Universidad Nacional de Mar del Plata

Facultad de Ingeniería

# Estadística Descriptiva

REPASO

# ¿Qué es la estadística?

"La estadística, como campo de estudio, es el arte y la ciencia de dar sentido a los datos numéricos"

*Hildebrand,* Estadística Aplicada a la Administración y a la economía.(1997)

# ¿Qué es la estadística?

" El contenido de la estadística moderna incluye la recopilación, presentación y caracterización de la información con el fin de auxiliar tanto en el análisis de datos como en el proceso de toma de decisiones"

*Berenson y Levine,* Estadística Básica en administración. (1992)

# Algunas definiciones

- *Población*: es el conjunto sobre el que se centra el objetivo de un análisis o investigación estadística. Esta compuesta por unidades elementales con características comunes observables.
- Una *unidad elemental* es cada objeto o sujeto que observamos de la población.
- Una *muestra* es un subconjunto «representativo» de unidades elementales tomadas de la población.
- Aquella característica que se observa o se mide sobre las unidades elementales, se denomina *variable estadística*.

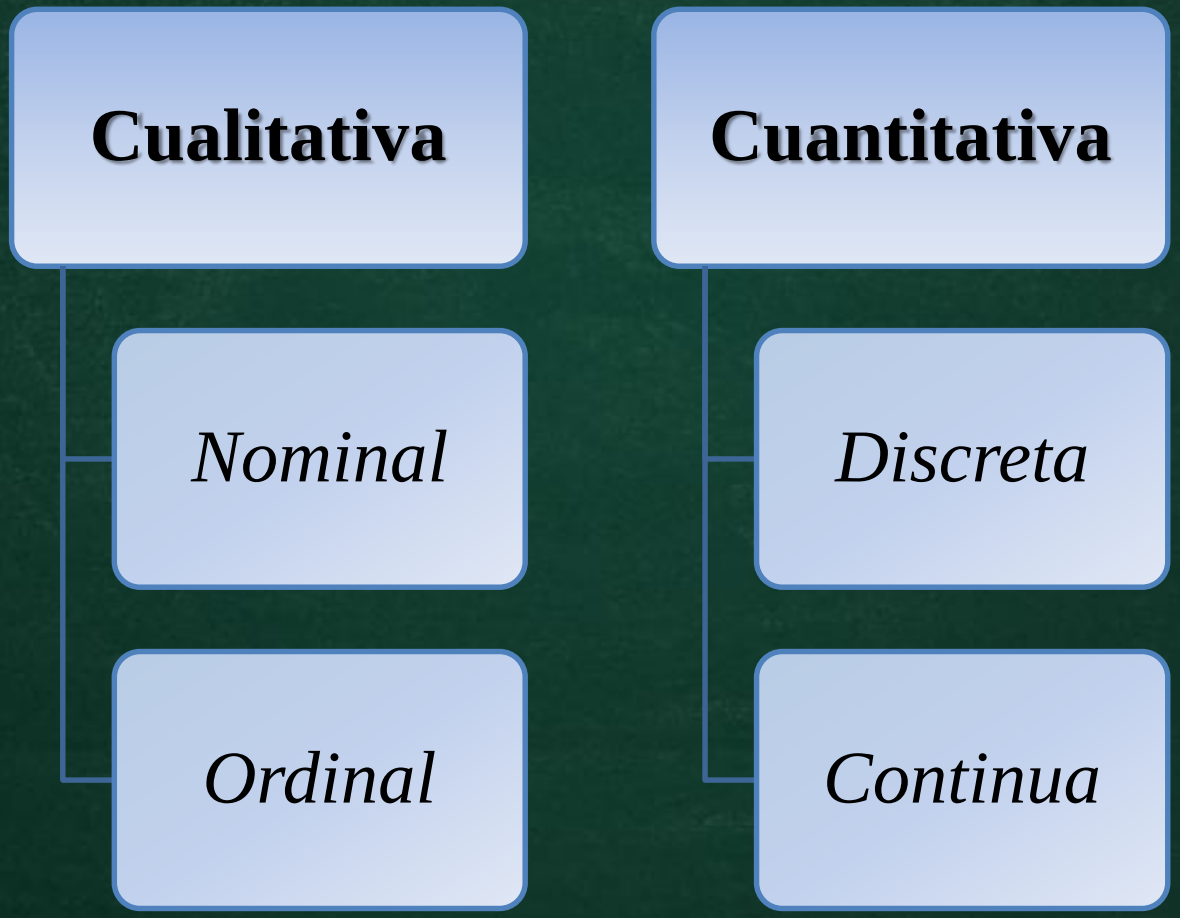
# Tipos de Variables

Variables: ***Cualitativas*** y ***Cuantitativas***.

- Las variables *cualitativas* son aquellas que permiten la expresión de una característica, una categoría, un atributo o una cualidad de los elementos de estudio.

- Las variables *cuantitativas* son aquellas cuyos datos son de tipo numérico.

# Clasificación de las Variables



# Organización de los datos:

Debido a la cantidad de datos que pueden ser generados en el proceso de investigación, es posible que a simple vista se dificulte la interpretación de los mismos. Por esta razón, es conveniente tabular los datos para facilitar su interpretación.

# ¿Cómo organizar los datos?

- Series Simples
- Series de Frecuencias
- Intervalos de Clases

# Organización de los datos:

- Serie de frecuencias

Ejemplo:

En una encuesta realizada a 1.509 familias se les pregunta por el número de hijos

¿Cuántas familias tienen menos de 2 hijos?
¿Qué porcentaje de familias tiene 6 hijos o menos?
¿Qué cantidad de hijos tiene al lo sumo el 95% de las familias encuestadas?

# Organización de los datos:

- Serie de frecuencias

| X | fi | fr = fi/n | (100.fr) % | Fa | Fa % |
|---|---|---|---|---|---|
| 0 | 419 | 0,278 | 27,8% | 419 | 0,278% |
| 1 | 255 | 0,169 | 16,9% | 674 | 44,7% |
| 2 | 375 | 0,249 | 24,9% | 1049 | 69,5% |
| 3 | 215 | 0,142 | 14.2% | 1264 | 83,8% |
| 4 | 127 | 0,084 | 8,4% | 1341 | 92,2% |
| 5 | 54 | 0,036 | 3,6% | 1445 | 95,8% |
| 6 | 24 | 0,016 | 1,6% | 1469 | 97,3% |
| 7 | 23 | 0,015 | 1,5% | 1492 | 98,9% |
| 8 o + | 17 | 0,011 | 1,1% | 1509 | 100% |
| $\sum$ | 1.509 | 1 | 100% | | |

¿Cuántas familias tienen menos de 2 hijos?

