 30$. Es decir, $n/2$ aparece en la columna de las frecuencias acumuladas:

Clase	f_i	F_i	Fr_i
5	8	8	0,1333
6	9	17	0,2833
Mediana →	7	30	0,5000
	8	40	0,6667
	9	46	0,7667
	10	60	1

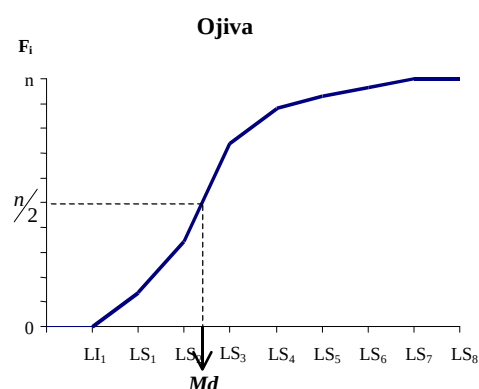
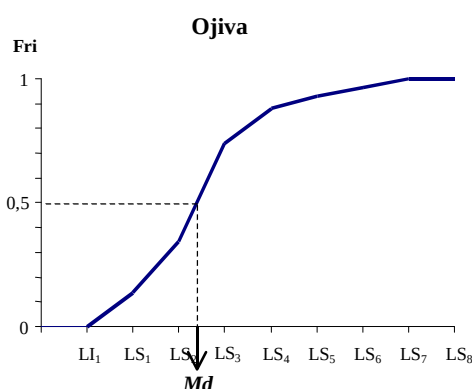
$n = 60$

Frecuencia acumulada que coincide con $\frac{n}{2} = 30$

Entonces la mediana viene dada por: $Md = \frac{7+8}{2} = 7,5$

La mediana gráficamente

Mediante la ojiva y a través del método de interpolación visto en esa sección se puede obtener de manera gráfica el valor de la mediana de una colección de datos agrupados en una distribución de frecuencias cuyas clases son intervalos. Si se usa la ojiva construida con la frecuencia acumulada F_i la mediana será aquél valor en el eje horizontal cuya ordenada sea $\frac{n}{2}$. En el caso de usar la ojiva construida con la frecuencia relativa acumulada F_{ri} (o $F_{ri} \cdot 100$), la mediana vendrá dada por el valor en el eje de las abscisas que corresponda a la ordenada 0,5 (o 50%).



Así aplicando el método de interpolación visto antes se obtiene la fórmula del cálculo de la mediana:

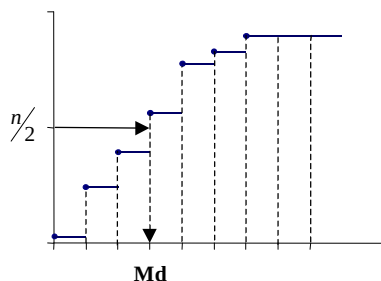
$$Md = LI_m + \left(\frac{\frac{n}{2} - F_{am}}{f_m} \right) * C_m$$

Ejercicio:

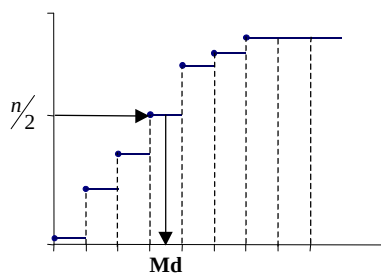
Obtenga gráficamente la fórmula anterior para el cálculo de la mediana.

Para el caso de distribuciones de frecuencias cuyas clases son valores individuales de la variable, se puede hallar gráficamente la mediana por medio del diagrama de frecuencias acumuladas. El procedimiento es similar que cuando se usa la ojiva. Se ubica en el eje vertical $\frac{n}{2}$ (o 50% si se usó $F_{ri} \cdot 100$) y se traza una línea paralela al eje horizontal, así se presentan las dos situaciones siguientes:

- ♦ Si la línea intercepta el gráfico, entonces la mediana viene dada por el valor en el eje de las abscisas que corresponde a la ordenada $\frac{n}{2}$ (o 50%).

F_i DIAGRAMA DE FRECUENCIAS ACUMULADAS

- ♦ Si la línea coincide con uno de los escalones del gráfico, la mediana vendrá dada por el punto medio de ese escalón.

F_i DIAGRAMA DE FRECUENCIAS ACUMULADAS

Propiedades de la mediana

- La mediana es una medida de tendencia central de fácil comprensión pero que solamente toma en cuenta la posición que ocupan las observaciones y no el valor en sí de las mismas. Esto hace que la mediana no sea susceptible de operaciones algebraicas y en consecuencia limita su utilidad, por ejemplo para fines de inferencia estadística.
- Puede calcularse en el caso de distribuciones de frecuencias con clases abiertas siempre y cuando se disponga de la información correspondiente a la clase mediana.
- No se ve afectada ante la presencia de unos pocos *valores atípicos* y es por ello que se recomienda su uso en el caso de distribuciones marcadamente asimétricas.

3. **Moda**

La moda es el valor que más se repite, es decir el que aparece con mayor frecuencia. En otras palabras la moda es el valor más común de los datos, se denota por Mo y viene expresada en las mismas unidades que los datos.

Ejemplo:

Calcule la moda de los siguientes datos: 5, 3, 6, 5, 4, 5, 2, 4.

En este caso el valor que más se repite es el 5, por tanto $Mo = 5$.

Ejemplo:

Calcule la moda de los siguientes datos: 5, 3, 6, 5, 4, 5, 2, 4, 4.

En este conjunto de datos existen dos valores que se repiten con la misma frecuencia: 4 y 5. Así, se tienen dos modas: $Mo_1 = 4$ y $Mo_2 = 5$.

Ejemplo:

Calcule la moda de los siguientes datos: 5, 3, 3, 5, 6, 2, 6, 4, 2, 4.

En este caso no existe la moda dado que no hay datos que se repitan más que otros.

En conclusión, una colección de datos puede que no tenga moda o puede ser que posea una o más modas.

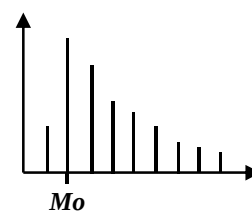
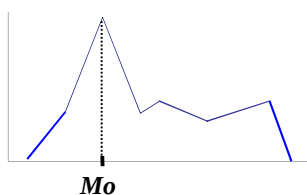
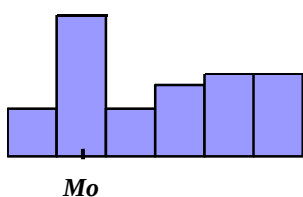
Nota:

Cuando hay una sola moda la distribución de datos se llama *unimodal*, con dos modas *bimodal*, con tres modas *trimodal* y con 4 o más modas se llama *polimodal* o *multimodal*. Si todos los valores se presentan la misma cantidad de veces, la distribución se llama *amodal*.

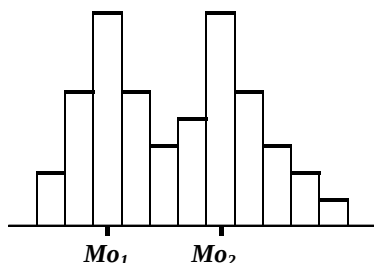
Cuando los datos están agrupados en distribuciones de frecuencias cuyas clases presenten igual amplitud, se toma el *punto medio* de la clase con mayor frecuencia absoluta como la moda.

Representación gráfica de la moda:

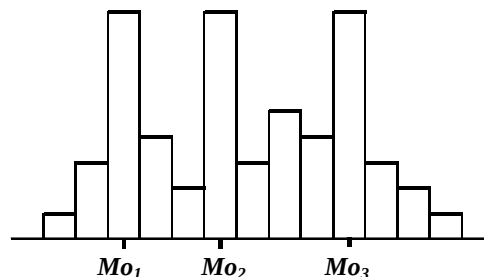
Distribuciones unimodales:



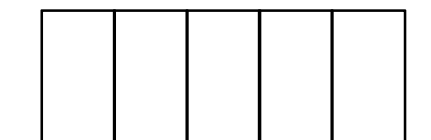
Distribución bimodal:



Distribución trimodal:



Distribución amodal:

**Ejercicio:**

Calcular la moda para las distribuciones de frecuencias correspondientes a las variables peso, número de hermanos, estatura, ingreso mensual del hogar, número de veces que visita la discoteca e índice académico.

