ESTADÍSTICA DESCRIPTIVA

Grado de ADE. Primer curso

Raquel Mª Álvarez Esteban

Medidas de posición, dispersión, forma y concentración

Tema 3

Descripción numérica de una variable

- **Objetivo:**

  Resumir distintos aspectos de las distribuciones de frecuencias

Interés de los resúmenes numéricos:

- Unos **pocos números** resumen toda la distribución
- **Complemento** natural de la descripción gráfica
- Facilitan la **comparación** de muestras con modelos de referencia y la comparación entre muestras

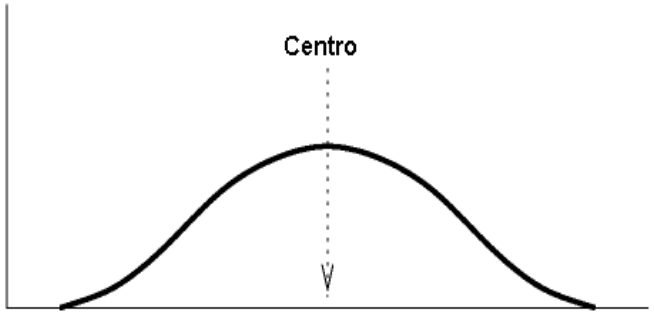
Descripción numérica de una variable

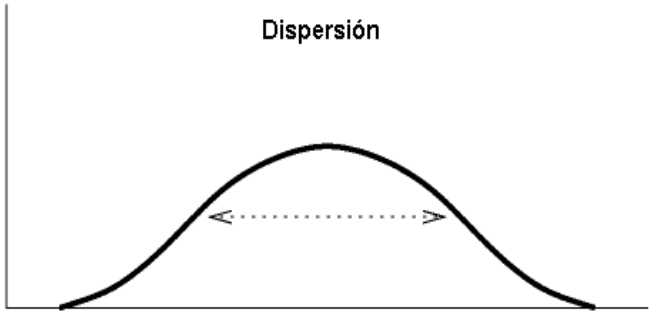
- **Medidas de posición:** Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
  - Cuartiles, deciles, cuantiles, percentiles

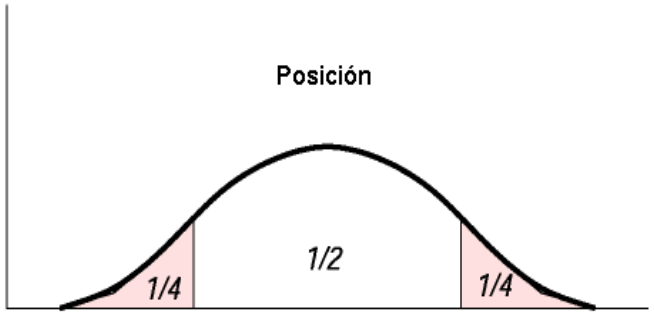
- **Medidas de centralización**: Indican valores con respecto a los que los datos parecen agruparse.
  - Media, mediana, moda, media geométrica, media armónica, …

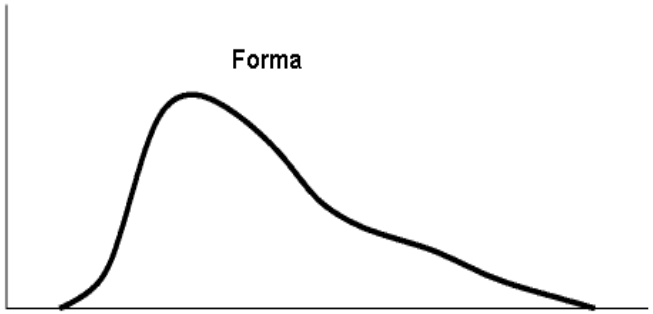
- **Medidas de dispersión**: Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
  - Desviación típica, coeficiente de variación, rango, varianza, recorrido intercuartílico..,

- **Medidas de forma**:
  - Asimetría
  - Apuntamiento o Curtósis


Centro


Dispersión


Posición
1/4
1/2
1/4


Forma

Parámetros y estadísticos

- **Parámetro**: es una cantidad numérica calculada sobre una población
  - La altura media de los individuos de un país
  - La idea es resumir toda la información que hay en la población en unos pocos números (parámetros).

- **Estadístico:** es una cantidad numérica calculada sobre una muestra (Cualquier función con los datos de la muestra destinada a cuantificar algún aspecto de la distribución de frecuencias.)
  - La altura media de los individuos de esta clase
    Somos una muestra (¿representativa?) de la población.
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar estimador.

- Normalmente interesa conocer un parámetro, pero por la dificultad que conlleva estudiar a **TODA** la población, calculamos un estimador sobre una muestra y "confiamos" en que sean próximos. Hay que elegir muestras representativas, para que el error sea "confiablemente" pequeño.

Estadísticos de posición y centralización:

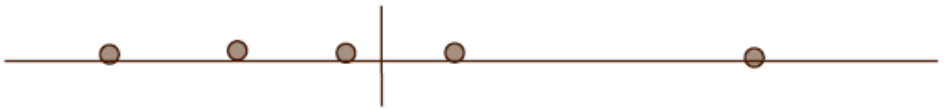
Media

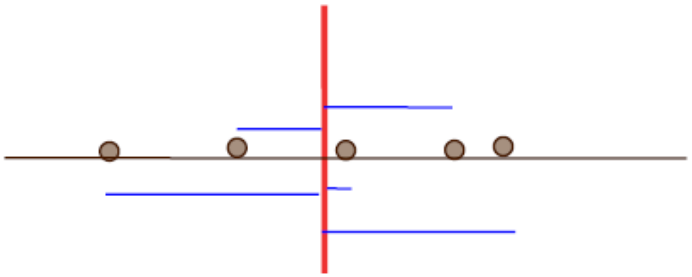
Media (media aritmética ó media muestral) Muestra: $x_1, x_2, \ldots, x_n$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Es el centro de gravedad de la distribución de frecuencias

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





La media es el valor A que hace mínima la suma de cuadrados de las desviaciones respecto a A

$$\min_{A} \sum_{i=1}^{n} (x_i - A)^2$$

Media

Muestra tabulada:

-Variable discreta: $\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \frac{n_i}{n} x_i = \sum_{i=1}^{k} f_i x_i$

-Variable continua: (datos agrupados en k clases) $\overline{X} \cong \frac{1}{n}\sum_{i=1}^{k} n_i m_i = \sum_{i=1}^{k} f_i m_i$

Ejemplo:

Muestra de tamaño 16, la media:

x_i	n_i
0	3
2	4
0,4	2
1	7

media=[(0*3) + (2*4) + (0,4*2) + (1*7)] /16 = 0,987

- Muestra de tamaño 100 se toma el peso en kg de las personas.

$L_{i-1} - L_i$	n_i
50-60	30
60-70	25
70-80	20
80-90	10
90-100	10
100-105	5

- Media?

Media

Muestra tabulada:

-Variable discreta: $\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \frac{n_i}{n} x_i = \sum_{i=1}^{k} f_i x_i$

-Variable continua: (datos agrupados en k clases) $\overline{X} \cong \frac{1}{n}\sum_{i=1}^{k} n_i m_i = \sum_{i=1}^{k} f_i m_i$

Ejemplo:

Muestra de tamaño 100 se toma el peso en kg de las personas.

$L_{i-1} - L_i$	n_i	m_i
50-60	30	55
60-70	25	65
70-80	20	75
80-90	10	85
90-100	10	95
100-105	5	105

$m_i = (L_{i-1} - L_i)/2$ **marcas de clase**

media = 71 kg

Media

Falta de robustez de la media

Ejemplo 1:

Media = [0(4) + 1(4) + 2(1)] / 9 = **0.6667**

X_i (valores)	n_i
0	4
1	4
2	1
	Total = 9

Media = [0(3) + 1(4) + 2(1) +6(1)] / 9 = **1.333**

X_i (valores)	n_i
0	3
1	4
2	1
6	1
	Total = 9

Ejemplo 2:

Datos: 1, 2, 3, 4, 7, 8, 9 n=7 media = 4.858

1, 2, 3, 4, 7, 8, 2450 n=7 media = 353.6

Media

Media ponderada

Media ponderada de k valores $(x_1, x_2, \ldots, x_k)$ con pesos $(w_1, w_2, \ldots, w_k)$:

$$\frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_k x_k}{w_1 + w_2 + \ldots + w_k} = \sum_{i=1}^{k} \left(\frac{w_i}{\sum_{j=1}^{k} w_j} \right) x_i$$

Si tomamos los pesos $(p_1, p_2, \ldots, p_k)$ de forma que sumen uno la media ponderada se calcula como $\sum_{i=1}^{k} p_i x_i$

Comprobamos como definiendo $p_i = \frac{w_i}{\sum_{j=1}^{k} w_j}$ ambas expresiones coinciden

Ejemplo: nota media de un alumno con calificaciones en tres asignaturas

A: 5, B:7, C: 9. Créditos de cada asignatura: A: 15, B: 7.5, C: 6

$$\frac{15(5) + 7.5(7) + \ldots + 6(9)}{15 + 7.5 + 6} = 6.368$$

Media

PROPIEDADES DE LA MEDIA:

- La suma de las desviaciones de la media a las observaciones es cero $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$
- Cambios de origen y escala en los datos $y_i = a + b\, x_i$, $i = 1, 2, \ldots, n$ conllevan los mismos cambios en la media $\bar{Y} = a + b\bar{X}$
- La media de una suma es la suma de las medias

$$x_1, x_2, \ldots, x_n \rightarrow \bar{X} \qquad y_1, y_2, \ldots, y_n \rightarrow \bar{Y}$$

$$x_1 + y_1,\ x_2 + y_2,\ \ldots,\ x_n + y_n \rightarrow \bar{X} + \bar{Y}$$

- La media es el valor A que hace mínima la suma de cuadrados de las desviaciones respecto a A $\sum_{i=1}^{n} (x_i - A)^2 \qquad \min_{x \in \Re} \sum_{i=1}^{n} (x_i - A)^2$
- Si la muestra esta dividida en dos grupos, la media de la muestra es la media ponderada (por los tamaños de los grupos) de las medias.

