Tema 2. Estadística descriptiva

(Actualizado el 22/09/23. Se han corregido errores)

Vicente Coll Serrano

Universitat de València

2023-09-28

Objetivo

- Introducir los conceptos básicos de estadística descriptiva como son: tabla de frecuencias, media, mediana, cuantiles, moda, dispersión, etc. Aprender a calcular estas medidas a mano y utilizando la hoja de cálculo Excel.

Introducción

¿Por qué analizar datos? (Pearson, 218:1)

- Para comprender lo que ha sucedido o está sucediendo.
- Para predecir lo que es probable que suceda (o en el futuro o bajo otras circunstancias que no se han dada aún).
- Para guiarnos en la toma de decisiones.

Definición y relaciones

Muchas definiciones de qué es la estadística. Una definición muy práctica: "Ciencia encargada de recoger, analizar, presentar e interpretar datos".

Definición y relaciones

ESTADÍSTICA DESCRIPTIVA

Recogida de información, descripción y análisis de un grupo de datos utilizando métodos numéricos y gráficos.

- Notas de un examen.
- Censo de población.

CÁLCULO DE PROBABILIDADES

(Juegos de azar).
La descripción de la realidad se modeliza utilizando métodos de Análisis Matemático.

INFERENCIA ESTADÍSTICA

Basándose en el cálculo de probabilidades, y a partir de los datos de una muestra, se efectúan estimaciones, decisiones y predicciones.

Permite generalizar sobre un conjunto mayor de datos (población).

- Estatura media de los ciudadanos de la Comunidad Valenciana.

Introducción de conceptos

En el proceso de observación y experimentación, se pueden encontrar o pueden acontecer dos tipos de fenómenos:

- **Fenómenos causales o deterministas**
  En idénticas condiciones dan los mismos resultados. Dadas unas causas, puede predecirse un resultado final.
- **Fenómenos inciertos, aleatorios o debidos al azar**
  Dadas unas causas no puede pronosticarse un resultado final. Influye el azar por lo que para idénticas condiciones nos encontramos con resultados distintos.

Introducción de conceptos

- **VARIABLE ESTADÍSTICA**
  Se refiere a una característica que puede tomar cualquier modalidad de un conjunto determinado o "dominio de la variable".
- **DATO**
  Par formado por una unidad observada y su correspondiente característica.
  Ejemplo: un individuo y su edad.

En general, cuando se habla de "datos" (ejemplo: "tenemos un conjunto de datos..") nos referimos a una matriz en la que las filas corresponden a los individuos (observaciones) y las columna a las variables.

Introducción de conceptos

Población: es el conjunto de las posibles observaciones de la característica (o variable) común que queremos analizar de un universo. La población puede ser finita o infinita. Si únicamente consideramos las observaciones de una característica del universo obtendremos una población unidimensional; pero si consideramos dos variables tendremos una población bidimensional, si consideramos tres variables una población tridimensional, etc.

Muestra: Es una parte (un subconjunto) representativa de la población. En muchas situaciones no es posible acceder al estudio de la población por diversas razones: imposibilidad, coste, etc. Por esta razón utilizamos la muestra para obtener información relevante acerca de la población.

Tipos de variables

Cualitativas: No se pueden medir. Pueden hacer referencia a:

- **Nominales o atributos**: Son variables de tipo nominal, se establecen diferentes categorías.
  - **Ejemplo**: estado civil, tipo de defecto en la fabricación de un producto, veredicto en un juicio, etc.
- **Ordinales**: Aquellas en las que es posible establecer cierta ordenación entre las diferentes categorías.
  - **Ejemplo**: motivación de los empleados, calidad del servicio prestado, grado de intensidad de la competencia, etc.

Tipos de variables

Cuantitativas: Son medibles. Pueden clasificarse en:

- **Discretas**: Cuando su dominio es un conjunto numerable de valores.
  - **Ejemplo**: número de defectos, volumen de ventas de una empresa, número de proveedores, etc.

Entre dos valores consecutivos la variable no puede tomar ningún valor intermedio.

