ANÁLISIS DE DATOS UNIDIMENSIONALES

TABLAS DE FRECUENCIAS Y REPRESENTACIONES GRÁFICAS
MEDIDAS DE POSICIÓN
- MEDIDAS DE TENDENCIA CENTRAL
  - MEDIA ARITMÉTICA
  - OTRAS MEDIAS: GEOMÉTRICA.ARMÓNICA.MEDIA GENERAL
  - MEDIANA (Me)
  - MODA (Mo).
- MEDIDAS DE POSICIÓN NO CENTRALES (CUANTILES)

MEDIDAS DE DISPERSIÓN
- MEDIDAS DE DISPERSIÓN ABSOLUTA VARIANZA
- MEDIDAS DE DISPERSIÓN RELATIVA

MOMENTOS
- MOMENTOS ORDINARIOS (RESPECTO AL ORIGEN)
- MOMENTOS CENTRALES
- RELACIÓN ENTRE MOMENTOS CENTRALES Y ORDINARIOS

TRANSFORMACIONES LINEALES DE UNA VARIABLE ESTADÍSTICA
- TIPIFICACIÓN

MEDIDAS DE FORMA
- MEDIDAS DE ASIMETRÍA
- MEDIDAS DE CURTOSIS

TABLAS DE FRECUENCIAS Y REPRESENTACIONES GRÁFICAS

Desarrollo con un ejemplo: gasolina repostada en una gasolinera por 16 clientes:

DATOS
TABLA DE FRECUENCIAS /
TABLA DE FRECUENCIAS DATOS AGRUPADOS
REPRESENTACIONES GRÁFICAS: diagrama de barras, acumulativo, histograma, polígono acumulativo.

Información transformada en tabla:

DATOS	
x (litros de gasolina)	n (número de clientes)
23	1
12	1
34	1
23	1
12	1
15	1
10	1
45	1
45	1
12	1
23	1
23	1
15	1
45	1
43	1
30	1
Total de clientes	16

Tabla de frecuencias

n = número de valores distintos de la variable.
$\mathbf{X_i}$ = cada uno de los n valores de la variable
$\mathbf{n_i}$ =frecuencias absolutas , nº de individuos que poseen valor de la variable igual a X_i
N= número total de individuos
$\mathbf{f_i}$ = frecuencias relativas ,tanto por uno de individuos que poseen valor de la variable igual a X_i ,
cálculo $f_i = n_i / N$
$\mathbf{N_i}$ = frecuencias absolutas acumuladas , numero de individuos que tienen un valor de la variable igual o inferior a X_i ,
cálculo : $N_1=n_1$, $N_2=N_1+n_2$, $N_i= N_{i-1}+n_i$,….. $N_n=N$
$\mathbf{F_i}$ = frecuencias relativas acumuladas, tanto por uno de individuos que tienen un valor

de la variable igual o inferior a X_i,
cálculo: $F_1=f_1$, $F_2=F_1+f_2$, $F_i= F_{i-1}+f_i$,….. $F_n=1$

Individuo /caso	X_i	n_i	f_i	N_i	F_i
1	10	1	0,0625	1	0,0625
2	12	3	0,1875	4	0,25
3	15	2	0,125	6	0,375
..	23	4	0,25	10	0,625
	30	1	0,0625	11	0,6875
6	34	1	0,0625	12	0,75
7	43	1	0,0625	13	0,8125
8=n	45	3	0,1875	16	1
	suma	N=16	1		

Tabla de frecuencias con los valores agrupados en intervalos

n = número de intervalos

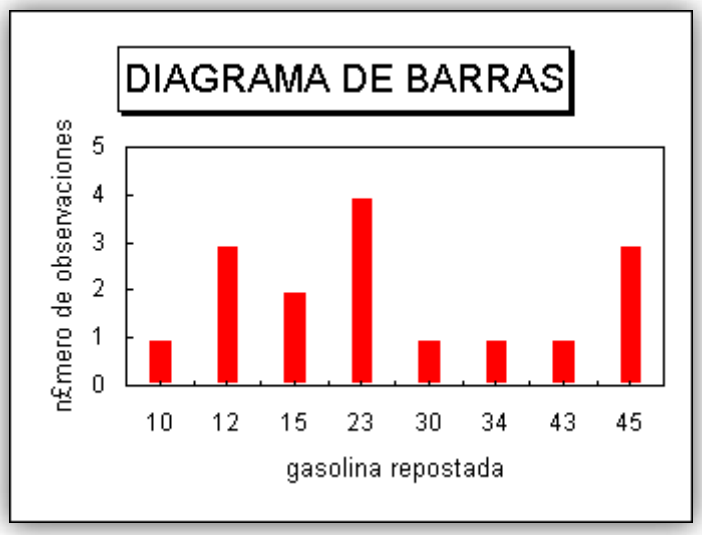
$\mathbf{c_i}$ =amplitud de intervalo, L_i menos L_{i-1}

