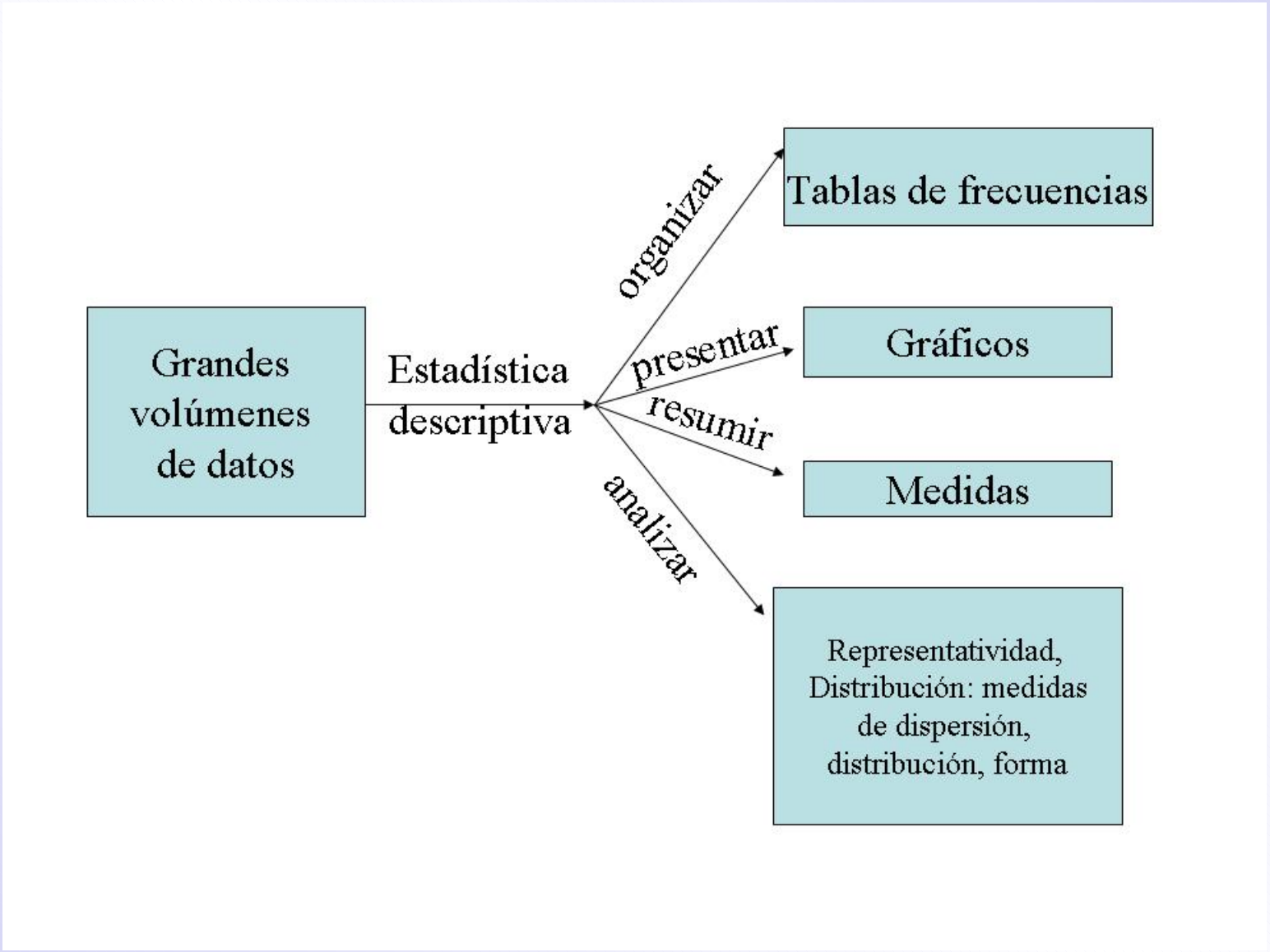
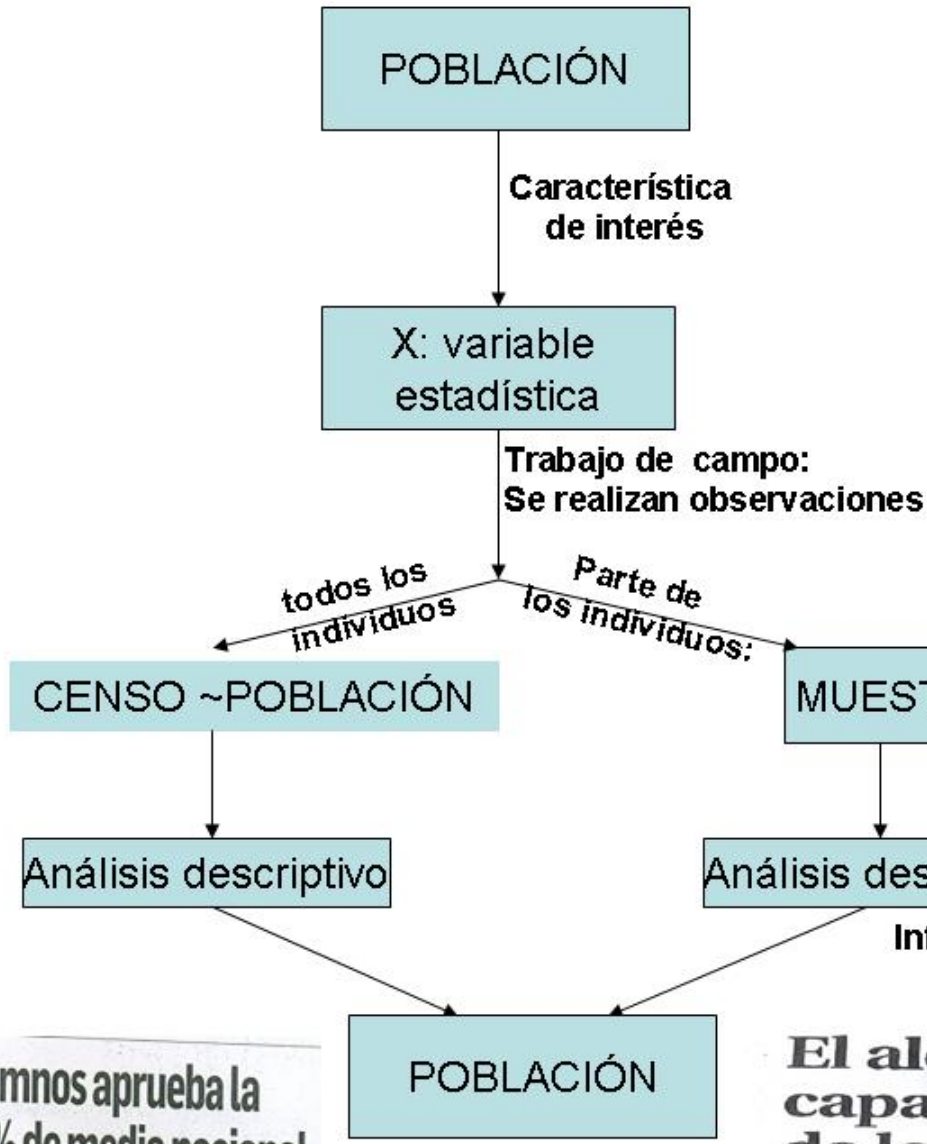
Estadística con R-commander

