Módulo de Estadística

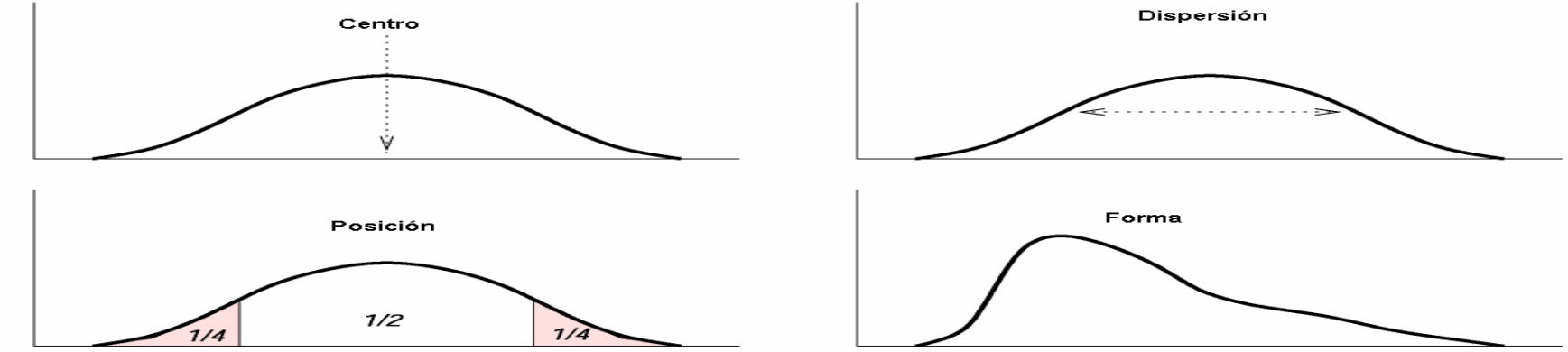
Tema 2: Estadística descriptiva

Medidas

- La finalidad de las **medidas de posición o tendencia central (centralización)** es encontrar unos valores que sinteticen o resuman las distribuciones de frecuencias

- Las **medidas de dispersión.** Estudian lo concretada que está la distribución de datos entorno a algún promedio.

- Las **medidas de asimetría** tienen como finalidad el elaborar un indicador que permita establecer el grado de simetría (o asimetría) que presenta una distribución sin necesidad de una representación gráfica.



