# Estadística I

## Profesor de teoría:

Andrés M. Alonso

Despacho 10.1.32
E. Mail: andres.alonso@uc3m.es
Web: www.est.uc3m.es/amalonso
Web docente: http://www.est.uc3m.es/amalonso/esp/docencia.html

## Profesores de práctica:

- Concepción Molina (Grupo 30)
- Javier Reques (Grupo 31)
- Natalia Fojo (Grupo 32)

# Estadística I

## Temario de la asignatura

- Análisis de datos univariantes.
- Análisis de datos bivariantes.
- Probabilidad.
- Variables aleatorias multidimensionales.
- Distribuciones muestrales.
- Estimación puntual.
- Estimación por intervalos.
- Contrastes de hipótesis.

# Estadística I

## Bibliografía básica

Newbold, P., Carlson, W.L. y Thorne, B. (2008) *Estadística para Administración y Economía*, Editorial Prentice Hall, Madrid.

Peña, D. (2001) *Fundamentos de Estadística*, Alianza Editorial, Madrid.

Peña, D. y Romo, J. (1997) *Introducción a la Estadística para las Ciencias Sociales*, Editorial McGraw Hill, Madrid.

## Bibliografía complementaria

Levin, R.I. y Rubin, D.S. (2004) *Estadística para Administración y Economía*, Editorial Prentice Hall, Madrid.

Newbold, P. (2001) *Estadística para los Negocios y la Economía*, Editorial Prentice Hall, Madrid.

Martín Pliego, F.J. (2004) *Introducción a la Estadística Económica y Empresarial*, Thomson Editores, Madrid.

Moore, D.S. (1998) *Estadística Aplicada Básica*, Editorial Antoni Bosch, Barcelona.

# Tema 1: Análisis de datos univariantes

1. Introducción

2. Representaciones y gráficos
   - Tablas de frecuencias
   - Diagrama de barras, Diagrama de sectores, Histograma, y Diagrama de caja

3. Resumen numérico
   - Medidas de localización
   - Medidas de dispersión
   - Medidas de forma

## Lecturas recomendadas:

- Capítulos 1 al 3 del libro de Newbold, Carlson, y Thorne (2008).
- Capítulos 1 y 2 del libro de Peña (2001).
- Capítulos 1 al 5 del libro de Peña y Romo (1997).

# Objetivos del tema

Despúes de estudiar este tema, se podrá:

- Explicar las siguientes definiciones básicas:
  - Población frente a Muestra
  - Parámetro frente Estadístico
  - Estadística Descriptiva frente a Estadística Inferencial
- Describir un muestreo aleatorio

# Objetivos del tema

Despúes de estudiar este tema, se podrá:

- Identificar tipos de datos y niveles de medidas.
- Crear e interpretar gráficos para describir variables categóricas:
  - Distribución de frecuencias, frecuencias absolutas y relativas, diagrama de barras, diagrama de tartas.
- Crear e interpretar gráficos para describir variables numéricas:
  - Distribución de frecuencias, frecuencias absolutas y relativas, frecuencias acumuladas absolutas y relativas, histograma, diagrama de cajas.

## Objetivos del tema

Despúes de estudiar este tema, se podrá:

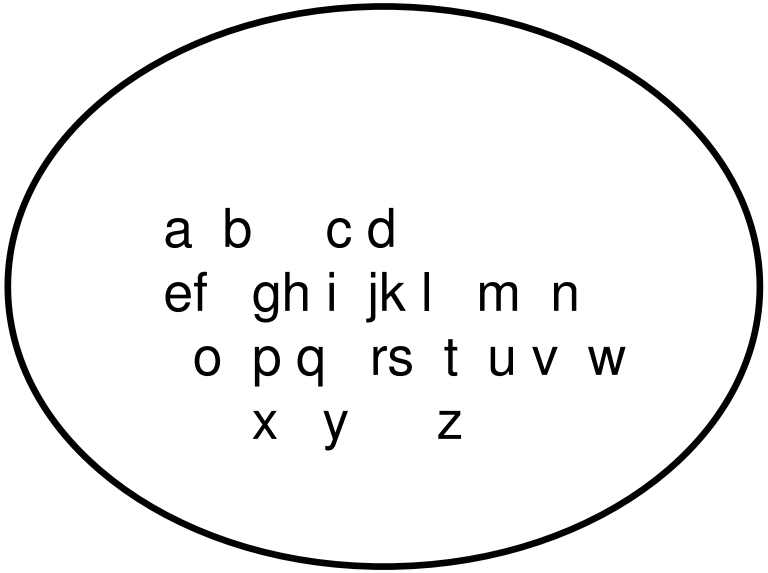
- Calcular e interpretar la media, mediana, y moda de un conjunto de datos.
- Calcular el rango, varianza, desviación estándar, y coeficiente de variación e interpretar dichos valores.

# Definiciones básicas

- Una población es la colección **completa** de **todos** los elementos de interés que se investigan.
    - $N$ representa el tamaño de la población
- Una muestra es un subconjunto observado de la población
    - $n$ representa el tamaño muestral
- Un parámetro es una característica específica de una población (fija)
- Un estadístico es una característica específica de una muestra (puede variar entre diferentes muestras)

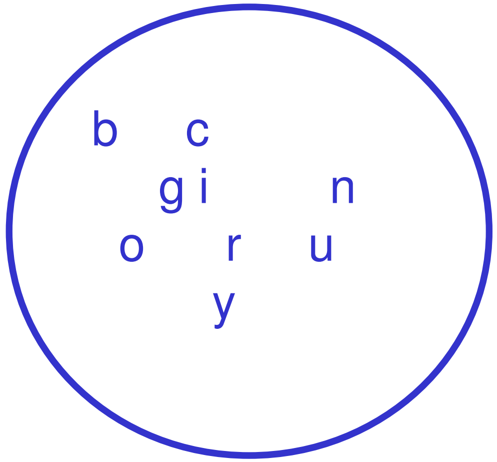
# Población frente a muestra

## Población




Valores calculados usando todos los elementos de la población se llaman **parámetros**

## Muestra




Valores calculados usando los elementos de la muestra se llaman **estadísticos**

# Ejemplos de poblaciones

- Nombres de todos los votantes de la Unión Europea
- Ingresos de todas las familias que viven en Getafe
- Indice anual de las acciones en la bolsa de Londres
- Nota media de todos los estudiantes de la universidad

# Muestreo aleatorio

El muestreo aleatorio simple es un procedimiento en el que

- Cada miembro de la población se elige al azar,
- Cada miembro de la población tiene la misma posibilidad de ser elegido,
- Cada posible muestra de $n$ elementos tiene la misma probabilidad de ser elegida

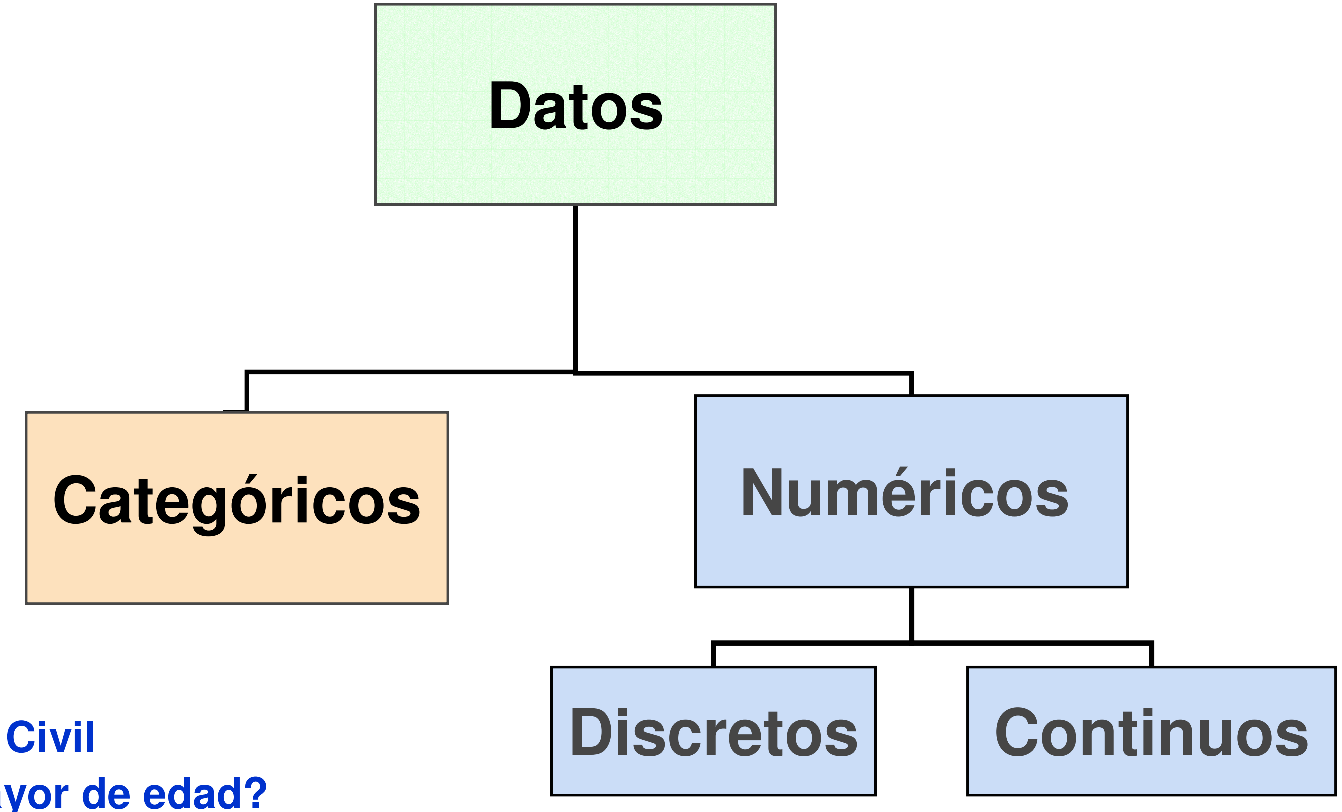
La muestra resultante se denomina *muestra aleatoria simple*