GRADOS DE INGENIERÍA

Estadística descriptiva





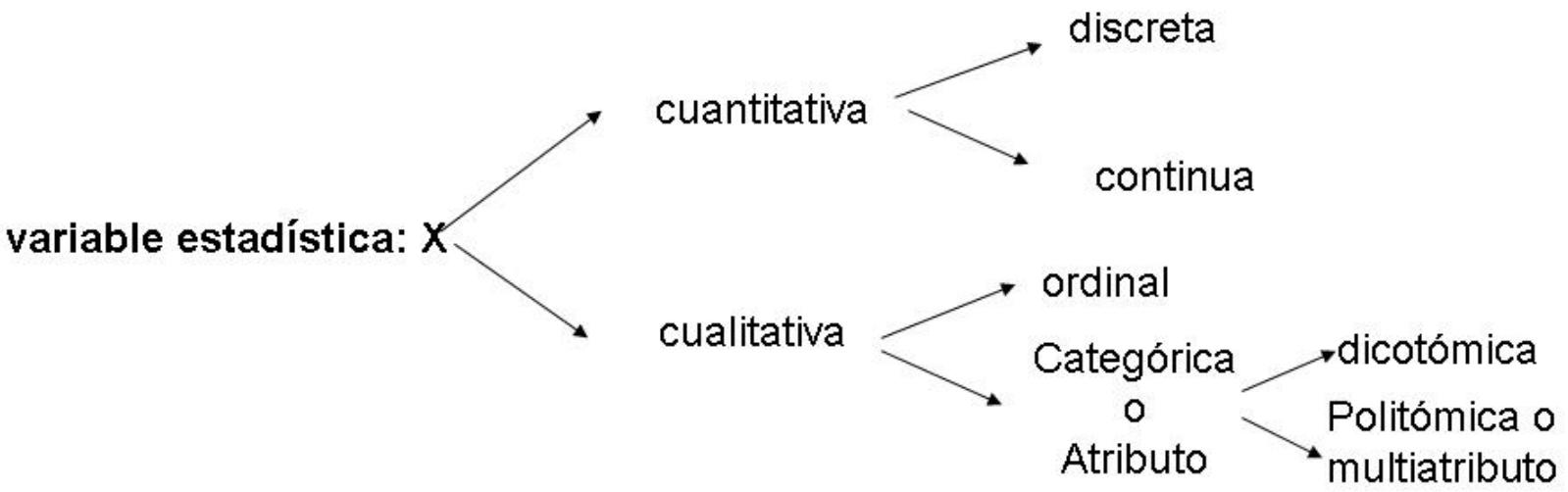
El 86% de los alumnos aprueba la ESO frente al 70% de media nacional

6 de cada 10 escolares terminan el ciclo con todo aprobado

El alcohol daña la capacidad mental de los adolescentes

■ Un estudio certifica que merma la memoria y la capacidad de aprendizaj


Estadística Descriptiva
Muestreo
Inferencia estadística
Probabilidad


variable estadística: X
cuantitativa
discreta
continua
cualitativa
ordinal
Categórica
o
Atributo
dicotómica
Politómica o
multiatributo

Análisis en una dimensión: Tipos de variables

Población: Centros de enseñanza secundaria de navarra.

Centro	Nº Alumnos	Nº profesores	Nota media	Tipo	Lengua	Zona
1	1250	42	6.28	1	1	Z1
2	528	26	5.72	2	2	Z3
3	950	35	6.44	1	2	Z2
4	1328	50	5.08	2	1	Z2
...	...	...	...	...	...	...

