Introducción a la Estadística

Tema 2

Distribuciones unidimensionales

2.1 INTRODUCCIÓN

A la chepa de cada distribución de frecuencias vamos a pegar un montón de números (medidas) que nos permitirán sacarle jugo a la información que ésta contiene.

1) Medidas de posición centrales

- Media aritmética
- Media geométrica
- Media armónica
- Fórmula de Foster
- Mediana
- Moda

2) Medidas de posición no centrales

- Cuantiles
- Cuartiles
- Deciles
- Percentiles

3) Momentos

- Respecto a un punto cualquiera
- Respecto al origen
- Respecto a la media aritmética

4) Medidas de dispersión absoluta

- Recorrido
- Recorrido intercuartílico
- Desviación media respecto a la media aritmética
- Desviación media respecto a la mediana
- Varianza
- Desviación típica

5) Medidas de dispersión relativa

- Coeficiente de apertura
- Recorrido relativo
- Recorrido semiintercuartílico
- Coeficiente de variación de Pearson
- Índice de variación respecto a la mediana

6) Medidas de forma (asimetría y apuntamiento o curtosis)

- Coeficiente de asimetría de Fisher
- Coeficiente de asimetría de Pearson
- Coeficiente de asimetría de Bowley
- Coeficiente de curtosis

7) Medidas de concentración

- Índice de Gini

Si $(x_i;n_i)$ es la distribución de frecuencias correspondiente a una muestra de tamaño "N" de una variable estadística "X", la **media aritmética** de la distribución se denota $\overline{x}$, y se obtiene multiplicando cada valor observado x_i por su frecuencia relativa, sumando después los productos realizados:

$$\overline{x} = \sum_{i=1}^{k} x_i . \frac{n_i}{N} = \frac{x_1.n_1 + ... + x_k.n_k}{N}$$

En una distribución con **datos agrupados** se toma la marca de clase de cada intervalo como punto representativo del intervalo, y la media aritmética se obtiene como acabamos de indicar.

Ejemplos

x_i	n_i
19	20
20	10
22	5
25	5
	40

$$\Rightarrow \overline{x} = \sum_{i=1}^{4} x_i . \frac{n_i}{40} = \frac{19.20 + 20.10 + 22.5 + 25.5}{40} = 20'375$$

Intervalo	x_i	n_i
(0;30]	15	8
(30;60]	45	4
(60;80]	70	5
(80;100]	90	3
		20

$$\Rightarrow \overline{x} = \sum_{i=1}^{4} x_i 2 \frac{n_i}{40} = \frac{15.8 + 45.4 + 70.5 + 90.3}{20} = 46$$

Media aritmética ponderada

A veces se asigna a cada x_i una **ponderación o peso** w_i distinto de su frecuencia relativa; en tal caso, la media ponderada es:

$$\overline{x} = \frac{x_1.w_1 + ... + x_n.w_n}{w_1 + ... + w_n}$$

Por ejemplo: un opositor hace 3 baterías de test que, debido a su distinta dificultad, ponderan 2, 3 y 5 respectivamente; así, si las respectivas puntuaciones que obtiene en cada test son 60, 80 y 40, su puntuación media es

$$\overline{x} = \frac{60.2 + 80.3 + 40.5}{2+3+5} = 56$$

PROPIEDADES

1) La suma de las desviaciones de los valores observados de la variable respecto de la media aritmética es cero.

$$\sum_{i=1}^{k}(x_i-\bar{x}).n_i=\left(\sum_{i=1}^{k}x_i.n_i\right)-\bar{x}.\sum_{i=1}^{k}n_i=$$

$$=N.\left(\sum_{i=1}^{k}\frac{x_i.n_i}{N}\right)-\bar{x}.\sum_{i=1}^{k}n_i=N.\bar{x}-\bar{x}.N=0$$

2) La media de los cuadrados de las desviaciones de los valores observados de la variable respecto de una constante "C" es mínima si $C=\bar{x}$.

$$D(C)=\sum_{i=1}^{k}(x_i-C)^2.\frac{n_i}{N}=\sum_{i=1}^{k}(x_i-C+\bar{x}-\bar{x})^2.\frac{n_i}{N}=$$

$$=\sum_{i=1}^{k}\left((x_i-\bar{x})-(C-\bar{x})\right)^2.\frac{n_i}{N}=$$

$$=\left(\sum_{i=1}^{k}(x_i-\bar{x})^2.\frac{n_i}{N}\right)-\frac{2.(C-\bar{x})}{N}.\sum_{i=1}^{k}(x_i-\bar{x}).n_i+(C-\bar{x})^2.\sum_{i=1}^{k}\frac{n_i}{N}=$$

$$\boxed{\sum_{i=1}^{k}(x_i-\bar{x}).n_i=0\ ;\ \sum_{i=1}^{k}\frac{n_i}{N}=1}$$

$$=\left(\sum_{i=1}^{k}(x_i-\bar{x})^2.\frac{n_i}{N}\right)+(C-\bar{x})^2$$

Obviamente, el valor de D(C) es mínimo si $C=\bar{x}$, pues en tal caso el sumando no negativo $(C-\bar{x})^2$ es cero.

3) **Cambio de escala:** si "K" es una constante y $z_i=K.x_i$, las distribuciones de frecuencias $(x_i;n_i)$ y $(z_i;n_i)$ son tales que $\bar{z}=K.\bar{x}$.

$$\bar{z}=\sum_{i=1}^{k}z_i.\frac{n_i}{N}=\sum_{i=1}^{k}(K.x_i).\frac{n_i}{N}=K.\left(\sum_{i=1}^{k}x_i.\frac{n_i}{N}\right)=K.\bar{x}$$

$$\boxed{\sum_{i=1}^{k}x_i.\frac{n_i}{N}=\bar{x}}$$

4) **Cambio de origen:** si "K" es una constante y $z_i=x_i+K$, las distribuciones de frecuencias $(x_i;n_i)$ y $(z_i;n_i)$ son tales que $\bar{z}=\bar{x}+K$.

$$\bar{z}=\sum_{i=1}^{k}z_i.\frac{n_i}{N}=\sum_{i=1}^{k}(x_i+K).\frac{n_i}{N}=\left(\sum_{i=1}^{k}x_i.\frac{n_i}{N}\right)+K.\sum_{i=1}^{k}\frac{n_i}{N}=\bar{x}+K$$

$$\boxed{\sum_{i=1}^{k}x_i.\frac{n_i}{N}=\bar{x}\ ;\ \sum_{i=1}^{k}\frac{n_i}{N}=1}$$

5) **Cambio de origen y de escala:** si K_1 y K_2 son constantes y $z_i = K_1.x_i + K_2$, las distribuciones $(x_i;n_i)$ y $(z_i;n_i)$ son tales que $\bar{z} = K_1.\bar{x} + K_2$.

$$\bar{z} = \sum_{i=1}^{k} z_i.\frac{n_i}{N} = \sum_{i=1}^{k} (K_1.x_i+K_2).\frac{n_i}{N} = K_1.\left(\sum_{i=1}^{k} x_i.\frac{n_i}{N}\right)+K_2.\sum_{i=1}^{n}\frac{n_i}{N} = K_1.\bar{x}+K_2$$

$$\boxed{\sum_{i=1}^{k} x_i.\frac{n_i}{N} = \bar{x} \ ; \ \sum_{i=1}^{k}\frac{n_i}{N} = 1}$$

6) **Media por estratos:** si de un conjunto de valores obtenemos dos o más subconjuntos disjuntos, la media aritmética de todo el conjunto puede expresarse en función de las medias aritméticas de dichos subconjuntos.

