



# Estadística Descriptiva

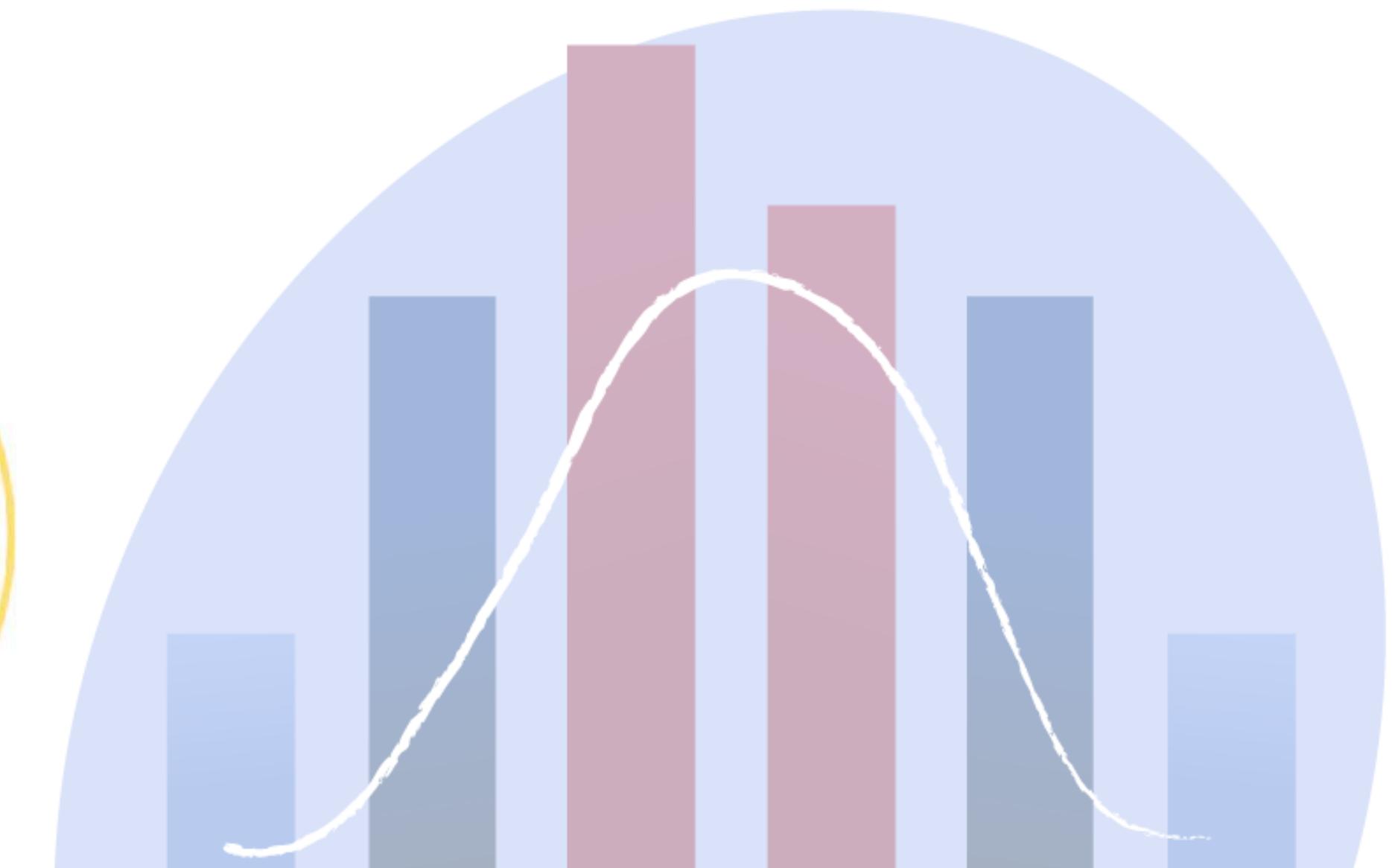
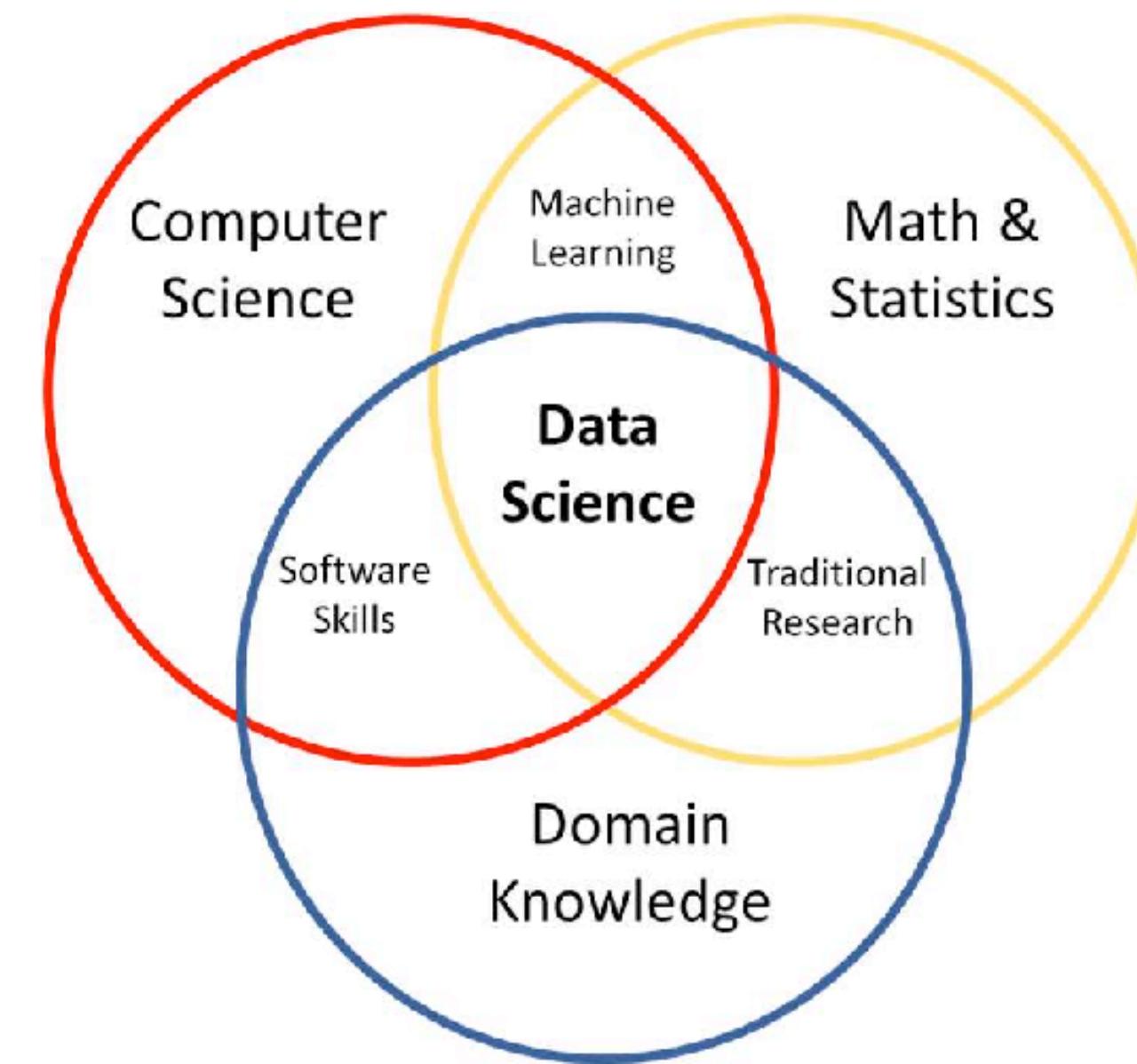
Propedéutico de la Maestría en Ciencia de Datos.  
CUCEA, Universidad de Guadalajara. 2024B

Guadalajara, Jal. Junio de 2025

# Estadística Descriptiva

## Overview

- Reunir y clasificar
- Organizar
- Resumir y analizar
- Interpretar



Se introducen métodos numéricos y gráficos para describir y mostrar datos de una muestra o población. Se aplican métodos gráficos para mostrar e interpretar datos, además de calcular medidas de tendencia central y de dispersión, e identifica valores atípicos en los datos con la finalidad de resumir y evaluar la información.

# Estadística Descriptiva

## Overview

- Utilizada para resumir y describir datos.
- Definir y recolectar información para describir variables de interés en una población o una muestra.
- Caracterizar los datos: tablas, gráficos o herramientas numéricas.
- Identificar patrones en los datos.
- Presentar, visualizar y exteriorizar la información.

# ¿Qué son los datos?

- Las observaciones recopiladas que se hace sobre un tema.
- ¿Son sólo números?. Pueden ser números, etiquetas, o incluso números que actúan como etiquetas (por ejemplo, tu número de estudiante).
- Son inútiles si no se les asigna un contexto. Para esto nos hacemos las siguientes preguntas: ¿quién?, ¿qué?, ¿Dónde?, ¿Cuándo? y ¿por qué?

# Five W's

## Quién?

- Nos dice los casos o individuos de los cuáles hemos recopilado información.
- Pueden ser participantes, o en el caso de objetos inanimados unidades experimentales.
- Normalmente se les llama observaciones a los valores de los datos.
- También se puede referir a quien creó la información.

# Five W's

## Qué?

- Las características recolectadas en cada observación se llaman **variables**. Éstas deben mencionar QUÉ es lo que se está midiendo. Éstas pueden ser clasificadas en variables cualitativas (categóricas) y cuantitativas.
  - Cualitativos
    - (Categóricos)
  - Datos
    - Cuantitativos
      - Discretos
      - Continuos

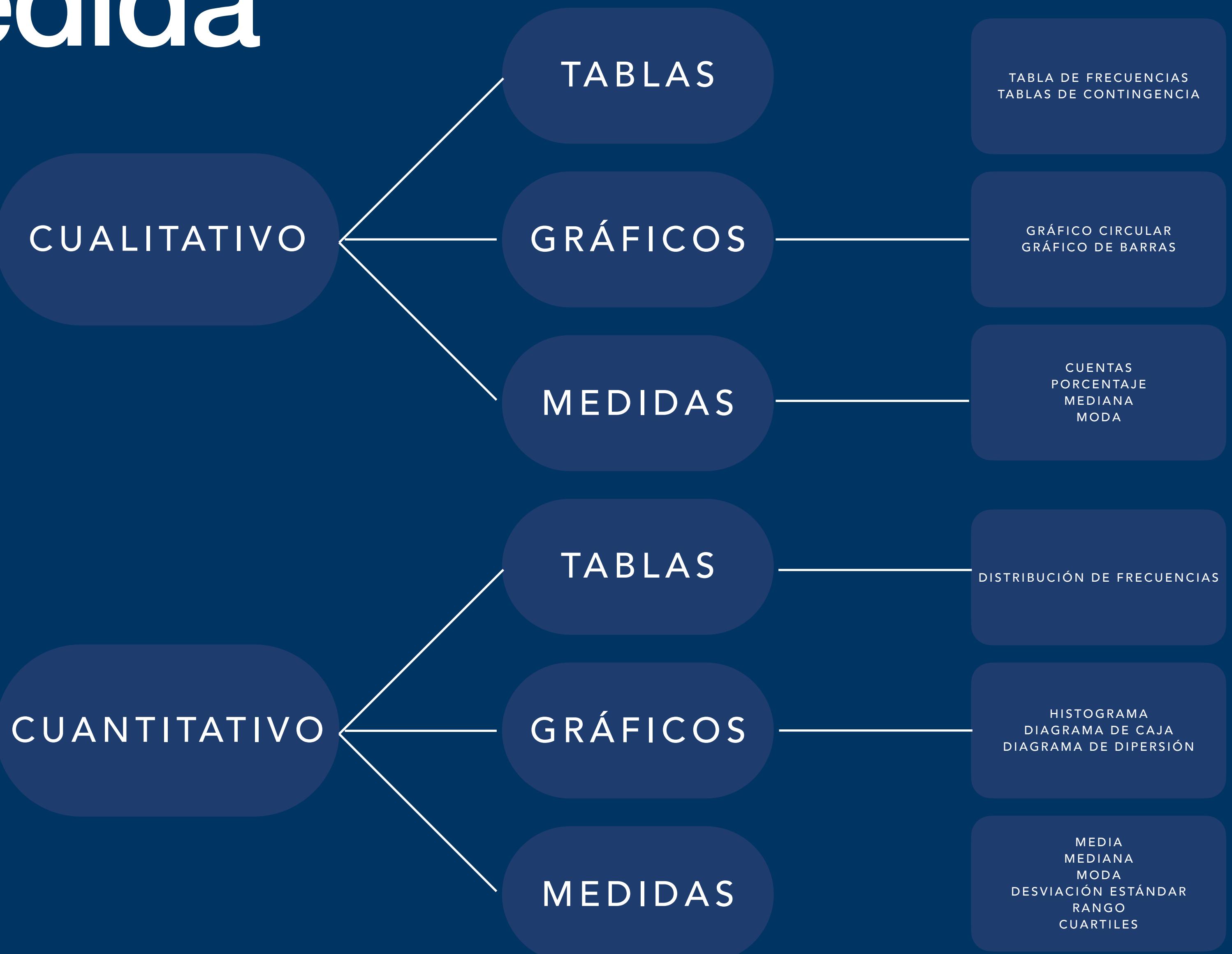
# Five W's

## ¿Dónde?, ¿Cuándo?, ¿Cómo?, ¿Por qué?

- CUÁNDO y DÓNDE nos mencionan información útil sobre el contexto de los datos. Algunos datos son más relevantes dentro de una época o lugar en específico. También se refiere a cuándo y donde la información fue archivada, o con acceso.
- El CÓMO los datos son registrados o medidos nos ayuda a entender la relevancia de los mismos. Datos mal tomados o con un método dudoso no tienen mucha relevancia.
- El POR QUÉ nos ayuda también a entender el contexto de los datos. Por qué existen estos datos. Datos que no son necesarios pueden eliminarse.

# Niveles de medida

MEDIDA	CUALITATIVO	CUANTITATIVAS
TANQUE DE GASOLINA	VACÍO/ CUARTO/ MEDIO/ TRES CUARTOS/ LLENO	VOLUMEN DE GASOLINA EN EL TANQUE
ALTURA	ALTO/ MEDIANO/ CHICO	METROS/ CENTIMETROS
RENDIMIENTO	BAJO/ MEDIO/ ALTO	RESULTADOS DE UN EXAMEN
ENTREGAS	A TIEMPO/ TARDÍO	TIEMPO DE ENTREGA



# Medidas de tendencia central

# Medidas de tendencia central

## Overview

- **Media** = “promedio calculado”
- **Mediana** = “valor de en medio”
- **Moda** = “el valor más común”
- Describen la “localización” general de los datos.
- No describen la forma de los datos.

