

# Bioestadística

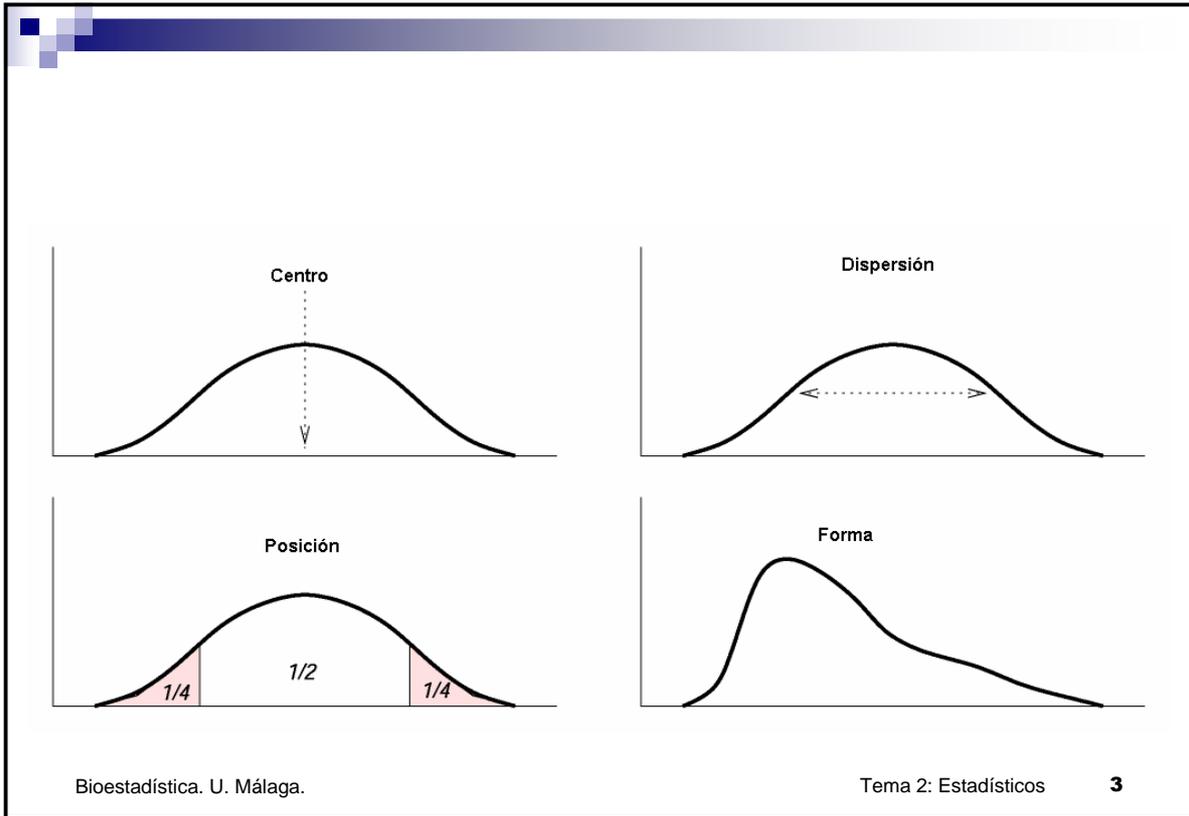
## Tema 2: Estadísticos

## 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:** Ídem (cambiar población por muestra)
  - La altura media de los que estamos en este aula.
    - Somos una muestra (¿representativa?) de la población.
  - Si un estadístico se usa para aproximar un parámetro también se le suele llamar **estimador**.



Normalmente nos 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. Más adelante veremos como elegir muestras para que el error sea “confiablemente” pequeño.



## Un brevísimo resumen sobre estadísticos

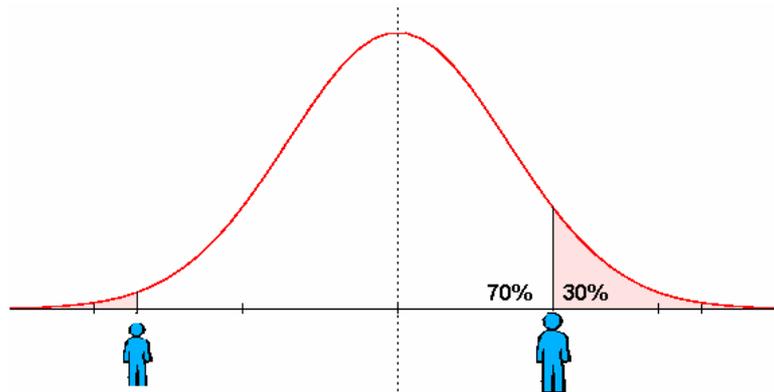
- **Posición**
  - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
    - Cuantiles, percentiles, cuartiles, deciles,...
- **Centralización**
  - Indican valores con respecto a los que los datos parecen agruparse.
    - Media, mediana y moda
- **Dispersión**
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
    - Desviación típica, coeficiente de variación, rango, varianza
- **Forma**
  - Asimetría
  - Apuntamiento o curtosis



Bioestadística. U. Málaga. Tema 2: Estadísticos **4**

## Estadísticos de posición

- Se define el **cuantil** de orden  $\alpha$  como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada  $\alpha$ .
- Casos particulares son los percentiles, cuartiles, deciles, quintiles,...



