## TEMA 1: Estadística descriptiva unidimensional

- Introducción.
- Distribuciones de frecuencia.
- Representaciones gráficas.
- Características unidimensionales:
  - Medidas de posición
  - Medidas de dispersión.
  - Momentos.
  - Medidas de forma.

# INTRODUCCIÓN

*La Estadística estudia los métodos científicos para recoger, organizar, resumir y analizar datos, así como para sacar conclusiones válidas y tomar decisiones razonables basadas en tal análisis* (M.R. Spiegel, 1961).

**Estadística Descriptiva** **Estadística Inferencial**

**OBJETIVO:** obtener conocimiento sobre algún tópico, situación o fenómeno real, recogiendo información sobre el mismo, sintetizando, analizando e interpretando dicha información.

*La Estadística es el estudio de cómo debe emplearse la información disponible sobre una situación práctica que envuelve incertidumbre, para dar una guía de acción en tal situación.* (V. Barnett, 1975).

- **Fenómenos determinísticos:** se concretan siempre en un mismo resultado cuando se realizan bajo las mismas condiciones.
- **Fenómenos aleatorios:** están sujetos a *incertidumbre*, en el sentido de que pueden concretarse en distintos resultados, incluso si las condiciones de realización son las mismas.

**Estadística Descriptiva**

*- Recoger, clasificar y representar datos.*
*- Describir la información contenida en los datos mediante ciertas medidas resumen.*
*- Interpretar las medidas descriptivas para explicar el comportamiento de los datos.*

**Cálculo de Probabilidades**

*Establecer y desarrollar modelos teóricos que cuantifican la incertidumbre, y permiten explicar el comportamiento de los fenómenos aleatorios.*

**Estadística Inferencial**

*Desarrollar métodos para ajustar un modelo probalístico teórico a una situación de incertidumbre, a partir de la información proporcionada por un conjunto de datos referidos a dicha situación.*

## CONCEPTOS BÁSICOS

- **Población (colectivo o universo):** conjunto de elementos sobre el que se desea analizar una o varias características. Cada elemento de la población se denomina *individuo* o *unidad estadística*.

  ***Muestra:*** subconjunto de la población en el que se estudian las propiedades de interés, con objeto de inferir conclusiones a la población.

- **Carácter:** propiedad que se desea estudiar en la población (observable en cada uno de los individuos que la componen).

  **Modalidad (categoría):** cada una de las diferentes formas en las que se manifiesta un carácter (*incompatibles y exhaustivas*).

  - ***Carácter cualitativo (atributo)***: modalidades no cuantificables numéricamente.
  - ***Carácter cuantitativo:*** modalidades cuantificables numéricamente.

- **Escalas de medida:** la identificación de las modalidades de un carácter mediante la asignación de símbolos o números se denomina *medición del carácter*. Según su naturaleza, la medición de un carácter se realiza en diferentes escalas:

  - Caracteres cualitativos:
    - *Escala nominal:* entre las diferentes modalidades sólo se dan relaciones de igualdad o desigualdad.
    - *Escala ordinal:* las diferentes modalidades presentan un orden lógico que las hace comparables dos a dos, pudiendo distinguirse si una es igual, mayor o menor que la otra.
  - Caracteres cuantitativos:
    - *Escala de intervalo:* además de existir una relación de orden entre las diferentes modalidades, se puede medir la distancia entre dos cualesquiera (hay unidad de medida), pero no hay origen preestablecido.
    - *Escala de razón*: hay unidad de medida y origen y, por tanto, se puede medir cuántas veces una modalidad es mayor o menor que otra.

- **Variable estadística:** representación numérica de un carácter cuantitativo, cuyos valores resultan de la medición de las modalidades.

  - ***Variable discreta:*** sus valores son puntos "aislados"(presenta, por tanto, un número finito o infinito numerable de valores).
  - ***Variable continua:*** puede tomar, *a priori*, cualquier valor en algún intervalo de la recta real.

# DISTRIBUCIONES DE FRECUENCIAS

Consideramos una población de $n$ individuos, en la que se estudia una variable (o atributo) que presenta los valores (o modalidades) $x_1,\ x_2,\cdots,x_k$.

- **Frecuencia absoluta del valor (modalidad)** $x_i,\ i=1,\ldots,k$**:**

  $n_i$=número de individuos de la población que presentan el valor $x_i$.

  *Distribución de frecuencias absolutas*: $\{(x_i,n_i);\ i=1,\ldots,k\} \longrightarrow \displaystyle\sum_{i=1}^{k} n_i = n.$

- **Frecuencia relativa del valor (modalidad)** $x_i,\ i=1,\ldots,k$**:**

  $f_i$=proporción de individuos que presentan el valor $x_i \longrightarrow f_i = \dfrac{n_i}{n}\cdot$

  *Distribución de frecuencias relativas*: $\{(x_i,f_i);\ i=1,\ldots,k\} \longrightarrow \displaystyle\sum_{i=1}^{k} f_i = 1.$

***Para caracteres medidos al menos en escala ordinal***, supuesto $x_1 < x_2 < \cdots < x_k$:

- **Frecuencia absoluta acumulada del valor (modalidad)** $x_i,\ i=1,\ldots,k$**:**

  $N_i$=número de individuos con valor menor o igual que $x_i \longrightarrow N_i = \displaystyle\sum_{j=1}^{i} n_j.$

  *Distribución de frecuencias absolutas acumuladas*: $\{(x_i,N_i);\ i=1,\ldots,k\} \longrightarrow N_k = n.$

- **Frecuencia relativa acumulada del valor (modalidad)** $x_i,\ i=1,\ldots,k$**:**