Propiedades de la Moda:

- i. La moda en realidad no es una medida de tendencia central, sino más bien indica punto(s) de concentración de datos.
- ii. No es susceptible de operaciones algebraicas y de allí que su uso es limitado.
- iii. Es la única de las medidas descriptivas que puede utilizarse para datos cualitativos de cualquier tipo.
- iv. Es posible su cálculo en algunos casos de distribuciones de frecuencias con intervalos de clase abiertos.
- v. Es una medida muy imprecisa e inestable. En una distribución de frecuencias depende de la forma en como se construyen las clases.

Ejemplo:

Considere la siguiente distribución de frecuencias:

Clases	f_i
[0 - 5)	3
[5 - 10)	5
[10 - 15)	6
[15 - 20)	6
[20 - 25)	4
[25 - 30)	7
[30 - 35)	2
Total	33

La clase modal es [25 - 30) y la moda es $Mo = 27,5$

Si se introduce una pequeña modificación en las clases, por ejemplo agrupando las dos primeras, se tiene:

Clases	f_i
[0 - 10)	8
[10 - 15)	6
[15 - 20)	6
[20 - 25)	4
[25 - 30)	7
[30 - 35)	2
Total	33

La clase modal pasa a ser [0 - 10) y $Mo = 5$. Obsérvese el cambio tan grande que se produce en la moda ya que pasa de 27,5 a 5.

vi. La moda es de utilidad en aquellos casos donde la naturaleza de los datos así lo indique.

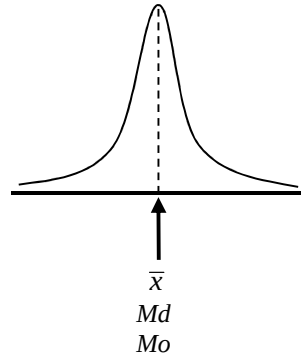
Ejemplo:

Para una fábrica de zapatos, el interés está en conocer la o las tallas más frecuentes en la población.

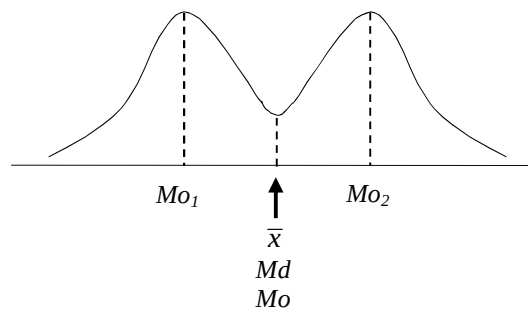
Relación entre la Media Aritmética, la Mediana y la Moda

En función de la simetría de una distribución se presentan las siguientes relaciones entre esas tres medidas:

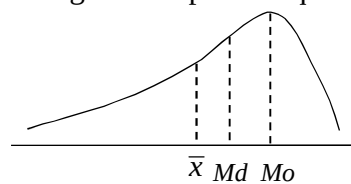
1. En distribuciones simétricas unimodales la media, la mediana y moda coinciden:



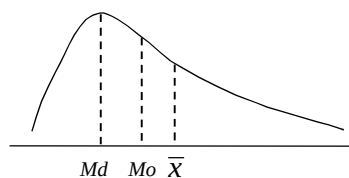
2. En distribuciones simétricas bimodales, la media y la mediana son iguales pero no coinciden con las modas.



3. En distribuciones asimétricas negativas o por la izquierda, se cumple que $\bar{x} < Md < Mo$



4. En distribuciones asimétricas positivas o por la derecha, se cumple que $\bar{x} > Md > Mo$



Selección de la Medida de Posición adecuada

Los siguientes factores deben tomarse en cuenta en el momento de la selección de la medida numérica apropiada para describir la posición o tendencia central de los datos:

1. De acuerdo al tipo de dato se puede utilizar una u otra medida de tendencia central. Las medidas que pueden aplicarse con cada tipo de dato son las siguientes:
 - i. Datos Nominales: Moda
 - ii. Datos Ordinales: Moda y Mediana
 - iii. Datos Discretos: Todas
 - iv. Datos Continuos: Todas
2. Teniendo en cuenta lo anterior se recomienda tener presente los siguientes aspectos:
 - a. **La naturaleza de la distribución de los datos.** Gráficamente se puede observar la forma general en que se distribuyen los datos. Esto es determinante en la selección del promedio adecuado.
 - Si se trata de una *distribución simétrica* o aproximadamente simétrica, se sabe que la media, la mediana y la moda coinciden y en consecuencia se puede utilizar cualquiera de ellas.
 - Si la *distribución es asimétrica*, la media aritmética no va a ser adecuada y es preferible inclinarse por la moda o la mediana.
 - b. **El concepto de tendencia central o de posición que interese reflejar en una situación dada.**
 - Si interesa conocer el valor más común de una serie de datos como por ejemplo la estatura típica de las personas que ingresan al ejército, es necesario usar *la moda*.
 - Si se desea ubicar a una persona en cuanto a su salario anual diciendo que gana por encima o por debajo de lo que gana la mitad de los trabajadores del país, entonces habrá que usar *la mediana*.
 - Cuando interesa el total de datos o reflejar el punto de equilibrio de los mismos se utiliza *la media aritmética*.
 - c. **Riesgos que se corren ante la presencia de valores atípicos.**

Si existen valores atípicos, hay que verificar si se incurrió en algún error en la recolección de la información o puede ser el alerta de alguna situación no esperada por el investigador. En todo caso hay que tener presente que la media aritmética se ve seriamente afectada ante la presencia de valores atípicos y será necesario recurrir a alguna de las otras medidas conocidas.
 - d. **Posibilidad de realizar inferencia estadística**

Cuando el análisis estadístico se realiza sobre una muestra de la población con la intención de generalizar a la totalidad, lo que se conoce como inferencia estadística, prácticamente la única medida de tendencia central utilizada hasta ahora satisfactoriamente es la media aritmética y esto se debe a que existe un fundamento teórico bien fundamentado que la respalda.

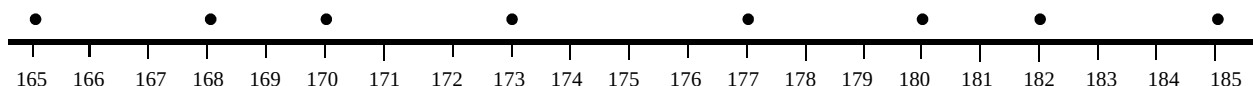
Medidas de Dispersión

Además de obtener la información que reúnen las medidas de tendencia central es muy conveniente tener conocimiento sobre el grado de dispersión o variabilidad que presentan los datos. Las *medidas de dispersión* indican si los valores están relativamente cercanos uno del otro o si se encuentran dispersos. Esta idea se ilustra en las siguientes figuras.

Recuérdese que en el ejemplo de los pesos de los cerdos tenemos los siguientes datos: 172, 177, 178, 173, 177, 174, 176, 173. El diagrama de puntos para esos valores es:



Si los cerdos de otro granjero tienen los siguientes pesos: 165, 182, 185, 168, 170, 173, 180, 177. Entonces el diagrama de puntos está dado por:



Obsérvese que ambos grupos de datos poseen la misma media aritmética y la misma mediana, $Md = \bar{x} = 175$ Kg. Además, también se puede advertir como las observaciones en el primer gráfico tienen valores relativamente más cercanos entre sí en comparación con los pesos del segundo grupo de cerdos.

Por consiguiente, además de las medidas de tendencia central, siempre es importante contar con indicadores que midan la dispersión de los datos. Una medida de tendencia central, casi nunca es suficiente por sí sola, para resumir adecuadamente las características de un conjunto de datos. Por lo general, es necesario, adicionalmente, una medida de la *dispersión* de los datos.