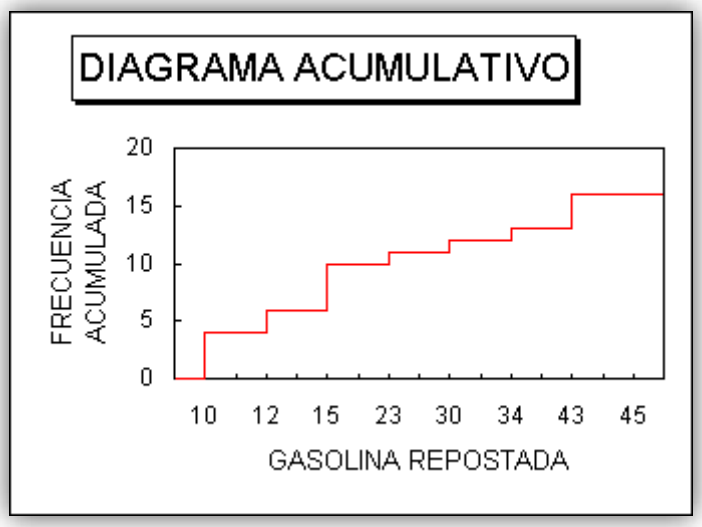
$\mathbf{x_i}$ = marca de clase, semisuma de los extremos de los intervalos

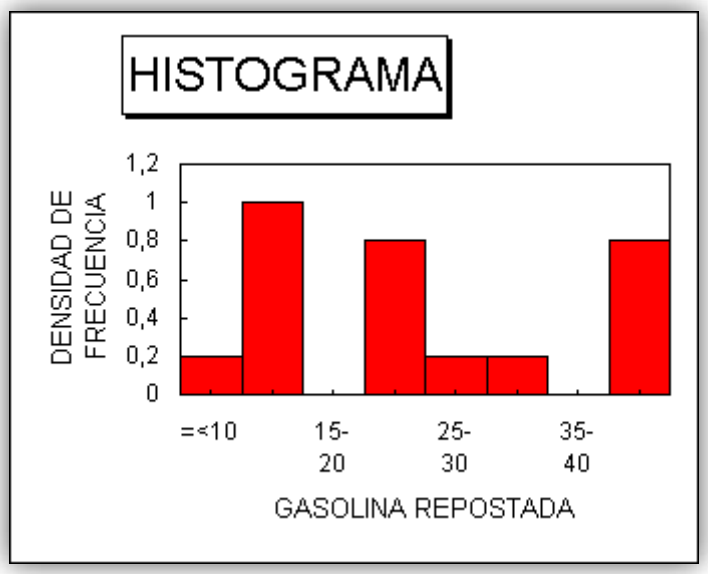
$\mathbf{h_i}$ = altura del histograma/densidad de frecuencia, frecuencia repartida en amplitud de intervalo, $h_i = n_i / c_i$

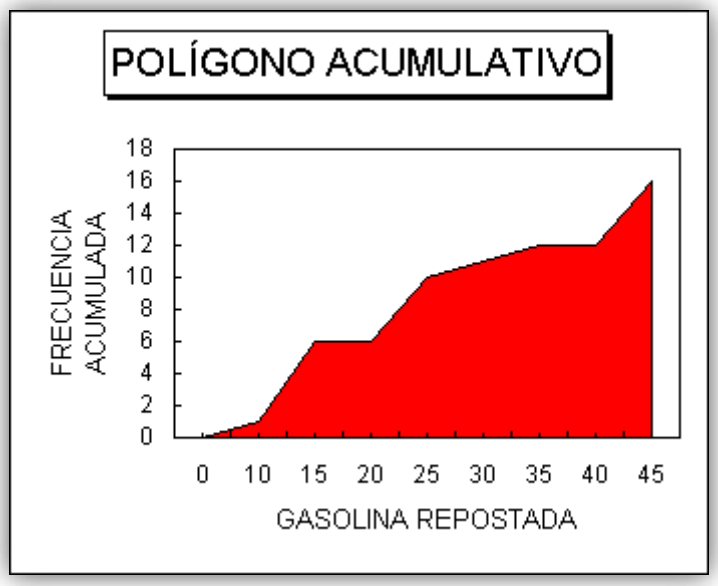
Intervalos de valores de la variable $]L_{i-1} – Li]$	c_i Amplitud del intervalo	Marca de clase x_i	n_i	f_i	N_i	F_i	h_i (Altura del histograma)
=<10]	tómese 5	7,5	1	0,0625	1	0,0625	0,2
]10-15]	5	12,5	5	0,3125	6	0,375	1
]15-20]	5	17,5	0	0	6	0,375	0
]20-25]	5	22,5	4	0,25	10	0,625	0,8
]25-30]	5	27,5	1	0,0625	11	0,6875	0,2
]30-35]	5	32,5	1	0,0625	12	0,75	0,2
]35-40]	5	27,5	0	0	12	0,75	0
]40-45]	5	42,5	4	0,25	16	1	0,8
SUMA			16	1			