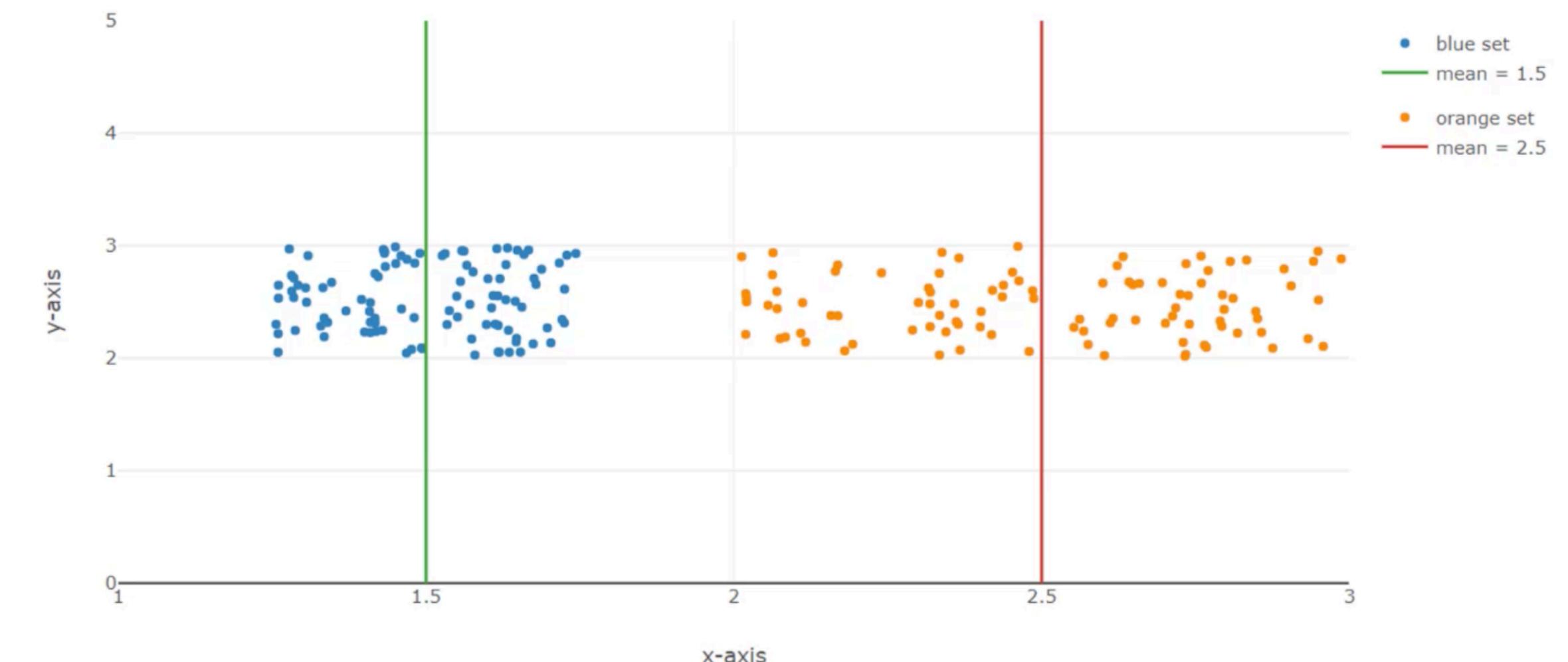
# Medidas de tendencia central

## Media aritmética

- Dado un conjunto de datos  $x = \{x_1, x_2, \dots, x_n\}$  donde  $n$  es el número de datos , la media aritmética está dada por:

$$\bullet \mu = \frac{\sum_{i=1}^N x_i}{N} \text{ para la población.}$$

$$\bullet \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \text{ para la muestra.}$$



# Medidas de tendencia central

## Mediana

- Esta medida se trabaja con datos ordenados.
- Por ejemplo, si tuviéramos un set de datos dado por  $x = \{9, 13, 44, 36, 34, 21, 28, 33, 15, 19, 10, 30, 16, 11, 10, 23, 19\}$ , para ordenarlo de menor a mayor se tendría lo siguiente:
  - $x = \{9, 10, 10, 11, 13, 15, 16, 19, 19, 21, 23, 28, 30, 33, 34, 36, 44\}$
  - Tenemos un número impar de valores de  $x$ , y la mediana es el valor central del conjunto ordenado.  $\text{Med} = 19$ .
- $\text{Median} = \left[ \frac{n+1}{2} \right]^{\text{th}} \text{ obs.}$  Datos impares.
- $\text{Median} = \frac{\frac{n}{2}^{\text{th}} \text{ obs.} + \left( \frac{n}{2} + 1 \right)^{\text{th}} \text{ obs.}}{2}$ . Datos pares.

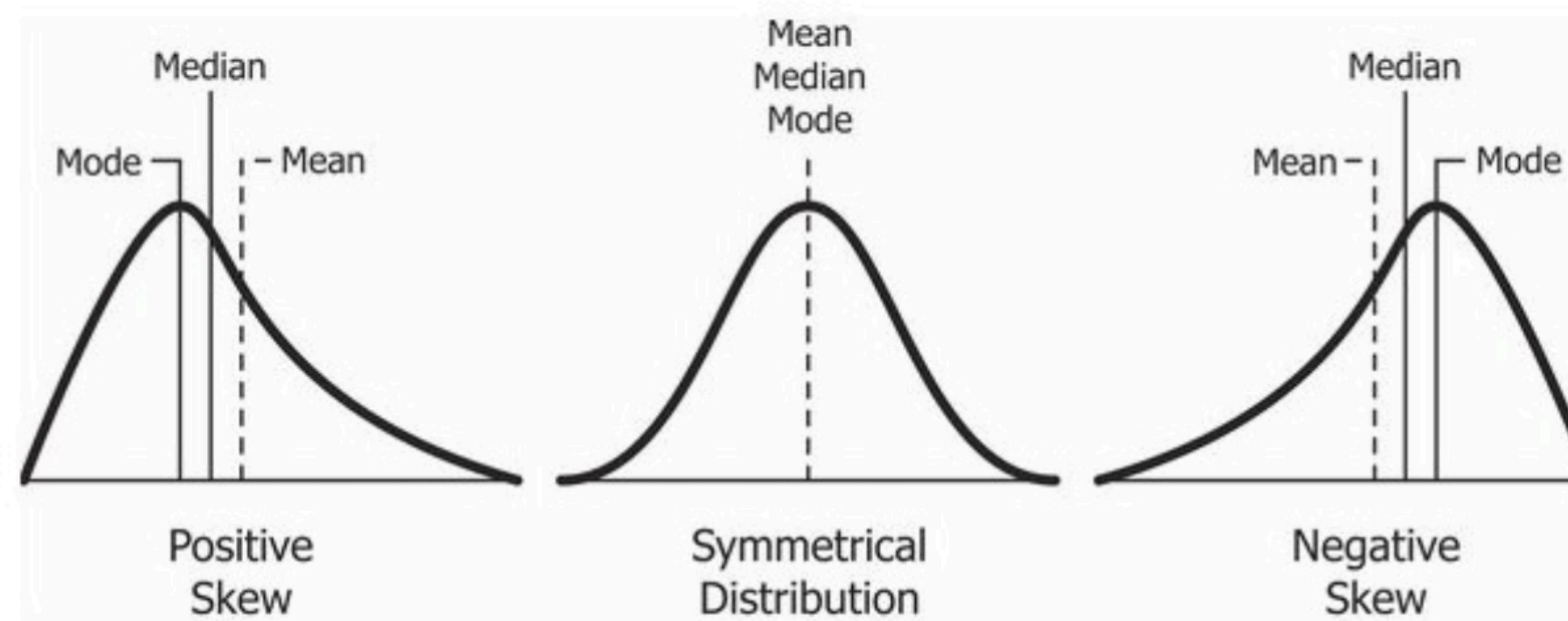
# Medidas de tendencia central

## Moda

- La moda es el valor que se repite más veces.
- $x = \{10, 10, 11, 13, 15, 16, 16, 16, 21, 23, 28, 30, 33, 34, 36, 44\}$
- La moda es 16
- Es útil para datos cualitativos.
- No se ve muy afectada por valores atípicos. Es buena medida de tendencia central para distribuciones con alta asimetría, pues nos dice el punto de máxima concentración de los datos.

# Medidas de tendencia central

## Comparación



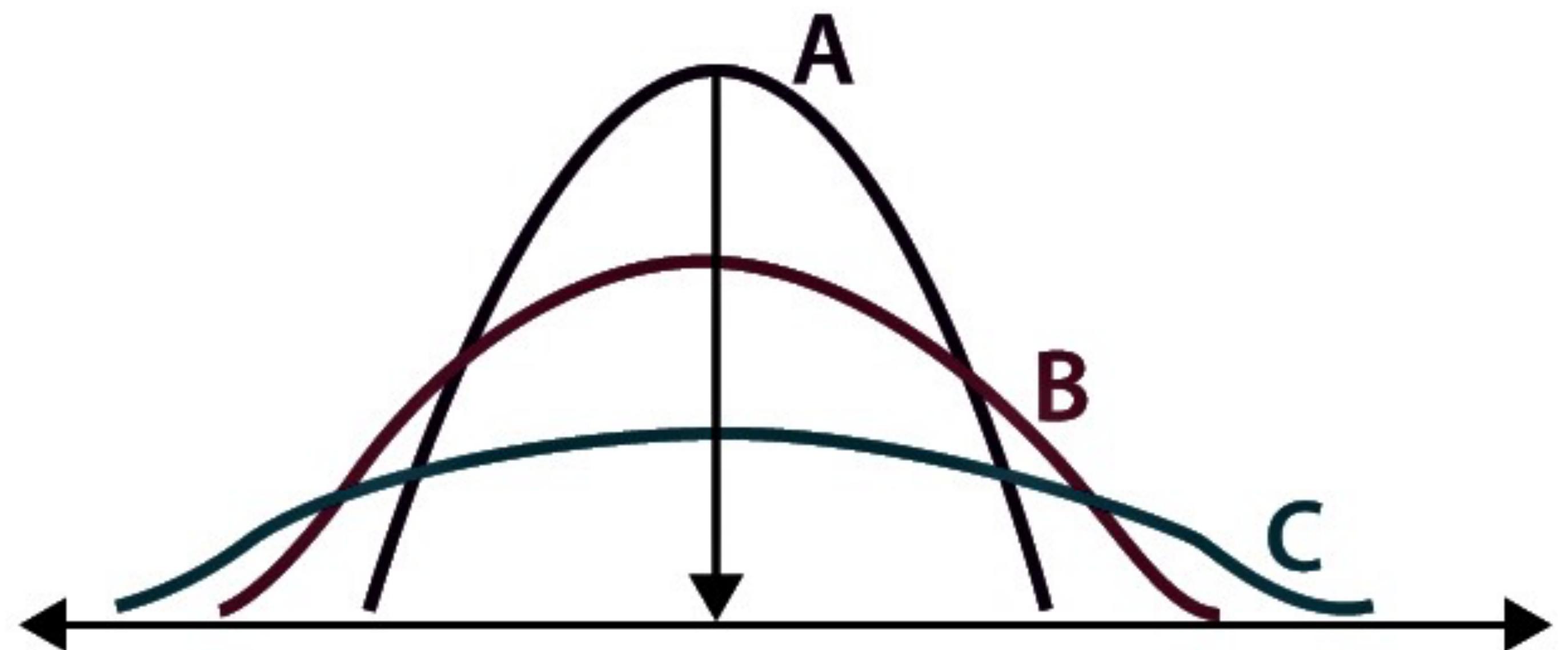
$$\text{Skew: } g = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{(n - 1)\sigma^3}$$

# Medidas de dispersión

# Medidas de dispersión

## Overview

- **Rango**
- **Varianza**
- **Desviación est<sup>á</sup>ndar**



# Medidas de dispersión

## Momentos estadísticos

- Ecuación de momento estadístico general:

$$\pi_k = \frac{1}{N} \sum_i^N \Psi^k$$

- Alrededor de la media:

$$m_r = \frac{\sum_{i=1}^n (x_i - \bar{x})^r}{n}.$$
 en el caso de muestras sería dividir entre  $n - 1$ .

- Como podemos observar, la varianza es el momento de orden 2.