Considera que la distribución

x_1	x_2	.....	x_h	x_{h+1}	x_{h+2}	.....	x_k
n_1	n_2	.....	n_h	n_{h+1}	n_{h+1}	.....	n_k

se particiona en las siguientes dos distribuciones:

x_1	x_2	.....	x_h
n_1	n_2	.....	n_h

$\Rightarrow$ su media es $\bar{u}_1 = \sum_{i=1}^{h} x_i.\frac{n_i}{N_1}$

$$\boxed{\text{siendo } \sum_{i=1}^{h} n_i = N_1}$$

x_{h+1}	x_{h+2}	.....	x_k
n_{h+1}	n_{h+2}	.....	n_k

$\Rightarrow$ su media es $\bar{u}_2 = \sum_{i=h+1}^{k} x_i.\frac{n_i}{N_2}$

$$\boxed{\text{siendo } \sum_{i=h+1}^{k} n_i = N_2}$$

Se tiene que:

$$\bar{x} = \sum_{i=1}^{k} x_i.\frac{n_i}{N} = \frac{\sum_{i=1}^{h} x_i.n_i + \sum_{i=h+1}^{k} x_i.n_i}{N} =$$

$$= \frac{N_1.\sum_{i=1}^{h} x_i.\frac{n_i}{N_1} + N_2.\sum_{i=h+1}^{k} x_i.\frac{n_i}{N_2}}{N} = \frac{N_1.\bar{u}_1 + N_2.\bar{u}_2}{N}$$

7) **Cálculo abreviado de la media aritmética:** si la distribución de frecuencias presenta numerosos valores o éstos son números grandes (en valor absoluto), para calcular $\bar{x}$ conviene hacer un cambio de origen y de escala:

$$z_i = (x_i - O)/c$$

donde "c" es el máximo común divisor de las diferencias entre cada dos valores consecutivos de la variable y "O" es un número arbitrario, que se procura sea un valor central de la distribución. Tras calcular $\bar{z}$ podremos determinar fácilmente $\bar{x}$:

$$z_i = \frac{x_i - O}{c} \Rightarrow x_i = c.z_i + O \Rightarrow \bar{x} = c.\bar{z} + O$$

Ejemplo 1

x_i	n_i
62341	4
62343	8
62345	4
62347	3
62349	8
	27

$\Rightarrow$

$z_i = (x_i - 62345)/2$	n_i
-2	4
-1	8
0	4
1	3
2	8
	27

$$\Rightarrow \bar{z} = \sum_{i=1}^{5} z_1 . \frac{n_i}{27} = \frac{3}{27} \Rightarrow$$

$\boxed{c = 2 \ ; \ O = 62345}$

$$\Rightarrow \bar{x} = c.\bar{z} + O = 2.\frac{3}{27} + 62345 = 62345'22$$

Ejemplo 2

Intervalo	x_i	n_i
$20 - 30$	25	7
$30 - 40$	35	11
$40 - 60$	50	16
$60 - 70$	65	10
$70 - 80$	80	6
		50

$\Rightarrow$

$z_i = (x_i - 50)/5$	n_i
-5	7
-3	11
0	16
3	10
6	6
	50

$$\Rightarrow \bar{z} = \sum_{i=1}^{5} z_1 . \frac{n_i}{50} = -\frac{2}{50} \Rightarrow$$

$\boxed{c = 5 \ ; \ O = 50}$

$$\Rightarrow \bar{x} = c.\bar{z} + O = 5.(-\frac{2}{50}) + 50 = 49'8$$

8) Ventajas e inconvenientes de la media aritmética

La media aritmética posee las tres características que se exigen a una medida de síntesis:

a) En su cálculo intervienen todos los valores de la distribución
b) Siempres es calculable
c) Es única

La media muestral viene a ser el **centro de gravedad** de la distribución de frecuencias, pues como hemos visto en la propiedad 1), es $\sum_{i=1}^{k} (x_i - \bar{x}).n_i = 0$.

También es una ventaja la propiedad 2):

$$\sum_{i=1}^{k} (x_i - \bar{x})^2 . \frac{n_i}{N} < \sum_{i=1}^{k} (x_i - C)^2 . \frac{n_i}{N}, \ \forall \ C \neq \bar{x}$$

La media aritmética puede conducir a conclusiones poco atinadas cuando la variable estadística presente valores anormalmente extremos, inconveniente que, como veremos en su momento, no tiene la **mediana.**

2.3 MEDIA GEOMÉTRICA

La **media geométrica** de la distribución de frecuencias $(x_i;n_i)$ correspondiente a una muestra de tamaño "N" se denota "G", y es la raíz N-ésima del producto de los valores observados:

$$G=\sqrt[N]{x_1^{n_1}.x_2^{n_2}.....x_k^{n_k}}$$

Ejemplo

x_i	n_i
100	10
120	5
125	4
140	3
	22

$$\Rightarrow G=\sqrt[22]{100^{10}.120^{5}.125^{4}.140^{3}}=10^{2'05554}=113'6$$

$$\boxed{\log G=\frac{1}{22}.(10.\log 100+5.\log 120+4.\log 125+3.\log 140)=2'05554}$$

PROPIEDADES

1) El logaritmo de la media geométrica es la media aritmética de los logaritmos de los valores de la variable:

$$\log G=\log\sqrt[N]{x_1^{n_1}.x_2^{n_2}.....x_k^{n_k}}=\frac{1}{N}.\log\left(\prod_{i=1}^{k}x_i^{n_i}\right)=$$

$$\boxed{\log a^{b}=b.\log a}$$

$$=\frac{1}{N}.\sum_{i=1}^{k}\log x_i^{n_i}=\frac{1}{N}.\sum_{i=1}^{k}(\log x_i).n_i$$

$$\boxed{\log(p.q)=(\log p)+(\log q)}$$

2) La suma $S=\sum_{i=1}^{k}((\log x_i)-C)^2.n_i$ es mínima si $C=\log G$. En efecto:

$$\frac{dS}{dC}=-2.\sum_{i=1}^{k}((\log x_i)-C).n_i=0\Rightarrow\left(\sum_{i=1}^{k}n_i.\log x_i\right)-C.\sum_{i=1}^{k}n_i=0\Rightarrow$$

$$\Rightarrow\left(\sum_{i=1}^{k}n_i.\log x_i\right)-C.N=0\Rightarrow C=\frac{1}{N}.\sum_{i=1}^{k}n_i.\log x_i=\log G$$