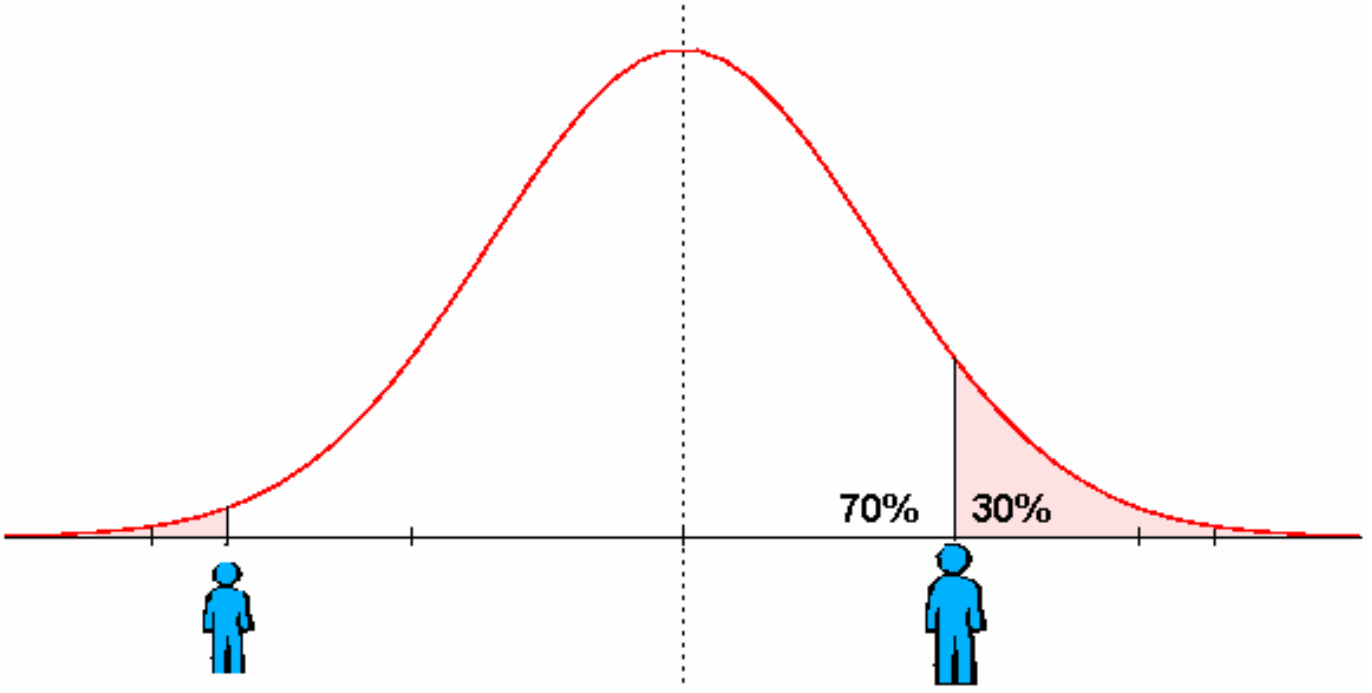
Un brevísimo resumen sobre medidas

- 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 (promedio), 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

Medidas de posición

- Se define el **cuantil** de orden α como un valor de la variable por debajo del cual se encuentra una frecuencia acumulada α.

- Casos particulares son los **percentiles**, **cuartiles, deciles, quintiles,...**