- **Continuas**: Cuando su dominio es continuo, es decir, cuando entre dos valores la variable puede tomar, al menos teóricamente, cualquiera de los infinitos valores existentes entre ellos.
  - **Ejemplo**: altura, peso, temperatura, potencia, etc.).

Representación de los datos

Tabla estadística o distribución de frecuencias (datos no agrupados)

x_i	n_i	N_i	f_i	F_i
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	N_i	f_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_l	n_l	N_l	f_l	F_l
	$\sum = n$		$\sum = 1$	

Representación de los datos

X: Variable estadística.

x_i son cada uno de los valores de la variable.

Frecuencia absoluta ordinaria (n_i): Número de veces que se repite cada valor de la variable.

Frecuencia absoluta acumulada (N_i): Número de veces que se repite un valor inferior o igual a x_i, es decir:

$$\mathrm{N_i} = \mathrm{n_1} + \mathrm{n_2} + \cdots + \mathrm{n_i} = \sum_{\mathrm{i}=1}^{\mathrm{i}} \mathrm{n_i}$$

Representación de los datos

Frecuencia relativa ordinaria (f_i): Proporción de veces que se repite cada valor de la variable. Es el cociente entre cada una de las frecuencias absolutas y el número total de observaciones, es decir:

$$f_i = \frac{n_i}{n}$$

Frecuencia relativa acumulada (F_i): Proporción de veces que se repite un valor inferior o igual a x_i, es decir:

$$\mathrm{F_i} = \mathrm{f_1} + \mathrm{f_2} + \cdots + \mathrm{f_i} = \sum_{\mathrm{i}=1}^{\mathrm{i}} \mathrm{f_i}$$

o también $F_i = \frac{N_i}{n}$

NOTA MUY IMPORTANTE

El sumatorio de las observaciones se puede identificar como n si trabajamos con una muestra o como N si trabajamos con todas las observaciones, la población.

Medidas de posición (1)

Medidas de tendencia central: Describen la localización central de un conjunto de observaciones numéricas.

- Media (aritmética)
- Mediana
- Moda
- Otras medidas: Media ponderada, Media geométrica, Media armónica, Rango medio, etc.

Medidas de tendencia no central.

- Cuantiles
  - Cuartiles
  - Deciles
  - Centiles/Percentiles

Medidas de posición (2)

Una medida de centralización es aquel valor que es capaz de representar todos los datos.

Media aritmética ($\bar{x}$): es la suma de todos los valores del conjunto de datos dividido entre el número total de observaciones.

- A partir de los datos brutos:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Si los valores de la variable están resumidos en una tabla estadística:

$$\bar{x} = \frac{\sum_{i=1}^{l} x_i \cdot n_i}{n} = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_l n_l}{n}$$

Medidas de posición (3)

Si $\bar{x}_i$ (i=1,2,..,k) corresponden a las medias de k grupos distintos de tamaño N_i (i=1,2,..,k), se cumple que la media del conjunto es:

$$\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i N_i}{\sum_{i=1}^{k} N_i} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \cdots + \bar{x}_k N_k}{N_1 + N_2 + \cdots + N_k}$$

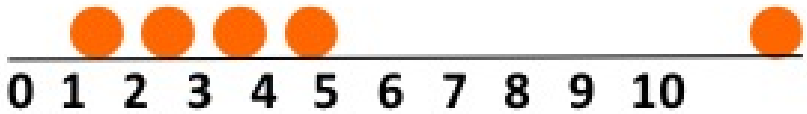
Medidas de posición (3)

Ventajas:

- Su cálculo es sencillo e intervienen todos los valores de la distribución.
- Resulta facil de interpretar.
- Es única.

