Parámetros Estadísticos

En el tema anterior, hemos visto una primera forma de reducir la complejidad de los datos estadísticos de una distribución, mediante la construcción de tablas y representaciones gráficas.

Las **tablas estadísticas** son una forma organizada de dar toda (o casi toda) la información, todos los datos de que disponemos. Con las **gráficas estadísticas** se pierde algo de información (mucho o poco, según los casos), pero el mensaje "entra por los ojos", que es lo que se pretende.

En cualquiera de los casos, la cantidad de datos que se dan es excesiva para que sea operativo, para poder hacer referencias concisas a esa distribución o comparaciones rápidas con otras distribuciones.

Esta es la razón de ser de los **parámetros estadísticos**, ***el resumir en un número un aspecto relevante de la distribución que pueda dar una idea de la misma o compararla en ese aspecto, con otras.***

Evidentemente, todo proceso de síntesis conlleva una pérdida de información; pero se gana en el hecho de que es más fácil trabajar con unos pocos parámetros con significado muy preciso que con la totalidad de los datos.

Los parámetros estadísticos suelen clasificarse, según el papel que juegan, en varios tipos:

Medidas de centralización	O de tendencia central. Las más importantes son: **moda** (*el valor que se presenta con mayor frecuencia*), **media aritmética** (*suma de todos los valores de una variable estadística dividido por el número de valores*), **media geométrica**, **media armónica** y **mediana** (*el valor del individuo que ocuparía el lugar central si se colocaran ordenados de menor a mayor*).
Medidas de dispersión	**Desviación media, varianza, desviación típica,...** *Sirven para medir el grado de alejamiento (dispersión) de los datos.*
Medidas de posición	**Cuartiles, deciles, percentiles:** *Señalan la situación de algunos valores importantes en la distribución*. Por ejemplo, los cuartiles son los valores que dejan a cada lado el 25% y el 75% de los demás.
Medidas de asimetría	*Con los coeficientes de asimetría se trata de medir si las observaciones están dispuestas simétrica o asimétricamente respecto a un valor central* (en general, la media aritmética) y el grado de esta asimetría.
Medidas de apuntamiento	O **curtosis,** *indican si la distribución es más o menos puntiaguda*.

Medidas de centralización

Se llama **medidas de centralización** a las medidas o parámetros que pretenden reflejar "en torno a qué valores se agrupan los datos reflejados", "qué valores son los más frecuentes", etc. A las medidas de centralización también se las llama medidas de tendencia central o promedios.

Las medidas de tendencia central más importantes son: **media aritmética, media geométrica, media armónica, mediana y moda.**

Media aritmética

Se llama media aritmética de una variable estadística a la suma de todos los valores de dicha variable dividido por el número de valores.

Sea x una variable estadística que toma los valores $x_1, x_2, x_3, ..., x_n$ con frecuencias absolutas $f_1, f_2, f_3, ..., f_n$ respectivamente. La media aritmética de la variable x se representa por $\overline{x}$, y viene dada por la siguiente expresión:

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots\cdots\cdots + x_n f_n}{f_1 + f_2 + \cdots\cdots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

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

Caso discreto

Ejemplo: Las calificaciones en la asignatura Historia del Arte de los 40 alumnos de una clase viene dada por la siguiente tabla:

Calificaciones	1	2	3	4	5	6	7	8	9
Número de alumnos	2	2	4	5	8	9	3	4	3

Hallar la calificación media.

En la práctica, los cálculos se disponen de la siguiente forma:

x_i	f_i	$x_i f_i$
1	2	2
2	2	4
3	4	12
4	5	20
5	8	40
6	9	54
7	3	21
8	4	32
9	3	27
	40	212

La media aritmética será:

$$\overline{x} = \frac{1\cdot 2+2\cdot 2+3\cdot 4+4\cdot 5+5\cdot 8+6\cdot 9+7\cdot 3+8\cdot 4+9\cdot 3}{40} = \frac{212}{40} = 5'3$$

Luego la calificación media en Historia del Arte es 5'3

Caso continuo

Si la variable x es continua, o aún siendo discreta, y por tratarse de muchos datos se encuentran agrupados en clases, se toman como valores $x_1, x_2, x_3, \ldots\ldots, x_n$ las marcas de clase.

Ejemplo: **Ejemplo:Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:**

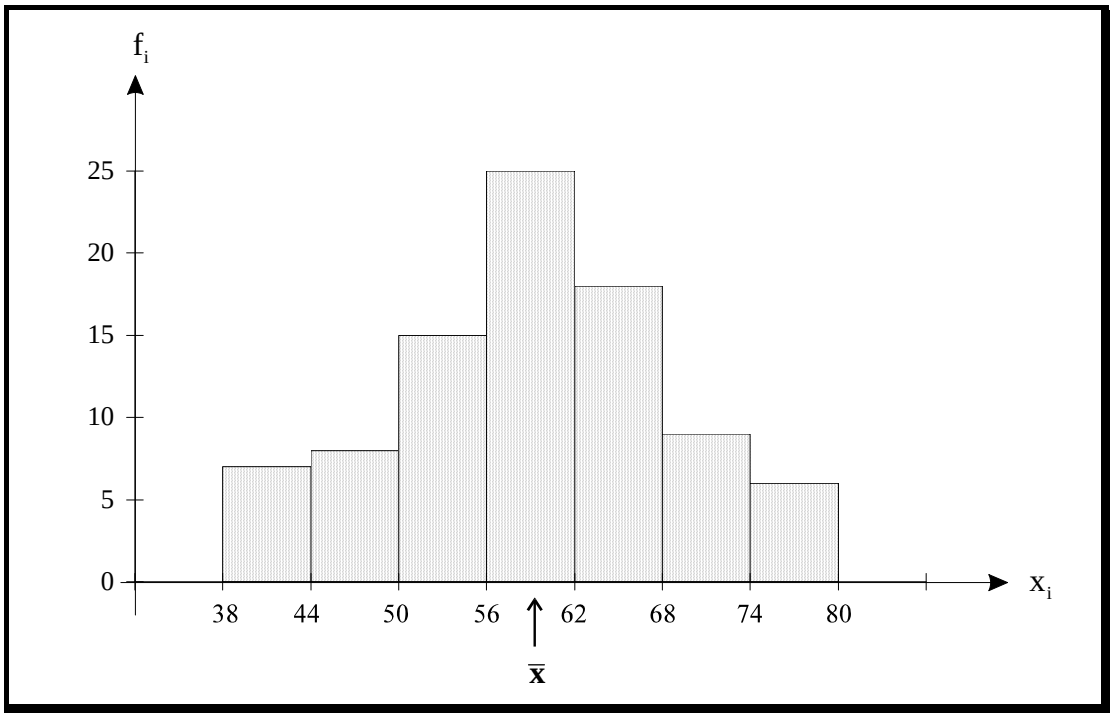
Puntuaciones	**Nú mero de trabajadores**
$[38-44)$	**7**
$[44-50)$	**8**
$[50-56)$	**15**
$[56-62)$	**25**
$[62-68)$	**18**
$[68-74)$	**9**
$[74-80)$	**6**

Calcular la puntuación media.

Clases o intervalos	Marcas de clase x_i	Frecuencias f_i	$x_i f_i$
$[38-44)$	41	7	287
$[44-50)$	47	8	376
$[50-56)$	53	15	795
$[56-62)$	59	25	1475
$[62-68)$	65	18	1170
$[68-74)$	71	9	639
$[74-80)$	77	6	462
		88	5.204

La puntuación media será

$$\bar{x} = \frac{5.204}{88} = 59'14$$



Observaciones:

- Geométricamente, la media aritmética está en el punto del eje x situado bajo "el centro de gravedad del histograma", es decir, en el punto donde debería apoyarse el histograma, para mantenerse en equilibrio.

- La media aritmética es la medida o parámetro de centralización que más se utiliza.

- Puede expresarse en las mismas unidades que la variable estudiada. Es más: Si cambiamos de unidad los valores de la variable, el cambio de valor experimentado por la media aritmética equivale a efectuar el mismo cambio de unidad con ella.

- Presenta la ventaja de tener en cuenta todos los datos de la distribución, además de resultar muy sencillo su cálculo.

- Tiene el grave inconveniente de que si la distribución posee valores extremos, excepcionalmente raros y poco significativos, éstos producen una distorsión sobre el valor de la media, alterando el significado de ésta. Por ejemplo, si se trata de hallar la talla media de los alumnos de una clase y hay un alumno de 2'05 m, este valor alterará considerablemente la talla media de la clase.

- No siempre es posible realizar el cálculo de la media aritmética. Por ejemplo:

  – Si los datos de la distribución no son cuantitativos sino cualitativos.

  – Cuando los datos de la distribución se encuentran agrupados en clases, estando algunas de ellas abiertas. Por ejemplo, en una encuesta sobre lectores de la prensa diaria, se obtuvo la siguiente distribución:

Grupos de edad	Número de personas
Menores de 18 años	264
Entre 18 y 40 años	1.367
Entre 40 y 60 años	825
Mayores de 60 años	341

En estos casos en los que no siempre es posible calcular la media, se utilizan otras medidas o parámetros de centralización, como por ejemplo la moda y la mediana.

Un error frecuente en el cálculo de la media. Media ponderada.

Hasta aquí hemos considerado todos los datos con la misma importancia, es decir, como si todos ellos tuvieran la misma fiabilidad. No obstante, puede suceder que en algún caso, debido al criterio adoptado o a las circunstancias en que se obtuvieron los datos, sea necesario dar más importancia a unos datos que a otros. En este caso, la media se llama **media ponderada** y su expresión es:

$$\overline{x} = \frac{x_1 f_1 a_1 + x_2 f_2 a_2 + \cdots\cdots\cdots + x_n f_n a_n}{f_1 a_1 + f_2 a_2 + \ldots\ldots\ldots\ldots + f_n a_n} = \frac{\sum_{i=1}^{n} x_i f_i a_i}{\sum_{i=1}^{n} f_i a_i}$$

$$\boxed{\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i a_i}{\sum_{i=1}^{n} f_i a_i}}$$

donde a_i son las distintas ponderaciones o pesos que se adjudican a los datos.

Ejemplo: **En un curso de C.O.U. los alumnos durante un período evaluativo han realizado las siguientes pruebas: un examen, dos controles y tres intervenciones en clase. Las pruebas, según acuerdo del Seminario, se valoran de la siguiente forma: 50% el examen, 30% los controles y 20% las intervenciones. Si un determinado alumno ha obtenido: 7 en el examen, 6 y 8 en los controles y 10, 5 y 2 en las intervenciones, obtener su nota media de la evaluación.**

$$\overline{x} = \frac{7 \cdot 0'5 + (6+8) \cdot \frac{0'30}{2} + (10+5+2) \cdot \frac{0'2}{3}}{1 \cdot 0'5 + 2 \cdot \frac{0'30}{2} + 3 \cdot \frac{0'20}{3}} = \mathbf{6'73}$$

Si en este mismo ejemplo hubiésemos considerado todas las notas con la misma importancia, la nota media obtenida sería:

$$\overline{x} = \frac{7+6+8+10+5+2}{6} = 6'33$$

Ejemplo: **En la tabla siguiente se da el consumo de combustible líquido, en litros por habitante, de los paises de la CEE en 1983 (Estadísticas de la Energía, OCDE 1985).**

País	Litros por habitante
Luxemburgo	2.518
Dinamarca	1.813
Bélgica	1.423
Alemania F.	1.415
Italia	1.258
Francia	1.234

País	Litros por habitante
Irlanda	1.076
G.Bretaña	1.031
Grecia	989
Holanda	820
España	814
Portugal	754

Hallar la media del consumo de combustible líquido por habitante para el global de la CEE.