$$m_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})}{n} = 0,$$

$$m_2 = \sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Moment

1st

Uncentered

$$E(X) = \mu$$

2nd

$$E(X^2)$$

Centered

$$E((X-\mu)^2)$$

3rd

$$E(X^3)$$

$$E((X-\mu)^3)$$

4th

$$E(X^4)$$

$$E((X-\mu)^4)$$

$$\text{Mean}(X)$$

$$= E(X)$$

$$\text{Var}(X)$$

$$= E((X-\mu)^2) = \sigma^2$$

$$\text{Skewness}(X)$$

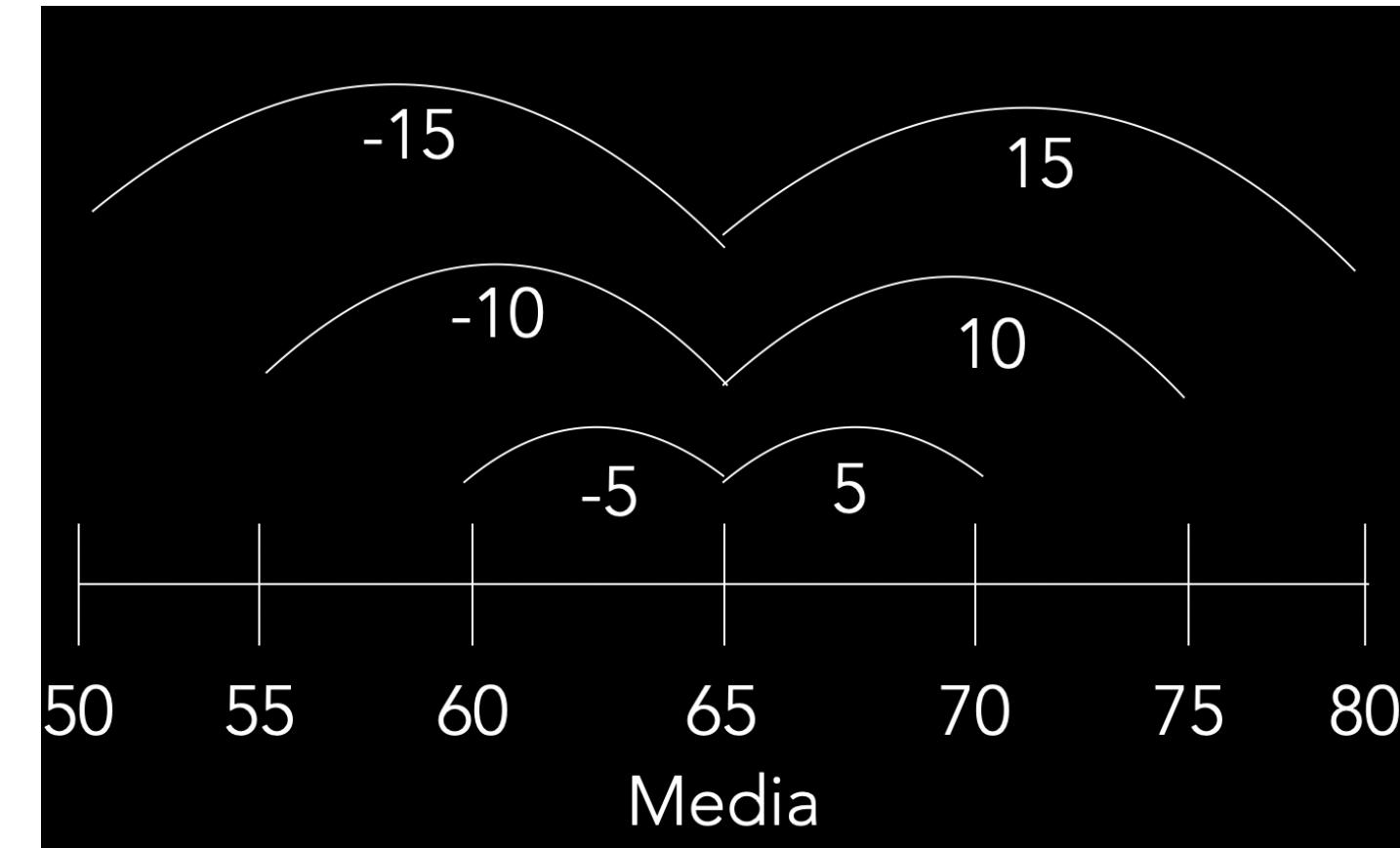
$$= E((X-\mu)^3) / \sigma^3$$

$$\text{Kurtosis}(X)$$

$$= E((X-\mu)^4) / \sigma^4$$

# Medidas de dispersión

## Varianza



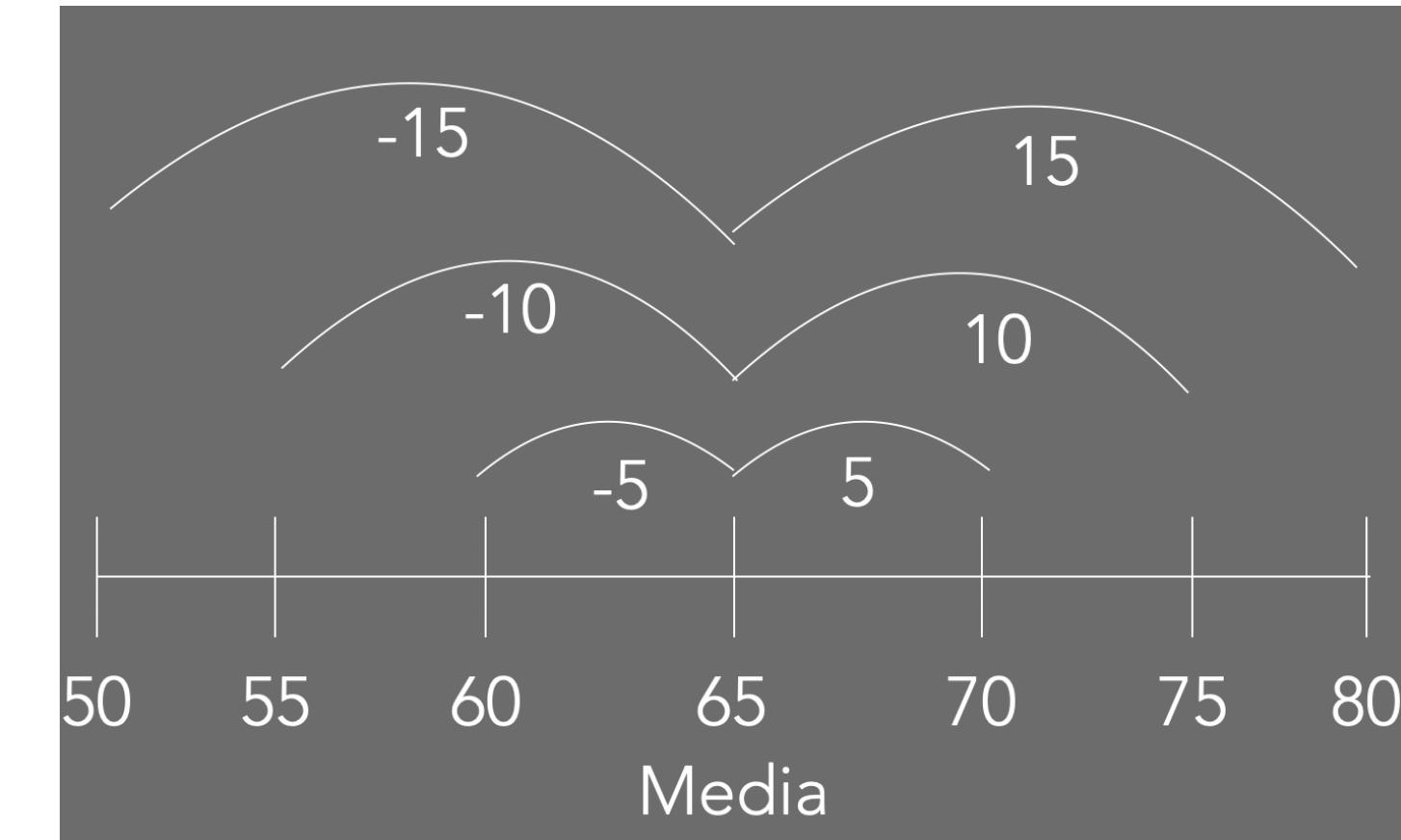
- En lugar de utilizar el valor absoluto para tener signo positivo, se utiliza la suma de los cuadrados de la distancia de cada punto con respecto a la media.
- IMPORTANTE: Se tiene que tomar en cuenta si lo que se analiza es una población o una muestra.
- Para esto se utiliza la llamada **corrección de Bessel** ( $n - 1$ ).

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}, \text{ Varianza de una muestra.}$$

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}, \text{ Varianza de una población.}$$

# Medidas de dispersión

## Desviación estándar

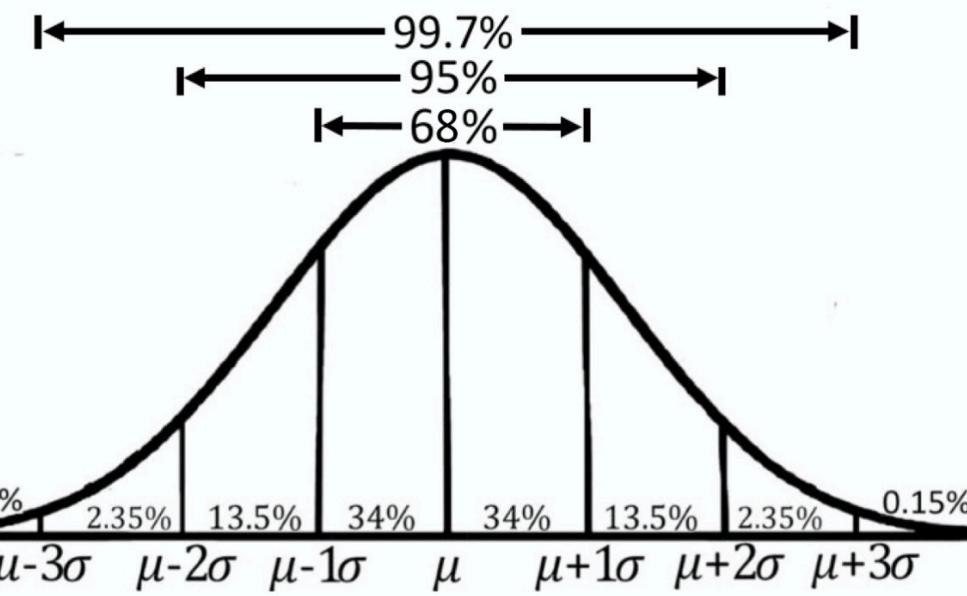


- Pero hay un problemita con la varianza. Al elevar al cuadrado las desviaciones, lo que tenemos son unidades originales al cuadrado, y es un poco complicado interpretar el resultado.
- La desviación estándar es la raíz cuadrada de la varianza.
- Utiliza las mismas unidades que la muestra (o población). Por tal razón, es significativo hablar de “valores que están dentro de tantas desviaciones estándar de la media”.

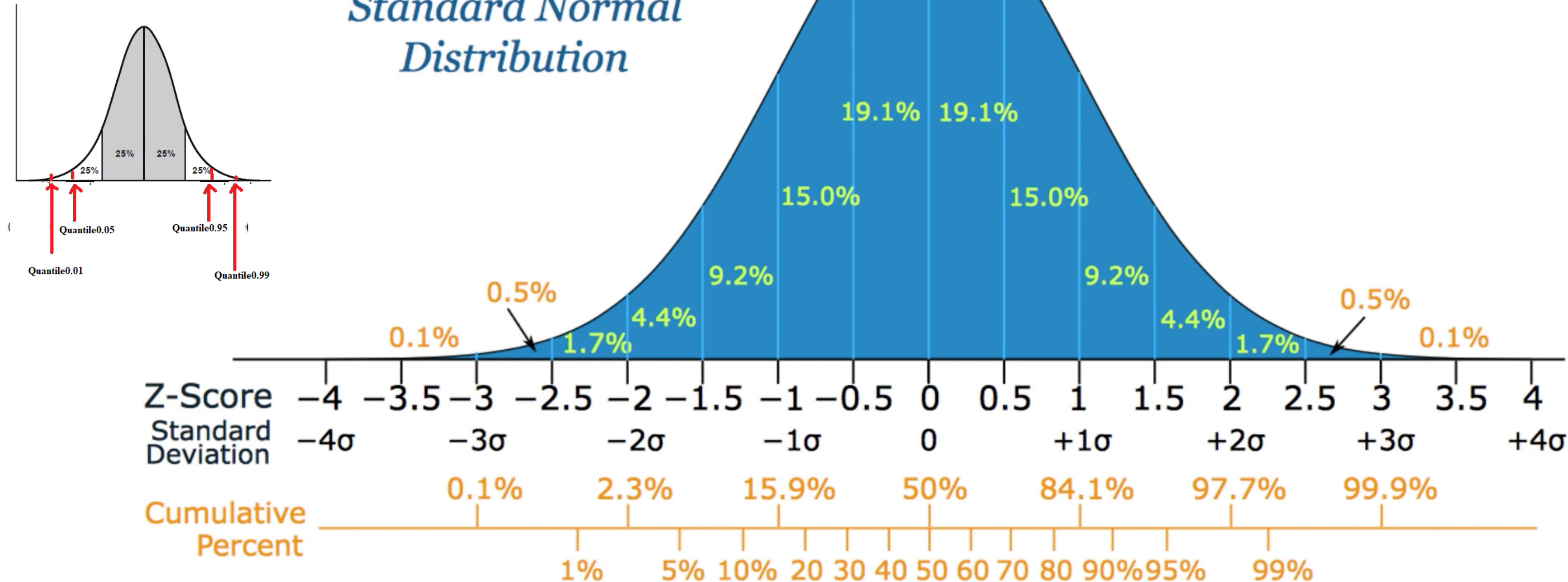
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}, \text{ Desviación estándar de una muestra.}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}, \text{ Desviación estándar de una población.}$$

## Empirical Rule



## *"Bell Curve"* Standard Normal Distribution

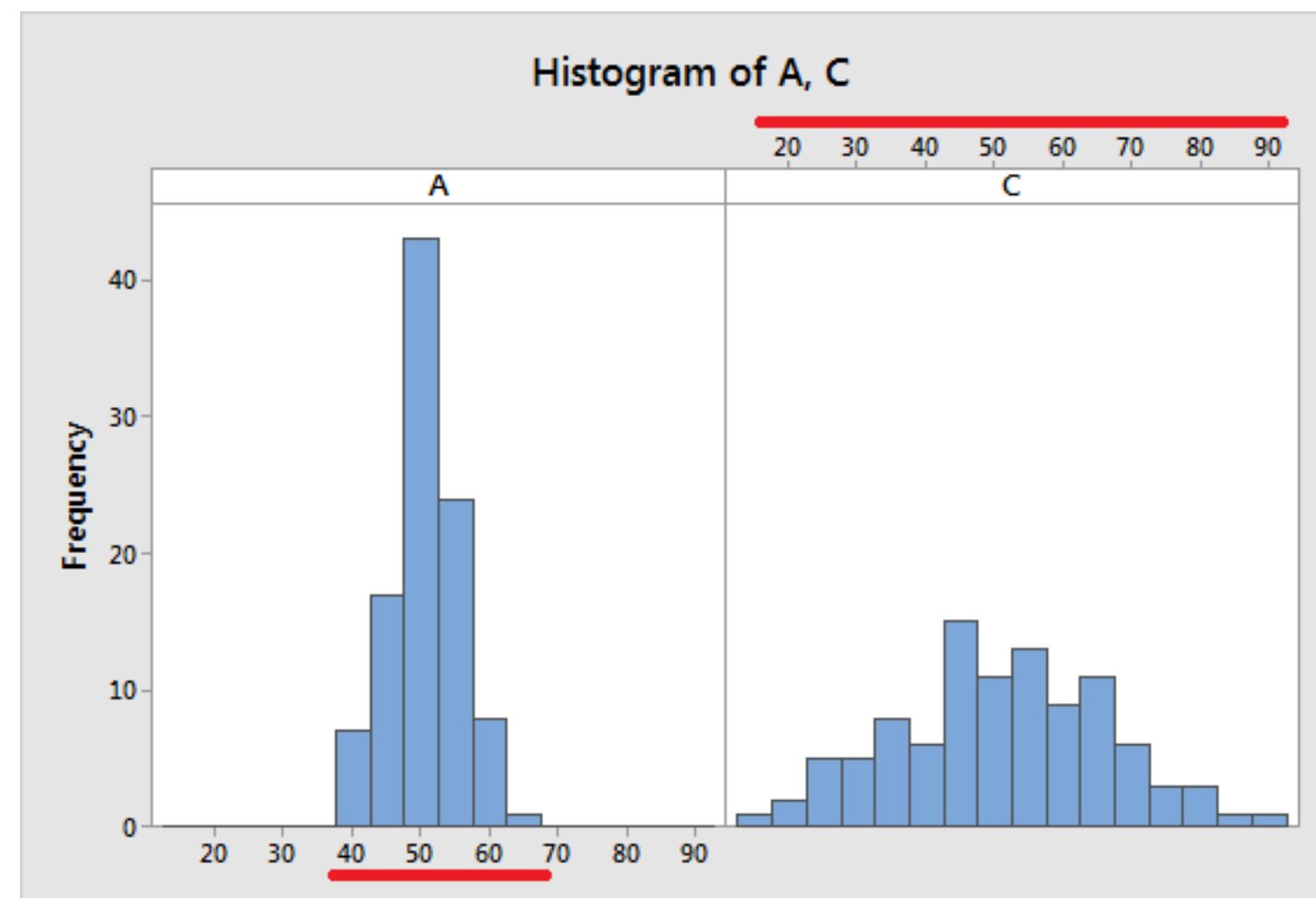


La Campana de Gauss

# Medidas de dispersión

## Rango

- Rango = max - min
- $x = \{9^*, 10, 11, 13, 15, 16, 19, 19, 21, 23, 28, 30, 33, 34, 36, 39^*\}$  tiene un minimo de 9 y un máximo de 39, por lo que el Rango( $x$ ) =  $39 - 9 = 30$ .



# Medidas de dispersión

## Coeficiente de variación

- Si se quieren comparar dos conjuntos de datos con unidades distintas, incluso las unidades de la desviación estándar pueden ser una limitante.
- El coeficiente de variación se define por:
  - $CV = \frac{s}{\bar{x}} \times 100$  para una muestra.
  - $CV = \frac{\sigma}{\mu} \times 100$  para una población.

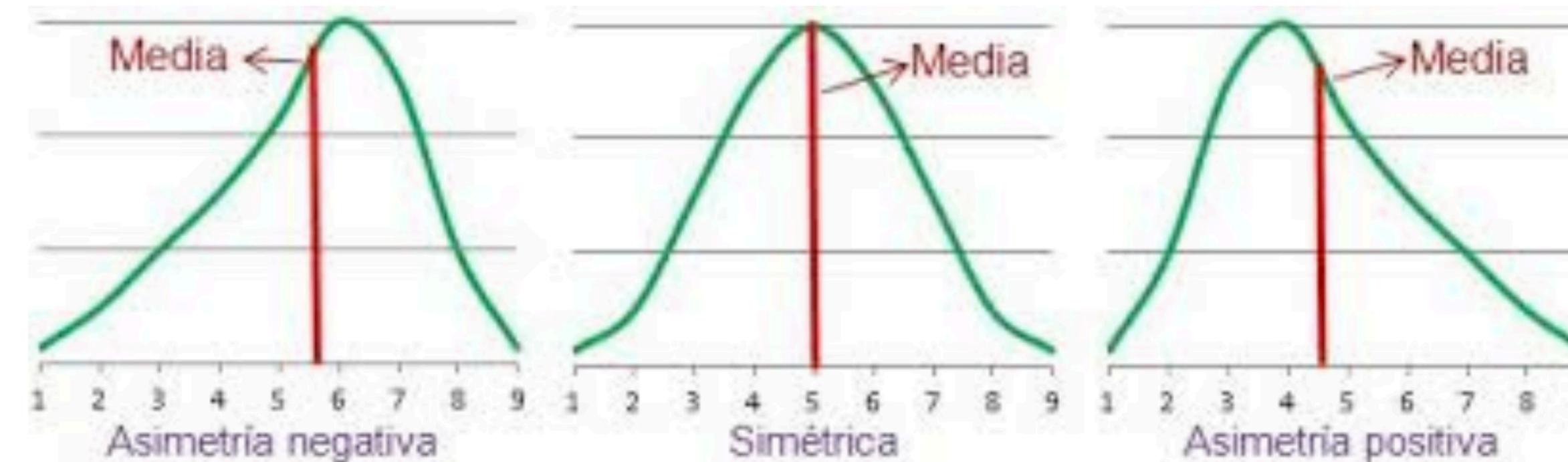
	ACCIÓN A	ACCIÓN B
ENERO	1.00	180.00
FEBRERO	1.50	175.00
MARZO	1.90	182.00
ABRIL	0.60	186.00
MAYO	3.00	188.00
JUNIO	0.40	190.00
JULIO	5.00	200.00
AGOSTO	0.20	210.00

- Esta es una muestra de datos de los últimos meses de las dos acciones.
- $s_A = 1.62$ ;  $s_B = 11.33$ . La variación de la acción es más grande.
- Sin embargo,  $CV_A = 95.3\%$  y  $CV_B = 6.0\%$ .

# Asimetría

## Coeficiente de asimetría de Fisher

- Definido por  $CA_F = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{ns^3}$ , es un momento de orden 3 dividido entre la desviación estándar elevada al cubo.
- Mediante este coeficiente observamos si una función es perfectamente simétrica (como una curva normal, por ejemplo). Cuanto más positivo más desplazada está hacia la izquierda (positiva), y viceversa.