Media

PROPIEDADES DE LA MEDIA:

❑ Cambios de origen y escala en los datos $y_i = a + b\,x_i$, $i = 1, 2, \ldots, n$ conllevan los mismos cambios en la media $\overline{Y} = a + b\overline{X}$

- cambio de escala: Y=bX

  ej. Cálculo de precios medios aplicando el IVA (21%)

precios					precios +IVA		
	xi	ni	xi*ni		1,21xi	ni	1,21xi*ni
	0,9	3	2,70		1,089	3	3,27
	1,3	6	7,80		1,573	6	9,44
	2,5	8	20,00		3,025	8	24,20
	3,9	5	19,50		4,719	5	23,60
	5,9	4	23,60		7,139	4	28,56
	7	2	14,00		8,47	2	16,94
total		28	87,60			28	106,00
		media	3,13		media		3,79
					media=	1,21*3,13	

- cambio de origen: Y=a+X

  ej. Sueldo medio de los trabajadores de una empresa después de un aumento de 50€ a todos

Media

PROPIEDADES DE LA MEDIA:

- Si la muestra esta dividida en dos grupos, la media de la muestra es la media ponderada (por los tamaños de los grupos) de las medias.

	x1	n1										
	x2	n2	h elementos		X1 es la media de esta poblacion							
	…	…										
	xh	nh										
				en total N elementos				media de los N elementos es=[(h*X1)+(N-h)*X2]/N				
	x(h+1)	n(h+1)										
	x(h+2)	n(h+2)	(N-h) elementos									
	…	…			X2 es la media de esta poblacion							
	xN	nN										

Ejemplo: salario medio en España a partir del salario medio de cada una de las provincias.

Media geométrica

$$M_G = \sqrt[n]{x_1 x_2 \ldots x_n}$$

$$\log M_G = \frac{1}{n}\sum_{i=1}^{n} \log(X_i)$$

- Muestra tabulada (discreta) $$M_G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}}$$

- Muestra tabulada (continua) $$M_G = \sqrt[n]{m_1^{n_1} m_2^{n_2} \ldots m_k^{n_k}}$$

donde m_i son las marcas de clase

Media geométrica:

$$M_G = \sqrt[n]{x_1 x_2 \ldots x_n}$$

- Desventajas:

Difícil de calcular (cifras muy elevadas: solución tomar logaritmos)

Si algún Xi =0, M_G se anula

No es relevante y puede ser un número complejo (número no real) si algún valor es negativo

- Usado para promediar %, tasas, números índices…(situaciones donde la vble presenta variaciones acumulativas).

  Ejemplo: C capital inicial colocado a tantos unitarios de interés i_i anual durante n años. Para calcular el tanto de interés medio del periodo, no es la media aritmética. Si Ci es el capital al final del año i, entonces

  $C_1=C(1+ i_1)$

  $C_2=C_1 (1+ i_2) =C(1+ i_1) (1+ i_2)$

  …

  $C_n=C_{n-1} (1+ i_n)= \ldots=C(1+ i_1) (1+ i_2)\ldots (1+ i_n)$

  El tanto de interés medio i será $C_n= C(1+ i)\ldots(1+i)$

  Igualando ambas expresiones y despejando i, se cumplirá $i= \sqrt[n]{(1+ i_1) (1+ i_2)\ldots (1+ i_n)} -1$

Media armónica

$$M_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}}$$

- Muestra tabulada (discreta)

$$M_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{k}\frac{1}{X_i^{n_i}}}$$

- Muestra tabulada (continua)

$$M_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{k}\frac{1}{m_i^{n_i}}}$$

donde m_i son las marcas de clase

Media armónica

Desventaja:

- influencia de los valores pequeños
- no se puede calcular si algún valor es 0

Usado para promediar aquello cuyas unidades sean el cociente de dos magnitudes simples: velocidades, rendimientos, etc.

Ejemplo: índice de Paasche es una media armónica

Relación entre las diferentes medias

$$M_H \leq M_G \leq \overline{X}$$

Mediana

- Estadísticos de orden

Muestra: $x_1, x_2, \ldots, x_n$

Muestra ordenada: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ $\quad x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

$x_{(1)}$: Mínimo $\quad x_{(n)}$: Máximo $\quad x_{(r)}$: Estadístico de orden r, $r = 1, \ldots, n$.

Me

Punto que parte la distribución en dos mitades del 50% a cada lado

Observación central en la muestra ordenada

Si n es impar $\quad Me = X_{((n+1)/2)}$

Si n es par $\quad Me \in (X_{(n/2)}, X_{(n/2+1)}) \quad Me = (X_{(n/2)} + X_{(n/2+1)})/2$

Mediana

Me

Punto que parte la distribución en dos mitades del 50% a cada lado

Observación central en la muestra ordenada

Si n es impar $Me = X_{((n+1)/2)}$

Si n es par $Me \in (X_{(n/2)}, X_{(n/2+1)})$ $Me = (X_{(n/2)} + X_{(n/2+1)})/2$

Ejemplos:

Datos 1: → 1, 2, 3, 4, 6, 7, 8 n=7 Me = 4 media = 4.4

Datos 2: → 1, 2, 3, 4, 5, 6, 7, 8 n=8 Me = 4.5 media = 4.5

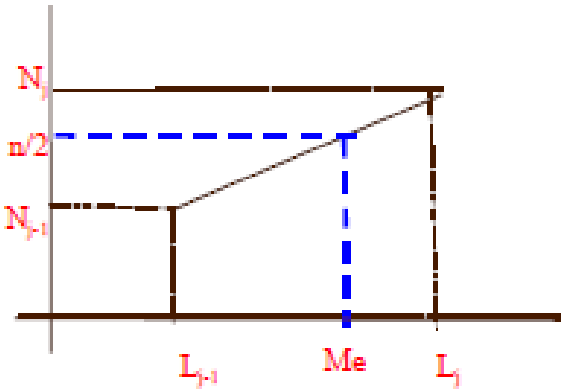
Datos 3: → 1, 3, 4, 2, 7, 2450, 8 n=7 Me = 4 media = 353.6

1, 2, 3, 4, 7, 8, 2450

Mediana

Cálculo de la mediana para datos agrupados:

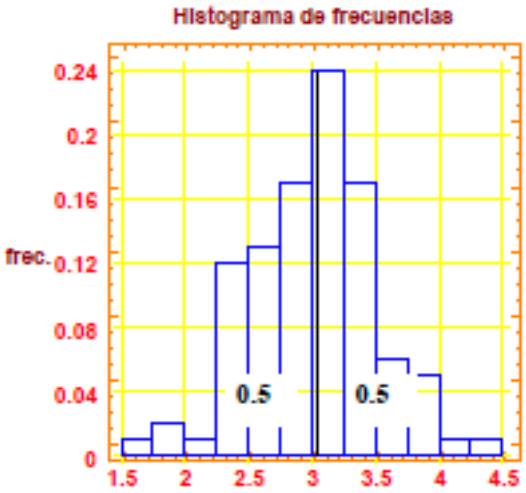
$$Me \cong L_{j-1} + \left(\frac{n}{2} - N_{j-1}\right)\frac{L_j - L_{j-1}}{N_j - N_{j-1}}$$

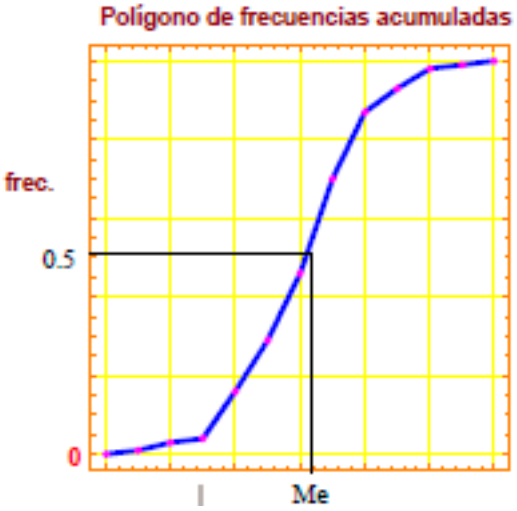


Misma formula:

$$Me = L_{j-1} + (L_j - L_{j-1})\frac{n/2 - N_{j-1}}{n_j}$$

Mediana

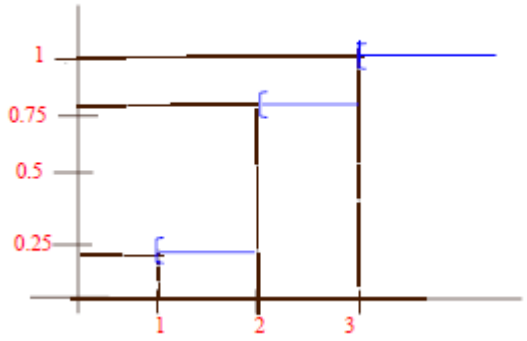

Histograma de frecuencias
frec.
0.24
0.2
0.16
0.12
0.08
0.04
0
0.5
0.5
1.5
2
2.5
3
3.5
4
4.5