Si calculamos el consumo en toda la CEE como media de los consumos en cada uno de los 12 paises miembros, se tiene:

$$\overline{x} = \frac{2518+1813+1423+1415+1258+1234+1076+1031+989+820+814+754}{12} =$$

$$1.262'08\ \frac{\text{litros}}{\text{habitante}}$$

Sin embargo, esta media ha sido erróneamente calculada, pues no se ha tenido en cuenta que, por ejemplo, hay muchos más alemanes que luxemburgeses, es decir, no se ha tenido en cuenta el peso de la población de cada país. La media correcta se llama **media ponderada** (de pondus = peso) de los datos según la población de cada país y daría la siguiente tabla:

País	Población (en millones)
Luxemburgo	0'4
Dinamarca	5'1
Bélgica	9'9
Alemania F.	61'4
Italia	56'6
Francia	54'7

País	Población (en millones)
Irlanda	3'5
G. Bretaña	55'6
Grecia	9'8
Holanda	14'4
España	38'1
Portugal	10'1

siendo el total de la población de la CEE 319'6 millones de habitantes.

La media ponderada se calcula multiplicando cada valor por los pesos (en este caso, poblaciones) respectivos, sumando estos productos y dividiendo por la suma de los pesos (en este caso población total).

$$\overline{x} = \frac{2518\cdot 0'4 + 1813\cdot 5'1 + 1423\cdot 9'9 + 1415\cdot 61'4 + 1258\cdot 56'6 + 1234\cdot 54'7 + }{0'4 + 5'1 + 9'9 + 61'4 + 56'6 + 54'7 + 3'5 + 55'6 + 9'8 + 14'4 + 38'1 + 10'1}$$

$$\frac{1076\cdot 3'5 + 1031\cdot 55'6 + 989\cdot 9'8 + 820\cdot 14'4 + 814\cdot 38'1 + 754\cdot 10'1}{} = 1161'26\,\frac{\text{litros}}{\text{hab}}$$

Uso de la calculadora para la obtención de $\overline{x}$

Los pasos a seguir son los siguientes:

1) Procura que la calculadora se encuentre en disposición de efectuar cálculos estadísticos. En tal caso suele presentar en la parte alta de la pantalla la notación **SD**. En cada modelo, esto se consigue de un modo distinto.

2) Comprueba que no hay nada acumulado. Para ello pulsa la tecla [**n**]. Si sale 0 en la pantalla, estás en condiciones de acumular los datos. Si no, borra lo que hay en memoria mediante la secuencia [**INV**] [**AC**].

3) Acumulación de datos:

1^{er} dato [**x**] 1^{a} frecuencia [**M+**] o [**DATA**]

2^{0} dato [**x**] 2^{a} frecuencia [**M+**] o [**DATA**]

Así sucesivamente hasta haber cargado todos los datos.

4) Pulsando cualquiera de las teclas [**n**] [$\sum x^2$] [$\sum x$] [$\bar{x}$] obtendremos el valor correspondiente, y esta consulta puede hacerse en cualquier momento del proceso. Después, si se quiere, se puede seguir introduciendo datos.

La tecla [**n**] nos da la suma de las frecuencias absolutas y por tanto equivale a $\sum_{i=1}^{n} f_i$. La tecla [$\bar{x}$] nos da la media.

5) Si se introduce erróneamente algún dato, se puede suprimir del siguiente modo:

Dato erróneo [**INV**] [**M+**] o [**DATA**]

Cálculo de $\sum_{i=1}^{n} x_i f_i$ usando la tabla de frecuencias

1) La calculadora debe de estar **en modo DEG** pero **no en modo SD**. Antes de comenzar, comprueba que no hay ningún dato acumulado en la memoria. Si lo hay, bórralo con la secuencia [**AC**] [**Min**] (modelos CASIO). Introduciremos los datos de la siguiente manera:

1^{er} dato [**x**] 1^{a} frecuencia [**M+**]	Aparece en la pantalla el resultado del primer producto, que anotaremos en la columna correspondiente.
2° dato [**x**] 2^{a} frecuencia [**M+**]	Aparece en la pantalla el resultado del segundo producto, que anotaremos debajo del anterior.

y así sucesivamente hasta haber ido anotando todos los productos.

2) Pulsando la tecla **MR** obtenemos la suma de todos los productos, es decir $\sum_{i=1}^{n} \mathbf{x_i f_i}$

Propiedades de la media

Las dos propiedades que se citan a continuación, además de simplificar los cálculos son de aplicación posterior.

1. **Si se suma una constante a todos los valores x_i, la media aumenta en el mismo número.**

   En efecto, si los nuevos valores son $x'_i = x_i + k$, se tiene que la nueva media, $\overline{x}'$, será:

$$\overline{x}' = \frac{\sum_{i=1}^{n}(x_i + k)\cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i f_i + \sum_{i=1}^{n} k f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} + \frac{k\cdot\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} f_i} = \overline{x} + k$$

2. **Si se multiplican todos los valores de la variable x_i por el mismo número, la media queda multiplicada por el mismo número.**

   Si tomamos $x'_i = k \cdot x_i$, la nueva media, $\overline{x}'$ será:

$$\overline{x}' = \frac{\sum_{i=1}^{n}(kx_i)\cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} kx_i f_i}{\sum_{i=1}^{n} f_i} = \frac{k\cdot\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = k\cdot\overline{x}$$

Media geométrica

La media geométrica se define como:

$$G = \sqrt[k]{x_1^{f_1}\cdot x_2^{f_2}\cdot\cdots\cdots x_n^{f_n}} = \sqrt[k]{\prod_{i=1}^{n} x_i^{f_i}}$$

$$\mathbf{G = \sqrt[k]{\prod_{i=1}^{n} x_i^{f_i}}}$$

siendo $k = \sum_{i=1}^{n} f_i$

El cálculo de la media geométrica resulta más fácil si en la expresión anterior se toman logaritmos:

$$\log G = \frac{1}{k} \cdot \left[f_1 \cdot \log x_1 + f_2 \cdot \log x_2 + \cdots\cdots\cdots + f_n \cdot \log x_n \right]$$

Es decir, el logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.

Ejemplo: Calcula la media geométrica de la siguiente distribución:

x_i	f_i	$\log x_i$	$f_i \cdot \log x_i$
1	**2**	**0**	**0**
2	**5**	**0'3010**	**1'5050**
3	**3**	**0'4771**	**1'4313**
	10		**2'9363**

$$\log G = \frac{\sum_{i=1}^{3} f_i \cdot \log x_i}{\sum_{i=1}^{3} f_i} = \frac{2'9363}{10} = 0'2936 \quad \Rightarrow \quad G = 10^{0'2936} = 1'96607$$

Media armónica

La media armónica se define como:

$$H = \frac{f_1 + f_2 + \cdots\cdots\cdots + f_n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \cdots\cdots\cdots + \frac{f_n}{x_n}} = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

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

Ejemplo: Calcula la media armónica de la siguiente distribución:

x_i	f_i	$\frac{f_i}{x_i}$
1	2	2
2	5	2'5
3	3	1
	10	5'5

$$H = \frac{\sum_{i=1}^{3} f_i}{\sum_{i=1}^{3} \frac{f_i}{x_i}} = \frac{10}{5'5} = 1'82$$

Moda

Se llama moda de una variable estadística al valor de dicha variable que presenta mayor frecuencia absoluta. Se representa por M_0.

La moda no tiene por qué ser única, puede haber varios valores de la variable con la mayor frecuencia. En este caso se dirá que la distribución es bimodal, trimodal, etc., según que sean 2, 3, etc, los valores de la variable que presentan mayor frecuencia. También se aplica este nombre a distribuciones en las que destacan varios valores con frecuencias muy altas, prácticamente iguales, aunque no todas sean máximas.

Variable estadística discreta

Ejemplo: Calcular la moda de la distribución

x_i	1	2	3	4
f_i	2	3	5	2

La Moda es $M_0 = 3$ ya que su frecuencia $f_i = 5$ es la mayor de la distribución.

Ejemplo: Calcular la moda de la distribución

x_i	16	17	18	19	20	21
f_i	1	8	3	2	8	2

Esta distribución es bimodal, ya que presenta dos modas:

$$M_0 = 17 \qquad M_0 = 20$$

que corresponden a dos valores que presentan idéntica frecuencia (8), que es la mayor de la distribución

Variable estadística continua

Ahora bien, en el caso de datos agrupados en intervalos tenemos que distinguir entre intervalos de amplitud constante e intervalos de amplitud variable.

Intervalos de amplitud constante:

En el caso de datos agrupados en intervalos es fácil determinar la **clase modal** (clase con mayor frecuencia), pero el valor dentro del intervalo que se presume tenga mayor frecuencia se obtiene a partir de la siguiente expresión:

$$\mathbf{M_0 = L_i + c \cdot \frac{D_1}{D_1 + D_2}}$$

L_i = límite inferior de la clase modal.

c = amplitud de los intervalos.

D_1 = diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase anterior.

D_2 = diferencia entre la frecuencia absoluta de la clase modal y la frecuencia absoluta de la clase siguiente.

Ejemplo: Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:

Puntuaciones	Nú mero de trabajadores
$[38-44)$	7
$[44-50)$	8
$[50-56)$	15
$[56-62)$	25
$[62-68)$	18
$[68-74)$	9
$[74-80)$	6

Calcular la moda.

La moda está en el intervalo $[56-62$ por ser el que presenta, mayor frecuencia. Al intervalo $[56-62$ que contiene la moda se le llama clase modal o intervalo modal.

Como primera aproximación de la moda se podría tomar la marca de la clase modal, es decir: 59.

Aplicando la expresión anterior y teniendo en cuenta que

$L_1 = 56 \qquad c = 6 \qquad D_1 = 25-15 = 10 \qquad D_2 = 25-18 = 7$

$$M_0 = 56 + 6 \cdot \frac{10}{10+7} = 59'52$$

Este es el valor que, teóricamente, se supone tiene mayor frecuencia.

Intervalos de amplitud variable:

En este caso, tenemos que sustituir la frecuencia de cada intervalo por su correspondiente **densidad de frecuencia**, que como sabemos es el cociente entre la frecuencia y la amplitud del intervalo. Aquí, la clase modal será la clase que tenga mayor densidad de frecuencia

$$\mathbf{M_0 = L_i + c \cdot \frac{d_1}{d_1 + d_2}}$$

L_i = límite inferior de la clase modal.

c = amplitud del intervalo que tiene mayor densidad de frecuencia.

d_1 = diferencia entre la densidad de frecuencia de la clase modal y la densidad de frecuencia de la clase anterior.

d_2 = diferencia entre la densidad de frecuencia de la clase modal y la densidad de frecuencia de la clase siguiente.