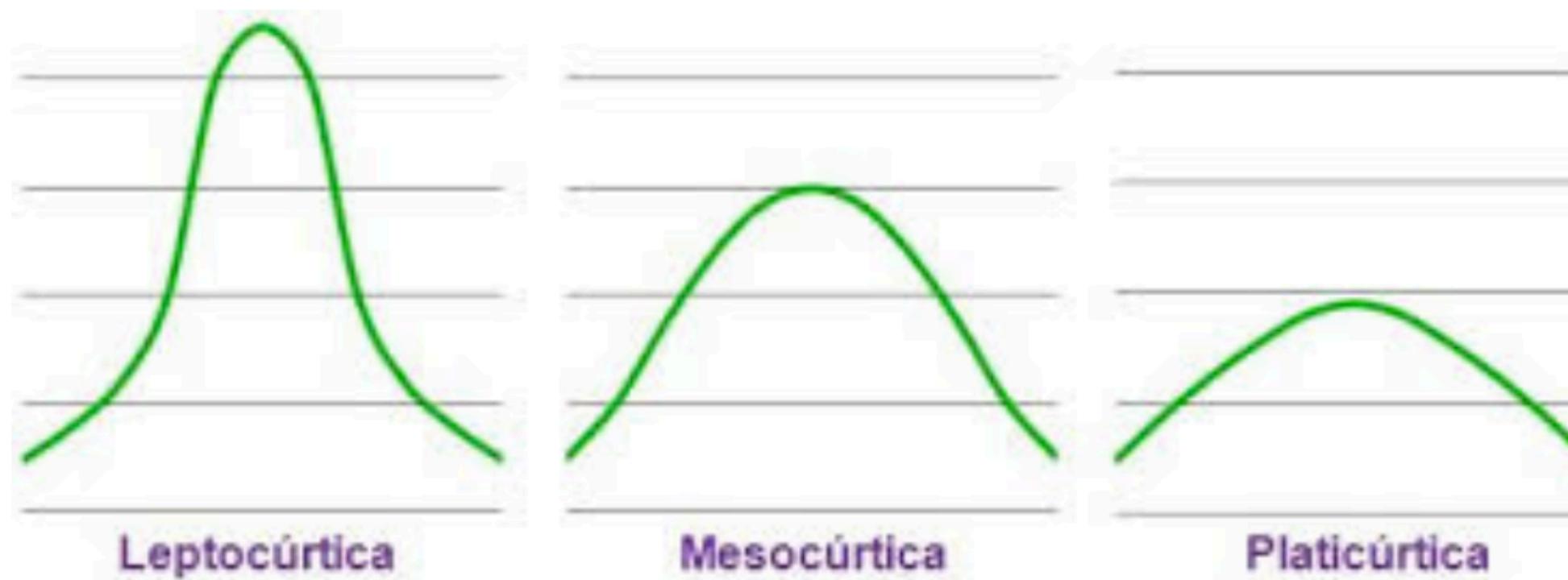
# Curtosis

## Coeficiente de curtosis

- Se utiliza el momento de orden 4 y se define por

$$\bullet \quad c = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{ns^4} - 3$$

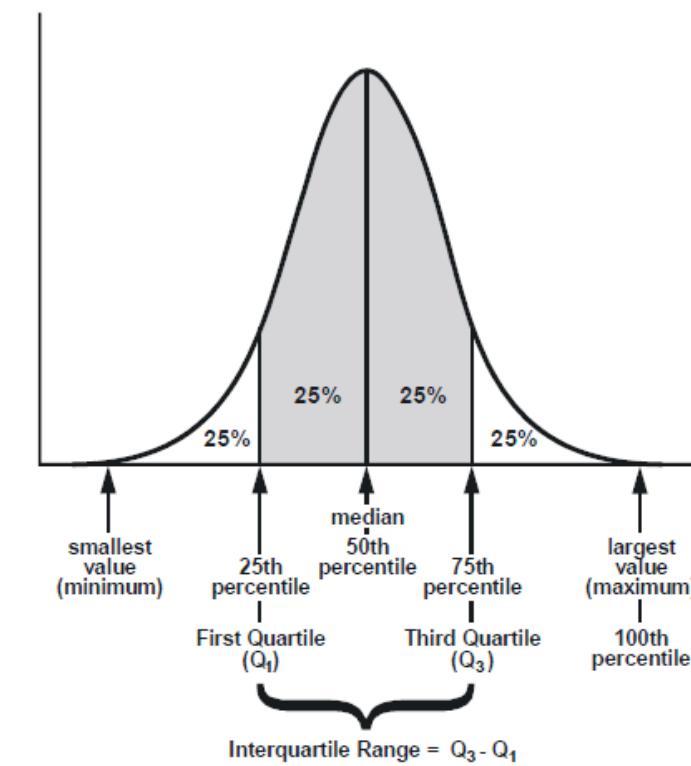
- Define una forma para la distribución de datos. En  $c=0$ , se asemeja mucho a una curva normal (aunque no en valores), con las colas y la mitad de la distribución balanceado (Mesocúrtica).
- $c > 0$ , Leptocúrtica. Los datos se concentran mucho en la región central y apenas tiene cola.
- $c < 0$ . Platicúrtica. Las colas de la distribución contienen muchos de los valores del conjunto.



# Cuartiles

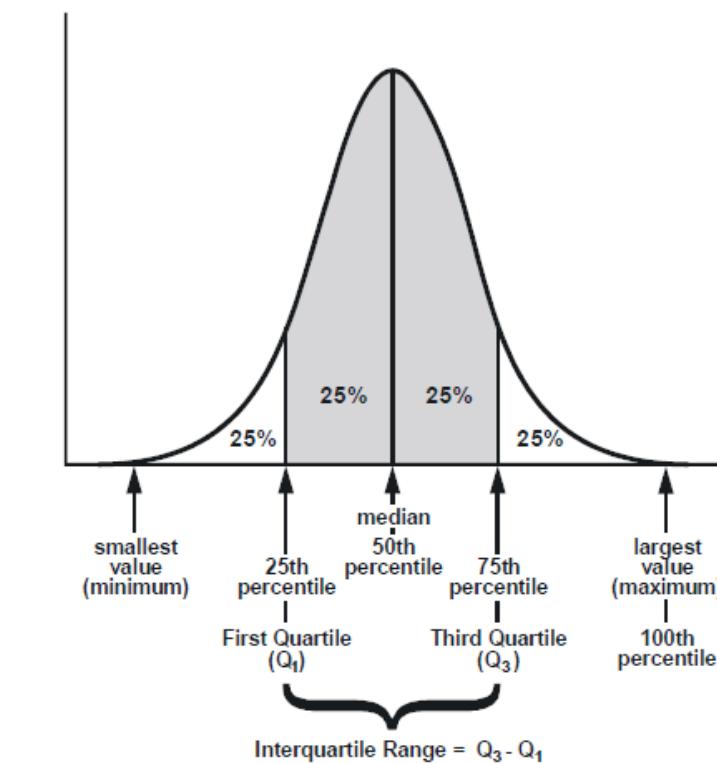
## Definición

- Otra forma de describir los datos es mediante cuartiles y el rango intercuartil (IQR).
- Consideremos un pequeño conjunto  $x = \{9, 10, 10, 11, 13, 15, 16, 19, 19, 21, 23, 28, 30, 33, 34, 36, 44, 45, 47, 60\}$ .
- Lo primero que hacemos es dividir nuestro set a la mitad (en términos del índice). La muestra consiste de 20 valores. (Evidentemente, esto es el mismo procedimiento para encontrar la mediana.)
- $[9, 10, 10, 11, 13, 15, 16, 19, 19, 21] [23, 28, 30, 33, 34, 36, 44, 45, 47, 60]$
- Volvemos a dividir cada subconjunto en dos.
- $[9, 10, 10, 11, 13] [15, 16, 19, 19, 21] [23, 28, 30, 33, 34] [36, 44, 45, 47, 60]$ 
  - Q1
  - Q2
  - Q3
- 14
- 22
- 35
- Esto se puede visualizar un un gráfico de caja.



# Cuartiles

## Rango intercuartil



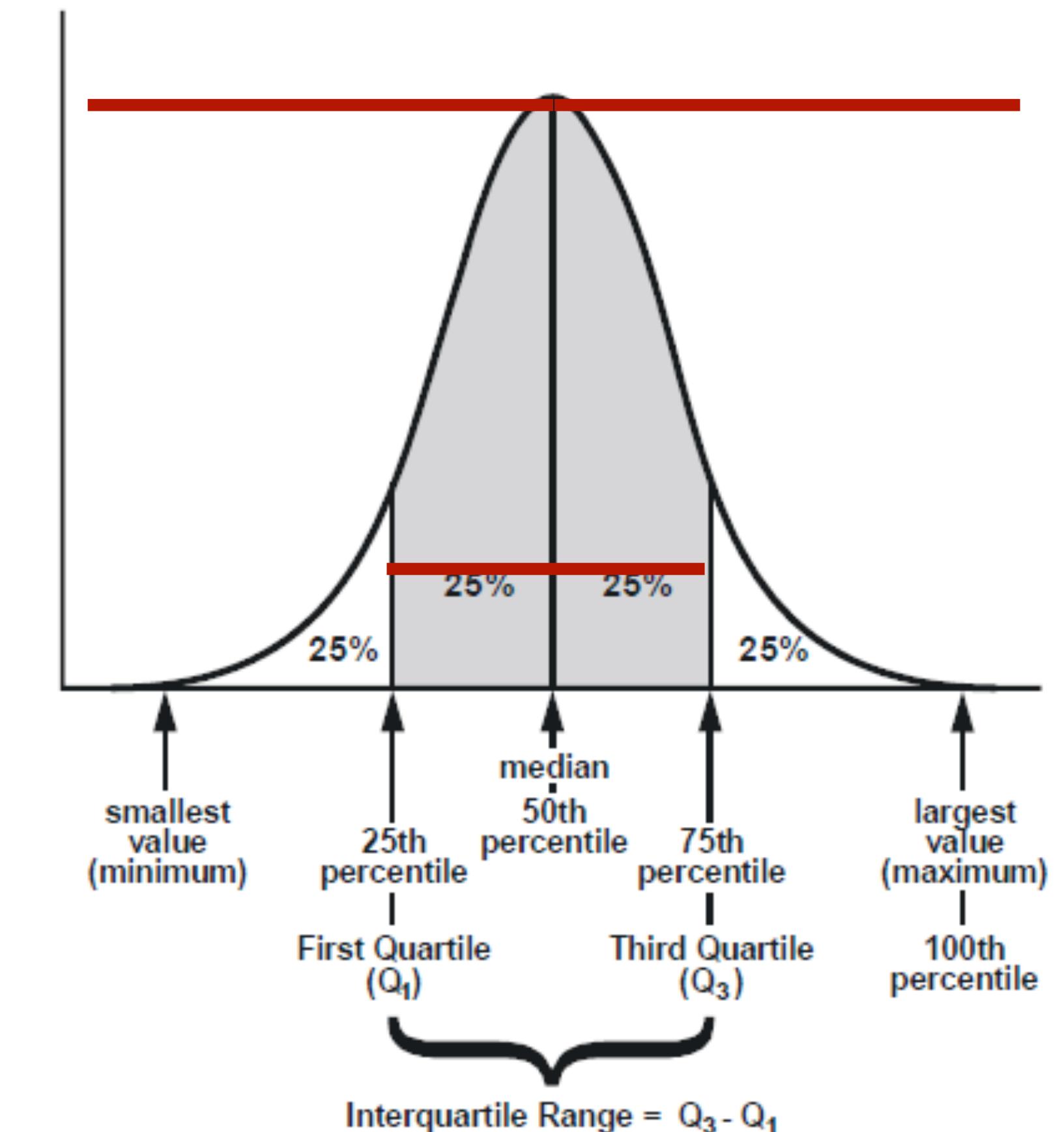
- ¿Qué es considerado un valor atípico?
- Una práctica común es poner un “límite” a 1.5 veces el rango intercuartil IQR a partir de Q1 y Q3.
- Todo valor fuera de este “límite” se considera valor atípico.
- Lo importante es que esto es determinado totalmente por los datos y no es un porcentaje arbitrario.

# Cuartiles

## Rango intercuartil

- El rango intercuartil está definido por  $IQR = Q_3 - Q_1$ .
- El límite a partir del cual los valores se consideran atípicos está dado por:
- $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ .

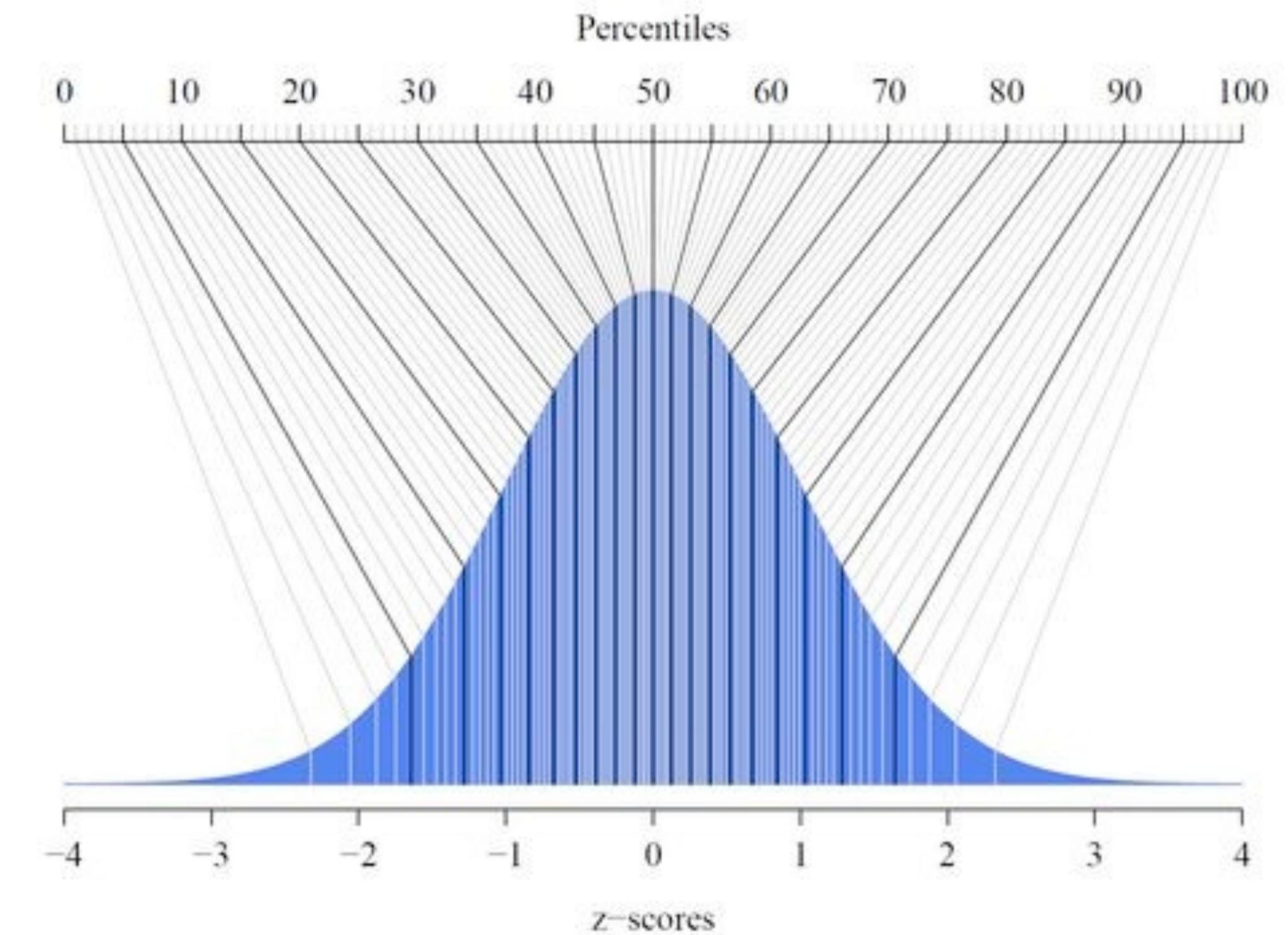
Cualquier valor fuera de estos límites es considerado atípico (outlier).