Representaciones gráficas






HISTOGRAMA
DENSIDAD DE FRECUENCIA
1,2
1
0,8
0,6
0,4
0,2
0
=<10
15-20
25-30
35-40
GASOLINA REPOSTADA


POLÍGONO ACUMULATIVO
FRECUENCIA ACUMULADA
18
16
14
12
10
8
6
4
2
0
0
10
15
20
25
30
35
40
45
GASOLINA REPOSTADA

MEDIDAS DE POSICIÓN: indicadores de conjunto que informan de "por dónde" se sitúan los datos. Las hay de tendencia central y medidas no centrales.

M. TENDENCIA CENTRAL

MEDIA ARITMÉTICA:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_n n_n}{N}$$

PROPIEDADES:

1) la media de las desviaciones de la variable respecto a la media aritmética es cero.

$$\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x}) n_i = 0 \quad \text{ya que :}$$

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

2) la media aritmética minimiza la expresión:

$$A(k) = \frac{1}{N}\sum_{i=1}^{n}(x_i - k)^2 n_i$$

que es mínima para el valor de *k*,

$$k = \bar{x}$$

3) la media aritmética conserva la linealidad:

Dada la distribución de la variable estadística x , y dada una nueva variable y construida como:

$$y_i = a + b x_i$$

la media de la variable y será:

$$\bar{y} = a + b\bar{x}$$

4) la media aritmética de la agrupación de 2 o más conjuntos de datos es la media (ponderada por las observaciones) de las medias de los distintos conjuntos.

Conjunto de datos	1º	2º	3º	Global (agregado de los tres)

Nº de observaciones	N_1	N_2	N_3	$N_1 + N_2 + N_3$
Media	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \bar{x}_3 N_3}{N_1 + N_2 + N_3}$

OTRAS MEDIAS: GEOMÉTRICA.ARMÓNICA.MEDIA GENERAL

M .GEOMÉTRICA. Tiene utilidad para promediar tasas, Números índices y porcentajes. Típicamente se emplea en el cálculo del tipo de interés medio (equivalente)

$$G = \sqrt[N]{\prod_{i=1}^{n} x_i^{n_i}} = \sqrt[N]{x_1^{n_1} + x_2^{n_2} + \ldots + x_n^{n_n}}$$

$$\log G = \overline{(\log X)} = \frac{1}{N}\sum_{i=1}^{n} (\log x_i)\, n_i$$

M . ARMÓNICA es útil para promediar velocidades, tiempos

$$H = \frac{N}{\frac{1}{x_1} n_1 + \frac{1}{x_2} n_2 + \ldots + \frac{1}{x_n} n_n}$$

$$\frac{1}{H} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{1}{x_i} n_i\right) = \overline{\left(\frac{1}{X}\right)}$$

M . GENERAL DE ORDEN m

$$M_{(m)} = \sqrt[m]{\frac{1}{N}\sum_{i=1}^{n} x_i^m\, n_i}$$

para m= -1 tenemos la media armónica , para m=0, la geométrica,

para m= 1, la media aritmética, para m=2 la media cuadrática ,etc.