En general, se pueden clasificar las medidas de dispersión en *absolutas* y *relativas*. Las *medidas de dispersión absolutas* son aquellas que vienen expresadas en las mismas unidades que los datos. Las *medidas de dispersión relativas* no vienen expresadas en las unidades de los datos sino en porcentaje.

A pesar de que existen diferentes medidas de dispersión, sólo se van a considerar las más usadas:

Medidas de dispersión absolutas:

- ♦ Rango o recorrido
- ♦ Varianza
- ♦ Desviación Estándar
- ♦ Basadas en Percentiles

Medida de dispersión relativa:

- ♦ Coeficiente de Variación

Todas estas medidas, excepto el rango, toman la media como punto de referencia. En cada caso un valor cero indica que no hay dispersión, mientras que la dispersión aumenta a medida que se incrementa el valor del indicador (varianza, coeficiente de variación, etc.)

1. Rango o recorrido

Esta es la medida más sencilla de calcular y comprender. Se concentra en el valor máximo y mínimo de la colección de datos y viene dada por:

$$R = \text{Valor máximo} - \text{Valor mínimo}$$

En el caso de distribuciones de frecuencias, el rango se obtiene restándole al límite superior de la última clase el límite inferior de la primera clase:

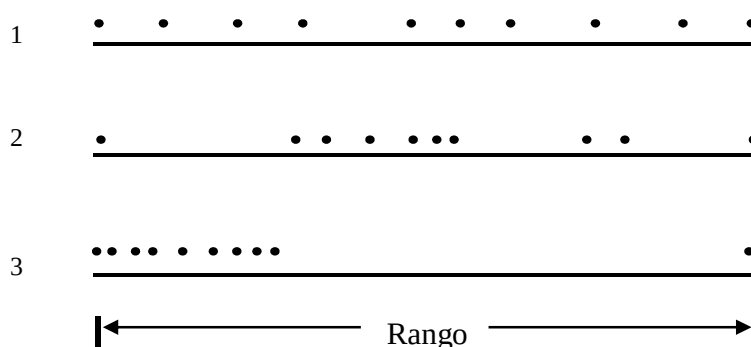
$$R = LS_K - LI_1$$

En los ejemplos anteriores para los dos grupos de cerdos se tiene que el recorrido para el grupo 1 es $R = 178 - 172 = 6$ Kg., y para el grupo 2 es $R = 185 - 165 = 20$ Kg.

La ventaja de utilizar el rango como medida de dispersión, es la sencillez de su cálculo, aun cuando se trate de un conjunto bastante grande de datos. Además, el significado de esta medida es fácil de comprender.

La principal limitación del rango es que considera solamente los valores extremos de los datos, y no proporciona información respecto a los demás valores.

En el siguiente ejemplo se presentan tres conjuntos de datos bastante diferentes, que poseen el mismo rango.



Nótese como en el primer grupo de datos, los valores se distribuyen en forma uniforme, y esta medida cumple con su objetivo. En el segundo conjunto, los datos se encuentran más agrupados y acá el rango mide de una "forma cruda" la dispersión. Sin embargo, la tercera colección demuestra cómo se puede influir fácilmente en el rango mediante unos cuantos *valores extremos*

(o valores atípicos), y presentar información bastante engañosa respecto a la dispersión de una colección de datos. Debido a estos problemas, el rango tiene una limitada utilidad ya que no resulta una medida de dispersión confiable.

2. Varianza

Supóngase que $x_1, x_2, \dots, x_n$ son las observaciones una muestra aleatoria, cuya media es $\bar{x}$. Dado que se está interesado en analizar la dispersión de estos valores, será natural fijarse en sus distancias con respecto a la media, esto es, en las diferencias:

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

Puesto que algunos valores de la muestra son mayores que la media y otros son menores, algunas de estas diferencias serán positivas y otras negativas. Es más, las diferencias están “equilibradas”, en el sentido de que su suma es 0 (por propiedad i. de la media aritmética, ver Pág.49)

Sin embargo, para analizar la dispersión de los datos, no interesa el signo de las diferencias, Así se tratará una diferencia negativa exactamente igual que una diferencia positiva de la misma cantidad. Por ejemplo, un salario que esté 100.000 bolívares por debajo de la media deberá ser tratado exactamente igual que uno que esté 100.000 bolívares por encima de la media. Una forma de conseguir este objetivo consiste en fijarse, no en las diferencias, sino en sus cuadrados:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$$

El promedio de los cuadrados de las diferencias proporciona una medida de la dispersión que se conoce con el nombre de *varianza*.

Este es un indicador que mide la dispersión de los datos con respecto a su media aritmética y se denota por S_*^2 .

Dada una colección de datos $x_1, x_2, \dots, x_n$, cuya media aritmética es $\bar{x}$, se define la *varianza* de esos datos como el promedio de las diferencias elevadas al cuadrado de cada uno de esos valores con respecto a su media. Es decir:

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

Nota

De la fórmula anterior se deduce que:

- i. Mientras más alejados estén los valores de su media mayor será el valor de la varianza y mientras más concentrados se encuentren alrededor de su media, menor será el valor de la varianza.
- ii. La varianza nunca es negativa, ya que se está sumando cantidades elevadas al cuadrado.
- iii. El valor mínimo que puede tomar es cero, el cual se logra cuando todos los valores son iguales entre sí, es decir, que no existe variabilidad entre ellos.

- iv. Si se desarrolla la fórmula anterior, se obtiene otra expresión equivalente de la varianza, más cómoda de calcular y además reduce los errores de redondeo:

$$S_*^2 = \frac{\sum_{i=1}^n x_i^2}{n} - (\bar{X})^2$$

Ejemplo:

Los datos de los pesos de los cerdos son: 172, 177, 178, 173, 177, 174, 176, 173 y la media es 175 Kg. Calcular la varianza.

$$S_*^2 = \frac{(172-175)^2 + (177-175)^2 + (178-175)^2 + (173-175)^2 + (177-175)^2 + (174-175)^2 + (176-175)^2 + (173-175)^2}{8}$$

$$S_*^2 = 4,5 \text{ Kg}^2$$

Por la otra fórmula:

$$S_*^2 = \frac{172^2 + 177^2 + 178^2 + 173^2 + 177^2 + 174^2 + 176^2 + 173^2}{8} - 175^2$$

$$S_*^2 = 4,5 \text{ Kg}^2$$

Nótese que la varianza viene expresada en las unidades de los datos pero elevadas al cuadrado. Por esta razón, la varianza resulta difícil de interpretar. Para solucionar esta situación se define la *desviación estándar*.

3. **Desviación Estándar (Desviación Típica):**

La *desviación estándar* o *desviación típica* de una colección de datos, denotada por S_* , se define como:

$$S_* = +\sqrt{S_*^2}$$

La cual viene dada en las mismas unidades de los datos.

Ejemplo:

Tomando el ejemplo anterior se tiene que:

$$S_* = \sqrt{4,5 \text{ Kg}^2}$$

$$= 2,12 \text{ Kg}$$

Para distribuciones de frecuencias

Para el caso de datos agrupados en distribuciones de frecuencias, las expresiones para la varianza y la desviación estándar son:

$$S_*^2 = \frac{\sum_{i=1}^k (m_i - \bar{x})^2 f_i}{n} = \frac{\sum_{i=1}^k m_i^2 f_i}{n} - (\bar{x})^2$$

$$\text{y } S_* = \sqrt{S_*^2} \quad \text{donde, } \bar{x} = \frac{\sum_{i=1}^n f_i m_i}{n}$$

Ejercicio:

Calcular la varianza y la desviación estándar para las distribuciones de frecuencias de las variables peso, número de hermanos, estatura, ingreso mensual del hogar, número de visitas a la discoteca y número de visitas al cine.

Propiedades de la varianza y la desviación estándar

Sea $x_1, x_2, \dots, x_n$ una colección de datos, cuya media, varianza y desviación estándar son $\bar{x}$, S_*^2 y S_* respectivamente.