¿Qué porcentaje de familias tiene 6 hijos o menos?

¿Qué cantidad de hijos tiene al lo sumo el 95% de las familias encuestadas?

# Organización de los datos:

- Serie de frecuencias

| X | fi | fr = fi/n | (100.fr) % | Fa | Fa % |
|---|---|---|---|---|---|
| 0 | 419 | 0,278 | 27,8% | 419 | 0,278% |
| 1 | 255 | 0,169 | 16,9% | 674 | 44,7% |
| 2 | 375 | 0,249 | 24,9% | 1049 | 69,5% |
| 3 | 215 | 0,142 | 14.2% | 1264 | 83,8% |
| 4 | 127 | 0,084 | 8,4% | 1341 | 92,2% |
| 5 | 54 | 0,036 | 3,6% | 1445 | 95,8% |
| 6 | 24 | 0,016 | 1,6% | 1469 | 97,3% |
| 7 | 23 | 0,015 | 1,5% | 1492 | 98,9% |
| 8 o + | 17 | 0,011 | 1,1% | 1509 | 100% |
| $\sum$ | 1.509 | 1 | 100% | | |

¿Cuántas familias tienen menos de 2 hijos?

# Organización de los datos: ▪ Serie de frecuencias

| X | fi | fr = fi/n | (100.fr) % | Fa | Fa % |
|---|---|---|---|---|---|
| 0 | 419 | 0,278 | 27,8% | 419 | 0,278% |
| 1 | 255 | 0,169 | 16,9% | 674 | 44,7% |
| 2 | 375 | 0,249 | 24,9% | 1049 | 69,5% |
| 3 | 215 | 0,142 | 14.2% | 1264 | 83,8% |
| 4 | 127 | 0,084 | 8,4% | 1341 | 92,2% |
| 5 | 54 | 0,036 | 3,6% | 1445 | 95,8% |
| 6 | 24 | 0,016 | 1,6% | 1469 | 97,3% |
| 7 | 23 | 0,015 | 1,5% | 1492 | 98,9% |
| 8 o + | 17 | 0,011 | 1,1% | 1509 | 100% |
| $\sum$ | 1.509 | 1 | 100% | | |

¿Cuántas familias tienen menos de 2 hijos?

frecuencia familias con 0 hijos
+
frecuencia familias con 1 hijo

**674 familias**

**Es decir, el 44,7 % de las familias encuestadas tiene menos de 2 hijos**

# Organización de los datos:

| X | fi | fr = fi/n | (100.fr) % | Fa | Fa % |
|---|---|---|---|---|---|
| 0 | 419 | 0,278 | 27,8% | 419 | 0,278% |
| 1 | 255 | 0,169 | 16,9% | 674 | 44,7% |
| 2 | 375 | 0,249 | 24,9% | 1049 | 69,5% |
| 3 | 215 | 0,142 | 14.2% | 1264 | 83,8% |
| 4 | 127 | 0,084 | 8,4% | 1341 | 92,2% |
| 5 | 54 | 0,036 | 3,6% | 1445 | 95,8% |
| 6 | 24 | 0,016 | 1,6% | 1469 | 97,3% |
| 7 | 23 | 0,015 | 1,5% | 1492 | 98,9% |
| 8 o + | 17 | 0,011 | 1,1% | 1509 | 100% |
| $\sum$ | 1.509 | 1 | 100% | | |

¿Qué porcentaje de familias tiene 6 hijos o menos?

# Organización de los datos:

| X | fi | fr = fi/n | (100.fr) % | Fa | Fa % |
|---|---|---|---|---|---|
| 0 | 419 | 0,278 | 27,8% | 419 | 0,278% |
| 1 | 255 | 0,169 | 16,9% | 674 | 44,7% |
| 2 | 375 | 0,249 | 24,9% | 1049 | 69,5% |
| 3 | 215 | 0,142 | 14.2% | 1264 | 83,8% |
| 4 | 127 | 0,084 | 8,4% | 1341 | 92,2% |
| 5 | 54 | 0,036 | 3,6% | 1445 | 95,8% |
| 6 | 24 | 0,016 | 1,6% | 1469 | 97,3% |
| 7 | 23 | 0,015 | 1,5% | 1492 | 98,9% |
| 8 o + | 17 | 0,011 | 1,1% | 1509 | 100% |
| $\sum$ | 1.509 | 1 | 100% | | |

¿Qué porcentaje de familias tiene 6 hijos o menos?

**Es decir, el 97,3% de las familias encuestadas tiene menos de 6 hijos.**

# Organización de los datos:

| X | fi | fr = fi/n | (100.fr) % | Fa | Fa % |
|---|---|---|---|---|---|
| 0 | 419 | 0,278 | 27,8% | 419 | 0,278% |
| 1 | 255 | 0,169 | 16,9% | 674 | 44,7% |
| 2 | 375 | 0,249 | 24,9% | 1049 | 69,5% |
| 3 | 215 | 0,142 | 14.2% | 1264 | 83,8% |
| 4 | 127 | 0,084 | 8,4% | 1341 | 92,2% |
| 5 | 54 | 0,036 | 3,6% | 1445 | 95,8% |
| 6 | 24 | 0,016 | 1,6% | 1469 | 97,3% |
| 7 | 23 | 0,015 | 1,5% | 1492 | 98,9% |
| 8 o + | 17 | 0,011 | 1,1% | 1509 | 100% |
| $\sum$ | 1.509 | 1 | 100% | | |

¿Qué cantidad de hijos tiene al lo sumo el 95% de las familias encuestadas?

**5 hijos**

# ¿Cómo organizar los datos?

Intervalos de Clases

# Organización de los datos:

¿Cómo saber cuántos intervalos considerar? ¿Cómo determinar su amplitud?