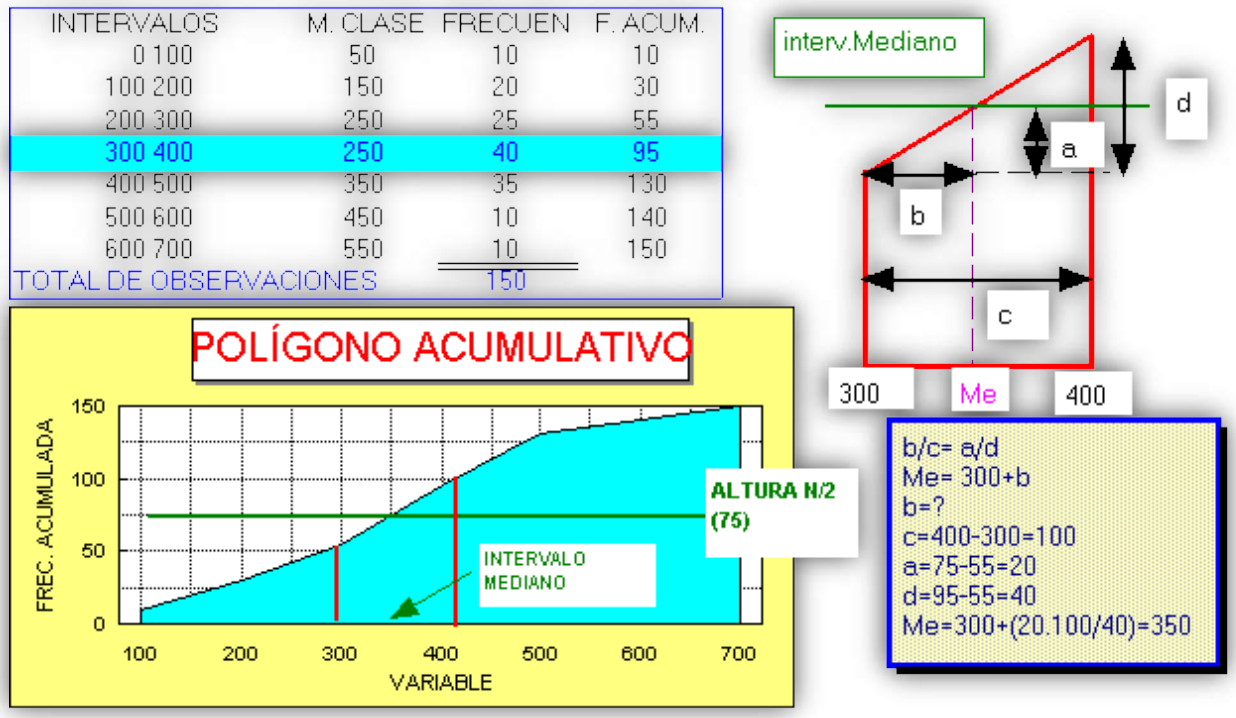
MEDIANA (Me) *es el valor de la variable que, ordenados los datos de menor a mayor, deja a izquierda y derecha el mismo número de observaciones. El valor de la variable que tiene una frecuencia acumulada de N/2.*

En el caso de una distribución "no agrupada" su determinación no presenta problemas.

En el caso de una distribución con los valores agrupados por intervalos: habrá de detectarse primero el "**intervalo mediano**" (aquel intervalo en el que se produzca una

acumulación de frecuencia de N/2).Después obtendremos el valor "intrapolando" gráficamente, suponiendo que la distribución de frecuencias dentro del intervalo es "**uniforme**":

Una vez detectado el intervalo mediano, aquél en el que la frecuencia acumulada llega a sobrepasar la mitad del total de las observaciones, consideraremos como valor de la mediana la abscisa correspondiente al punto de corte del polígono acumulativo y la recta $Y=N/2$ .La determinación de ese valor puede resolverse fácilmente por semejanza de triángulos:



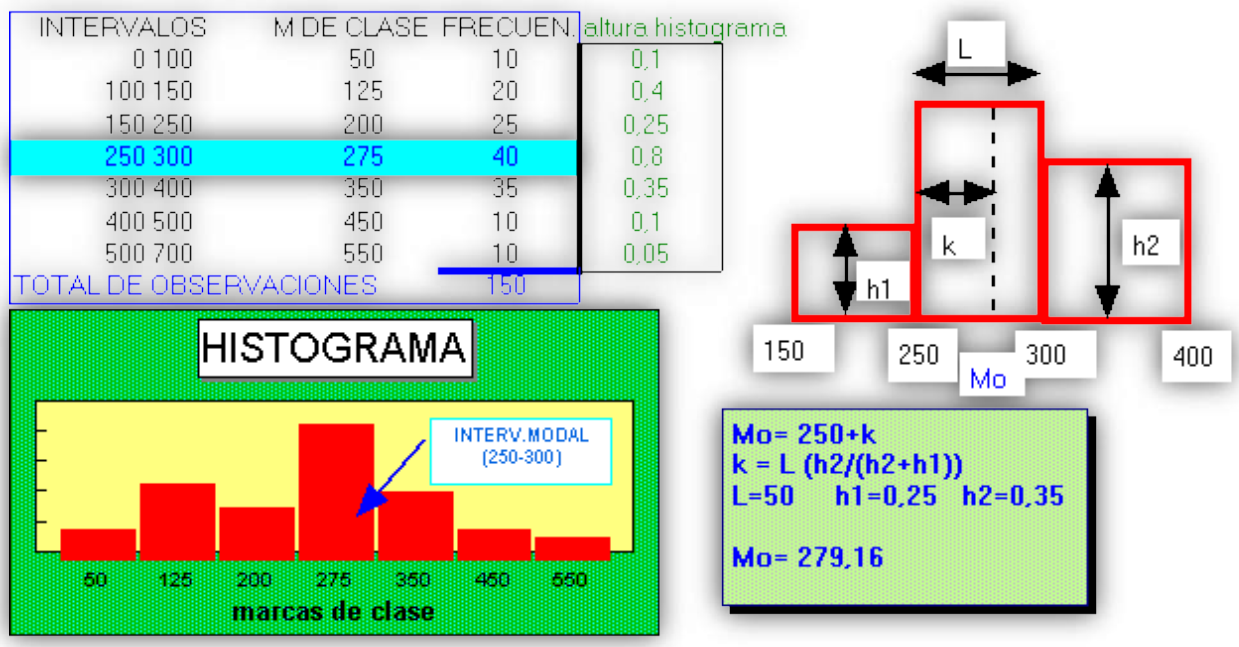
MODA (Mo). *La moda es el valor de la variable que tiene mayor frecuencia. (Puede haber más de una)*

