ESTADÍSTICA

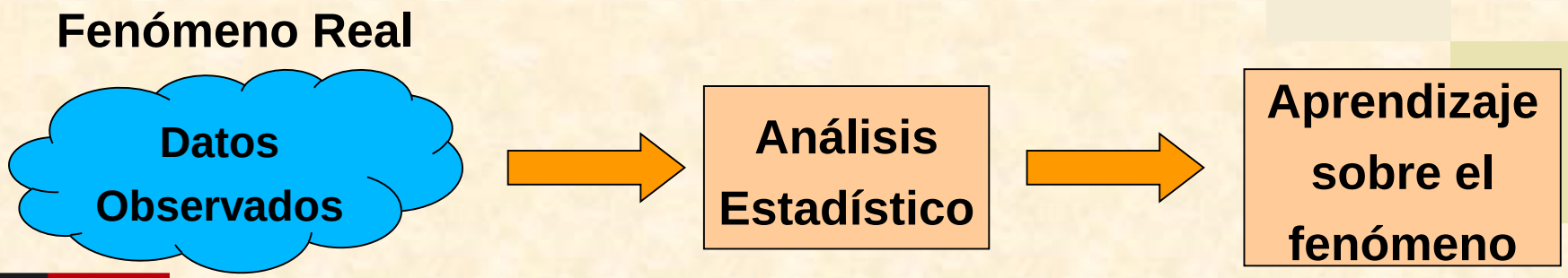
INTRODUCCIÓN

- ¿Qué es la **estadística**?.

  - Es la rama de las matemáticas que estudia la recolección, análisis e interpretación de datos.

- ¿Por qué estudiamos **estadística**?

  - Aprender sobre fenómenos físicos o naturales con el objetivo de obtener conclusiones y por tanto, poder tomar decisiones.



INTRODUCCIÓN

- La estadística se divide en dos grandes áreas:

1. ESTADÍSTICA DESCRIPTIVA

- Visualización/representación y resumen de datos.

2. ESTADÍSTICA INFERENCIAL.

- Creación de modelos en base a los datos observados con el objetivo de poder hacer predicciones.

ESTADÍSTICA DESCRIPTIVA UNIVARIADA

ESTADÍSTICA DESCRIPTIVA UNIVARIADA

ÍNDICE

- DEFINICIONES.
- REPRESENTACIÓN DE DATOS.
  - Tabla de frecuencias.
  - Representaciones gráficas.
- MEDIDAS DE LOS DATOS.
  - Centralización y posición.
  - Dispersión.
  - Forma.

DEFINICIONES

Población: conjunto de seres, medidas u objetos acerca de los que se desea tener información.

Elemento/Individuo: cada uno de los miembros de la población.

Muestra: subconjunto de individuos de la población.

DEFINICIONES

VARIABLE ESTADÍSTICA

Característica que se mide/observa en los individuos de una población.

Ejemplo: la altura, la edad, el peso, el sexo, número de hermanos, etc

Cualitativa: Describe una cualidad.

Ejemplo: color de ojos, preferencias, etc.

Cuantitativa: Toma valores numéricos.

Ejemplo: altura, peso, etc.

Discreta: conjunto numerable (enteros)

Continua: Valores en un intervalo (toma decimales)

REPRESENTACIÓN DE LOS DATOS

OBJETIVO

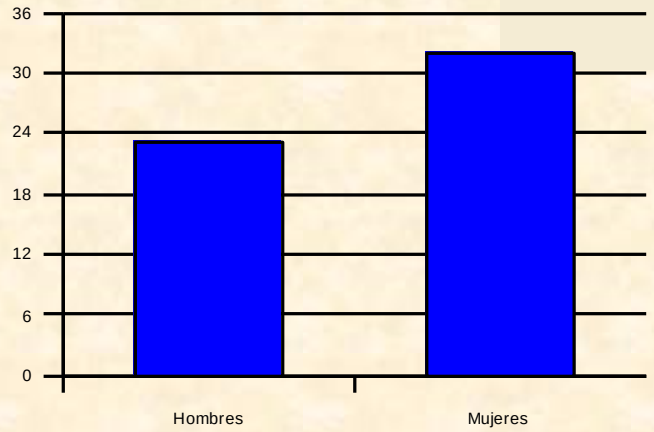
Resumir información contenida en los datos para facilitar su análisis

- Hay dos maneras equivalentes de presentar la información contenida en un conjunto de datos:

TABLAS DE FRECUENCIAS

Sexo	Frecuencia
Hombres	23
Mujeres	32

REPRESENTACIONES GRÁFICAS



REPRESENTACIÓN DE LOS DATOS

TABLA DE FRECUENCIA

Muestra la frecuencia de cada valor observado.

- **Cuando los Datos son cualitativos o cuantitativos discretos con pocos valores distintos** → Damos la frecuencia con que aparece en la muestra para cada valor o categoría.

- **Datos cuantitativos continuos o discretos con muchos valores distintos** → Damos la frecuencia con los datos agrupados en clases o intervalos.

TABLA DE FRECUENCIA

- **Frecuencias absolutas (F_i):** Contabilizan el número total de elemento de cada clase.

- **Frecuencias relativas (f_i):** Es la proporción (porcentaje) de individuos de cada tipo que pertenecen a cada clase sobre el total de individuos de la muestra. Se obtiene dividiendo la frecuencia absoluta entre el total de individuos.

- **Frecuencias acumuladas (F_{ac} y f_{ac}):** Se obtienen sumando las frecuencias de las clases anteriores.

REPRESENTACIÓN DE LOS DATOS

TABLA DE FRECUENCIA

EJEMPLO 1

Se ha contado el número de hijos de 100 matrimonios que llevan casados más de 15 años. Obteniendo los siguientes resultados:

0	0	1	1	2	0	3	0	2	4
2	1	0	5	5	2	2	3	1	1
1	2	2	4	5	0	3	2	2	2
2	4	3	1	1	1	0	0	2	3
1	4	0	0	1	1	2	2	3	2
3	1	1	0	0	1	2	0	2	2
0	0	0	0	1	1	4	3	3	2
1	6	3	1	3	2	1	2	3	0
1	3	0	2	3	2	1	3	4	0
6	2	1	3	0	3	1	0	2	2

REPRESENTACIÓN DE LOS DATOS

TABLA DE FRECUENCIA

EJEMPLO 1

Nº de hijos	F_i	f_i	F_{ac}	f_{ac}
0	22	0.22	22	0.22
1	24	0.24	46	0.46
2	26	0.26	72	0.72
3	17	0.17	89	0.89
4	6	0.06	95	0.95
5	3	0.03	98	0.98
6	2	0.02	100	1

100 1

REPRESENTACIÓN DE LOS DATOS

TABLA DE FRECUENCIA

Hay muchos valores → **Se agrupan en clase o intervalos**

- **¿Cuántas clases elegir?**
  - **Pocas** → Se pierde mucha información de los datos.
  - **Muchas.** → La frecuencia resultante en cada una puede ser pequeña y poco útil para el estudio

- **¿Qué longitud elegir para cada clase?**

  Se suelen elegir intervalos de igual longitud

REPRESENTACIÓN DE LOS DATOS

TABLA DE FRECUENCIA

EJEMPLO 2

Los siguientes datos muestran los niveles de colesterol en la sangre de 40 estudiantes de primer año de una universidad.