  $F_i$= proporción de individuos con valor menor o igual que $x_i \longrightarrow F_i = \dfrac{N_i}{n} = \displaystyle\sum_{j=1}^{i} f_j.$

  *Distribución de frecuencias relativas acumuladas*: $\{(x_i,F_i);\ i=1,\ldots,k\} \longrightarrow F_k = 1.$

## Tabla de frecuencias para variables discretas y atributos

| Valores | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| $x_1$ | $n_1$ | $f_1$ | $N_1 = n_1$ | $F_1 = f_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_i$ | $n_i$ | $f_i$ | $N_i = \sum_{j=1}^{i} n_j$ | $F_i = \sum_{j=1}^{i} f_j$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_k$ | $n_k$ | $f_k$ | $N_k = n$ | $F_k = 1$ |
| *Suma* | $n$ | $1$ | | |

## Tabla de frecuencias para variables continuas

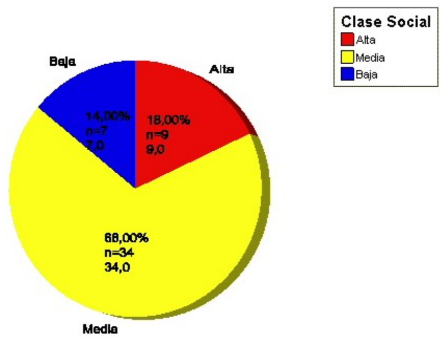
| Intervalos | $n_i$ | $f_i$ | $N_i$ | $F_i$ | $x_i$ | $a_i$ | $h_i$ |
|---|---|---|---|---|---|---|---|
| $(e_0, e_1]$ | $n_1$ | $f_1$ | $N_1 = n_1$ | $F_1 = f_1$ | $x_1$ | $a_1$ | $h_1$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $(e_{i-1}, e_i]$ | $n_i$ | $f_i$ | $N_i = \sum_{j=1}^{i} n_j$ | $F_i = \sum_{j=1}^{i} f_j$ | $x_i = \frac{e_{i-1} + e_i}{2}$ | $a_i = e_i - e_{i-1}$ | $h_i = \frac{n_i}{a_i}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $(e_{k-1}, e_k]$ | $n_k$ | $f_k$ | $N_k = n$ | $F_k = 1$ | $x_k$ | $a_k$ | $h_k$ |
| *Suma* | $n$ | $1$ | | | | | |

- $x_i = \dfrac{e_{i-1} + e_i}{2}, \quad i = 1, \ldots, k \rightarrow$ Marcas de clase
- $a_i = e_i - e_{i-1}, \quad i = 1, \ldots, k \rightarrow$ Amplitudes
- $h_i = \dfrac{n_i}{a_i}, \quad i = 1, \ldots, k \rightarrow$ Densidades de frecuencia

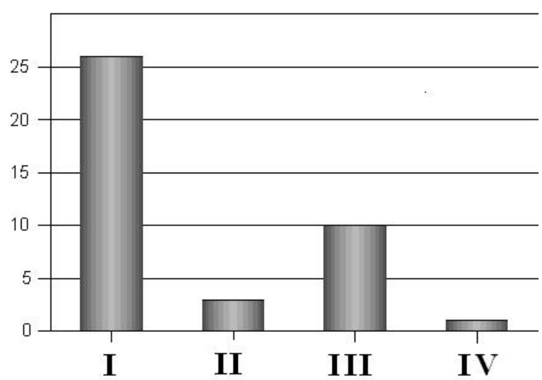
# REPRESENTACIONES GRÁFICAS I

## (caracteres cualitativos o atributos)

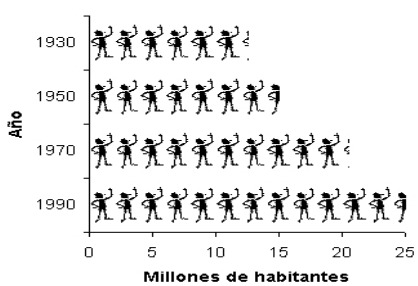
- **Diagrama de sectores**: Es un círculo dividido en tantos sectores circulares como modalidades tenga el carácter, siendo el área de cada uno proporcional a la frecuencia (absoluta o relativa, ya que son proporcionales) de la modalidad.

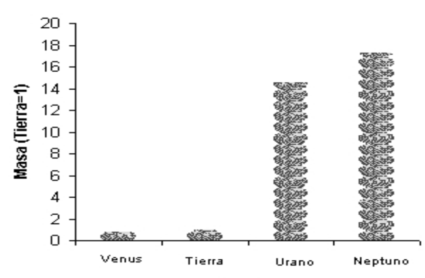



- **Diagrama de rectángulos o barras**: Está formado por rectángulos (uno por modalidad) de base constante y alturas proporcionales a las frecuencias (absolutas o relativas) de cada modalidad.




- **Pictograma**: Se dibujan figuras, normalmente alusivas al carácter que se está estudiando, bien una para cada modalidad con tamaño proporcional a su frecuencia, o bien repitiendo la figura tantas veces como requieran las frecuencias.

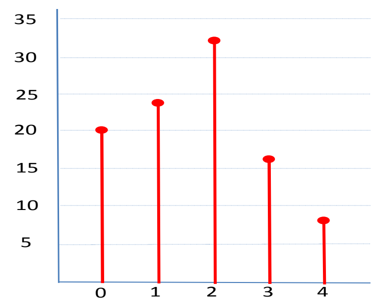


# REPRESENTACIONES GRÁFICAS II
## (variables discretas y continuas con valores no agrupados)

- **Diagrama de barras**: En un sistema de ejes cartesianos se representan los valores de la variable en el eje de abcisas, y se trazan barras verticales con longitudes proporcionales a sus frecuencias (absolutas o relativas).