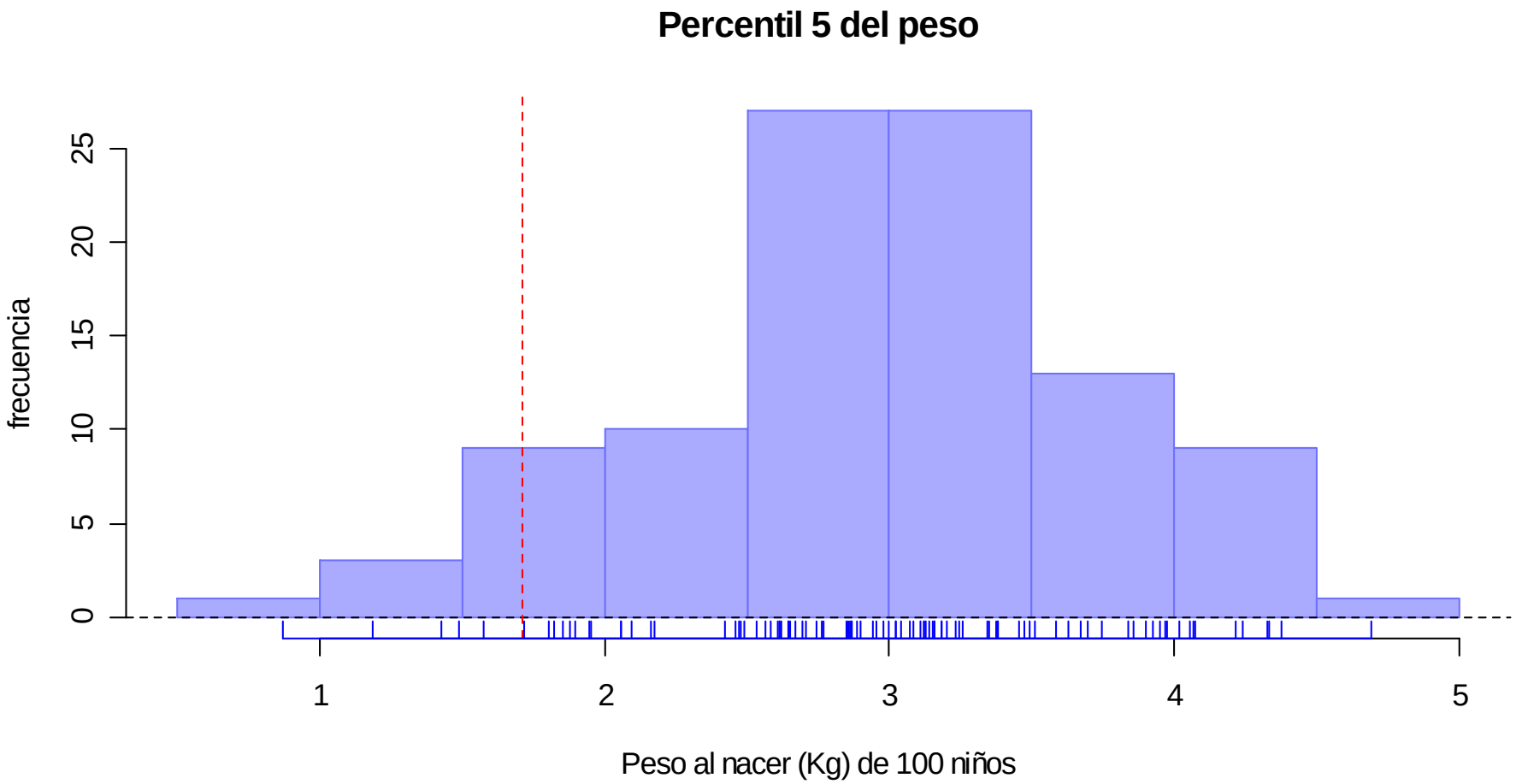
Medidas de posición

- **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

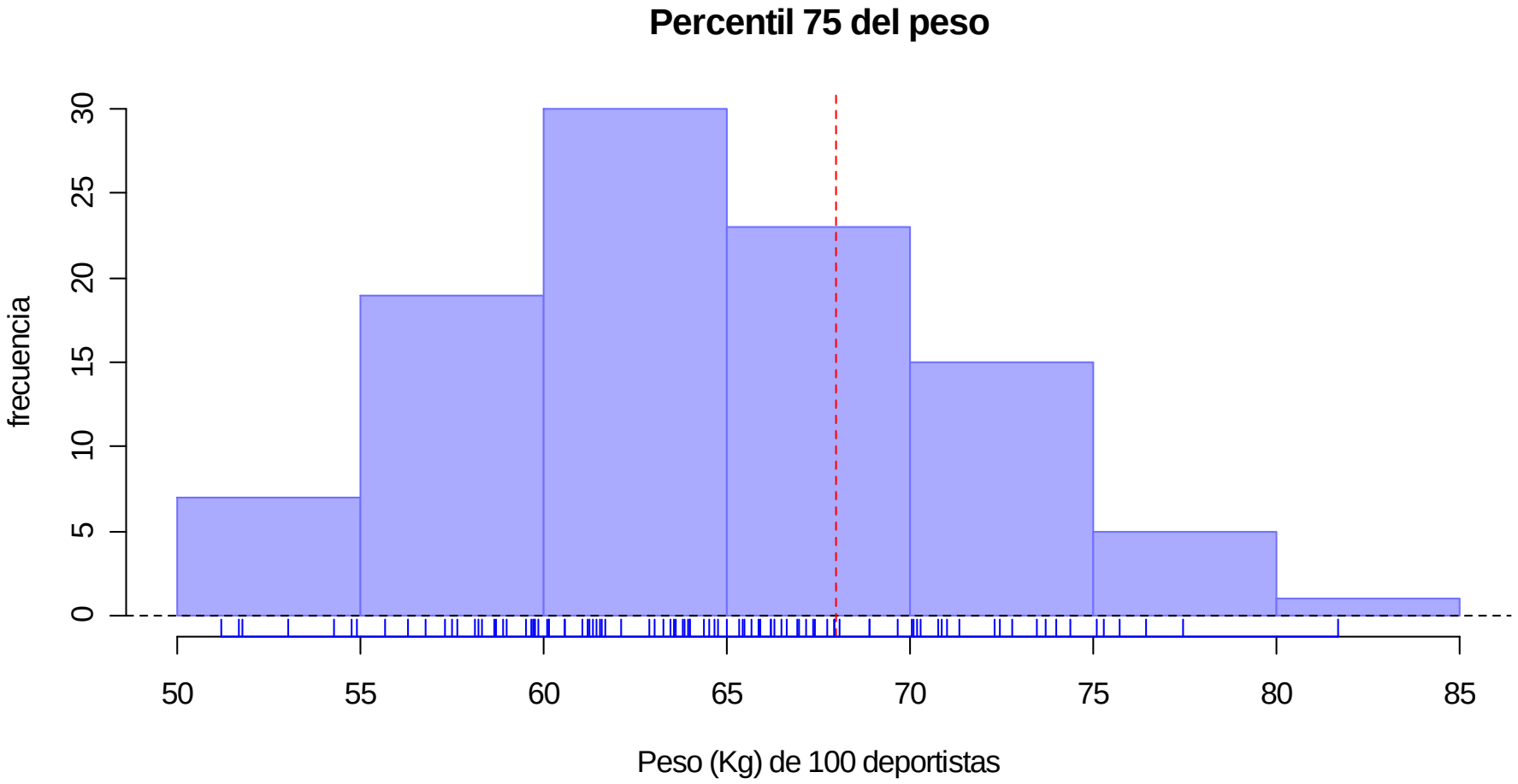
- El 5% de los recién nacidos tiene un peso demasiado bajo. ¿Qué peso se considera “demasiado bajo”?
  - Percentil 5 o cuantil 0,05