213	173	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

REPRESENTACIÓN DE LOS DATOS

TABLA DE FRECUENCIA

EJEMPLO 2

Tabla de frecuencias de los niveles de colesterol en sangre

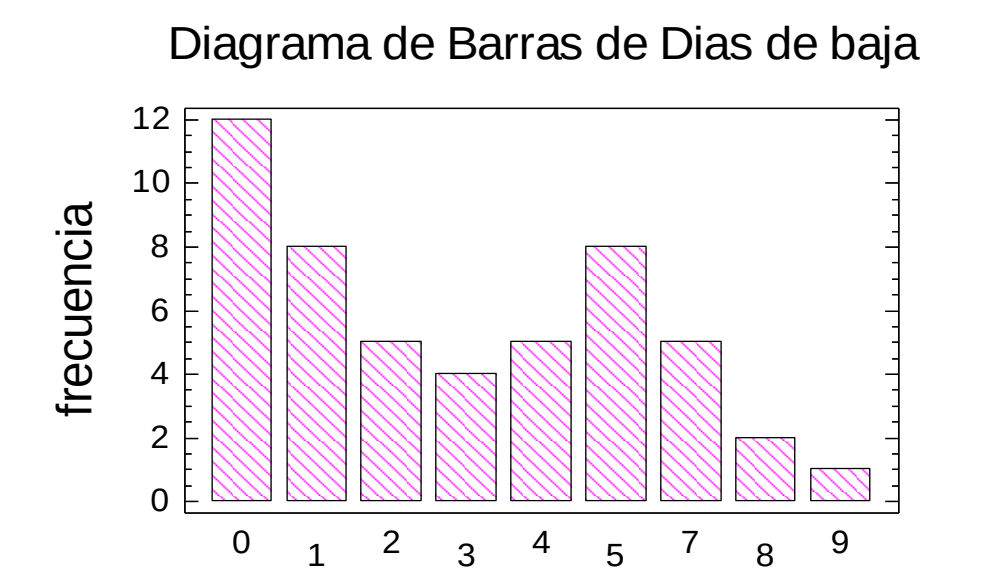
Intervalo de clase	Frecuencia absoluta	Frecuencia relativa
170 – 180	3	0,075
180 – 190	7	0,175
190 – 200	13	0,325
200 – 210	8	0,200
210 - 220	5	0,125
220 – 230	4	0,100

REPRESENTACIÓN DE LOS DATOS

GRÁFICAS

1. Diagrama de Barras

- Gráficos de frecuencias para datos cualitativos.
- Barras separadas para cada valor.
- La altura de las barras representa la frecuencia absoluta o relativa de cada valor.

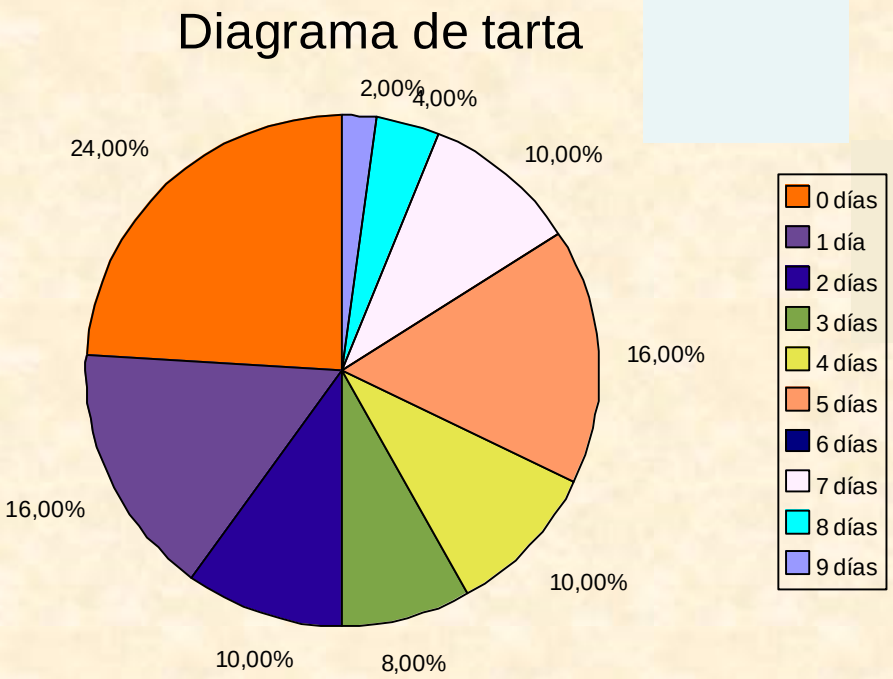


REPRESENTACIÓN DE LOS DATOS

GRÁFICAS

2. Diagrama de tarta o sectores

- Gráficos de frecuencias para datos cualitativos.
- El área de cada sector representa la frecuencia relativa de cada valor.



REPRESENTACIÓN DE LOS DATOS

GRÁFICAS

3. Histograma

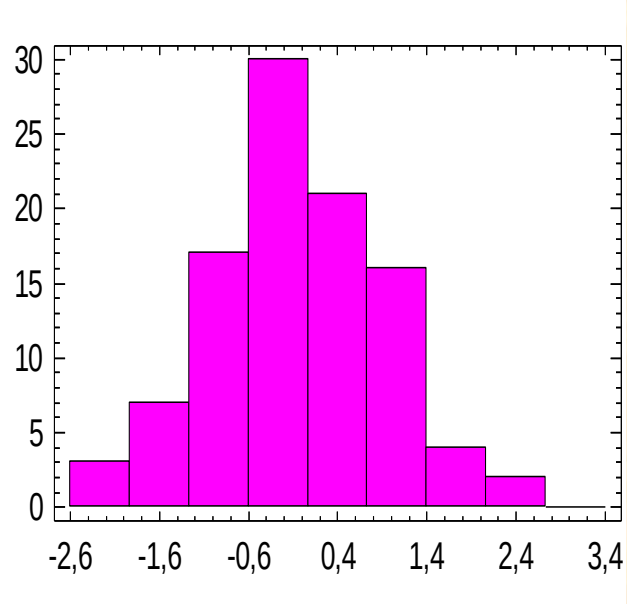
- Gráficos de frecuencias para datos cuantitativos.
- Cada barra representa una **clase**. No hay hueco entre barras.
- Las bases son iguales a la amplitud de cada clase.
- La altura corresponde a la frecuencia absoluta o relativa de la clase.
- **Marca** de clase: Es el valor medio de cada clase.
- El área que hay bajo el histograma es proporcional a la cantidad de individuos del intervalo.

REPRESENTACIÓN DE LOS DATOS

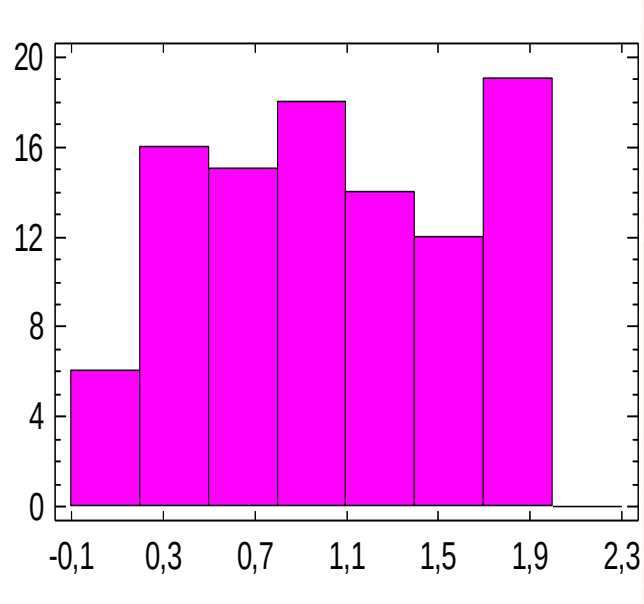
GRÁFICAS

El histograma da información sobre:

- La simetría de los datos y la dispersión de los mismos



Simétricos



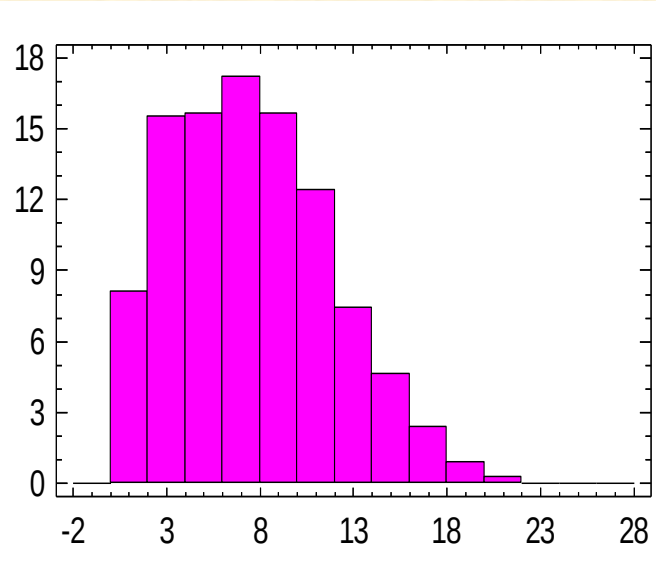
Dispersos

REPRESENTACIÓN DE LOS DATOS

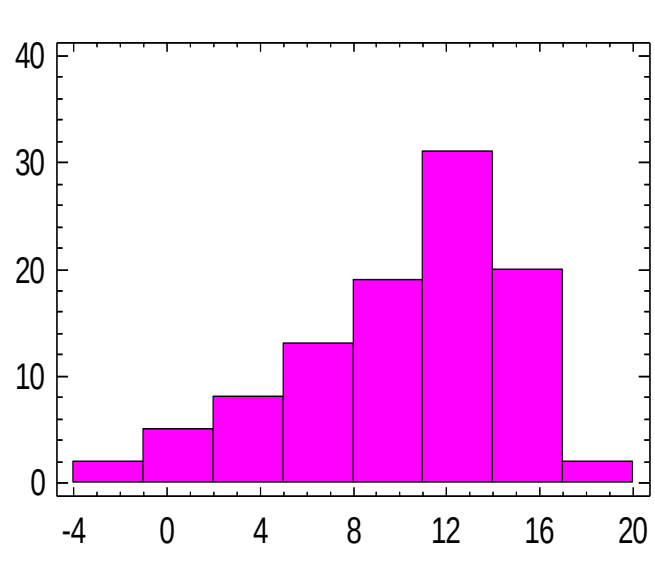
GRÁFICAS

El histograma da información sobre:

- La forma de la distribución



Asimétricos a la derecha



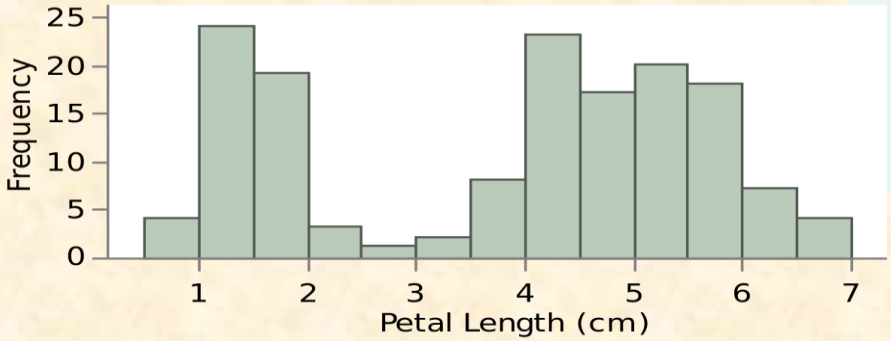
Asimétricos a la izquierda

REPRESENTACIÓN DE LOS DATOS

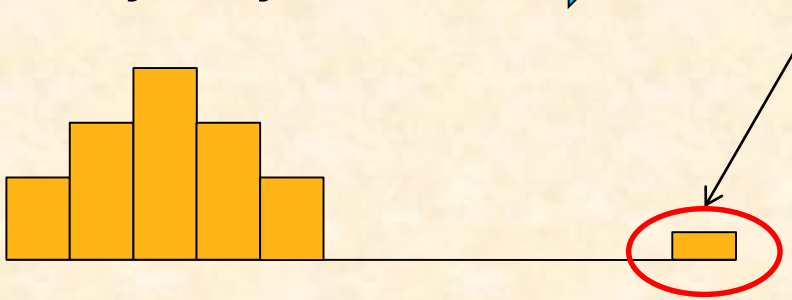
GRÁFICAS

El histograma da información sobre:

- Si existen brechas entre los datos → posibles dos poblaciones.



- Si hay valores muy alejados → valores atípicos.



REPRESENTACIÓN DE LOS DATOS

EJEMPLO 3

La variable representa el peso (en gr.) de 191 monedas de 100 pesetas.

TABLAS DE FRECUENCIAS

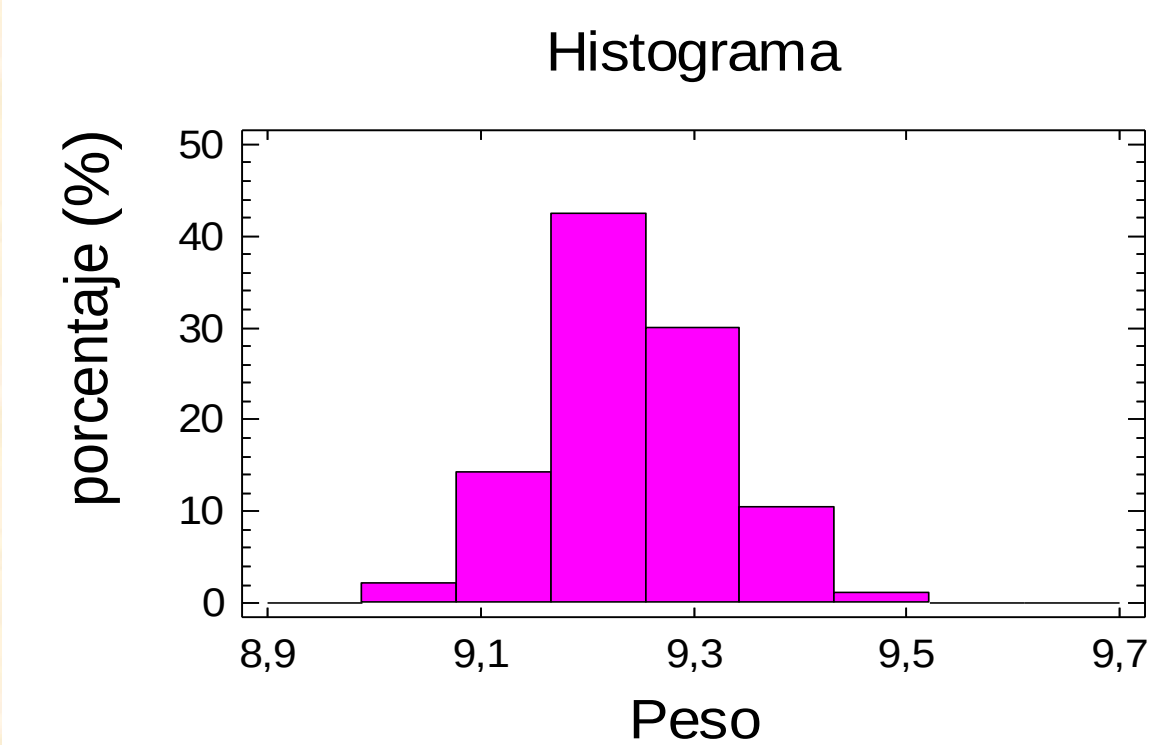
Clase	Límite Inferior	Límite Superior	Marca	Frecuencia	Frecuencia Relativa	Frecuencia Acumulada	Frecuencia Relat. Acumulada
menor que 8,9				0	0	0	0
1	8,9	8,98	8,95	0	0	0	0
2	8,98	9,07	9,03	4	0,02	4	0,02
3	9,07	9,16	9,12	27	0,14	31	0,16
4	9,16	9,25	9,21	81	0,42	112	0,59
5	9,25	9,34	9,3	57	0,3	169	0,88
6	9,34	9,43	9,39	20	0,1	189	0,99
7	9,43	9,52	9,47	2	0,01	191	1
8	9,52	9,61	9,56	0	0	191	1
9	9,61	9,7	9,65	0	0	191	1
mayor que 9,7				0	0	191	1