Ejemplo: Consideremos la siguiente distribución:

Intervalos	**[0,4)**	**[4,10)**	**[10,20)**	**[20,40)**	**[40,70)**
$\mathbf{f_i}$	**20**	**100**	**180**	**260**	**240**

Construimos la tablas con las densidades de frecuencia

Intervalos	f_i	$d_i = \frac{f_i}{c_i}$
$[0,4)$	20	5
$[4,10)$	100	16'6
$[10,20)$	180	18
$[20,40)$	260	13
$[40,70)$	240	8

El intervalo que tiene mayor densidad de frecuencia es $[10,20$, por tanto

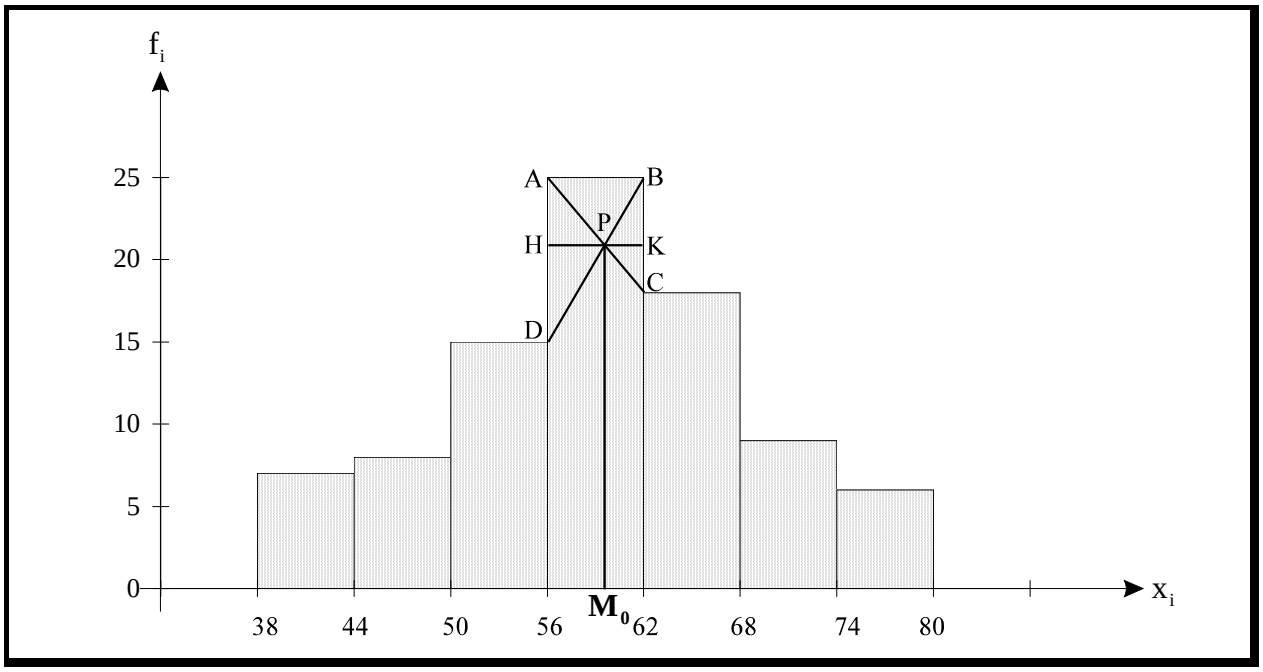
$$M_0 = 10 + 10 \cdot \frac{18 - 16'6}{(18 - 16'6) + (18 - 13)} = 12'19$$

Cálculo de la moda por el método gráfico

Para las distribuciones cuyos datos se encuentran clasificados en intervalos existe un método gráfico que permite obtener la moda con cierta aproximación. Para ello se representa el histograma de frecuencias absolutas a ser posible sobre papel milimetrado con el fin de poder obtener una mayor precisión.

Seguidamente se unen, con líneas de puntos, los extremos de la clase modal con las contiguas como en el diagrama adjunto (rectas AC y BD). La moda M_0 viene dada por la abscisa P del punto de corte.

El diagrama adjunto permite calcular la moda del ejemplo anterior del test sobre satisfación en el trabajo $(M_0 = 59'5)$ por el método gráfico utilizando la semejanza de triángulos.



Los triángulos ADP y CPB son semejantes y por tanto se tiene $\frac{PH}{DA} = \frac{PK}{CB}$

Conocemos $PK = 6 - PH$ $DA = 10$ $CB = 7$

luego

$$\frac{PH}{10} = \frac{6 - PH}{7} \Rightarrow 7PH = 60 - 10PH \Rightarrow PH = \frac{60}{17} = 3'52$$

Por tanto la moda será $M_0 = 56 + 3'52 = 59'52$

Observaciones

- Puede ocurrir que algunas distribuciones no tengan moda; eso ocurre cuando las frecuencias de todos los datos son iguales.

- La moda es menos representativa que la media aritmética, pero en algunas ocasiones no ocurre así. Es útil para describir situaciones en las cuales intervienen variables cualitativas, como por ejemplo el sexo, el grupo sanguíneo o el nivel de estudios. No tiene sentido hablar de la media del sexo o de la media del nivel de estudios de los miembros del gobierno, por ejemplo, y en cambio sí tiene sentido decir que la moda es el sexo masculino y el nivel de estudios el universitario. También tiene interés el valor de la moda en los juegos de azar, en los cuales determinados valores aparecen más veces que otros. Por ejemplo, si en el lanzamiento de dos dados obtenemos las siguientes sumas: 2, 5, 8, 7, 7, 8, 7, 9, 11, 8, 12, 7, 5, 7, está claro que saber cual es la moda es importante, y en cambio, conocer la media no tiene utilidad en un juego de apuestas.

- En la moda no intervienen todos los datos de la distribución.

- Aún cuando la moda se considera una medida o parámetro de centralización, no siempre tiene por qué situarse en la zona central; es frecuente encontrar la moda próxima a los valores extremos de la distribución.

- En definitiva, la moda representa el valor dominante de la distribución; así, por ejemplo, en unas elecciones la moda es el partido más votado.

Mediana

Si los datos de la muestra estudiada se ordenan siguiendo un criterio de crecimiento o decrecimiento, se denomina mediana al valor del dato que ocupa el lugar central, o dicho de otro modo, la mediana es el valor que divide a la serie de datos en dos partes exactamente iguales. Se representa con el símbolo M_e.

Variable estadística discreta. Datos simples.

En este caso se ordenan los datos de menor a mayor y la media será el término o valor central.

- **Si el número de datos es impar, el valor central de la variable es único**.

  Ejemplo: La mediana de la serie estadística 2, 3, 5, 6, 9, 11, 12 es $M_e = 6$

- Si el número de datos es par, no existe término central, sino dos términos centrales, y por ello se dice que hay dos medianas, que son los dos valores centrales. No obstante, acostumbra tomarse como **mediana la semisuma de los dos valores centrales**, aunque dicho valor no pertenezca al conjunto de datos.

  Ejemplo: En la serie estadística 2, 3, 5, 6, 9, 11, 12, 13 es $M_e = \frac{6+9}{2} = 7'5$

Variable estadística discreta. Datos agrupados.

Para la determinación de la mediana se calculan, en primer lugar, las frecuencias acumuladas, F_i, y la mitad de las frecuencias totales. A continuación se observa **cuál es la primera F_i que supera o iguala a la mitad de las frecuencias totales**, distinguiéndose dos casos:

1) **Si ese primer F_i supera a la mitad de las frecuencias totales, entonces la mediana es el valor de la variables x_i que corresponde a dicho F_i.**

   Ejemplo: Las calificaciones en la asignatura de historia del arte de los 40 alumnos de una clase vienen dadas por la siguiente tabla:

Calificaciones	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Número de alumnos	**2**	**2**	**4**	**5**	**8**	**9**	**3**	**4**	**3**

Calculamos las frecuencias absolutas acumuladas:

x_i	f_i	F_i
1	2	2
2	2	4
3	4	8
4	5	$13 < 20$
5	8	$21 > 20$
6	9	30
7	3	33
8	4	37
9	3	40
	40	

La mitad del número de datos es $\frac{40}{2} = 20$

Entonces la mediana es $M_e = 5$, dado que es el primer valor de la variable cuya frecuencia absoluta acumulada, 21, excede a la mitad del número de datos, 20.

2) Si el primer F_i iguala exactamente a la mitad de las frecuencias totales, entonces se toma convencionalmente como mediana la media aritmética de los valores de la variable, x_i y x_{i+1}, que corresponden a dicha F_i y a F_{i+1}.

Ejemplo: Consideremos la siguiente tabla de frecuencias

x_i	**3**	**6**	**7**	**8**	**9**
f_i	**15**	**20**	**15**	**40**	**10**

x_i	f_i	F_i
3	15	15
6	20	35
7	15	$50 \leq 50$
8	40	$90 > 50$
9	10	100
	100	

La mitad del número de datos es $\frac{100}{2} = 50$

Como 50 coincide con la frecuencia absoluta acumulada del valor 7, la mediana vendrá dada por la semisuma de 7 y el valor siguiente 8.

$$M_e = \frac{7+8}{2} = 7'5$$

Variable estadística continua

En este caso, procediendo de forma análoga a como acabamos de hacer, resulta fácil detectar cuál es la **clase mediana (donde se alcanzan la mitad de los datos)**, pero para obtener el valor concreto de la variable que deja a su izquierda igual número de datos que a su derecha, aplicaremos la expresión:

$$M_e = L_i + c \cdot \frac{\left(\frac{\sum f_i}{2}\right) - F_{i-1}}{f_i}$$

L_i = límite inferior de la clase mediana.

c = amplitud del intervalo.

$\sum f_i$ = número total de datos.

F_{i-1} = frecuencia absoluta acumulada de la clase anterior a la clase mediana.

f_i = frecuencia absoluta de la clase mediana.

Ejemplo: Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:

Puntuaciones	Número de trabajadores
$[38-44)$	7
$[44-50)$	8
$[50-56)$	15
$[56-62)$	25
$[62-68)$	18
$[68-74)$	9
$[74-80)$	6

Calcular la mediana.

Para calcular la mediana formamos la tabla estadística con las frecuencias absolutas acumuladas F_i.

Clases	f_i	F_i
$[38-44)$	7	7
$[44-50)$	8	15
$[50-56)$	15	30
$[56-62)$	25	**55** > **44**
$[62-68)$	18	73
$[68-74)$	9	82
$[74-80)$	6	88
	88	

Al primer intervalo cuya frecuencia absoluta acumulada exceda a la mitad del número de datos, se le llama clase mediana o intervalo mediana.

Así pues, se tiene

$$L_i = 56 \qquad c = 6 \qquad \frac{\sum f_i}{2} = 44 \qquad F_{i-1} = 30 \qquad \text{y} \qquad f_i = 25$$

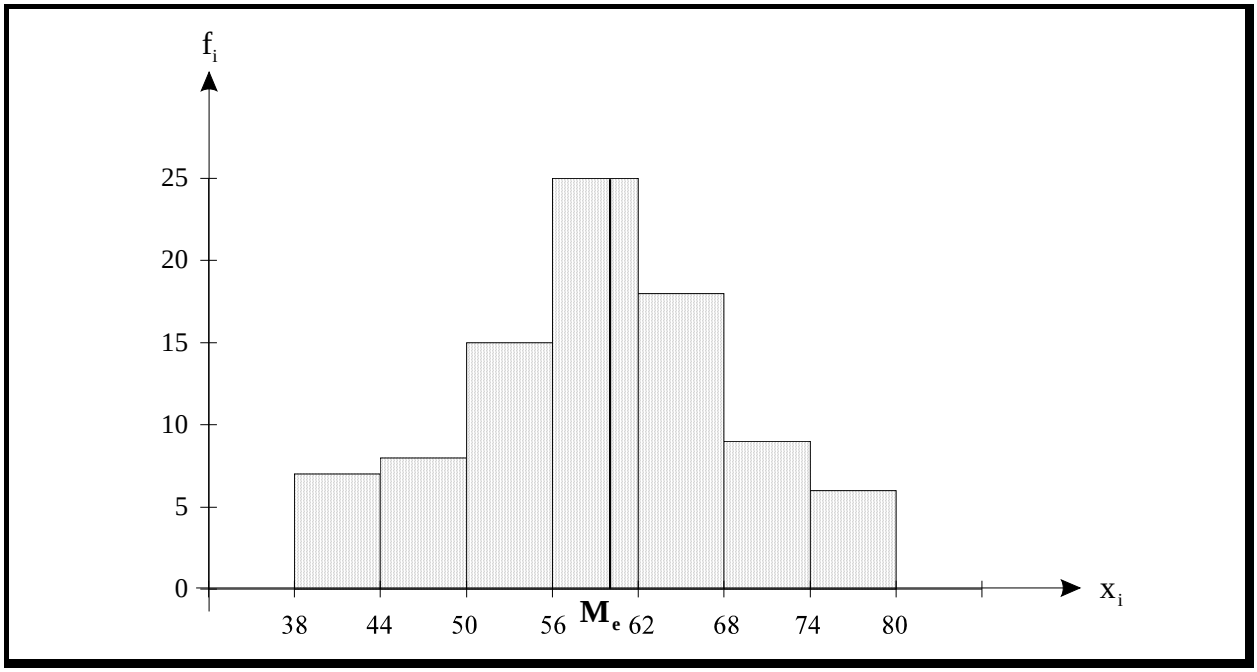
Sustituyendo en la expresión de la mediana resulta:

$$M_e = 56 + 6 \cdot \frac{44 - 30}{25} = 59'36$$

Cálculo de la mediana por el método gráfico

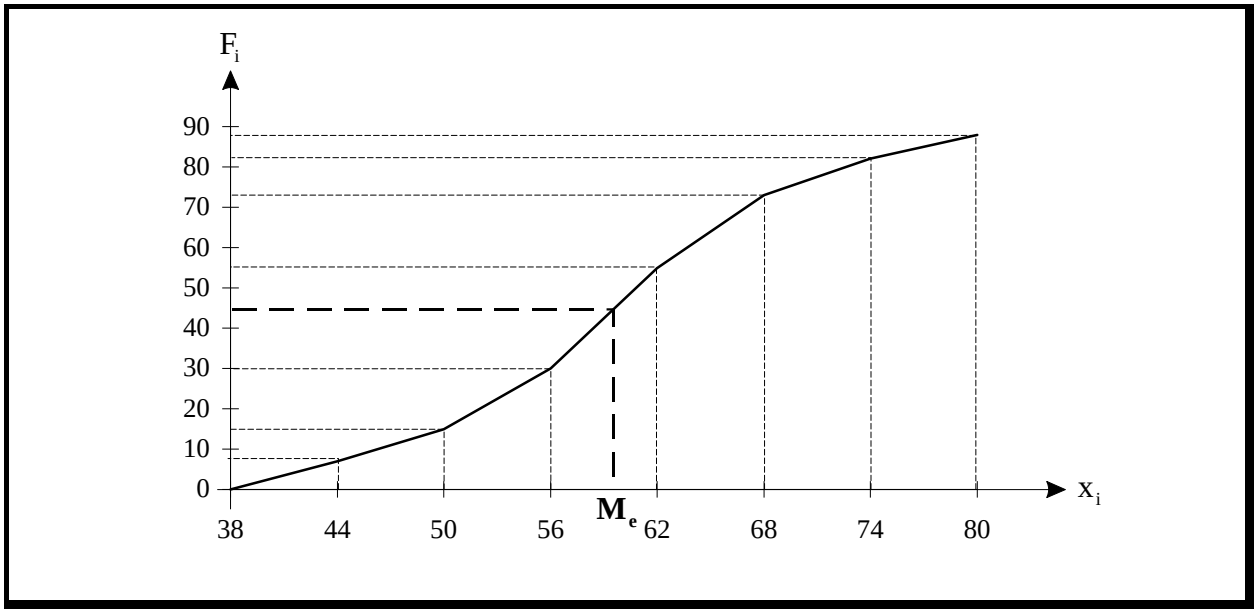
A través del histograma de frecuencias

Por debajo de la mediana se encuentran la mitad de los valores observados, y la otra mitad se hallan por encima (de ahí su nombre). Por tanto, divide el histograma en dos zonas de igual área.

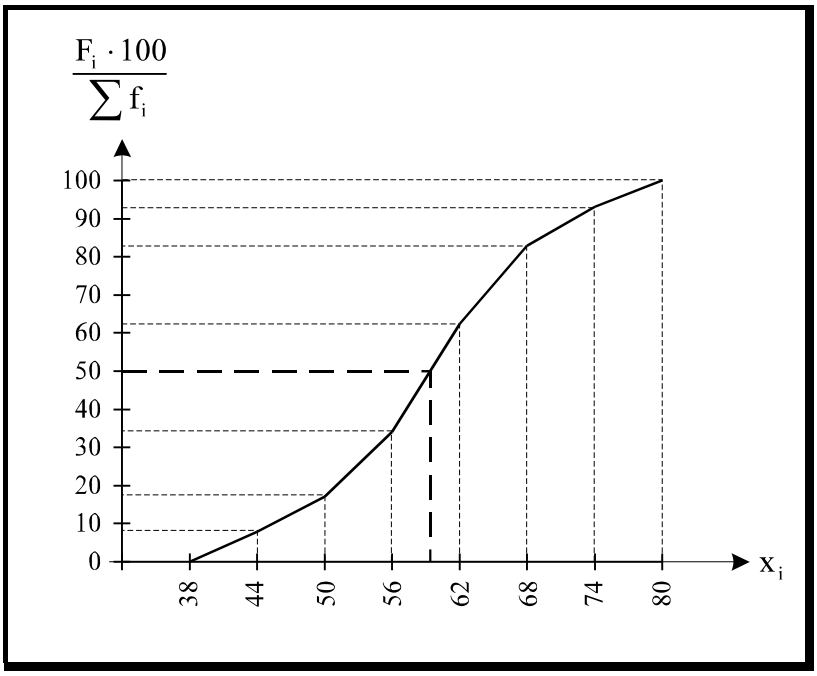


A través del polígono de frecuencias acumuladas

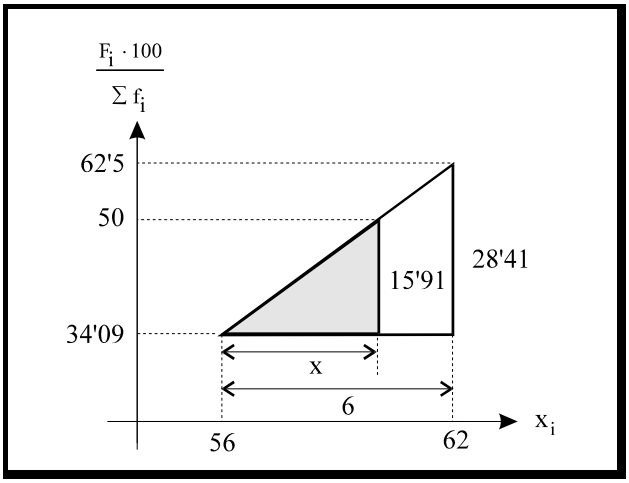
También podemos determinar gráficamente la mediana utilizando el polígono acumulativo de frecuencias. Obsérvese cómo, al unir por una recta el límite inferior y superior de cada intervalo, estamos admitiendo implícitamente que la distribución en el interior de cada intervalo es uniforme.



A través de la semejanza de triángulos en el polígono de porcentajes acumulados



Clases	f_i	F_i	$\frac{F_i \cdot 100}{\sum f_i}$
$[38-44)$	7	7	7'95
$[44-50)$	8	15	17'04
$[50-56)$	15	30	34'09
$[56-62)$	25	55	62'50
$[62-68)$	18	73	82'95
$[68-74)$	9	82	93'18
$[74-80)$	6	88	100
	88		



$$\frac{28'41}{6} = \frac{15'91}{x} \Rightarrow x = 3'36$$

Por lo tanto la mediana será:

$$\mathbf{M_e} = 56 + 3'36 = \mathbf{59'36}$$

Observaciones

- La mediana es particularmente útil en los siguientes casos:

  a) Cuando entre los datos existe alguno ostensiblemente extremo que, como hemos visto, afecta a la media.

  b) Cuando los datos están agrupados en clases y alguna de ellas es abierta.

- Como consecuencia de la definición de mediana, se tiene que el 50% de los datos son menores o iguales a ella y el 50% restante son mayores o iguales.

- La mediana es el primer parámetro de centralización que depende del orden de los datos y no de su valor.

- Geométricamente, y para distribuciones que se puedan representar mediante un histograma de frecuencias, la mediana es un valor de la variable, tal que la vertical levantada sobre el mismo divide al histograma en dos partes de igual área.

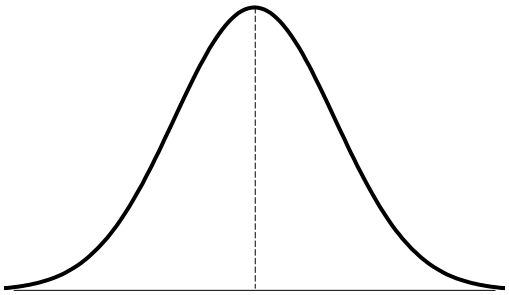
Relación entre media, moda y mediana

Para distribuciones unimodales, si al construir el polígono de frecuencias se observa que la distribución es simétrica o ligeramente asimétrica, entonces es posible comprobar experimentalmente la siguiente relación:

$$\textbf{Media} - \textbf{Moda} = 3(\textbf{Media} - \textbf{Mediana})$$

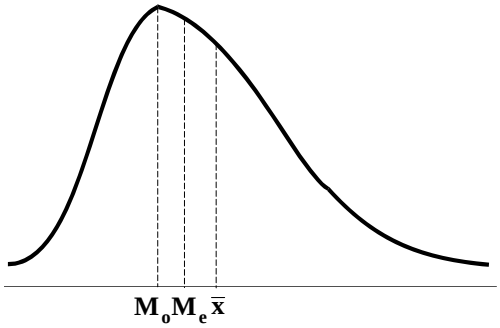
Gracias a esta relación se puede obtener, con un cierto error, algunos de estos parámetros en función de los otros, siempre y cuando se compruebe que la distribución es simétrica o ligeramente asimétrica.

A continuación representamos tres distribuciones estadísticas, en las que se sitúan los parámetros de centralización:



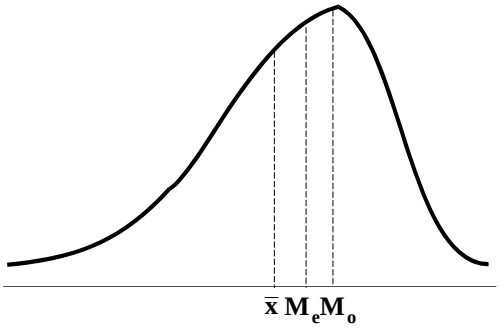
$M_e = \bar{x} = M_o$

Curva simétrica



$M_o M_e \bar{x}$

Curva asimétrica



$\bar{x}\, M_e M_o$

Curva asimétrica

¡OJO!

Queremos destacar que, a veces, las medidas de centralización no siempre son una descripción adecuada de todos los datos de una distribución. Para ello veamos a continuación, con alguna variante, un anecdótico ejemplo expuesto por J. C. Stanley en Measurement in Today's Schools.

Ejemplo: **En cierta ocasión se sentaron cinco hombres en un banco de un parque. Dos de ellos eran vagabundos y todos sus bienes ascendían a 1000 pts cada uno. El tercero era un obrero que no tenía más propiedades que una cuenta bancaria con 50.000 pts. El cuarto era un administrativo que entre su vivienda y su cuenta bancaria tenía unos bienes valorados en 6.000.000 de pts. El quinto era un agraciado de la Loto que tenía un capital igual a 400.000.000 pts.**

Calcular las medidas de centralización

La serie estadística es la siguiente: 1.000; 1.000; 50.000; 6.000.000; 400.000.000

Si calculamos las medidas de centralización de esta serie estadística se obtiene:

$$\overline{x} = \frac{1.000 + 1.000 + 50.000 + 6.000.000 + 400.000.000}{5} = 81.210.310 \text{ pts}$$

$$M_0 = 1.000 \text{ pts} \qquad M_e = 50.000 \text{ pts}$$

Vemos que la media no da una idea de cómo es la distribución; tampoco la moda permite asegurar nada, pues si bien 1.000 pts son los bienes del 40% de la distribución (los dos vagabundos), este valor se encuentra muy lejos y es prácticamente insignificante para el multimillonario de la Loto. Por último, con el conocimiento de la mediana, que describe muy bien el capital del obrero, nada permite afirmar de los capitales de los otros cuatro señores.

Con el fin de evitar contradicciones como la presente en la anécdota, se deben evitar grandes diferencias numéricas entre los datos de una distribución. Por otra parte, los parámetros estadísticos de una distribución informan mejor de ésta cuanto mayor es el número de datos.

Medidas de posición. Cuantiles

Al estudiar la Mediana hemos visto que, una vez ordenados de menor a mayor los datos de una distribución, la mediana divide a éstos en partes iguales. Análogamente, tiene interés estudiar otros parámetros que dividan a los datos de la distribución en función de otras cuantías.

Reciben genéricamente la denominación de **cuantiles** ***aquellos valores que dividen la distribución en intervalos, de forma que cada uno de ellos tenga la misma frecuencia.*** Los cuantiles toman denominaciones específicas según sea el número de intervalos en que se divide la distribución. así:

❒ Cuartiles

Se llama cuartiles a tres valores que dividen a la serie de datos en cuatro partes iguales, conteniendo cada una de ellas el 25% de la población.