En el caso de una distribución de valores sin agrupar su determinación no presenta problemas.

En el caso de una distribución con los valores agrupados por intervalos se requiere:

1) determinar el "**intervalo modal**" que es el de **mayor densidad de frecuencia** (mayor altura en el histograma) (sólo coincide con el de mayor frecuencia si los intervalos tienen la misma amplitud)

2) se determina el valor concreto de la moda, considerando que será un valor interno del intervalo modal que estará más cerca del intervalo vecino con mayor densidad de frecuencia y más lejos del intervalo vecino con menor densidad de frecuencia (de forma proporcional):



MEDIDAS DE POSICIÓN NO CENTRALES (CUANTILES)

CUARTILES HAY 3: Qj es el valor de la variable cuya frecuencia acumulada es j N/4

DECILES HAY 9: Dj es el valor de la variable cuya frecuencia acumulada es j N/ 10

CENTILES (PERCENTILES) HAY 99: Cj es el valor de la variable cuya frecuencia acumulada es j N/ 100

Su determinación es en todo idéntica a la de la mediana.

MEDIDAS DE DISPERSIÓN miden cuánto se dispersan (en términos globales) los valores de la variable respecto de alguna medida de tendencia central.

Además de indicar la variabilidad (dispersión) de la distribución sirven para matizar la representatividad de las medidas de tendencia central).

Las hay absolutas y relativas.

RECORRIDO: $R = x_n - x_1$

RECORRIDO INTERCUARTÍLICO: $RI = Q_3 - Q_1$

DESVIACIÓN MEDIA RESPECTO A LA MEDIA ARITMÉTICA

$$D_M = \frac{1}{N}\sum_{i=1}^{n} |x_i - \bar{x}|\, n_i$$

DESVIACIÓN MEDIA RESPECTO A LA MEDIANA, O, SIMPLEMENTE DESVIACIÓN MEDIA:

$$D = \frac{1}{N}\sum_{i=1}^{n} |x_i - Me|\, n_i$$

VARIANZA (S^2): está considerada (junto con la desviación típica) el mejor indicador de la variabilidad global de la distribución, mide la dispersión de los datos respecto a la media aritmética, de hecho, nos suministra el valor medio del cuadrado de las desviaciones de los valores respecto de la media :

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

Tiene el inconveniente de medirnos la dispersión en términos del cuadrado de las unidades en las que esté medida la variable, por lo que muy a menudo se utiliza su raíz cuadrada: la DESVIACIÓN TÍPICA:

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

PRINCIPALES PROPIEDADES:

1ª) $S^2 \geq 0$

2ª) Dada una transformación lineal *y* de una variable *x*: $y_i = a + b x_i$ la varianza de la nueva variable *y* queda como $S_y^2 = b^2 S_x^2$ Lógicamente la desviación típica de la nueva variable será: $S_y = b S_x$

3ª) Su expresión de cálculo es ya que: $S^2 = \left(\frac{\sum_{i=1}^{n} x_i^2\, n_i}{N} \right) - \bar{x}^2$

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

$$= \frac{\sum_{i=1}^{n} x_i^2\, n_i}{N} - 2\,\bar{x}^2 + \bar{x}^2 = \frac{\sum_{i=1}^{n} x_i^2\, n_i}{N} - \bar{x}^2$$

MEDIDAS DE DISPERSIÓN RELATIVAS

Son indicadores de la dispersión de la distribución que se han relativizado, para que no afecten las unidades de medida de la variable y para que puedan hacerse comparaciones entre las dispersiones de conjuntos de datos dispares.