# Percentiles

## Definición

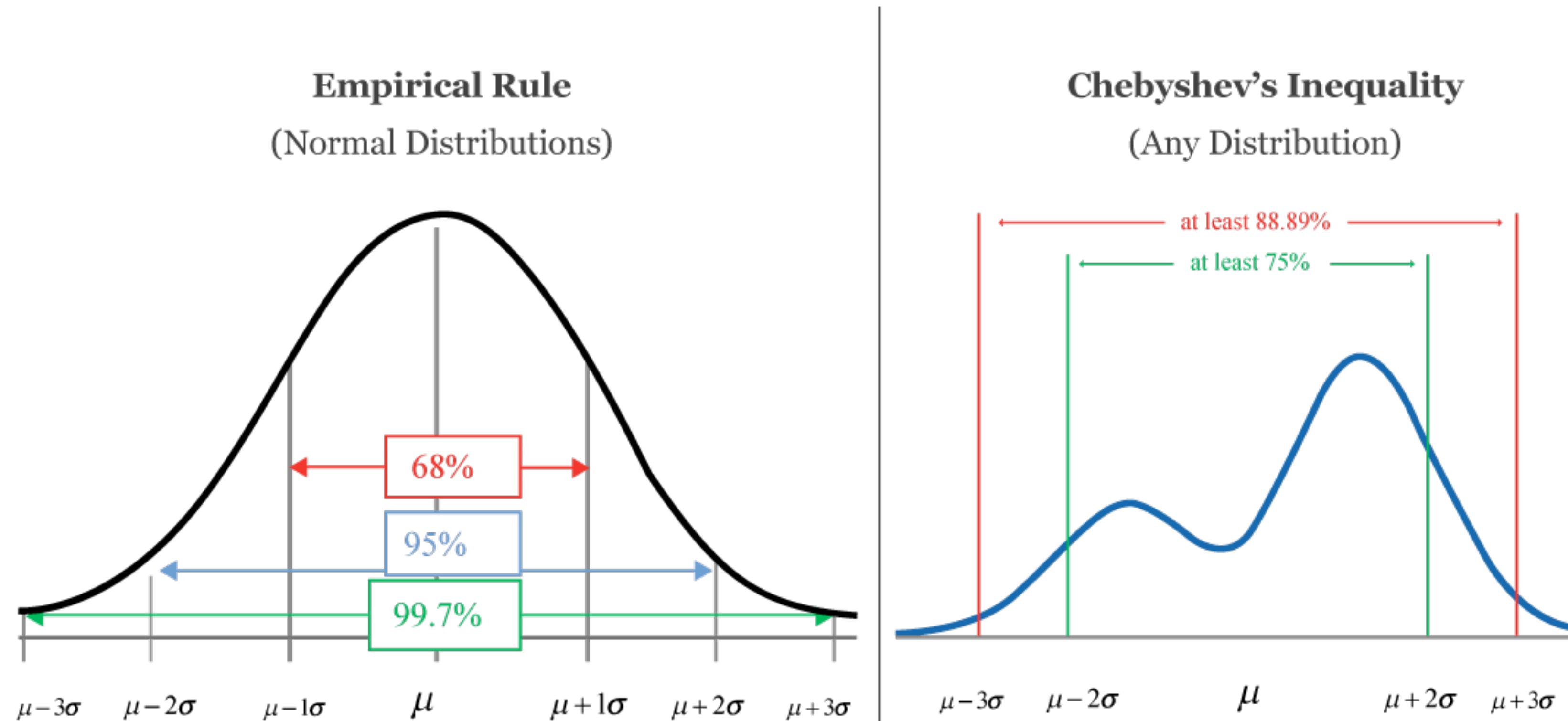
- Para dividir nuestro conjunto de datos en 100 partes iguales se utilizan los percentiles, definidos por
- $L_p = (n) \frac{p}{100}$
- Donde  $n$  es el número de observaciones (datos) y  $p$  es el percentil deseado.



# Teorema de Chebyshev

## Definición

- Por cada población de  $n$  valores y un valor real  $k > 1$ , la proporción de valores dentro de  $k$  desviaciones estándar de la media es **al menos**  $1 - \frac{1}{k^2}$



# Teorema de Chebyshev

- Suponer una población de  $n$  valores que consisten de  $n_1$  para  $x_1$ ,  $n_2$  valores para  $x_2$  y así. Tenemos  $n_i$  valores. Suponer que  $\sigma$  es la desviación estándar de población,  $\mu$  la media, y  $k > 1$  un número real positivo
- $$\sigma^2 = \frac{\sum (x_i - \mu)^2 n_i}{n} \geq \frac{\sum (x_i - \mu)^2 n_i}{n}$$
 donde la última suma sólo considera los valores donde se cumple que  $|x_i - \mu| \geq k\sigma$
- Si  $|x_i - \mu| \geq k\sigma \Rightarrow (x_i - \mu)^2 \geq k^2\sigma^2$
- $$\frac{\sum (x_i - \mu)^2 n_i}{n} \geq \frac{\sum (k^2\sigma^2)n_i}{n}$$

# Teorema de Chebyshev

- $\Rightarrow \geq \frac{\sum k^2 \sigma^2 n_i}{n} = k^2 \sigma^2 \frac{\sum n_i}{n}$
- Donde esta ultima suma  $\frac{\sum n_i}{n} = p_{fuera}$  es la proporción fuera de  $k$  desviaciones estándar de la media.
- $\sigma^2 \geq k^2 \sigma^2 \frac{\sum n_i}{n} \Rightarrow 1 \geq k^2 p_{fuera} \Rightarrow p_{dentro} \geq 1 - \frac{1}{k^2}$
- Donde  $p_{fuera} = 1 - p_{dentro}$

# Teorema de Chebyshev

## Ejemplo

- Ejercicio: La calificación media de un examen es de 75, con desviación estándar de 5. Cuál porcentaje de los datos queda entre 50 y 100 según el teorema de Chebyshev.

$$k = 5 \quad p_{dentro} \geq 1 - \frac{1}{25} = 0.96$$

- La media de la edad de las empleadas de una tienda es de 30, con desviación estándar de 6. Entre qué edades se encuentran el 75% de los datos, alrededor de la media.

$$p_{dentro} = 0.75 \geq 1 - \frac{1}{k^2}$$

$$\frac{1}{k^2} \geq 0.25$$

$$1 \geq 0.25k^2 \Rightarrow \frac{1}{0.25} = 4 \geq k^2 \Rightarrow k = 2$$

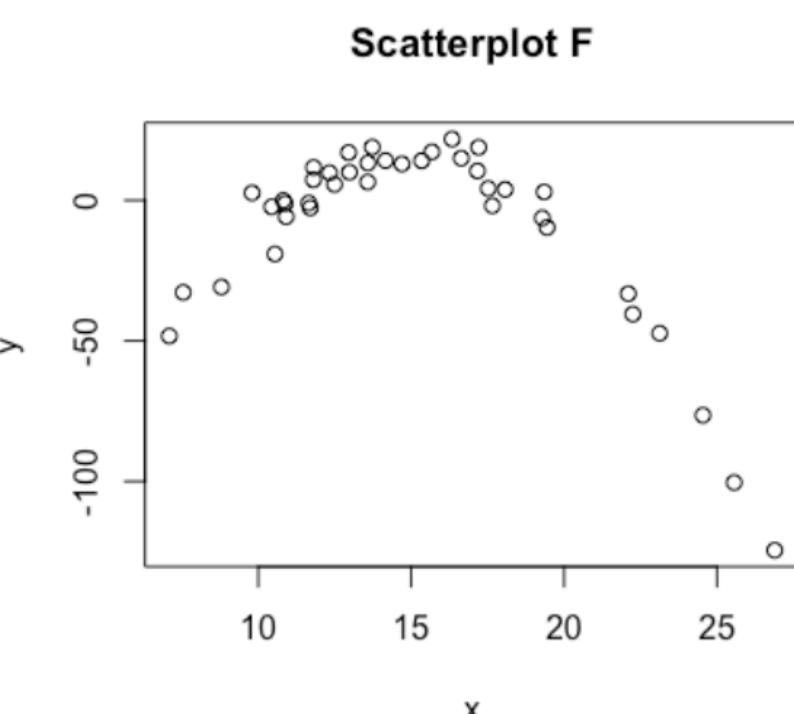
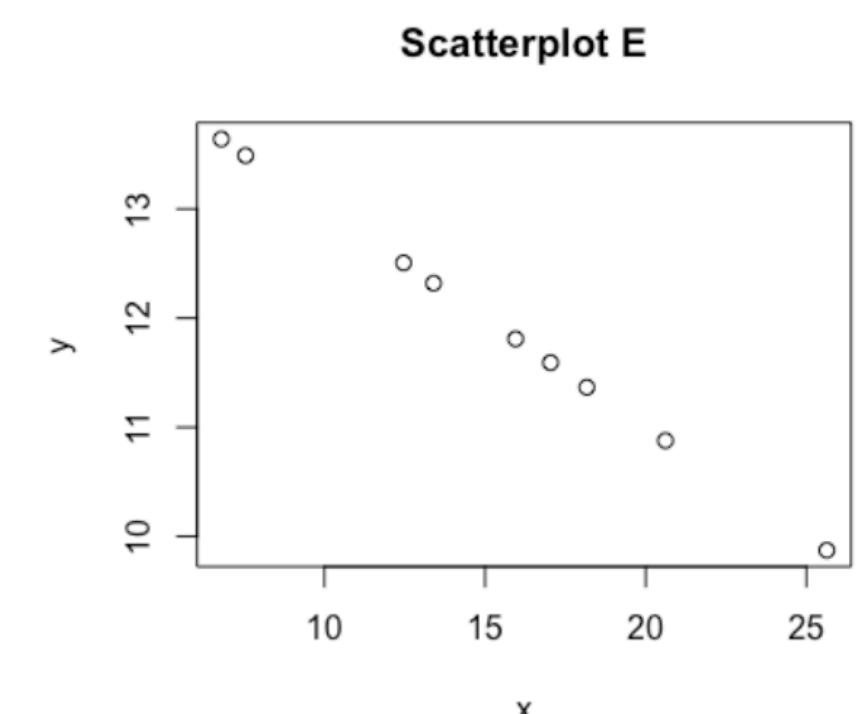
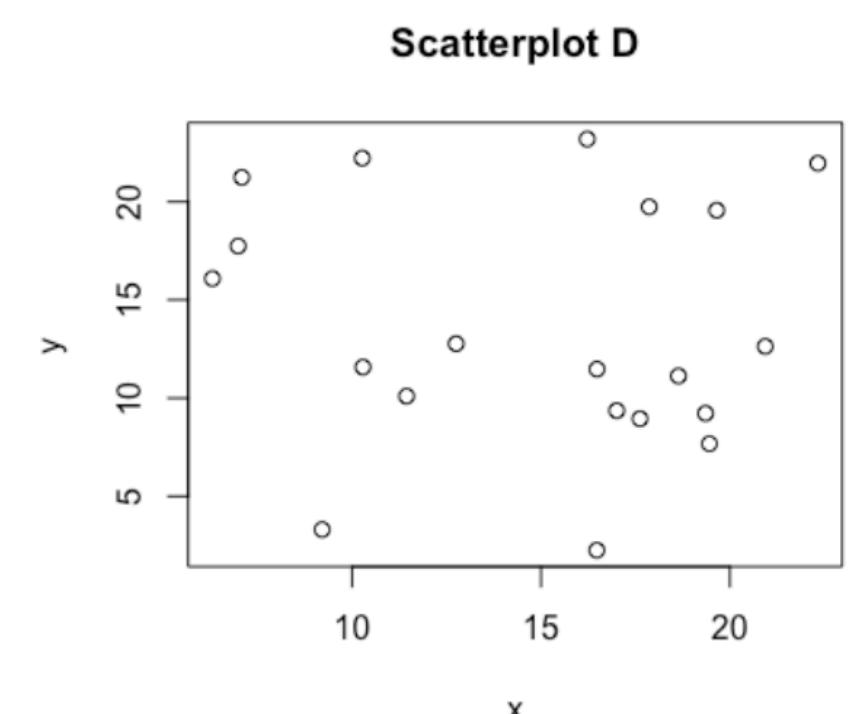
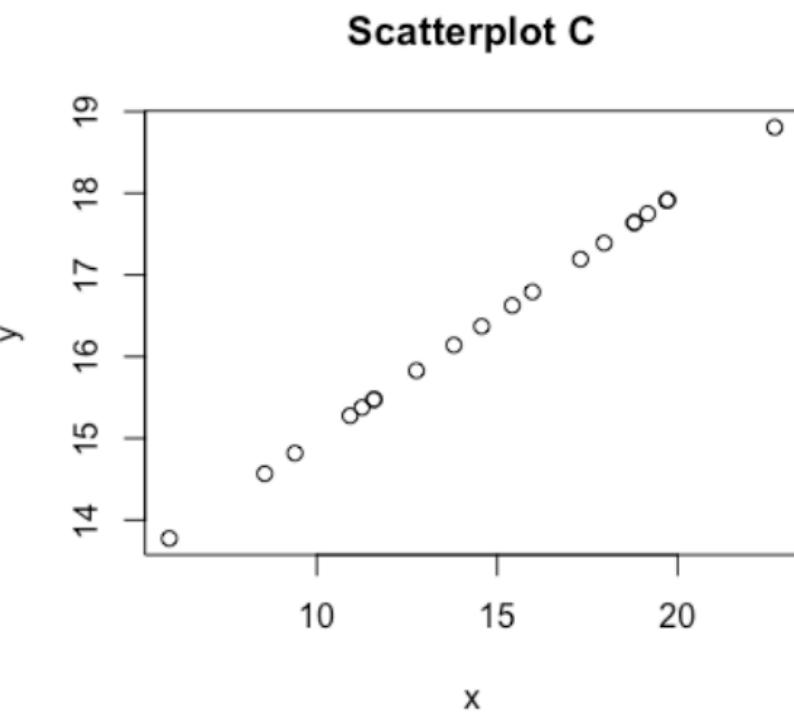
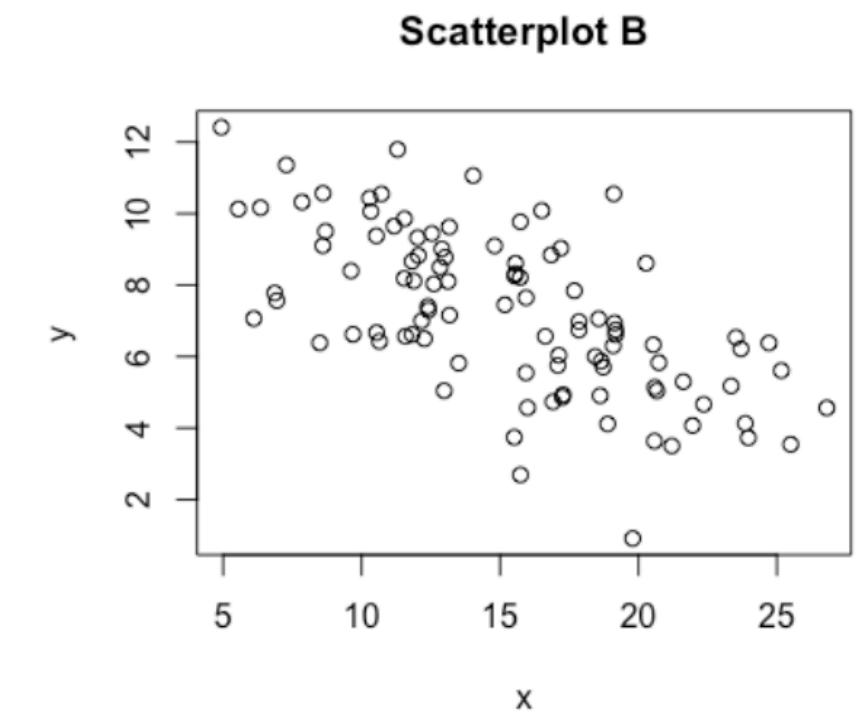
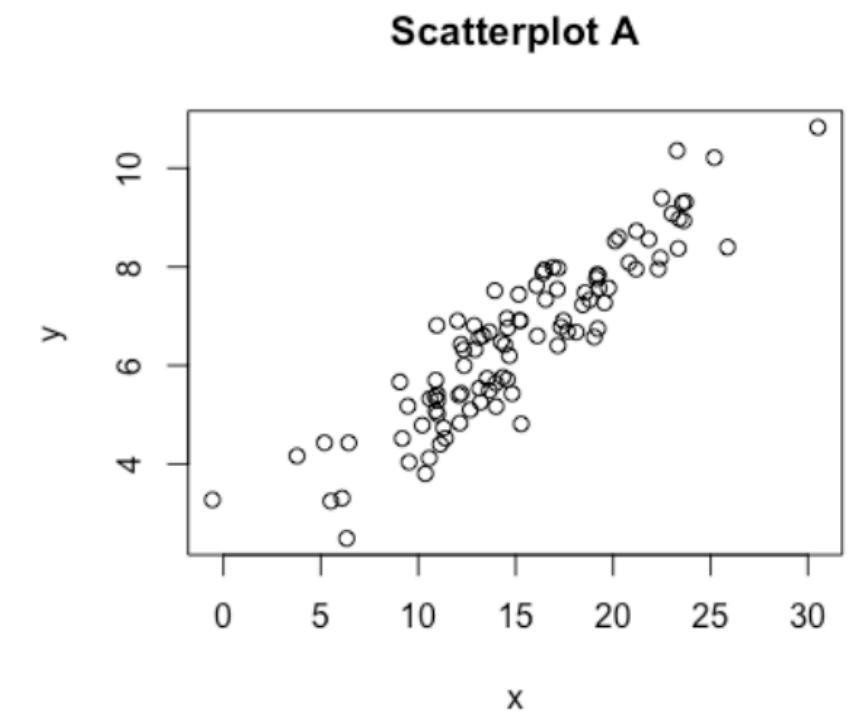
# Datos bivariados

# Datos bivariados

## Diagrama de dispersión

- Datos Bivariados es simplemente cuando comparamos dos variables diferentes.
- Por convención, el eje x se fija como la variable independiente.
- Se fija el eje y como variable dependiente, aquella que se mide relativa a x.
- Una técnica gráfica que se usa para mostrar la relación entre variables es llamada diagrama de dispersión.
- En esta, puntos (x,y) (observaciones) son graficados.