Tipo { 1: público; 2: privado }

Lengua { 1: Castellano; 2: Euskera }

Zona { Z1: Sur; Z2: Media; Z3: Norte }

Análisis de variables cualitativas

1. **Tablas de frecuencias. Modalidades, frecuencias absolutas y relativas.**

2. **Gráficos: diagrama de sectores y diagrama de barras.**

Análisis de variables cualitativas

- **Distribuciones de frecuencias**

Modalidades	**Frecuencias**	**Frec. relativas**
Modalidad 1	n_1	$f_1 = n_1/N$
Modalidad 2	n_2	$f_2 = n_2/N$
$\ldots$	$\ldots$	$\ldots$
Modalidad k	n_k	$f_k = n_k/N$
Total	N	**1**

Distribución de frecuencias de la variable: Principal confesión religiosa.

Confes. Religiosa	Frecuencias	Frec. relativas
Musulmán	5	5/30=0,167
Católicos	8	8/30=0,267
Cristianos no católicos	11	11/30=0,367
Otras religiones	6	6/30=0,200
Total	30	1

- Menú *distribuciones de frecuencias*
- Gráficos: diagrama de barras y de sectores
- Menú *Gráficas* en R

Análisis de variables cuantitativas

1. Tablas de frecuencias. Valores (discretas) o intervalos (continuas)

2. Gráficos: Histogramas y diagramas de barras (discretas)

3. Estadísticos:

   - Percentiles (medidas de posición)
   - Medidas de tendencia central: media y mediana (y moda)
   - Medidas de dispersión: Rango intercuártílico, varianza (desviación típica) y coeficiente de variación.

Variables discretas

- **Distribuciones de frecuencias**

Valores	**Frecuencias**	**Frec. relativas**	**Frec. Acumuladas**	**Frec. Acum. relativas**
x_1	n_1	$f_1 = n_1/N$	N_1	$F_1 = N_1/N$
x_2	n_2	$f_2 = n_2/N$	$N_2 = n_1 + n_2$	$F_2 = N_2/N$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_k	$f_k = n_k/N$	$N_k = n_1 + n_2 + \cdots n_k = N$	$F_k = N_k/N = 1$
Total	N	**1**		

Ejemplo: Centros de enseñanza secundaria

Variable: Número de aulas de ordenadores.

N. aulas	Frecuencias	Frec. relativas	Frec. Acumuladas	Frec. Acum. relativas
1	6	0,086=6/70	6	0,086
2	12	0,171=12/70	18	0,257
3	25	0,357=25/70	43	0,614
4	18	0,257=18/70	61	0,871
5	9	0,129=9/70	70	1
Total	70	1		

- El procedimiento frecuencias en el menú resúmenes
- Gráficos: diagrama de barras (en R histograma)

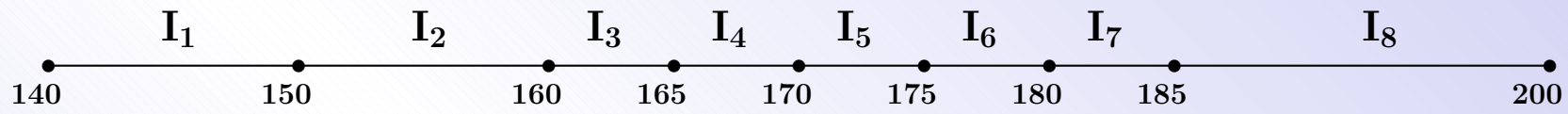
Variables continuas

- **Distribuciones de frecuencias**

Se divide el recorrido de la variable en tramos o intervalos,

y se hace el conteo de frecuencias en cada intervalo, es decir, número de elementos en la población con valor dentro de cada intervalo.

Ejemplo, intervalos para la variable Estatura(cm).



Estatura	Frecuencias
140–150	2
150–160	6
160–165	14
165–170	22
170–175	16
175–180	12
180–185	5
185–200	3
Total	80

Tabla de frecuencias:

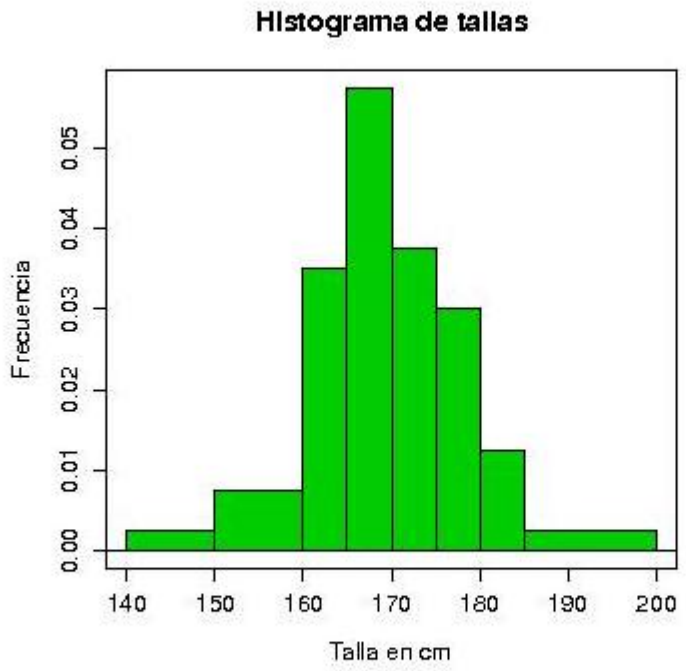
Intervalos	**Frecuencias**	**Frec. relativas**	**Frec. Acumuladas**	**Frec. Acum. relativas**
I_1	n_1	$f_1 = n_1/N$	N_1	$F_1 = N_1/N$
I_2	n_2	$f_2 = n_2/N$	$N_2 = n_1 + n_2$	$F_2 = N_2/N$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
I_k	n_k	$f_k = n_k/N$	$N_k = n_1 + n_2 + \cdots n_k = N$	$F_k = N_k/N = 1$
Total	N	**1**		