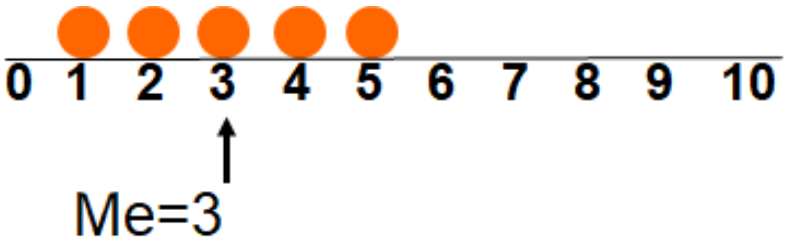
Inconvenientes:

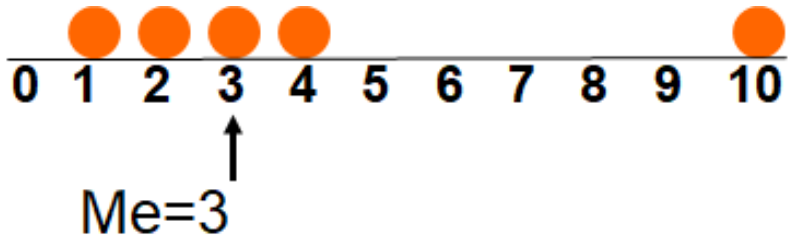
La media no es la mejor medida para describir o resumir un conjunto de datos que tiene valores extremos.

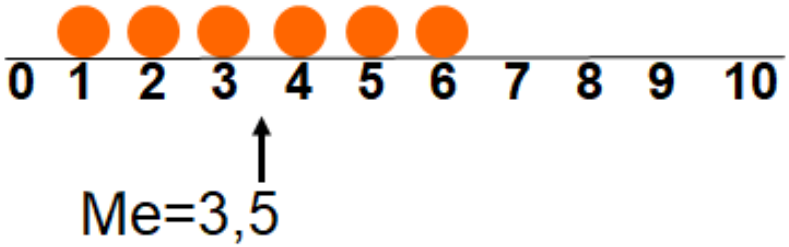


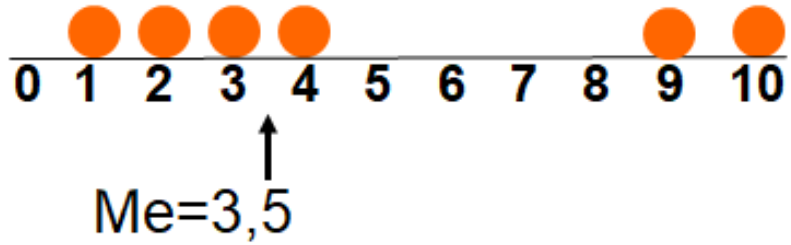
Medidas de posición (4)

Mediana (*Me*): valor del recorrido de la variable que deja aproximadamente el mismo número de observaciones a su izquierda y a su derecha.









Medidas de posición (5)

La Media es la medida de tendencia central que más se usa, pero como se ve influenciada por valores extremos, en estos casos con frecuencia la Mediana es preferida.

Medidas de posición (6)

Moda (*Mo*)

Distribución de frecuencias de valores sin agrupar: Valor de la variable de mayor frecuencia (n_i o f_i).

Distribución de frecuencias de valores agrupados:

- Misma amplitud: intervalo modal en el intervalo de mayor frecuencia o altura.
- Diferente amplitud: intervalo modal en el intervalo de mayor altura.

Medidas de posición (7)

Cuantiles:

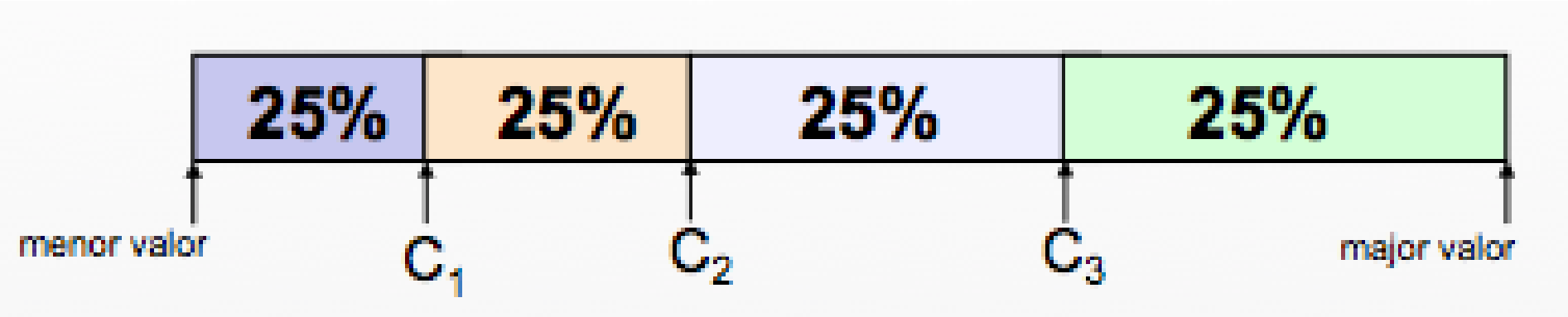
- CUARTILES (k=4): C_s (s=1,2,3,4)
- DECILES (k=10): D_s (s=1,2,3,..,10)
- PERCENTILES (k=100): P_s (s=1,2,3,..,100)

Expresión general de cálculo:

$$\text{Si} \begin{cases} n_{i-1} < \dfrac{s \cdot N}{k} < N_i \Rightarrow Q_{\frac{s}{k}} = x_i \\ n_i = \dfrac{s \cdot n}{k} \Rightarrow Q_{\frac{s}{k}} = \dfrac{x_i + x_{i+1}}{2} \end{cases}$$

Medidas de posición (8)

cuartil (k=4 , s=1, 2, 3)



$Q_{1/4}=C_1$: Valor del recorrido de la variable para el cual el 25% de las observaciones son más pequeñas y el 75% son mayores.

¿Cómo interpretamos el valor que toma el $P_{40}(= Q_{40/100})$?

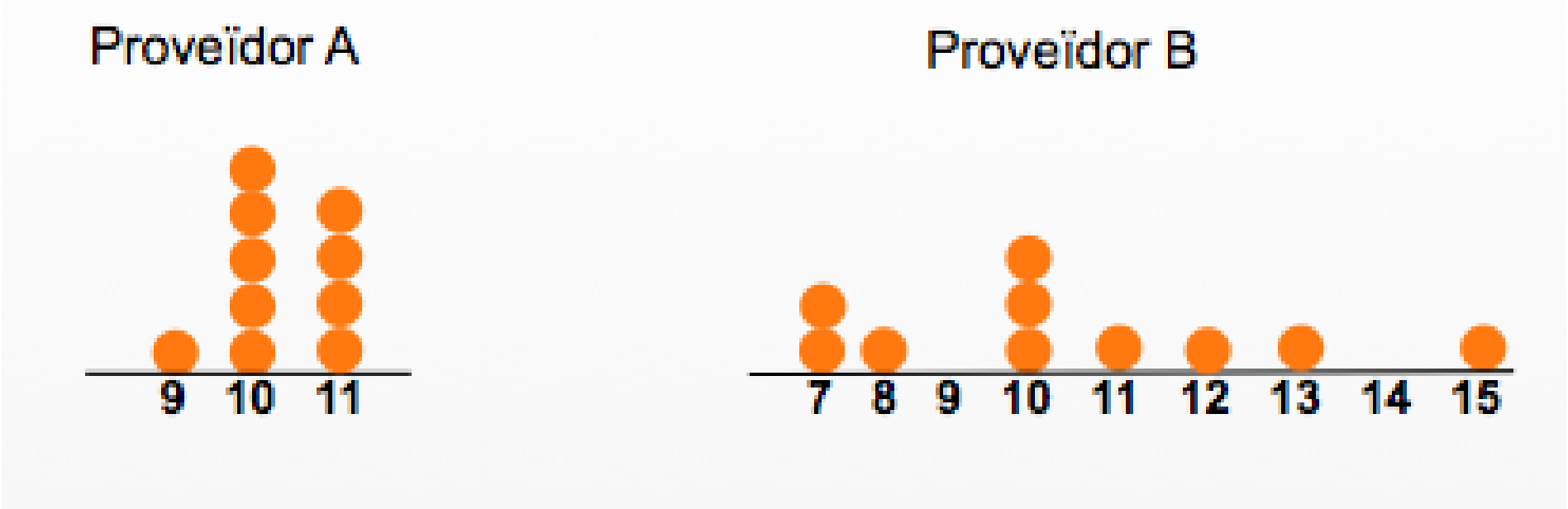
¿Cómo calculamos los diferentes Cuantiles?