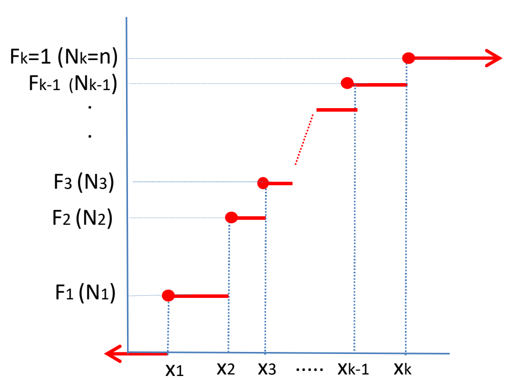
| $x_i$ | $n_i$ | $f_i = n_i/100$ |
|---|---|---|
| 0 | 20 | 0.2 |
| 1 | 24 | 0.24 |
| 2 | 32 | 0.32 |
| 3 | 16 | 0.16 |
| 4 | 8 | 0.08 |
| | 100 | 1 |



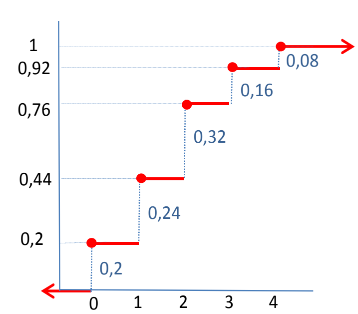

- **Curva acumulativa o de distribución**: Es la representación de la denominada *función acumulativa, de repartición o de distribución*, definida, para cada número real $x$, como la proporción de datos menores o iguales que $x$. Esto es, si $x_1 < x_2 < \cdots < x_k$ son los valores de la variable ordenados y $x_i \le x < x_{i+1}$, $F(x) = F_i$:

$$F : \mathbb{R} \longrightarrow [0,1] \;\text{ definida como }\; F(x) = \begin{cases} 0, & x < x_1 \\ \dfrac{N_i}{n} = F_i, & x_i \le x < x_{i+1} \\ 1, & x \ge x_k. \end{cases}$$

Su representación gráfica es una curva en escalera que parte de 0 y acaba en 1; los saltos corresponden a los valores de la variable, y la longitud de cada salto es la frecuencia relativa de cada valor $(F(x_i) - F(x_{i-1}) = F_i - F_{i-1} = f_i)$.




| $x_i$ | $f_i$ | $F_i$ |
|---|---|---|
| 0 | 0.2 | 0.2 |
| 1 | 0.24 | 0.44 |
| 2 | 0.32 | 0.76 |
| 3 | 0.16 | 0.92 |
| 4 | 0.08 | 1 |




**Nota**: Una curva similar con las frecuencias absolutas acabaría en $n$ (número de datos) y las longitudes de los saltos serían las frecuencias absolutas de cada valor.

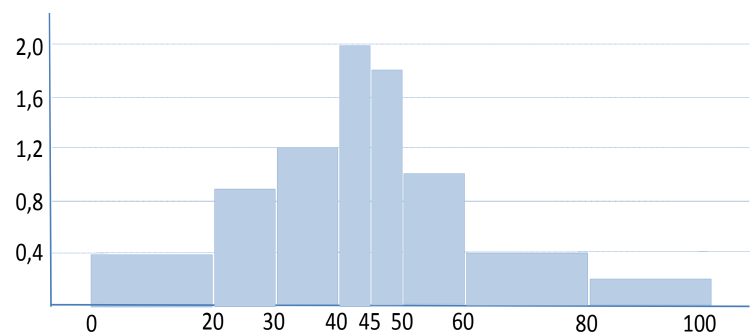
# REPRESENTACIONES GRÁFICAS III

## (variables continuas con valores agrupados en intervalos)

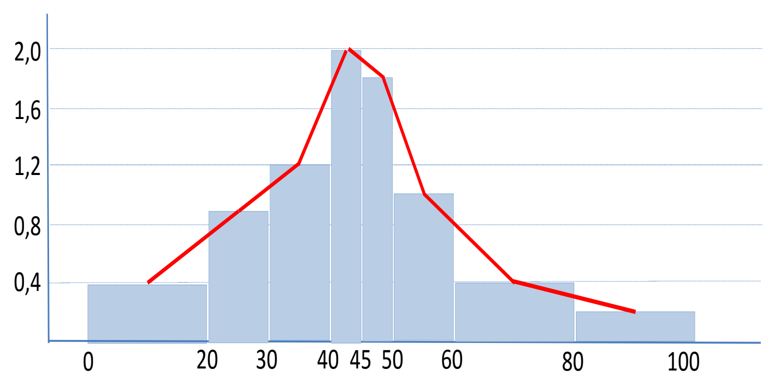
- **Histograma**: Si los valores de una variable continua se presentan agrupados en intervalos de distinta amplitud, las frecuencias de distintos intervalos no son directamente comparables. Se trabaja entonces con la *densidad de frecuencia* de cada intervalo, $h_i = n_i/a_i$, que especifica la *frecuencia media por unidad de amplitud.*

  El histograma está formado por rectángulos yuxtapuestos, cuyas bases son los intervalos de agrupación de los valores, y sus alturas son iguales (o proporcionales) a las densidades de frecuencia; así, el área de cada rectángulo es igual (o proporcional) a la frecuencia del intervalo correspondiente.

| $I_i$ | $n_i$ | $h_i$ |
|---|---|---|
| $(0, 20]$ | 8 | 0.4 |
| $(20, 30]$ | 9 | 0.9 |
| $(30, 40]$ | 12 | 1.2 |
| $(40, 45]$ | 10 | 2 |
| $(45, 50]$ | 9 | 1.8 |
| $(50, 60]$ | 10 | 1 |
| $(60, 80]$ | 8 | 0.4 |
| $(80, 100]$ | 4 | 0.2 |




- **Poligonal de frecuencias**: Es la que resulta al unir los puntos correspondientes a las marcas de clase de los intervalos en el histograma.