REPRESENTACIÓN DE LOS DATOS

EJEMPLO 3

HISTOGRAMA

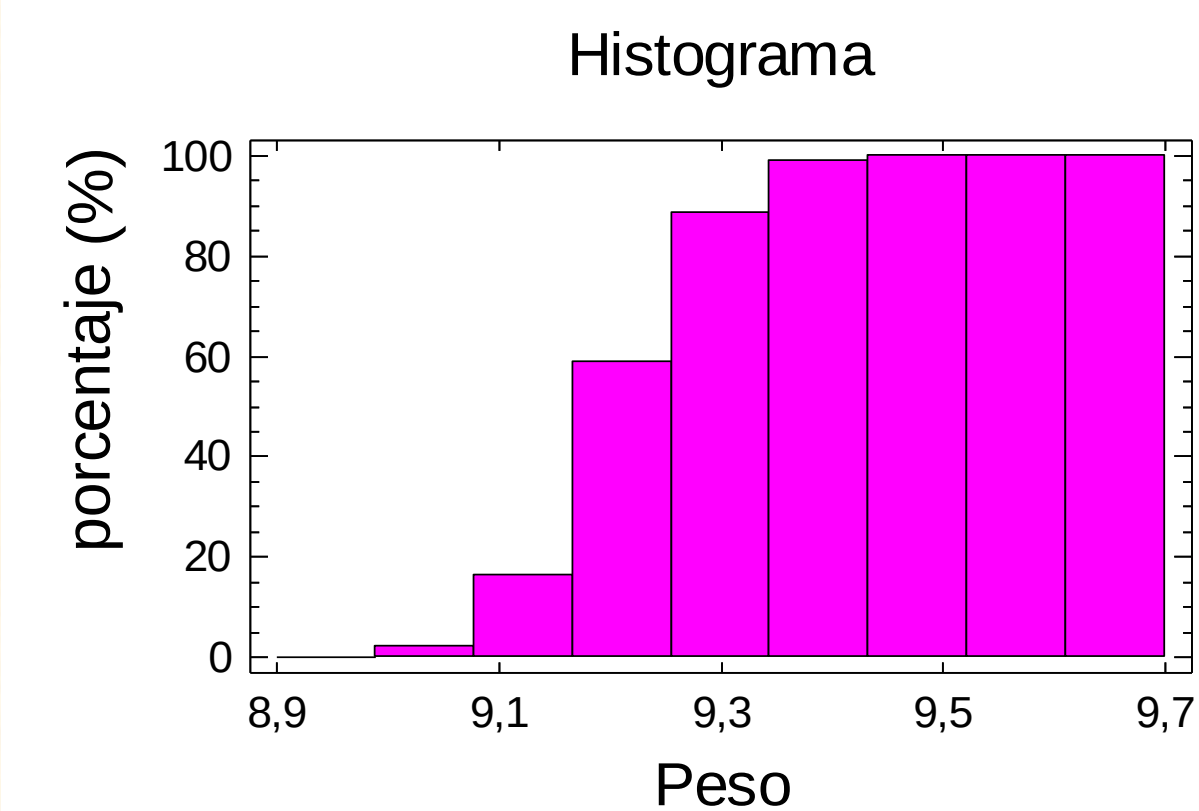
- Representación de las frecuencias relativas:



REPRESENTACIÓN DE LOS DATOS

EJEMPLO 3

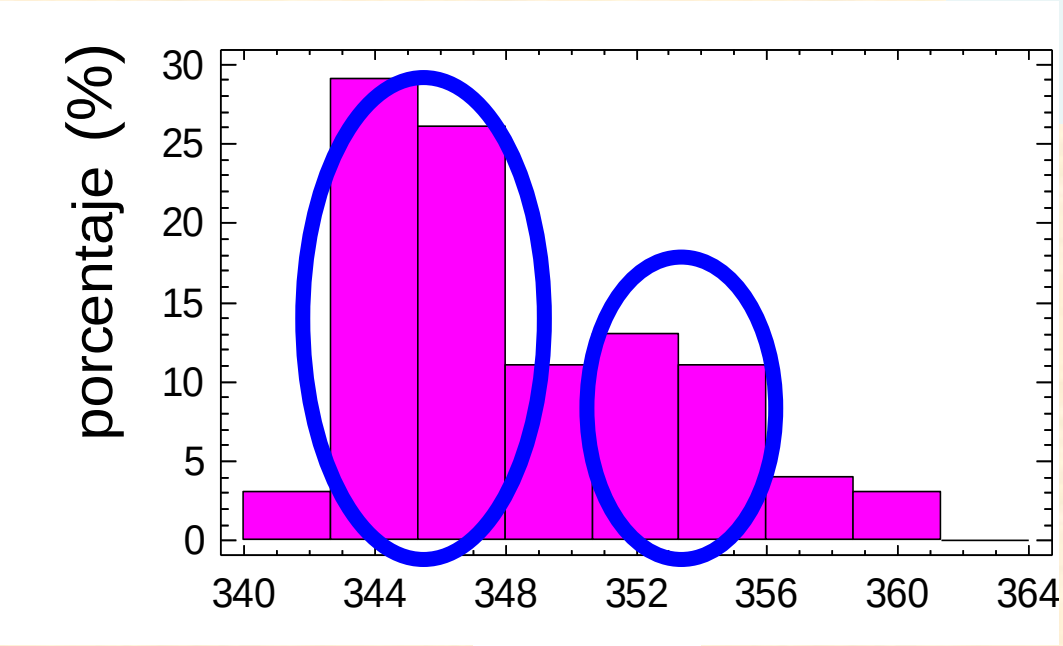
- Representación de las frecuencias relativas acumuladas:



REPRESENTACIÓN DE LOS DATOS

EJEMPLO 4

Datos correspondientes a las longitudes ($\times\ 10^{-4}m$) de 100 clavos del mismo tipo, medidos por dos personas, 50 clavos cada una, que usaron calibres diferentes.

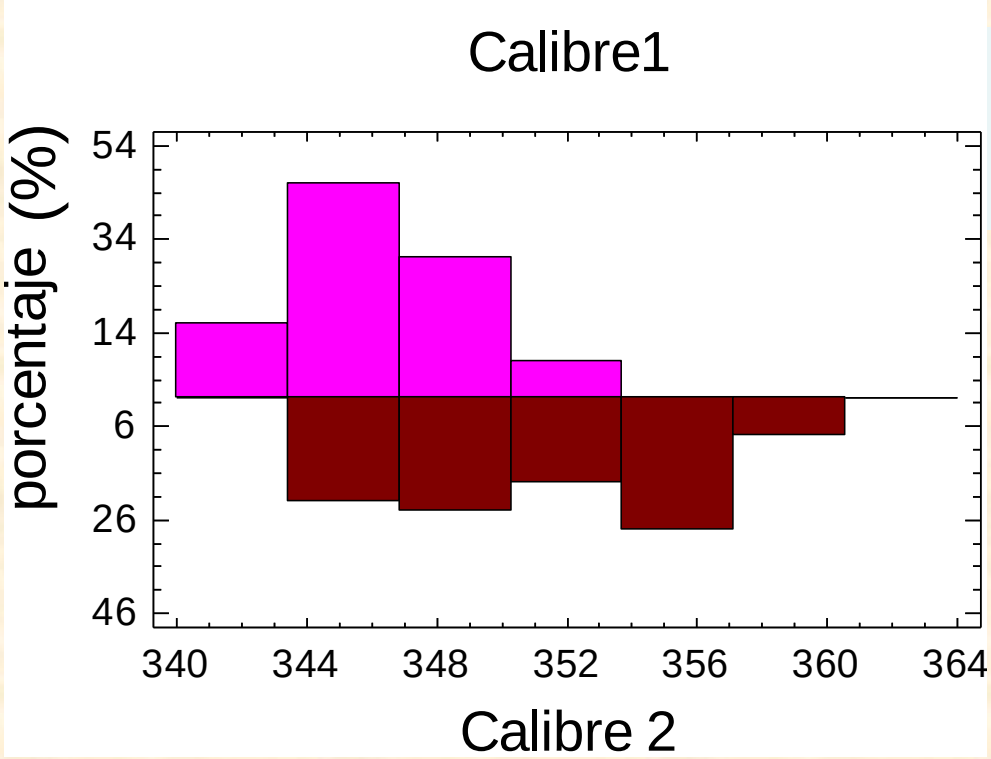


Posible presencia de 2 poblaciones

REPRESENTACIÓN DE LOS DATOS

EJEMPLO 4

Comparación de los histogramas separando los datos según el calibre utilizado:



REPRESENTACIÓN DE LOS DATOS

OTRAS REPRESENTACIONES GRÁFICAS

- **Polígono de frecuencias**
  - Se representa los puntos medios de cada clase (marcas) frente a la frecuencia de la clase correspondiente y se unen estos puntos por líneas rectas.
  - Útiles para comparar conjuntos de datos.
- **Diagrama de tallos y hojas**
  - Para conjunto de datos pequeño o moderado.
  - Los datos se separan en “un tallo” y “hojas”
  - Ventaja: no se pierden los datos.