Medidas de dispersión o variación (1)

Medidas de posición: valores alrededor de los cuales se sitúa un grupo de observacioness.

Información insuficiente: no tienen en cuenta la situación relativa de los datos.

Ejemplo: Somos el responsable de compras de una empresa. Buscamos información sobre 2 proveedores y determinamos que ambos tardan por término medio 10 días en servir el pedido.



¿Qué proveedor preferimos?
¿Qué proveedor es más consistentes/confiable?

Medidas de dispersión o variación (2)

Medidas de dispersión absoluta.

- Rango o recorrido.
- Recorrido intercuartílico.
- Varianza.
- Desviación típica.

Medidas de dispersión relativa.

- Coeficiente de variación de Pearson.
- Otras medidas: Recorrido intercuartílico relativo, coeficiente de variación mediano, recorrido relativo, etc.

Medidas de dispersión o variación (3)

La Varianza, a diferencia del Rango y el Recorrido Intercuartílico, consideran cómo se distribuyen o agrupan las observaciones.

La **varianza (S^2)**: es la media de los cuadrados de las diferencias entre los valores de la variable y su media.

Si calculamos la varianza a partir de los datos brutos:

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

Si hacemos los cálculos a partir de una tabla de frecuencias:

$$S_x^2 = \frac{\sum_{i=1}^{I} (x_i - \bar{x})^2 \cdot n_i}{n}$$

Evalúa la manera en que fluctúan los valores de la variable respecto a la media.

¿Por qué se utiliza el cuadrado de las diferencias?

Medidas de dispersión o variación (4)

Al hacer el cuadrado, las observaciones que se encuentran más lejos de la media adquieren más peso que las más cercanas. Cuanto mayor sea la varianza (S^2) más dispersión de los datos.

Si en la expresión anterior de la S^2 desarrollamos el cuadrado:

$$S_x^2 = \frac{\sum_{i=1}^{n} {x_i}^2}{n} - \bar{x}^2 \quad o \quad S_x^2 = \frac{\sum_{i=1}^{l} {x_i}^2 \cdot n_i}{n} - \bar{x}^2$$

- Utiliza sólo los valores de la variable
- Más sencillo y rápido de calcular

Medidas de dispersión o variación (6)

- Rango o recorrido: $\text{Re} = x_{\max} - x_{\min}$
- Varianza (o varianza poblacional)

$$S_x^2 = \frac{\sum_{i=1}^{I} (x_i - \bar{x})^2 \cdot n_i}{n} \quad o \quad S_x^2 = \frac{\sum_{i=1}^{I} {x_i}^2 \cdot n_i}{n} - \bar{x}^2$$