Se representan por $\mathbf{Q_1, Q_2}$ **y** $\mathbf{Q_3}$ y se designan cuartil primero, segundo y tercero respectivamente. Los cuartiles Q_1, Q_2, y Q_3 son los valores que superan, exactamente, al 25%, 50% y 75% de los valores de la distribución respectivamente. El cuartil Q_2 coincide con la mediana de la distribución.

Hay dos valores, uno que separa a la población en un 25% por debajo y un 75% por encima, y el otro que deja por debajo al 75% y por encima al 25% de la población. Se llaman **cuartil inferior (CI) y cuartil superior (CS)**, y corresponden a Q_1 y Q_3 respectivamente.

Si el problema que estudiamos son las notas en una determinada asignatura "Estar por encima del cuartil superior" significa estar entre el 25% de los mejores.

❒ Quintiles

Se llaman quintiles a cuatro valores que dividen a la serie de datos en cinco partes iguales, conteniendo cada una de ellas el 20% de la población.

Se representan por $\mathbf{K_1, K_2, K_3}$ **y** $\mathbf{K_4}$ y se designan quintil primero, segundo, tercero y cuarto respectivamente.

❒ Deciles

Se llama deciles a nueve valores que dividen a la serie de datos en diez partes iguales, conteniendo cada una de ellas la décima parte de la población.

Se representan por $\mathbf{D_1, D_2, \cdots\cdots, D_9}$ y se designan decil primero, segundo, tercero,cuarto, ..., y noveno respectivamente. Hablar del decil 4 significa dejar por debajo del valor que representa al 40% de la población.

❒ Centiles o Percentiles

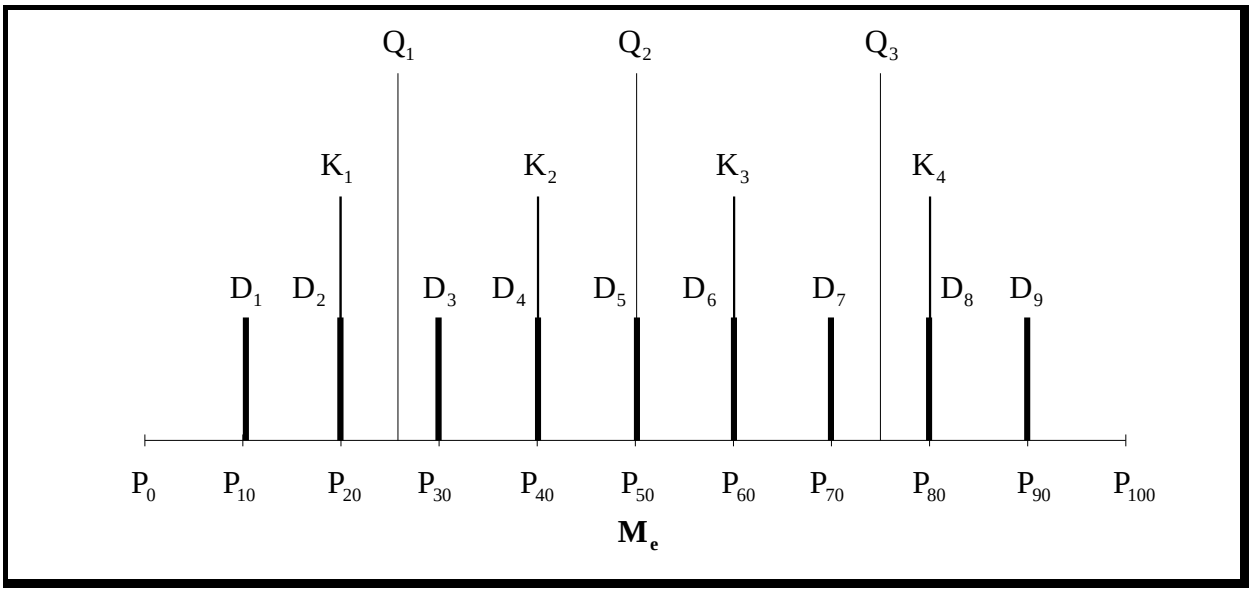
Se llaman centiles o percentiles a 99 valores que dividen a la serie de datos en cien partes iguales.

Se representan por $\mathbf{P_1, P_2, \cdots\cdots, P_{99}}$ y se designan percentil primero, segundo, tercero, cuarto,....., y nonagésimo noveno respectivamente. Hablar del centil 38 significa dejar por debajo del valor que representa al 38% de la población. Cuando se dice "Según su inteligencia abstracta este chico está en el centil 85" significa que su inteligencia abstracta, es superior a la del 85% de la población e inferior al 15% restante. Los centiles son muy utilizados por los psicólogos para dar los resultados de los tests

Ejemplo: Al analizar los resultados de una evaluación de COU, el profesor informa a un determinado padre que su hijo se encuentra en el decil 8 respecto del resultado total de la clase.

Esto significa que el alumno deja al 80% de la clase por debajo de su puntuación, o, lo que es igual, se encuentra dentro del 20% de los que obtuvieron mejores calificaciones.

A continuación se representa un gráfico donde se muestran las relaciones entre los distintos cuantiles.



Obsérvese que la Mediana coincide con el cuartil segundo $\left(Q_2\right)$, el decil quinto $\left(D_5\right)$ y el percentil de orden cincuenta $\left(P_{50}\right)$, es decir:

$$M_e = Q_2 = D_5 = P_{50}$$

Cálculo de los Cuantiles

Debido a que los Cuantiles son parámetros del tipo de la Mediana, su cálculo se realiza de forma análoga.

Variable discreta

Ejemplo: Las calificaciones en la asignatura de Historia del Arte de los 40 alumnos de una clase vienen dadas por la siguiente tabla:

Calificaciones	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Número de alumnos	**2**	**2**	**4**	**5**	**8**	**9**	**3**	**4**	**3**

Calcular: a) Los cuartiles primero y tercero

b) Los percentiles de orden 30 y 70

Formemos la tabla de frecuencias incluyendo las absolutas acumuladas.

x_i	f_i	F_i
1	2	2
2	2	4
3	4	8
4	5	13
5	8	21
6	9	30
7	3	33
8	4	37
9	3	40
	40	

a) Q_1 deja la cuarta parte de la distribución a la izquierda; como $\frac{\sum f_i}{4} = 10$, se verifica que el cuartil primero es 4, por ser éste el primer valor de la variable cuya frecuencia absoluta acumulada excede a la cuarta parte del número de datos:

$$Q_1 = 4$$

Q_3 deja las tres cuartas partes de la distribución a la izquierda; como $3 \cdot \frac{\sum f_i}{4} = 30$, se verifica que el cuartil tercero es 6'5, por ser éste el primer valor de la variable cuya frecuencia absoluta acumulada excede a las tres cuartas partes:

$$Q_3 = 6'5$$

b) P_{30} deja el 30% de la distribución a la izquierda; como $30 \cdot \frac{\sum f_i}{100} = 12$, se verifica que el percentil de orden 30 es 4, por ser éste el primer valor de la variable cuya frecuencia absoluta acumulada excede al 30% del total del número de datos:

$$P_{30} = 4$$

P_{70} deja el 70% de la distribución a la izquierda; como $70 \cdot \frac{\sum f_i}{100} = 28$, se verifica que el percentil de orden 70 es 6, por ser éste el primer valor de la variable cuya frecuencia absoluta acumulada excede al 70% del total del número de datos:

$$P_{70} = 6$$

Variable continua

Se tiene para los cuartiles las fórmulas siguientes:

Para el cuartil inferior:

$$\mathbf{Q_1 = L_i + c \cdot \frac{\frac{\sum f_i}{4} - F_{i-1}}{f_i}}$$

Para el cuartil superior:

$$\mathbf{Q_3 = L_i + c \cdot \frac{3 \cdot \frac{\sum f_i}{4} - F_{i-1}}{f_i}}$$

Se tiene para los deciles la fórmula siguiente:

Para el decil de orden k (D_k):

$$\mathbf{D_k = L_i + c \cdot \frac{k \cdot \frac{\sum f_i}{10} - F_{i-1}}{f_i}}$$

Se tiene para los centiles o percentiles la fórmula siguiente:

Para el centil de orden k (P_k):

$$\mathbf{P_k = L_i + c \cdot \frac{k \cdot \frac{\sum f_i}{100} - F_{i-1}}{f_i}}$$

Ejemplo: Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:

Puntuaciones	Nú mero de trabajadores
$[38-44)$	7
$[44-50)$	8
$[50-56)$	15
$[56-62)$	25
$[62-68)$	18
$[68-74)$	9
$[74-80)$	6

Calcular:

a) Los cuartiles primero y tercero.

b) Los percentiles de orden 40 y 90.

Formamos la tabla estadística incluyendo las frecuencias absolutas acumuladas:

Clases	f_i	F_i
$[38-44)$	7	7
$[44-50)$	8	15
$[50-56)$	15	30
$[56-62)$	25	55
$[62-68)$	18	73
$[68-74)$	9	82
$[74-80)$	6	88
	88	

a) Q_1 deja la cuarta parte de la distribución a la izquierda; como $\frac{\sum f_i}{4} = 22$, resulta que la clase que contiene el primer cuartil (mirando la columna de frecuencias absolutas acumuladas) es la que tiene por límites $[50-56$. Aplicando una expresión análoga a la de la mediana para datos agrupados en intervalos, se tiene:

$$Q_1 = L_i + c \cdot \frac{\left(\frac{\sum f_i}{4}\right) - F_{i-1}}{f_i} = 50 + 6 \cdot \frac{22-15}{15} = 52'8$$

Q_3 deja las tres cuartas partes de la distribución a la izquierda; como $3 \cdot \frac{\sum f_i}{4} = 66$, resulta que la clase que contiene el tercer cuartil (mirando la columna de frecuencias absolutas acumuladas) es la que tiene por límite $[62-68$. Aplicando una expresión análoga a la de la mediana para datos agrupados en intervalos, se tiene:

$$Q_3 = L_i + c \cdot \frac{\left(3 \cdot \frac{\sum f_i}{4}\right) - F_{i-1}}{f_i} = 62 + 6 \cdot \frac{66-55}{18} = 65'67$$

b) P_{40} deja el 40% de la distribución a la izquierda; como $40 \cdot \frac{\sum f_i}{100} = 40 \cdot \frac{88}{100} = 35'2$, resulta que la clase que contiene el percentil de orden 40 (mirando la columna de frecuencias absolutas acumuladas) es la que tiene los límites $[56-62$. Aplicando una expresión análoga a la de la mediana, se tiene:

$$P_{40} = L_i + c \cdot \frac{\left(40 \cdot \frac{\sum f_i}{100}\right) - F_{i-1}}{f_i} = 56 + 6 \cdot \frac{35'2 - 30}{25} = 57'25$$

P_{90} deja el 90% de la distribución a la izquierda.

Como $90 \cdot \frac{\sum f_i}{100} = 90 \cdot \frac{88}{100} = 79'2$, resulta que la clase que contiene el percentil de orden 90 (mirando la columna de frecuencias absolutas acumuladas) es la que tiene los límites $[68-74$. Aplicando una expresión análoga a la de la mediana, se tiene:

$$P_{90} = L_i + c \cdot \frac{\left(90 \cdot \frac{\sum f_i}{100}\right) - F_{i-1}}{f_i} = 68 + 6 \cdot \frac{79'2 - 73}{9} = 72'13$$

Observaciones

- Los cuantiles, preferentemente los deciles y percentiles, son parámetros estadísticos muy utilizados en las ciencias sociales.

- A los cuantiles se les suele denominar parámetros de estructura, ya que nos proporcionan información acerca de la estructura o distribución interna de los datos.

Cálculo gráfico de los cuantiles

Para calcular gráficamente los cuantiles de una distribución existe un método muy sencillo que consiste en representar el polígono de porcentajes acumulados, situando en el eje "x" los valores de la variable (si es discreta), o los intervalos (si es continua), y en el eje "y" la frecuencia absoluta acumulada en porcentaje, es decir, la obtenida al multiplicar la frecuencia absoluta acumulada por el cociente entre 100 y la suma de todas las frecuencias absolutas. ***Conviene realizar la representación sobre papel milimetrado, a fin de poder obtener una mayor precisión***.

$$\text{Porcentajes de frecuencias absolutas acumuladas} = F_i \cdot \frac{100}{\sum f_i}$$

Para obtener el cuantil de que se trate, se traza una paralela al eje "x" por el punto correspondiente al cuantil deseado. Ésta corta al polígono de frecuencias absolutas acumuladas en un punto; por éste se traza una paralela al eje "y", que corta al eje "x" en el punto buscado.