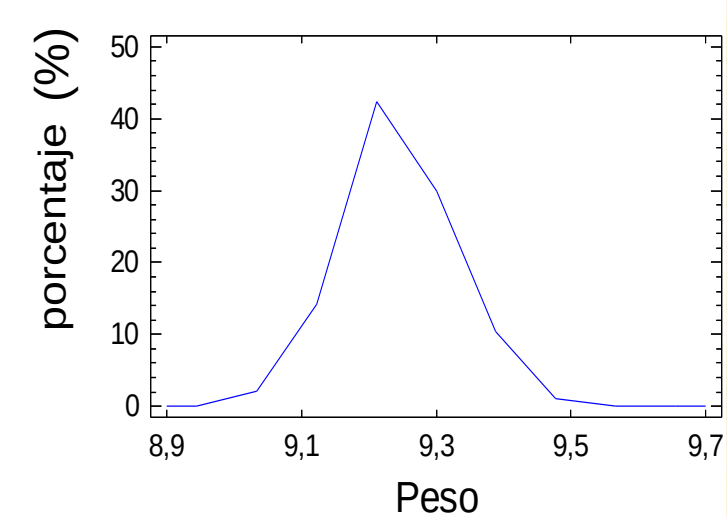
REPRESENTACIÓN DE LOS DATOS

OTRAS REPRESENTACIONES GRÁFICAS

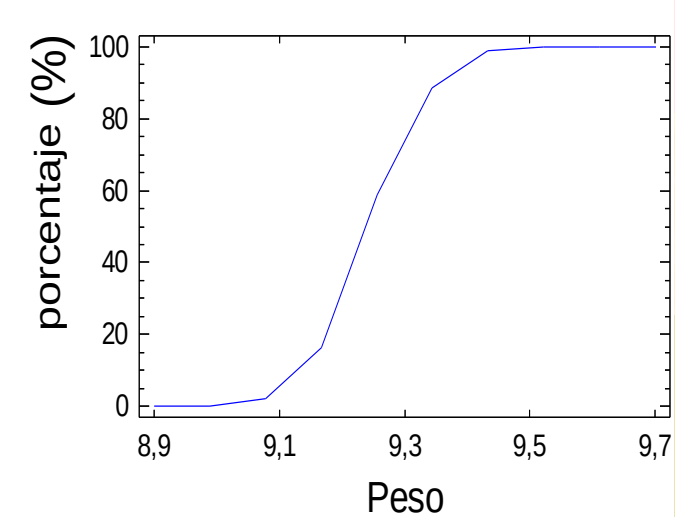
EJEMPLO 5

- La variable representa el peso (en gr.) de 191 monedas de 100 pesetas.

POLÍGONO DE FRECUENCIAS



Polígono de frecuencias relativas



Polígono de frecuencias acumuladas

REPRESENTACIÓN DE LOS DATOS

OTRAS REPRESENTACIONES GRÁFICAS

EJEMPLO 5

DIAGRAMA DE TALLOS Y HOJAS

unidad = 0,01 1|2 representa 0,12

```
89|
90|123
90|9
91|0000111222333334444
91|555566677777777778888888888999999
92|000000001111111122222222222333333333333444444
92|55555556666666667788888888889999999
93|00000001111222223333333444
93|55555667788899
94|001233
94|68
```

MEDIDAS DE UN CONJUNTO DE DATOS

INTRODUCCIÓN

- **Medidas de centralización y posición**
  - Valor que representa a todo el conjunto de datos: **media**, **mediana**, **moda** y **cuantiles**.
- **Medidas de dispersión**
  - Valor que cuantifica cómo están distribuidos los datos con respecto a la media: **varianza** (desviación típica).
- **Medidas de forma**
  - Valores que miden lo simétrica o “apuntada/picuda” que es la distribución de nuestros datos: **coeficiente de asimetría**, **coeficiente de apuntamiento** (curtosis)

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE CENTRALIZACIÓN

MEDIA ARITMÉTICA

- Para datos no agrupados la **media aritmética** de un conjunto de datos $x_1, x_2, x_3, \ldots x_n$ es

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

- Para datos agrupados en tablas de frecuencias:

$$\bar{x} = \frac{\sum x_i \cdot F_i}{n} = \sum x_i \cdot \left(\frac{F_i}{n} \right) = \sum x_i \cdot f_i$$

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE CENTRALIZACIÓN

EJEMPLO 1

Una pequeña empresa tiene cinco trabajadores. Sus salarios mensuales son: 510, 560, 575, 600 y 800 Euros. Calcular el salario medio.

$$\bar{x} = \frac{510 + 560 + 575 + 600 + 800}{5} = 609$$

¿Qué ocurriría si el valor 800 fuera 5000?.

510, 560, 575, 600 y 5000

$$\bar{x} = \frac{510 + 560 + 575 + 600 + 5000}{5} = 1449$$

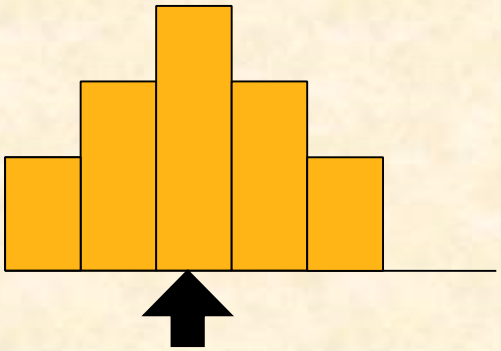
La medida pierde representatividad

MEDIDAS DE UN CONJUNTO DE DATOS

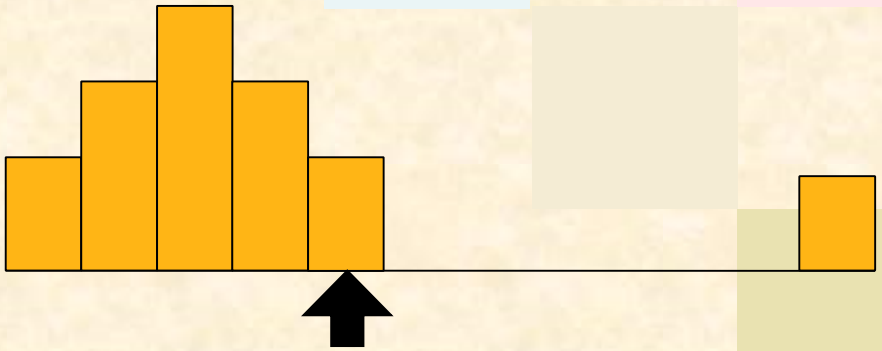
MEDIDAS DE CENTRALIZACIÓN

EJEMPLO 1

- Actúa como centro geométrico o como centro de masas del conjunto de datos.
- Es muy sensible a valores extremos (atípicos).



MEDIA



MEDIA

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE CENTRALIZACIÓN

MEDIANA

- Es un valor que divide a los datos en dos grupos con el mismo número de individuos.
- Es conveniente cuando los datos son asimétricos.
- No es sensible a valores extremos.
- Para calcularla:
  - ordenamos los datos de menor a mayor.
  - si el número de datos es impar, la mediana es el dato del medio
  - si el número de datos es par, la mediana es la media de los datos centrales.

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE CENTRALIZACIÓN

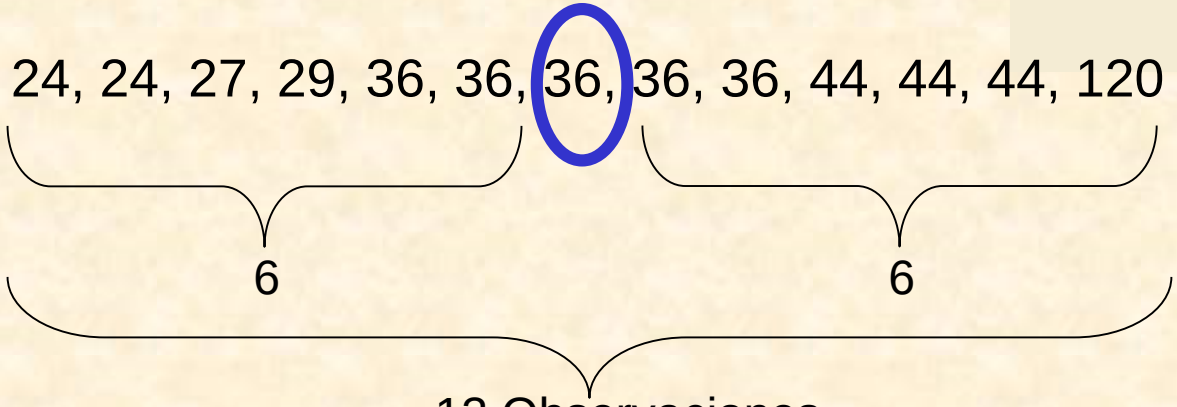
EJEMPLO 2

13 ovejas comieron una hierba venenosa. Las horas que tardaron en morir fueron:

44, 27, 24, 24, 36, 36, 44, 44, 120, 29, 36, 36 y 36.