- Desviación típica

$$S_X = +\sqrt{S_X^2} \geqslant 0$$

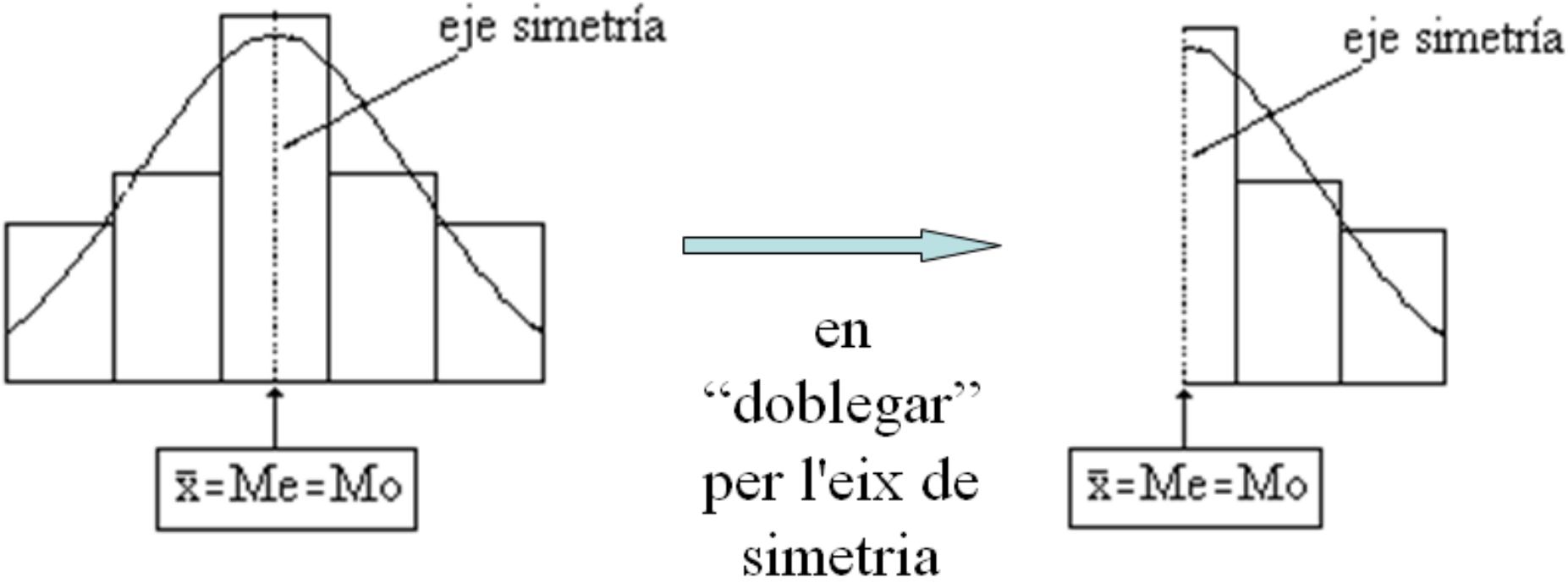
- Cuasivarianza (o varianza muestral)

$$S_x^{2*} = \frac{n}{n-1} \cdot S_X^2$$

- Coeficiente de variación (de Pearson) (PARA COMPARAR)

$$g_0(X) = \frac{S_X}{|\bar{x}|}$$

Medidas de forma: simetria (1)



Coeficiente de asimetría (de Fisher):

$$g_1(X) = \frac{\sum_{i=1}^{I} (x_i - \bar{x})^3 \cdot n_i}{\frac{n}{S_X^3}}$$

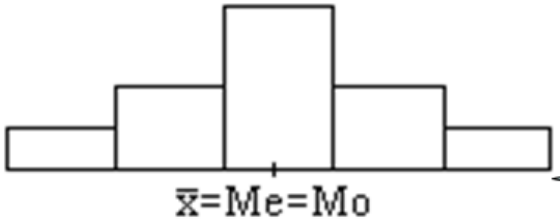
Son muchos cálculos para hacerlos a mano (lo haremos con el ordenador).

Medidas de forma: simetria (2)