El valor obtenido de "C" corresponde a un mínimo, pues:

$$\frac{d^2S}{dC^2}=2.\sum_{i=1}^{k}n_i=2.N>0.$$

3) El cuadrado de la media geométrica es la media geométrica de los cuadrados de los valores de la variable:

$$G^2=\left(\sqrt[N]{x_1^{n_1}.x_2^{n_2}.....x_k^{n_k}}\right)^2=\sqrt[N]{(x_1^2)^{n_1}.(x_2^2)^{n_2}.....(x_k^2)^{n_k}}$$

Ventajas e inconvenientes de la media geométrica

- En su cálculo intervienen todos los valores de la distribución de frecuencias.
- Por ser un producto, es menos sensible a los valores extremos que $\bar{x}$.
- Su significado estadístico es menos intuitivo que la media aritmética.
- La media geométrica es 0 si algún x_i es 0. Si hay valores negativos de x_i no siempre se puede calcular "G".

x_i	-2	2	3
n_i	3	2	5

$$\Rightarrow G = \sqrt[10]{(-2)^3 . 2^2 . 3^5} \equiv \sqrt[\text{par}]{\text{negativo}} \notin \Re$$

Suele emplearse "G" para promediar variables que tienen variaciones acumulativas (porcentajes, tasas, números índices): un capital C_0 colocado "n" años en capitalización compuesta a tantos unitarios de interés anual $i_1, \ldots, i_n$ produce un montante $C_0.(1+i_1).....(1+i_n)$; así el **tanto medio** "i" al que debe colocarse C_0 para producir igual montante en "n" años ha de ser tal que:

$$C_0.(1+i)^n = C_0.(1+i_1).....(1+i_n) \Rightarrow i = \sqrt[n]{(1+i_1).....(1+i_n)} - 1$$

2.4 MEDIA ARMÓNICA

La **media armónica** de la distribución de frecuencias $(x_i; n_i)$ correspondiente a una muestra de tamaño "N" se denota "H", y es la inversa de la media aritmética de los inversos de los valores observados:

$$H = \frac{N}{\frac{1}{x_1}.n_1 + ... + \frac{1}{x_k}.n_k} = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i}.n_i}$$

Suele emplearse para promediar rendimientos, velocidades, tiempos.

Ejemplo

x_i	n_i
20	10
21	5
23	4
25	3
	22

$$\Rightarrow H = \frac{22}{\frac{1}{20}.10 + \frac{1}{21}.5 + \frac{1}{23}.4 + \frac{1}{25}.3}$$

Ventajas e inconvenientes de la media armónica

- En su cálculo intervienen todos los valores de la distribución de frecuencias.
- A veces es más significativa que la media aritmética.
- Le afectan mucho los valores pequeños de la variable; por ello no debe emplearse en tal caso.
- No está definida si algún x_i es 0.

2.5 FÓRMULA DE FOSTER DE LOS PROMEDIOS

La **media de orden "r"** de la distribución de frecuencias $(x_i; n_i)$ correspondiente a una muestra de tamaño "N" se denota M(r), siendo:

$$M(r) = \sqrt[r]{\frac{1}{N} \cdot \sum_{i=1}^{k} x_i^r . n_i} = \left(\frac{1}{N} \cdot \sum_{i=1}^{k} x_i^r . n_i \right)^{1/r}$$

Ejemplo

x_i	n_i
2	5
3	7
4	8
	20

$$\Rightarrow M(6) = \left(\frac{1}{20} . \left(2^6 .5 + 3^6 .7 + 4^6 .8 \right) \right)^{1/6}$$

La media M(r) un promedio, pues su valor está comprendido entre el menor y el mayor valor observado en la muestra:

$$M(r) = \sqrt[r]{\frac{1}{N} \cdot \sum_{i=1}^{k} x_i^r . n_i} \leq \sqrt[r]{\frac{1}{N} \cdot \sum_{i=1}^{k} x_p^r . n_i} = \sqrt[r]{\frac{1}{N} . x_p^r \cdot \sum_{i=1}^{k} n_i} = \sqrt[r]{x_p^r} = x_p$$

si el mayor valor observado es x_p

$$M(r) = \sqrt[r]{\frac{1}{N} \cdot \sum_{i=1}^{k} x_i^r . n_i} \geq \sqrt[r]{\frac{1}{N} \cdot \sum_{i=1}^{k} x_q^r . n_i} = \sqrt[r]{\frac{1}{N} . x_q^r \cdot \sum_{i=1}^{k} n_i} = \sqrt[r]{x_q^r} = x_q$$

si el menor valor observado es x_q

- Es $M(-1) = \left(\frac{1}{N} \cdot \sum_{i=1}^{k} x_i^{-1} . n_i \right)^{-1} = \dfrac{N}{\sum_{i=1}^{k} \frac{1}{x_i} . n_i} = H$
- Es $M(1) = \frac{1}{N} \cdot \sum_{i=1}^{k} x_i . n_i = \bar{x}$
- Aunque M(r) no está definida si $r = 0$, su límite cuando $r \to 0$ es la media geométrica "G".
- Puede demostrarse que para cualquier distribución de frecuencias sucede que

$$H \leq G \leq \bar{x}$$

dándose la igualdad sólo si todos los valores observados son iguales.

Si la muestra es de tamaño "N" y los valores observados están ordenados de modo creciente, la **mediana** es el valor de la variable que a su derecha y a su izquierda deja al menos la mitad de la distribución. Se denota "Me".

• Cálculo de la mediana en tablas tipo I

Si "N" es impar, la mediana es el valor que ocupa el lugar $(N+1)/2$.
Si "N" es par ($\Rightarrow N/2$ es par), se suelen dar como medianas los valores de la variable correspondientes a $N/2$ y $(N/2)+1$ .... o su media aritmética, si ésta es un valor que puede tomar la variable estadística.

x_i	n_i
11	1
13	1
14	1
15	1
18	1
	$5 \equiv$ impar

$\Rightarrow Me = 14$;

x_i	n_i
11	1
13	1
14	1
15	1
18	1
21	1
	$6 \equiv$ par

$\Rightarrow Me = 14, 15$

• Cálculo de la mediana en tablas tipo II

Si **no existe** la frecuencia absoluta acumulada $N/2$, la mediana es el primer valor cuya frecuencia absoluta acumulada es mayor o igual a $N/2$.

x_i	n_i	N_i
12	15	15
13	20	35
14	25	60
15	20	80
16	20	100
	100	

$\Rightarrow Me = 14$

No existe la frecuencia absoluta acumulada $N/2 = 50$.
La primera frecuencia absoluta acumulada mayor que $N/2 = 50$
es la que corresponde a 14

Si **existe** la frecuencia absoluta acumulada $N/2$, la mediana es la media aritmética entre el correspondiente valor de la variable y el siguiente, si dicha media aritmética es un posible valor de la variable. En caso contrario, se toman como medianas los valores de la variable correspondientes a $N/2$ y $(N/2)+1$