- **Gráficos: histograma**
- **Menú gráficas/histograma**

Ejemplo para la variable Estatura:

Estatura	Frecuencias	Frec. relativas	Frec. Acumuladas	Frec. Acum. relativas
140–150	2	0,025	2	0,025
150–160	6	0,075	8	0,100
160–165	14	0,175	22	0,275
165–170	22	0,275	44	0,550
170–175	16	0,200	60	0,750
175–180	12	0,150	72	0,900
180–185	5	0,063	77	0,963
185–200	3	0,038	80	1'000
Total	80	1		

- **Gráficos: histograma**



Medidas de posición

- **Percentiles (Cuartiles)**
- **El diagrama de caja y bigotes**

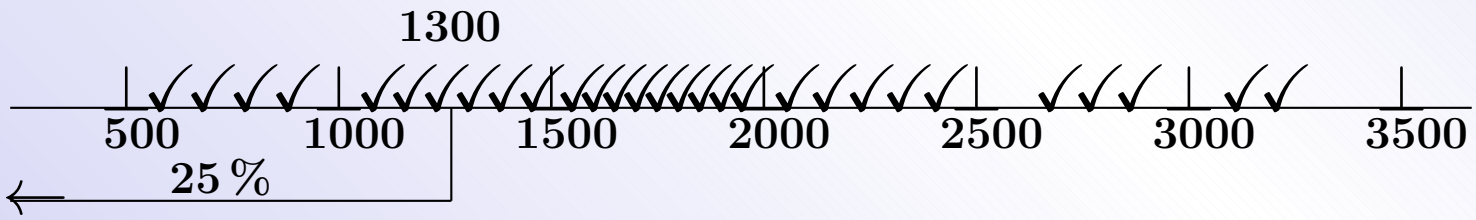
Se dispone de los datos de salario mensual en euros de un conjunto de trabajadores.

Salario de 29 trabajadores

600	700	800	900	1100
1200	1250	1300	1350	1400
1575	1600	1650	1700	1725
1750	1800	1850	1950	2100
2200	2300	2400	2450	2700
2800	2900	3100	3200	