- i. S_*^2 y S_* son no negativas, es decir $S_*^2 \geq 0$ y $S_* \geq 0$ para cualquier conjunto de datos.
- ii. Si cada uno de los datos $x_1, x_2, \dots, x_n$ es igual a un mismo valor fijo o constante c , entonces la varianza S_*^2 y la desviación estándar S_* son iguales a cero.

En el diagrama de puntos que se muestra en la página 50 se puede observar esta situación.

- iii. Si a cada uno de los datos originales se le suma un mismo número real c , positivo o negativo, entonces la nueva colección de datos que se origina $x_1+c, x_2+c, \dots, x_n+c$ tiene la misma S_*^2 y S_* que los datos originales.

Obsérvese el gráfico correspondiente a la propiedad iv de la media aritmética en la página 50. Nótese como la dispersión se mantiene invariante al sumar o restar una constante.

- iv. Si cada uno de los datos se multiplica por un mismo número real cualquiera c , la varianza y la desviación estándar de los "nuevos datos" $x_1*c, x_2*c, \dots, x_n*c$ vienen dadas por $c^2 S_*^2$ y $|c| S_*$ respectivamente.

Esto se ilustra en los gráficos que corresponden a la propiedad v de la media aritmética en la página 51. Obsérvese como se produce una alteración en la dispersión de los nuevos datos, ya sea disminuyendo o aumentando dependiendo del valor de c .

Ejercicio:

En un estudio realizado en un hospital se determinó que se gastaba en medicinas un promedio de Bs. 80.000 semanalmente por paciente con una desviación estándar de Bs. 15.000.

- a. Si se produce un aumento del 100% en el precio de las medicinas, cuanto será el gasto promedio por paciente y cuanto será la varianza.
- b. Cuanto será el gasto promedio por paciente y cuanto será la desviación estándar si el aumento es del 20%.

Varianza muestral y varianza poblacional

Al definir la varianza S_*^2 se ha estado suponiendo que la colección de datos $x_1, x_2, \dots, x_n$ constituye una muestra de tamaño n de una población y que $\bar{x}$ es la media de esa muestra. La *varianza poblacional*, denotada por σ^2 , de una población de N elementos cuya media poblacional es μ , se define por:

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

y la *desviación estándar poblacional* es:

$$\sigma = +\sqrt{\sigma^2}$$

Nota:

La varianza muestral también puede definirse como:

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

Se utiliza con la finalidad de, además de tener fines descriptivos, realizar inferencias sobre una población usando S^2 y no S_*^2 por cuanto se demuestra que S^2 es un mejor estimador de la varianza poblacional σ^2 que S_*^2 como se verá en el tema de *estimación*.

Coefficiente de variación

La *medida de dispersión relativa* más conocida es el *coeficiente de variación*. En algunas ocasiones es de interés comparar la dispersión de dos colecciones de datos. Si los datos están medidos en las mismas unidades y las respectivas medias aritméticas son iguales o muy parecidas es posible utilizar la desviación estándar. Si esto no se cumple, no se puede utilizar la desviación estándar para comparar las dispersiones de los dos grupos de datos.

Una medida de dispersión que permite la comparación de la dispersión en cualquier situación, que no viene expresada en ninguna unidad es el *coeficiente de variación*.

El *coeficiente de variación* se define como:

$$CV = \frac{S_*}{\bar{x}} * 100\%$$

El *coeficiente de variación* es la proporción o porcentaje de la media que representa la desviación estándar. Obsérvese como la fórmula anterior proviene de una regla de tres simple:

$$\bar{x} \rightarrow 100\%$$

$$S_* \rightarrow ?$$

Si por ejemplo el $CV=20\%$, significa que la desviación estándar representa el 20% del valor de la media aritmética.

Ejercicio:

Supóngase que se desea comparar las dispersiones de los sueldos de los empleados de las empresas "Cervecería El Cóndor" y "Aguardiente Tropical". Los sueldos promedio para estas empresas son Bs. 7000 y Bs. 2500 respectivamente; las desviaciones estándar correspondientes son Bs. 3000 y Bs. 300.

Ejercicio:

En una encuesta sobre bienes raíces en la Urbanización Santa Cecilia de una ciudad, se obtiene entre otras cosas, información sobre el valor actual de la casa y el tamaño del lote de terreno. Se está interesado en determinar si el valor de avalúo tiene mayor variabilidad que el tamaño del lote. De la mencionada encuesta se consigue lo siguiente:

Valor de la casa	Tamaño del terreno
$\bar{x} = 1.550.000$ Bs.	$\bar{x} = 650$ mts ²
$S_* = 500.000$ Bs.	$S_* = 350$ mts ²

Ejercicio:

Compare la dispersión de la distribución de frecuencias del peso de los varones con la dispersión de la distribución del peso de las hembras.

Percentiles, Deciles y Cuartiles

Además de las medidas de tendencia central, dispersión y forma, también existen algunas medidas interesantes de posición que se utilizan al resumir y analizar las características o propiedades de grandes colecciones de datos.

1. Percentiles

Los *percentiles* son aquellos valores que dividen a los datos *ordenados* de forma creciente, en cien partes iguales. Existen noventa y nueve percentiles que se denotan por $P_1, P_2, \dots, P_{99}$. Entre dos percentiles consecutivos se encuentra el 1% de los datos. Así, por ejemplo, entre los percentiles P_{10} y P_{20} se encuentran 10% de los datos.

Para denotar un percentil cualquiera usamos P_h , donde $h = 1, 2, 3, \dots, 99$. Así, la definición formal de percentil es la siguiente:

El percentil P_h de una colección de datos que previamente han sido ordenados (de forma creciente), es un valor tal que como máximo el $h\%$ de los datos son menores que él, y también como máximo un $(100-h)\%$ de los datos son mayores que él.

Como en el caso de la mediana, si dos valores consecutivos del conjunto de datos cumplen con la definición anterior, se conviene en tomar como percentil al promedio de ellos dos.

Ejemplo:

Suponga que los pesos de ocho personas (en Kg) son: 52, 97, 108, 63, 90, 74, 86, 73. Hallar lo percentiles: P_{20}, P_{50} y P_{80} .

En primer lugar se deben ordenar de forma creciente los datos:

52 63 73 74 86 90 97 108

El P_{20} es el valor tal que el 20% de los datos, es decir el 20% de 8 = 1,6 datos como máximo son menores que él, y también como máximo el 80% de 8 = 6,4 datos son mayores que él.

Observe que el valor 63 cumple con estas condiciones. Por tanto, $P_{20} = 63$ Kg.

Ahora, en el cálculo de P_{50} se observa que existen dos valores 74 y 86, que cumplen con la definición. De esta manera, $P_{50} = (74 + 86) / 2 = 80$ Kg.

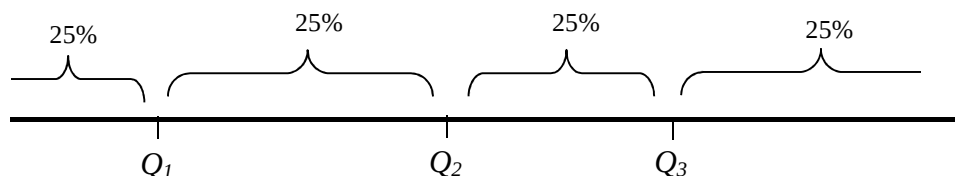
Para estos datos, P_{80} tiene como máximo 6,4 datos por debajo de él y a lo sumo 1,6 datos por encima. El valor 97 satisface esto, así $P_{80} = 97$ Kg. Nótese que ni el valor 90 ni 108 cumple con las condiciones. Por ejemplo, el valor 90 tiene cinco datos por debajo que cumple con lo que se exige pero por encima tiene a dos datos (el 25% de los datos), lo que no satisface los requerimientos para ser percentil 80.

2. Deciles

Los *Deciles* son los valores que dividen a los datos ordenados (de forma creciente) en diez partes iguales. Existen nueve deciles que se denotarán por $D_1, D_2, \dots, D_9$. Entre dos deciles consecutivos se encuentra un 10% de los datos.