Ejemplo: Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:

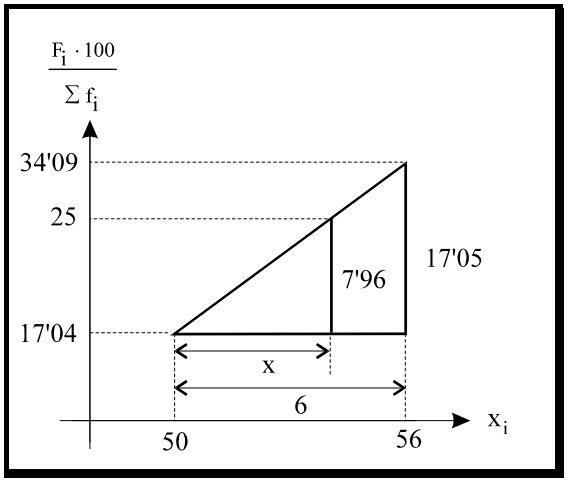
Puntuaciones	Nú mero de trabajadores
$[38-44)$	7
$[44-50)$	8
$[50-56)$	15
$[56-62)$	25
$[62-68)$	18
$[68-74)$	9
$[74-80)$	6

Calcular el cuartil inferior, el decil 7 y el percentil de orden 90.

¿Qué centil corresponde a 45 puntos?

Clases	f_i	F_i	$\frac{F_i \cdot 100}{88}$
$[38-44)$	7	7	7'95
$[44-50)$	8	15	17'04
$[50-56)$	15	30	34'09
$[56-62)$	25	55	62'5
$[62-68)$	18	73	82'95
$[68-74)$	9	82	93'18
$[74-80)$	6	88	100
	88		

Cuartil inferior

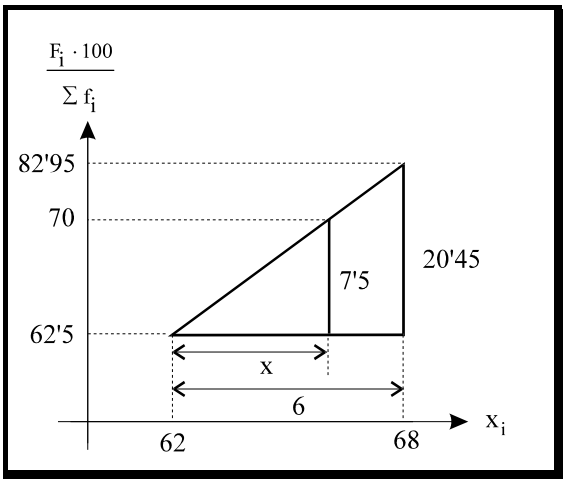


El cuartil inferior corresponde al percentil 25.

$$\frac{17'05}{6} = \frac{7'96}{x} \Rightarrow x = 2'8011$$

$$Q_1 = 50 + 2'8011 = 52'8011$$

Decil 7

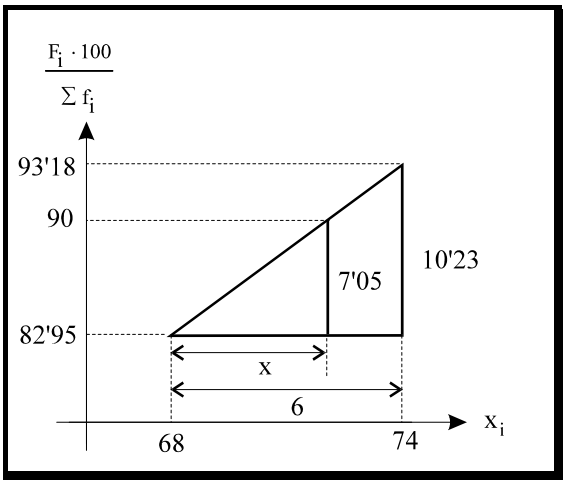


El decil 7 corresponde al percentil 70.

$$\frac{20'45}{6} = \frac{7'5}{x} \Rightarrow x = 2'20$$

$$D_7 = 62 + 2'2 = 64'2$$

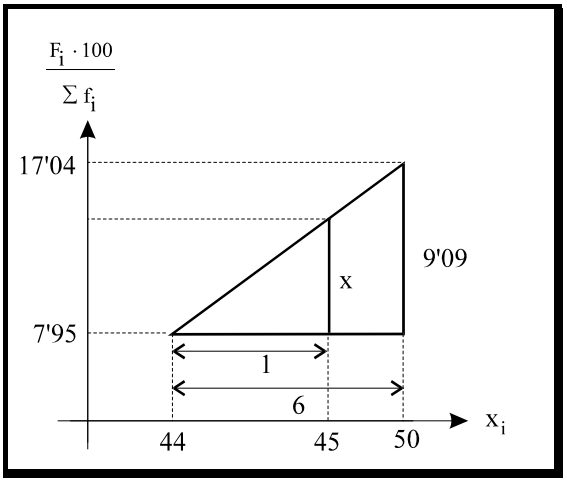
Percentil 90



$$\frac{10'23}{6} = \frac{7'05}{x} \Rightarrow x = 4'13$$

$$P_{90} = 68 + 4'13 = 72'13$$

¿Qué centil corresponde a 45 puntos?



$$\frac{9'09}{6} = \frac{x}{1} \Rightarrow x = 1'51$$

$$7'95 + 1'51 = 9'46$$

Corresponde aproximadamente al centil 9.

Medidas de dispersión

Ejemplo: Se ha aplicado a dos grupos de 8 alumnos de 8º de EGB un test de 100 preguntas sobre capacidad numérica, obteniéndose los siguientes resultados:

GrupoA	**46**	**48**	**49**	**50**	**50**	**51**	**52**	**54**
GrupoB	**10**	**18**	**30**	**50**	**50**	**70**	**82**	**90**

Las medias de cada una de las dos distribuciones son:

$$\bar{x}_A = 50 \qquad \bar{x}_B = 50$$

y sin embargo, los dos grupos de alumnos son bien distintos. Mientras que en el grupo A la mayoría de los alumnos han contestado prácticamente a la mitad de las preguntas, en el grupo B hay alumnos que casi han contestado a la totalidad, y otros que han contestado a muy pocas preguntas. Por tanto las puntuaciones del grupo A están muy concentradas, poco dispersas; en cambio las del grupo B se encuentran poco concentradas en torno a la media y diremos que se encuentran muy dispersas.

Así pues, la investigación acerca de una distribución queda incompleta si solo se estudian las medidas de centralización, siendo imprescindible conocer si los datos numéricos están agrupados o no alrededor de los valores centrales. A esto es a lo que se llama dispersión, y **a los parámetros que miden estas desviaciones respecto a la media se les llama medidas de dispersión o parámetros de dispersión.**

Las medidas de dispersión más importantes son: el **rango** o **recorrido**, la **desviación media**, la **varianza** y la **desviación típica**.

Rango o recorrido

Se llama rango o recorrido de una distribución a la diferencia entre el mayor y el menor valor de la variable estadística. Bajo el supuesto de que los valores de la variable estén ordenados en sentido creciente, su expresión matemática sería:

$$\mathbf{R = x_n - x_1}$$

En el ejemplo anterior, $R_A = 54 - 46 = 8$ y $R_B = 90 - 10 = 80$. En consecuencia, al tener el mismo número de datos ambas distribuciones y ser el recorrido de la distribución del grupo A mucho más pequeño, diremos que está más concentrada, o menos dispersa, que la distribución del grupo B.

Observaciones

- Cuanto menor es el recorrido de una distribución mayor es el grado de representatividad de los valores centrales.

- El recorrido tiene la ventaja de su sencillez de cálculo.

- Tiene gran aplicación en procesos de control de calidad, y de una manera general, en aquellos procesos que se pretenda verificar longitudes, pesos y volúmenes, estando prefijados de antemano los límites permitidos.

- El recorrido presenta el inconveniente de que solo depende de los valores extremos. De esta forma basta que uno de ellos se separe mucho, para que el recorrido se vea sensiblemente afectado. Es por tanto muy sensible a la fluctuación de estos valores extremos.

- Para paliar en alguna medida este inconveniente se utilizan en ocasiones otros dos rangos:

$$\textbf{Rango intercuartílico: } \mathbf{Q = Q_3 - Q_1}$$

$$\textbf{Rango entre percentiles: } \mathbf{P = P_{90} - P_{10}}$$

  Estos dos valores son más estables que el rango, ya que tienden a eliminar aquellos valores extremadamente alejados.

Desviación media

Si $\overline{x}$ es la media aritmética del conjunto de valores $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ de una variable estadística, se llama **desviación** del valor x_i, respecto de la media, a la diferencia $x_i - \overline{x}$, y **desviación absoluta** (respecto de la media), al valor absoluto de las desviaciones, esto es, a $|x_i - \overline{x}|$. La primera es positiva, o negativa, según que x_i sea superior o inferior a la media; la segunda es positiva o nula.

Parece lógico suponer que la suma de las desviaciones de todos los valores debe reflejar el grado de dispersión de la distribución estudiada. Sucede, no obstante, que esa suma siempre da cero, cuando se toman las desviaciones "tal cual" (con signo), por lo que, en todo caso, habría que tomar la suma de las desviaciones absolutas como medida de dispersión, o, aún mejor, **la media aritmética de los valores de las desviaciones absolutas**, parámetro que recibe el nombre de **desviación media**. Viene dada por la expresión:

$$\mathbf{DM} = \frac{\sum_{i=1}^{n} |\mathbf{x_i} - \overline{\mathbf{x}}|\ \mathbf{f_i}}{\sum_{i=1}^{n} \mathbf{f_i}}$$

En el caso continuo, x_i son las marcas de clase.

Ejemplo: Calcular la desviación media de la siguiente distribución:

x_i	2	9	11	12
f_i	1	2	4	3

Construimos la tabla con todos los datos que intervienen en la fórmula

x_i	f_i	$x_i f_i$	$\lvert x_i - \overline{x}\rvert$	$\lvert x_i - \overline{x}\rvert f_i$
2	1	2	8	8
9	2	18	1	2
11	4	44	1	4
12	3	36	2	6
	10	100		20

$$\overline{x} = \frac{100}{10} = 10$$

$$DM = \frac{20}{10} = 2$$

La desviación media nos indica que, como promedio, los valores se apartan dos unidades respecto de la media.

Varianza y Desviación típica

La desviación media tiene el inconveniente de no destacarnos suficientemente cuándo un valor se separa de la media, y, en cambio, destaca excesivamente las pequeñas diferencias que otros valores puedan tener respecto de la media. Para evitar esto y aumentar los contrastes entre las dispersiones de los valores, se toman los cuadrados de las desviaciones en lugar de las desviaciones simples.

Se llama varianza de una variable a la media aritmética de los cuadrados de las desviaciones respecto a la media. La varianza se representa por s^2.

Sea x una variable estadística que toma los valores $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ con frecuencias absolutas $f_1, f_2, \ldots\ldots\ldots\ldots, f_n$ respectivamente.