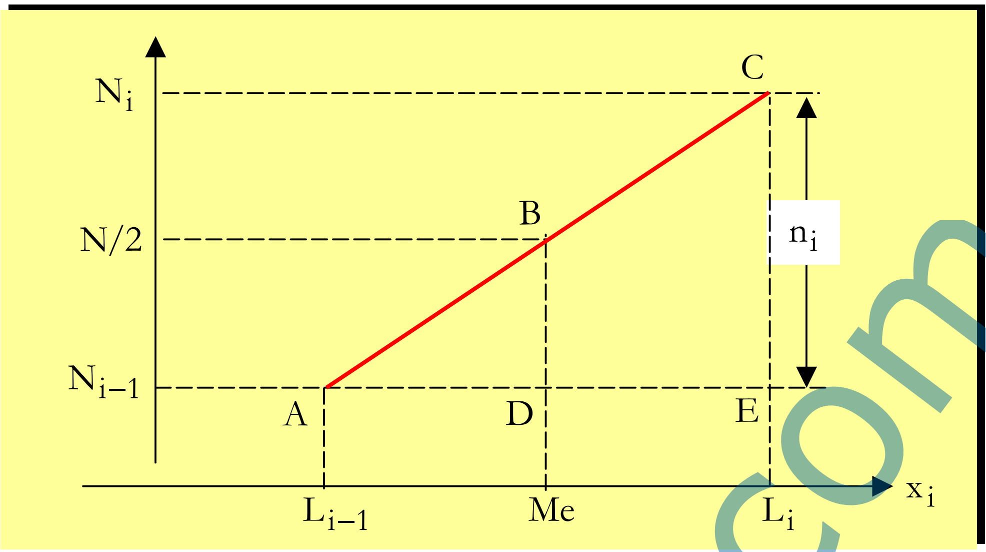
x_i	n_i	N_i
12	25	25
13	25	50
14	15	65
15	30	95
16	5	100
	100	

$\Rightarrow Me = (14+15)/2 = 14'5$

Hay una frecuencia absoluta acumulada que coincide con $N/2 = 50$
es la que corresponde a 14

• Cálculo de la mediana en tablas tipo III

Si los datos se agrupan en intervalos, determinamos el **intervalo mediano** siguiendo los criterios establecidos con anterioridad; después, para determinar la mediana "Me" suponemos que todos los valores del intervalo mediano se distribuyen de modo uniforme a lo largo de él; o sea, hacemos una **interpolación lineal.**



Por semejanza de triángulos, es $\frac{\overline{BD}}{\overline{AD}} = \frac{\overline{CE}}{\overline{AE}}$; o sea:

$$\frac{\frac{N}{2} - N_{i-1}}{Me - L_{i-1}} = \frac{n_i}{L_i - L_{i-1}} \Rightarrow Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i}.(L_i - L_{i-1})$$

Ejemplo

Intervalo	n_i	N_i
20 – 25	50	50
25 – 30	65	115
30 – 40	100	215
40 – 50	75	290
50 – 60	20	310
	310	

$$\Rightarrow Me = 30 + \frac{155 - 115}{100}.10 = 34$$

El intervalo mediano es el tercero, pues contiene a $N/2 = 310/2 = 155$

PROPIEDADES

1) La suma $\sum_{i=1}^{k} |x_i - C|.n_i$ de los valores absolutos de las desviaciones de los valores observados de la variable respecto de "C" es mínima si $C = Me$.

2) **Cambio de origen:** si "K" es una constante y $z_i = x_i + K$, la mediana de $(z_i; n_i)$ se obtiene al sumar "K" a la mediana de $(x_i; n_i)$.

3) **Cambio de escala:** si "K" es una constante y $z_i = K.x_i$, la mediana de $(z_i; n_i)$ se obtiene al multiplicar por "K" la mediana de $(x_i; n_i)$.

4) La mediana no depende de los valores extremos.

La **moda** es el valor de la variable que más veces se repite. Se denota "Mo".

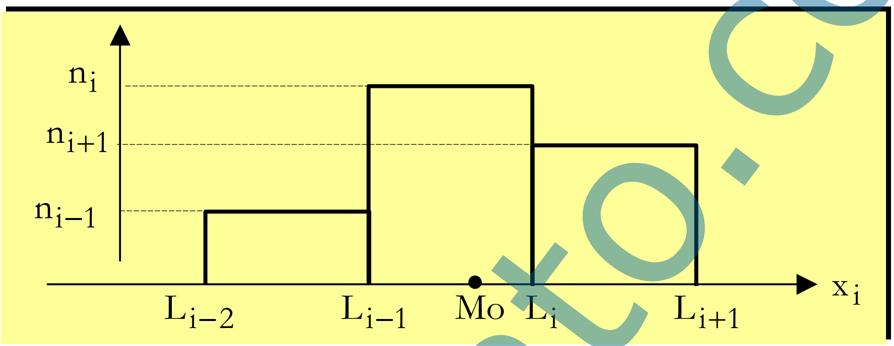
• Tablas tipo I y II

En la columna de frecuencias absolutas buscamos la mayor, y el correspondiente x_i es la moda, que puede no ser única.

• Tablas tipo III con intervalos de igual amplitud

1) Si los intervalos tienen igual amplitud, determinamos el **intervalo modal,** que es el de mayor frecuencia absoluta, y elegimos la moda atendiendo a cualquiera de los siguientes criterios:

   a) Tomamos como moda el extremo inferior del intervalo modal.
   b) Tomamos como moda el extremo superior del intervalo modal.
   c) Tomamos como moda la marca de clase del intervalo modal.
   d) Tomamos como moda el punto del intervalo modal $(L_{i-1};L_i]$ cuya distancia a los intervalos contiguos sea inversamente proporcional a la frecuencia absoluta de éstos.



$$\frac{Mo - L_{i-1}}{L_i - Mo} = \frac{n_{i+1}}{n_{i-1}} \Rightarrow$$

$$\Rightarrow (Mo - L_{i-1}).n_{i-1} = (L_i - Mo).n_{i+1} \Rightarrow$$

$$\Rightarrow Mo.(n_{i-1} + n_{i+1}) = L_i.n_{i+1} + L_{i-1}.n_{i-1} \Rightarrow$$

$$\Rightarrow Mo = \frac{L_i.n_{i+1} + L_{i-1}.n_{i-1}}{n_{i-1} + n_{i+1}} \Rightarrow$$

$$\boxed{\text{sumando y restando } L_{i-1}}$$

$$\Rightarrow Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}}.(L_i - L_{i-1})$$

Ejemplo

Intervalo	n_i
0 – 25	20
25 – 50	40
50 – 75	80
75 – 100	70

$$\Rightarrow Mo = 50 + \frac{70}{70 + 40}.(75 - 50)$$

$$\boxed{\text{El intervalo modal (el de mayor frecuencia absoluta) es el tercero}}$$

• Tablas tipo III con intervalos de distinta amplitud

Si los intervalos de clase son de distinta amplitud, la frecuencia absoluta de cada intervalo $(L_{i-1};L_i]$ no es representativa, pero si lo es la correspondiente densidad de frecuencia $d_i = n_i/(L_i - L_{i-1})$, que indica el número de observaciones por cada unidad de amplitud de $(L_{i-1};L_i]$. Así, el intervalo modal es el de mayor densidad de frecuencia ... y una vez determinado éste, para seleccionar la moda pueden aplicarse los criterios establecidos para intervalos de igual amplitud.