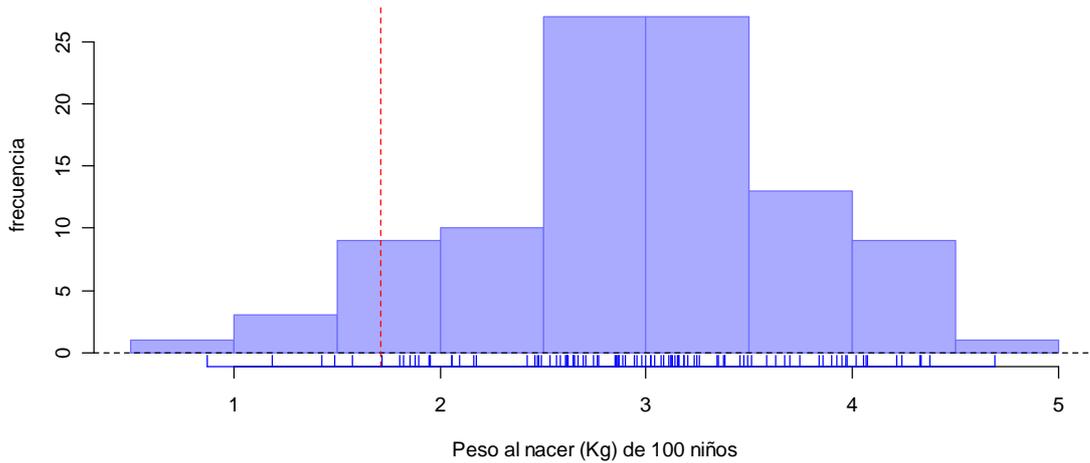
## Estadísticos de posición

- **Percentil** de orden  $k$  = cuantil de orden  $k/100$ 
  - 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 = Percentil 25 = Cuantil 0,25
  - Segundo cuartil = Percentil 50 = Cuantil 0,5 = mediana
  - Tercer cuartil = Percentil 75 = cuantil 0,75

## Ejemplos

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

Percentil 5 del peso



Bioestadística. U. Málaga.

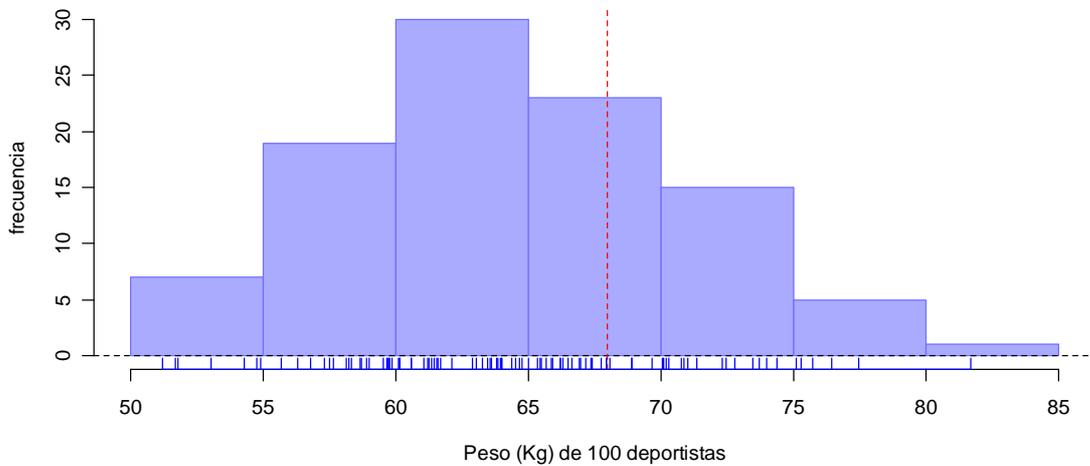
Tema 2: Estadísticos

7

## Ejemplos

- ¿Qué peso es superado sólo por el 25% de los individuos?
  - **Percentil 75 o tercer cuartil**

Percentil 75 del peso



Bioestadística. U. Málaga.