Ejemplos

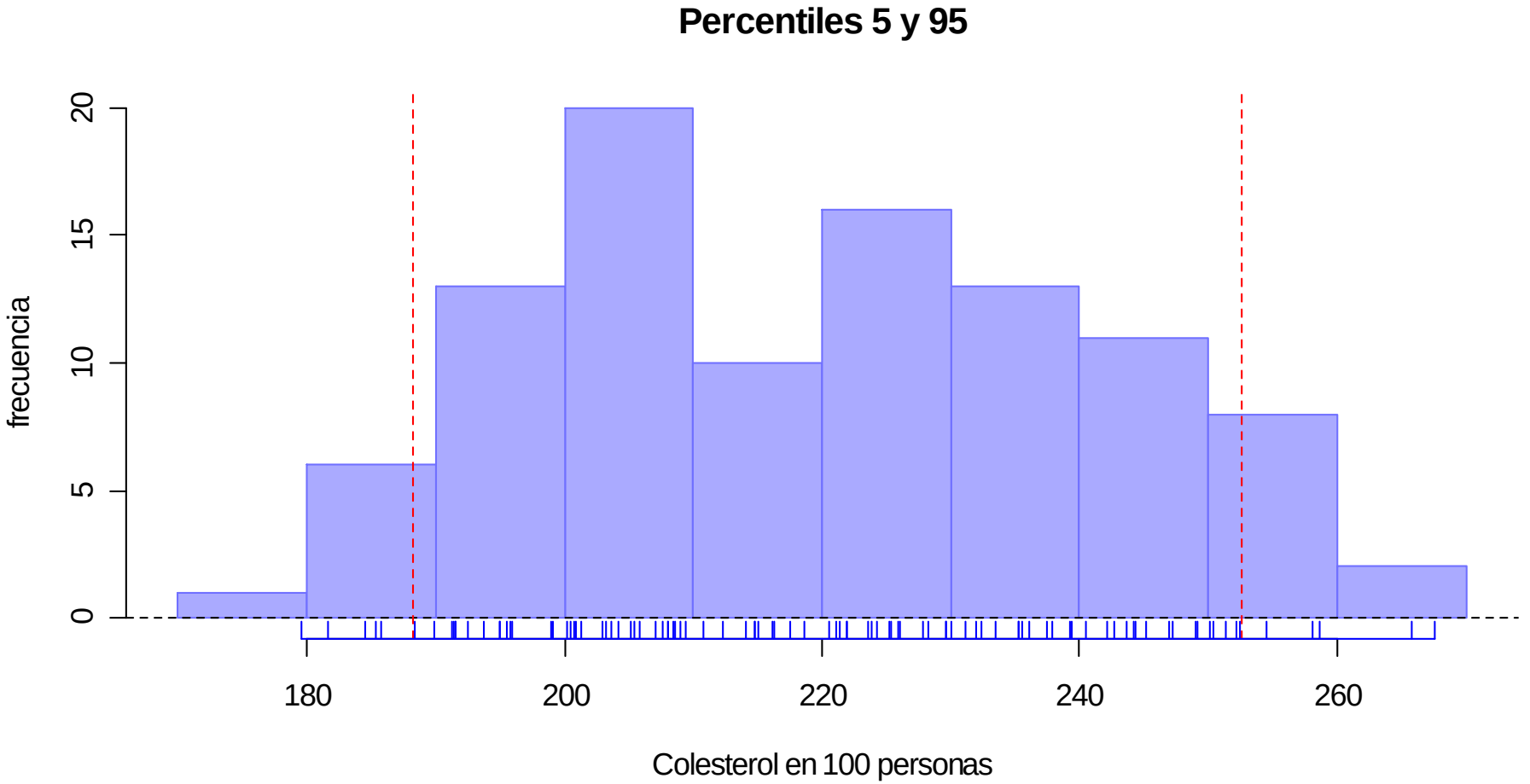
¿Qué peso es superado sólo por el 25% de los individuos?