Calcular **la mediana**.

→ Ordenamos los valores de menor a mayor:

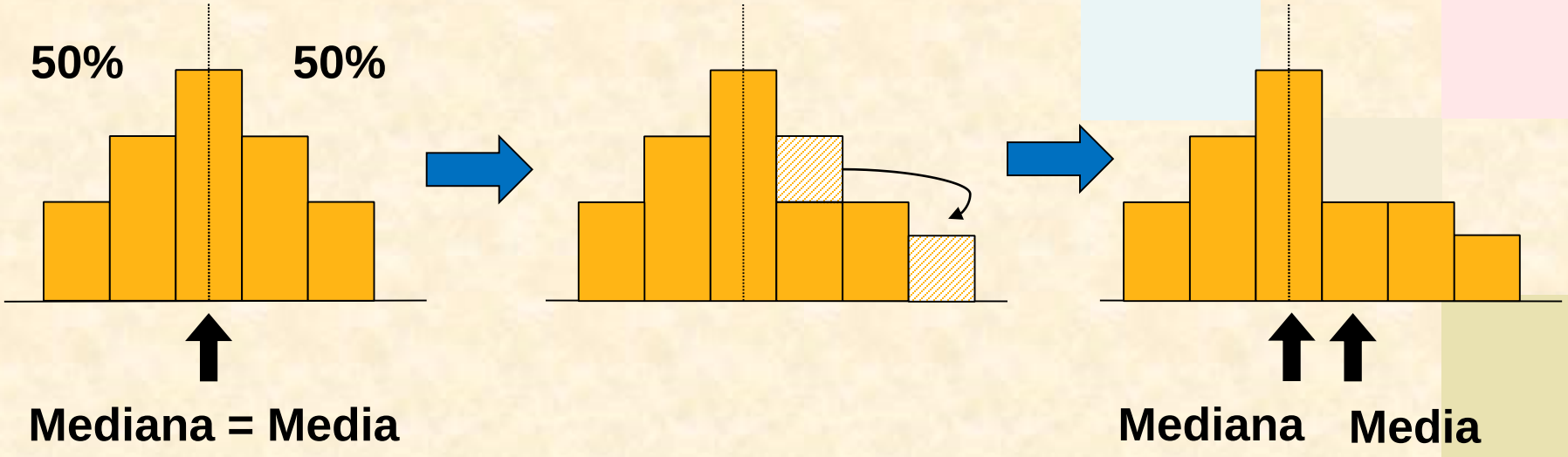


MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE CENTRALIZACIÓN

MEDIANA

- Poco sensible a las **asimetrías** del histograma.

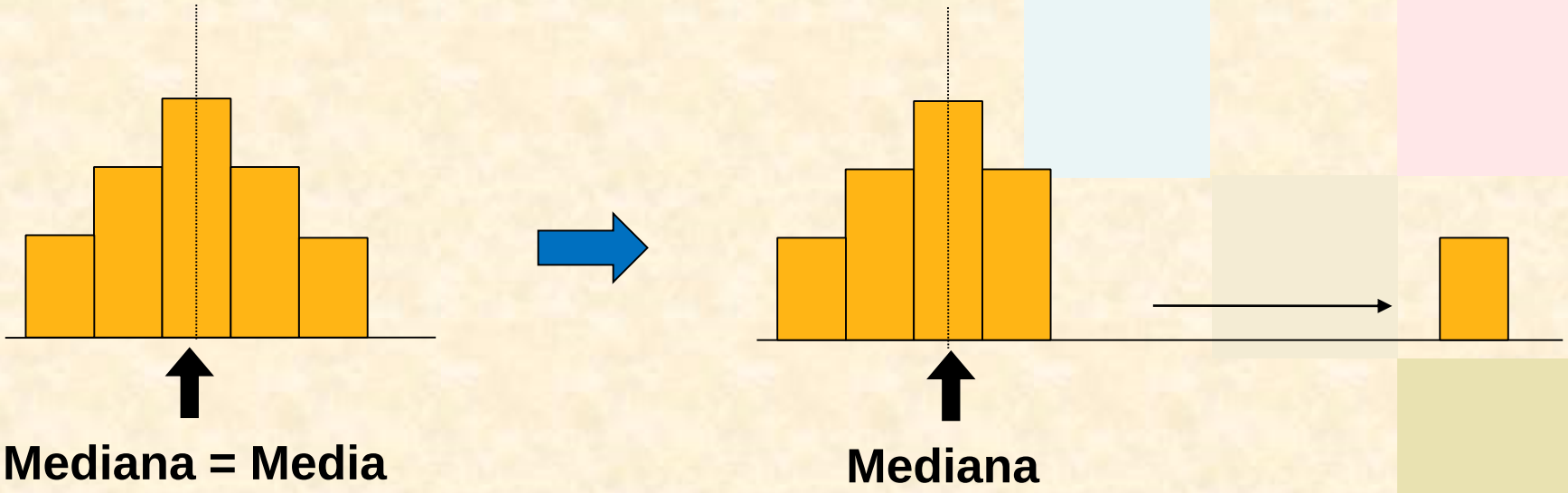


MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE CENTRALIZACIÓN

MEDIANA

- Poco sensible a **valores atípicos**.

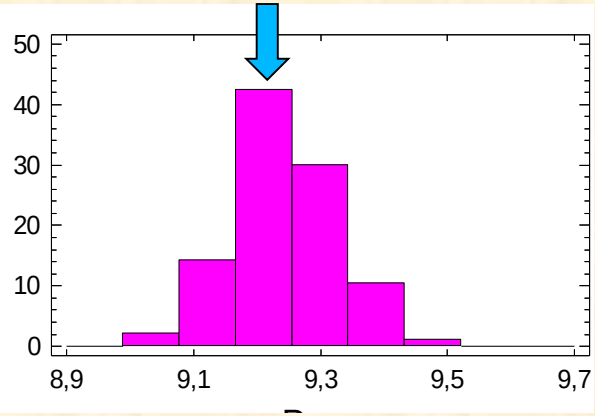


MEDIDAS DE UN CONJUNTO DE DATOS

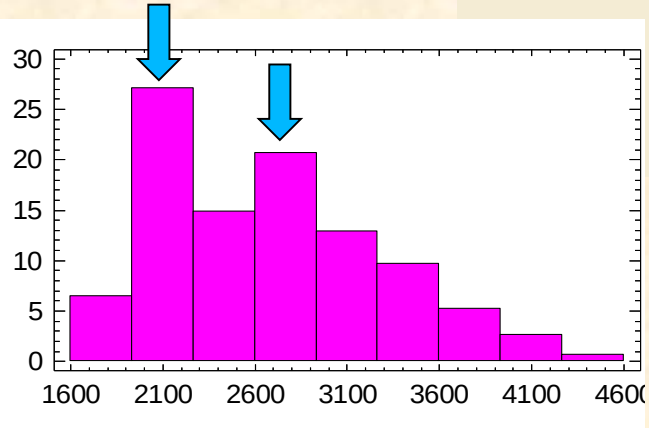
MEDIDAS DE CENTRALIZACIÓN

MODA

- Es el valor más frecuente, el que más se repite.
- En datos agrupados, es la clase más frecuente.
- La presencia de varias modas puede indicar la existencia de varios grupos.