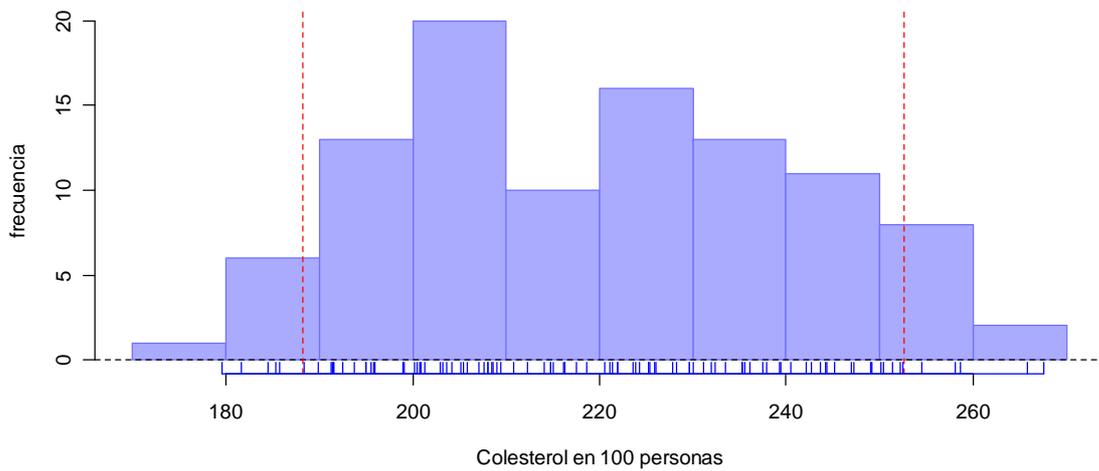
Tema 2: Estadísticos

8

## 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?

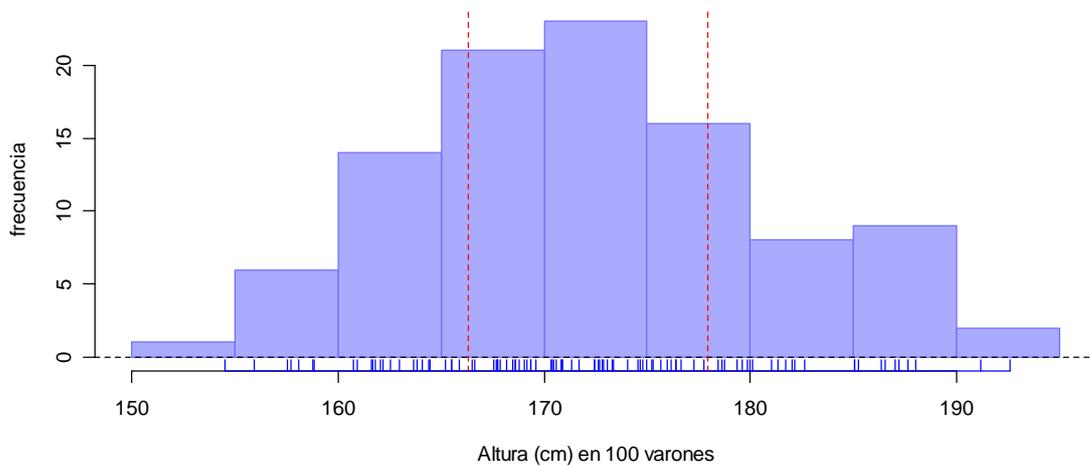
Percentiles 5 y 95



## Ejemplos

- ¿Entre qué valores se encuentran la mitad de los individuos “más normales” de una población?
  - Entre el cuartil 1º y 3º

Percentiles 25 y 75



# Diagramas de Tukey

- **Resumen con 5 números:**
  - Mínimo, cuartiles y máximo.
  - Suelen dar una buena idea de la distribución.
- La zona central, 'caja', contiene al 50% central de las observaciones.
  - Su tamaño se llama 'rango intercuartílico' (R.I.)
- Es costumbre que 'los bigotes', no lleguen hasta los extremos, sino hasta las observaciones que se separan de la caja en no más de 1,5 R.I.
  - Más allá de esa distancia se consideran anómalas, y así se marcan.

Diagrama de cajas de Tukey: Resumen en 5 números

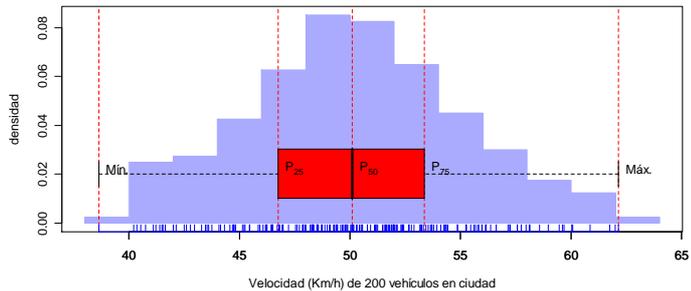
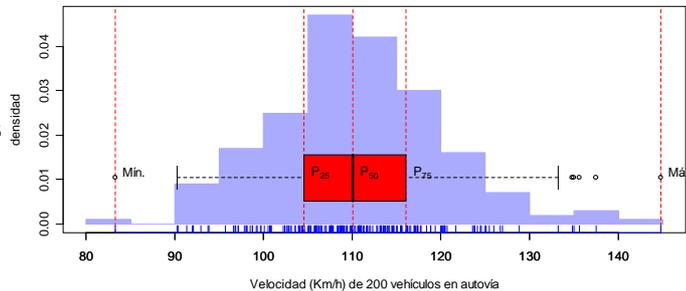