Primero debemos determinar el rango de los datos, que es la diferencia entre el mayor y el menor de los valores obtenidos.

$$\text{Rango} = X_{máx} - X_{mín}$$

Luego debemos establecer el número de intervalos (K) y determinar la amplitud (A) de los mismos.

$$K = 1 + 3{,}3 \cdot \log n$$ (regla de **Sturges**)

$$A = \text{Rango} / K$$

# Gráficos

# ¿Cómo organizar los datos?

**Variables Cualitativas**

- Barras Simples
- Barras Proporcionales
- Barras Agrupadas
- Diagramas Sectoriales

**Variables Cuantitativas Discretas**

- Bastones

**Variables Cuantitativas Continuas**

- Histograma
- Polígono de Frecuencias Simples
- Polígono de Frecuencias Acumuladas

# Estadísticos

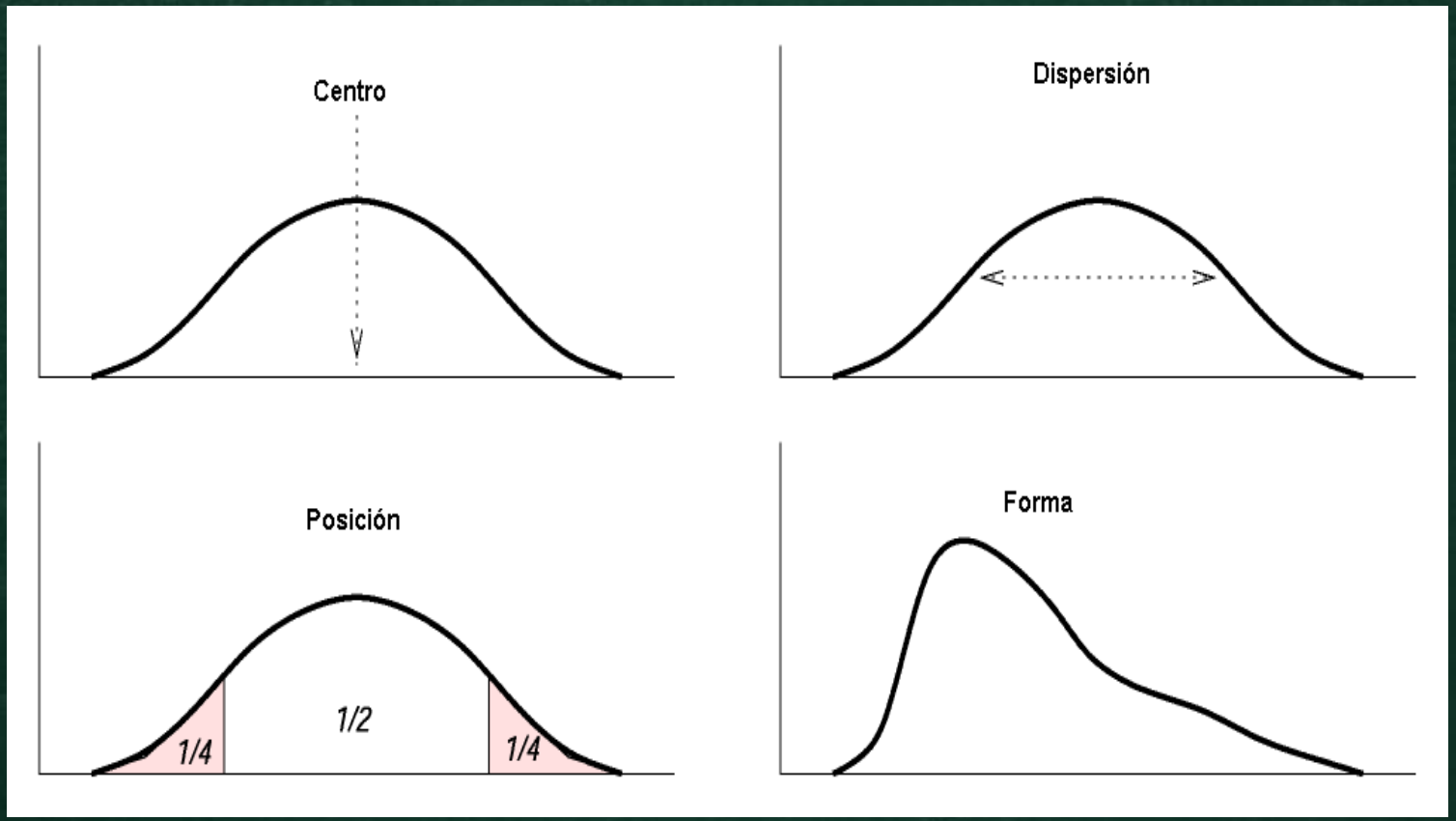
En todo análisis y/o interpretación de datos es necesario disponer de «valores» numéricos para extraer y resumir las principales características de los mismos.

Existen diversas medidas descriptivas que representan las propiedades de tendencia central, dispersión y forma.

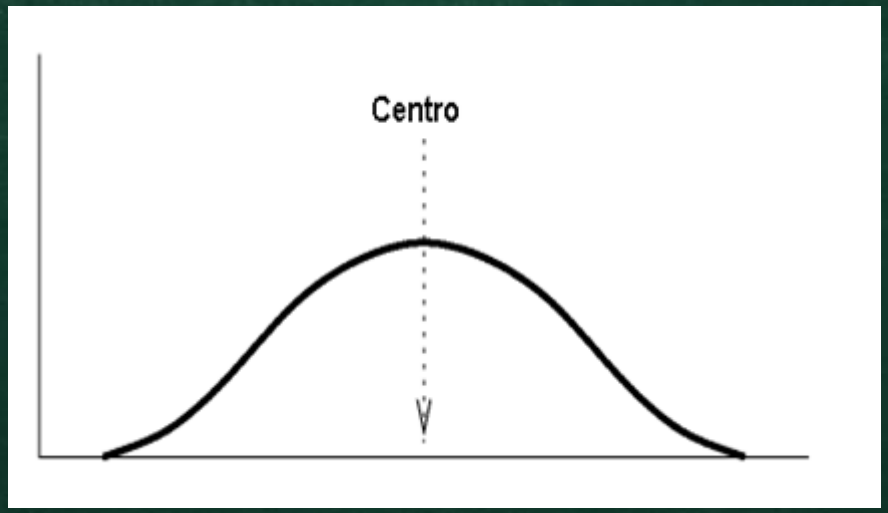
# Estadísticos

- **Centralización**
  - ✓ Indican valores con respecto a los que los datos parecen agruparse.
    - ✓ Media, mediana y moda
- **Posición**
  - ✓ Dividen un conjunto ordenado de datos en grupos de la misma cantidad de individuos.
    - ✓ Cuartiles, deciles, percentiles
- **Dispersión**
  - ✓ Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
    - ✓ Rango, Varianza, Desviación estándar, Coeficiente de Variación.