Polígono de frecuencias acumuladas
frec.
0.5
0
Me

Mediana

Ejemplo de cálculo para vble discreta, a partir de la tabla de frecuencias:

X_i	n_i	N_i	f_i	F_i
1	20	20	0.2	0.2
2	60	80	0.6	0.8
3	20	100	0.2	1



Me = 2

Mediana

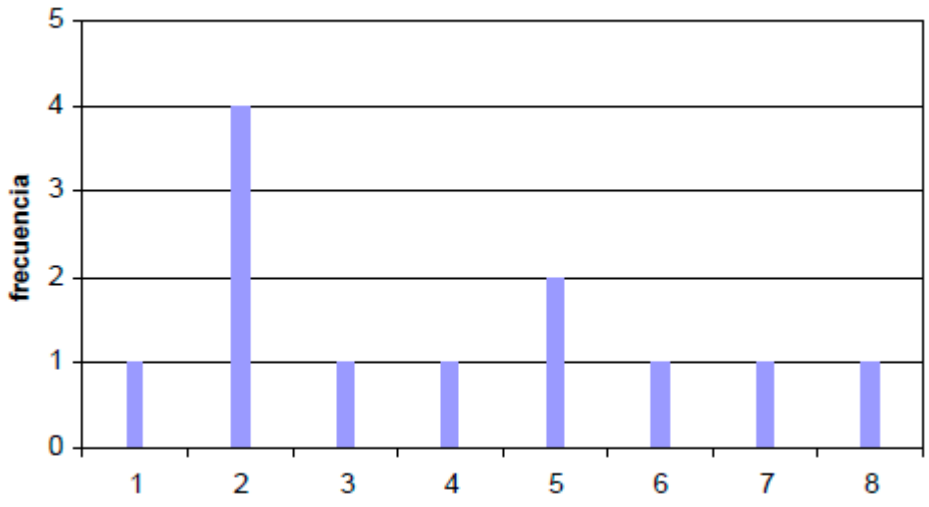
Datos agrupados en clases: buscar el intervalo en el que se alcanza la mediana $[L_{i-1}, L_i]$

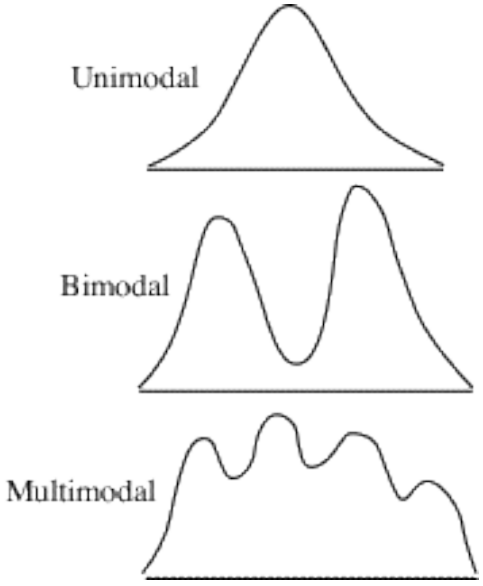
$$Me = L_{i-1} + \frac{N\frac{1}{2} - N_{i-1}}{n_i} c_i$$

Moda

Mo= Punto donde se alcanza el máximo de la distribución de frecuencias

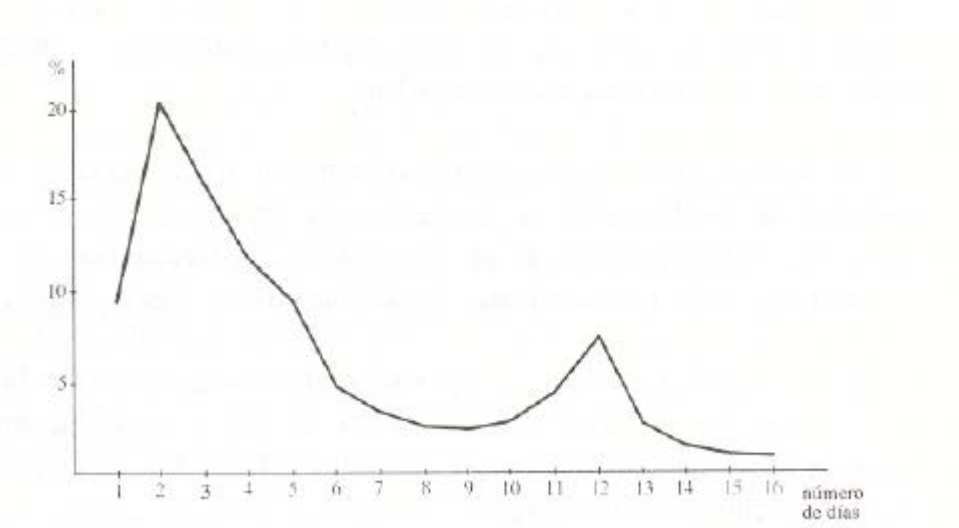
Hay distribuciones con varias modas (bimodales o multimodales)





Moda

Mo= Punto donde se alcanza el máximo de la distribución de frecuencias



Distribución del número de días de baja anuales en una plantilla de 9.870 trabajadores de entre los que tuvieron al menos una baja

Moda absoluta, moda relativa, intervalo modal

Moda

Mo= Punto donde se alcanza el máximo de la distribución de frecuencias

En distribuciones agrupadas, el intervalo modal es el de mayor densidad

Si queremos seleccionar un pto, diferentes criterios:

[Mo= L_{i-1}] ;

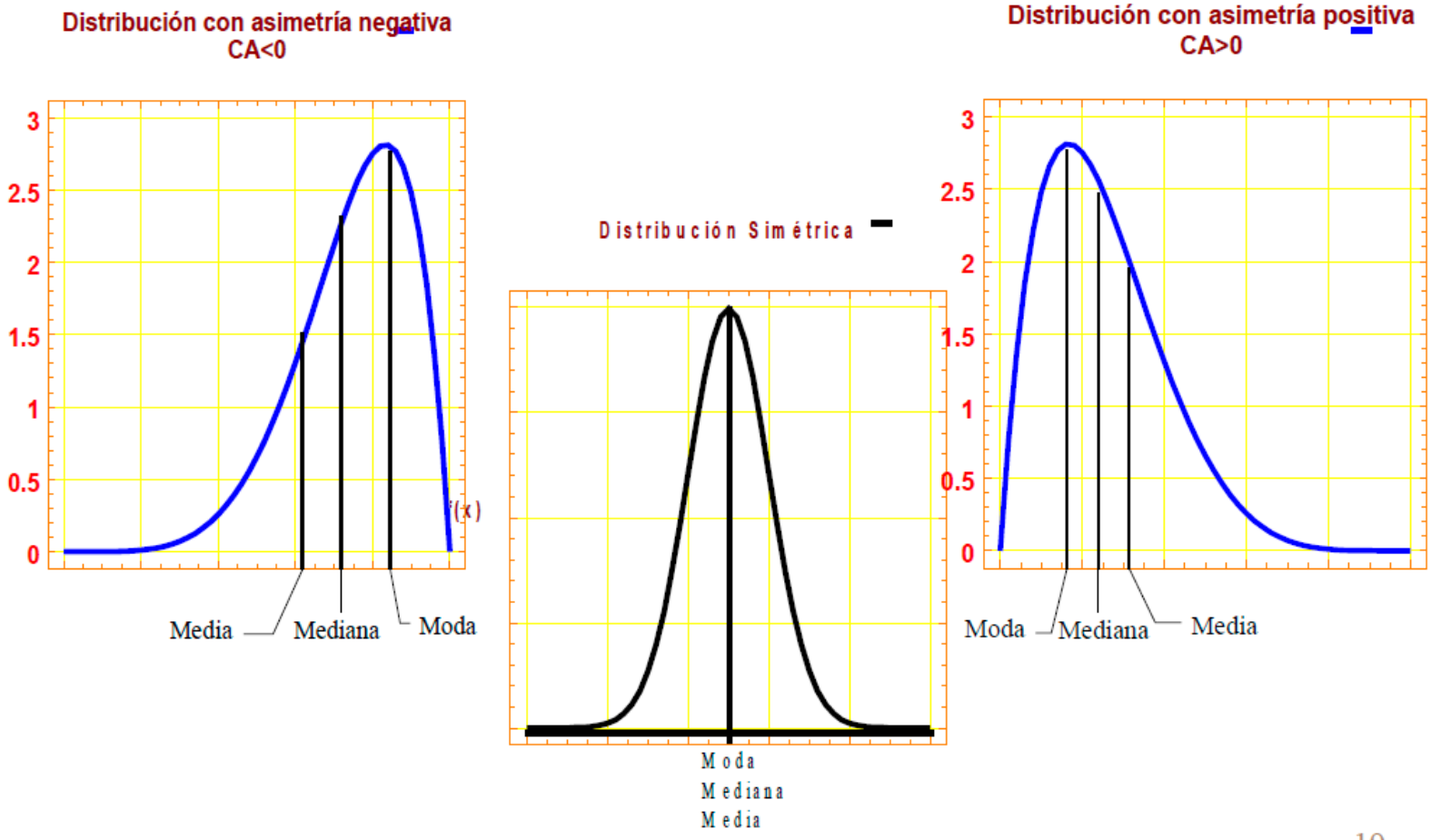
[Mo= Li];