RECORRIDO SEMI-INTERCUARTÍLICO $\mathbf{R_s = Q_3\text{-}Q_1/(Q_3+Q_1)}$

COEFICIENTE DE VARIACIÓN DE PEARSON:

$$V = g_0 = \frac{S}{\bar{x}}$$

Mide, en términos relativos, la dispersión alrededor de la media aritmética, es el más utilizado, aunque presenta el inconveniente que es sensible a los cambios de "origen" en los valores de la variable.

ÍNDICE DE DISPERSIÓN RESPECTO A LA MEDIANA

$$V_{Me} = D/Me$$

MOMENTOS:

Son indicadores genéricos de una distribución. Se basan en una generalización de la idea de media, concretamente se tratará de la media aritmética de la r-sima potencia de los valores de la variable (o de sus desviaciones respecto a la media aritmética).Es interesante el hecho de que si dos distribuciones son iguales todos sus infinitos momentos coinciden.

MOMENTOS ORDINARIOS (RESPECTO AL ORIGEN)

$$a_r = \frac{1}{N}\sum_{i=1}^{n} x_i^r \cdot n_i$$

Se cumple que $a_0 = 1$ para cualquier distribución.

a_1 = la media aritmética

además son interesantes a_2 ,a_3 ,a_4 .

los momentos ordinarios se ven afectados por los cambios de origen y de escala (unidad) en los valores de la variable.

MOMENTOS CENTRALES

(RESPECTO A LA MEDIA ARITMÉTICA)

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

Se cumple que: $m_0 = 1$; $m_1 = 0$; $m_2 = S^2$; y son interesantes los de tercer y cuarto orden.

Los momentos centrales sólo se ven afectados por los cambios de escala.

RELACIÓN ENTRE MOMENTOS CENTRALES Y ORDINARIOS:

$$m_r = (a - \overline{x})^{(r)}$$

Donde la potenciación simbólica (r) debe entenderse que afecta como subíndice a la **a** y como potencia a la media:

$$m_2 = S^2 = (a - \overline{x})^{(2)} = a_2 - 2a_1\overline{x} + \overline{x}^2 = a_2 - \overline{x}^2$$

$$m_3 = (a - \overline{x})^{(3)} = a_3 - 3a_2\overline{x} + 3a_1\overline{x}^2 - \overline{x}^3 = a_3 - 3a_2\overline{x} + 2\overline{x}^3$$

$$m_4 = (a - \overline{x})^{(4)} = a_4 - 4a_3\overline{x} + 6a_2\overline{x}^2 - 4a_1\overline{x}^3 + \overline{x}^4 = a_4 - 4a_3\overline{x} + 6a_2\overline{x}^2 - 3\overline{x}^4$$

TRANSFORMACIONES LINEALES DE UNA VARIABLE ESTADÍSTICA

Dada una variable estadística x una nueva variable estadística y, será una transformación lineal de x cuando cada valor de y (y_i) dependa de cada observación de x (x_i) según una función lineal:

$y_i = a + b\, x_i$, manteniéndose la asignación de frecuencias.

El comportamiento de los momentos centrales de orden r es: $m_r(y) = b^r \cdot m_r(x)$

TIPIFICACIÓN (canónica): Es la operación de realizar una transformación lineal en la variable para que la (nueva) variable transformada, tenga por media, **cero,** y por desviación típica, **uno**.

Consiste en restar a la variable la media y dividirla por la desviación típica:

$$t_i = \frac{x_i - \overline{x}}{S}$$

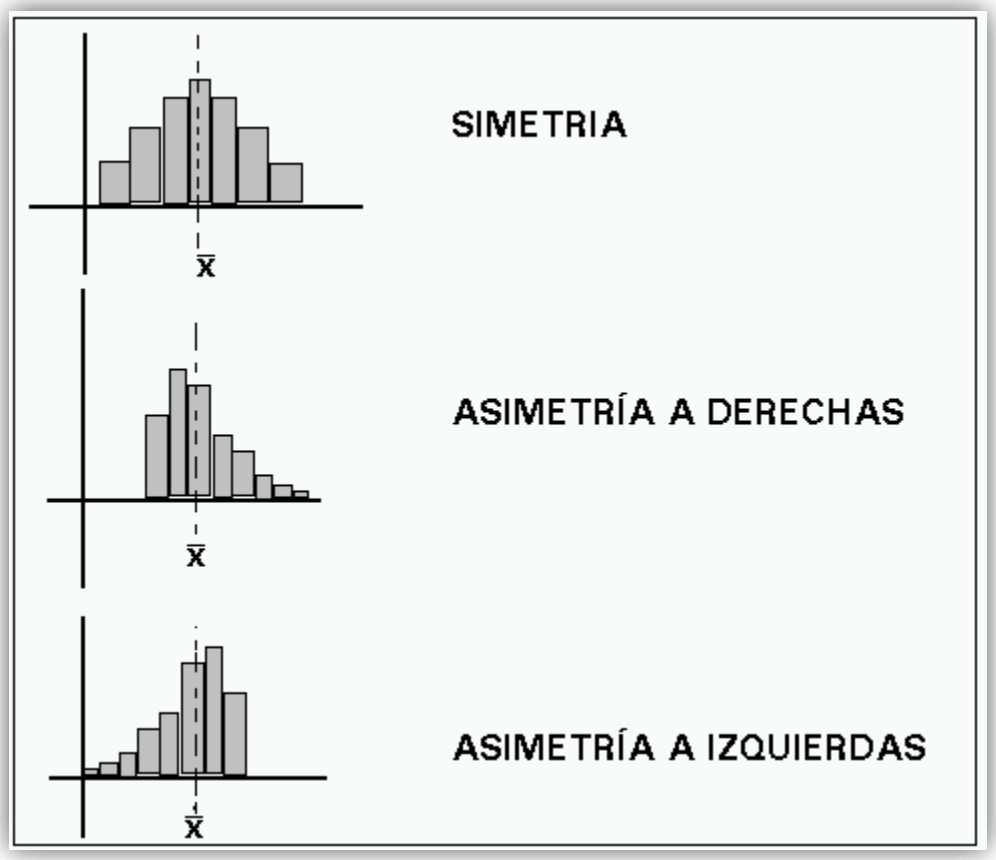
La tipificación permite la comparición de valores puntuales (puntuaciones) de dos distribuciones distintas de distintas medias y varianzas.

Es extraordinariamente útil para construir indicadores de comparación universal (univariantes: coeficiente de asimetría y coeficiente de curtosis;; y multivariantes : coeficientes de correlación).

MEDIDAS DE FORMA (ASIMETRÍA Y CURTOSIS)

A través de las medidas de posición y de dispersión, podemos hacernos una idea de por donde se sitúan los valores de la variable y cuánto se dispersan en términos globales .Pero si queremos conocer algo más de la forma en que se distribuye los valores necesitamos otros indicadores.

Los indicadores de **SIMETRÍA/ ASIMETRÍA** deberán informarnos de si los valores de la distribución se disponen simétricamente alrededor de la media, o bien si se decantan en mayor medida hacia la derecha (asimetría a derechas, o positiva) o hacia la izquierda (asimetría a izquierdas, o negativa), sin necesidad de representar gráficamente la distribución de frecuencias.



Como se trata de determinar si la disposición se decanta hacia un lado u otro de la media será necesario trabajar con un indicador que nos considere las diferencias de los de valores y de la media (con su signo).Por tanto habrá que considerar un momento central de orden impar. El de orden uno no es útil porque siempre se anula.

Habrá que considerar: m_3 :

Si $m_3 < 0$ la distribución será asimétrica negativa.

Si $m_3 > 0$ la distribución será asimétrica positiva.

Si $m_3 = 0$ la distribución será simétrica.

Pero si estamos interesados en encontrar un indicador la simetría/ asimetría, que no dependa de las unidades (del cubo de las unidades) y que nos permita hacer comparaciones de carácter universal, m3 no nos es útil. Por esta razón se define el coeficiente de asimetría como:

el momento central de tercer orden de la variable tipificada:

$$g_1 = m_3(t) = \frac{m_3(x)}{S_x^3}$$

MEDIDAS DE CURTOSIS.(COEFICIENTE DE CURTOSIS)

Dependiendo del número de observaciones que haya en la zona central de la distribución y del que haya en las zonas alejadas dos distribuciones con la misma varianza pueden tener dos perfiles distintos, con mayor o menor forma " de punta ".Al mayor o menor "apuntamiento" que puede tener una distribución con independencia del valor que tome su varianza se le llama CURTOSIS (o APUNTAMIENTO). [Ver gráfico]