# Estadísticos



# Centralización



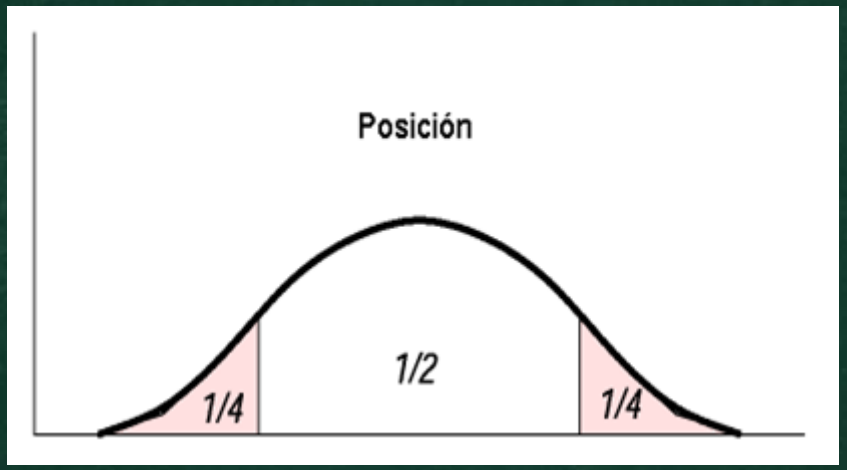
# Medidas de Centralización

En la mayoría de los casos, el conjunto de datos obtenidos, ya sea de una muestra o de una población, tienden a reunirse alrededor de un valor central. De esta manera, es posible obtener un valor típico o representativo de todo el conjunto de datos, el cual se denomina medida de tendencia central.

Las medidas de tendencia central más representativas son:

- ✓ Media aritmética,
- ✓ Mediana,
- ✓ Moda.

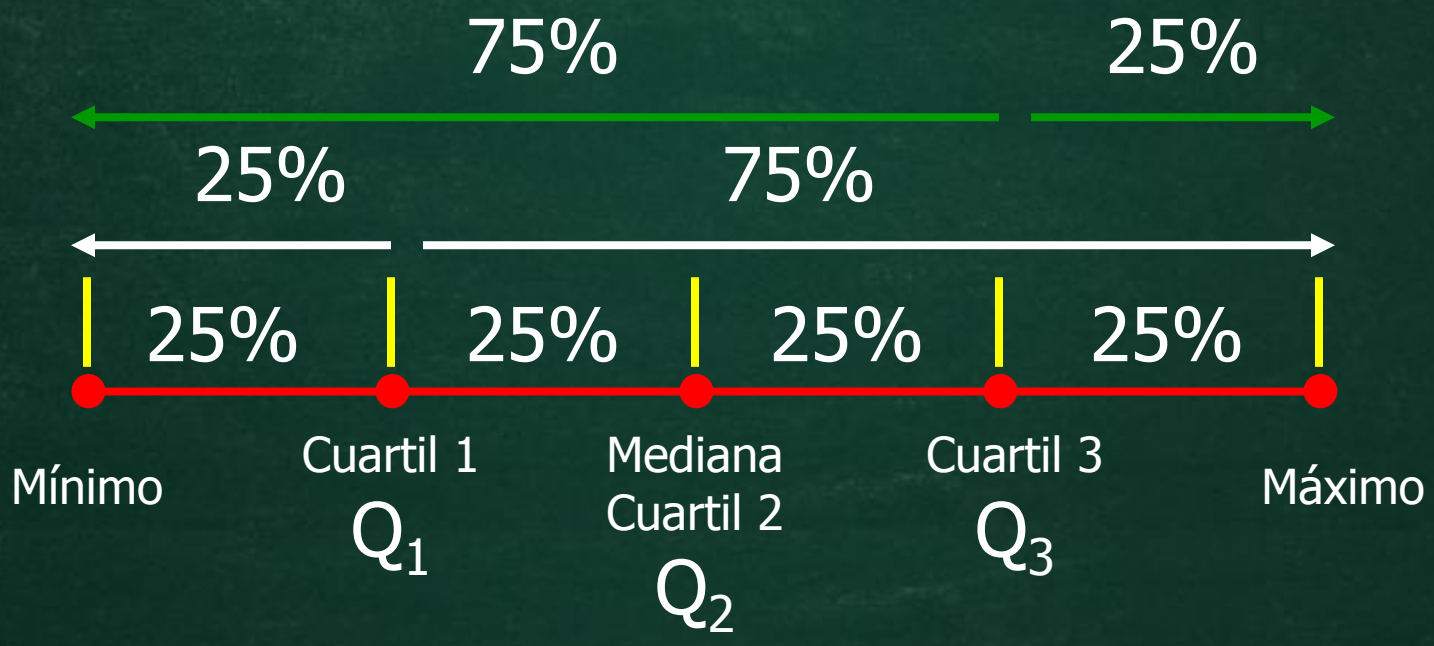
# Medidas de Posición




- ✓ Dividen un conjunto ordenado de datos en grupos.
  - ✓ Cuartiles, deciles, percentiles

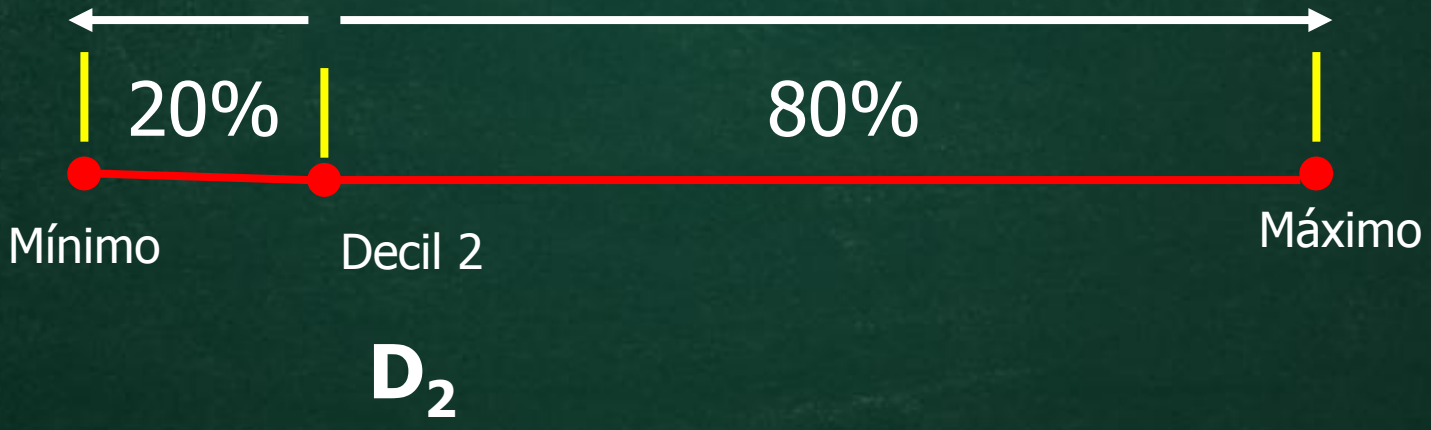
# Cuartiles

Los cuartiles ($Q_k$) son valores que fraccionan la distribución de los datos en cuatro partes iguales. Existen tres cuartiles y cada una de las partes representa un 25% de los datos.