El criterio más utilizado:

$$\mathbf{Mo = L_{i\_1} + \frac{d_{i+1}}{d_{i+1} + d_{i\_1}} c_i}$$

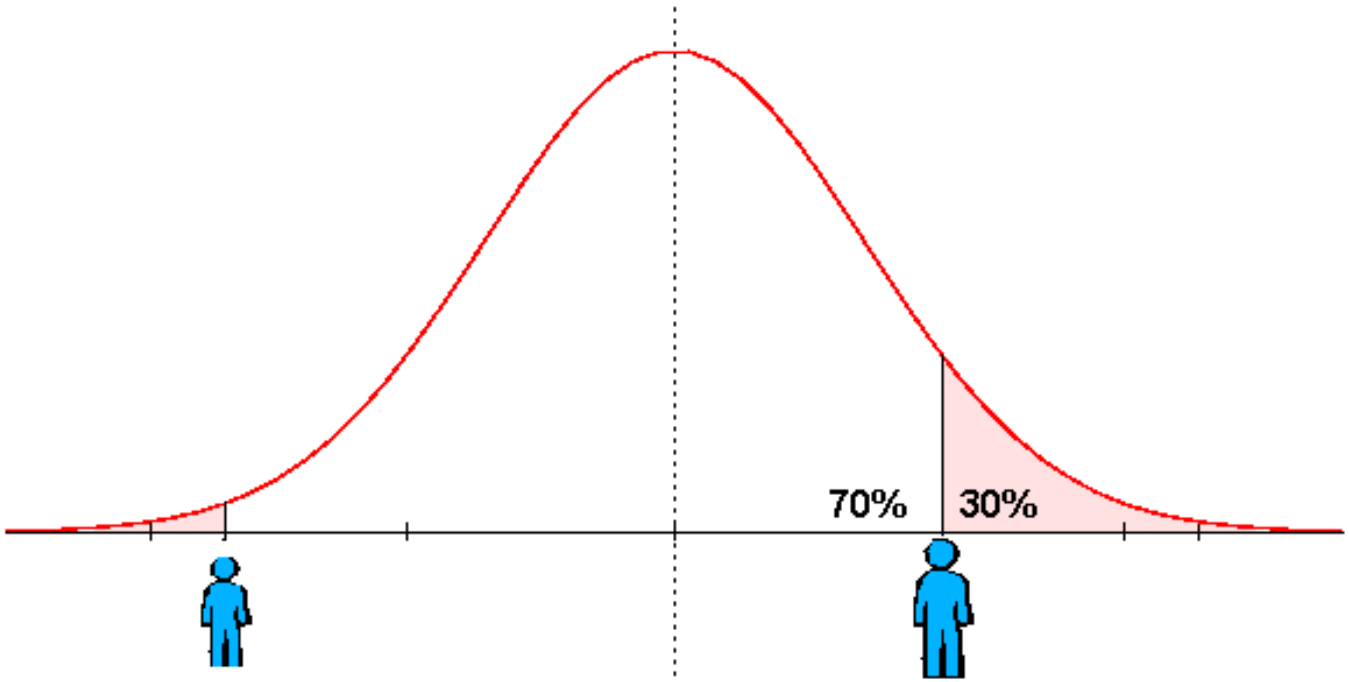
$$\boxed{d_i = \frac{n_i}{c_i}}$$

Posición relativa de media, mediana y moda



Cuantiles

- **Cuantil** de orden α es el valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.
- Casos particulares son los percentiles, cuartiles, deciles, quintiles,...

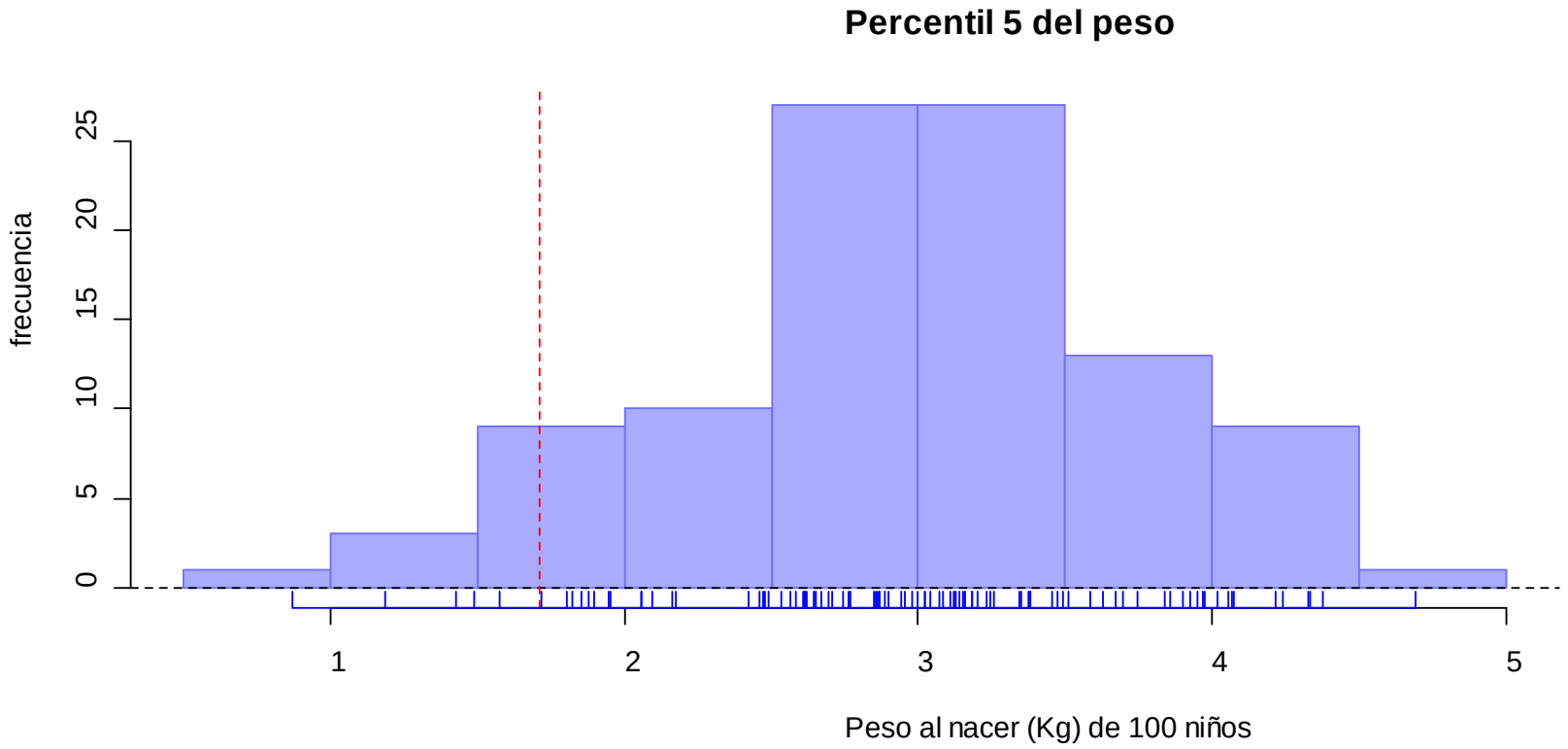


Cuantiles:

- **Percentil** de orden p (P_p)= cuantil de orden p/100 (me deja a la izda el p% de la distribución de frecuencias y a la drcha el (1-p)%
  - La mediana es el percentil 50
  - El percentil de orden 15 deja por debajo al 15% de las observaciones. Por encima queda el 85%
- **Cuartiles**: Dividen a la muestra en 4 grupos con frecuencias similares.
  - Primer cuartil Q1= Percentil 25 (p=0,25)= Cuantil 0,25
  - Segundo cuartil Q2= Percentil 50 (p=0, 5)= = Cuantil 0,5 = mediana
  - Tercer cuartil Q3= Percentil 75 (p=0,75)= = cuantil 0,75