- Percentil 75 o tercer cuartil



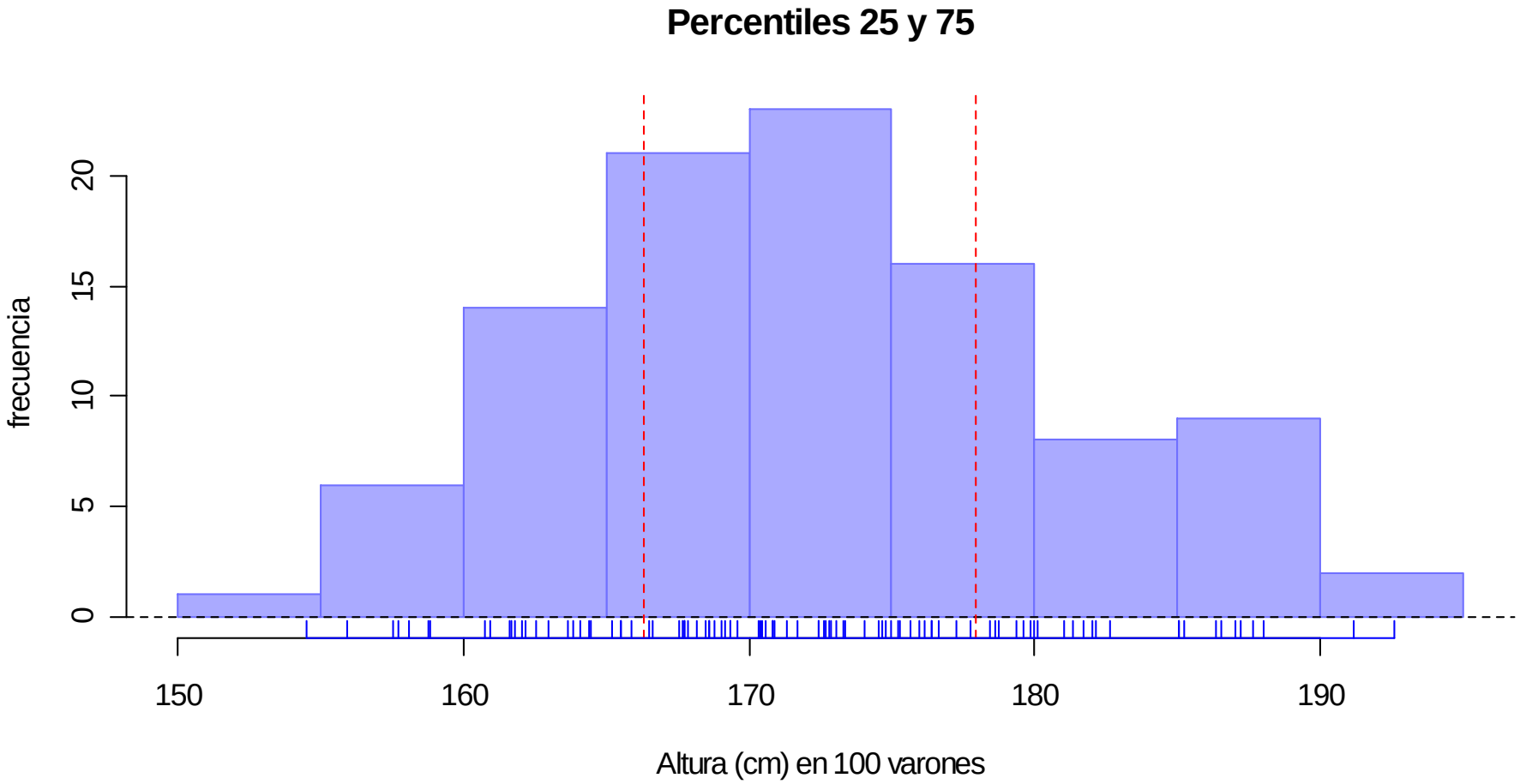
Ejemplos

- El colesterol (mg/100ml) 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?