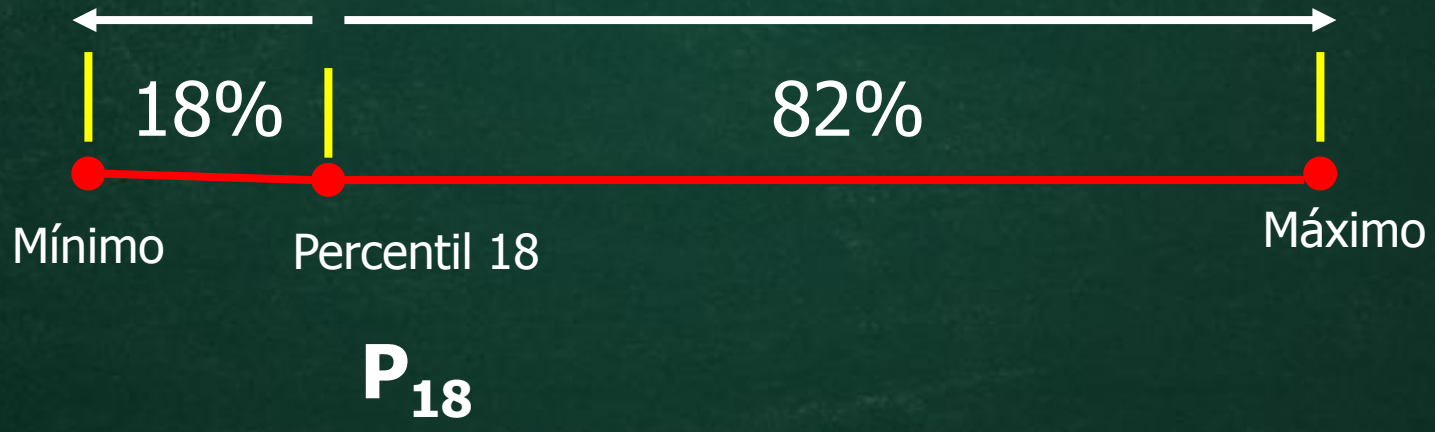
# Deciles

Cuando se divide un conjunto ordenado de datos en diez partes iguales, los puntos de división se conocen como deciles.

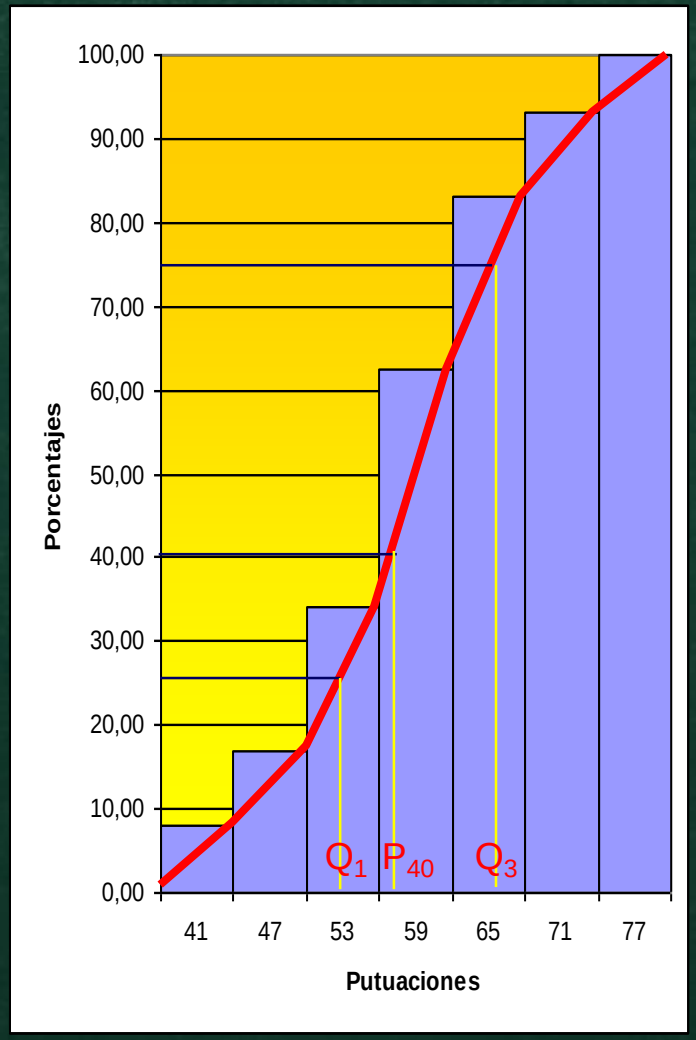


# Percentiles

Cuando se divide un conjunto ordenado de datos en cien partes iguales, los puntos de división se conocen como percentiles.