Unimodal



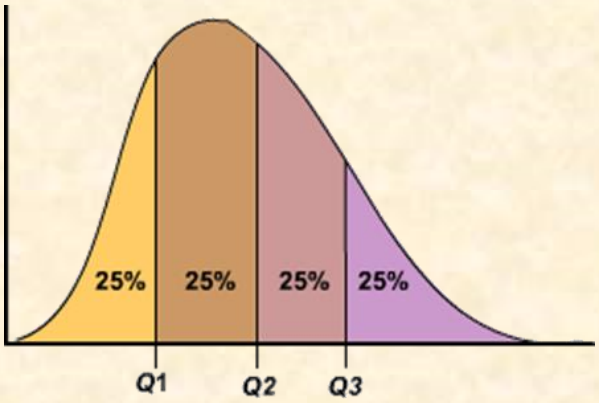
Bimodal

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE POSICIÓN (CUANTILES)

CUARTILES

- Son valores no centrales muy importantes de las distribuciones.
- Son valores de la variable ($Q_1, Q_2\ y\ Q_3$) que dividen a la distribución en 4 partes, cada una de las cuales engloba el 25 % de las mismas.



Q_1= Primer cuartil

Q_2= Segundo cuartil = Mediana

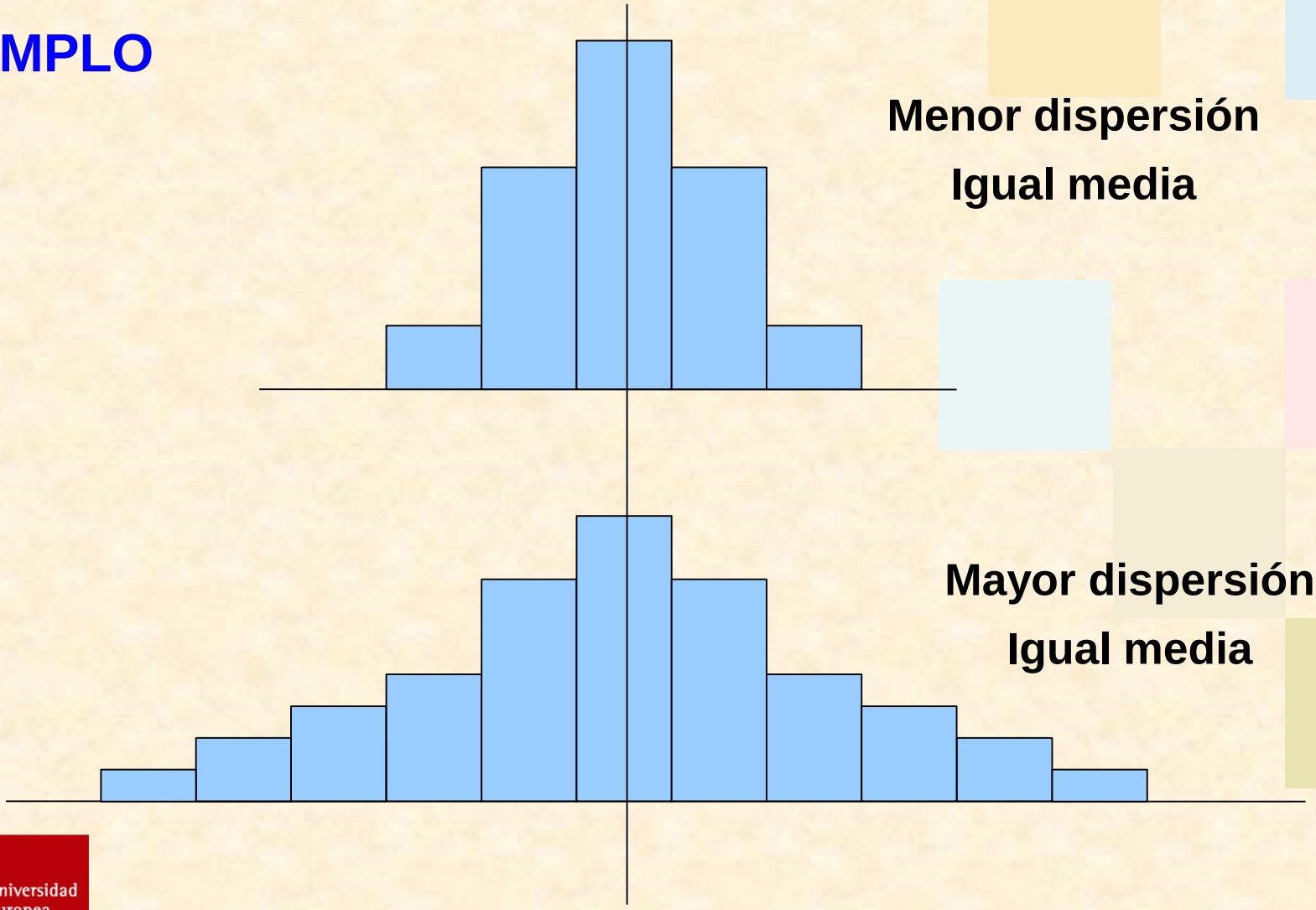
Q_3= Tercer cuartil

$Q_3 - Q_1$= Rango intercuartílico

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN

EJEMPLO



Menor dispersión

Igual media

Mayor dispersión

Igual media

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN

EJEMPLO

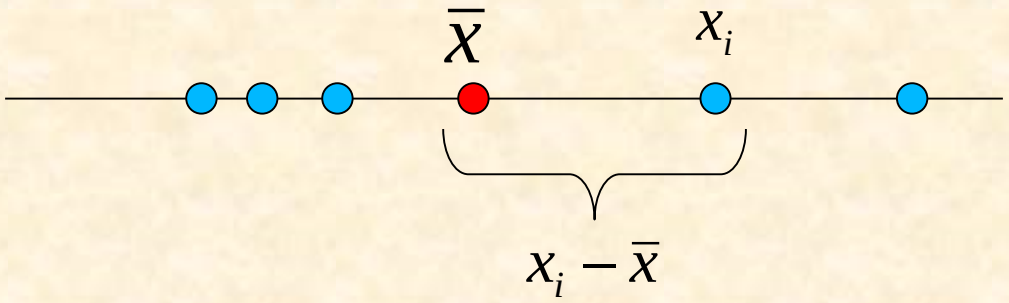
- Los valores 0 y 10 tienen como media 5.
- Los valores 5 y 5 tienen como media 5.
- En ambos casos tienen la misma media, sin embargo los conjuntos son diferentes.

- En ocasiones, conocer sólo la media **no** nos da una idea de cómo están repartidos el resto de valores entorno a ella.

¿están cerca o lejos de la media?

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN



Distancia entre un valor y la media

- Calculamos la **distancia media** como el promedio de las distancias de todos los valores al valor medio:

$$D = \frac{1}{n}\sum (x_i - \bar{x})$$

Problema: La distancia media puede salir 0 sin que los puntos sean todos igual que la media, puesto que los valores se pueden cancelar.

Solución: Quitar los signos negativos elevando al cuadrado.

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN

VARIANZA

- Mide el promedio de las desviaciones (al cuadrado) de las observaciones con respecto a la media.
- Para datos no agrupados la **varianza** de un conjunto de datos $x_1, x_2, x_3, \ldots x_n$ es:

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Distancia cuadrática

- Una manera fácil de calcular la varianza es utilizar el desarrollo:

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \bar{x}^2$$

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN

VARIANZA

- Si los datos están agrupados en una tabla de frecuencias:

$$\sigma^2 = \left(\frac{1}{n} \sum F_i x_i^2 \right) - \bar{x}^2$$

$$\sigma^2 = \left(\sum f_i x_i^2 \right) - \bar{x}^2$$

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN

VARIANZA POBLACIONAL VS MUESTRAL

- Para datos no agrupados la **varianza poblacional** de un conjunto de datos $x_1, x_2, x_3, \dots x_N$ (toda la población) es:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Para datos no agrupados la **varianza muestral** de un conjunto de datos $x_1, x_2, x_3, \dots x_n$ (muestra de la población) es:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Estimador no sesgado de la varianza poblacional

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE DISPERSIÓN

DESVIACIÓN TÍPICA

- Para lograr una medida de la distancia media calculamos la raíz cuadrada de la varianza:

$$\sigma = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

- Tiene las mismas unidades que la variable estadística y es en general más “tangible”.

MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE FORMA

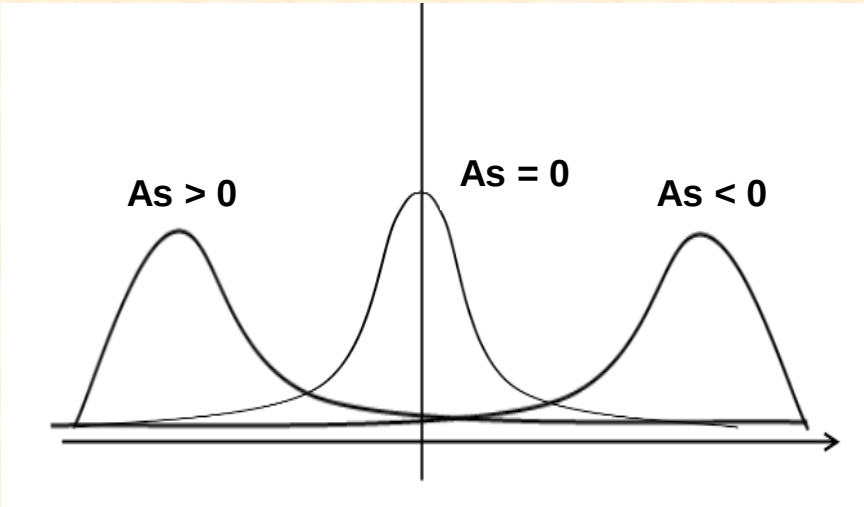
COEFICIENTE DE ASIMETRÍA

- El **coeficiente de asimetría** mide la simetría de los datos respecto de la media. Se define como:

$$As = \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{\sigma^3}$$

$x_1, x_2, x_3, \ldots x_N$ constituye toda la población

σ es la desviación típica



MEDIDAS DE UN CONJUNTO DE DATOS

MEDIDAS DE FORMA

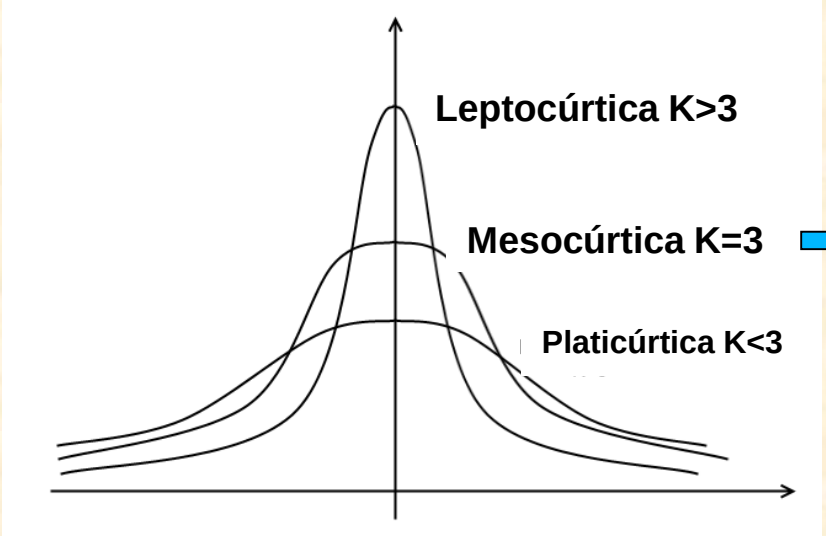
CURTOSIS

- El coeficiente de **curtosis** mide el apuntamiento de los datos. Se define como:

$$K = \frac{1}{N}\frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{\sigma^4}$$

$x_1, x_2, x_3, \ldots x_N$ constituye toda la población

σ es la desviación típica

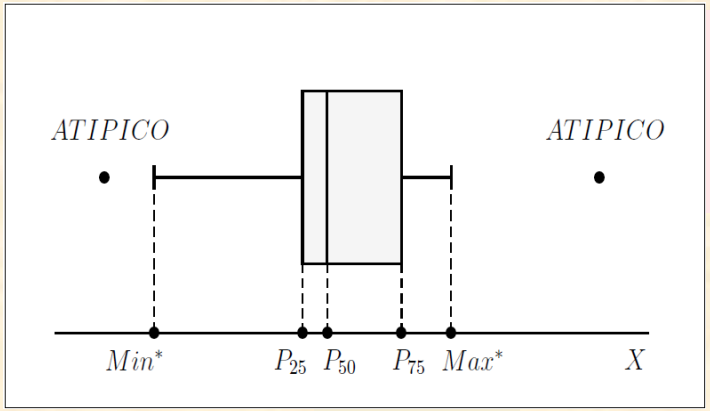


Distribución Gaussiana

MEDIDAS DE UN CONJUNTO DE DATOS

DIAGRAMA DE CAJA Y BIGOTES

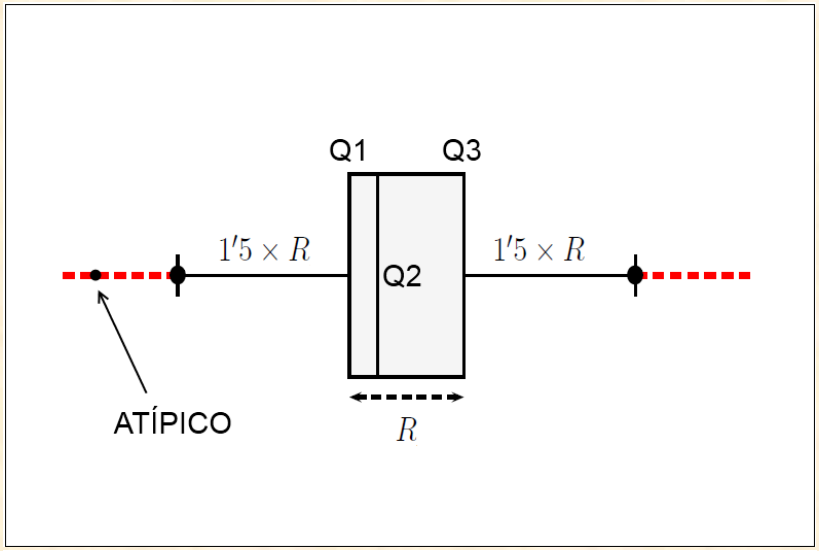
- Es una representación gráfica de un conjunto de datos que consta de dos partes, la caja y los bigotes.

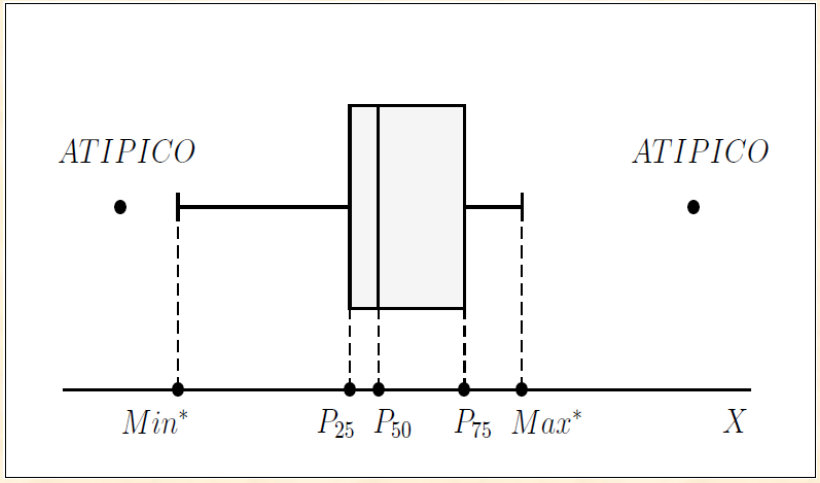


- Ofrece información acerca de:
  - ❑ La simetría de los datos.
  - ❑ La concentración de los datos.
  - ❑ La dispersión de los datos.
  - ❑ La presencia de puntos atípicos y su desvío respecto de generalidad.

MEDIDAS DE UN CONJUNTO DE DATOS

DIAGRAMA DE CAJA Y BIGOTES





- Q_1= Primer cuartil
- Q_2= Segundo cuartil = Mediana
- Q_3= Tercer cuartil
- R = $Q_3 - Q_1$

(Rango intercuartílico)

- L_{max} de los bigotes = 1.5 x R

- **Valores atípicos**:

Valores tal que $> Q_3 + 1.5 \times R$

Valores tal que $< Q_1 - 1.5 \times R$

- **Extremos del bigote**:

Máx valor tal que $< Q_3 + 1.5 \times R$

Mín valor tal que $> Q_1 - 1.5 \times R$