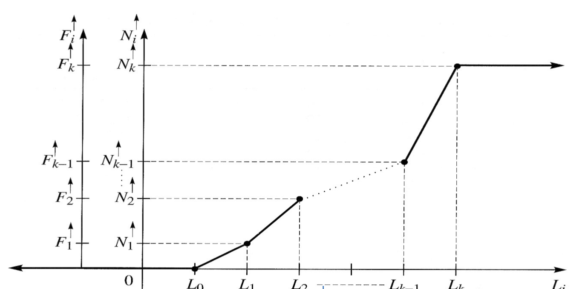
- **Curva acumulativa o de distribución**: Como en el caso de variables discretas, la *función de distribución*, $F$, es la que asigna a cada punto $x \in \mathbb{R}$ la proporción de datos menores o iguales que $x$; esto es, la frecuencia relativa acumulada hasta $x$. Es, como en el caso de variables discretas, una función monótona no decreciente, con $\lim\limits_{x\to-\infty} F(x) = 0$ y $\lim\limits_{x\to+\infty} F(x) = 1$.

  En este caso, como los valores de la variable se presentan agrupados en intervalos, sólo conocemos las frecuencias acumuladas hasta $e_0$, a partir de $e_k$, y en los extremos superiores de los intervalos:

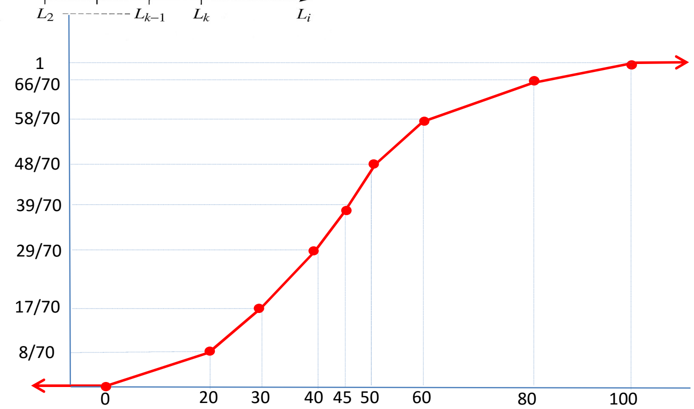
| $I_i$ | $n_i$ | $f_i$ | $F_i$ |
|---|---|---|---|
| $(e_0, e_1]$ | $n_1$ | $f_1$ | $F_1$ |
| $(e_1, e_2]$ | $n_2$ | $f_2$ | $F_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $(e_{i-1}, e_i]$ | $n_i$ | $f_i$ | $F_i$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $(e_{k-1}, e_k]$ | $n_k$ | $f_k$ | $F_k = 1$ |

$$F(x) = 0,\ x \le e_0;$$
$$F(e_i) = \sum_{j=1}^{i} f_j = F_i, \quad i = 1, \ldots, k;$$
$$F(x) = 1,\ x \ge e_k.$$

  En los puntos intermedios de cada intervalo, la función $F$ se aproxima mediante una recta. Así, la curva acumulativa, o curva de distribución, se define como la poligonal obtenida partiendo de cero y uniendo los puntos correspondientes a los extremos superiores de los intervalos y sus frecuencias relativas acumuladas (si se representan las frecuencias absolutas acumuladas, se obtiene una curva similar).




| $I_i$ | $n_i$ | $f_i$ | $F_i$ |
|---|---|---|---|
| $(0, 20]$ | 8 | $8/70$ | $8/70$ |
| $(20, 30]$ | 9 | $9/70$ | $17/70$ |
| $(30, 40]$ | 12 | $12/70$ | $29/70$ |
| $(40, 45]$ | 10 | $10/70$ | $39/70$ |
| $(45, 50]$ | 9 | $9/70$ | $48/70$ |
| $(50, 60]$ | 10 | $10/70$ | $58/70$ |
| $(60, 80]$ | 8 | $8/70$ | $66/70$ |
| $(80, 100]$ | 4 | $4/70$ | 1 |




**Nota:** Los diagramas que representan frecuencias acumuladas (tanto en variables discretas como continuas) suelen denominarse *diagramas integrales*, mientras que los que representan frecuencias sin acumular se denominan *diagramas diferenciales*.

# MEDIDAS DESCRIPTIVAS

*Resúmenes de los datos que reflejan la información contenida en ellos, facilitando su interpretación, así como la comparación con otros conjuntos de datos.*

**Propiedades de Yule** :

- *Estar definidas de manera objetiva.*
- *Tener un significado concreto.*
- *Usar todos los datos.*
- *Ser sencillas de calcular y prestarse fácilmente al cálculo algebraico.*
- *Ser poco sensibles a cambios en los valores extremos.*

## I: Medidas de posición (tendencia): *valores representativos del centro u otras partes de la distribución, que informan sobre la localización del conjunto de datos en la recta real.*

**Posición central**

- Medias (aritmética, geométrica, armónica, cuadrática)
- Mediana
- Moda

**Posición no central**

- Cuantiles

## II: Medidas de dispersión: *miden el grado de separación entre los datos y sirven también para medir la representatividad de las medidas de posición.*

**Absolutas** (con unidad de medida)

- Desv. absoluta media respecto de la media aritmética
- Desv. absoluta media respecto de la mediana
- Varianza y desviación típica
- Recorridos