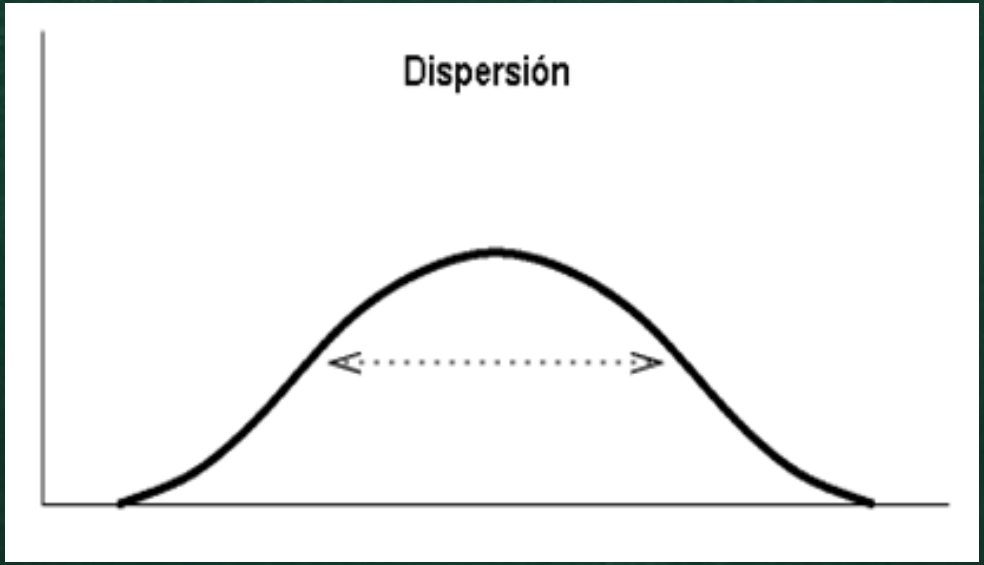


# Mediana y Cuartiles representados en el polígono de frecuencias acumuladas

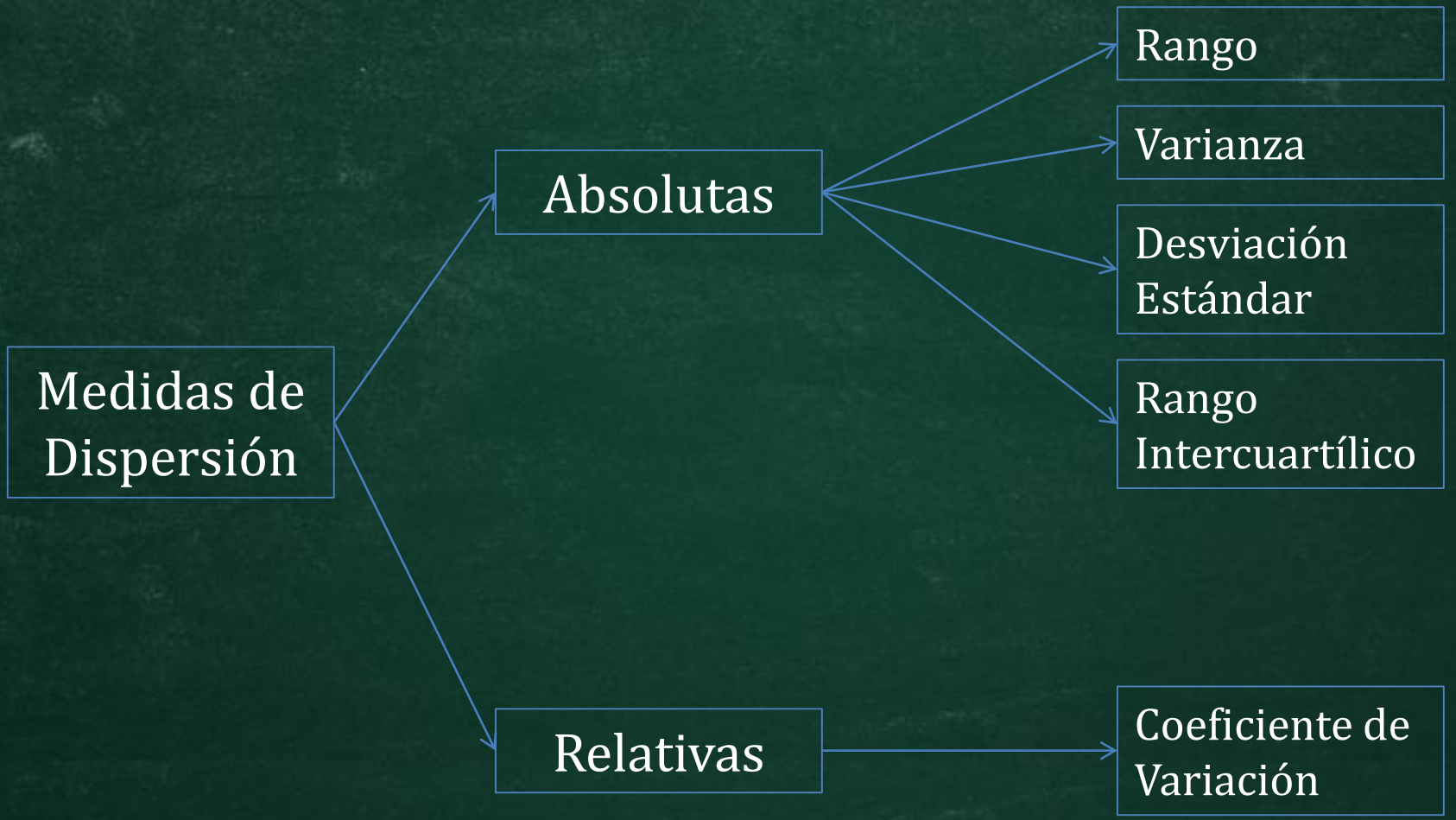


# Medidas de Dispersión

Además de las medidas de tendencia central que posibilitan la representación del conjunto de datos por medio de un valor, es necesario conocer la variabilidad o la dispersión que los datos pueden tener en relación a una medida de tendencia central.



# Medidas de Dispersión



# Rango

**El rango** se define como la diferencia entre la observación más grande y la más pequeña :

$$r = x_{\max} - x_{\min}$$

# Rango Intercuartílico (RIC)

$$RIC = Q_3 - Q_1$$

Los valores extremos NO influyen en el conjunto de datos.

# Varianza

Para el conjunto de datos $x_1, x_2, \ldots, x_n$ de una población de tamaño N. Las diferencias de cada dato y la media, determinan los desvíos o desviaciones. Dado que la suma de estas desviaciones es cero, se utiliza como medida de variabilidad el promedio de los cuadrados de tales desvíos.

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

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

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad (3)$$

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 . f_i}{n-1} \quad (4)$$

**Varianza Poblacional**

siendo N el tamaño de la población.

Para datos sin agrupar (1) y agrupados (2)

**Varianza muestral**

siendo *n* el tamaño de la muestra. Para datos sin agrupar (3) y agrupados (4)

Si los datos se agrupan por intervalos, usamos $x_{mi}$ en lugar de $x_i$

# Desvío estándar muestral

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

Para datos sin agrupar

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \, f_i}$$

Para datos agrupados por frecuencias

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_{mi} - \overline{X})^2 \, f_i}$$

Para datos agrupados por Intervalos

# Coeficiente de variación

El coeficiente de variación ($CV$) es una medida que relaciona la desviación estándar con la media aritmética para determinar qué tan homogénea o dispersa es la información.