La varianza viene dada por la siguiente expresión:

$$s^2 = \frac{f_1 (x_1 - \overline{x})^2 + f_2 (x_2 - \overline{x})^2 + \cdots\cdots\cdots + f_n (x_n - \overline{x})^2}{f_1 + f_2 + \cdots\cdots\cdots + f_n} = \frac{\sum_{i=1}^{n} f_i (x_i - \overline{x})^2}{\sum_{i=1}^{n} f_i}$$

con frecuencia $\overline{x}$ no es un número entero, por lo que las desviaciones $(x_i - \overline{x})$ suelen ser números decimales. Las operaciones de elevar al cuadrado cada una de las desviaciones y multiplicarlas por las frecuencias respectivas peden resultar sumamente laboriosas, por ello veamos otra expresión equivalente a la anterior en la que se evitan estos cálculos:

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

$$\frac{\sum_{i=1}^{n} x_i^2\, f_i}{\sum_{i=1}^{n} f_i} - 2\overline{x}\cdot\frac{\sum_{i=1}^{n} x_i\, f_i}{\sum_{i=1}^{n} f_i} + \overline{x}^2\cdot\frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i^2\, f_i}{\sum_{i=1}^{n} f_i} - 2\,\overline{x}\,\overline{x} + \overline{x}^2\cdot 1 = \frac{\sum_{i=1}^{n} x_i^2\, f_i}{\sum_{i=1}^{n} f_i} - \overline{x}^2$$

$$s^2 = \frac{f_1 x_1^2 + f_2 x_2^2 + \cdots\cdots\cdots + f_n x_n^2}{f_1 + f_2 + \cdots\cdots\cdots + f_n} - \overline{x}^2 = \frac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - \overline{x}^2$$

$$\boxed{s^2 = \frac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - \overline{x}^2}$$

Se llama desviación típica de una variable estadística a la raíz cuadrada positiva de la varianza.

Como consecuencia de la definición, la desviación típica viene dada por la siguiente expresión:

$$\boxed{s = \sqrt{\frac{\sum_{i=1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - \overline{x}^2}}$$

Uso de la calculadora para la obtención de s

Los pasos a seguir son los mismos que para el cálculo de la media aritmética. Una vez introducidos todos los datos, pulsando la tecla $\boxed{\sigma_n}$ (o $\boxed{\sigma}$ según el modelo de la calculadora) obtenemos en pantalla directamente el valor de la desviación típica.

Ejemplo: **El número de horas que dedica un alumno de COU al estudio durante la semana es el siguiente: 3'5, 5, 4, 6, 5'5, 3. Calcular el rango, la varianza y la desviación típica.**

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
3	1	3	9
3'5	1	3'5	12'25
4	1	4	16
5	1	5	25
5'5	1	5'5	30'25
6	1	6	36
	6	27	128'5

Rango: $6-3=3$

Media: $\bar{x}=\dfrac{\sum x_i f_i}{\sum f_i}=\dfrac{27}{6}=4'5$

Varianza: $s^2=\dfrac{x_i^2 f_i}{f_i}-\bar{x}^2=\dfrac{128'5}{6}-4'5^2=1'16$

Desviación típica: $s=\sqrt{1'16}=1'08$

Ejemplo: **Las calificaciones en la asignatura de Historia del Arte de los 40 alumnos de una clase vienen dadas por la siguiente tabla:**

Calificaciones	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Número de alumnos	**2**	**2**	**4**	**5**	**8**	**9**	**3**	**4**	**3**

Calcular el rango, la varianza y la desviación típica.

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
1	2	2	2
2	2	4	8
3	4	12	36
4	5	20	80
5	8	40	200
6	9	54	324
7	3	21	147
8	4	32	256
9	3	27	243
	40	212	1296

Rango: $9-1=8$

Media: $\bar{x}=\dfrac{212}{40}=5'3$

Varianza: $s^2=\dfrac{1296}{40}-(5'3)^2=4'31$

Desviación típica: $s=\sqrt{4'31}=2'08$

Ejemplo: **Se ha aplicado un test sobre satisfacción en el trabajo a 88 empleados de una fábrica, obteniéndose los siguientes resultados:**

Puntuaciones	**Número de trabajadores**
$[38-44)$	**7**
$[44-50)$	**8**
$[50-56)$	**15**
$[56-62)$	**25**
$[62-68)$	**18**
$[68-74)$	**9**
$[74-80)$	**6**

Calcular el rango, la varianza y la desviación típica.

Clases	Marcas de clase x_i	f_i	$x_i f_i$	$x_i^2 f_i$
$[38-44)$	41	7	287	11.767
$[44-50)$	47	8	376	17.672
$[50-56)$	53	15	795	42.135
$[56-62)$	59	25	1.475	87.025
$[62-68)$	65	18	1.170	76.050
$[68-74)$	71	9	639	45.369
$[74-80)$	77	6	462	35.574
		88	5.204	315.592

Rango: $80-38=42$

Media: $\bar{x}=\dfrac{5204}{88}=59'14$

Varianza: $s^2=\dfrac{315.592}{88}-(59'14)^2=88'73$

Desviación típica: $s=\sqrt{88'73}=9'4$

Observaciones

- Tanto la varianza como la desviación típica dependen de todos los valores de la distribución así como de la media.

- En los casos en que no sea posible calcular la media aritmética, no será posible tampoco obtener la varianza y la desviación típica por ser funciones de la media aritmética.

- *La varianza tiene el inconveniente que no viene expresada en las misma unidades que los datos, debido a que las desviaciones van elevadas al cuadrado. Así, por ejemplo, si los datos son metros, la varianza vendrá dada en metros cuadrados. En cambio, la desviación típica viene expresada en las mismas unidades que los datos de la distribución, de ahí que la desviación típica resulte más interesante que la varianza*.

Utilización conjunta de $\overline{x}$ y s

Podemos asegurar que, salvo para distribuciones muy estrafalarias, el porcentaje de individuos que se encuentran en el intervalo $(\overline{x}-s,\overline{x}+s)$ oscila entre el 60% y el 80% y suele ser próximo al 68%. Digamos para acordarnos, que en ese intervalo están aproximadamente los 2/3 de la población.

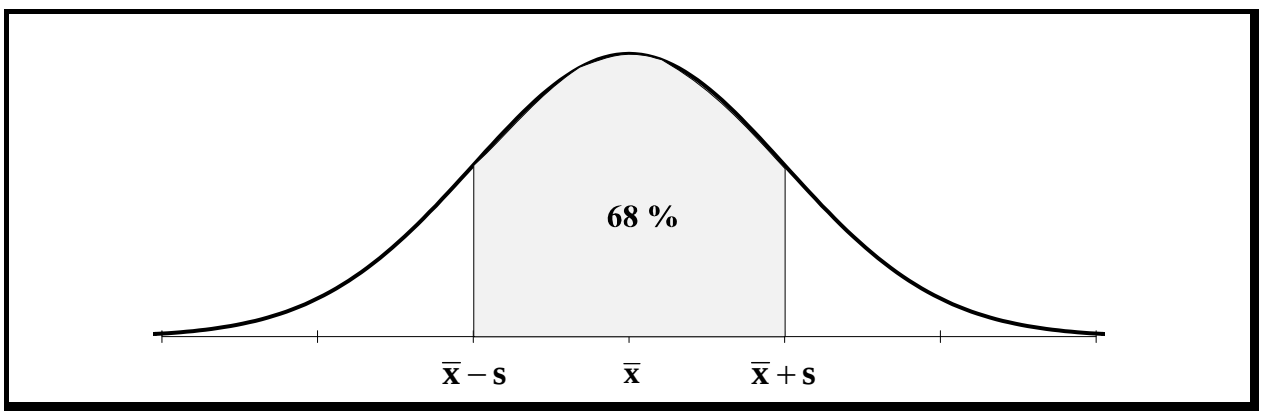
En las distribuciones unimodales, simétricas o ligeramente asimétricas, se verifica que:

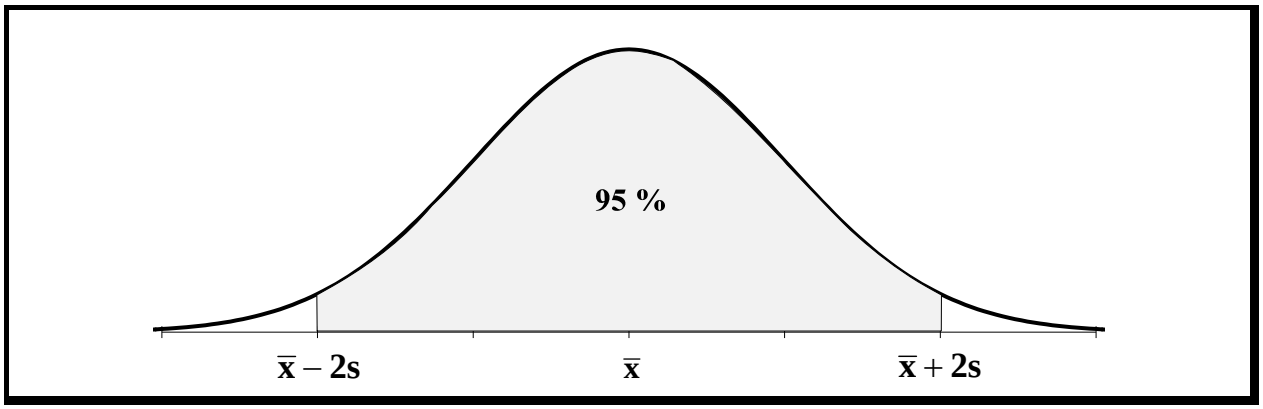
1. **En el intervalo $(\overline{x}-s,\overline{x}+s)$ se encuentra el 68% de los datos.**

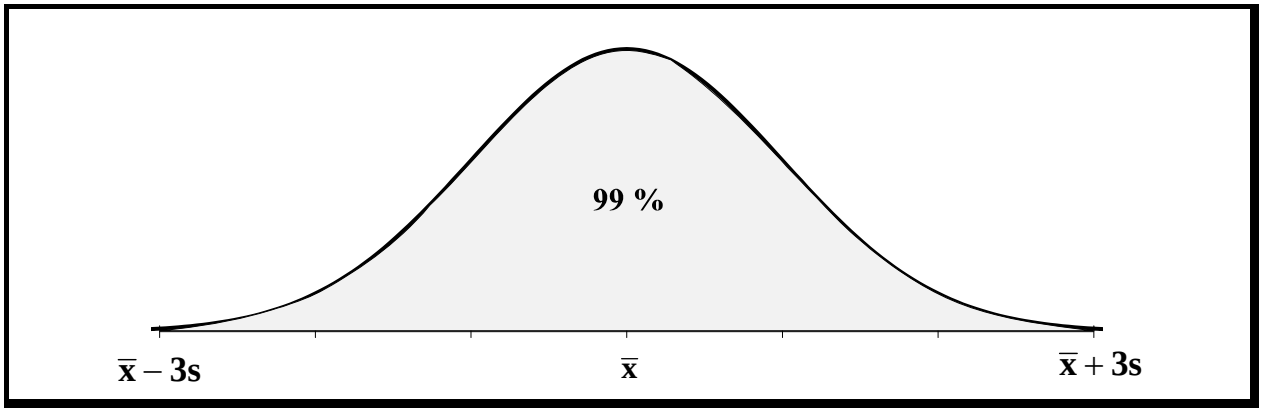
2. **En el intervalo $(\overline{x}-2s,\overline{x}+2s)$ se encuentra el 95% de los datos.**

3. **En el intervalo $(\overline{x}-3s,\overline{x}+3s)$ se encuentra el 99% de los datos.**

Cuando la distribución no es totalmente simétrica calcularemos los extremos de los intervalos, y para hallar el porcentaje en cada intervalo **haremos un recuento a ojo** de las cantidades que hay entre los extremos del mismo con los datos que tengamos en la tabla de frecuencias, y luego calcularemos la proporción







El uso conjunto de la media y la desviación típica nos permite comparar valores de una misma variable en distribuciones distintas, o bien ambas distribuciones. A tal fin, son útiles los siguientes conceptos:

Coeficiente de variación de Pearson

La desviación típica es un término absoluto del que se pueden sacar conclusiones erróneas sobre la dispersión de la muestra. Para comparar las desviaciones de poblaciones muy distintas, en vez de la desviación típica se usa el coeficiente de variación de Pearson. Mide la **variación relativa**.

Se llama coeficiente de variación de Pearson de una distribución de media $\bar{x}$ y desviación típica s, al número

$$CV = \frac{s}{\bar{x}}$$

Ejemplo: De dos muestras, la primera con media 30 y desviación típica 4 y la segunda con media 60 y desviación típica 6, ¿cuál es la que aparece más dispersa?