**Relativas** (adimensionales)

- Coeficiente de variación de Pearson
- Índice de dispersión respecto a la mediana
- Coeficiente de apertura
- Recorridos relativos

## III: Medidas de forma: *informan sobre distintos aspectos de la forma de una distribución (del diagrama de barras, en el caso discreto, y del histograma en el caso continuo).*

**Asimetría**

- Coeficiente de Fisher
- Coeficientes de Pearson

**Curtosis**

- Coeficiente de Fisher
- Coeficiente de Kelley

# MEDIA ARITMÉTICA

**Suma de todos los datos dividida por el número total de datos**

Se requiere que los datos sean numéricos, y adquiere sentido sólo si éstos son de naturaleza aditiva

- **Variables discretas:** $\{(x_i, n_i);\ i = 1, \ldots, k\},\ n = \sum_{i=1}^{k} n_i,\quad f_i = \frac{n_i}{n}$
- **Variables continuas:** $\{((e_{i-1},\ e_i], n_i);\ i = 1, \ldots, k\},\ n = \sum_{i=1}^{k} n_i,\quad f_i = \frac{n_i}{n},\quad x_i = \frac{e_{i-1} + e_i}{2}$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

- *Si existen al menos dos valores distintos, la media aritmética está estrictamente comprendida entre los valores extremos*:

$$x_1 < x_k \Longrightarrow x_1 < \overline{x} < x_k.$$

- *La suma de las desviaciones de los datos respecto de su media aritmética es nula; por ello, la media aritmética suele interpretarse como el* **centro de gravedad** *de los datos*:

$$\sum_{i=1}^{k} n_i (x_i - \overline{x}) = 0.$$

- *Si se someten los datos a una transformación lineal afín, la media aritmética queda afectada por la misma transformación:*

$$y_i = a x_i + b,\quad i = 1, \ldots, k \ \Longrightarrow\ \overline{y} = a\overline{x} + b.$$

  - La media aritmética de una constante es la propia constante ($a = 0,\ b \in \mathbb{R}$).
  - Si se multiplican los datos por una constante (*cambio de escala*), la media aritmética queda multiplicada por la misma constante ($a \neq 0,\ b = 0$).
  - Si se suma una constante a todos los datos (*traslación, o cambio de origen*), la nueva media se obtiene sumando a la original la misma constante ($a = 1,\ b \neq 0$).

- *La media aritmética de las desviaciones cuadráticas respecto a la media aritmética es mínima:*

$$\sum_{i=1}^{k} f_i (x_i - \overline{x})^2 < \sum_{i=1}^{k} f_i (x_i - a)^2,\quad \forall a \neq \overline{x}.$$

# OTRAS MEDIAS

Valores numéricos o marcas de clase: $\{(x_i, n_i);\ i = 1, \ldots, k\}$, $n = \sum_{i=1}^{k} n_i$, $f_i = \dfrac{n_i}{n}$

- **Media geométrica:** $G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}} \longrightarrow \log G = \dfrac{\sum_{i=1}^{k} n_i \log x_i}{n} = \sum_{i=1}^{k} f_i \log x_i.$

  $\downarrow$

  Datos con efectos multiplicativos acumulativos (positivos)

- **Media armónica:** $H = \dfrac{n}{\dfrac{n_1}{x_1} + \dfrac{n_2}{x_2} + \ldots + \dfrac{n_k}{x_k}} \longrightarrow H^{-1} = \dfrac{\sum_{i=1}^{k} n_i x_i^{-1}}{n} = \sum_{i=1}^{k} f_i x_i^{-1}.$

  $\downarrow$

  Datos de magnitudes relativas (no nulos)

- **Media cuadrática:** $Q = \sqrt{\dfrac{\sum_{i=1}^{k} n_i x_i^2}{n}} \longrightarrow Q^2 = \dfrac{\sum_{i=1}^{k} n_i x_i^2}{n} = \sum_{i=1}^{k} f_i x_i^2.$

  $\downarrow$

  Datos con efectos cuadráticos sobre un total (elimina los efectos del signo)

$$x_i > 0,\ i = 1, \ldots, k \ y\ \exists x_i \neq x_j \longrightarrow \boxed{H < G < \overline{x} < Q}$$

# MEDIANA

**Valor o valores (modalidad o modalidades) que ocupan la posición, o posiciones centrales al ordenar los datos; esto es:**

- **al menos la mitad de los datos son menores o iguales que la mediana y**
- **al menos la mitad de los datos son mayores o iguales que la mediana.**

Sólo se requiere que los datos sean de naturaleza ordinal (no necesariamente numéricos)

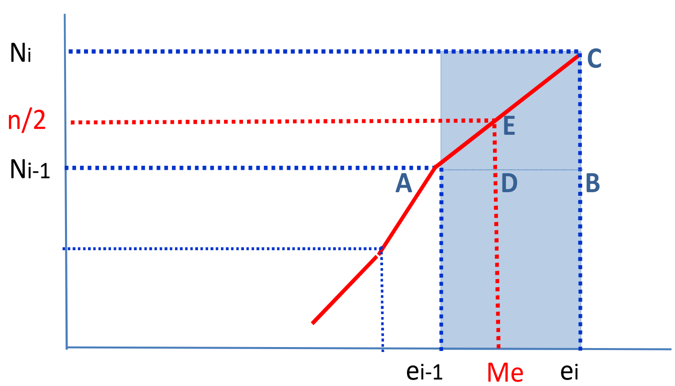
- **Variables discretas y atributos ordinales:** $\{(x_i, n_i);\ i = 1, \dots, k\}$, $n = \sum_{i=1}^{k} n_i$, $N_i = \sum_{j=1}^{i} n_j$

$$x_i \ / \ N_{i-1} < \frac{n}{2} \leq N_i \longrightarrow \begin{cases} N_i > \dfrac{n}{2} \ \Rightarrow \ Me = x_i. \\ N_i = \dfrac{n}{2} \ \Rightarrow \ Me = x_i, x_{i+1} \end{cases}$$