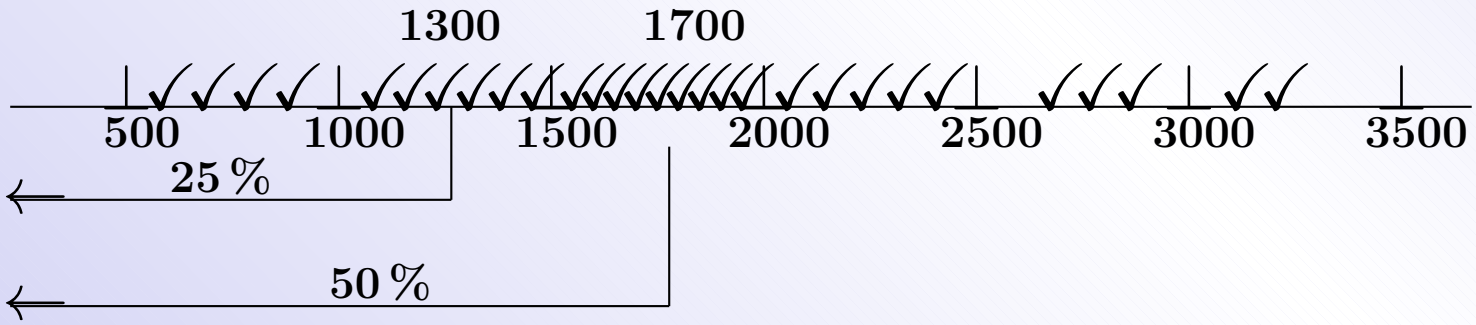
Percentiles

Percentiles



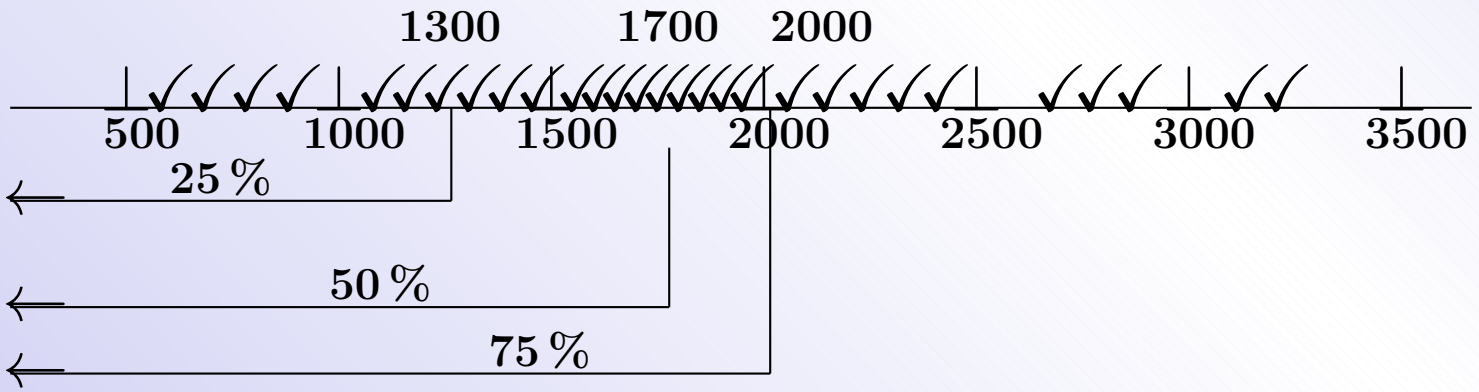
$\mathbf{P_{25} = 1300}$

Percentiles



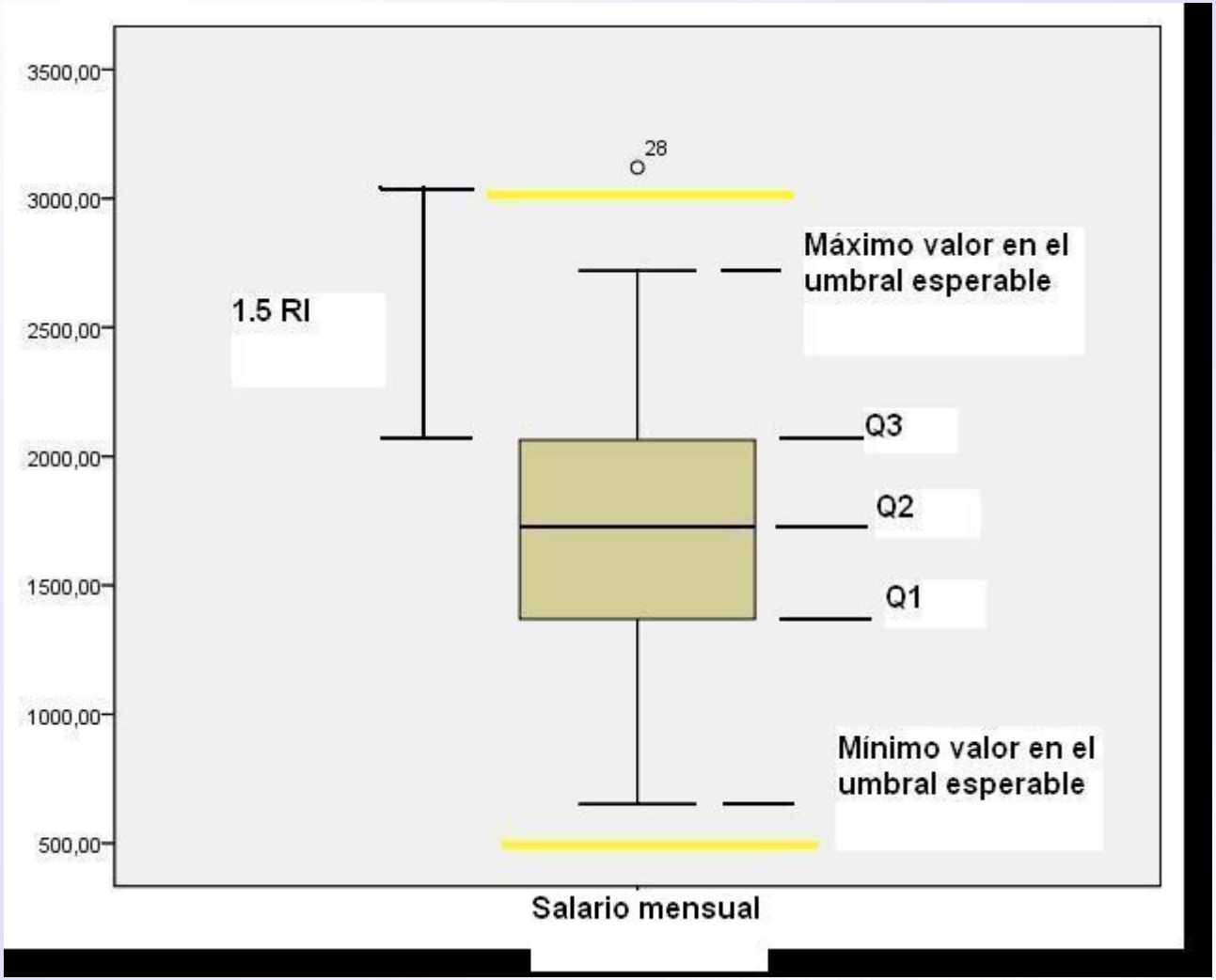
$\mathbf{P_{25} = 1300,\ P_{50} = 1700}$

Percentiles



$\mathbf{P_{25} = 1300,\ P_{50} = 1700,\ \ P_{75} = 2000, P_{10} = 950,\ P_{80} = 2150,\ \ P_{90} = 2500}$

Diagrama de cajas y bigotes.



Resumen numérico de aspectos importantes de una variable cuantitativa

Hay tres aspectos importantes que pueden describirse mediante valores numéricos en un conjunto de datos de una variable cuantitativa:

Centro de los datos: medidas de centralización

Grado de dispersión (o, por contra, concentración) de los datos: medidas de dispersión

La forma de distribución: medidas de forma

Medidas de posición

- **Percentiles (cuartiles)**
- **El diagrama de caja y bigotes**

Medidas de centralización

- **Media** $\overline{x}$
- **Mediana** Me
- **Moda** Mo

Centro de los datos.

Medidas de centralización: Media, moda y mediana

- **Media=** $\dfrac{\text{Suma de todos los datos}}{\text{El número de elementos}}$ $\qquad\qquad \overline{x} = \dfrac{x_1+x_2+\cdots x_N}{N}$

- **Mediana, es el punto de la recta real que deja al 50 % de los valores por debajo de él y al otro 50 % por encima de él. En el histograma es el valor de la horizontal que divide al histograma en dos partes de áreas iguales.**

La mediana es un estadístico robusto: muy poco sensible a datos extremos. La media en cambio es muy sensible a este tipo de datos. Es habitual calcular la media recortada, por ejemplo al 5 %, para evitar el arrastre de datos extremos.