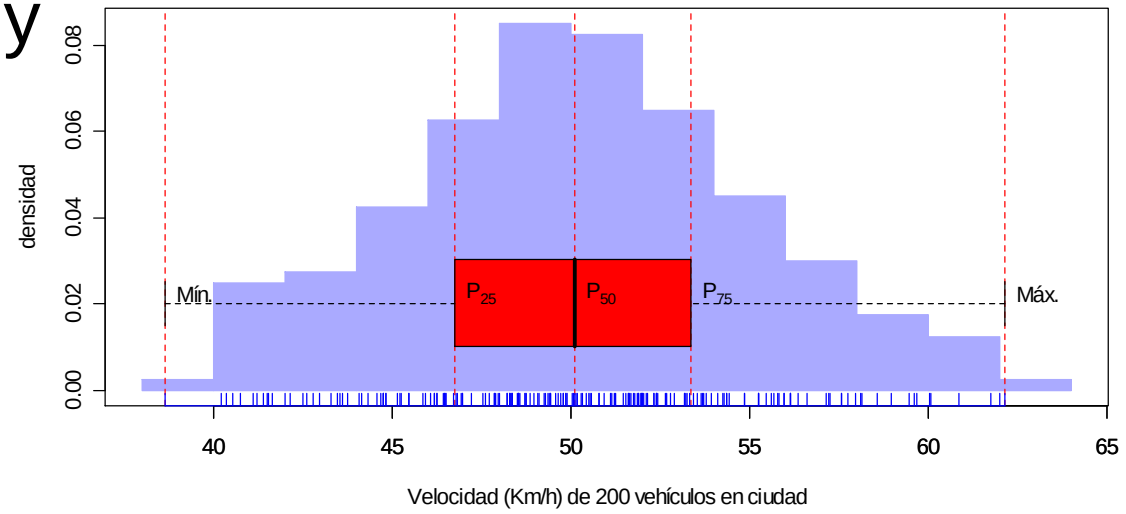
Ejemplos

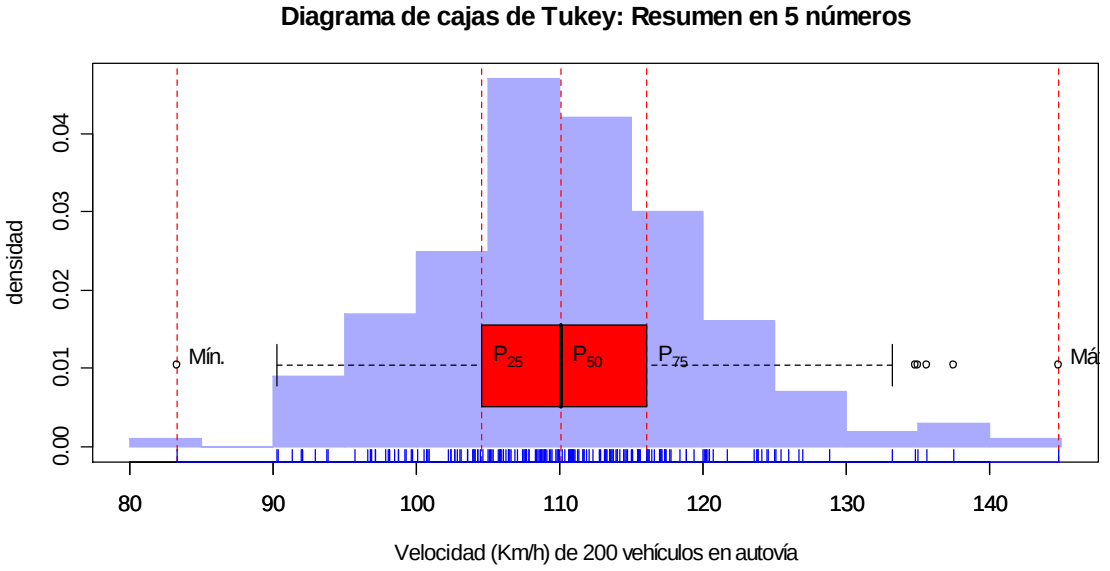
- ¿Entre qué valores se encuentran la mitad de los individuos “más normales” de una población?
  - Entre el cuartil 1º y 3º



Diagramas de Tukey (1997)

- 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.