(en datos numéricos suele tomarse el punto medio)

- **Variables continuas agrupadas:** $\{((e_{i-1}, e_i], n_i);\ i = 1, \dots, k\}$, $n = \sum_{i=1}^{k} n_i$, $N_i = \sum_{j=1}^{i} n_j$

$$Intervalo\ mediano: (e_{i-1},\ e_i] \ / \ N_{i-1} < \frac{n}{2} \leq N_i \ \longrightarrow \ Me = e_{i-1} + \frac{n/2 - N_{i-1}}{n_i}(e_i - e_{i-1}).$$




$$\frac{AD}{AB} = \frac{DE}{BC} \longrightarrow \frac{Me - e_{i-1}}{e_i - e_{i-1}} = \frac{n/2 - N_{i-1}}{n_i}.$$

## Propiedades de la mediana:

- *Si se realizan cambios de escala y origen en los datos, la mediana queda afectada por el mismo cambio.*
- *La desviación absoluta media respecto a la mediana es mínima:*

$$\sum_{i=1}^{k} f_i |x_i - Me| < \sum_{i=1}^{k} f_i |x_i - a|, \quad \forall a \neq Me.$$

# MODA

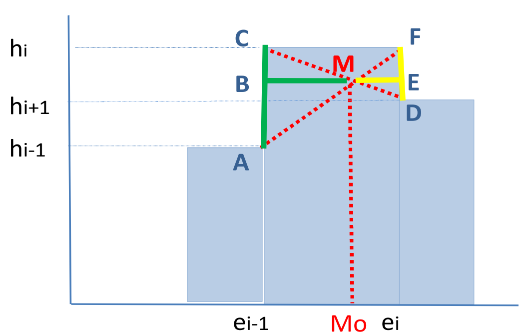
**Valor o valores (modalidad o modalidades) más frecuentes**

Puede calcularse para datos de cualquier tipo, basta una escala nominal

- **Variables discretas y atributos:** $\{(x_i, n_i);\ i = 1, \dots, k\} \longrightarrow Mo = x_i \ / \ n_i \geq n_j, \quad \forall j \neq i.$

- **Variables continuas agrupadas:** $\{((e_{i-1},\ e_i], n_i);\ i = 1, \dots, k\}, \quad a_i = e_i - e_{i-1}, \quad h_i = \dfrac{n_i}{a_i}$

$$\textit{Intervalo modal}: \ (e_{i-1},\ e_i] \ / \ h_i \geq h_j, \quad \forall j \neq i \ \longrightarrow \ Mo = e_{i-1} + \frac{h_i - h_{i-1}}{2h_i - h_{i-1} - h_{i+1}} a_i.$$




$$\frac{BM}{AC} = \frac{ME}{DF} \longrightarrow \frac{Mo - e_{i-1}}{h_i - h_{i-1}} = \frac{e_i - Mo}{h_i - h_{i+1}} = \frac{a_i - (Mo - e_{i-1})}{h_i - h_{i+1}}.$$

## Propiedades de la moda:

- *Para caracteres cuantitativos, si se realizan cambios de escala y origen en los datos, la moda queda afectada por el mismo cambio.*

# CUANTILES

$C_q,\ q \in (0,1)$: **Valor o valores (modalidad o modalidades) tales que:**

- **al menos el** $100q\,\%$ **de los datos** ($nq$ **datos) son menores o iguales que** $C_q$ **y**
- **al menos el** $100(1-q)\,\%$ **de los datos** ($n(1-q)$ **datos) son mayores o iguales que** $C_q$.

Sólo se requiere que los datos sean de naturaleza ordinal (no necesariamente numéricos)

- **Variables discretas y atributos ordinales:** $\{(x_i, n_i);\ i = 1, \ldots, k\}, \quad n = \sum_{i=1}^{k} n_i, \quad N_i = \sum_{j=1}^{i} n_j$

$$x_i \ /\ N_{i-1} < nq \leq N_i \longrightarrow \begin{cases} N_i > nq \ \Rightarrow \ C_q = x_i. \\ \\ N_i = nq \ \Rightarrow \ C_q = x_i, x_{i+1} \ \text{(en datos numéricos suele tomarse el punto medio)} \end{cases}$$

- **Variables continuas agrupadas:** $\{((e_{i-1}, e_i], n_i);\ i = 1, \ldots, k\}, \quad n = \sum_{i=1}^{k} n_i, \quad N_i = \sum_{j=1}^{i} n_j$

$$(e_{i-1},\ e_i] \ /\ N_{i-1} < nq \leq N_i \ \longrightarrow \ C_q = e_{i-1} + \frac{nq - N_{i-1}}{n_i}(e_i - e_{i-1}).$$

$$\downarrow$$

**Percentiles:** $P_r = C_{r/100},\ r = 1, \ldots, 99:$ *Valor o valores (modalidad o modalidades) tales que:*

- *al menos el* $r\,\%$ *de los datos* ($\frac{nr}{100}$ *datos) son menores o iguales que* $P_r$ *y*
- *al menos el* $(100-r)\,\%$ *de los datos* ($n\left(1 - \frac{r}{100}\right)$ *datos) son mayores o iguales que* $P_r$.