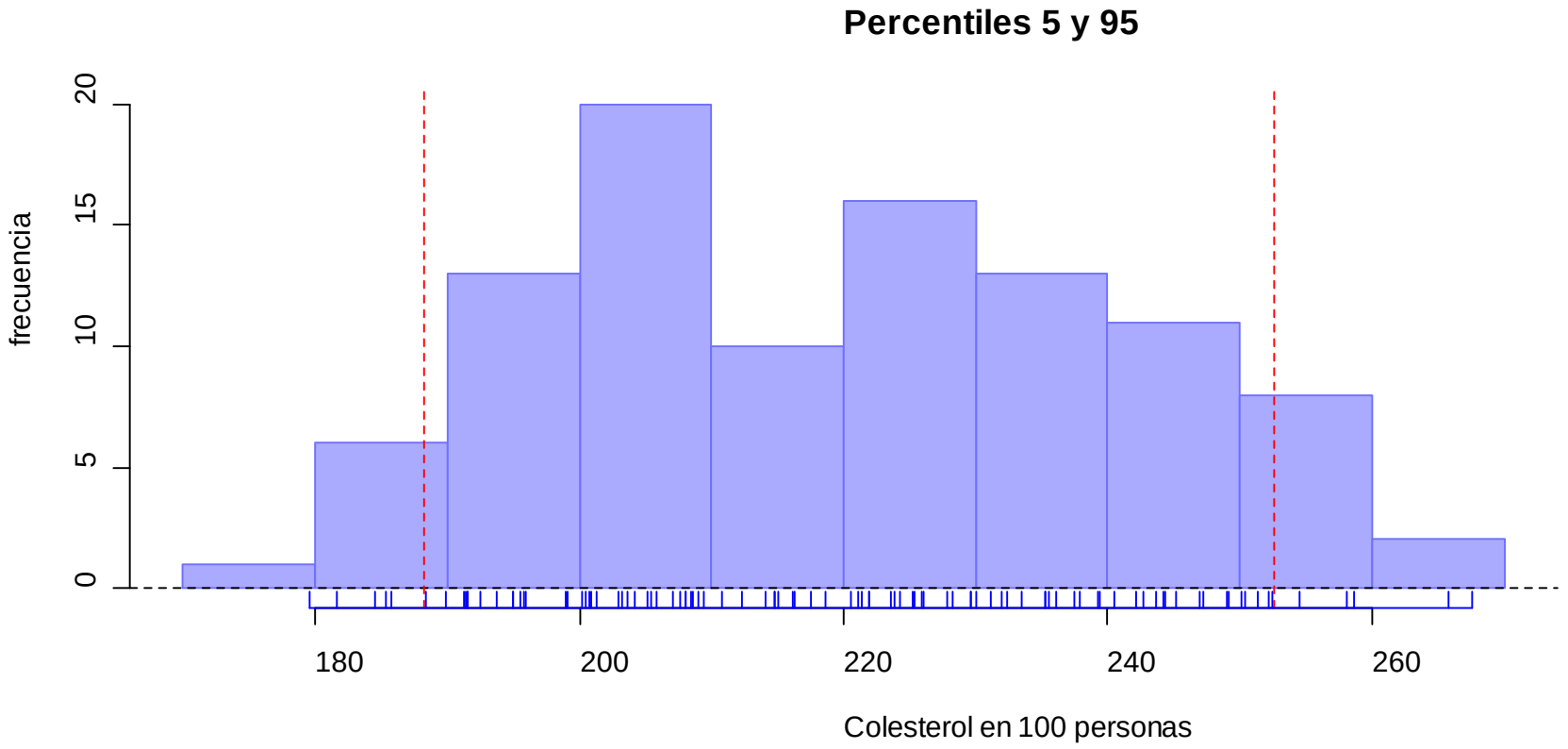
Cuantiles: ejemplos

- El 5% de los recién nacidos tiene un peso demasiado bajo. ¿Qué peso se considera “demasiado bajo”?
    - Percentil 5 o cuantil 0,05



Cuantiles: ejemplos

El colesterol se distribuye simétricamente en la población. Supongamos que se consideran patológicos los valores extremos. El 90% de los individuos son normales ¿Entre qué valores se encuentran los individuos normales?



Cuantiles: ejemplo

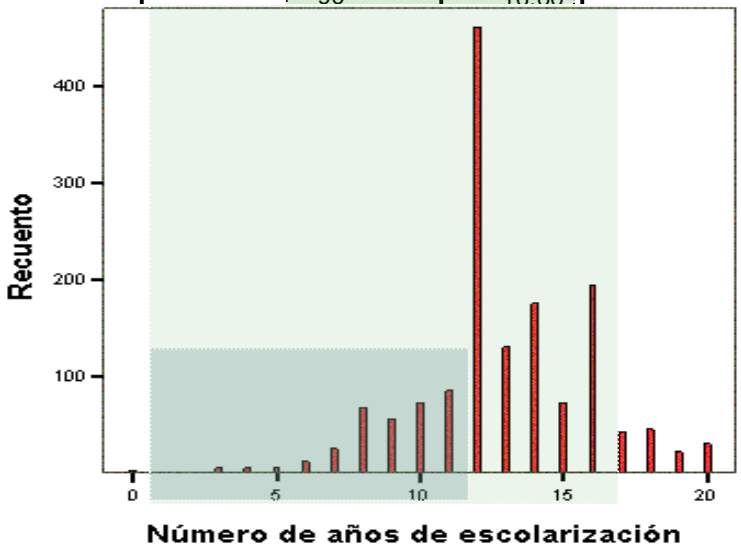
Número de años de escolarización

	Frecuencia	Porcentaje	Porcentaje acumulado	
3	5	,3	,3	
4	5	,3	,7	
5	6	,4	1,1	
6	12	,8	1,9	
7	25	1,7	3,5	
8	68	4,5	8,0	
9	56	3,7	11,7	
10	73	4,8	16,6	
11	85	5,6	22,2	≥20%?
12	461	30,6	52,8	
13	130	8,6	61,4	
14	175	11,6	73,0	
15	73	4,8	77,9	
16	194	12,9	90,7	≥ 90%?
17	43	2,9	93,6	
18	45	3,0	96,6	
19	22	1,5	98,0	
20	30	2,0	100,0	
Total	1508	100,0		

Estadísticos

Número de años de escolarización

N	Válidos	1508
	Perdidos	0
Media		12,90
Mediana		12,00
Moda		12
Percentiles	10	9,00
	20	11,00
	25	12,00
	30	12,00
	40	12,00
	50	12,00
	60	13,00
	70	14,00
	75	15,00
	80	16,00
	90	16,00



Cuantiles

Para calcularlo el cuantil de orden r/k se hallan las frecuencias acumuladas y se busca el valor que ocupe el lugar **(r/k)*N** de la distribución.

Si K=4, entonces tendremos los cuartiles:

r=1 primer cuartil

r=2 segundo cuartil (mediana)

…

Si K=10, entonces tendremos los deciles:

r=1, primer decil

…

Si K=100, entonces tendremos los percentiles…

Cuantiles: ejemplos

Datos→ 70, 24, 68, 20, 50, 30,53, 40, 36, 57, 66 n=11

20	24	30	36	40	50	53	57	66	68	70
$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$X_{(6)}$	$X_{(7)}$	$X_{(8)}$	$X_{(9)}$	$X_{(10)}$	$X_{(11)}$

$n\frac{1}{4} = n(0.25) = 11(0.25) = 2.75$ $\quad Q_1 = X_{(3)} = 30$

$n\frac{1}{2} = n(0.5) = 11(0.5) = 5.5$ $\quad Q_2 = Me = X_{(6)} = 50$

$n\frac{3}{4} = n(0.75) = 11(0.75) = 8.25$ $\quad Q_3 = X_{(9)} = 66$

¿¿Percentil 10 ?? $\quad n(0.10) = 11(0.10) = 1.1$

Percentil 10 $= X_{(2)} = 24$

Percentiles: cálculo para vbles contínuas con datos agrupados

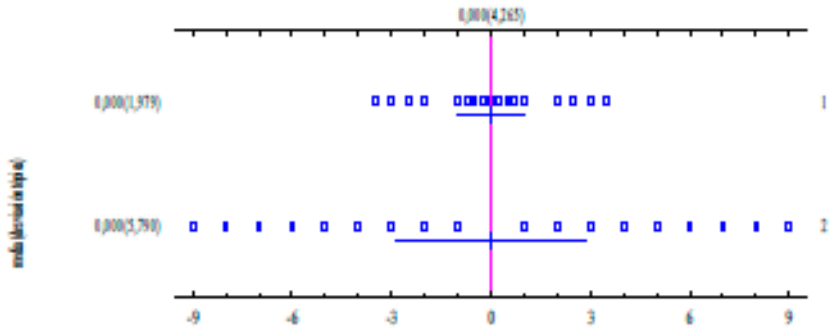
Datos agrupados en clases: buscar el intervalo en el que se alcanza dicha frecuencia : $[L_{i\_1}, L_i]$

$$C_{r,k} = L_{i\_1} + \frac{N \cdot \frac{r}{k} - N_{i\_1}}{n_i} c_i$$

Si K=4, entonces tendremos los cuartiles:
Si K=10, entonces tendremos los deciles:
Si K=100, entonces tendremos los percentiles…

Medidas de dispersión

Ejemplo de dos muestras con la misma media y distinta dispersión, n = 18 en ambas



Medidas de dispersión

Miden el grado de dispersión (variabilidad) de los datos, independientemente de su causa.

Medidas de dispersión absolutas:

- Rango
- Recorrido intercuartílico
- Varianza
- Desviación típica o desviación estandar
- Otras (desviación respecto de la media, desviación respecto de la mediana)

Medidas de dispersión relativas:

- Coeficiente de variación
- Recorrido semi-intercuartilico

Medidas de dispersión

- **Rango (o Amplitud)** : Máximo – Mínimo= $X_{(n)} - X_{(1)}$

Diferencia entre observaciones extremas.