Medidas de centralización

- 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

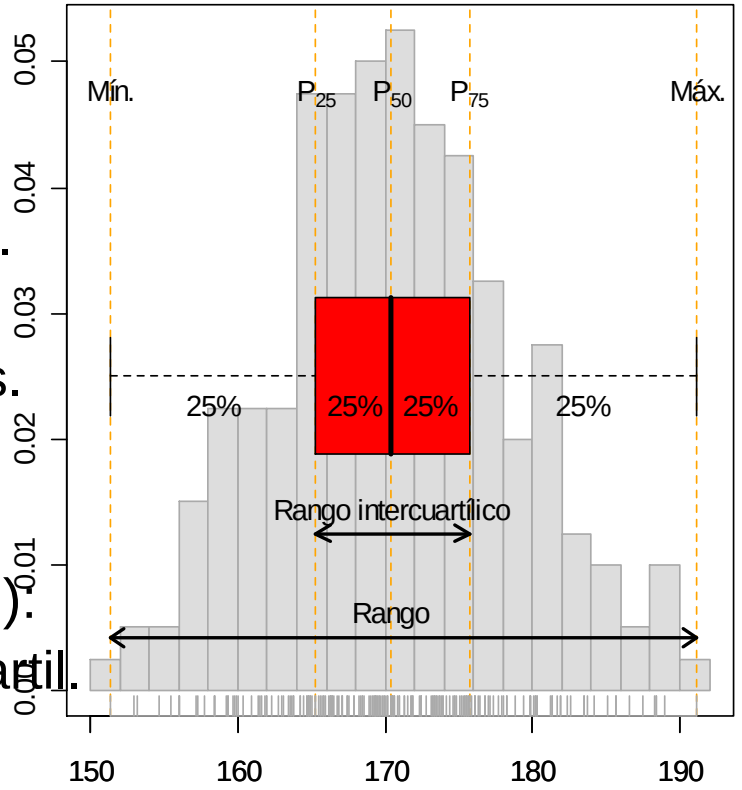
- Mediana (‘median’) Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1,2,4,**5**,6,6,8 es 5
  - Mediana de 1,2,4,**5**,**6**,6,8,9 es (5+6)/2=5,5
  - Es 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. El valor que mas se repite

Medidas de dispersión

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

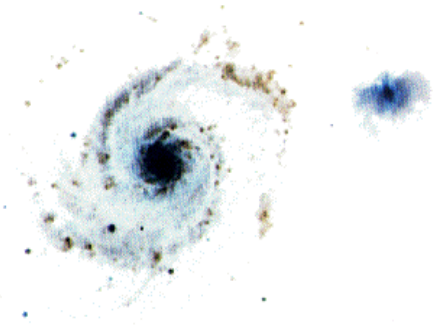
- Amplitud o Rango ('range'): Diferencia entre observaciónes extremas.
  - 2,**1**,4,3,**8**,4. El rango es 8-1=7
  - Es muy sensible a los valores extremos.



- 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 me

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



  - 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.

- **Desviación típica** (‘standard deviation’)
  Es la raíz cuadrada de la varianza

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

  - Tiene las misma dimensionalidad (unidades) que la variable. Versión ‘estética’ de la varianza

- **Coeficiente de variación**

  Es un estadístico de dispersión que tiene la ventaja de que no lleva asociada ninguna unidad, por lo que nos permitirá decir entre dos muestras, cual es la que presenta mayor dispersión.

$$CV = \frac{s}{x}(x100)$$

Apuntamiento o curtosis

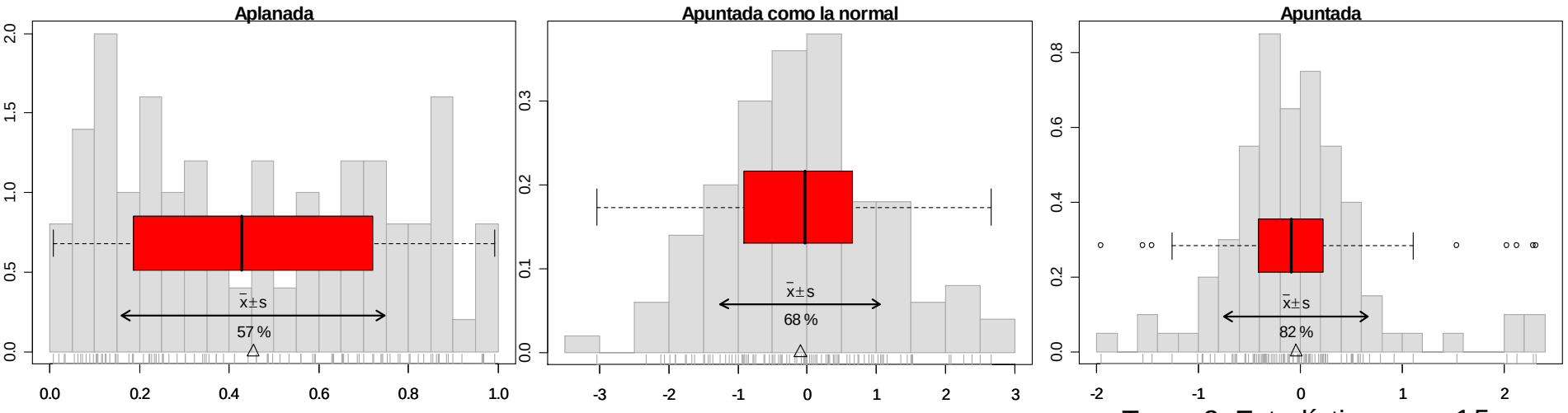
La curtosis nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.