3. Cuartiles

Los *cuartiles* son los valores que dividen a una colección de datos que previamente han sido ordenados en forma creciente, en cuatro partes iguales. De esta manera, existen tres cuartiles que se denotan Q_1, Q_2 y Q_3 . Nótese que entre dos cuartiles consecutivos se encuentra un 25% de los datos. Además, por debajo de Q_1 , se encuentra un 25% de los datos y por encima un 75%; mientras por debajo del cuartil tres, se encuentra un 75% de los datos y por encima de él existe un 25% de los datos.



Nótese que el segundo cuartil, Q_2 , es igual a la mediana. Además, puede dejarse ver las siguientes relaciones entre los cuartiles deciles y percentiles:

$$\begin{aligned}
 Q_1 &= P_{25} \\
 Q_2 &= D_5 = P_{50} = Md \\
 Q_3 &= P_{75} \\
 D_1 &= P_{10} \\
 D_2 &= P_{20} \\
 &\vdots \\
 D_9 &= P_{90}
 \end{aligned}$$

Nota:

A los cuartiles, deciles y percentiles en general se les denominan *cuantiles*

Ejercicio:

Para los datos no agrupados de estatura, calcular e interpretar: los cuartiles, el decil tres y el percentil diez.

Cálculo de Percentiles en distribuciones de frecuencias

1. En el caso de distribuciones de frecuencias cuyas clases son intervalos, los percentiles, de la misma manera como se hizo con la mediana, se pueden calcular mediante un método de interpolación tanto de forma algebraica como gráfica.

Algebraicamente, para el cálculo del percentil h -ésimo, P_h , se sigue el siguiente procedimiento:

- a) Se ubica la clase del percentil h , que es aquella que contiene la frecuencia acumulada $n * \left(\frac{h}{100} \right)$.
- b) Una vez ubicada la clase del percentil h , mediante un proceso de interpolación se puede obtener la siguiente fórmula para el cálculo de los percentiles:

$$P_h = LI_p + \left[\frac{n * \left(\frac{h}{100} \right) - F_{ap}}{f_p} \right] * C_p$$

en donde,

LI_p : Límite inferior de la clase del percentil h .

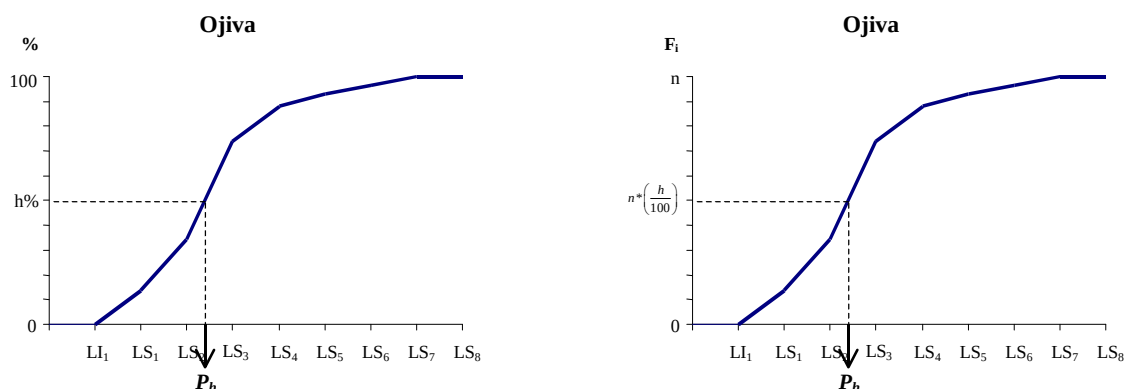
n : No. total de observaciones o datos

F_{ap} : Frecuencia acumulada anterior a la clase del percentil h .

f_p : Frecuencia absoluta de la clase del percentil h .

C_p : Amplitud de la clase del percentil h .

Como se vio antes con la mediana, se pueden obtener gráficamente los percentiles utilizando la ojiva:



Nótese que por el mismo método de interpolación gráfico de la ojiva para distribuciones de frecuencias con intervalos; si se conoce algún valor de los datos, digamos P_h , entonces puede ser encontrada la proporción (o porcentaje) de datos, h , que son menores (o puede ser, mayores o iguales) que el valor P_h . Simplemente, despejando h de la fórmula para calcular percentiles en distribuciones de frecuencias. Así de ese despeje queda que:

$$h = \left[(P_h - LI_p) \frac{f_p}{C_p} + F_{ap} \right] \frac{100}{n}$$

Ejercicio:

- i. Obtenga la fórmula anterior:
 - a. Despejándola de la fórmula para el cálculo de los percentiles.
 - b. Deduciéndola mediante el método gráfico de interpolación, con la ojiva.
 - ii. Del ejercicio para la ojiva de la página 44, obtenga las respuestas de los ítems 1 a 4, usando la última fórmula.
2. Si las clases de la distribución de frecuencias son *valores individuales* de la variable en estudio, se procede similarmente a como se hizo con la mediana. En este caso, no hace falta hacer alguna interpolación. Se puede encontrar cualquier percentil mediante la definición.

Ejercicio:

Calcular los cuartiles Q_1 y Q_3 , el decíl D_9 y el percentil P_{90} en las distribuciones de frecuencias de las variables peso y estatura usando el método algebraico y el método gráfico.

Ejercicio:

Calcular los percentiles P_{15} y P_{80} para las distribuciones de frecuencias correspondientes a las variables número de hermanos y número de visitas al cine usando el método algebraico y el método gráfico.

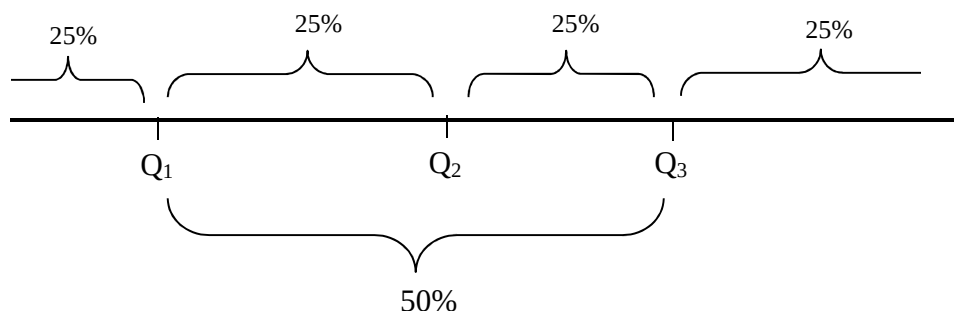
Los percentiles también son utilizados como indicadores de la dispersión de los datos. Con ellos se construyen algunas medidas de dispersión. Veamos algunas de ellas:

Recorrido Intercuartil

El *recorrido intercuartil*, viene dado por:

$$RQ = Q_3 - Q_1$$

Esta medida refleja la dispersión de la parte central de la distribución ya que toma en cuenta al 50% de los datos del centro de la distribución:



Desviación Cuartil ó Recorrido Semi-Intercuartil

La *desviación cuartil* se obtiene mediante la siguiente expresión:

$$Q = \frac{Q_3 - Q_1}{2}$$

Si se calcula:

$$Md \pm Q$$

se obtiene un intervalo que contiene aproximadamente el 50% de los datos.

Fácilmente puede notarse que las dos medidas anteriores no toman en cuenta a todos los datos, lo cual puede representar una seria desventaja ya que es posible que por debajo de Q_1 o por encima de Q_3 , los datos se encuentren muy concentrados o muy dispersos y el efecto sobre RQ y Q será el mismo. Aunque por otro lado, y por la misma razón, el recorrido intercuartil y la desviación cuartil no son afectados por valores atípicos.

Recorrido Percentil

Es una medida basada en la misma idea que el RQ , la cual viene dada por:

$$RP = P_{90} - P_{10}$$

Este indicador refleja el 80% de los datos ubicados en la parte central de la distribución

Ejercicio:

Para las distribuciones de frecuencia correspondientes a las variables peso e ingreso hallar:

- a) RQ
- b) RP
- c) El intervalo que contiene aproximadamente el 50% de los datos de la parte central de la distribución.