Ej: 2,**1**,4,3,**8**,4. El rango es 8-1=7

Es muy sensible a los valores extremos.

- **Rango/recorrido intercuartílico:**

Es la distancia entre primer y tercer cuartil.

Rango intercuartílico = $P_{75} - P_{25} = Q_3 - Q_1$

Parecida al rango, pero eliminando las observaciones más extremas inferiores y superiores. (el 50% de los valores centrales están ahí)

No es tan sensible a valores extremos.

Medidas de dispersión

Varianza S^2 Miden el promedio de las desviaciones (al cuadrado) de las observaciones con respecto a media.

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

- Es **sensible** a valores extremos (alejados de la media).

- Sus unidades son el cuadrado de las de la variable.
- Muestra tabulada
  - Vble discreta

$$S^2 = \frac{1}{n}\sum_i n_i (x_i - \bar{x})^2 = \sum_i f_i (x_i - \bar{x})^2$$

  - Vble contínua

$$S^2 = \frac{1}{n}\sum_i n_i (m_i - \bar{x})^2 = \sum_i f_i (m_i - \bar{x})^2$$

donde mi son las marcas de clase

Medidas de dispersión

Varianza S^2

Otra expresión para la varianza

$$S^2 = \left(\frac{1}{n}\sum_i x_i^2\right) - \bar{x}^2$$

Medidas de dispersión

Desviación típica S Es la raíz cuadrada de la varianza

$$S = \sqrt{S^2}$$

Mismas unidades que la variable.

Varianza: propiedades

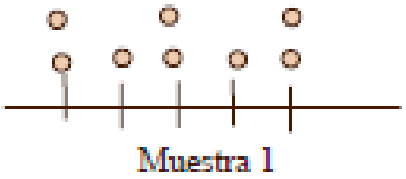
$$S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

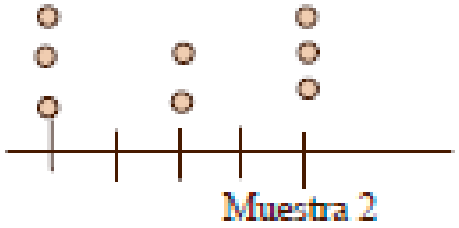
- Siempre es positiva.
- Varianza=0 si y solo si todos los valores son iguales (es decir, la variable es constante)
-

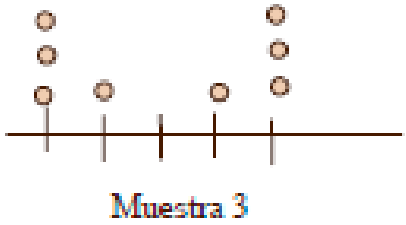
$$S^2_{a+bx} = b^2 S^2_x$$

$$S_{a+bx} = |b| S_x$$

ejemplos

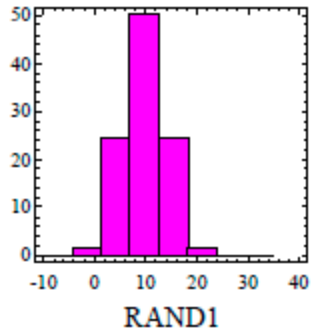


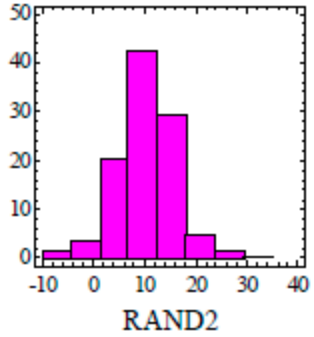


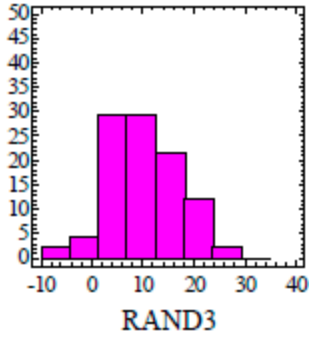


	Muestra 1	Muestra 2	Muestra 3
	6	6	6
	6	6	6
	7	6	6
	8	8	7
	8	8	9
	9	10	10
	10	10	10
	10	10	10
N=	8	8	8
Mediana=	8	8	8
Rango=	4	4	4
S=	1,50	1,73	1,80
S*2=Var=	2,25	3,00	3,25

ejemplos







Rand1

Varianza = 14,6842

S = 3,832

Rand2

Varianza = 24,5909

S = 4,95892

Rand3

Varianza = 52,9725

S = 7,27822

ejemplos

Un empresario paga a sus vendedores asalariados un salario fijo más una comisión. Los salarios fijos son:

Meses como empleado	Salarios (en miles de $)	Sexo
6	7.5	Hombres
10	8.6	Hombres
12	9.1	Hombres
18	10.3	Hombres
30	13.0	Hombres
5	6.2	Mujeres
13	8.7	Mujeres
15	9.4	Mujeres
21	9.8	Mujeres

	Hombres	Mujeres
n	5	4
media	9.7	8.525
varianza	4.415	2.60917
Desviación típica	2.10119	1.61529

Media = 9,17778

Varianza = 3,56944

S = 1,8893

Medidas de dispersión

Desviación media respecto a la media:

$$D_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Desviación media respecto a la mediana:

$$D_{Me} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Me\right|$$

Medidas de dispersión:

Coeficiente de variación CV: cociente entre desviación típica y la media

$$CV = \frac{S}{\overline{x}}$$

- Mide la desviación típica en forma de "qué tamaño tiene con respecto a la media" (número de veces que la desviación típica contiene a la media aritmética.)
- También se la denomina variabilidad relativa.
- Es frecuente mostrarla en porcentajes
  - Si la media es 80 y la desviación típica 20 entonces CV=20/80=0,25=25% (variabilidad relativa)

- Es una cantidad **adimensional**. Interesante para comparar la variabilidad de diferentes variables.
  - Si el peso tiene CV=30% y la altura tiene CV=10%, los individuos presentan más dispersión en peso que en altura.

- Solo para vbles positivas
- Facilita la comparación

(homogeneidad)

Medidas de dispersión:

Recorrido semi-intercuartilico R_s:

cociente entre el recorrido intercuartílico y la suma del primer y tercer cuartil

$$R = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Momentos

- Momentos muestrales
  - Momentos respecto al origen

Momento de orden k=1,2…

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i{}^k n_i$$

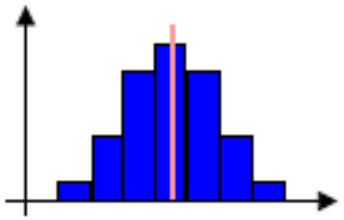
  - Momentos respecto a la media

Momento de orden k=1,2..

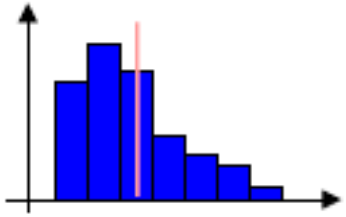
$$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})^k n_i$$

La distribución de frecuencias queda definida por todos sus momentos

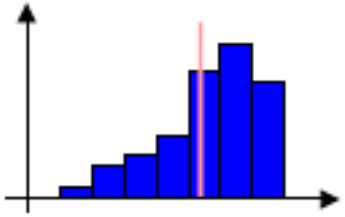
Medidas de simetría



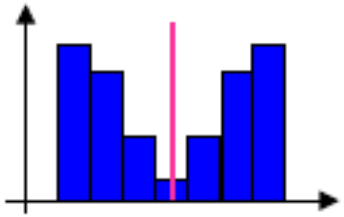
Distribución simétrica.

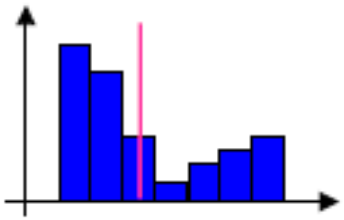


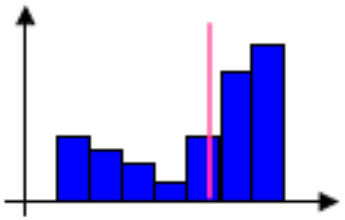
Distribución asimétrica positiva o a la derecha.



Distribución asimétrica negativa o a la izquierda.