Si tomamos como moda el punto del intervalo modal $(L_{i-1};L_i]$ cuya distancia a los intervalos contiguos sea inversamente proporcional a la densidad de frecuencia absoluta de éstos, resulta ser:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}}.(L_i - L_{i-1})$$

Ejemplo

Intervalo	n_i	Amplitud	Densidad
0 – 25	20	25	$20/25 = 0'8$
25 – 50	140	25	$140/25 = 5'6$
50 – 100	180	50	$180/50 = 3'6$
100 – 150	40	50	$40/50 = 0'8$
150 – 200	20	50	$20/50 = 0'4$

$$\Rightarrow Mo = 25 + \frac{3'6}{0'8 + 3'6}.(50 - 25)$$

El intervalo modal (el de mayor densidad de frecuencia) es el segundo

La moda es la medida más representativa en distribuciones en escala nominal; o sea, con datos no susceptibles de ordenación

Obvio:

1) **Cambio de origen:** si "K" es una constante y $z_i = x_i + K$, la moda de $(z_i;n_i)$ se obtiene al sumar "K" a la moda de $(x_i;n_i)$.

2) **Cambio de escala:** si "K" es una constante y $z_i = K.x_i$, la moda de $(z_i;n_i)$ se obtiene al multiplicar por "K" la moda de $(x_i;n_i)$.

2.8 MEDIDAS DE POSICIÓN NO CENTRALES

El **cuantil** r-ésimo de orden "k" se denota $C_{r/k}$, y es el valor de la variable tal que hay r/k partes de la distribución que son menores o iguales que él y las restantes $(k-r)/k$ partes son mayores o iguales que él. Así, si la muestra es de tamaño "N", el cuantil $C_{r/k}$ deja a su izquierda al menos $r.N/k$ valores y a su derecha deja al menos $N-(r.N/k)$ valores.

Los **cuartiles** son los 3 valores de la variable que dividen a la distribución en 4 partes de frecuencia $N/4$. Se denotan Q_i $(i=1,2,3)$. Es $C_2 \equiv Me$.

Los **deciles** son los 9 valores de la variable que dividen a la distribución en 10 partes de frecuencia $N/10$. Se denotan D_i $(i=1,2,....,9)$. Es $D_5 \equiv Me$.

Los **percentiles** son los 99 valores de la variable que dividen a la distribución en 100 partes de frecuencia $N/100$. Se denotan P_i $(i=1,2,....,99)$. Es $P_{50} \equiv Me$.

• Tablas tipo II

Tara determinar $C_{r/k}$ determinamos la frecuencias absolutas acumuladas y seleccionamos el valor que ocupa el lugar $r.N/k$: Así:

$C_1 \equiv$ valor que ocupa el lugar $N/4$
$C_2 \equiv$ valor que ocupa el lugar $2.N/4$
$C_3 \equiv$ valor que ocupa el lugar $3.N/4$
$D_1 \equiv$ valor que ocupa el lugar $N/10$
$D_2 \equiv$ valor que ocupa el lugar $2.N/10$
::
$D_9 \equiv$ valor que ocupa el lugar $9.N/10$
$P_1 \equiv$ valor que ocupa el lugar $N/100$
::
$P_{99} \equiv$ valor que ocupa el lugar $99.N/100$

Ejemplo

x_i	n_i	N_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	

La frecuencia acumulada $3.40/4 = 30$ está en la tabla

$$\Rightarrow \begin{cases} C_1 = 4,\ C_2 = 5,\ C_3 = (6+7)/2 = 6'5 \\ D_1 = (2+3)/2 = 2'5,\ D_2 = (3+4)/2 = 3'5,\ D_3 = 4 \\ P_{40} = 5,\ P_{60} = P_{70} = 6 \end{cases}$$

• Tablas tipo III

El cuantil $C_{r/k}$ está en el primer intervalo cuya frecuencia absoluta acumulada sea mayor o igual a $r.N/k$. Una vez determinado dicho intervalo, el valor de $C_{r/k}$ se determina de igual modo que hicimos con la mediana:

$$C_{r/k} = L_{i-1} + \frac{\frac{r.N}{k} - N_{i-1}}{n_i}.(L_i - L_{i-1})$$

Ejemplo

Intervalo	n_i	N_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	

* Como $1.N/4 = 88/4 = 22$, el **primer cuarttil** está en el intervalo $50 - 56$, que es el primero con frecuencia absoluta acumulada mayor o igual a 22:

$$C_1 = L_{i-1} + \frac{\frac{1.N}{4} - N_{i-1}}{n_i}.(L_i - L_{i-1}) = 50 + \frac{22-15}{15}.(56-50) = 52'8$$

* Como $3.N/4 = 3.88/4 = 66$, el **tercer cuarttil** está en el intervalo $62 - 68$, que es el primero con frecuencia absoluta acumulada mayor o igual a 66:

$$C_3 = L_{i-1} + \frac{\frac{3.N}{4} - N_{i-1}}{n_i}.(L_i - L_{i-1}) = 62 + \frac{66-55}{18}.(68-62) = 65'67$$

* Como $40.N/100 = 40.88/100 = 35'2$, el **percentil 40** está en intervalo $56 - 62$, que es el primero con frecuencia absoluta acumulada mayor o igual a 35'2:

$$P_{40} = L_{i-1} + \frac{\frac{40.N}{100} - N_{i-1}}{n_i}.(L_i - L_{i-1}) =$$

$$= 56 + \frac{35'2-30}{25}.(62-56) = 57'25$$

* Como $90.N/100 = 90.88/100 = 79'2$, el **percentil 90** está en el intervalo $68 - 74$, que es el primero con frecuencia absoluta acumulada mayor o igual a 79'2:

$$P_{90} = L_{i-1} + \frac{\frac{90.N}{100} - N_{i-1}}{n_i}.(L_i - L_{i-1}) =$$

$$= 68 + \frac{79'2-73}{9}.(74-68) = 72'13$$

2.9 MOMENTOS

Los momentos de una distribución de frecuencias son unos números que la caracterizan plenamente: dos distribuciones son iguales si tienen iguales todos sus momentos .... y son tanto más parecidas cuanto mayor sea el número de momentos coincidentes en una y otra.