Medidas de Forma

En una sección anterior se examinó la forma en que se distribuyen los datos analizando el respectivo gráfico. Se observó la simetría (o asimetría) que presentan los datos y también se podía percibir el grado de apuntamiento (o achatamiento) del gráfico que representa la distribución de los datos.

Existen indicadores que cuantifican la asimetría y el apuntamiento de una distribución, los cuales son de utilidad cuando no se dispone del gráfico o para confirmar las conclusiones obtenidas gráficamente.

Tanto las medidas de asimetría como las de apuntamiento son indicadores relativos ya que no vienen expresados en alguna unidad de medida.

1. Medidas de Asimetría

Los resultados que se discutirán se refieren a distribuciones unimodales:

a. Coeficiente de Asimetría de Pearson

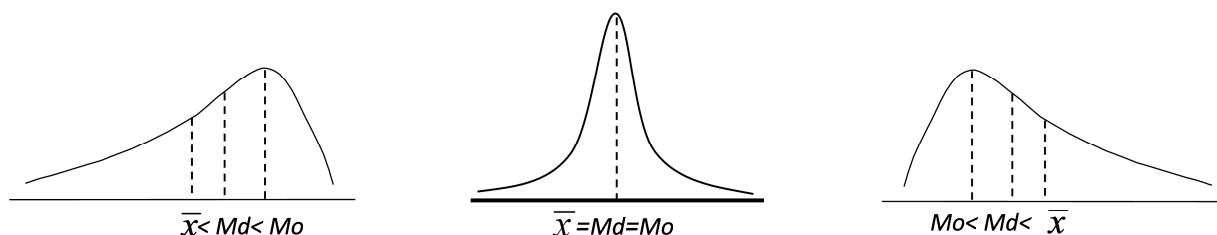
Este indicador se basa en la relación existente entre la media y la mediana:

$$ASP = \frac{3(\bar{x} - Md)}{S_*}$$

Obsérvese que si la distribución es:

- ♦ Simétrica $\Rightarrow ASP = 0$, ya que en este caso $\bar{x} = Md$
- ♦ Asimétrica por la derecha $\Rightarrow ASP > 0$, debido a que $\bar{x} > Md$
- ♦ Asimétrica por la izquierda $\Rightarrow ASP < 0$, porque $\bar{x} < Md$

El coeficiente de asimetría de Pearson toma valores en el intervalo $(-3, 3)$



b. Coeficiente de Asimetría de Fisher

Se denota por γ_1 y viene dado por la siguiente fórmula:

- ♦ Datos no agrupados:

$$\gamma_1 = \frac{\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n} \right)}{S_*^3}$$

- ♦ Datos agrupados:

$$\gamma_1 = \frac{\left(\frac{\sum_{i=1}^k (m_i - \bar{x})^3 f_i}{n} \right)}{S_*^3}$$

El coeficiente γ_1 está basado en la media aritmética e indica de que lado las diferencias respecto de éstas son mayores.

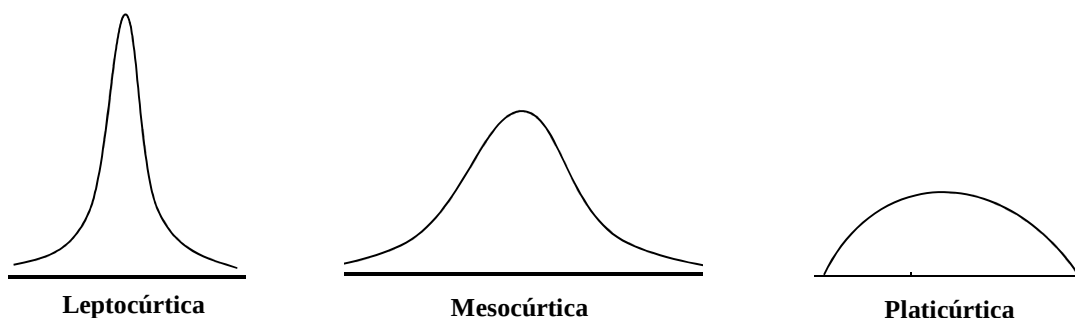
Su interpretación es similar a la del coeficiente de Asimetría de Pearson

Ejercicio:

Calcule e interprete los coeficientes de asimetría ASP y γ_1 para las distribuciones de frecuencias correspondientes a las variables peso, estatura, ingreso y número de hermanos.

2. Medidas de Apuntamiento o Curtosis

Estas medidas indican el grado de apuntamiento o achatamiento del gráfico. La medición del apuntamiento de un gráfico se hace tomando como referencia la curva normal (curva de campana o curva de Gauss). Por tanto, el gráfico al que se le desee medir el apuntamiento, debe ser al menos aproximadamente simétrico. A la curva normal se le llama mesocúrtica, si es más puntiaguda se le llama leptocúrtica y si es más achatada platicúrtica.



Nótese que los indicadores de curtosis, miden el nivel de concentración de datos en la región central.

Coefficiente de Pearson

El coeficiente β_2 de Pearson es el más utilizado de las medidas de apuntamiento y viene dado por:

- ♦ Datos no agrupados:

$$\beta_2 = \frac{\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n} \right)}{S_*^4}$$

- ♦ Datos agrupados:

$$\beta_2 = \frac{\left(\frac{\sum_{i=1}^k (m_i - \bar{x})^4 f_i}{n} \right)}{S_*^4}$$

- Si la curva es normal (mesocúrtica), $\beta_2 = 3$
- Si la curva es leptocúrtica, $\beta_2 > 3$
- Si la curva es platicúrtica, $\beta_2 < 3$

Ejercicio:

Calcule e interprete el coeficiente β_2 de Pearson para las distribuciones de frecuencia correspondientes a las variables peso, estatura y número de hermanos.

DIAGRAMAS DE CAJA

El diagrama de tallo y hoja y el histograma proporcionan una impresión visual general del conjunto de datos, mientras que las cantidades numéricas tales como $\bar{X}$ o S brindan información sobre una sola característica de los datos. El **diagrama de caja** es una presentación visual que describe al mismo tiempo varias características importantes de un conjunto de datos, tales como el centro, la dispersión, la simetría o asimetría y la identificación de observaciones atípicas.

El diagrama de caja representa los tres cuartiles, y los valores mínimo y máximo de los datos sobre un rectángulo (caja), alineado horizontal o verticalmente.

Construcción:

1. El rectángulo delimita el rango intercuartílico con la arista izquierda (o inferior) ubicada en el primer cuartil Q_1 , y la arista derecha (o superior) en el tercer cuartil Q_3 .
2. Se dibuja una línea a través del rectángulo en la posición que corresponde al segundo cuartil (que es igual al percentil 50 o a la mediana), $Q_2 = Md$.
3. De cualquiera de las aristas del rectángulo se extiende una línea, o *bigote*, que va hacia los valores extremos (valor mínimo y valor máximo). Estas son observaciones que se encuentran entre cero y 1.5 veces el rango intercuartílico a partir de las aristas del rectángulo.
4. Las observaciones que están entre 1.5 y 3 veces el rango intercuartílico a partir de las aristas del rectángulo reciben el nombre de *valores atípicos*. Las observaciones que están más allá de tres veces el rango intercuartílico a partir de las aristas del rectángulo se conocen como *valores atípicos extremos*. En ocasiones se emplean diferentes símbolos (como círculos vacíos o llenos), para identificar los dos tipos de valores atípicos.

A veces, los diagramas de caja reciben el nombre de *diagramas de caja y bigotes*. Nótese que el rectángulo o caja representa el 50% de los datos que particularmente están ubicados en la zona central de la distribución. La caja representa el cuerpo de la distribución y los bigotes sus colas.

La Figura 1 presenta esquemáticamente un diagrama de caja indicando sus partes. Del diagrama se interpreta que la distribución de los datos es asimétrica por la derecha, ya que la longitud de los rectángulos por debajo y por encima de la mediana así como los bigotes indican que los datos están más agrupados en sus valores inferiores que en los superiores y además se observa que $\bar{X} > Md$. También destaca la existencia de dos valores atípicos en el extremo superior de los datos.