Medidas de simetría

- Coeficiente de asimetría de Fisher (g_1 ó CA)

$$CA = g_1 = \frac{m_3}{S_x^3}$$

$$CA = g_1 = \frac{m_3}{S_x^3} = \frac{\frac{1}{n}\sum_{i=1}(x_i - \overline{X})^3}{S_x^3} = \frac{1}{n}\sum_{i=1}(\frac{x_i - \overline{X}}{S_x})^3$$

Si la distribución está tabulada

$$CA = \frac{\frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{X})^3 n_i}{S_x^3} = \frac{1}{n}\sum_{i=1}^{k}(\frac{x_i - \overline{X}}{S_x})^3 n_i$$

Distribución simétrica: CA≅0

Distribución asimétrica positiva: CA>0 (cola derecha más pesada)

Distribución asimétrica negativa: CA<0 (cola izquierda más pesada)

Medidas de simetría

- Coeficiente de asimetría de Pearson A_p

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

Distribución simétrica $A_p=0$
Distribución asimétrica positiva $A_p>0$
Distribución asimétrica negativa $A_p<0$

Medidas de simetría

- Coeficiente de asimetría de Bowley:

$$A_B = CAB = \frac{(Q_3 + Q_1 - 2Me)}{Q_3 - Q_1}$$

Distribución simétrica $A_B=0$
Distribución asimétrica positiva $A_B>0$
Distribución asimétrica negativa $A_B<0$

Medidas de forma

- Coeficiente de **curtosis o apuntamiento** g_2

$$g_2 = \frac{m_4}{S_x^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^4}{S_x^4} - 3$$

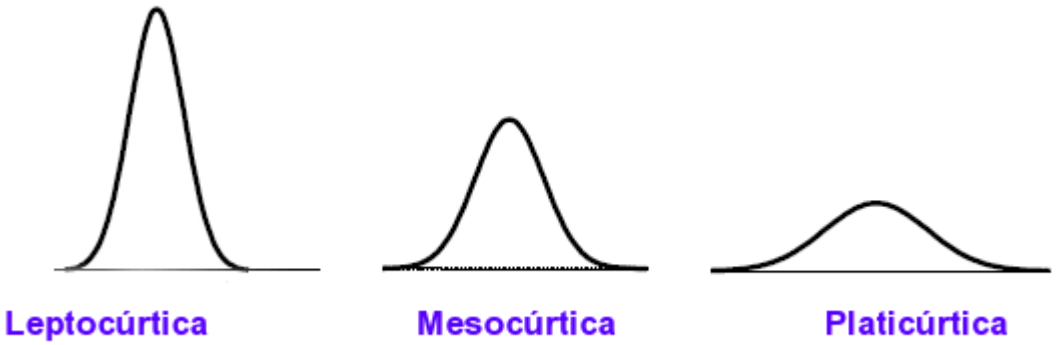
Si $g_2 > 0$, está más apuntada que la distribución normal, se denomina leptocúrtica.

Si $g_2 < 0$, está menos apuntada que la distribución normal, se denomina platicúrtica.

$g_2 = 0$, mismo apuntamiento que la normal: mesocúrtica.

Medidas de forma

- Coeficiente de **curtosis o apuntamiento** g_2



Si $g_2 > 0$, está más apuntada que la distribución normal, se denomina leptocúrtica.

Si $g_2 < 0$, está menos apuntada que la distribución normal, se denomina platicúrtica.

$g_2 = 0$, mismo apuntamiento que la normal: mesocúrtica.

Tipificar/estandarizar una variable

- Una variable se denomina tipificada o estandarizada, si su media es cero y su varianza es uno.
- Para tipificar una variable se hace la siguiente transformación:

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

CONCENTRACIÓN

- Estudiamos la CONCENTRACIÓN para variables cuantitativa positivas en las cuales la suma de los valores individuales tiene el sentido de un "todo" del cual cada individuo participa con una "parte". La idea es analizar el grado de igualdad o falta de esta en el reparto del "todo".
- Ejemplos:
  - La riqueza de la población de un país
  - Los salarios de los empleados de una empresa o de un sector
  - La población de los municipios de una provincia

  No tiene sentido en variables como la altura, el número del pie, etc…

- La concentración oscila entre una situación en la cual un individuo tiene el "todo" y el resto no tiene nada (máxima concentración) y una situación en la que todos los individuos tienen exactamente la misma cantidad (concentración mínima).

- Construiremos el Índice de Gini para medir estas situaciones y la curva de Lorenz para visualizar el grado de concentración

Concentración: índice de Gini

Se ordenan los valores de la variable (ó las clases) de menor a mayor.

Se comparan cantidades acumuladas por los individuos (o clases) con frecuencias acumuladas de individuos

$P_i = N_i/N$ proporciones acumuladas de población

$q_i = u_i/u_n$ proporciones acumuladas de la vble, donde $ui = \sum_{j=1}^{i} x_j n_j$

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

- ✓ El índice de Gini está entre 0 y 1
- ✓ El índice de Gini toma el valor 0 cuando hay igualdad, es decir si todos los individuos disponen de igual "parte" del "todo" (mínima concentración)
- ✓ El índice de Gini toma el valor 1 cuando hay máxima desigualdad, es decir, un individuo dispone del "todo" y el resto de individuos no tienen ninguna "parte" (máxima concentración)

Concentración: índice de Gini

Cliente	Ventas	ni
A	100	1
B	200	1
C	625	1
D	675	1
total	1600	4

$$I_G = \frac{0.672}{1.5} = 0.448$$

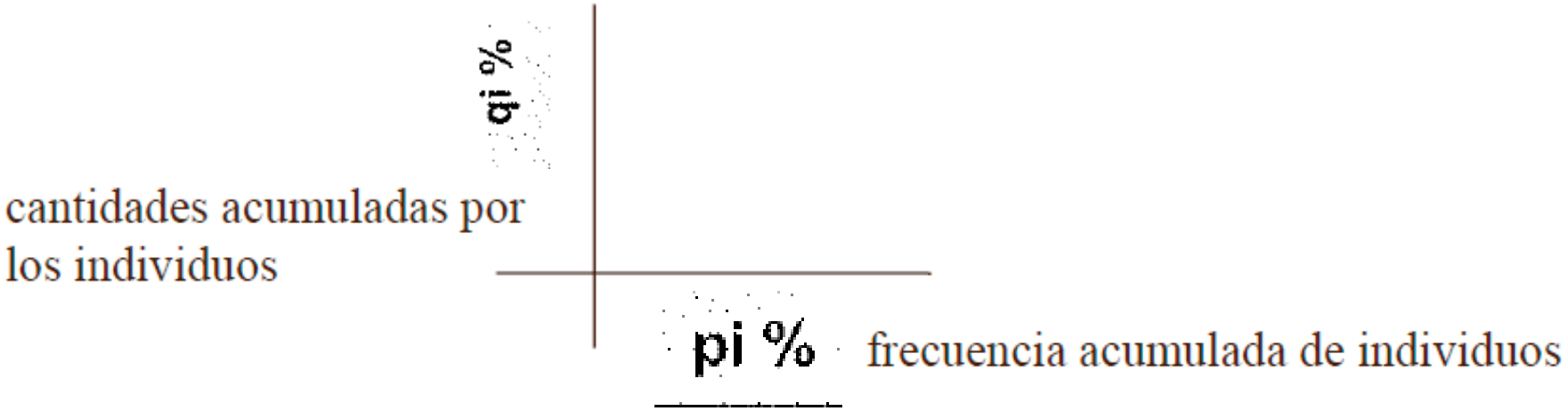
Cliente	Ventas	ni
A	400	1
B	400	1
C	400	1
D	400	1
total	1600	4

$$I_G = \frac{0}{1.5} = 0$$

Cliente	Ventas	ni
A	0	1
B	0	1
C	0	1
D	1600	1
total	1600	4

$$I_G = \frac{1.5}{1.5} = 1$$

Concentración: curva de Lorenz

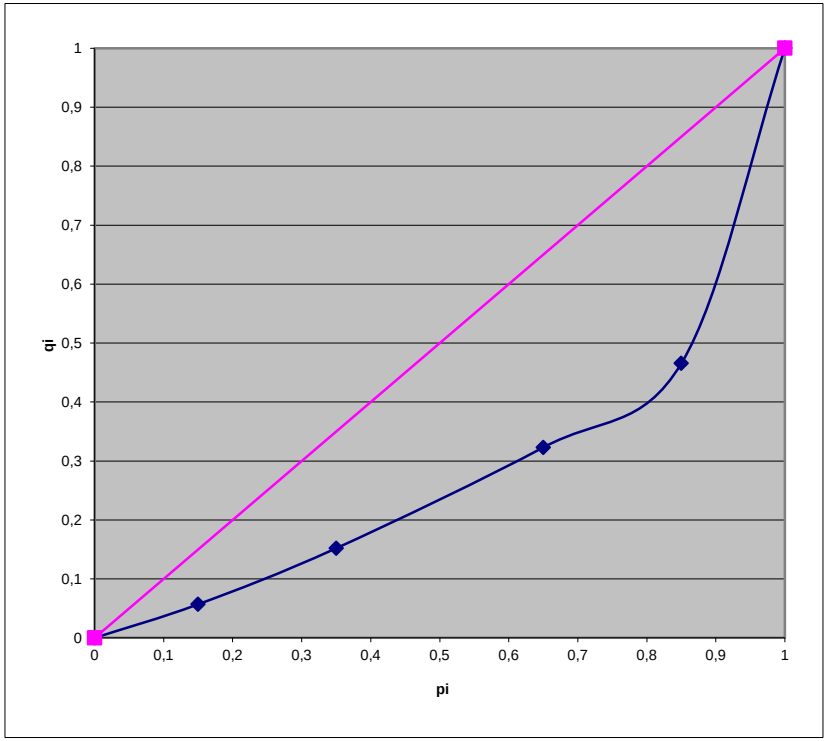


Ejemplo:

Empresa A	
nº personas	salario percibido (€)
15	800
20	1000
30	1200
20	1500
15	7500