Diagrama de cajas de Tukey: Resumen en 5 números



# 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
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
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	

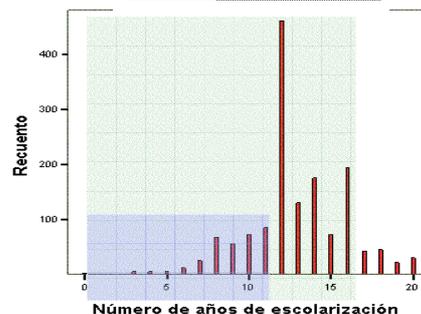
≥20%?

≥90%?

Estadísticos

Número de años de escolarización

N	Válidos	Perdidos
N	1508	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

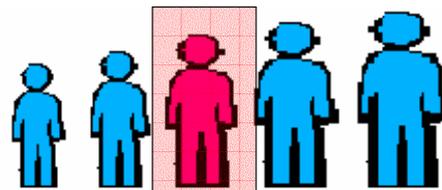
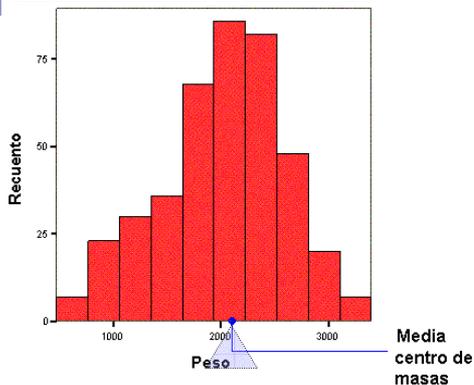


# Estadísticos de centralización

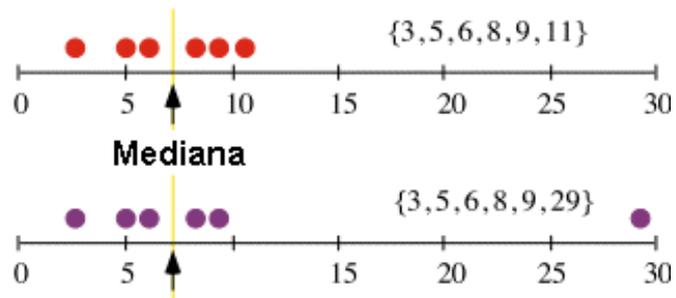
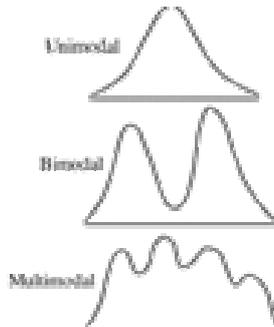
Añaden unos cuantos casos particulares a las medidas de posición. En este caso son medidas que buscan posiciones (valores) con respecto a los cuales los datos muestran tendencia a agruparse.



- **Media** ('mean') Es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
  - Media de 2,2,3,7 es  $(2+2+3+7)/4=3,5$
  - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
  - Centro de gravedad de los datos
- **Mediana** ('median') Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1,2,4,5,6,6,8 es 5
  - Mediana de 1,2,4,5,6,6,8,9 es  $(5+6)/2=5,5$
  - Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
    - Mediana de 1,2,4,5,6,6,800 es 5. ¡La media es 117,7!
- **Moda** ('mode') Es el/los valor/es donde la distribución de frecuencia alcanza un máximo.



Altura mediana



## Algunas fórmulas

- **Datos sin agrupar:**  $x_1, x_2, \dots, x_n$

- Media

$$\bar{x} = \frac{\sum_i x_i}{n}$$

- **Datos organizados en tabla**

- si está en intervalos usar como  $x_i$  las marcas de clase. Si no ignorar la columna de intervalos.