## Estadística descriptiva e Inferencia estadística

Dos ramas de la estadística:

- Estadística Descriptiva
  - Recoger, resumir y procesar los datos para transformar dichos datos en *información*
- Inferencia Estadística
  - Proporciona las bases para predicciones y estimaciones para convertir la *información* en *conocimiento*.

# Tipos de variables o datos




**Ejemplos:**

- Estado Civil
- ¿Es mayor de edad?
- Color de Ojos

**(Categorías definidas o grupos)**

**Ejemplos:**

- Número de hijos
- Defectos por hora

**(recuento de elementos)**

**Ejemplos:**

- Peso
- Voltaje

**(Características Medibles)**

# Niveles de medida

| | | |
|---|---|---|
| Diferencias entre medidas, dado que existen ceros verdaderos | **Datos de Razón** | |
| | ↑ | Datos Cuantitativos |
| Diferencias entre medidas pero no existe un cero verdadero | **Datos de Intervalo** | |
| | ↑ | |
| Categorías ordenadas (rangos, orden, o escalamientos) | **Datos Ordinales** | |
| | ↑ | Datos Cualitativos |
| Categorías (no hay orden o dirección) | **Datos Nominales** | |

## Representaciones y gráficos

- Datos *en bruto* en forma de listas no son fáciles de usar para tomar decisiones
- Se necesita algún tipo de organización:
  - Tablas
  - Gráficos
- El tipo de gráfico depende de la variable que se va a resumir

# Representaciones y gráficos

## Técnicas que se presentan en este tema

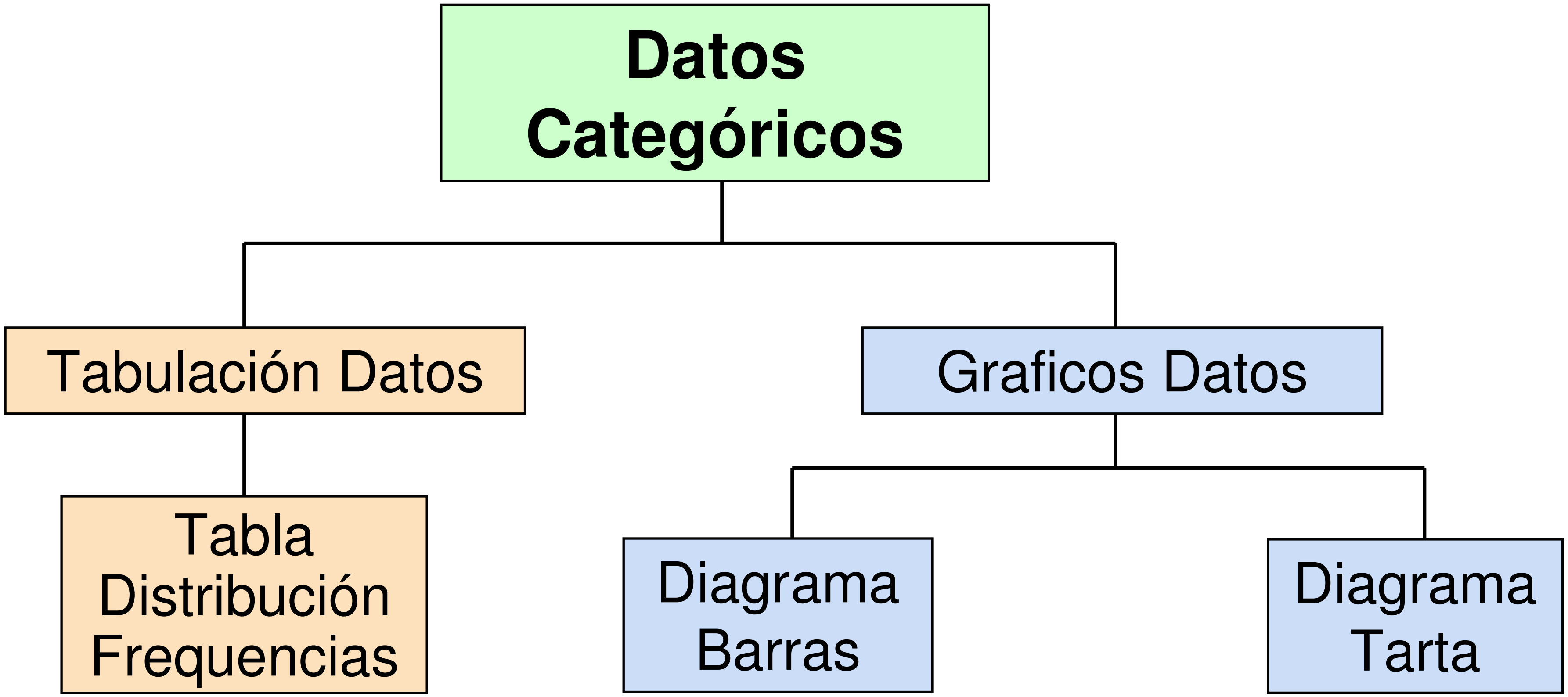
**Variables Categóricas**

- Distribución Frecuencias
- Diagrama de Barras
- Diagrama de Tarta

**Variables Numéricas**

- Distribución Frecuencias
- Histograma
- Diagrama de Caja

# Tablas y gráficos para variables categóricas



# Tabla de frecuencias

## Resumir datos por categorías

*Ejemplo*: Pacientes de un Hospital según Servicio

| Indice Clase<br>$i = 1,\ldots,k$ | Servicio Hospital<br>Clase<br>$c_i$ | Número de Pacientes<br>Frequencia Absoluta<br>$n_i$ = número de observaciones clase $c_i$ | Proporción de Pacientes<br>Frequencia Relativa<br>$f_i = n_i / n$ |
|---|---|---|---|
| 1 | Cardiología | 1052 | 0.12 |
| 2 | Emergencias | 2245 | 0.25 |
| 3 | *UCI* | 340 | 0.04 |
| 4 | Maternidad | 552 | 0.06 |
| 5 (=$k$) | Cirugía | 4630 | 0.53 |
| | | $n_1 + n_2 + \ldots + n_k = n = 8819$ | $f_1 + \ldots + f_k = 1.00$ |