• **Moda, es el valor más repetido. En las variables discretas de pocos valores puede ser una buena medida para representar el centro de los datos, pero en los demás casos no lo suele ser.**

Fórmulas para la media.

- **Datos no agrupados:** $\overline{x} = \dfrac{x_1 + x_2 + \cdots x_N}{N}$

- **Datos agrupados en frecuencias:**

$$\overline{x} = \frac{x_1\, n_1 + x_2\, n_2 + \cdots + x_k\, n_k}{N}$$

Valores	**Frecuencias**
x_1	n_1
x_2	n_2
$\cdots$	$\cdots$
x_k	n_k
	N

• **Datos de intervalos agrupados en frecuencias: x es la marca de clase, el punto medio del intervalo.**

$$\overline{x} = \frac{x_1\, n_1 + x_2\, n_2 + \cdots + x_k\, n_k}{N}$$

Intervalos	**Marca de clase**	**Frecuencias**
I_1	x_1	n_1
I_2	x_2	n_2
$\ldots$	$\ldots$	$\ldots$
I_k	x_k	n_k
		N

Medidas de dispersión.

- **Medidas de dispersión**
  - **Rango**
  - **Rango intercuartílico** RI
  - **basadas en dist. al centro**
    - **Desv. media mediana** DM_{Me}
    - **Desv. media media** $DM_{\overline{x}}$
    - **Varianza** S^2
    - **Desv. típica** S
    - **Coeficiente de variación** CV

Medidas basadas en distancias hasta la media.

Se calcula para cada valor la distancia hasta la media del conjunto de datos,

$$\mathbf{x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_N - \overline{x},}$$

y a continuación se promedian.

Si el promedio de distancias es grande, los datos están poco concentrados alrededor de la media. Si el promedio es pequeño, los datos están bastante concentrados en la media.

No se pueden promediar así

$$\frac{(\mathbf{x_1 - \overline{x}}) + (\mathbf{x_2 - \overline{x}}) + \cdots + (\mathbf{x_N - \overline{x}})}{\mathbf{N}}$$

porque por definición de $\overline{x}$, este promedio siempre vale cero. Se compensan distancias positivas con negativas.

Para evitar el signo lo que se hace es elevar al cuadrado las distancias.

$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}}$$

A este promedio se le llama varianza y se le denota por S^2.

Tiene la ventaja de eliminar el signo y de exagerar distancias pequeñas y grandes.

Para el cálculo suele utilizarse esta fórmula:

$$\mathbf{S^2 = \frac{x_1^2 + x_2^2 + \cdots + x_N^2}{N} - \overline{x}^2}$$

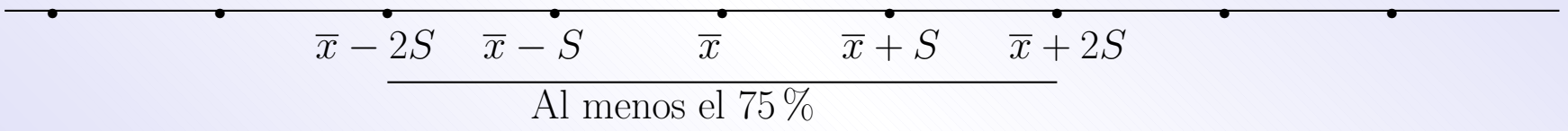
Su raíz se denomina desviación típica y se denota por S. S = $\sqrt{S^2}$

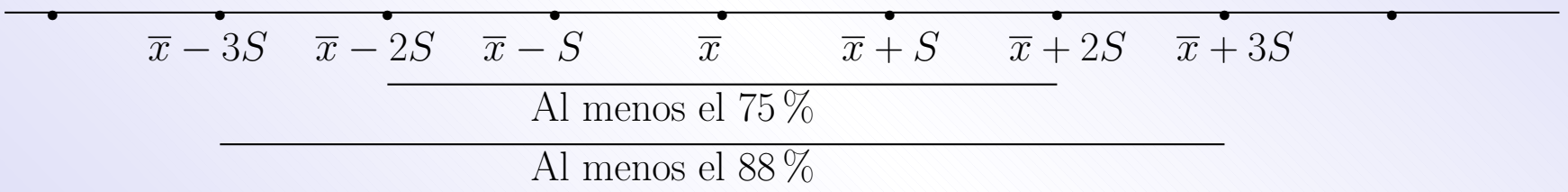
Aunque la desviación devuelve a la unidad de medida en que se esté trabajando, la desviación es mejor desligarla de la unidad de medida, es mejor hallar un coeficiente. Si un conjunto de datos de estaturas tiene una desviación de 120(en cm) tendrá una desviación de 1,2 (en m) pero dispersión tienen igual, es el mismo conjunto de datos.

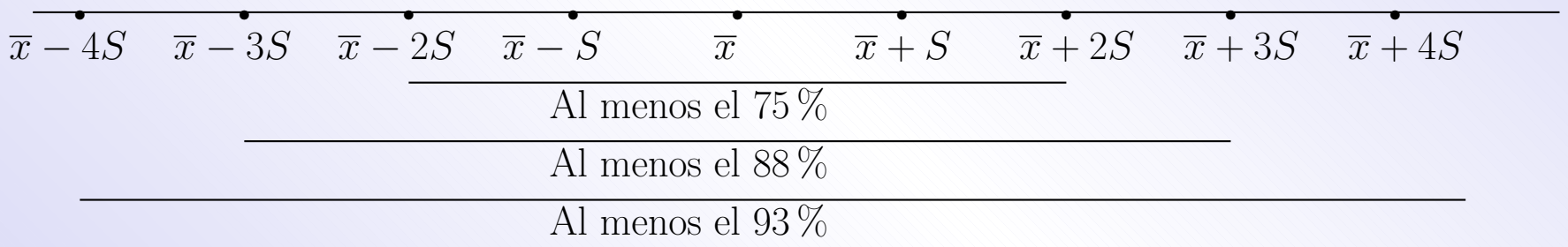
Por otro lado, debería de tenerse en cuenta el centro del conjunto de datos, no es lo mismo una desviación de 10 en un grupo de datos de media 15 que en otro de media 100.