- Media

$$\bar{x} = \frac{\sum_i x_i n_i}{n}$$

Variable	fr.	fr. ac.
$L_0 - L_1$	$x_1$	$n_1$
$L_1 - L_2$	$x_2$	$n_2$
...		
$L_{k-1} - L_k$	$x_k$	$n_k$
n		

- Cuantil de orden  $\alpha$

- $i$  es el menor intervalo que tiene frecuencia acumulada superior a  $\alpha \cdot n$
- $\alpha=0,5$  es mediana

$$C_\alpha = L_{i-1} + \frac{\alpha \cdot n - N_{i-1}}{n_i} (L_i - L_{i-1})$$

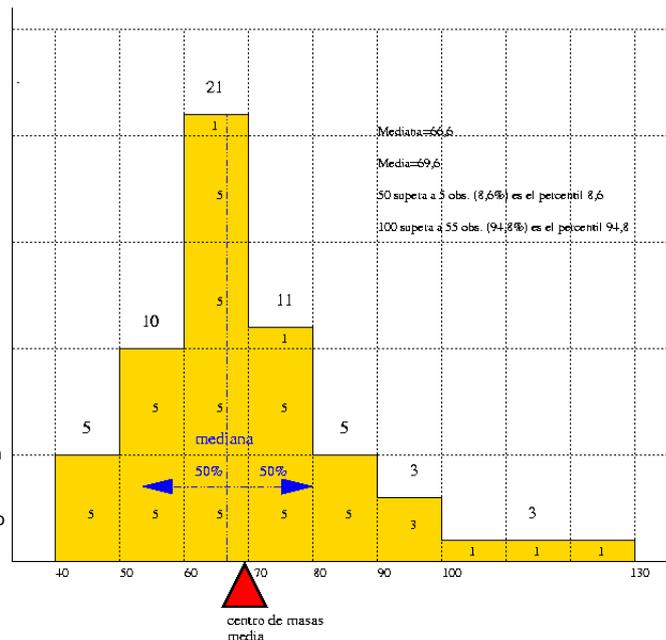
## Ejemplo con variable en intervalos

Peso	M. Clase	frec	Fr. acum.
40 - 50	45	5	5
50 - 60	55	10	15
60 - 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 - 130	115	3	58

En el histograma se identifica "unidad de área" con "individuo".

Para calcular la media es necesario elegir un punto representante del intervalo: La marca de clase.

La media se desplaza hacia los valores extremos. No coincide con la mediana. Es un punto donde el histograma "estaría en equilibrio" si tuviese masa.



## Ejemplo (continuación)

Peso	M. Clase	Fr.	Fr. ac.
40 - 50	45	5	5
50 - 60	55	10	15
60 - 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 - 130	115	3	58
			58

$$\bar{x} = \frac{\sum_i x_i n_i}{n} = \frac{45 \cdot 5 + 55 \cdot 10 + \dots + 115 \cdot 3}{58} = 69,3$$

$$\begin{aligned} \text{Mediana} &= C_{0,5} = L_{i-1} + \frac{0,5 \cdot 58 - N_{i-1}}{n_i} (L_i - L_{i-1}) \\ &= 60 + \frac{0,5 \cdot 58 - 15}{21} (70 - 60) = 66,6 \end{aligned}$$

$$P_{75} = C_{0,75} = L_{i-1} + \frac{0,75 \cdot 58 - N_{i-1}}{n_i} (L_i - L_{i-1}) = 70 + \frac{43,5 - 36}{11} (80 - 70) = 76,8$$

- Moda = marca de clase de (60,70] = 65
  - Cada libro ofrece una fórmula diferente para la moda (difícil estar al día.)

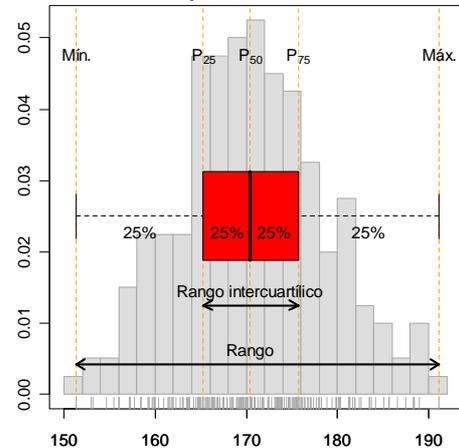
## Variabilidad o dispersión

- Los estudiantes de Bioestadística reciben diferentes calificaciones en la asignatura (**variabilidad**). ¿A qué puede deberse?
  - **Diferencias individuales en el conocimiento** de la materia.
- ¿Podría haber otras razones (**fuentes de variabilidad**)?
- Por ejemplo supongamos que todos los alumnos poseen el mismo nivel de conocimiento. ¿Las notas serían las mismas en todos? Seguramente No.
  - Dormir poco el día del examen, el croissant estaba envenenado...
    - **Diferencias individuales en la habilidad** para hacer un examen.
  - El examen no es una medida perfecta del conocimiento.
    - **Variabilidad por error de medida.**
  - En alguna pregunta difícil, se duda entre varias opciones, y al azar se elige la mala
    - **Variabilidad por azar, aleatoriedad.**

## Medidas de dispersión

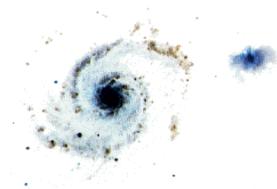
Miden el grado de dispersión (variabilidad) de los datos, independientemente de su causa.