Distribució asimètrica negativa

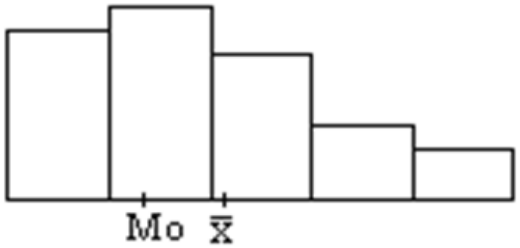
cua més llarga a l'esquerra de la moda

$g_1(X) < 0$



Distribució simètrica

$g_1(X) = 0$

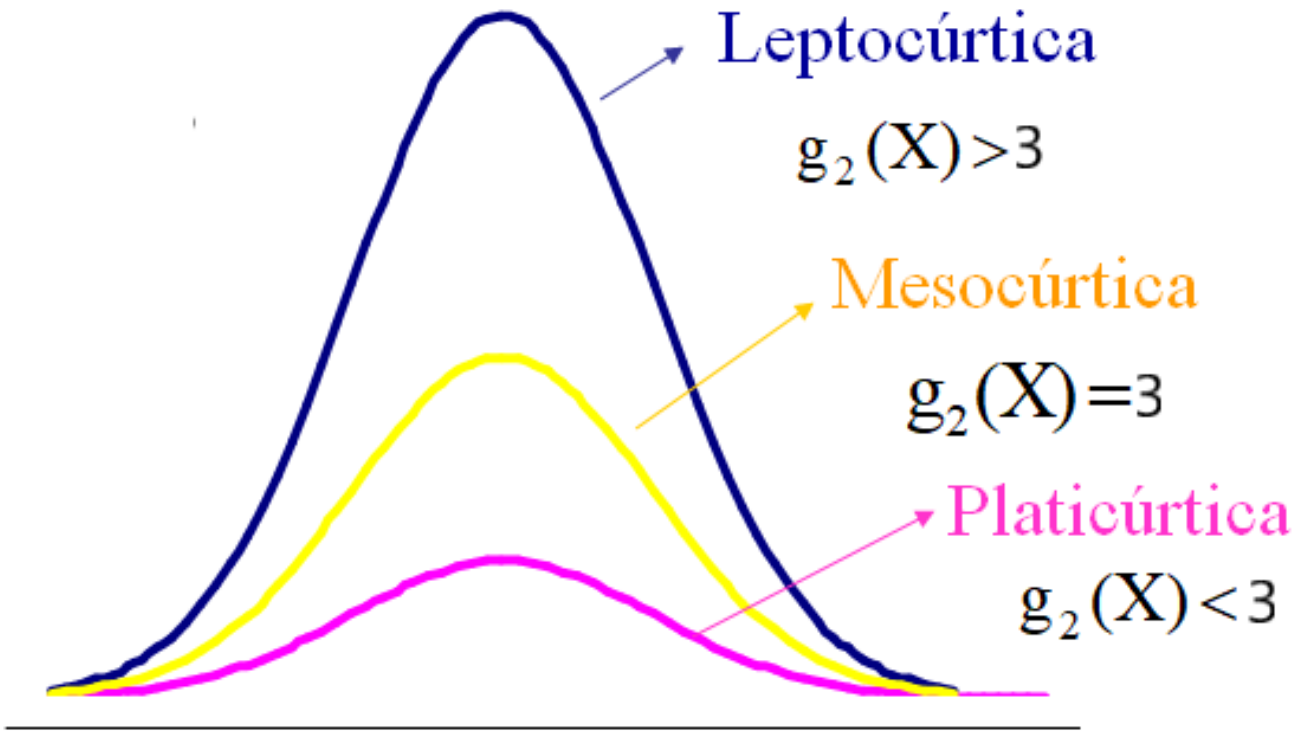


Distribució asimètrica positiva

cua més llarga a la dreta de la moda

$g_1(X) > 0$

Medidas de forma: apuntamiento



Coeficiente de Curtosis:

$$g_2(X) = \frac{\sum_{i=1}^{I} (x_i - \bar{x})^4 \cdot n_i}{\frac{n}{S_X^4}}$$

Son muchos cálculos para hacerlos a mano (lo haremos con el ordenador).