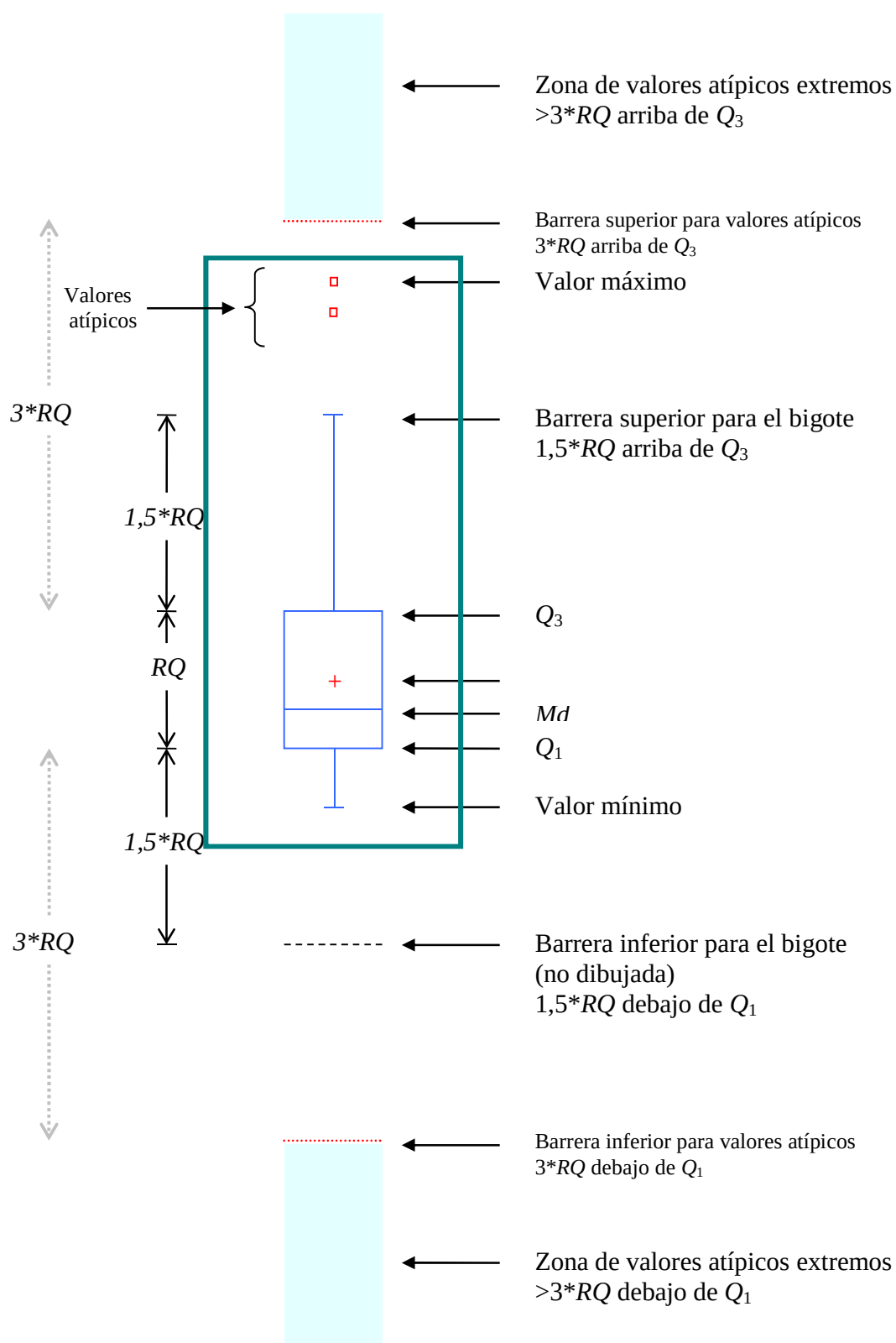
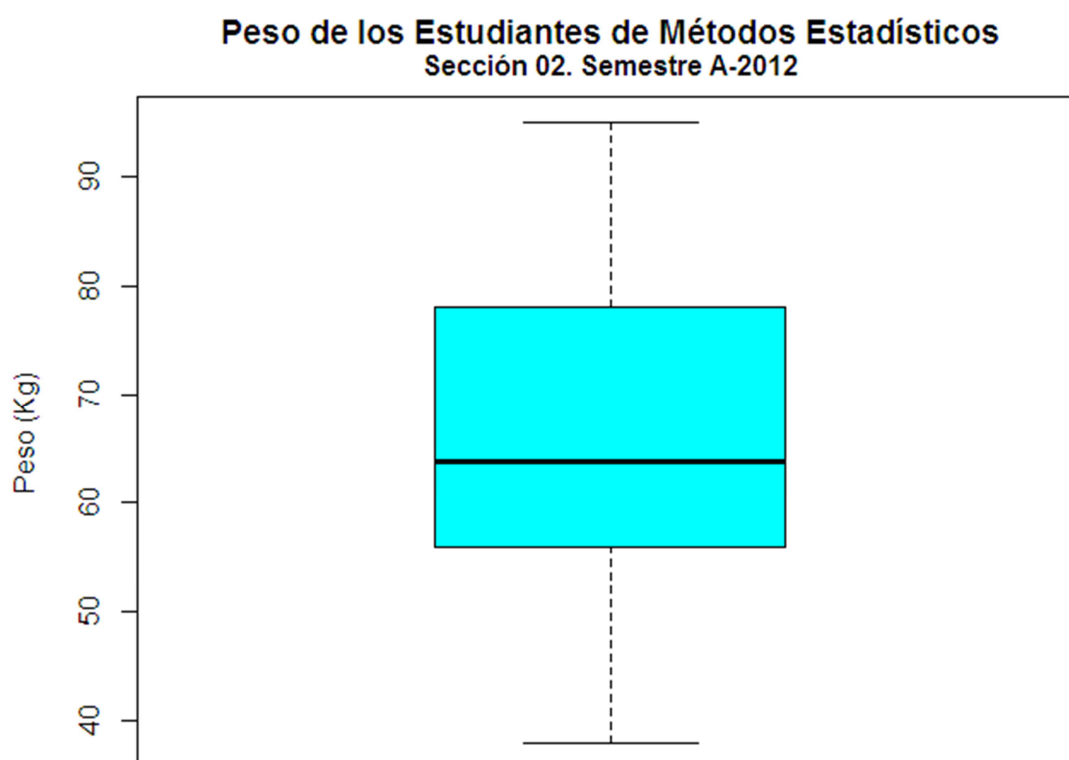


Figura 1. Partes de un diagrama de Caja

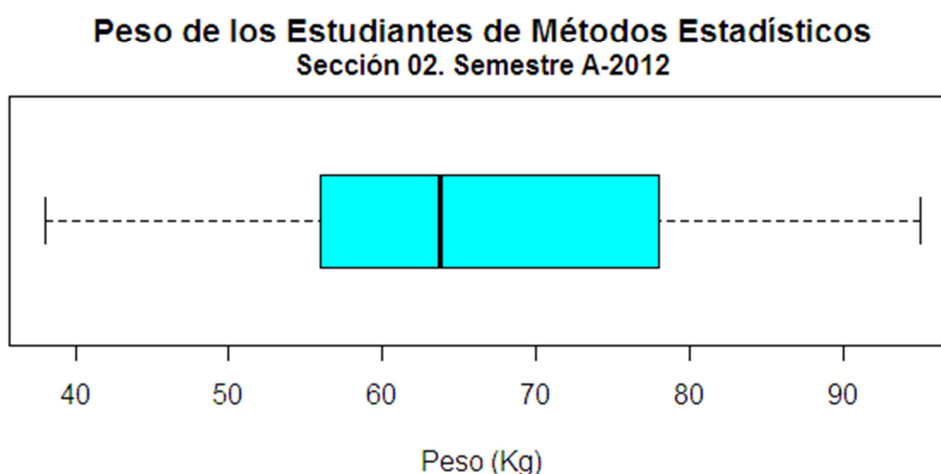
En la Figura 2, se muestra el diagrama de caja para la variable *peso* de los estudiantes de Métodos Estadísticos I, sección 02 cursantes del semestre A-2012. Analizando este diagrama se observa que la distribución de los pesos es asimétrica por la derecha, no existen valores atípicos y que por debajo del primer cuartil se encuentra aproximadamente la misma cantidad de datos que por arriba del tercer cuartil. Asimismo, se nota que la mitad de los pesos correspondientes a la parte central de su distribución, se encuentran entre un valor cercano a los 60 kilos y un valor cercano a los 80 kg. También se puede observar que el rango de los pesos varía entre un valor mínimo cercano a los 40 kg y un valor máximo cercano a los 100 kg. Este diagrama de caja fue generado mediante el uso del software estadístico R.



Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas-FACES-ULA. Abril 2012

Figura 2. Diagrama de caja (vertical) para los datos de peso de los estudiante de Métodos Estadísticos I Sem-A2012

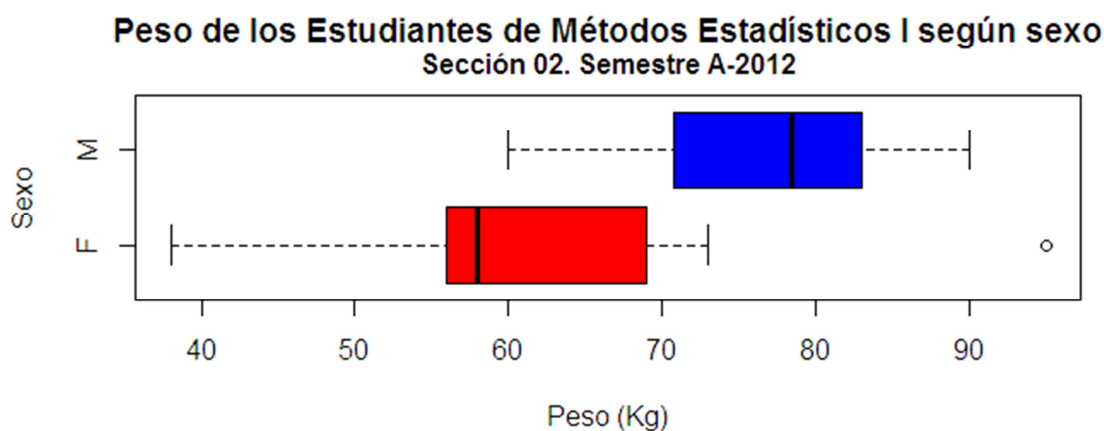
Nótese en la Figura 3 el mismo diagrama de caja para los datos de pesos, pero ahora dibujado de forma horizontal. Los análisis obtenidos con el diagrama de caja vertical son los mismos que se obtendrían al analizar el diagrama orientado horizontalmente.



Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas-FACES-ULA. Abril 2012

Figura 3. Diagrama de caja (horizontal) para los datos de peso de los estudiante de Métodos Estadísticos I Sem-A2012

Los diagramas de caja son muy útiles al hacer comparaciones gráficas entre conjuntos de datos, ya que tienen un gran impacto visual y son fáciles de comprender. Por ejemplo, la Figura 4 presenta los diagramas de caja comparativos para la variable *peso* de los estudiantes de Métodos Estadísticos I clasificados por sexo. El examen de este diagrama revela que el peso de los varones es mayor que el de las hembras. También se observa que la variabilidad de los pesos de las hembras es mayor a la de los varones. Sin embargo, la variabilidad en la parte central de la distribución de los pesos tanto de las féminas como de los masculinos es muy similar. Se nota la existencia de un valor atípico en la distribución de las mujeres, que es un peso muy alto (el valor máximo de todos los pesos) en comparación a los pesos del resto de las muchachas. La distribución del peso de los varones es asimétrico por la izquierda mientras que las hembras presentan una distribución asimétrica por la derecha influenciada por el valor atípico.



Fuente: Encuesta realizada por la cátedra de Estadísticas Básicas-FACES-ULA. Abril 2012

Figura 4 Diagramas de Caja comparativos para la variable Índice Académico clasificado por Sexo

Los datos de la Figura 5, muestran diferentes tipos de distribuciones. Se colocan de manera comparativa los diagramas de caja con los histogramas del mismo conjunto de datos.

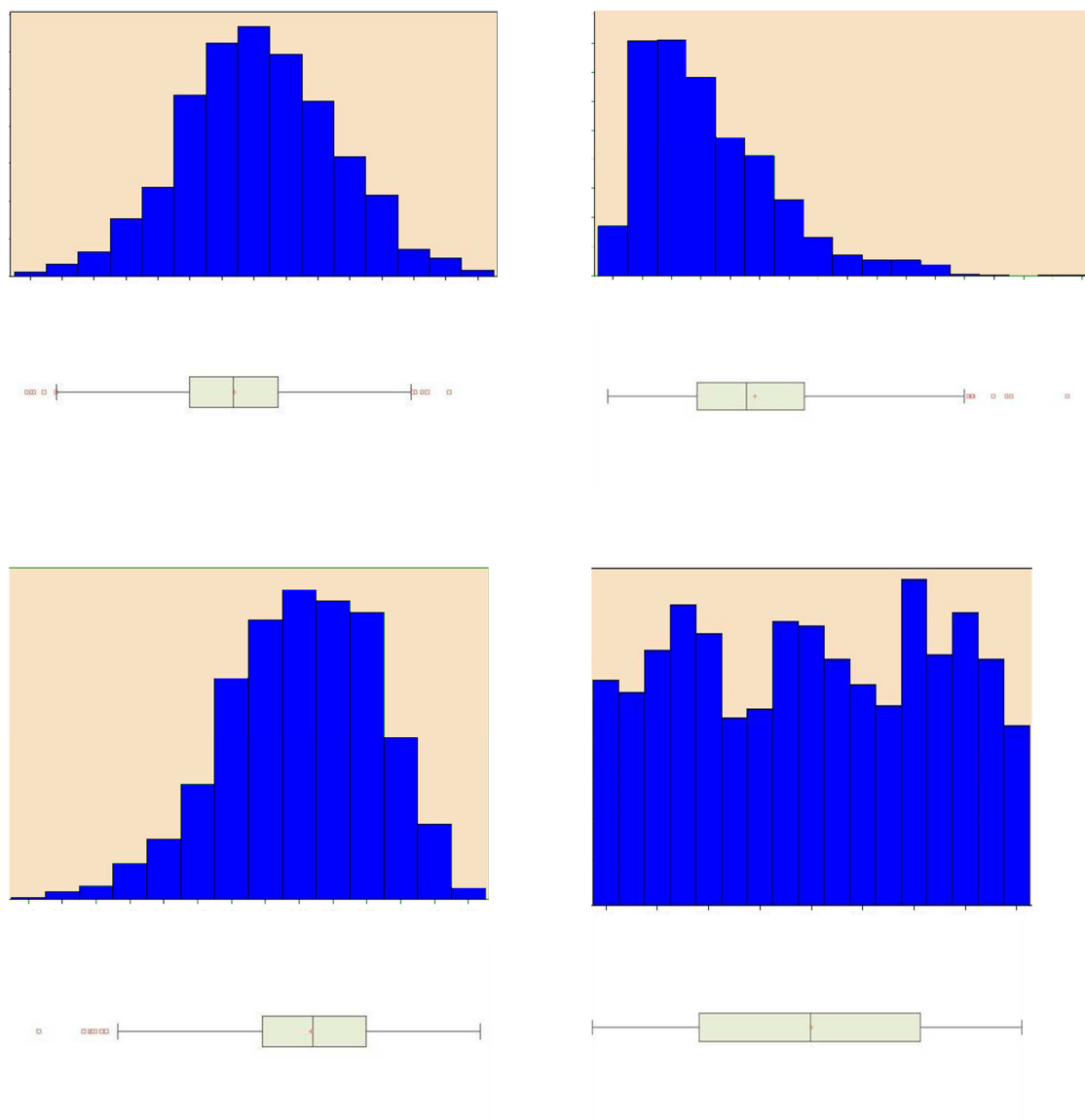


Figura 5. Histogramas y Diagramas de Caja para 1000 observaciones de cuatro distribuciones