¿Cuánto es de grande o pequeña una desviación de 3? ¿Será un conjunto de datos homogéneo o disperso?

La desigualdad de Chebychev, permite asegurar que:


$\overline{x} - 2S$
$\overline{x} - S$
$\overline{x}$
$\overline{x} + S$
$\overline{x} + 2S$
Al menos el 75 %


$\overline{x}-3S$
$\overline{x}-2S$
$\overline{x}-S$
$\overline{x}$
$\overline{x}+S$
$\overline{x}+2S$
$\overline{x}+3S$
Al menos el 75 %
Al menos el 88 %


$\overline{x}-4S$
$\overline{x}-3S$
$\overline{x}-2S$
$\overline{x}-S$
$\overline{x}$
$\overline{x}+S$
$\overline{x}+2S$
$\overline{x}+3S$
$\overline{x}+4S$
Al menos el 75 %
Al menos el 88 %
Al menos el 93 %

¿Cuánto es de grande o pequeña una desviación de 3? ¿Será un conjunto de datos homogéneo o disperso?

Los datos son pesos de 100 varones.

Coeficiente de variación:

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

Este coeficiente arregla los dos aspectos, se desliga de la unidad de medida y enfrenta el valor de la desviación al de la media.

Fórmulas

$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N} = \frac{x_1^2 + x_2^2 + \cdots + x_N^2}{N} - \overline{x}^2}$$

Si los datos están agrupados en frecuencias:

$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2\, n_1 + (x_2 - \overline{x})^2\, n_2 + \cdots + (x_k - \overline{x})^2\, n_k}{N} = \frac{x_1^2\, n_1 + x_2^2\, n_2 + \cdots + x_k^2\, n_k}{N} - \overline{x}^2}$$

La cuasivarianza $\widehat{S}^2$ y cuasidesviación típica $\widehat{S} = \sqrt{\widehat{S}}$.

$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}}$$

$$\mathbf{\widehat{S}^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N - 1}}$$

Fórmulas

$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N} = \frac{x_1^2 + x_2^2 + \cdots + x_N^2}{N} - \overline{x}^2}$$

Si los datos están agrupados en frecuencias:

$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2\, n_1 + (x_2 - \overline{x})^2\, n_2 + \cdots + (x_k - \overline{x})^2\, n_k}{N} = \frac{x_1^2\, n_1 + x_2^2\, n_2 + \cdots + x_k^2\, n_k}{N} - \overline{x}^2}$$

La cuasivarianza $\widehat{S}^2$ y cuasidesviación típica $\widehat{S} = \sqrt{\widehat{S}}$.

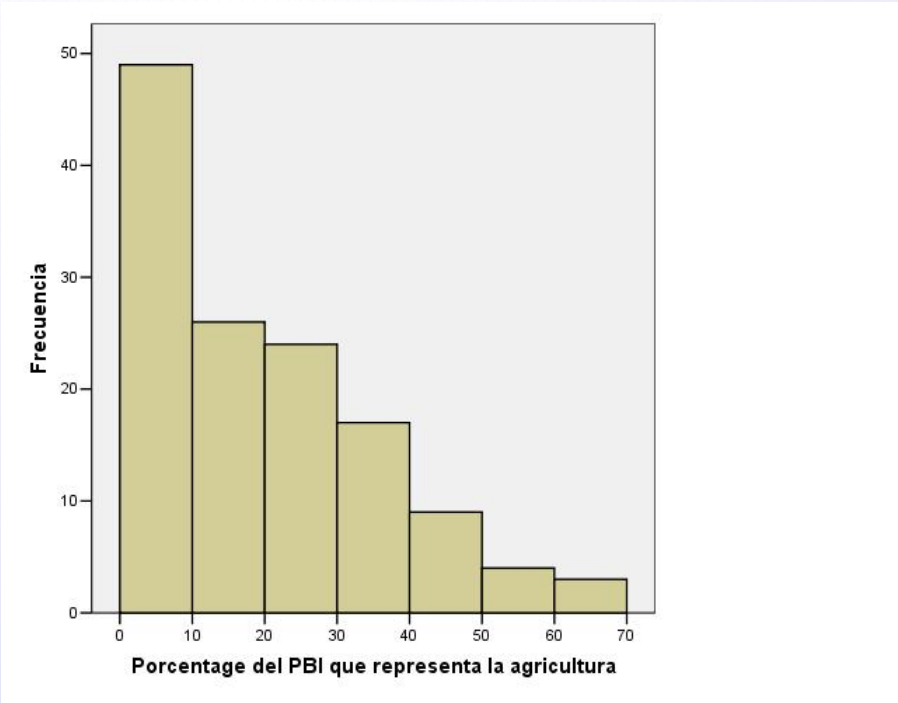
$$\mathbf{S^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N}}$$