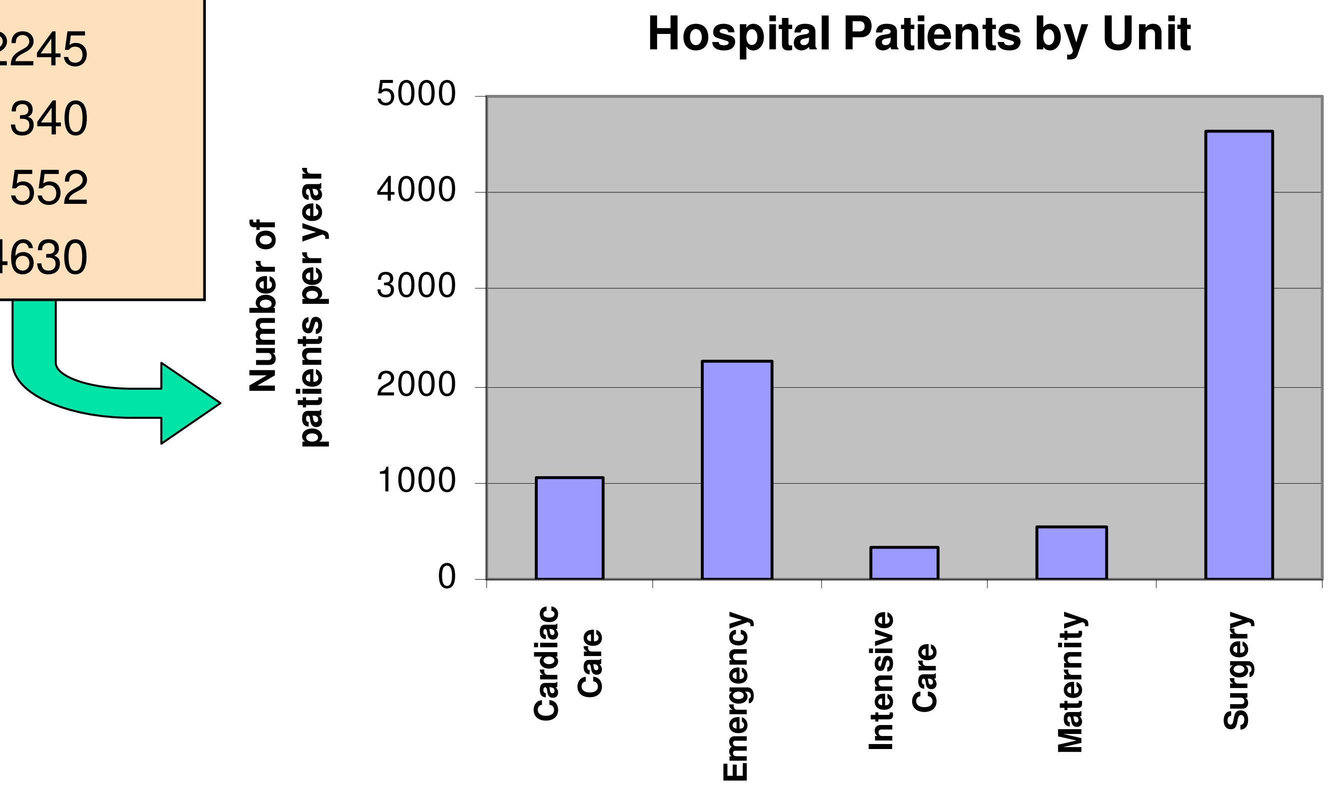
## Diagrama de Barras y de Sectores

- Los Diagramas de Barras y los Diagramas de Sectores o Tartas se usan a menudo para datos cualitativos (categóricos)

- La altura de la barra, o el tamaño de la porción de tarta, muestran la frecuencia o porcentaje de cada categoría

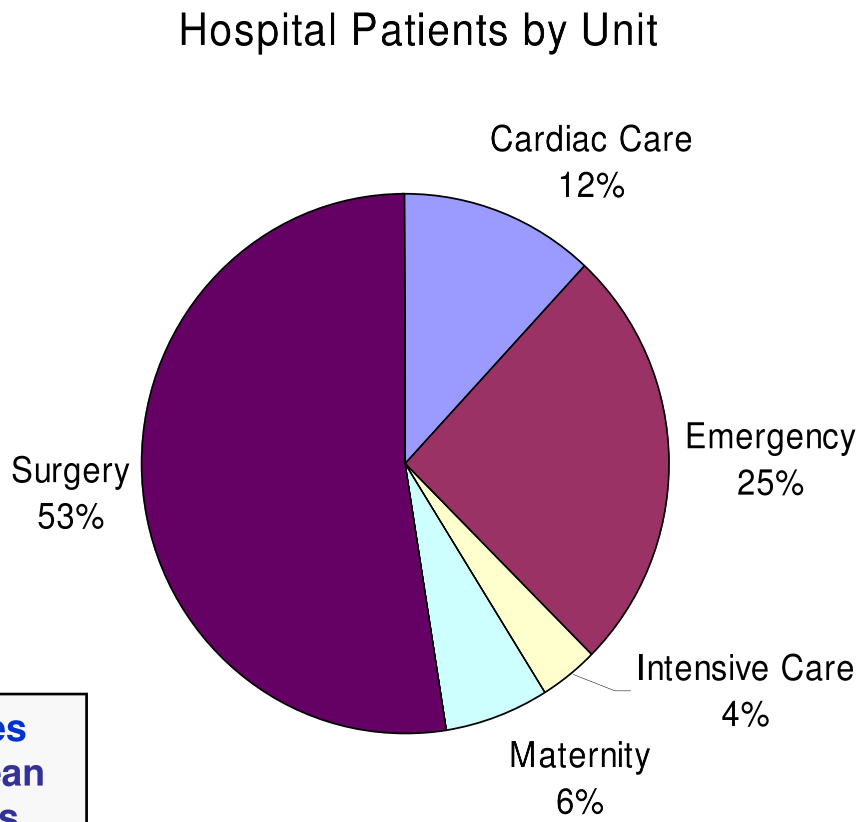
# Ejemplo de Diagrama de Barras

| Hospital Unidad | Número Pacientes |
|---|---|
| Cardiac Care | 1052 |
| Emergency | 2245 |
| Intensive Care | 340 |
| Maternity | 552 |
| Surgery | 4630 |



# Ejemplo de Diagrama de Sectores

| Hospital Unidad | Numero Pacientes | % de Total |
|---|---|---|
| Cardiac Care | 1052 | 11.93 |
| Emergency | 2245 | 25.46 |
| Intensive Care | 340 | 3.86 |
| Maternity | 552 | 6.26 |
| Surgery | 4630 | 52.50 |




**(Porcentajes se redondean al valor más cercano)**

# Tablas y gráficos para variables cuantitativas

**Datos Numéricos**

**Distribuciones de Frecuencias y Distribuciones Acumuladas**

**Histogramas**

## Distribución de frecuencias

¿Qué es una Distribución de Frecuencias?

- Una distribución de frecuencias es una tabla
- que contiene agrupamientos en clases (categorías o rangos en donde *caen* los datos)
- y las frecuencias correspondientes con las que se presentan los datos en cada clase o categoría

# ¿Por qué usar tablas de frecuencias?

- Una distribución de frecuencias es una manera de resumir los datos.
- La distribución condensa la lista de datos *en bruto* de una forma más útil que
- permite una interpretación visual rápida de los datos
- y permite la comparación con otros conjuntos de datos

# Intervalos y extremos de clase

- Cada clase de agrupamiento tiene, generalmente, la misma anchura.
- Determinar la anchura de cada intervalo por:

$$A = \text{anchura de intervalo} = \frac{\text{Número mayor} - \text{Número menor}}{\text{Número deseado de intervalos}}$$

- Usar al menos 5, pero no más de 15-20 intervalos
- Intervalos **nunca** solapan.
- Se redondea la anchura de los intervalos para obtener los extremos de más facil manejo

## Ejemplo de distribución de frecuencias

Ejemplo: un fabricante de *aislamientos* selecciona al azar 20 días de invierno y recoge las temperaturas máximas diarias:

**24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27**

# Ejemplo de distribución de frecuencias

- Ordenar los datos *en bruto* en orden ascendente :
  **12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58.**
- Calcular el rango: **58 - 12 = 46**
- Seleccionar número de clases: **5** (usualmente entre 5 y 15)
- Calcular anchura de intervalos: **10** (46/5 por lo que se redondea)
- Determinar extremos de intervalos: **10 pero menos que 20, 20 pero menos que 30, . . . , 60 pero menos que 70**
- Contar las observaciones y asignarlas a las clases.