- El **momento de orden "r" respecto a un punto arbitrario** "O" se denota M_r:

$$M_r = \frac{1}{N}.\sum_{i=1}^{k} (x_i - O)^r.n_i$$

- El **momento de orden "r" respecto al origen** se denota a_r, y corresponde al caso en que el punto "O" es el origen:

$$a_r = \frac{1}{N}.\sum_{i=1}^{k} x_i^r.n_i$$

Por tanto:

$$a_0 = \frac{1}{N}.\sum_{i=1}^{k} x_i^0.n_i = \frac{1}{N}.\sum_{i=1}^{k} n_i = 1$$

$$a_1 = \frac{1}{N}.\sum_{i=1}^{k} x_i.n_i = \overline{x}$$

$$a_2 = \frac{1}{N}.\sum_{i=1}^{k} x_i^2.n_i$$

$$a_3 = \frac{1}{N}.\sum_{i=1}^{k} x_i^3.n_i$$

- El **momento central de orden "r"** es el momento de orden "r" respecto a la media aritmética $\overline{x} \equiv a_1$, y se denota m_r; o sea, corresponde al caso en que "O" es la media aritmética:

$$m_r = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \overline{x})^r.n_i$$

Por tanto:

$$m_0 = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \overline{x})^0.n_i = \frac{1}{N}.\sum_{i=1}^{k} n_i = 1$$

$$m_1 = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \overline{x}).n_i = 0$$

$$m_2 = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \overline{x})^2.n_i$$

$$m_3 = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \overline{x})^3.n_i$$

- **Los momentos respecto a la media pueden expresarse en función de los momentos respecto al origen:**

$$m_r = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^r.n_i = \frac{1}{N}.\sum_{i=1}^{k} \left(\sum_{h=0}^{r} (-1)^h.\binom{r}{h}.x_i^{r-h}.\bar{x}^h \right).n_i =$$

$$\boxed{\text{Newton: } (x_i - \bar{x})^r = \sum_{h=0}^{r} (-1)^h.\binom{r}{h}.x_i^{r-h}.\bar{x}^h}$$

$$= \frac{1}{N}.\sum_{h=0}^{r} \left(\sum_{i=1}^{k} (-1)^h.\binom{r}{h}.x_i^{r-h}.\bar{x}^h \right).n_i =$$

$$= \frac{1}{N}.\sum_{h=0}^{r} \left((-1)^h.\binom{r}{h}.x_i^{r-h}.\bar{x}^h.\left(\sum_{i=1}^{k} x_i^{r-h} \right).n_i \right) =$$

$$\boxed{\frac{1}{N}.\sum_{i=1}^{k} x_i^{r-h}.n_i = a_{r-h} \quad ; \quad \bar{x} = a_1}$$

$$= \sum_{h=0}^{r} (-1)^h.\binom{r}{h}.a_1^h.a_{r-h}$$

<u>Por ejemplo:</u>

$$m_2 = \sum_{h=0}^{2} (-1)^h.\binom{2}{h}.a_1^h.a_{2-h} = a_2 - 2.a_1^2 + a_1^2 = a_2 - a_1^2$$

$$m_3 = \sum_{h=0}^{3} (-1)^h.\binom{3}{h}.a_1^h.a_{3-h} = a_3 - 3.a_2.a_1 + 3.a_1^3$$

$$m_4 = \sum_{h=0}^{4} (-1)^h.\binom{4}{h}.a_1^h.a_{4-h} = a_4 - 4.a_3.a_1 + 6.a_2.a_1^2 - 3.a_1^4$$

2.10 MEDIDAS DE DISPERSIÓN

Las medidas de dispersión, que **permiten saber en qué grado las medidas de posición centrales son representativas de la información contenida en la distribución de frecuencias,** expresan la separación de los valores de la distribución respecto de la medida de posición central.

Se dice que una medida de dispersión es **absoluta** si se expresa en las mismas unidades que la variable estadística. Si una medida de dispersión es adimensional, se dice que es **relativa.**

Medidas de dispersión absolutas

1) **Recorrido:** es la diferencia entre el mayor y el menos valor observado.

$$R = \text{Máx.}\{x_i\} - \text{Min.}\{x_i\}$$

Por ejemplo, si observamos la edad de 50 personas y $R = 5$, podemos decir que, en principio, hay poca dispersión en las edades.

2) **Recorrido intercuartílico:** es la diferencia entre los cuartiles tercero y primero.

$$R_I = C_3 - C_1$$

Indica que la mitad de los valores centrales de la distribución están en un intervalo de amplitud R_I. Si R_I es pequeño, podemos intuir que la dispersión es pequeña.

> El recorrido y el recorrido intercuartílico son medidas de dispersión que **no están referidas a ningún promedio,** al contrario que las siguientes.

3) **Desviación media respecto a la media aritmética:** es la media aritmética de los valores absolutos de las desviaciones de la variable respecto a la media aritmética.

$$D_{\bar{x}} = \frac{1}{N} . \sum_{i=1}^{k} |x_i - \bar{x}| . n_i$$

La media aritmética es más representativa cuanto menor sea $D_{\bar{x}}$.

Por ejemplo, si en una muestra de tamaño 300 de las notas (de 0 a 10) en el examen de Estadística es $D_{\bar{x}} = 0'7$, eso indica que hay muchos alumnos con nota próxima a la media, pero no así si $D_{\bar{x}} = 2'7$.

4) **Desviación media respecto a la mediana (desviación mediana):** es la media aritmética de los valores absolutos de las desviaciones de la variable respecto a la mediana.

$$D_{Me} = \frac{1}{N} . \sum_{i=1}^{k} |x_i - Me| . n_i$$

Sabemos que $\frac{1}{N} . \sum_{i=1}^{k} |x_i - C| . n_i$ es mínimo si $C = Me$; por tanto: $D_{Me} \le D_{\bar{x}}$.

La mediana es más representativa cuanto menor sea D_{Me}.

5) **Varianza:** es el momento de segundo orden respecto de la media aritmética.

$$S^2 \equiv m_2 = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^2.n_i$$

La media aritmética es más representativa cuanto menor sea la varianza.

Propiedades

a) Es $S^2 \geq 0$, siendo $S^2 = 0$ sólo si todos los valores observados son iguales.

b) La varianza es la medida cuadrática de dispersión óptima, pues sabemos que $\frac{1}{N}.\sum_{i=1}^{k} (x_i - C)^2.n_i$ es mínimo si $C = \bar{x}$.

c) Es $S^2 = a_2 - a_1^2$.

d) La varianza no es sensible a los cambios de origen, pues si $z_i = x_i + K$, es:

$$S_Z^2 = \frac{1}{N}.\sum_{i=1}^{k} (z_i - \bar{z})^2.n_i = $$

$$\boxed{\text{es } z_i = x_i + K \text{ y } \bar{z} = \bar{x} + K}$$

$$= \frac{1}{N}.\sum_{i=1}^{k} (x_i + K - (\bar{x} + K))^2.n_i = \frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^2.n_i = S_X^2$$

e) La varianza es sensible a los cambios de escala, pues si $z_i = K.x_i$, es:

$$S_Z^2 = \frac{1}{N}.\sum_{i=1}^{k} (z_i - \bar{z})^2.n_i = $$

$$\boxed{z_i = K.x_i \ ; \ \bar{z} = K.\bar{x}}$$

$$= \frac{1}{N}.\sum_{i=1}^{k} (K.x_i - K.\bar{x})^2.n_i = K^2.\left(\frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^2.n_i\right) = K^2.S_X^2$$