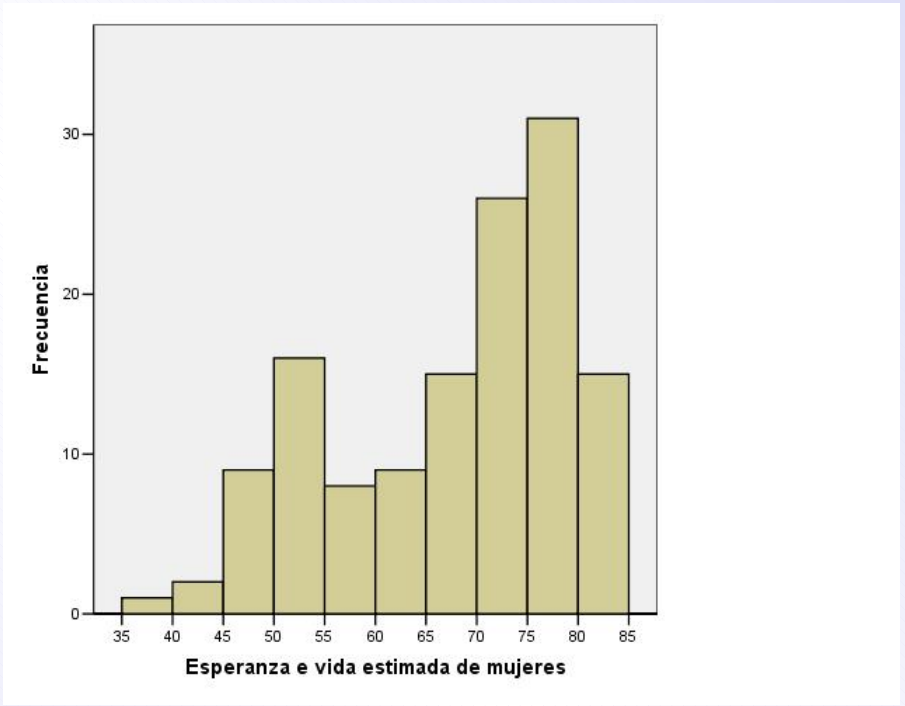
$$\mathbf{\widehat{S}^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N - 1}}$$

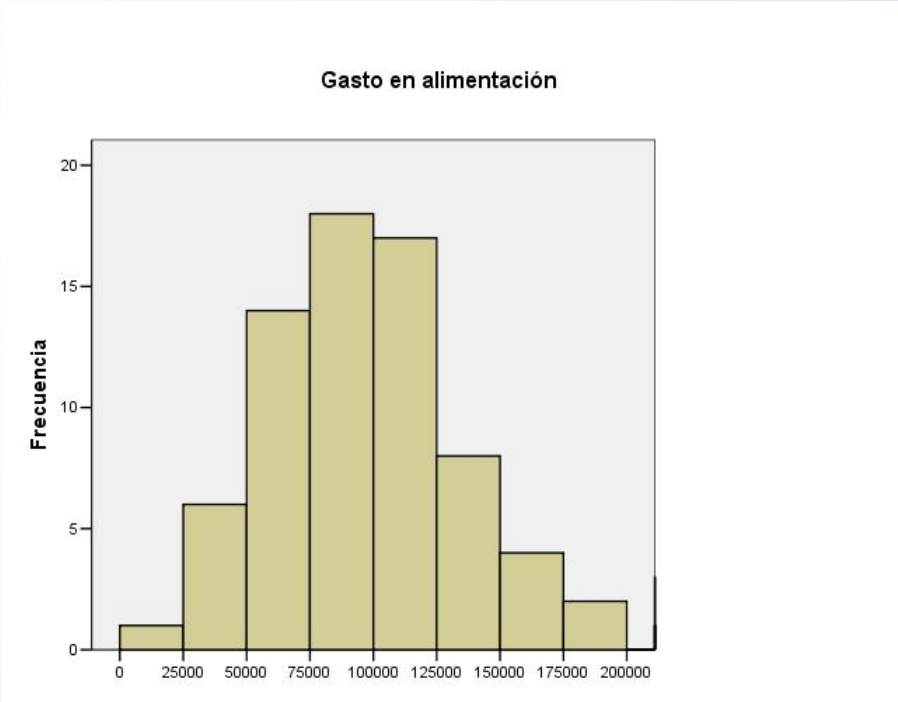
Medidas de Forma.

Medidas de forma
- Coeficiente de Asimetría
- Coeficiente de curtosis

Coeficiente de asimetría.




Frecuencia
30
20
10
0
35
40
45
50
55
60
65
70
75
80
85
Esperanza e vida estimada de mujeres



Se basa en la distancia hasta la media del conjunto de datos,

$$\mathbf{x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_N - \overline{x}.}$$

El coeficiente es:

$$\mathbf{g_1 = \frac{(x_1 - \overline{x})^3 + (x_2 - \overline{x})^3 + \cdots + (x_N - \overline{x})^3}{N\, S^3}}$$

Si $g_1 < 0$ la distribución es asimétrica hacia la izquierda.

Si $g_1 = 0$ la distribución es simétrica.

Si $g_1 > 0$ la distribución es asimétrica hacia la derecha.

Coeficiente de curtosis.

Se trata de comparar el apuntamiento del histograma respecto del considerado normal, el de la distribución Normal. También se basa en la distancia hasta la media del conjunto de datos,

$$\mathbf{x_1 - \overline{x}, x_2 - \overline{x}, \ldots, x_N - \overline{x}.}$$

El coeficiente es:

$$\mathbf{g_2 = \frac{(x_1 - \overline{x})^4 + (x_2 - \overline{x})^4 + \cdots + (x_N - \overline{x})^4}{N\ S^4} - 3}$$

Si $g_2 < 0$ la distribución es menos apuntada que la normal.

Si $g_2 = 0$ la distribución es igual de apuntada que la normal.

Si $g_2 > 0$ la distribución es más apuntada que la normal.