# Ejemplo de distribución de frecuencias

**Datos ordenados:**

**12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58**

| Intervalos | Frecuencias | Freq. Relativas | Porcentaje |
|---|---|---|---|
| 10 *y menos que* 20 | **3** | **.15** | **15** |
| 20 *y menos que* 30 | **6** | **.30** | **30** |
| 30 *y menos que* 40 | **5** | **.25** | **25** |
| 40 *y menos que* 50 | **4** | **.20** | **20** |
| 50 *y menos que* 60 | **2** | **.10** | **10** |
| **Total** | **20** | **1.00** | **100** |

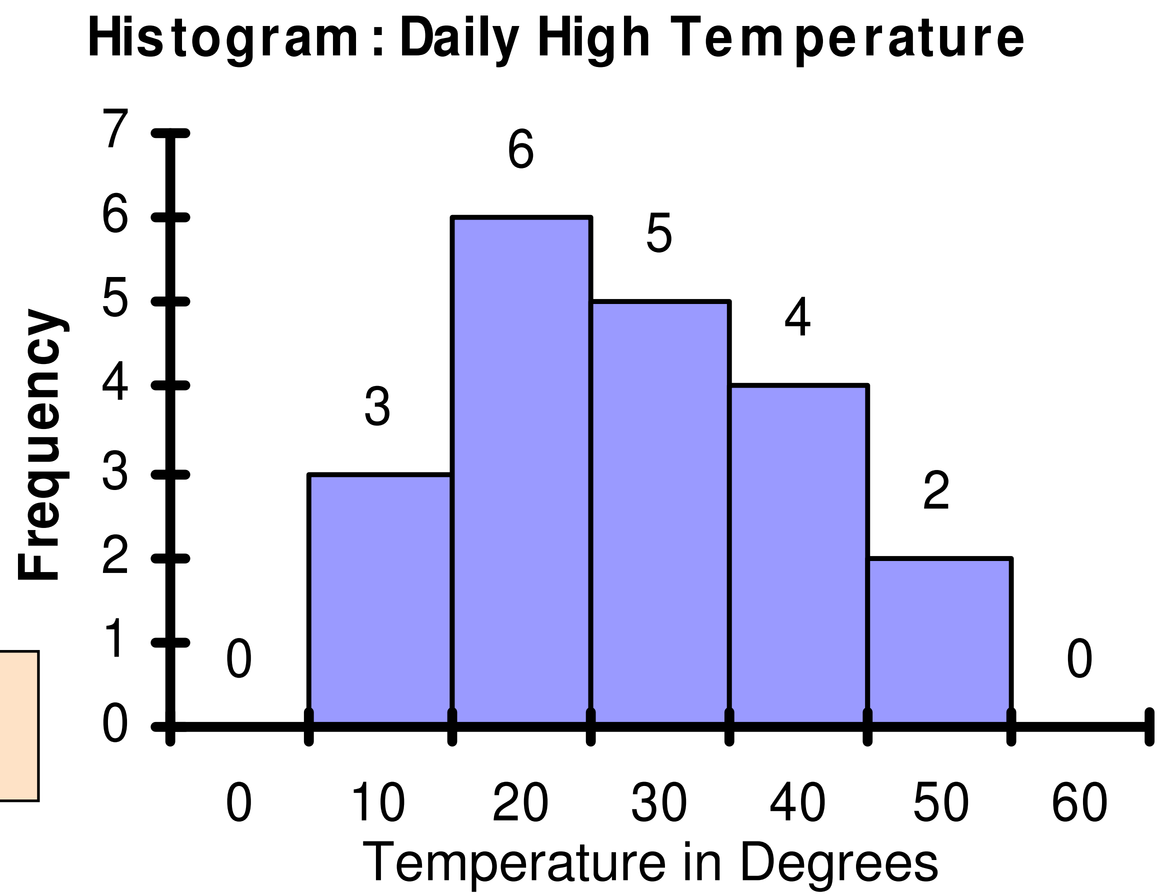
# Histograma

- El gráfico de los datos en una distribución de frecuencias se llama un **histograma**.
- Los **extremos de intervalos** aparecen en el eje horizontal.
- El eje vertical representa **frecuencias, frecuencias relativas, ó porcentajes**.
- Se usan barras de alturas adecuadas para representar el número de observaciones dentro de cada clase.

# Ejemplo de Histograma

| Intervalo | Frecuencia |
| --- | --- |
| 10 y menos que 20 | 3 |
| 20 y menos que 30 | 6 |
| 30 y menos que 40 | 5 |
| 40 y menos que 50 | 4 |
| 50 y menos que 60 | 2 |

(Sin *huecos* entre barras)

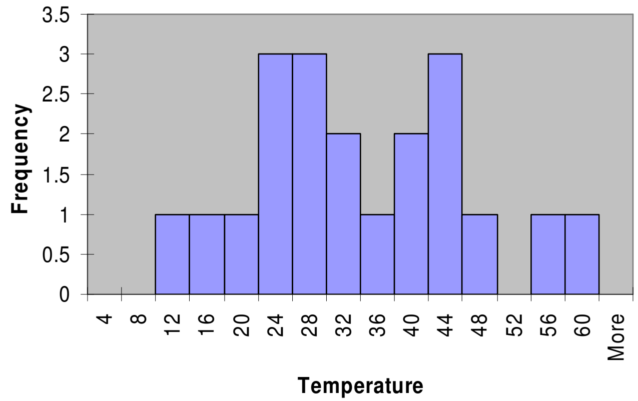


## ¿Cómo agrupar los datos?

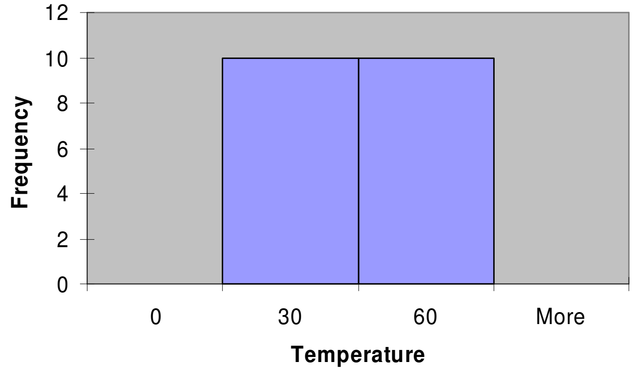
- ¿Cuantas clases deben usarse?
  - Si $n$ no es demasiado grande tomar $\sqrt{n}$, en caso contrario tomar $1+3.22 \ln(n)$
  - A menudo se responde por *prueba y error*, sujeto al juicio del investigador
  - El objetivo es crear una distribución que no sea ni demasiado *dentada* ni demasiado *en bloques*
  - El objetivo es mostrar apropiadamente el patrón de variación de los datos.

# ¿Cuantos intervalos de clase?

- **Muchos (Intervalos de clase Estrechos)**
  - Puede dar lugar a una distribución dentada con huecos de clases vacías
  - Puede ocultar cómo varía la frecuencia entre las clases




- **Pocos (Intervalos de clase Anchos)**
  - Puede comprimir mucho la variación y originar una distribución en bloque.
  - Puede oscurecer patrones importantes de variación.



# Distribución de frecuencias acumuladas

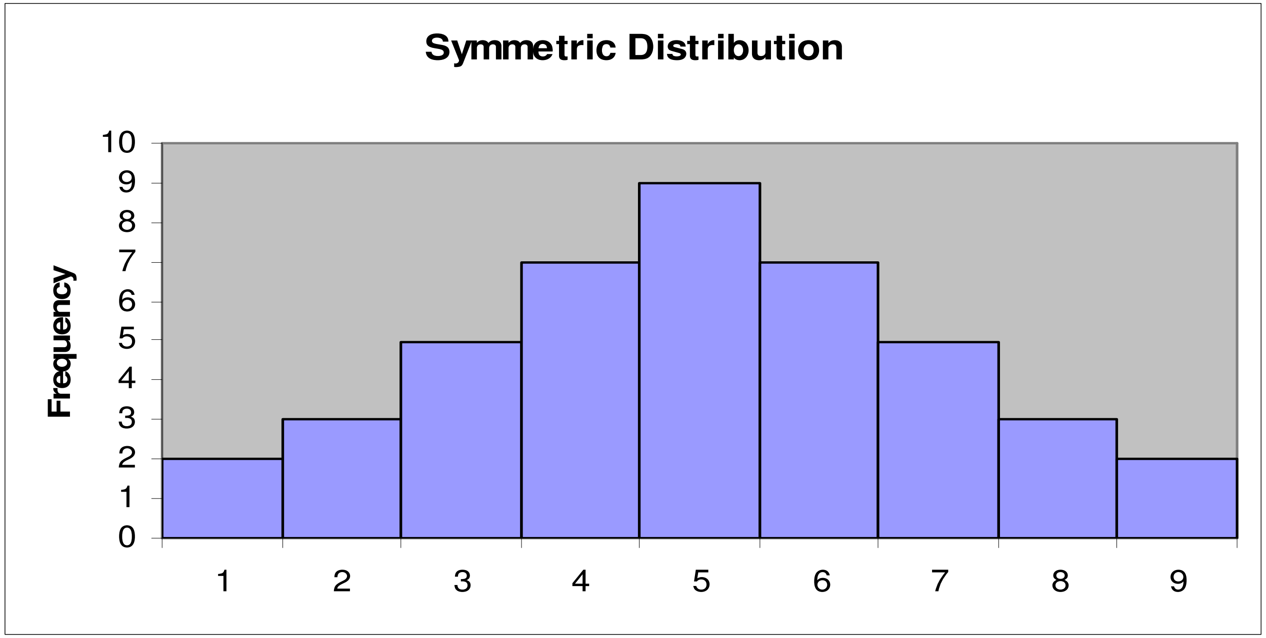
**Datos ordenados:**

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

| Clase | Frecuencia | Porcentaje | Frecuencia Acumulada | Porcentaje Acumulada |
|---|---|---|---|---|
| 10 y menos que 20 | 3 | 15 | 3 | 15 |
| 20 y menos que 30 | 6 | 30 | 9 | 45 |
| 30 y menos que 40 | 5 | 25 | 14 | 70 |
| 40 y menos que 50 | 4 | 20 | 18 | 90 |
| 50 y menos que 60 | 2 | 10 | 20 | 100 |
| Total | 20 | 100 | | |