Normalmente, las dos variables están relacionadas

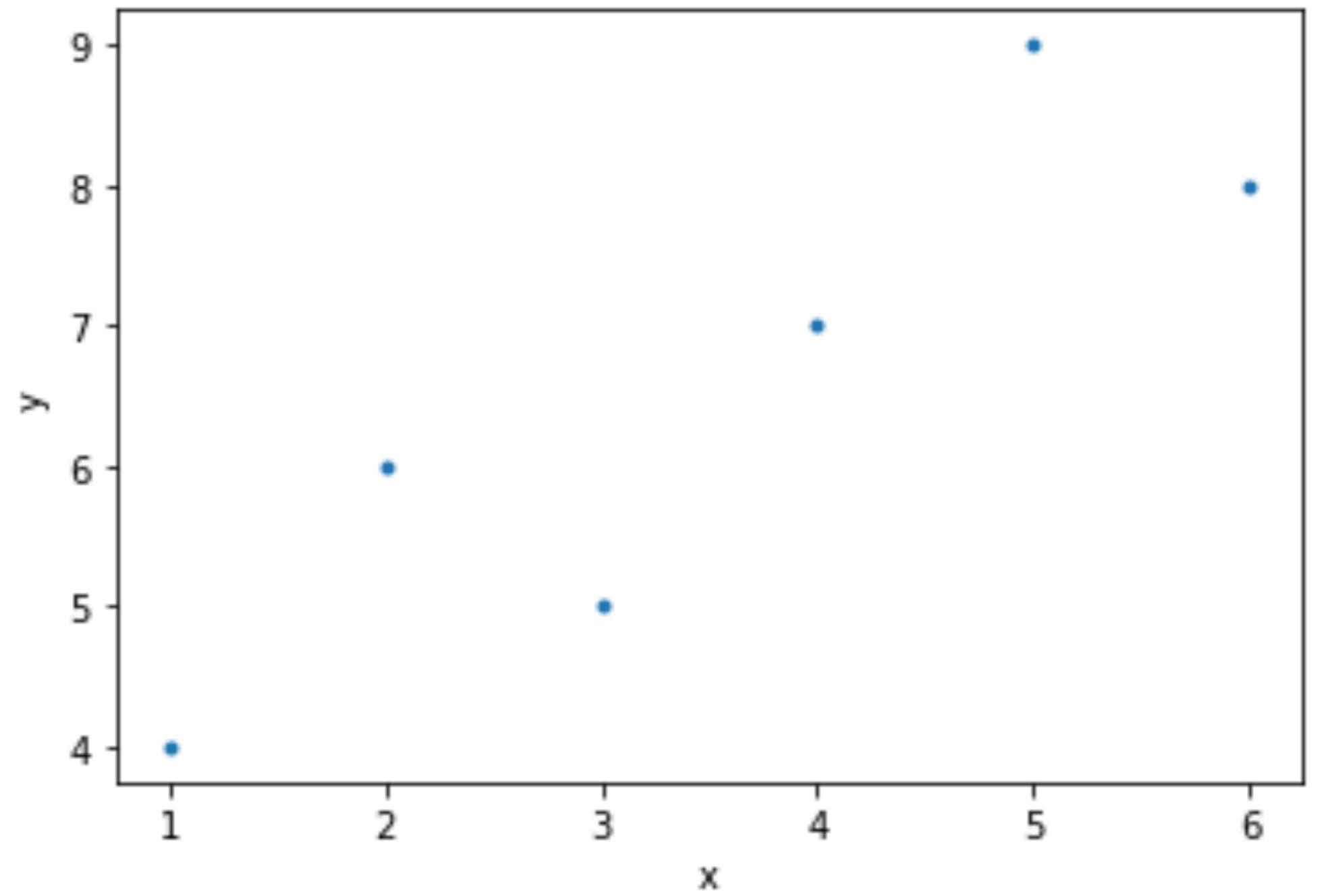


# Datos bivariados

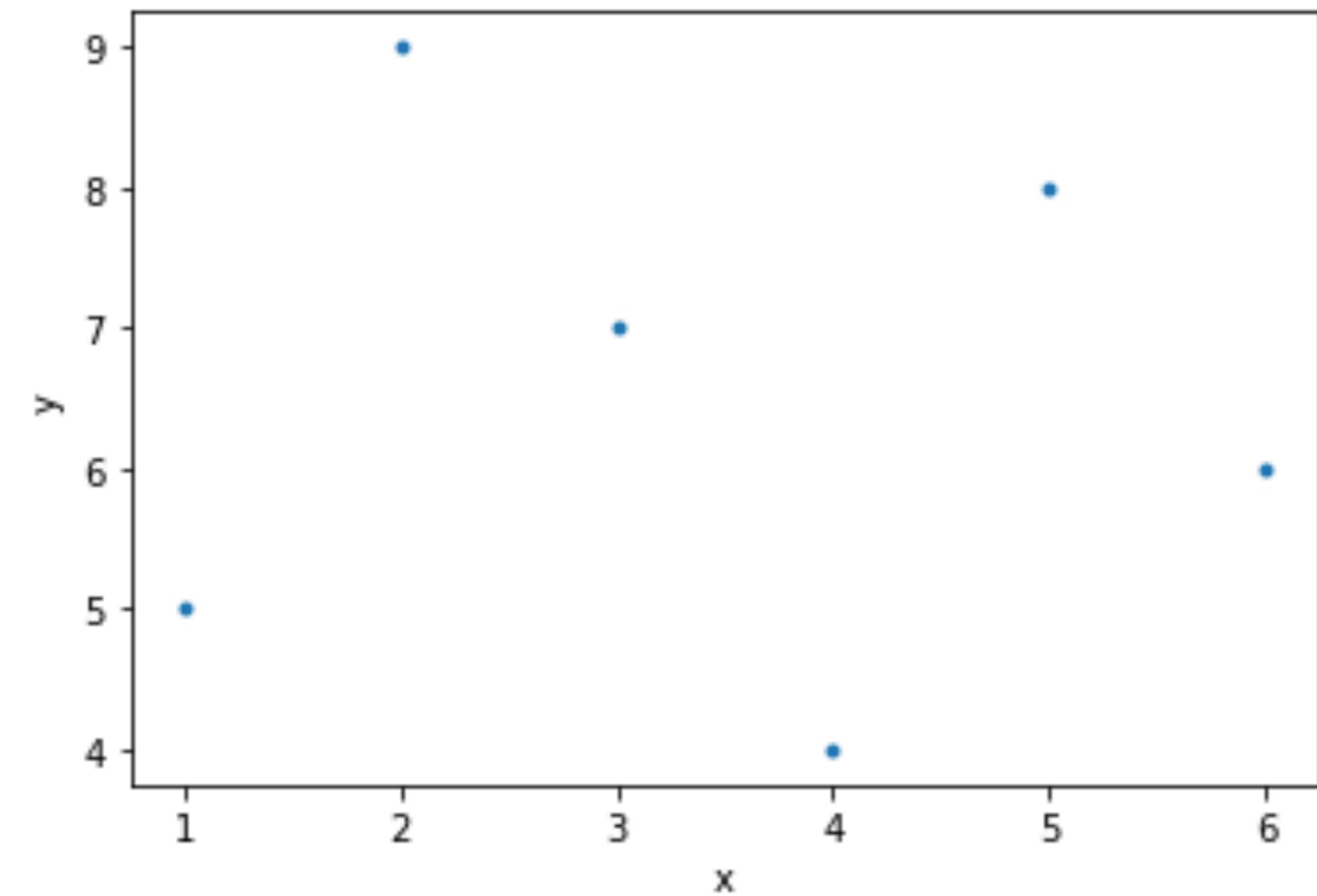
## Covarianza

- Una forma común de comparar dos variables es comparando sus variantes- qué tan lejos está cada valor de su media correspondiente.
- La covarianza de una población, aunque usando la notación de la media de muestra, definida por:

$$cov(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N} \sum_{i=1}^N x_i y_i - \bar{x} \bar{y}$$



$$Cov(x, y) = 2.583$$



$$Cov(x, y) = -0.083$$

# Datos bivariados

## Factor de correlación de Pearson

- Para solucionar el problema de las unidades, normalizamos los valores que vienen de dos distribuciones distintas y definimos el factor de correlación de Pearson.

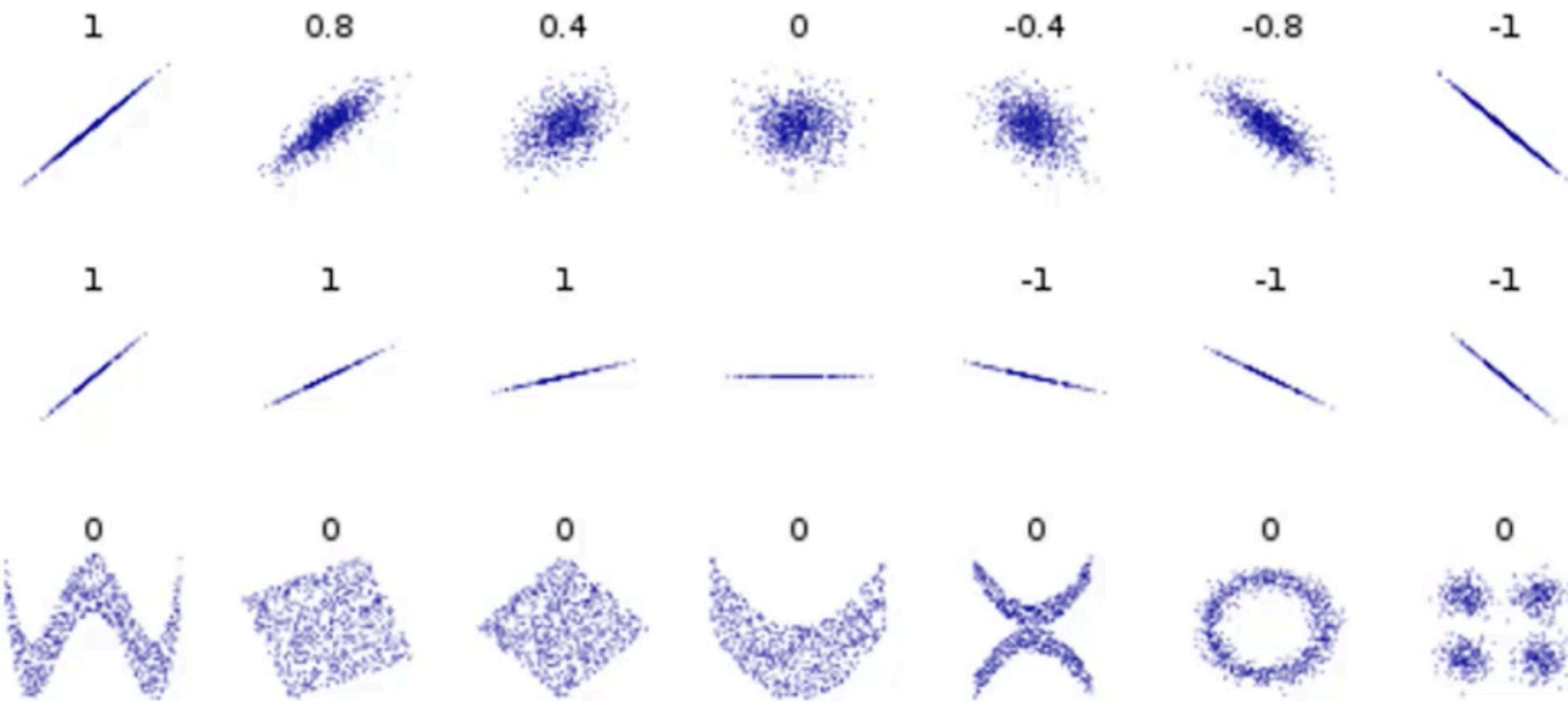
$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{N} \sum_{i=1}^N (x - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}}} = \frac{\sum_{i=1}^N (x - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

Los valores del factor de correlación caen entre -1 y 1.

= 1 Significa una correlación total positiva.

= 0 Significa que no existe una correlación lineal.

= -1 Significa una correlación total negativa.



# Datos bivariados

## Otros coeficientes de correlación

### Spearman's Rank Correlation Coefficient

$$\rho = \frac{\sum_{i=1}^n (R(x_i) - \bar{R(x)})(R(y_i) - \bar{R(y)})}{\sqrt{\sum_{i=1}^n (R(x_i) - \bar{R(x)})^2 \cdot \sum_{i=1}^n (R(y_i) - \bar{R(y)})^2}} = 1 - \frac{6 \sum_{i=1}^n (R(x_i) - R(y_i))^2}{n(n^2 - 1)}$$

Where,  $R(x_i)$  = rank of  $x_i$

$R(y_i)$  = rank of  $y_i$

$\bar{R(x)}$  = mean rank of  $x$

$\bar{R(y)}$  = mean rank of  $y$

$n$  = number of pairs

### Kendall's Tau Coefficient

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n - 1)/2}$$

Where,  $n_c$  = number of concordant pairs

$n_d$  = number of discordant pairs

$n$  = number of pairs

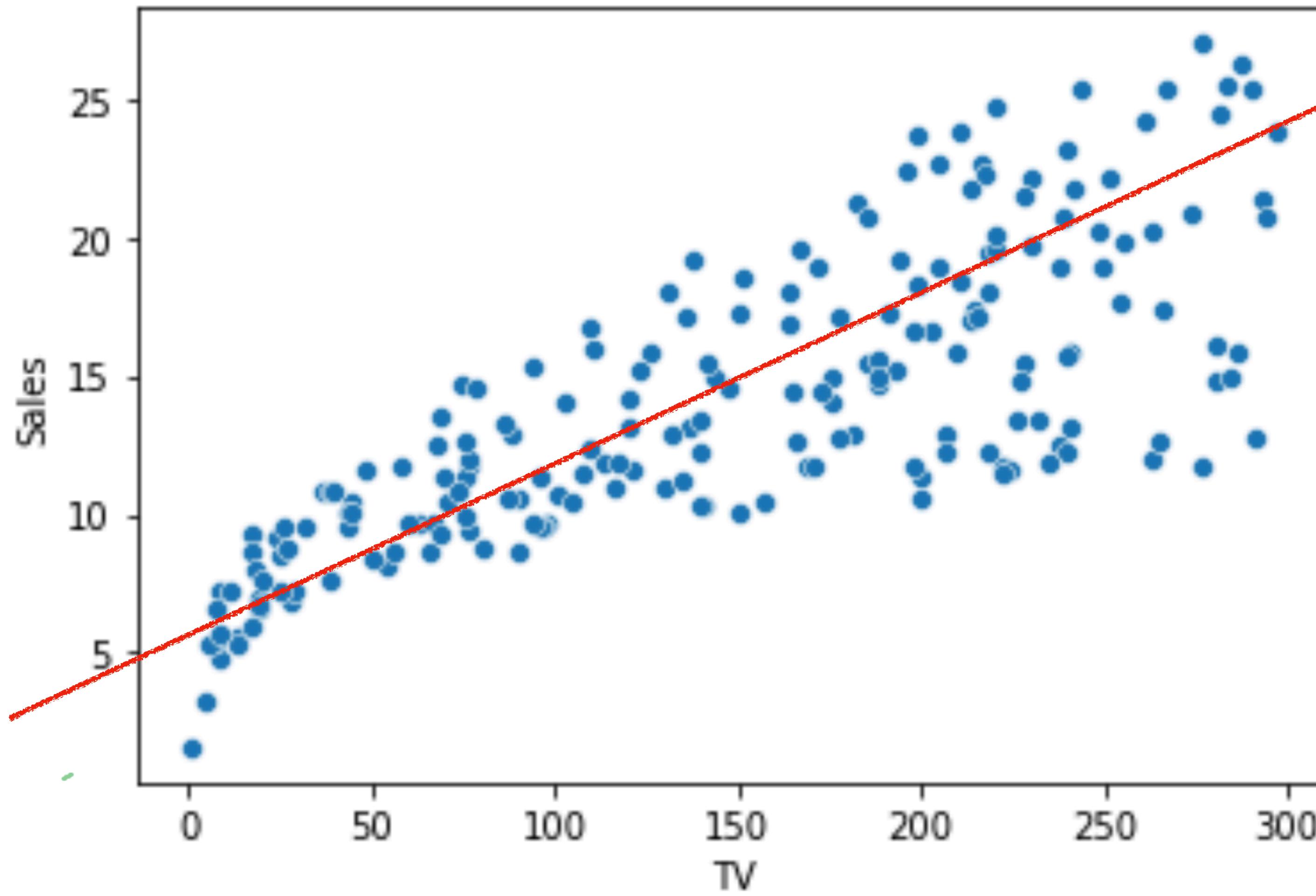
# Regresión lineal

# Regresión lineal simple

## Mínimos cuadrados

- ¿Cómo podemos encontrar los parámetros del modelo lineal?
- La diferencia entre el valor real y el estimado (residuales) se puede escribir como:
- $E = (\hat{y}_i - y_i)$ , donde  $\hat{y}_i$  son los valores predecidos por cada  $x_i$ , y  $y_i$  son los valores históricos (asociados a cada  $x_i$ ).
- El objetivo es minimizar la suma de errores al cuadrado sobre todos los puntos del data set  $X = \{(x_i, y_i)\}_{i=1}^n$
- $$\min \sum_{i=1}^n E^2 = \sum_{i=1}^n (\hat{y}_i(x_i) - y_i)^2 = \sum_{i=1}^n ((\beta_0 + \beta_1 x_i) - y_i)^2 = e$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$