- **Amplitud o Rango** ('range'): Diferencia entre observaciones extremas.
  - 2, 1, 4, 3, 8, 4. El rango es  $8-1=7$
  - Es muy sensible a los valores extremos.
- **Rango intercuartílico** ('interquartile range'):
  - Es la distancia entre primer y tercer cuartil.
    - Rango intercuartílico =  $P_{75} - P_{25}$
  - Parecida al rango, pero eliminando las observaciones más extremas inferiores y superiores.
  - No es tan sensible a valores extremos.



- **Varianza  $S^2$**  ('Variance'): Mide el promedio de las desviaciones (al cuadrado) de las observaciones con respecto a la media.

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



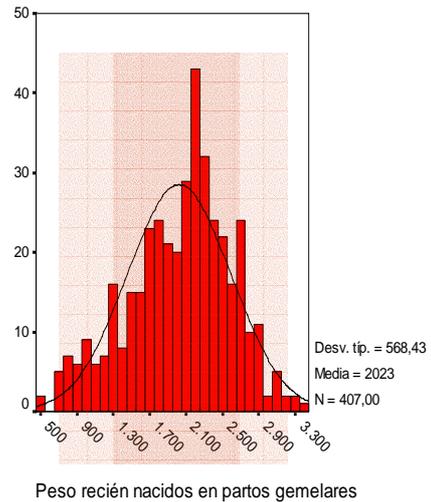
- Es sensible a valores extremos (alejados de la media).
- Sus unidades son el cuadrado de las de la variable. De interpretación difícil para un principiante.
- La expresión es fea, pero de gran belleza 'natural' (físicamente). Contiene la información geométrica relevante en muchas situaciones donde la energía interna de un sistema depende de la posición de sus partículas.
  - Energía de rotación (vía el coeficiente de inercia): patinadores con brazos extendidos (dispersos) o recogidos (poco dispersos)
  - Energía elástica: Muelles 'estirados' con respecto a su posición de equilibrio (dispersos) frente a muelles en posición cercana a su posición de equilibrio (poco dispersos)

## Desviación típica ('standard deviation')

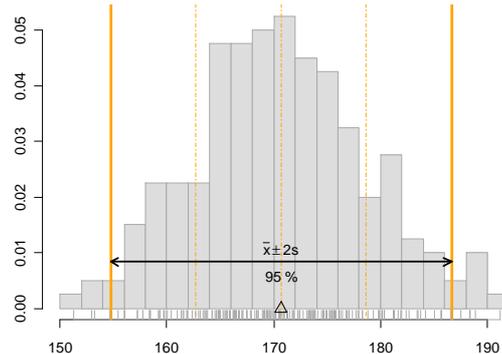
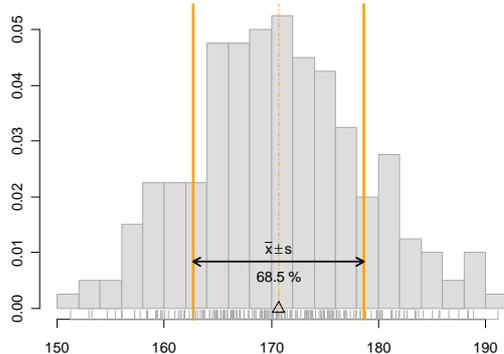
Es la raíz cuadrada de la varianza

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

- Tiene la misma dimensionalidad (unidades) que la variable. Versión 'estética' de la varianza.
- Cierta distribución que veremos más adelante (**normal o gaussiana**) quedará completamente determinada por la media y la desviación típica.
  - A una distancia de una desv. típica de la media hay más de la 'más de la mitad'.
  - A una distancia de dos desv. típica de la media las tendremos casi todas.