# Forma de la distribución

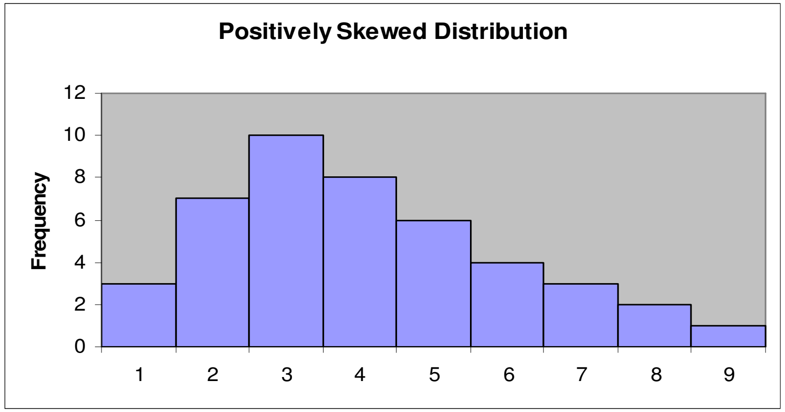
- La forma de la distribución se dice que es ***simétrica*** si las observaciones están equilibradas, o distribuidas simétricamente respecto al centro.



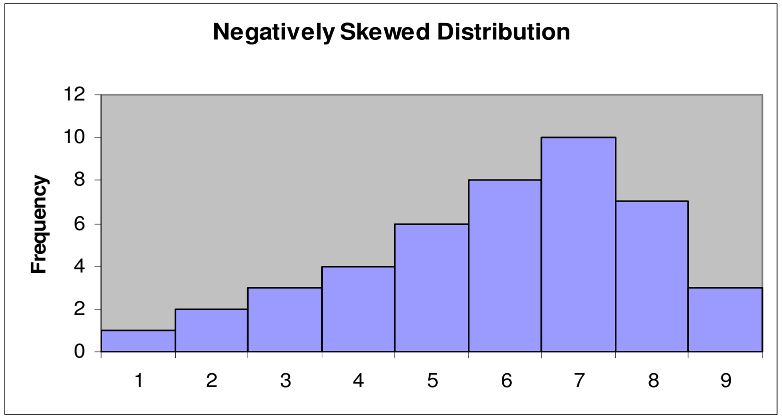
# Forma de la distribución

- La forma de la distribución se dice que es ***asimétrica*** si las observaciones **NO** están equilibradas, o distribuidas simétricamente respecto al centro.

Una distribución asimétrica positiva (asimétrica a la derecha) tiene una cola que se extiende a la derecha en dirección de los valores positivos.




Una distribución asimétrica negativa (asimétrica a la izquierda) tiene una cola que se extiende a la izquierda en dirección de los valores negativos.



# Resumen numérico

- **Descripción numérica de datos**
  - **Tendencia Central**
    - Media Aritmética
    - Mediana
    - Moda
  - **Variación**
    - Rango
    - Rango Intercuartílico
    - Varianza
    - Desviación estándar
    - Coeficiente de Variación

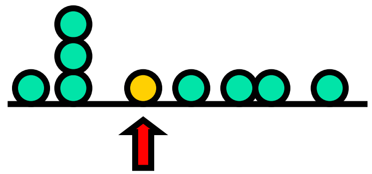
# Medidas de tendencia central

**Tendencia Central**

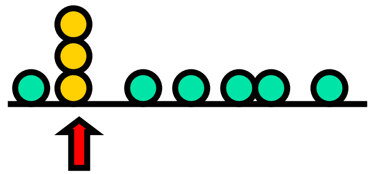
| Media | Mediana | Moda |
| --- | --- | --- |

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

Media Aritmética



Punto central de los valores ordenados



Valor observado más frecuente

## Media aritmética

- La media aritmética (media) es la medida más común de tendencia central

  – Para una población de *N* valores:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Valores Población

Tamaño Población

  – Para una muestra de *n* valores:

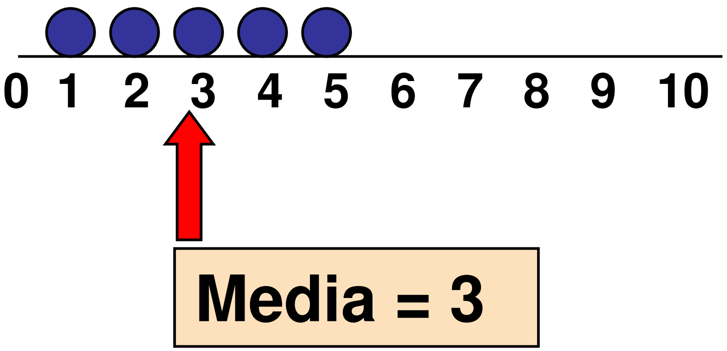
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Valores observados

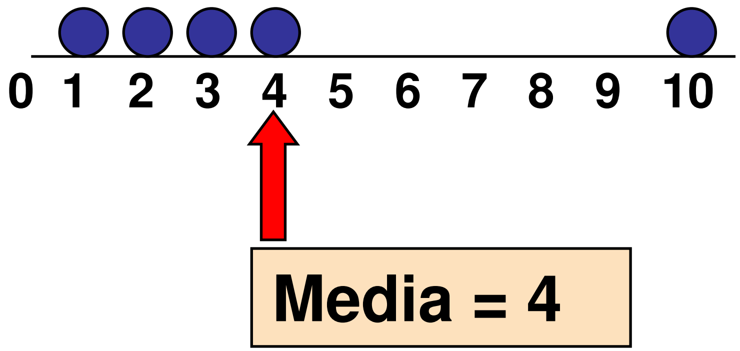
Tamaño muestra

# Media aritmética

- Es la medida más común de tendencia central
- Es una medida de fácil cálculo
- Afectada por valores extremos (*outliers*)



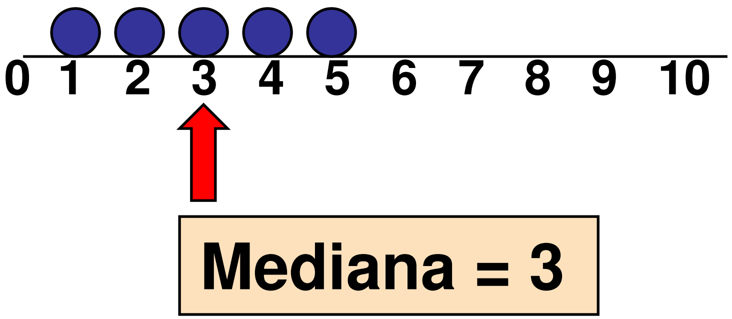

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

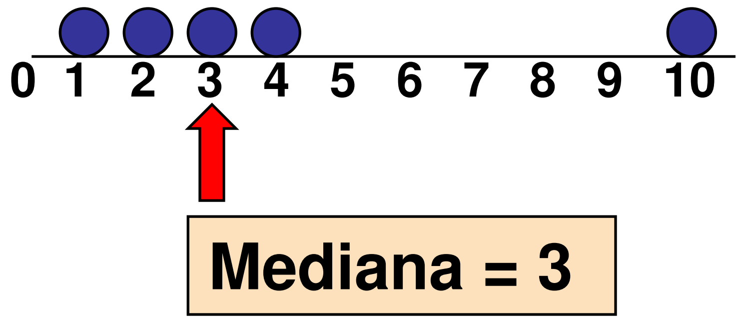



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

# Mediana

- En una lista ordenada, la mediana es valor central (50% por encima, 50% por debajo)




No resulta afectada por valores extremos

# Cálculo de la mediana

- La localización de la mediana:

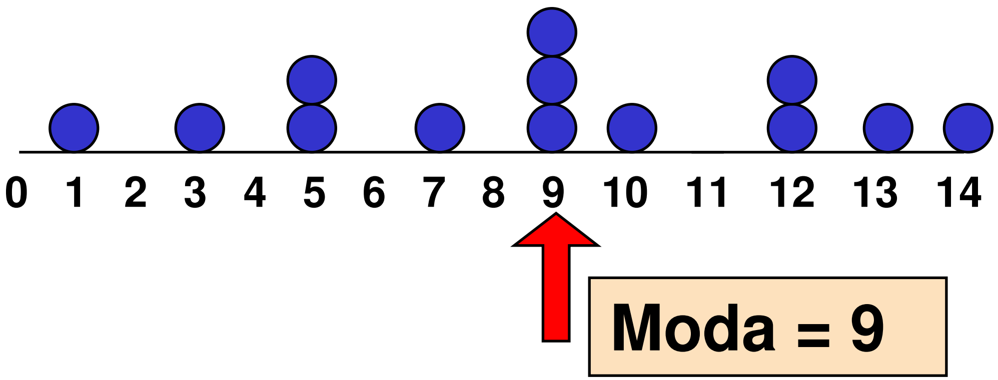
$$\text{Posición Mediana} = \frac{n+1}{2} \text{ posición en los datos ordenados}$$

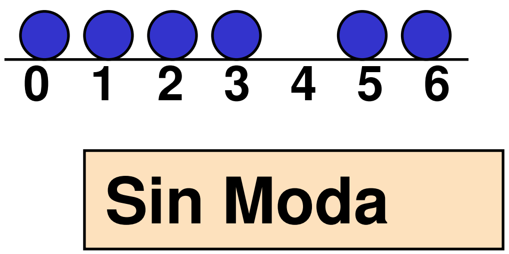
  – Si el número de valores es impar, la mediana es la observación central
  – Si el número de valores es par, la mediana es la media de las dos observaciones centrales

- Nótese que $\frac{n+1}{2}$ no es el *valor* de la mediana, sólo es la *posición* de la mediana en los datos ordenados

# Moda

- Es una medida de tendencia central
- Valor que aparece más en la muestra
- No afectada por valores extremos
- Usada para valores numéricos o categóricos
- Puede no haber una moda
- Puede haber varias modas



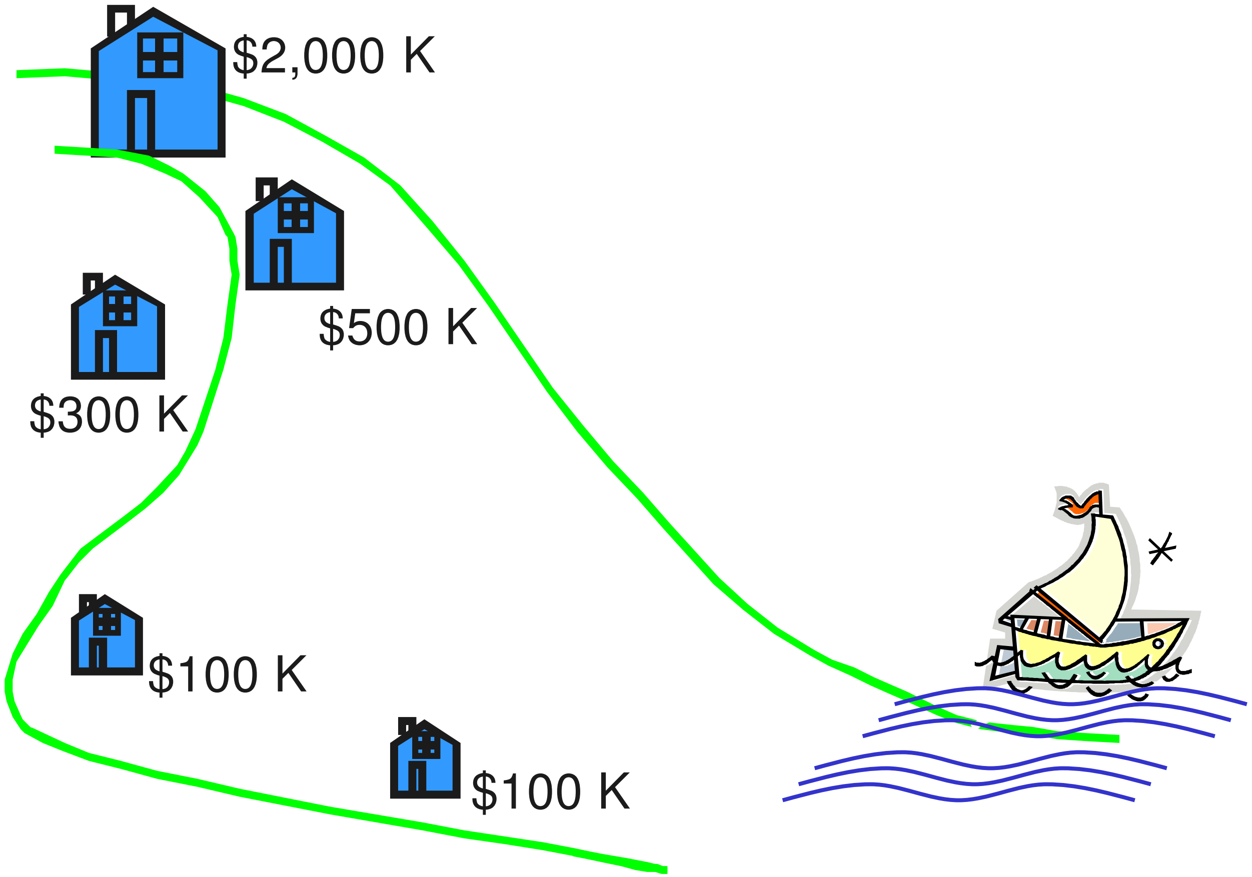



## Ejemplo

- Cinco casas en una colina cerca de la playa

**Precios Casas:**

**€2.000.000**
**500.000**
**300.000**
**100.000**
**100.000**



# Ejemplo

**Precios Casas:**

| |
|---|
| €2.000.000 |
| 500.000 |
| 300.000 |
| 100.000 |
| 100.000 |
| Suma **3.000.000** |

- **Media:** (€3.000.000/5)
  = **€600.000**

- **Mediana:** valor medio de los datos ordenados
  = **€300.000**

- **Moda:** valor más frecuente
  = **€100.000**

## ¿Cual es la mejor medida de centralidad?

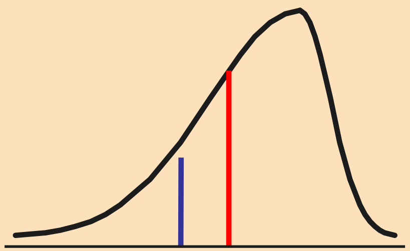
- **Media:** se usa generalmente, salvo que existan valores extremos (*outliers*).
- En ese caso se usa la **mediana**, porque no es sensible a valores extremos.
  - *Ejemplo*: Mediana de los precios de inmuebles para una región: es menos sensible a outliers.

# Forma de la distribución

- Describe cómo se distribuyen los datos
- Medidas de **forma**
  - Simétrica o asimétrica

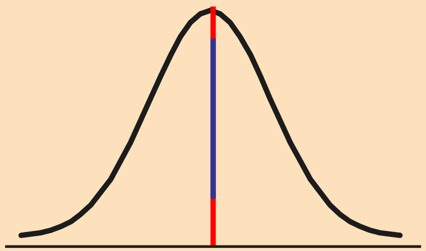
**Asim. Izquierda**

Media < Mediana



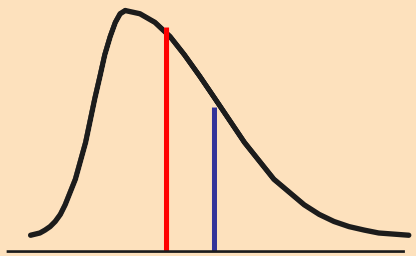
**Simétrica**

Media = Mediana



**Asim. Derecha**

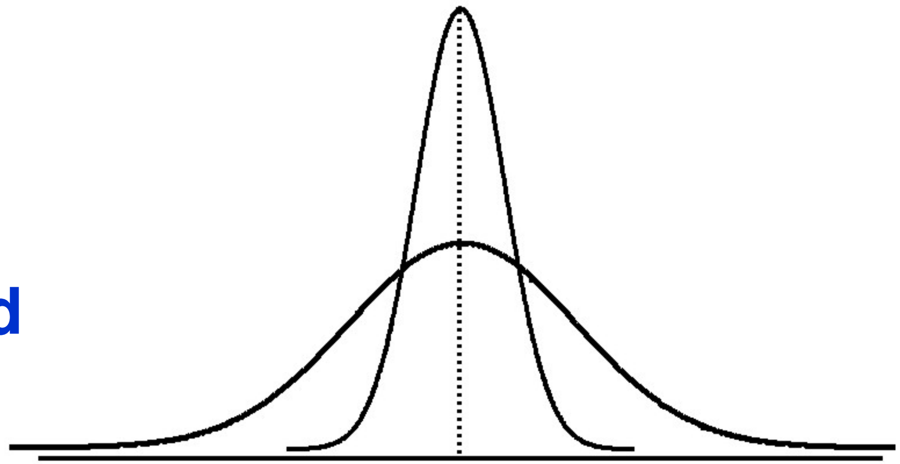
Mediana < Media



# Medidas de variación

**Variación**

- **Rango**
- **Rango Intercuartílico**
- **Varianza**
- **Desviación estándar**
- **Coeficiente de Variación**



- Las medidas de variación dan información sobre la **dispersión** o **variabilidad** de los datos.