**Cuartiles:** $Q_1 = P_{25},\ Q_2 = P_{50} = M_e,\ Q_3 = P_{75}.$

**Deciles:** $D_1 = P_{10},\ D_2 = P_{20}, \ldots,\ D_9 = P_{90}.$

## CÁLCULO DE PERCENTILES

- **Variables discretas y atributos ordinales:** $\{(x_i, n_i);\ i = 1, \ldots, k\}, \quad n = \sum_{i=1}^{k} n_i, \quad N_i = \sum_{j=1}^{i} n_j$

$$x_i \ /\ N_{i-1} < \frac{nr}{100} \leq N_i \longrightarrow \begin{cases} N_i > \dfrac{nr}{100} \ \Rightarrow \ P_r = x_i. \\ \\ N_i = \dfrac{nr}{100} \ \Rightarrow \ P_r = x_i, x_{i+1} \ \text{(en datos numéricos suele tomarse el punto medio)} \end{cases}$$

- **Variables continuas agrupadas:** $\{((e_{i-1}, e_i], n_i);\ i = 1, \ldots, k\}, \quad n = \sum_{i=1}^{k} n_i, \quad N_i = \sum_{j=1}^{i} n_j$

$$(e_{i-1},\ e_i] \ /\ N_{i-1} < \frac{nr}{100} \leq N_i \ \longrightarrow \ P_r = e_{i-1} + \frac{\frac{nr}{100} - N_{i-1}}{n_i}(e_i - e_{i-1}).$$

# MEDIDAS DE DISPERSIÓN ABSOLUTAS

Valores numéricos o marcas de clase: $\{(x_i, n_i);\ i = 1, \ldots, k\}$, $\ n = \sum_{i=1}^{k} n_i$, $\ f_i = \dfrac{n_i}{n}$

▶ **Desviación absoluta media respecto de la media aritmética**: $D_{\overline{x}} = \sum_{i=1}^{k} f_i |x_i - \overline{x}|$.

▶ **Desviación absoluta media respecto de la mediana**: $D_{Me} = \sum_{i=1}^{k} f_i |x_i - Me|$.
↓
desviación absoluta media mínima ($\Longrightarrow D_{Me} < D_{\overline{x}}$ si $Me \neq \overline{x}$).

▶ **Varianza**

$$Var(X) = \sigma_x^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2$$

▶ **Desviación típica**

$$\sigma_x = \sqrt{\sum_{i=1}^{k} f_i (x_i - \overline{x})^2}$$

- $\sigma_x^2,\ \sigma_x \geq 0$ y $\sigma_x^2 = 0$ $(\sigma_x = 0)$ *si y sólo si todos los datos coinciden.*
- *Teorema de König* : $\sum_{i=1}^{k} f_i (x_i - a)^2 = \sum_{i=1}^{k} f_i (x_i - \overline{x})^2 + (a - \overline{x})^2$ $\left(\Rightarrow\ \sigma_x^2 = \sum_{i=1}^{k} f_i {x_i}^2 - \overline{x}^2\right)$.
- $\sigma_x^2$ *es la desviación cuadrática media mínima:* $\sum_{i=1}^{k} f_i (x_i - \overline{x})^2 < \sum_{i=1}^{k} f_i (x_i - a)^2$, $\ \forall a \neq \overline{x}$.
- $\sigma_x^2$ *está acotada por las desviaciones cuadráticas mínima y máxima:*

$$\min_{1 \leq i \leq k} (x_i - \overline{x})^2 \leq \sigma_x^2 \leq \max_{1 \leq i \leq k} (x_i - \overline{x})^2.$$

- *La varianza y la desviación típica son invariantes frente a traslaciones, pero se ven afectadas por cambios de escala en el siguiente sentido:*

$$y_i = a x_i + b,\ i = 1, \ldots, k \implies \sigma_y^2 = a^2 \sigma_x^2, \quad \sigma_y = |a| \sigma_x.$$

- $\overline{x} \neq Me \Rightarrow D_{Me} < D_{\overline{x}} < \sigma_x$.
- ***Desigualdad de Tchebychev*** : *el porcentaje de datos en cualquier intervalo de la forma* $(\overline{x} - k\sigma_x,\ \overline{x} + k\sigma_x)$ *es, como mínimo, el* $100\left(1 - \dfrac{1}{k^2}\right)\%$.
- ***Tipificación:*** $z_i = \dfrac{x_i - \overline{x}}{\sigma_x}$, $\ i = 1, \ldots, k \implies \overline{z} = 0, \quad \sigma_z^2 = 1$.

▶ **Recorrido (rango)**: $R = \max_{1 \leq i \leq k} x_i - \min_{1 \leq i \leq k} x_i$.

▶ **Recorrido (rango) intercuartílico**: $R_I = Q_3 - Q_1$.

# MEDIDAS DE DISPERSIÓN RELATIVAS

▶ **Coeficiente de variación de Pearson** : $CV_x = \dfrac{\sigma_x}{\overline{x}}$.

▶ **Indice de dispersión respecto a la mediana** : $V_{Me} = \dfrac{D_{Me}}{Me}$.

▷ *Coeficiente de apertura* : $C_A = \dfrac{\max\limits_{1\leq i\leq k} x_i}{\min\limits_{1\leq i\leq k} x_i}$.

▷ *Recorrido relativo* : $R_R = \dfrac{R}{\overline{x}} = \dfrac{\max\limits_{1\leq i\leq k} x_i - \min\limits_{1\leq i\leq k} x_i}{\overline{x}}$.

▷ *Recorrido semi-intercuartílico* : $R_{SI} = \dfrac{R_I}{Q_3 + Q_1} = \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$.