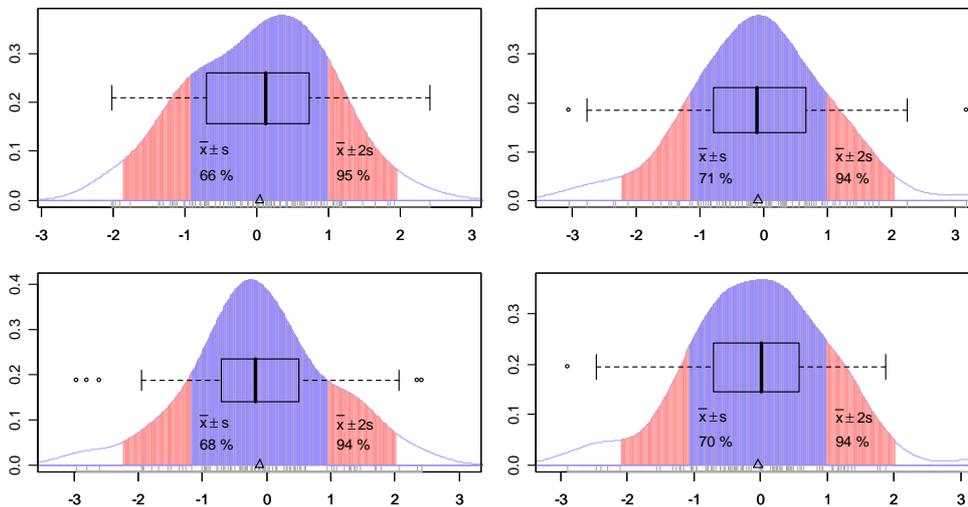


## Dispersión en distribuciones 'normales'



- Centrado en la media y a una desv. típica de distancia hay aproximadamente el 68% de las observaciones.
- A dos desviaciones típicas tenemos el 95% (aprox.)

- Datos 'casi normales'. Eje 'x' medido en desviaciones típicas...
  - ¿Encuentras relación entre rango intercuartílico y desviación típica?
  - ¿Y entre los 'bigotes' y dos desviaciones típicas? ¿Podrías caracterizar las observaciones anómalas?



## Coeficiente de variación

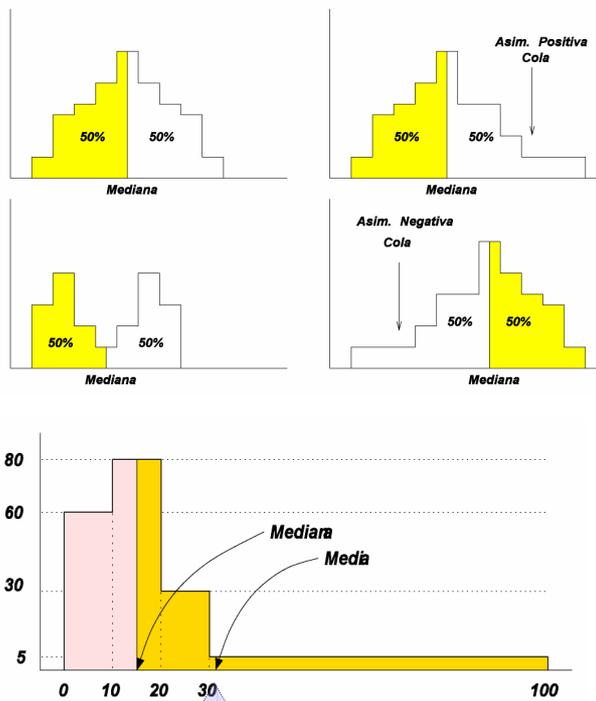
Es la razón entre la desviación típica y la media.

- Mide la desviación típica en forma de "qué tamaño tiene con respecto a la media"
- 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.
- No debe usarse cuando la variable presenta valores negativos o donde el valor 0 sea una cantidad fijada arbitrariamente
  - Por ejemplo  $0^{\circ}C \neq 0^{\circ}F$
- Los ingenieros electrónicos hablan de la razón 'señal/ruido' (su inverso).

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

## Asimetría o Sesgo

- Una distribución es simétrica si la mitad izquierda de su distribución es la imagen especular de su mitad derecha.
- En las distribuciones simétricas media y mediana coinciden. Si sólo hay una moda también coincide
- La asimetría es positiva o negativa en función de a qué lado se encuentra la cola de la distribución.
- La media tiende a desplazarse hacia los valores extremos (colas).
- Las discrepancias entre las medidas de centralización son indicación de asimetría.

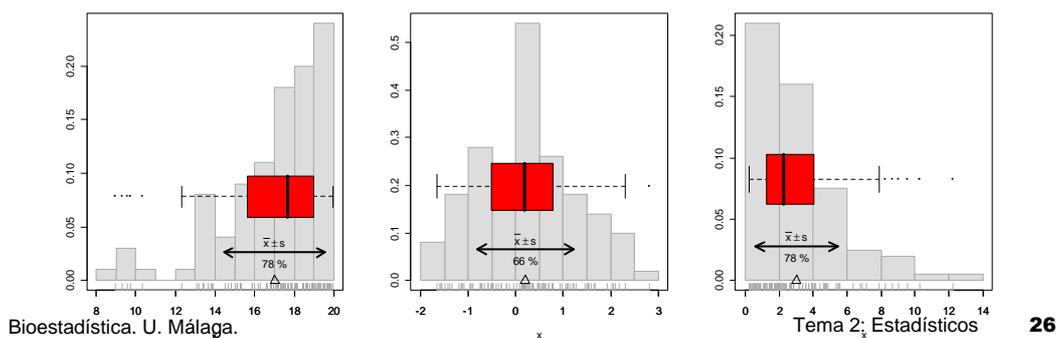


Bioestadística. U. Málaga.