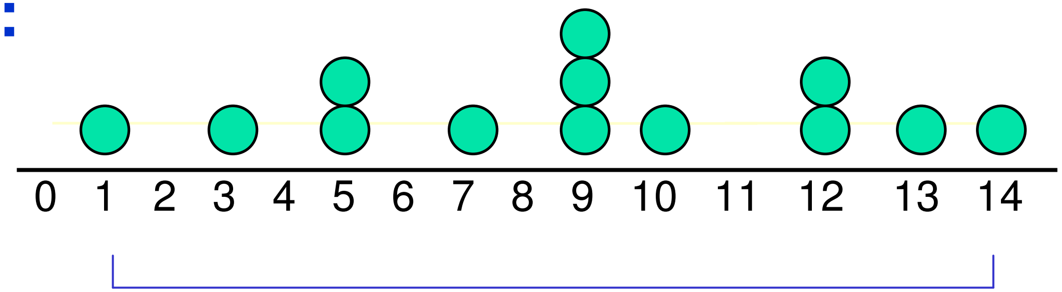
**Mismo centro, diferente variación**

# Rango

- Medida más simple de variación
- Diferencia entre la mayor y la menor de las observaciones:

$$\text{Rango} = X_{mayor} - X_{menor}$$

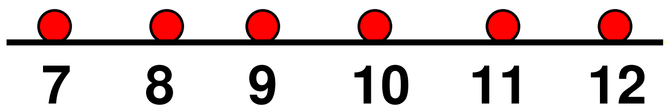
**Ejemplo:**



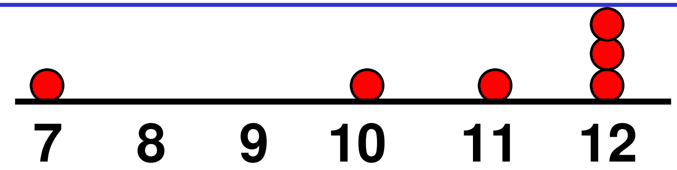

**Rango = 14 - 1 = 13**

# Desventajas del rango

- Ignora el modo en el que se distribuyen los datos




**Rango = 12 - 7 = 5**




**Rango = 12 - 7 = 5**

- Muy sensible a *outliers*

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,5

**Rango = 5 - 1 = 4**

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120

**Rango = 120 - 1 = 119**

## Rango intercuartílico

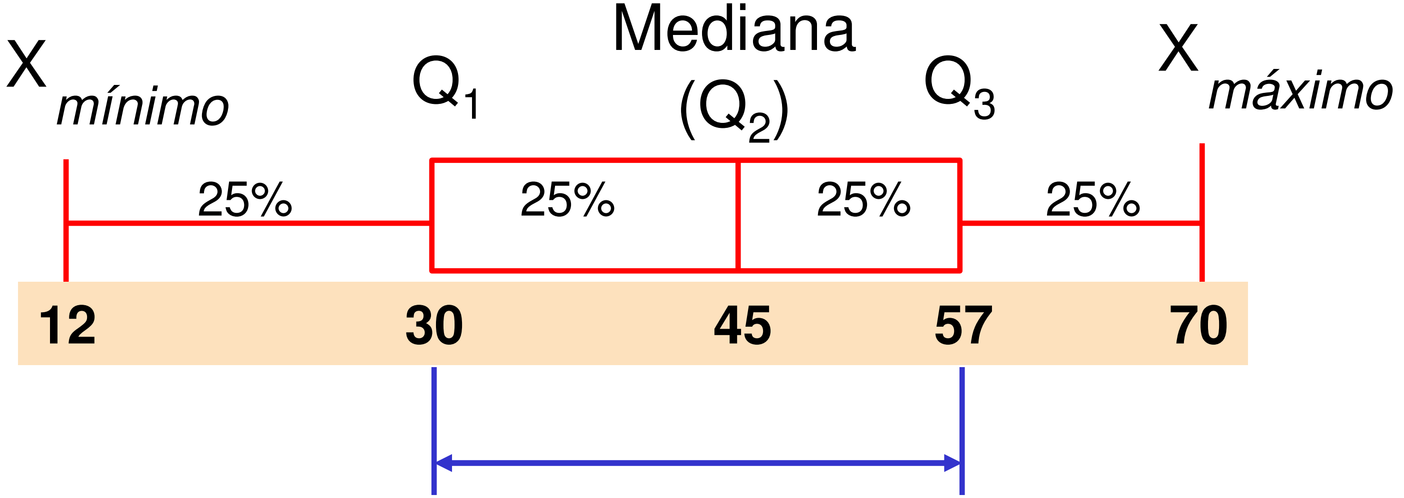
- Se pueden eliminar algunos problemas de *outliers* usando el **rango intercuartílico**
- Elimina valores muy grandes y muy pequeños calculando el rango de la *parte central* formada por el 50% de los datos

- Rango Intercuartílico = 3$^{er}$ cuartil – 1$^{er}$ cuartil

$$IQR = Q_3 - Q_1$$

# Rango intercuartílico

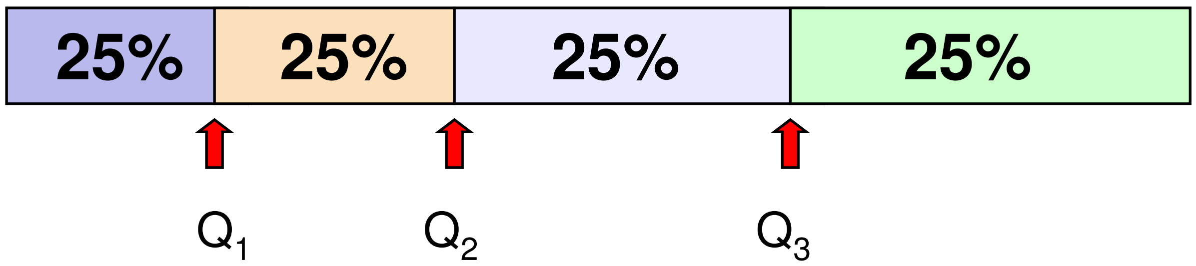
Ejemplo:




Rango Intercuartílico
$= 57 – 30 = 27$

# Cuartiles

- Cuartiles dividen los datos ordenados en 4 segmentos con igual número de valores por segmento




- Primer cuartil, $Q_1$, es el valor tal que el 25% de las observaciones son menores y el 75% son mayores
- $Q_2$ es la mediana (50% son menores, 50% son mayores)
- Sólo el 25% de las observaciones son mayores que el tercer cuartil

# Cálculo de los cuartiles

- Calcular un cuartil determinando el valor en la posición adecuada en los datos ordenados:

- Posición primer cuartil : $Q_1 = 0.25(n+1)$

- Posición segundo cuartil: $Q_2 = 0.50(n+1)$

- Posición tercer cuartil: $Q_3 = 0.75(n+1)$

donde $n$ es el número de valores observados

## Cálculo de los cuartiles

- *Ejemplo*: Calcular el primer cuartil

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Datos Muestrales Ordenados:** | **11** | **12** | **13** | **16** | **16** | **17** | **18** | **21** | **22** |

$Q_1$ = está en la posición 0.25(9+1) = 2.5 de los datos ordenados

Así, se usa el valor intermedio entre los valores 2º y 3ero :

$$\boxed{Q_1 = 12.5}$$

# Varianza poblacional

- Media de las desviaciones al cuadrado de los valores a la media

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

donde $\mu$ = media población

$N$ = tamaño población

$x_i$ = $i^{esimo}$ valor de la variable x

## Varianza muestral

- Promedio de las desviaciones al cuadrado de los valores a la media

  - **Varianza Muestral:**

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

donde $\bar{x}$ = media aritmética

$n$ = tamaño muestral

$x_i$ = $i^{esimo}$ valor de la variable x

## Desviación estándar poblacional

- Medida de variación más comúnmente usada
- Muestra la variación alrededor de la media
- Tiene las mismas unidades de medida que los datos originales