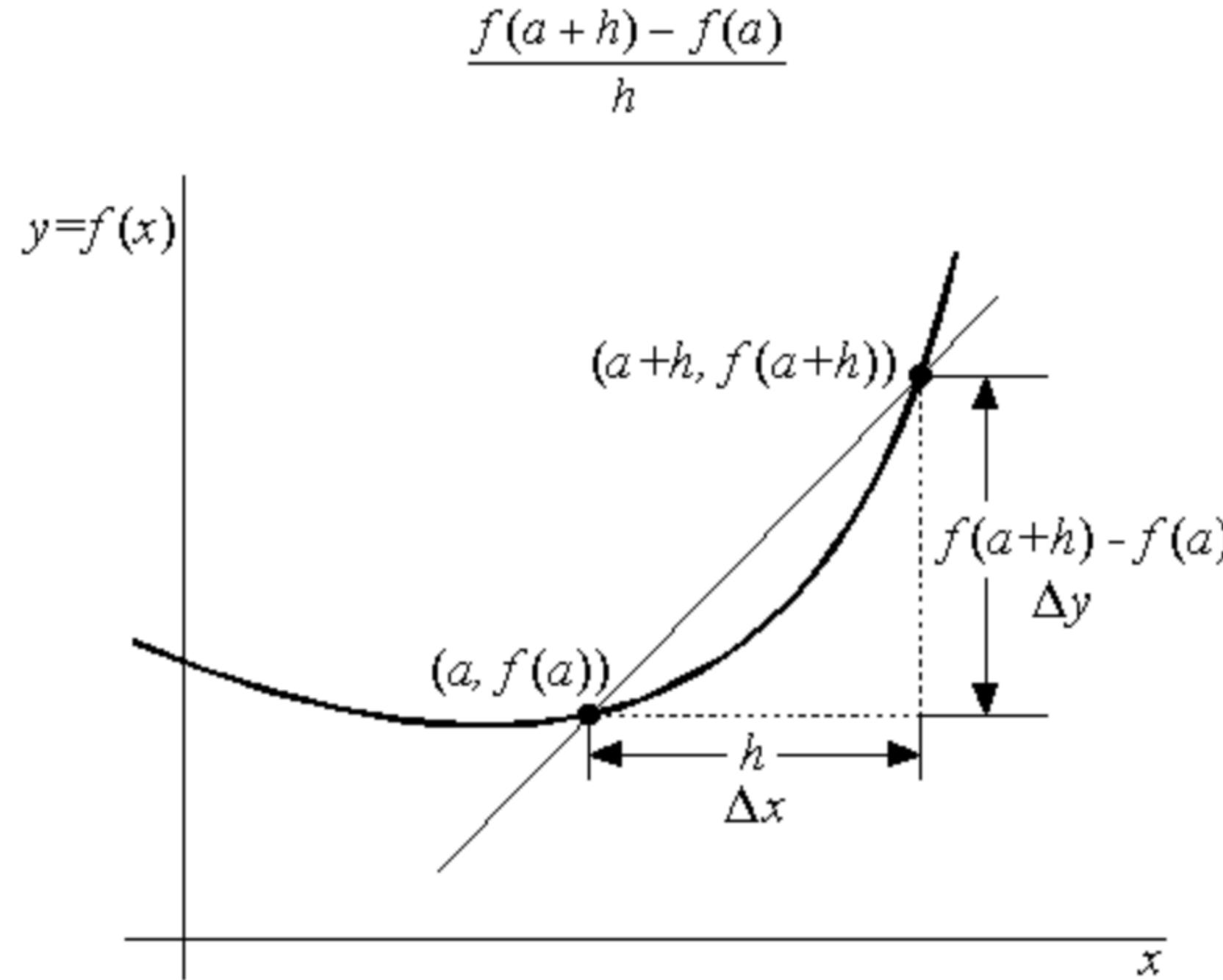
$$y = b + mx$$

Sales = 5 + 0.1 TV

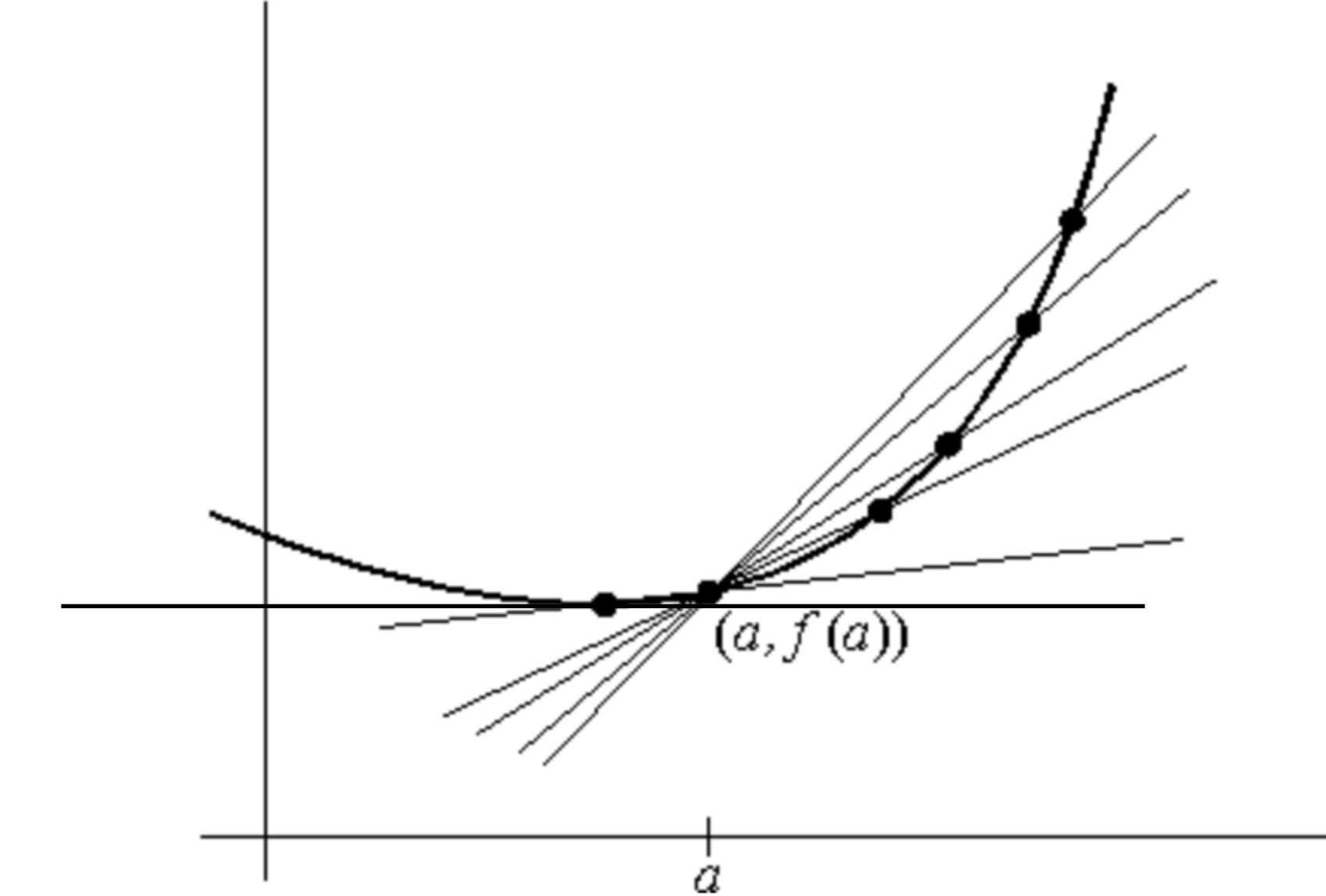
# Regresión lineal simple

## Mínimos cuadrados

- La suma de residuales al cuadrado se tiene que minimizar, y para eso derivamos para encontrar los parámetros óptimos  $\beta_0$  y  $\beta_1$



$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$



# Regresión lineal simple

## Parámetros óptimos

- Los parámetros que minimizan la suma de los residuos al cuadrado son:

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$A = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

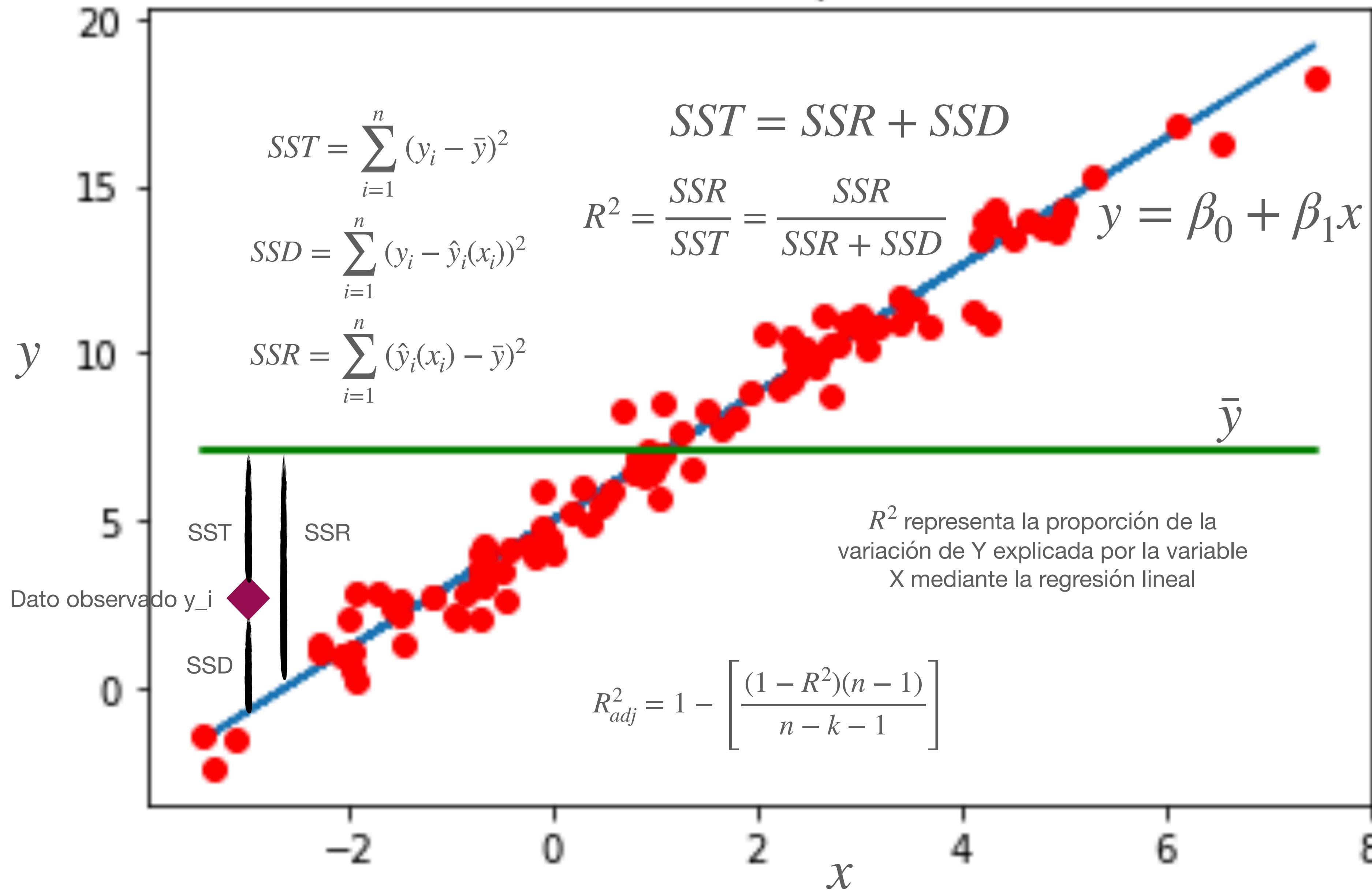
$$B = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

# Regresión lineal

## La componente de error

- En un mundo ideal, el modelo sería perfectamente lineal  $\hat{y} = \beta_0 + \beta_1 x$ .
- Pero en realidad, siempre tendremos una componente de error o residuo  $\epsilon$ .
- $y = \beta_0 + \beta_1 x + \epsilon$
- El residuo  $\epsilon$  será una variable aleatoria con distribución normal.

## Valor actual vs predicción



# Regresión lineal simple

## Suposiciones

- **Libre de error.** Las variables predictoras son libre de error, i. e., que no son variables aleatorias.
- **Linealidad.**
- **Varianza constante.** El error en la variable de respuesta es constante sobre los valores de entrada.
- **No hay multicolinealidad.** Ninguno de los parámetros de entrada son redundantes uno de otro.

# Regresión lineal múltiple

## Overview

- La regresión lineal que hemos visto tiene la forma
- $y_{model} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$
- Se podría pensar que todas las variables toman valores numéricos.
- Se pueden usar variables dummy para variables categóricas.