Considerando las desviaciones típicas diríamos que la segunda. Sin embargo, reduciendo los datos a una misma escala (puesto que una media es el doble de la otra), ocurre justamente lo contrario.
Si analizamos el coeficiente de variación de Pearson para las dos muestras tenemos que para la muestra de media 30 es $\frac{4}{30} = 0'13$, es decir el 13%, y para la muestra de media 60 es $\frac{6}{60} = 0'1$, es decir el 10%, lo que nos muestra que la primera tiene, relativamente más desviación típica que la segunda.

El coeficiente de variación de Pearson suele expresarse porcentualmente y sólo se usa en variables que no tomen valores negativos y no tengan medias próximas a cero, pues un denominador pequeño distorsiona el cociente.

Puntuación típica

En una distribución de media $\bar{x}$ y desviación típica s, se llama puntuación típica del valor x_i de la variable al número

$$Z = \frac{x_i - \bar{x}}{s}$$

La puntuación típica mide la desviación respecto de la media del dato considerado, tomando como unidad la desviación típica, por lo que refleja cuán desviado se halla éste respecto de la media, independientemente de la unidad utilizada.

Observaciones

- Las puntuaciones típicas son muy usadas en las ciencias sociales.
- La media aritmética de las puntuaciones típicas es 0.
- La desviación típica de las puntuaciones típicas es 1.
- **Las puntuaciones típicas se utilizan para comparar las puntuaciones obtenidas en distintas distribuciones. A mayor puntuación típica mejor puntuación en su distribución respecto de la otra distribución con quien se compara.**
- Conviene no confundir puntuación típica, que se refiere a puntuaciones obtenidas por cada individuo del grupo, con desviación típica, que se refiere a un parámetro obtenido para todo el grupo.

Ejemplo: **Una empresa textil tiene unos beneficios netos de 6.000.000 de pts en un año, y otra empresa del sector químico tiene unos beneficios de 12.000.000 de pts anuales. Sabiendo que los beneficios del sector textil tuvieron una media por empresa de 5.000.000 de pts con una desviación típica de 1.000.000 de pts, y los beneficios del sector químico tuvieron una media de 10.000.000 de pts por empresa, con una desviación típica de de 3.000.000 de pts ¿Qué empresa está mejor gestionada?**

Para contestar a esta pregunta tendremos que comparar los beneficios de cada empresa con los de su sector.

Empresa textil $\longrightarrow$ $Z = \dfrac{6.000.000 - 5.000.000}{1.000.000} = 1$

Empresa química $\longrightarrow$ $Z = \dfrac{12.000.000 - 10.000.000}{3.000.000} = 0'66$

Esto nos indica que, dentro de su sector, la industria textil tuvo un mejor rendimiento que la química

Ejemplo: **Un alumno ha contestado a dos test, obteniendo las siguientes puntuaciones:**

TEST A: 50 puntos **TEST B: 32 puntos**

El profesor ha calculado que el grupo de alumnos que ha contestado a cada uno de los tests tiene las siguientes medias y desviaciones típicas:

Para el TEST A: $\bar{x}_A = 45$ **y** $s_A = 6$ **; Para el TEST B:** $\bar{x}_B = 26$ **y** $s_B = 2$

¿En cuál de los dos tests ha obtenido, comparativamente con el grupo, mejor resultado el alumno?

Para poder contestar a esta pregunta tendremos que comparar las puntuaciones del alumno con las del grupo; para ello, restamos a cada una de las puntuaciones del alumno la media del grupo y dividimos por la desviación típica.

TEST A: $Z = \dfrac{50 - 45}{6} = 0'83$ TEST B: $Z = \dfrac{32 - 26}{2} = 3$

Así pues, si bien la puntuación directa del test A ha sido mayor que la obtenida en B, comparativamente con el grupo es mucho mayor la obtenida en el test B que en el A.

Momentos de una distribución de frecuencias

Los momentos son medidas que caracterizan a una distribución de frecuencias, existiendo una relación biunívoca entre una distribución y el conjunto de todos sus momentos.

Se llama **momento central de orden k** al parámetro estadístico:

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

Observa que:

- el momento central de orden 1 es siempre 0 $(\mu_1 = 0)$.
- el momento central de orden 2 es la varianza $(\mu_2 = s^2)$
- cuanto mayor sea k más influyen en el valor del momento correspondiente los valores muy alejados de la medida.

El momento central de orden 3 sirve para medir la asimetría de la distribución y el de orden 4 su grado de apuntamiento o aplastamiento, es decir, el que la gráfica sea más o menos picuda. Basándose en los momentos, se definen los siguientes parámetros:

Coeficiente de Asimetría

Con los coeficientes de asimetría se trata de medir si las observaciones están dispuestas simétrica o asimétricamente respecto a un valor central (en general, la media aritmética) y el grado de esta asimetría.

Todos los coeficientes que se utilizan son números abstractos y, por tanto, sin dimensiones. el más utilizado de todos es debido a Fisher y tiene la siguiente expresión:

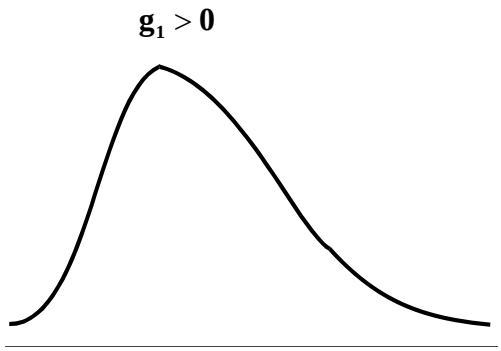
$$g_1 = \frac{\mu_3}{s^3} = \frac{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^3 \cdot f_i}{\sum_{i=1}^{n} f_i}}{s^3}$$

Puesto que $\left(x_i - \overline{x}\right)^3$ puede ser positivo o negativo, el coeficiente de asimetría puede ser positivo o negativo, teniéndose, según los casos, la siguiente interpretación:

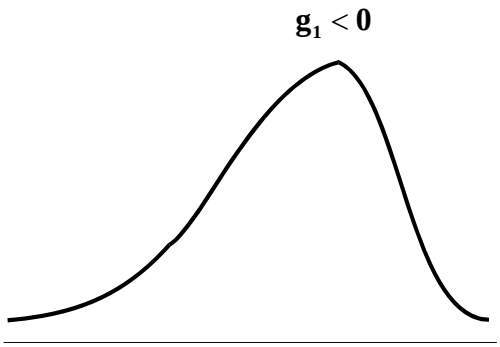
Si $g_1 < 0$ La distribución es asimétrica a la izquierda. Los valores a la izquierda de la media "pesan" más que los que están a la derecha. A la derecha, los valores caen con más rapidez: la curva es **sesgada a la izquierda**.

Si $g_1 = 0$ La distribución es simétrica

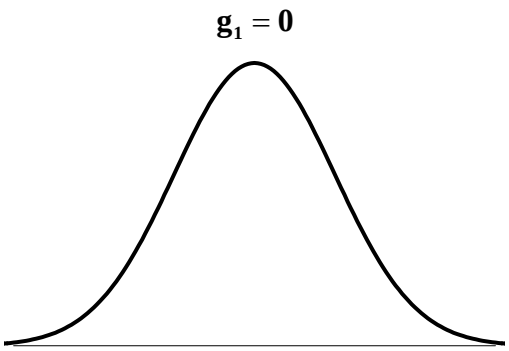
Si $g_1 > 0$ La distribución es asimétrica a la derecha. La curva es **sesgada a la derecha**.



sesgada a la derecha



sesgada a la izquierda



normal

Coeficiente de Apuntamiento o Curtosis

Con el coeficiente de apuntamiento o curtosis se trata de medir el grado de apuntamiento de una distribución respecto a la distribución **normal**, que se toma como patrón, y cuyo coeficiente de curtosis es 0. La distribución normal es la más importante, tanto en la teoría de la probabilidad como en le práctica de los trabajos estadísticos. Se caracteriza por ser simétrica respecto al eje $x = \overline{x}$. Su función de densidad, en la práctica histograma de frecuencias, tiene forma de campana. su importancia es debida en gran medida a que a ella convergen un conjunto importante de distribuciones estadísticas: Binomial, Poisson, χ^2, *t* de Student, etc.

Se llama coeficiente de apuntamiento o curtosis al parámetro:

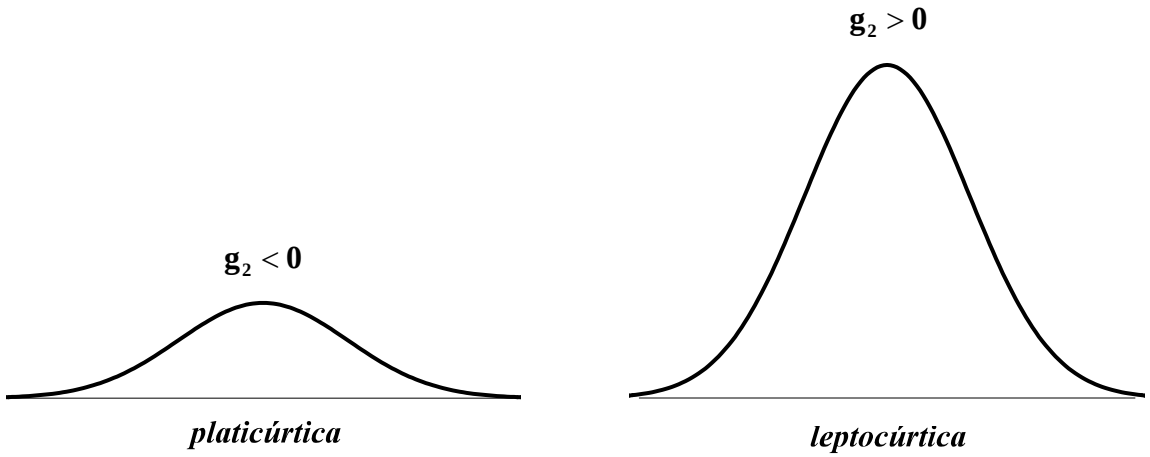
$$g_2 = \frac{\mu_4}{s^4} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \cdot f_i}{\sum_{i=1}^{n} f_i}}{s^4} - 3$$

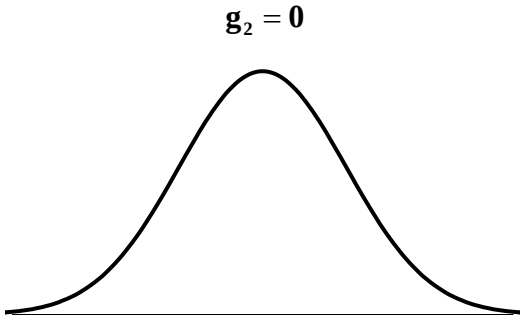
Este coeficiente, siempre positivo, tiene un valor crítico, el valor 0, considerado normal. Los casos son los siguientes:

Si $g_2 > 0$ La distribución es **leptocúrtica** (más apuntada que la normal.

Si $g_2 = 0$ La distribución es **normal**.

Si $g_2 < 0$ La distribución es **platicúrtica** (menos apuntada que la normal.





Ejemplo: Determinar los coeficientes de asimetría y curtosis de la siguiente distribución de frecuencias:

x_i	1	2	3	4
f_i	6	10	2	2

Construimos la tabla con todos los datos que vamos a necesitar:

x_i	f_i	$x_i f_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot f_i$	$(x_i - \bar{x})^3 \cdot f_i$	$(x_i - \bar{x})^4 \cdot f_i$
1	6	6	–1	1	6	–6	6
2	10	20	0	0	0	0	0
2	2	6	1	1	2	2	2
4	2	8	2	4	8	16	32
	20	40			16	12	40

$$\bar{x} = \frac{40}{20} = 2 \quad g_1 = \frac{\frac{12}{20}}{\left(\sqrt{\frac{16}{20}}\right)^3} = \frac{0'6}{0'72} = 0'83 \quad g_2 = \frac{\frac{40}{20}}{\left(\frac{16}{20}\right)^4} - 3 = \frac{50}{16} - 3 = \frac{2}{16} = 0'125$$