Tema 2: Estadísticos 25

## Estadísticos para detectar asimetría

- Hay diferentes estadísticos que sirven para detectar asimetría.
  - Basado en diferencia entre estadísticos de tendencia central.
  - Basado en la diferencia entre el 1º y 2º cuartiles y 2º y 3º.
  - Basados en *desviaciones con signo al cubo con respecto a la media*.
    - Los calculados con ordenador. Es pesado de hacer a mano.
- En función del signo del estadístico diremos que la asimetría es *positiva* o *negativa*.
  - Distribución simétrica → asimetría nula.



Bioestadística. U. Málaga.

Tema 2: Estadísticos 26

## Apuntamiento o curtosis

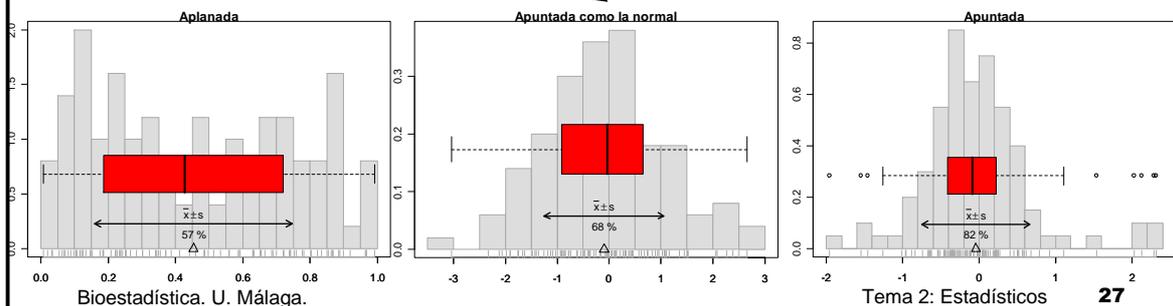
La **curtosis** nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.

**Platicúrtica (aplanada):** curtosis  $< 0$

**Mesocúrtica (como la normal):** curtosis  $= 0$

**Leptocúrtica (apuntada):** curtosis  $> 0$

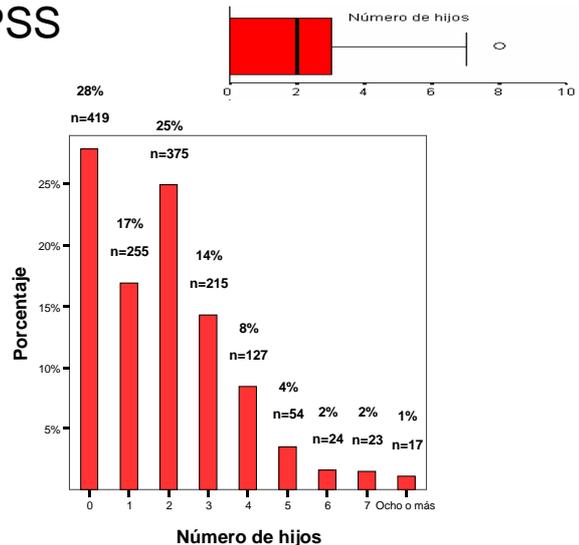
En el curso serán de especial interés las mesocúrticas y simétricas (parecidas a la normal).



## Ejercicio: descriptiva con SPSS

Descriptivos para Número de hijos

		Estadístico	Error típ.
Media		1,90	,045
Intervalo de confianza para la media al 95%	Límite inferior	1,81	
	Límite superior	1,99	
Media recortada al 5%		1,75	
Mediana		2,00	
Varianza		3,114	
Desv. típ.		1,765	
Mínimo		0	
Máximo		8	
Rango		8	
Amplitud intercuartil		3,00	
Asimetría		1,034	,063
Curtosis		1,060	,126



- Está sombreado lo que sabemos interpretar hasta ahora. Verifica que comprendes todo. ¿Qué unidades tiene cada estadístico? ¿Variabilidad relativa?
- Calcula los estadísticos que puedas basándote sólo en el gráfico de barras.

## ¿Qué hemos visto?

- Parámetros
- Estadísticos y estimadores
- Clasificación
  - Posición (cuantiles, percentiles,...)
    - Diagramas de cajas
  - Medidas de centralización: Media, mediana y moda
    - Diferenciar sus propiedades.
  - Medidas de dispersión
    - con unidades: rango, rango intercuartílico, varianza, desv. típica
    - sin unidades: coeficiente de variación
      - ¿Qué usamos para comparar dispersión de dos poblaciones?
  - Asimetría
    - positiva
    - negativa
      - ¿Podemos observar asimetría sin mirar la gráfica?
      - ¿Cómo me gustan los datos?
  - Medidas de apuntamiento (curtosis)
    - ¿Cómo me gustan los datos?