# Pruebas de hipótesis

# Prueba de hipótesis

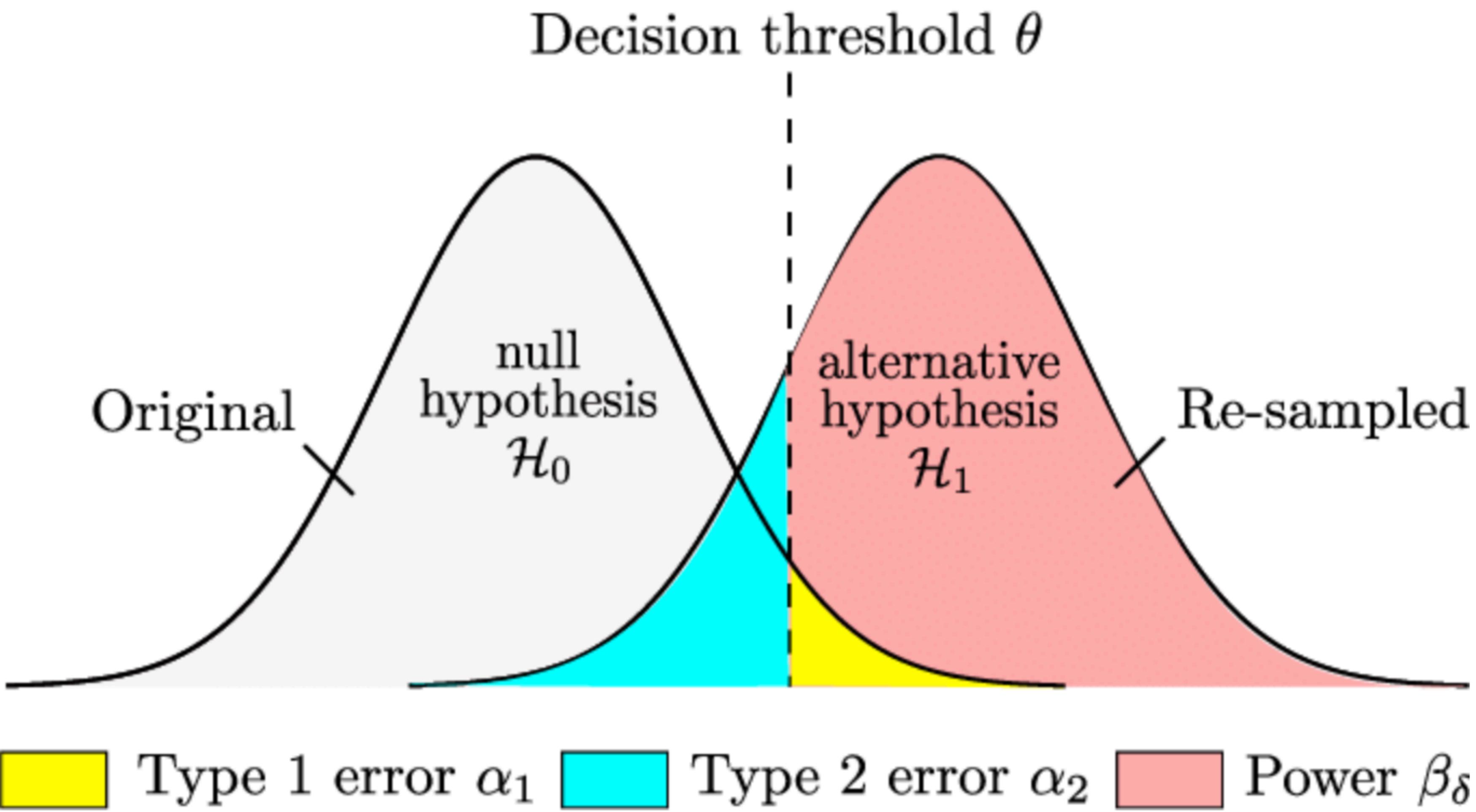
## Overview

- Redactar la hipótesis **nula**  $H_0$  y **alternativa**  $H_1$  según lo que se pretende analizar.
- Decidir un nivel de **significancia** para la prueba.
- Realizar un experimento. Recolectar información de una **muestra** de datos.
- Calcular una **estadística de prueba** según el modelo utilizado.

# Prueba de hipótesis

## Overview

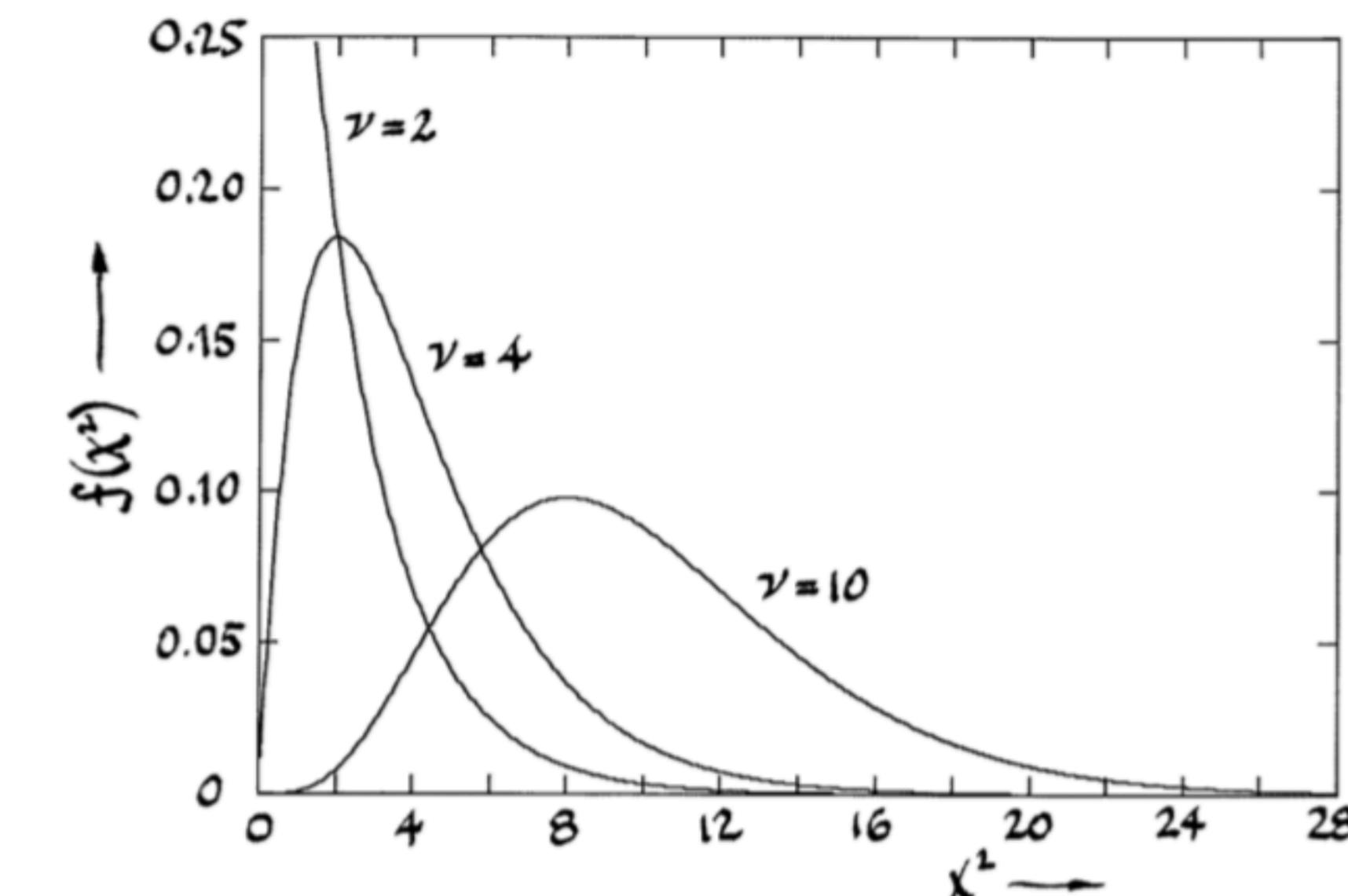
- Comparar la estadística de prueba con la **distribución asociada** bajo la **suposición** que se hizo sobre la hipótesis nula.
- Calcular el **p-valor** o de manera equivalente, **la región crítica**.
- **Rechazar** la hipótesis nula si el p-valor es menor que el nivel de significancia, o equivalentemente, si la estadística de prueba se encuentra en la región crítica.
- En caso contrario, se concluye que no hay suficiente información estadística para rechazar la hipótesis nula.



# Prueba de hipótesis

## Prueba chi cuadrada para independencia entre variables

- **Ejemplo:** Relación entre educación y un buen salario. Una no siempre causa la otra, pero en general se tiene una asociación entre ellas.
- Esta prueba comienza con la hipótesis de no asociación entre las dos variables a considerar. Esto quiere decir que un cambio en una no se traduce en un cambio predictivo en la otra variable.



# Prueba de hipótesis

## Prueba chi cuadrada para independencia entre variables

- Un experimento de psicología se realiza para investigar los efectos de la ansiedad en los deseos de una persona por estar sola o en compañía.. Una muestra de 30 personas es dividida al azar es dos grupos de tamaños 13 (Ansiedad **Baja**) y 17 (Ansiedad **Alta**).
- Al grupo de 17 personas se le dice que va a experimentar unos shocks eléctricos.
- Al otro grupo de 13, se les dice que van a experimentar también shocks eléctricos, pero que estos serán sin dolor y casi notorios.
- A ambos grupos se les dijo que se tienen que esperar 10 minutos antes de comenzar la prueba, y a cada participante de le dio la opción de esperar **solos o acompañado** de los demás participantes. Los datos recolectado del experimento son mostrados en la siguiente **tabla de contingencia**:

	ACOMPAÑADO (S)	SOLO (N)
Ansiedad Alta (A)	12	5
Ansiedad Baja (B)	4	9

# Pruebas de hipótesis

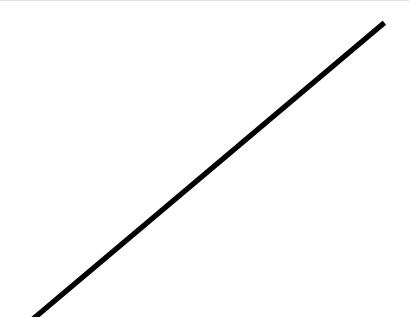
## Hipótesis nula y alternativa

- La hipótesis en esta situación es si la ansiedad sentida está asociada con el deseo de una persona de estar solo o acompañado. Podríamos pensar que la proporción de personas que piden estar acompañados sean más aquellos que experimentan mayor nivel de ansiedad.
- $H_0$  : Las dos variables son independientes
- $H_1$  : Las dos variables no son independientes
- Pedimos un nivel de significancia del 5%,  $\alpha = 0.05$ .

	ACOMPAÑADO (S)	SOLO (N)	
Ansiedad Alta (A)	$(17/30)(16/30)$	$(17/30)(14/30)$	<b><math>P(A) = 17/30</math></b>
Ansiedad Baja (B)	$(13/30)(16/30)$	$(13/30)(14/30)$	<b><math>P(B) = 13/30</math></b>
	<b><math>P(S) = 16/30</math></b>	<b><math>P(N) = 14/30</math></b>	

	ACOMPAÑADO (S)	SOLO (N)	
Ansiedad Alta (A)	0.3022	0.2644	<b><math>P(A) = 17/30</math></b>
Ansiedad Baja (B)	0.2311	0.2022	<b><math>P(B) = 13/30</math></b>
	<b><math>P(S) = 16/30</math></b>	<b><math>P(N) = 14/30</math></b>	

	ACOMPAÑADO (S)	SOLO (N)
Ansiedad Alta (A)	9.07	7.93
Ansiedad Baja (B)	6.93	6.07



# Pruebas de hipótesis

## Prueba de independencia

ACOMPA $\sim$ SOLO (N)	
Ansiedad	12
Ansiedad	5

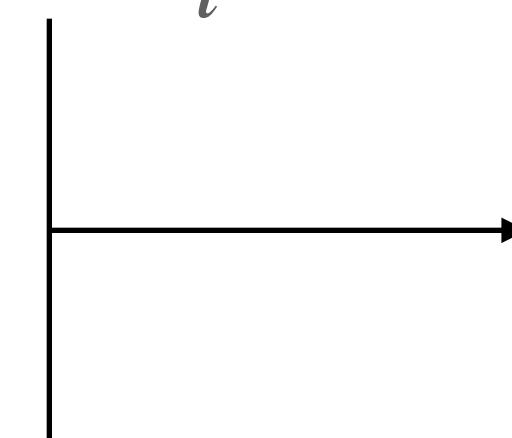
ACOMPAÑ SOLO (N)	
Ansieda	9.07
Ansieda	7.93

ACOMPAÑ SOLO (N)	
Ansieda	6.93
Ansieda	6.07

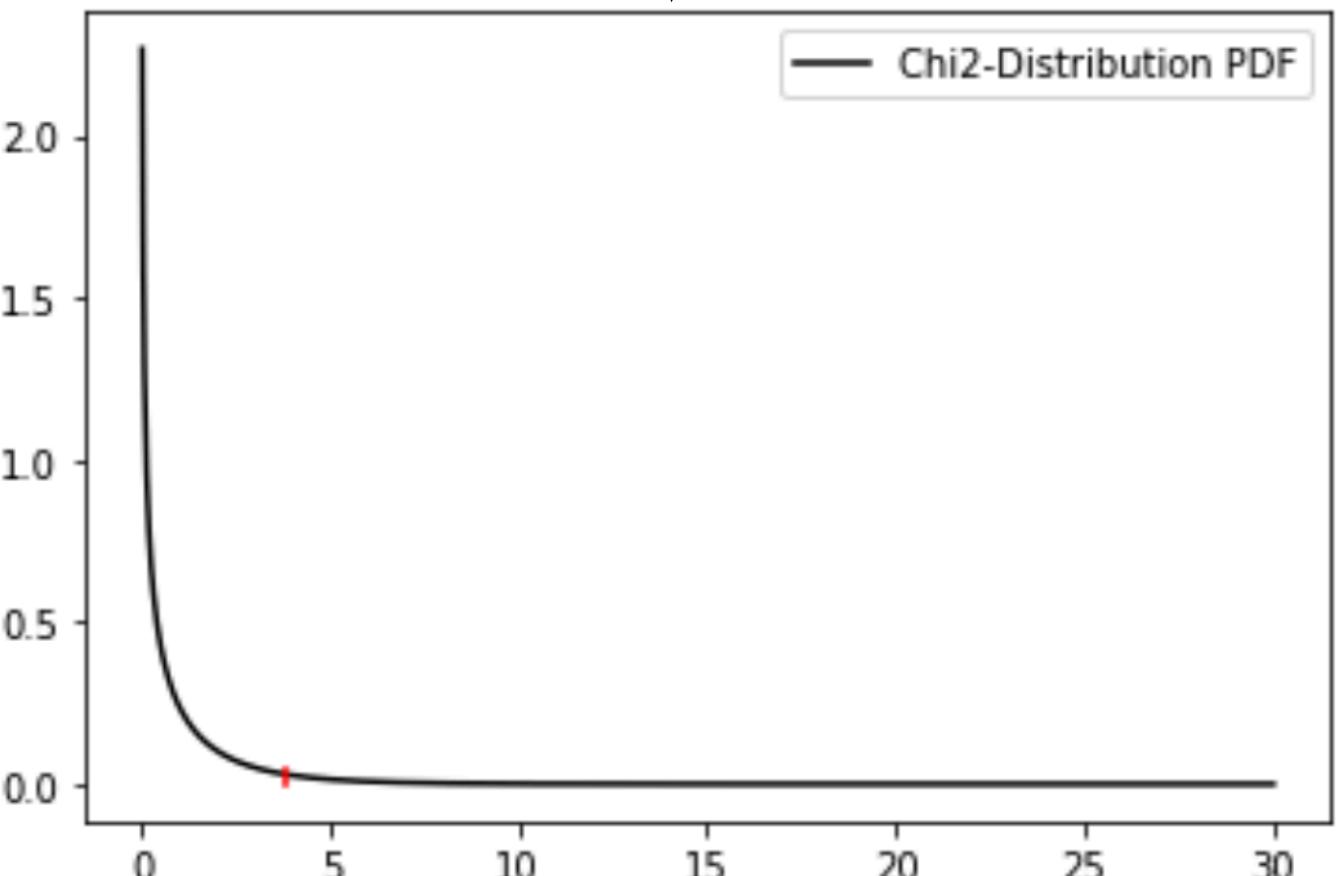
- El valor obtenido  $\chi^2 = 4.6822$  está por encima del valor crítico 3.84. para un nivel de significancia del 5%. Sólo se tiene 1 grado de libertad.
- **Concluimos que hay suficiente evidencia estadística para rechazar la hipótesis nula con un nivel del 5%. Esto apoya la teoría que estas variables tienen algún tipo de relación.**

Valores observados  $O_i$



Valores esperados  $E_i$

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 4.6822$$



# Pruebas de hipótesis

## Coeficiente de contingencia

- El coeficiente de contingencia es un coeficiente de asociación que nos dice si dos variables o dos datasets son dependientes o independientes uno del otro. Está basado en el estadístico chi square, y se define por:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

- Si  $C$  es cercano (o igual) a cero, se concluye que las variables son independientes una de la otra; que no hay asociación entre ellas.
- Si  $C$  se encuentra lejano a cero, existe algún tipo de relación entre las variables.  $C$  sólo puede tener valores positivos.
- **Coeficiente V de Cramer**

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

- **Coeficiente Phi**

$$\phi = \sqrt{\frac{\chi^2}{N}}$$