Concentración: curva de Lorenz

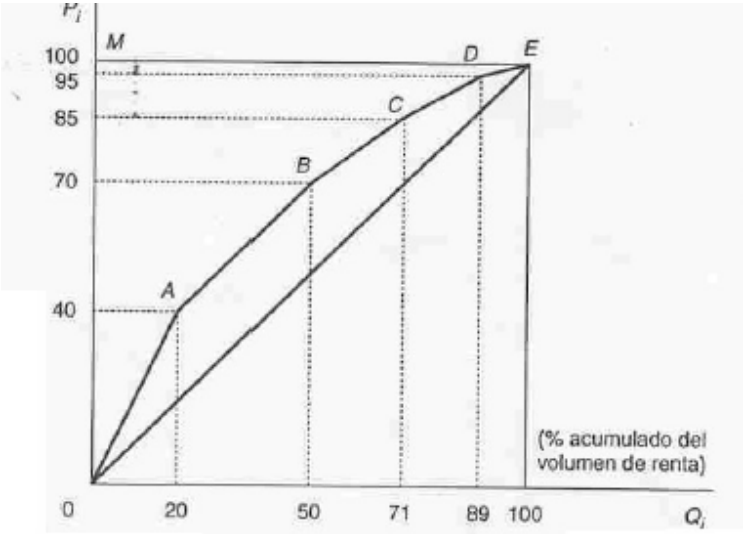
Empresa A							
xi	ni	Ni	xi*ni	ui: Acum xi*ni	pi	qi	pi-qi
					0	0	
800	15	15	12000	12000	0,15	0,057	0,093
1000	20	35	20000	32000	0,35	0,152	0,198
1200	30	65	36000	68000	0,65	0,323	0,327
1500	20	85	30000	98000	0,85	0,466	0,384
7500	15	100	112500	210500	1	1,000	0,000
			Sumas n-1		2		1,0023753
IGA	0,50118765´=1,0023753/2						



ejemplo: curva de Lorenz

Distribución de frecuencias absolutas de la renta de 20.000 familias

Renta en millones de pesetas $[L_{i-1}, L_i)$	*Número de familias* (× 1.000) f_i
[1, 4)	8
[4, 6)	6
[6, 8)	3
[8, 10)	2
[10, 12]	1
Total	20



Curva de Lorenz (datos de la tabla

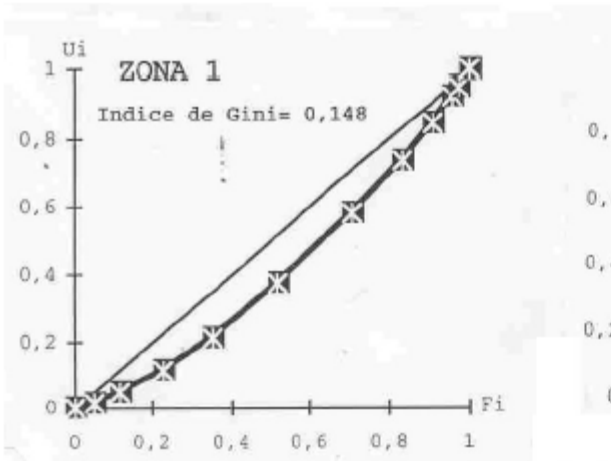
Ejemplo concentración

Ingresos en los hogares por zonas (millones pts). Concentración de los ingresos.

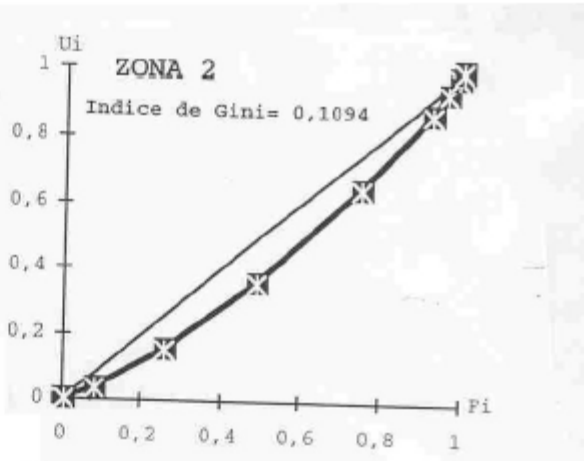
		ZONA 1			ZONA 2		
Intervalo	x_{ci}	n_i	F_i	U_i	n_i	F_i	U_i
0,25 – 0,74	0,5	0	0,0000	0,0000	1	0,0083	0,0018
0,75 – 1,24	1,0	6	0,0500	0,0157	9	0,0833	0,0350
1,25 – 1,74	1,5	8	0,1167	0,0472	21	0,2583	0,1510
1,75 – 2,24	2,0	13	0,2250	0,1153	28	0,4917	0,3573
2,25 – 2,74	2,5	15	0,3500	0,2136	31	0,7500	0,6427
2,75 – 3,24	3,0	20	0,5167	0,3709	21	0,9250	0,8748
3,25 – 3,74	3,5	23	0,7083	0,5819	5	0,9667	0,9392
3,75 – 4,24	4,0	15	0,8333	0,7392	3	0,9917	0,9834
4,25 – 4,74	4,5	9	0,9083	0,8453	1	1,0000	1,0000
4,75 – 5,24	5,0	6	0,9583	0,9240	0		
5,25 – 5,74	5,5	2	0,9750	0,9528	0		
5,75 – 6,24	6,0	3	1,0000	1,0000	0		
	Suma	120	6,6417	5,8060	120	5,4750	4,9853

ZONA 1	IG= 0,1481
ZONA 2	IG= 0,1094

¿conclusión?



Ingresos en los hogares por zonas (millones pts). Curva de Lorenz en la ZONA 1



Ingresos en los hogares por zonas (millones pts). Curva de Lorenz en la ZONA 2

Concentración baja en las dos zonas.

Menor concentración en zona 2

3.9 Establezca, con base estadística, en cuál de las siguientes empresas el salario está repartido de forma más equitativa.

$$ui = \sum_{j=1}^{i} x_j n_j$$

$P_i = N_i / N$

$q_i = u_i / u_n$

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

Empresa A		Empresa B	
nº personas	salario percibido (€)	nº personas	salario percibido (€)
15	800	10	800
20	1000	30	1000
30	1200	35	1200
20	1500	24	1500
15	7500	1	7500

¿Qué conclusiones puede obtener del análisis de las correspondientes curvas de Lorenz?

Empresa A

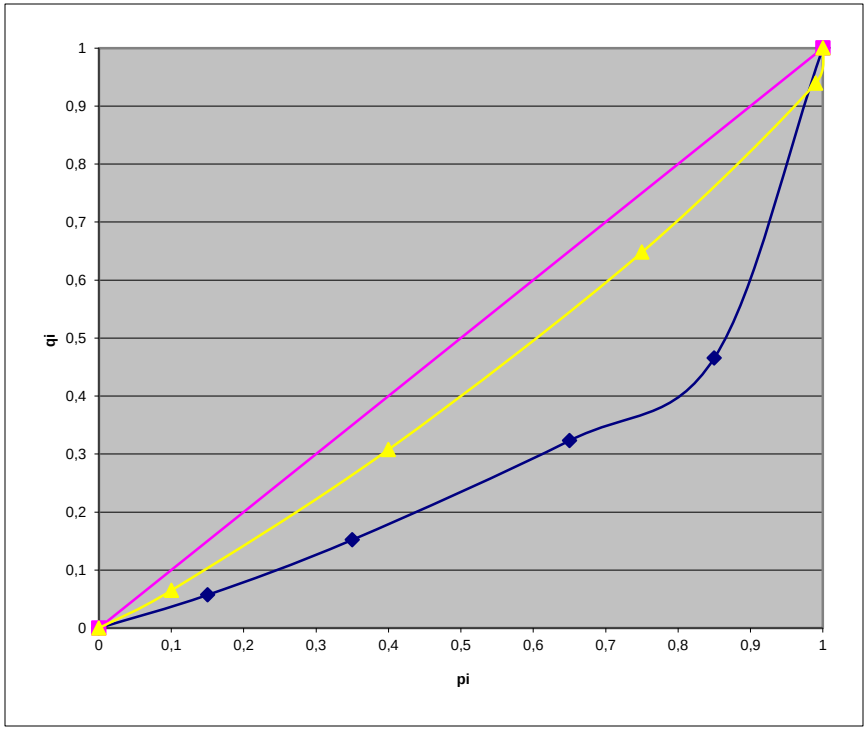
xi	ni	Ni	xi*ni	ui: Acum xi*ni	pi	qi	pi-qi
					0	0	
800	15	15	12000	12000	0,15	0,057	0,093
1000	20	35	20000	32000	0,35	0,152	0,198
1200	30	65	36000	68000	0,65	0,323	0,327
1500	20	85	30000	98000	0,85	0,466	0,384
7500	15	100	112500	210500	1	1,000	0,000
			Sumas n-1		2		1,0023753

IGA 0,50118765´=1,0023753/2

Empresa B

xi	ni	Ni	xi*ni	ui: Acum xi*ni	pi	qi	pi-qi
					0	0	
800	10	10	8000	8000	0,1	0,065	0,035
1000	30	40	30000	38000	0,4	0,308	0,092
1200	35	75	42000	80000	0,75	0,648	0,102
1500	24	99	36000	116000	0,99	0,939	0,051
7500	1	100	7500	123500	1	1,000	0,000
			Sumas n-1		2,24		0,28048583

IGB 0,12521689´=0,28048583/2,24


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
qi
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
pi