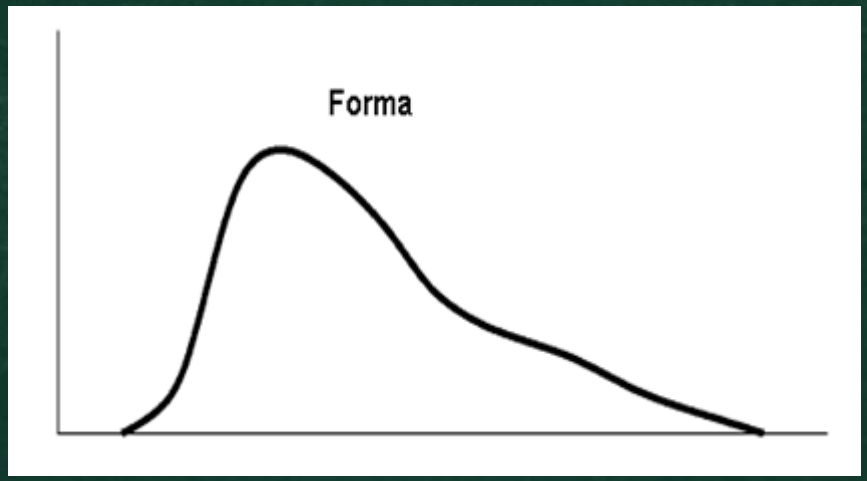
$$CV = \frac{S}{\overline{X}}$$

- Mide el grado de variabilidad en una muestra o población.
- Está desprovisto de unidades. Permite comparar la variabilidad entre distintas variables y poblaciones.
- El valor expresado en términos porcentuales, se llama coeficiente de variación porcentual.

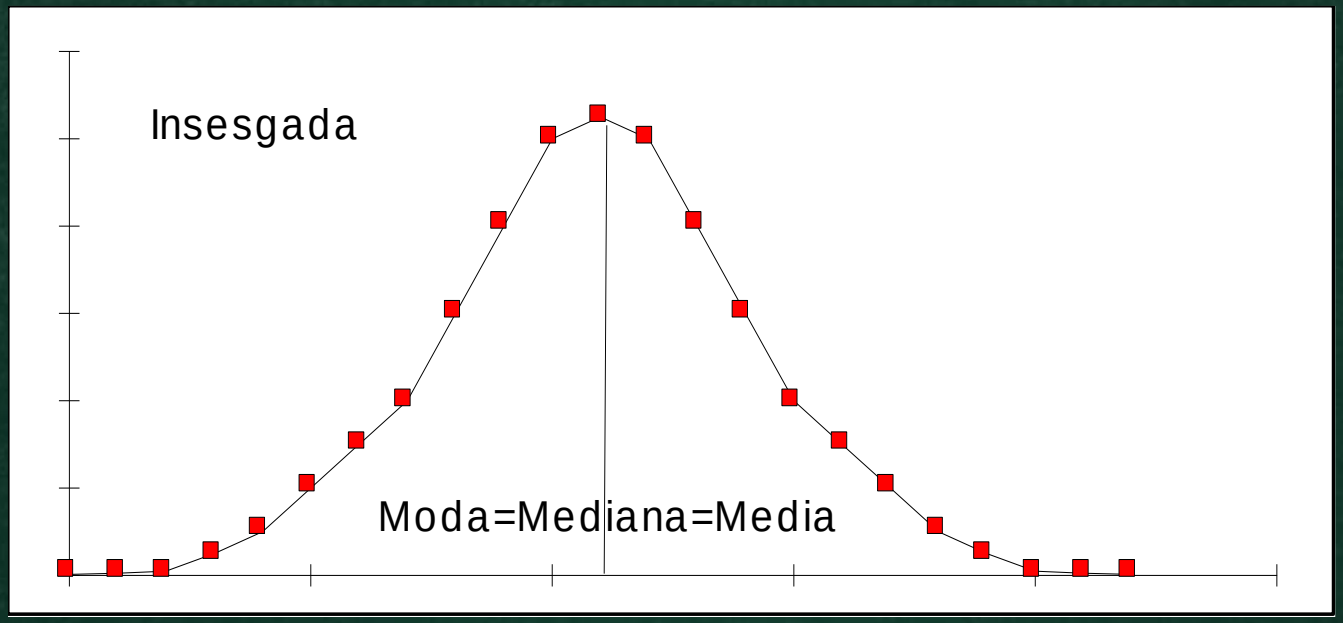
$$CV\% = \frac{S}{\overline{X}} \times 100\%$$