# MOMENTOS

Valores numéricos o marcas de clase : $\{(x_i, n_i);\ i = 1, \ldots, k\}, \quad n = \sum_{i=1}^{k} n_i, \quad f_i = \frac{n_i}{n}$

**Momentos no centrados**
*(centrados en el origen)*

$$m_r = \frac{\sum_{i=1}^{k} n_i x_i^r}{n} = \sum_{i=1}^{k} f_i x_i^r, \quad r \in \mathbb{N}$$

- $m_1 = \overline{x}$
- $m_2 = \sigma_x^2 + \overline{x}^2$

**Momentos centrados**
*(centrados en media)*

$$\mu_r = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x})^r}{n} = \sum_{i=1}^{k} f_i (x_i - \overline{x})^r, \quad r \in \mathbb{N}$$

- $\mu_1 = 0$
- $\mu_2 = \sigma_x^2$

- *Momentos centrados en función de los no centrados:*

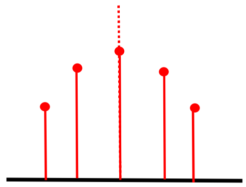
$$\mu_r = \sum_{t=0}^{r} (-1)^t \binom{r}{t} m_1^t \, m_{r-t} \longrightarrow \begin{cases} \mu_2 = m_2 - m_1^2 \\ \mu_3 = m_3 - 3\, m_2\, m_1 + 2\, m_1^3 \\ \mu_4 = m_4 + 4\, m_3\, m_1 + 6\, m_1^2\, m_2 - 3\, m_1^4 \\ \vdots \end{cases}$$

- *Momentos no centrados en función de los centrados y de $m_1$:*

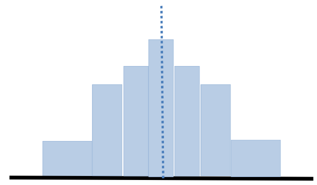
$$m_r = \sum_{t=0}^{r} \binom{r}{t} m_1^t \, \mu_{r-t} \longrightarrow \begin{cases} m_2 = \mu_2 + m_1^2 \\ m_3 = \mu_3 + 3\, \mu_2\, m_1 + m_1^3 \\ m_4 = \mu_4 + 4\, \mu_3\, m_1 + 6\, \mu_2\, m_1^2 + m_1^4 \\ \vdots \end{cases}$$

# MEDIDAS DE ASIMETRÍA

**Distribución simétrica:** *Para cada dato de la forma $\overline{x} - c$ existe otro de la forma $\overline{x} + c$:*



$\overline{x} = Me = Mo$



$\overline{x} = Me = Mo$

**Propiedades de las distribuciones simétricas:**

- *La media, la mediana y la moda (si es única) coinciden con el centro de simetría.*
- *Los momentos centrados de orden impar son nulos.*

▶ **Coeficiente de asimetría de Fisher** : $\gamma_1 = \dfrac{\mu_3}{\sigma_x^3} = \sum_{i=1}^{k} f_i \left( \dfrac{x_i - \overline{x}}{\sigma_x} \right)^3.$

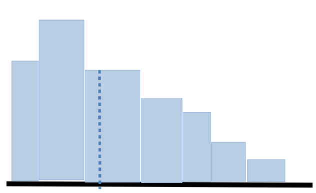
- $\gamma_1 = 0$ en distribuciones simétricas
- $\gamma_1 > 0 \rightarrow$ Asimetría a la derecha
- $\gamma_1 < 0 \rightarrow$ Asimetría a la izquierda

**Propiedades** : $\gamma_1$ *es adimensional, invariante por traslaciones e invariante, salvo el signo, por cambios de escala:*

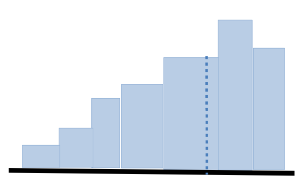
$$\gamma_{1,ax+b} = \pm \gamma_{1,x} \text{ (según el signo de } a).$$

▶ **Coeficientes de asimetría de Pearson:**

$$\left.\begin{array}{l} \textit{Distribuciones unimodales}: \ A_p = \dfrac{\overline{x} - Mo}{\sigma_x} \\ \\ \textit{Distribuciones plurimodales}: \ A_p^* = \dfrac{3(\overline{x} - Me)}{\sigma_x} \end{array}\right\} \begin{array}{l} = 0 \text{ en distribuciones simétricas} \\ \\ > 0 \rightarrow \text{Asimetría a la derecha} \\ \\ < 0 \rightarrow \text{Asimetría a la izquierda} \end{array}$$



Asimetría por la derecha



Asimetría por la izquierda

# MEDIDA DE CURTOSIS (APUNTAMIENTO)

*Distribuciones unimodales simétricas o moderadamente asimétricas*

▶ **Coeficiente de curtosis de Fisher** : $\gamma_2 = \dfrac{\mu_4}{\sigma_x^4} - 3 = \displaystyle\sum_{i=1}^{k} f_i \left( \frac{x_i - \overline{x}}{\sigma_x} \right)^4 - 3$

- $\gamma_2 = 0 \to$ Distribución mesocúrtica
- $\gamma_2 > 0 \to$ Distribución leptocúrtica
- $\gamma_2 < 0 \to$ Distribución platicúrtica

**Propiedades** : $\gamma_2$ *es adimensional e invariante por traslaciones y por cambios de escala:*

$$\gamma_{2,ax+b} = \gamma_{2,x}.$$

▶ **Coeficiente de curtosis de Kelley** : $K = \dfrac{1}{2}\dfrac{Q_3 - Q_1}{D_9 - D_1} - 0.263.$

- $K = 0 \to$ Distribución mesocúrtica
- $K > 0 \to$ Distribución leptocúrtica
- $K < 0 \to$ Distribución platicúrtica