6) **Desviación típica:** es la raíz cuadrada (con signo positivo) de la varianza.

$$S = +\sqrt{S^2} = +\sqrt{\frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^2.n_i}$$

Propiedades

a) Es $S \geq 0$, siendo $S = 0$ sólo si todos los valores observados son iguales.

b) La desviación típica es una medida cuadrática de dispersión óptima.

c) Es $S = +\sqrt{a_2 - a_1^2}$.

d) La desviación típica no es sensible a los cambios de origen.

e) La varianza es sensible a los cambios de escala, pues si $z_i = K.x_i$, es:

$$S_Z = K.S_X$$

7) **Varianza por estratos**

Considera que una distribución que se particiona en las siguientes dos

x_1	x_2	.....	x_h
n_1	n_2	.....	n_h

;

x_{h+1}	x_{h+2}	.....	x_k
n_{h+1}	n_{h+2}	.....	n_k

siendo $N_1 = \sum_{i=1}^{h} n_i$ y $N_2 = \sum_{i=h+1}^{k} n_i$. Sabemos que si $(\overline{x})_r$ es la media aritmética correspondiente a la distribución de frecuencias del r-ésimo estrato $(r = 1,2)$, la media aritmética $\overline{x}$ de la distribución dada es $\overline{x} = \dfrac{N_1.(\overline{x})_1 + N_2.(\overline{x})_2}{N}$.

Denotemos $(a_2)_r$ al momento de orden 2 que respecto al origen tiene la distribución de frecuencias del i-ésimo estrato $(r = 1,2)$:

$$(a_2)_1 = \sum_{i=1}^{h} x_i^2.\frac{n_i}{N_1} \;;\; (a_2)_2 = \sum_{i=h+1}^{k} x_i^2.\frac{n_i}{N_2}$$

El momento de orden 2 de la distribución dada respecto al origen es:

$$a_2 = \sum_{i=1}^{k} x_i^2.\frac{n_i}{N} = \frac{\sum_{i=1}^{h} x_i^2.n_i + \sum_{i=h+1}^{k} x_i^2.n_i}{N} =$$

$$= \frac{N_1.\sum_{i=1}^{h} x_i^2.\frac{n_i}{N_1} + N_2.\sum_{i=h+1}^{k} x_i^2.\frac{n_i}{N_2}}{N} = \frac{N_1.(a_2)_1 + N_2.(a_2)_2}{N}$$

Lo mismo vale para todo orden: $a_r = \sum_{i=1}^{k} x_i^r.\dfrac{n_i}{N} = \dfrac{N_1.(a_r)_1 + N_2.(a_r)_2}{N}$

Así, la varianza de la distribución dada, es:

$$S^2 = a_2 - (\overline{x})^2 = \frac{N_1.(a_2)_1 + N_2.(a_2)_1}{N} - \left(\frac{N_1.(\overline{x})_1 + N_2.(\overline{x})_2}{N}\right)^2$$

Ejemplo

	Tamaño	Media	Varianza
Muestra 1	100	11	180
Muestra 2	150	12	200

$$S^2 = a_2 - (\overline{x})^2 = 326'8 - 11'6^2 = 192'24$$

$$\overline{x} = \frac{N_1.(\overline{x})_1 + N_2.(\overline{x})_2}{N} = \frac{100.11 + 150.12}{100 + 150} = 11'6$$

$$a_2 = \frac{N_1.(a_2)_1 + N_2.(a_2)_2}{N} = \frac{100.301 + 150.344}{100 + 150} = 326'8$$

$$(a_2)_1 = S_1^2 + (\overline{x})_1^2 = 180 + 11^2 = 301$$

$$(a_2)_2 = S_2^2 + (\overline{x})_2^2 = 200 + 12^2 = 344$$

Medidas de dispersión relativas (índices de dispersión)

Son adimensionales y miden la dispersión sin tener en cuenta en qué unidades se expresa la variable estadística.

1) **Coeficiente de apertura:** $A = \dfrac{\text{Máx.}\{x_i\}}{\text{Min.}\{x_i\}}$

2) **Recorrido relativo:** $R = \dfrac{Re}{\overline{x}} = \dfrac{\text{Máx.}\{x_i\} - \text{Min.}\{x_i\}}{\overline{x}}$

3) **Recorrido semiintercuartílico:** $R_s = \dfrac{C_3 - C_1}{C_3 + C_1}$

4) **Coeficiente de variación de Pearson:** $V = \dfrac{S}{\overline{x}}$ ó $V = \dfrac{S}{\overline{x}} \times 100$

Es la medida de dispersión relativa más empleada, y expresa el número de veces que la desviación típica contiene a la media aritmética. Cuanto mayor sea "V", menor es la representatividad de $\overline{x}$. Si $S = 0$ ($\Leftrightarrow$ todos los valores observados son iguales), es $V = 0$ y la representatividad de $\overline{x}$ es máxima.

Si $\overline{x}$ es próxima a cero, el valor de "V" se dispara hacia infinito, por lo que no conviene emplear el coeficiente de variación de Pearson como medida de dispersión relativa, ya que podría conducir a conclusiones equivocadas.

Observa: en la determinación de "V" intervienen todos los valores de muestra.

Hay quien dice que el coeficiente de variación de Pearson es **insensible a los cambios de escala,** salvo en el signo: si $z_i = K.x_i$, es:

$$V_Z = \frac{S_Z}{\overline{z}} = \frac{|K|.S_X}{K.\overline{x}} = \begin{cases} V_X & \text{si } K > 0 \\ -V_X & \text{si } K < 0 \end{cases}$$

$$\boxed{S_Z = |K|.S_X \;\; ; \;\; \overline{z} = K.\overline{x}}$$

El coeficiente de variación de Pearson es **sensible a los cambios de origen,** pues si $z_i = x_i + K$, es:

$$V_Z = \frac{S_Z}{\overline{z}} = \frac{S_X}{\overline{x} + K}$$

$$\boxed{S_Z = S_X \;\; ; \;\; \overline{z} = \overline{x} + K}$$

5) **Índice de variación respecto a la mediana:**

$$V_{Me} = \frac{D_{Me}}{Me} = \frac{\frac{1}{N}.\sum_{i=1}^{k} |x_i - Me|.n_i}{Me}$$

Cuanto mayor sea V_{Me}, menor es la representatividad de la mediana.

Medidas de asimetría

Permiten establecer el grado de simetría de la distribución de frecuencias.