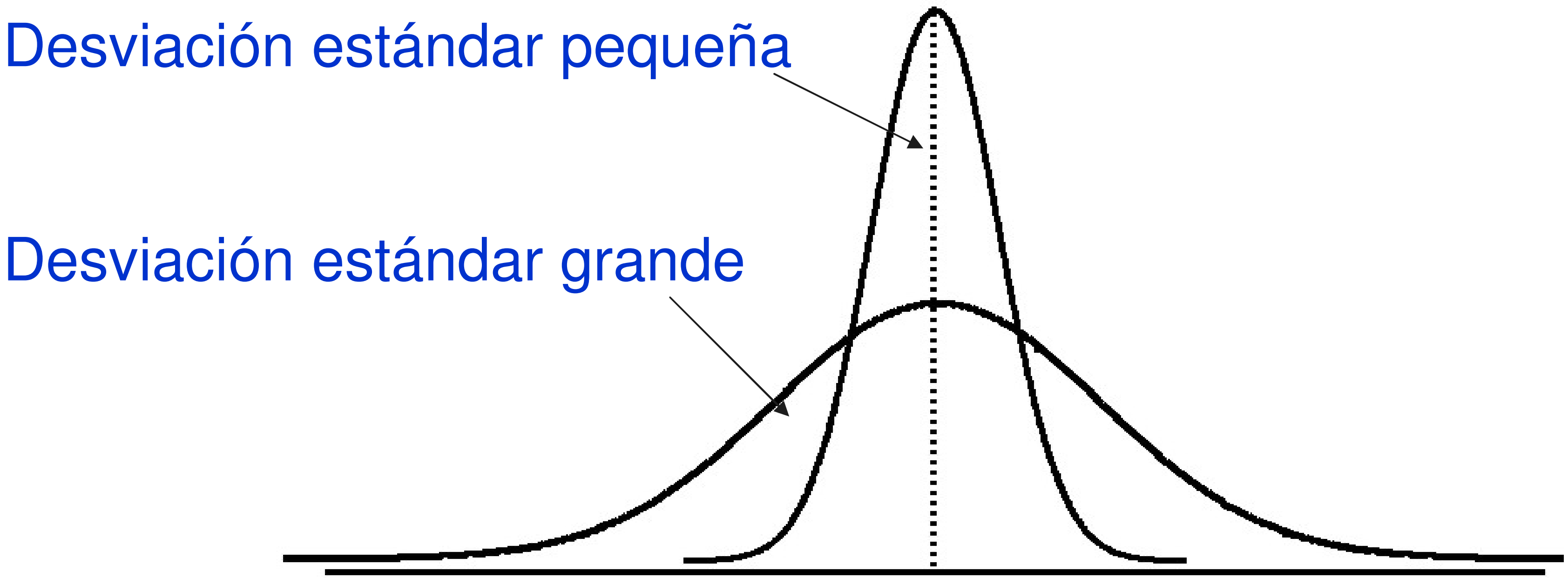
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

# Desviación estándar muestral

- La medida de variación usada más común
- Muestra la variación respecto a la media
- Tiene las mismas unidades de medida que los datos originales

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

# Medida de variación



# Ejemplo

Datos Muestrales ($x_i$):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 10 | 12 | 14 | 15 | 17 | 18 | 18 | 24 |

$n = 8$ Media $= \overline{x} = 16$

$$s = \sqrt{\frac{(10 - \overline{x})^2 + (12 - \overline{x})^2 + (14 - \overline{x})^2 + \cdots + (24 - \overline{x})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \cdots + (24 - 16)^2}{7}}$$

$$= \sqrt{\frac{126}{7}} = \boxed{4.2426}$$

Medida del *promedio* de la dispersión alrededor de la media

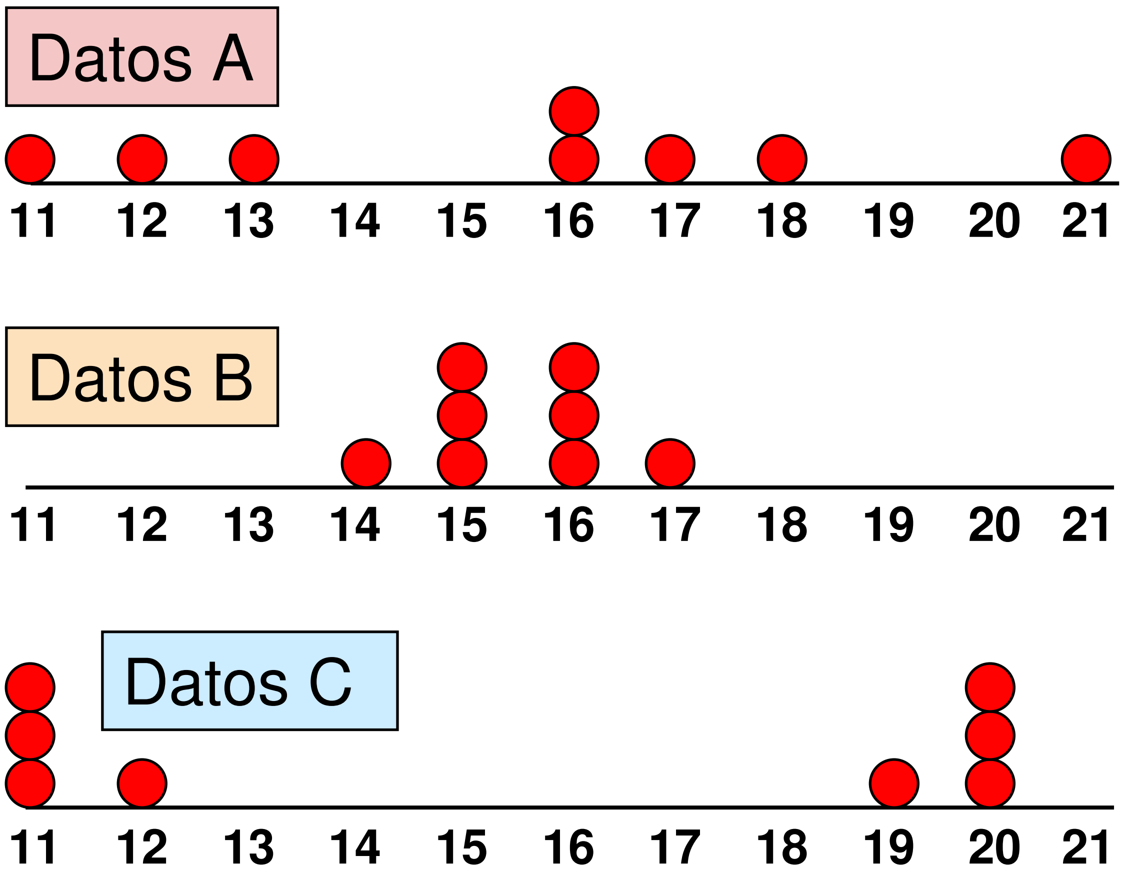
# Cálculo de la desviación estándar

- SC = Suma de Cuadrados:

$$SC(x) = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}}$$

# Comparación de desviaciones estándar




Media = 15.5
s = 3.122

Media = 15.5
s = 0.866

Media = 15.5
s = 4.275

## Ventajas de la varianza y de la desviación estándar

- Se usan todos los valores del conjunto de datos en los cálculos.
- A los valores alejados de la media se les asigna un peso extra (porque las desviaciones a la media se elevan al cuadrado)

# Coeficiente de variación

- Medida de la variación relativa
- Se expresa en porcentaje (%)
- Muestra la variación relativa respecto a la media
- Se puede usar para comparar dos o más conjuntos de datos, medidos en diferentes unidades

$$CV = \left( \frac{s}{\bar{X}} \right) \cdot 100\%$$

# Comparación de coeficientes de variación

- *Stock* A:
  - Precio medio último año = €50
  - Desviación estándar = €5

$$CV_A = \left( \frac{s}{\bar{x}} \right) \cdot 100\% = \frac{€5}{€50} \cdot 100\% = 10\%$$

- *Stock* B:
  - Precio medio último año = €100
  - Desviación estándar = €5

$$CV_B = \left( \frac{s}{\bar{x}} \right) \cdot 100\% = \frac{€5}{€100} \cdot 100\% = 5\%$$

Ambos stocks tienen la misma desviación estándar, pero el stock B es menos variable en relación a su precio