Tipificación de variables

Variable tipificada:

$$Z = \frac{X - \bar{x}}{S_X} \quad con \quad \bar{z} = 0 \quad y \quad S_z = 1$$

Valores tipificados:

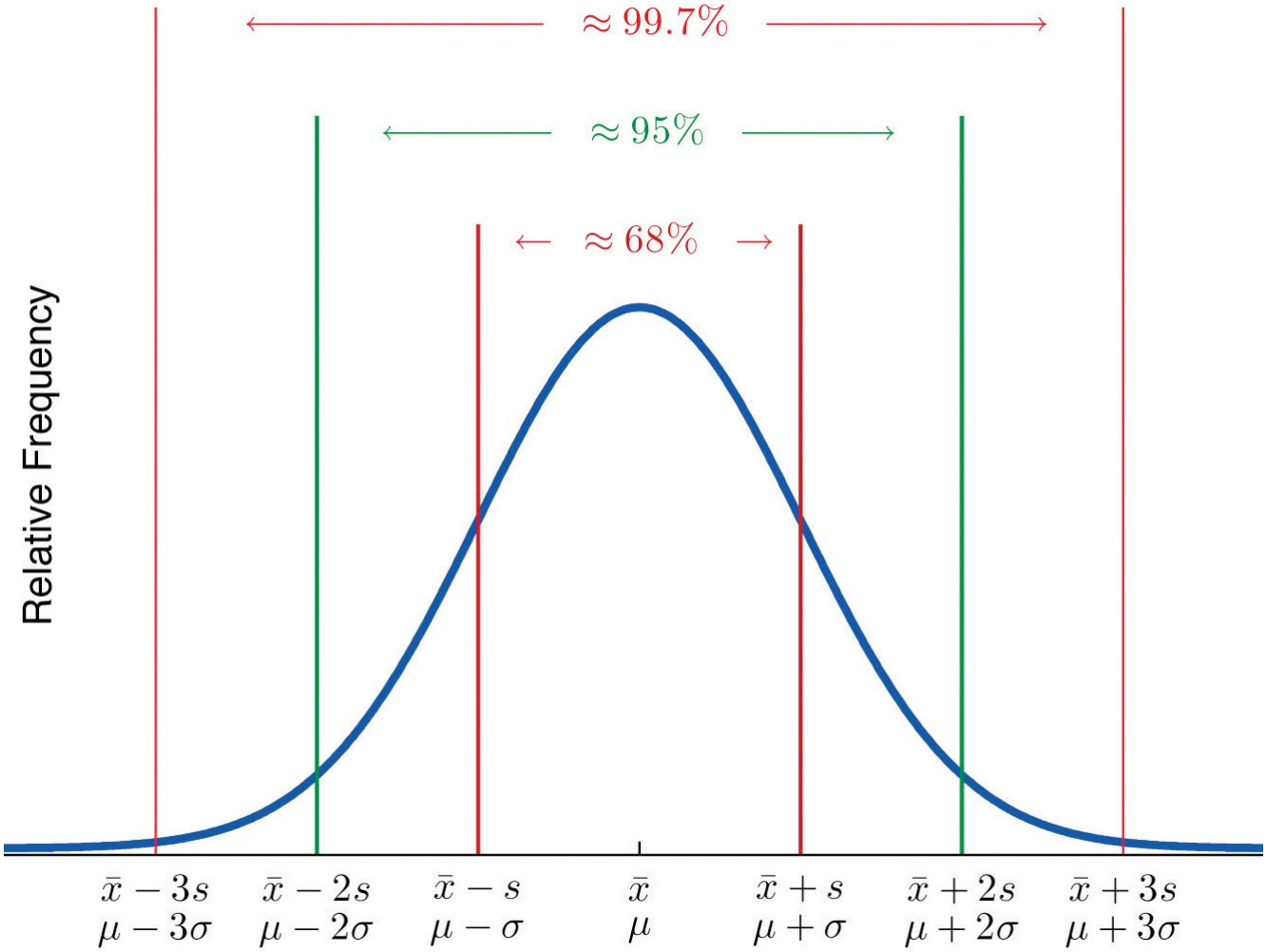
$$z_i = \frac{x_i - \bar{x}}{S_X}$$

Utilidad:

- Las variables pasan a ser adimensionales.
- Permite **comparar** utilizando como distancia el número de desviaciones típicas respecto a la media.
- Para la detección de valores atípicos (outliers o anómalos).

Valores atípicos (1)

Regla Empírica



Valores atípicos (2)

Diagrama de Caja (Box plot)