1) **Coeficiente de asimetría de Fisher:** $g_1 = \dfrac{m_3}{S^3} = \dfrac{\frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^3.n_i}{\left(\frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^2.n_i\right)^{3/2}}$

- Si $g_1 = 0$, la distribución es simétrica respecto a $\bar{x}$; así: $\bar{x} = Mo = Me$.
- Si $g_1 > 0$, la distribución es asimétrica positiva o hacia la derecha (se **extiende** hacia la derecha) ..... y en tal caso es $Mo < Me < \bar{x}$.
- Si $g_1 < 0$, la distribución es asimétrica negativa o hacia la izquierda (se **extiende** hacia la izquierda) ..... y en tal caso es: $\bar{x} < Me < Mo$.

Ojo!, puede ser $g_1 = 0$ sin que la distribución sea simétrica.
El coeficiente g_1 es **insensible a cambios de origen y de escala.**

2) **Coeficiente de asimetría de Pearson**

Se emplea para distribuciones campaniformes (es decir, unimodales), y moderadamente asimétricas

$$A_p = \frac{\bar{x} - Mo}{S}$$

- Si $A_p > 0$, la distribución es asimétrica positiva o hacia la derecha.
- Si $A_p < 0$, la distribución es asimétrica negativa o hacia la izquierda.

3) **Coeficiente de asimetría de Bowley:** $A_B = \dfrac{C_1 + C_3 - 2.Me}{C_3 - C_1}$

Es nulo si la distribución es simétrica. Su signo coincide con el de la asimetría.

Medidas de apuntamiento o curtosis

Se emplea para medir el **aplastamiento/apuntamiento** de distribuciones campaniformes (es decir, unimodales), simétricas o moderadamente asimétricas.

Coeficiente de apuntamiento o curtosis

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{\frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^4.n_i}{\left(\frac{1}{N}.\sum_{i=1}^{k} (x_i - \bar{x})^2.n_i\right)^{2}} - 3$$

- Si $g_2 = 0$, la distribución se dice mesocúrtica.
- Si $g_2 > 0$, la distribución se dice leptocúrtica (más apuntada que la normal).
- Si $g_2 < 0$, la distribución se dice platicúrtica (menos apuntada que la normal).

El coeficiente de variación g_2 es **insensible a cambios de origen y de escala.**

2.12 MEDIDAS DE CONCENTRACIÓN

Permiten establecer el **grado de equidad** en el reparto se salarios, rentas, etc.

Sea $(x_i;n_i)$ la distribución de frecuencias de los salarios de una empresa, estando los salarios ordenados de modo creciente.

Construimos las siguiente tabla:

$x_i.n_i \equiv$ salario total percibido por los n_i individuos con salario x_i

$u_i \equiv$ Salario total percibido por los N_i individuos con salario $\leq x_i$

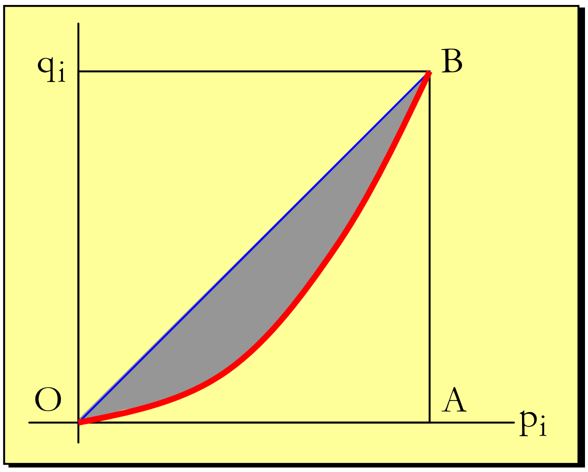
Porcentaje de la masa salarial u_k percibido por los N_i individuos con salario $\leq x_i$

Porcentaje de individuos con salario $\leq x_i$

x_i	n_i	$x_i.n_i$	N_i	$u_i = x_1.n_1 + ... + x_i.n_i$	$p_i = \frac{N_i}{N}.100$	$q_i = \frac{u_i}{u_k}.100$
x_1	n_1	$x_1.n_1$	N_1	u_1	p_1	q_1
x_2	n_2	$x_2.n_2$	N_2	u_2	p_2	q_2
x_3	n_3	$x_3.n_3$	N_3	u_3	p_3	q_3
:	:	:	:	:	:	:
x_i	n_i	$x_i.n_i$	N_i	u_i	p_i	q_i
:	:	:	:	:	:	:
x_k	n_k	$x_k.n_k$	N_k	u_k	p_k	q_k
	N	u_k				

$u_k \equiv$ masa salarial total

Dibujando un cuadrado de lado unidad, al representar los puntos $(p_1;q_i)$ obtenemos la llamada **curva de Lorenz,** que está por debajo de la diagonal OB del cuadrado, pues los salarios están ordenados de modo creciente. El **índice de concentración de Gini** es



$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

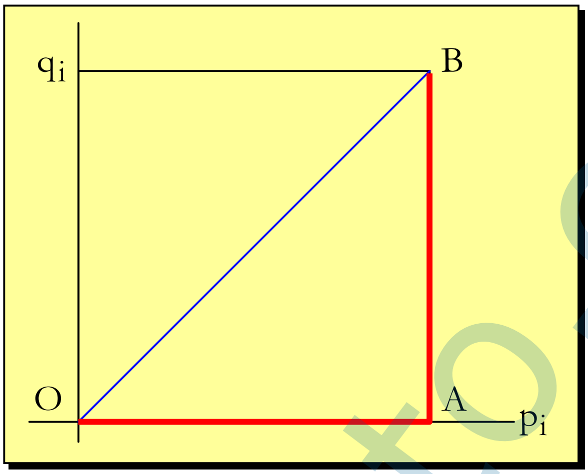
El valor de I_G coincide aproximadamente con el cociente entre el área sombreada y el área del triángulo OAB.

Si la equidad del reparto de la masa salarial es máxima (o sea, todos los trabajadores tienen igual sueldo), entonces $p_i = q_i$, por lo que $I_G = 0$. Así, la curva de Lorenz coincide con la diagonal OB.

Si la equidad del reparto es mínima $(x_1 = x_2 = \ldots = x_{k-1} = 0 \text{ y } x_k \neq 0)$, entonces $q_1 = q_2 = \ldots = q_{k-1} = 0$, por lo que:

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{\sum_{i=1}^{k-1} (p_i - 0)}{\sum_{i=1}^{k-1} p_i} = 1$$

En tal caso, la curva de Lorenz es como se indica a continuación.



El índice de Gini es insensible a los cambios de escala, no cambia si todos los salarios varían en la misma proporción.

Ejemplo

x_i	n_i	$x_i.n_i$	N_i	u_i	$p_i = \frac{N_i}{N}.100$	$q_i = \frac{u_i}{u_k}.100$	$p_i - q_i$
800	20	16000	20	16000	20	3'74	16'26
1000	10	10000	30	26000	30	6'07	23'93
1200	10	12000	40	38000	40	8'88	31'12
1500	10	15000	50	53000	50	12'38	37'62
7500	50	375000	100	428000	×	×	×
	100	428000			140		108'93

$$I_G = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i} = \frac{108'93}{140} = 0'78$$