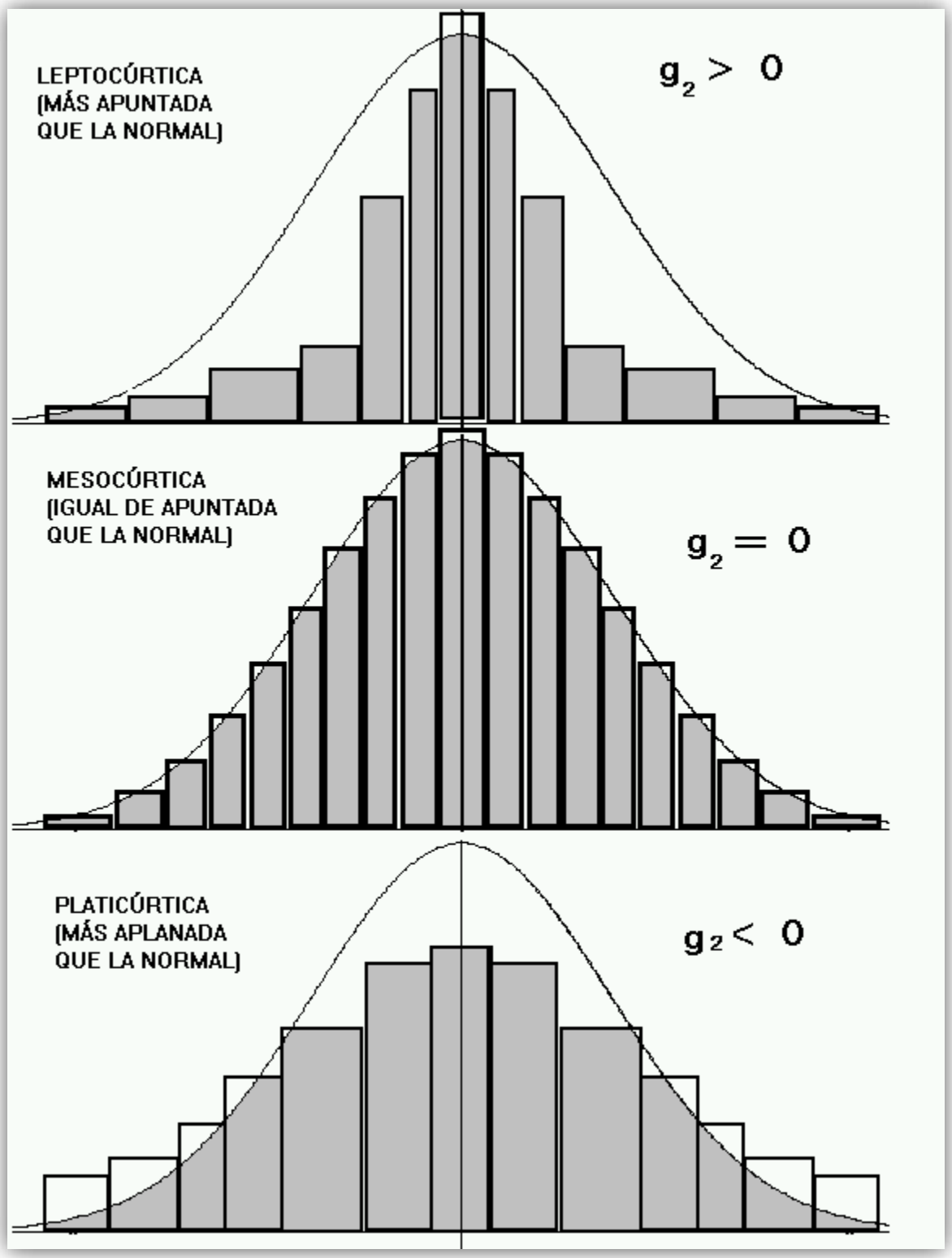
Como nos interesa comparar (ponderadamente) el número de observaciones cercanas a la media con el número de observaciones lejanas (con independencia del signo de su distancia a la media), para medir la curtosis, deberemos considerar un momento central de orden par; pero como la curtosis es el mayor o menor apuntamiento con independencia de la varianza, deberemos considerar el momento central de orden 4:

$$m_4 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^4 n_i$$

Pero si queremos disponer de una medida valida para la comparación universal, el hecho de que m_4 dependa de las unidades (de la cuarta potencia de las unidades) es un inconveniente, por lo que deberemos considerar como indicador de la curtosis el momento de cuarto orden la variable tipificada: $m_4(t)$

Por último suele considerarse el coeficiente de curtosis "relativizado" para permitir la comparación del apuntamiento de la distribución con el apuntamiento "patrón" que es el que tiene (el modelo Normal) la DISTRIBUCIÓN NORMAL DE PROBABILIDAD (campana de Gauss), cuyo momento de cuarto orden tipificado es tres. Por ello se define el coeficiente de curtosis como el momento central de cuarto orden de la variable tipificada menos tres unidades:

$$g_2 = m_4(t) - 3 = \frac{m_4(x)}{S_x^4} - 3$$


LEPTOCÚRTICA
(MÁS APUNTADA
QUE LA NORMAL)
$g_2 > 0$
MESOCÚRTICA
(IGUAL DE APUNTADA
QUE LA NORMAL)
$g_2 = 0$
PLATICÚRTICA
(MÁS APLANADA
QUE LA NORMAL)
$g_2 < 0$