Platicúrtica (aplanada): curtosis < 0

Mesocúrtica (como la normal): curtosis = 0

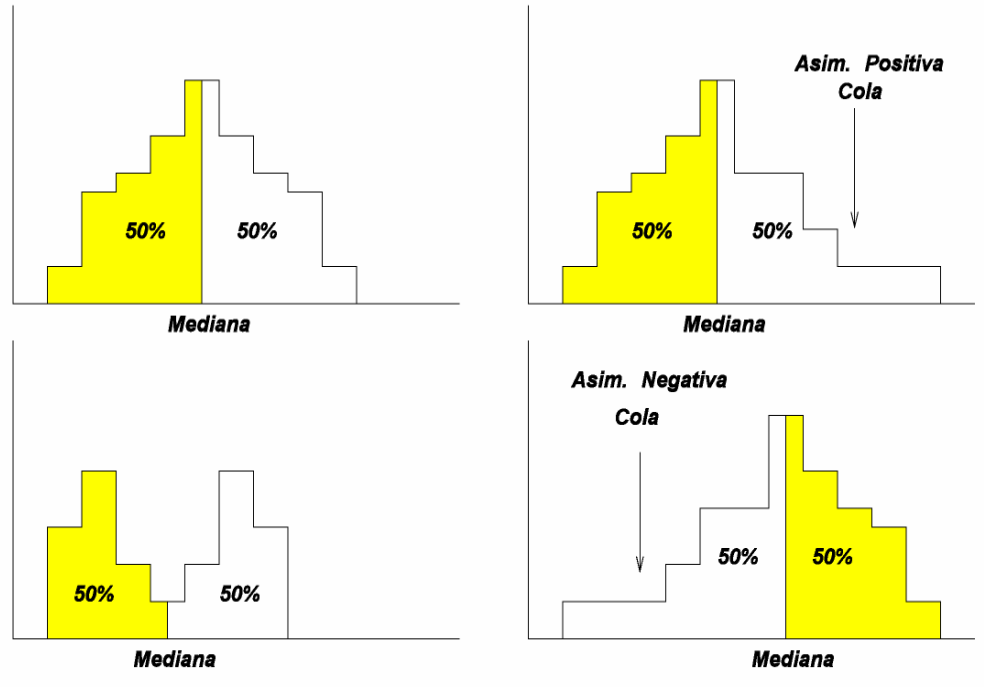
Leptocúrtica (apuntada): curtosis > 0

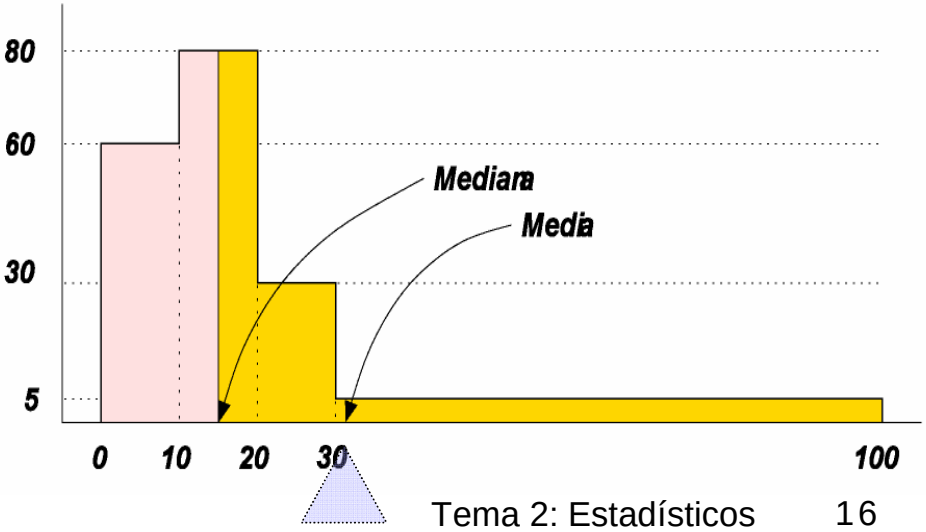
Son de especial interés las mesocúrticas y simétricas (parecidas a la normal).



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 las valores extremos (colas).
- Las discrepancias entre las medidas de centralización son indicación de asimetría.





Asimetría o Sesgo

Asimétrica negativa izquierda asimetría (As) < 0

Simétrica (como la normal): asimetría (As) = 0

Asimétrica positiva derecha asimetría (As) > 0