Consideraremos poca variabilidad, si el CV% es a lo sumo del 30 %

# Análisis de la Forma



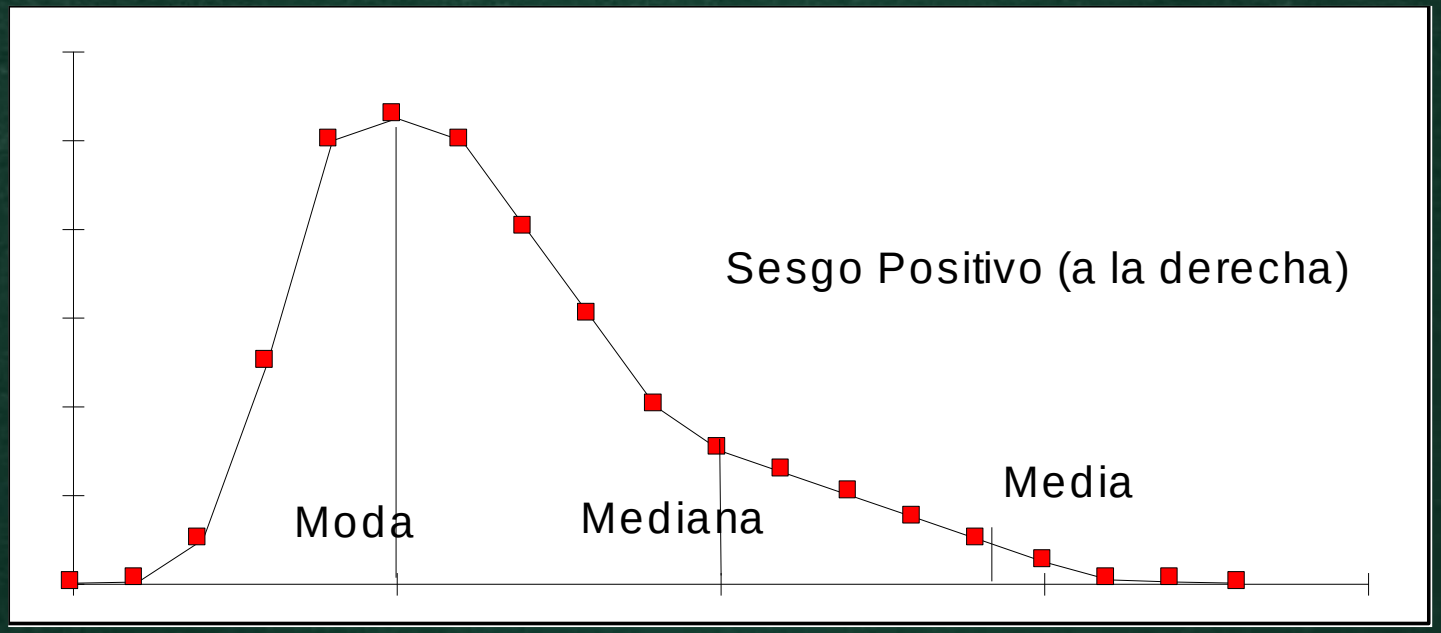
# Distribuciones Simétricas




$$\overline{X} = Me = Mo$$

La distribución de los datos es simétrica

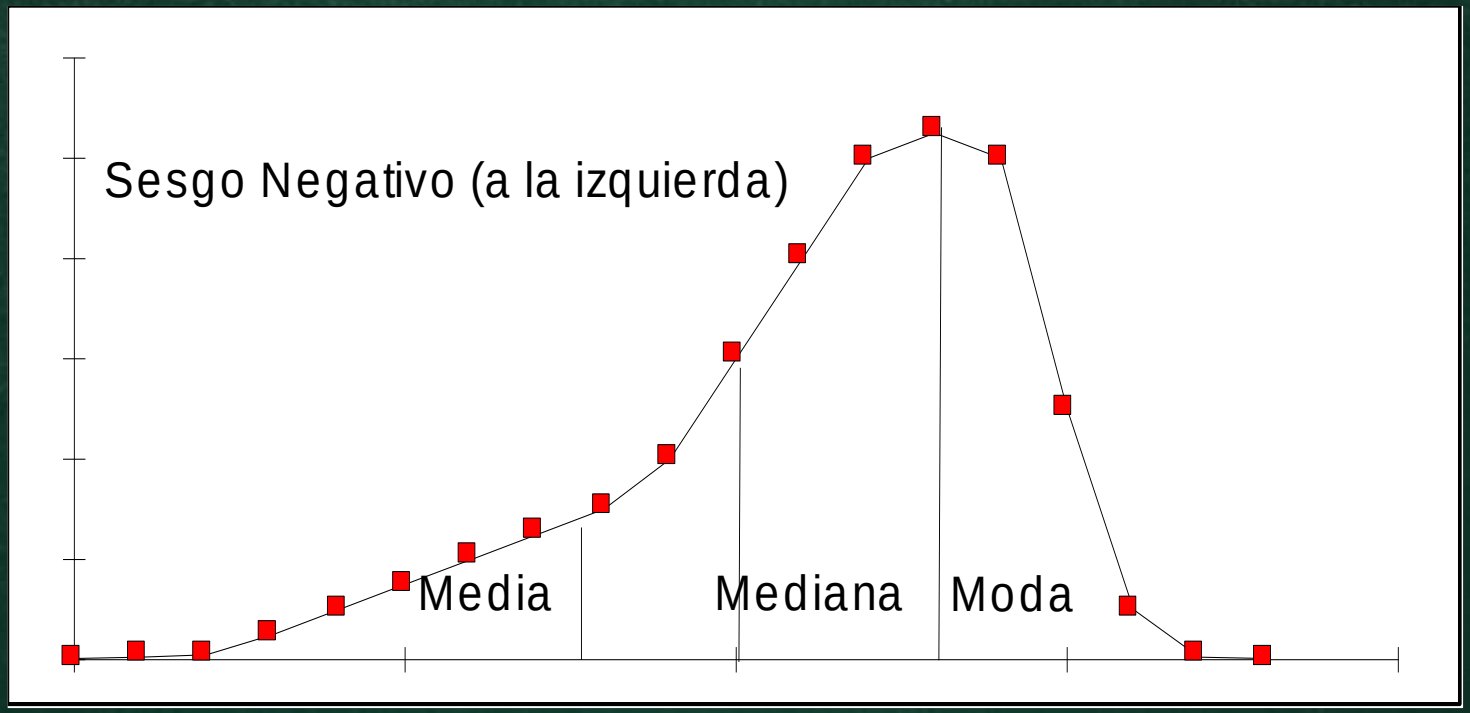
# Distribuciones Asimétricas




$$Si\ \ Mo < Me < \overline{X}\ :\ \ Asimétrica\ Positiva$$

Si la distribución es asimétrica positiva, la media no representa al conjunto de datos.

# Distribuciones Asimétricas



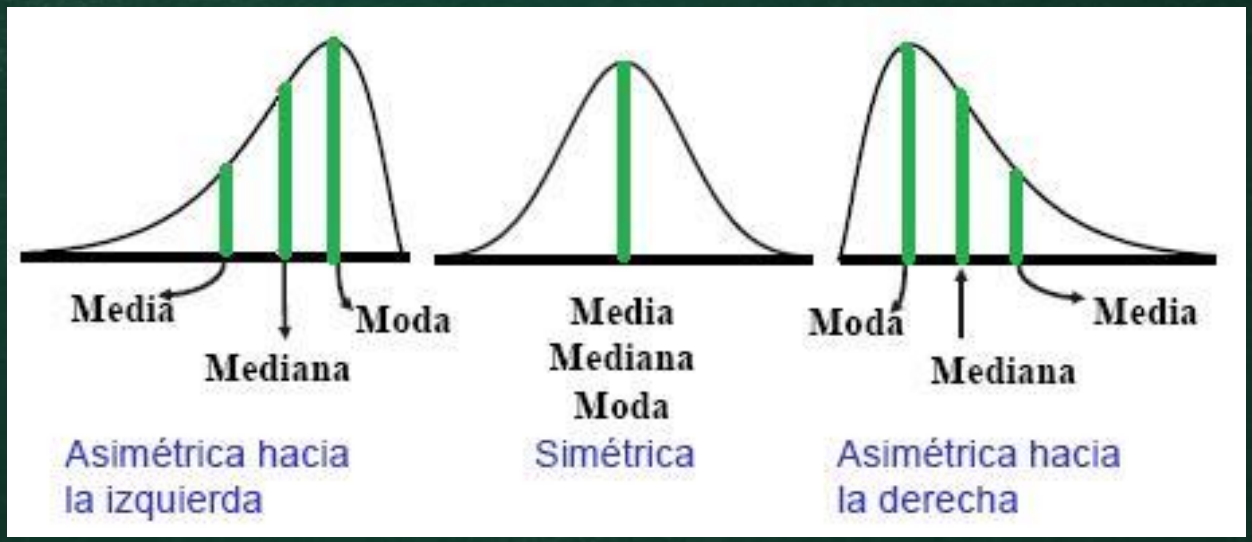

$$Si\ \ \overline{X} < Me < Mo :\ \ Asimétrica\ Negativa$$

Si la distribución es asimétrica negativa,
la media no representa al conjunto de datos.

# Análisis de la simetría

Coeficiente de asimetría

$$A_s = \frac{\bar{x} - M_e}{s}$$

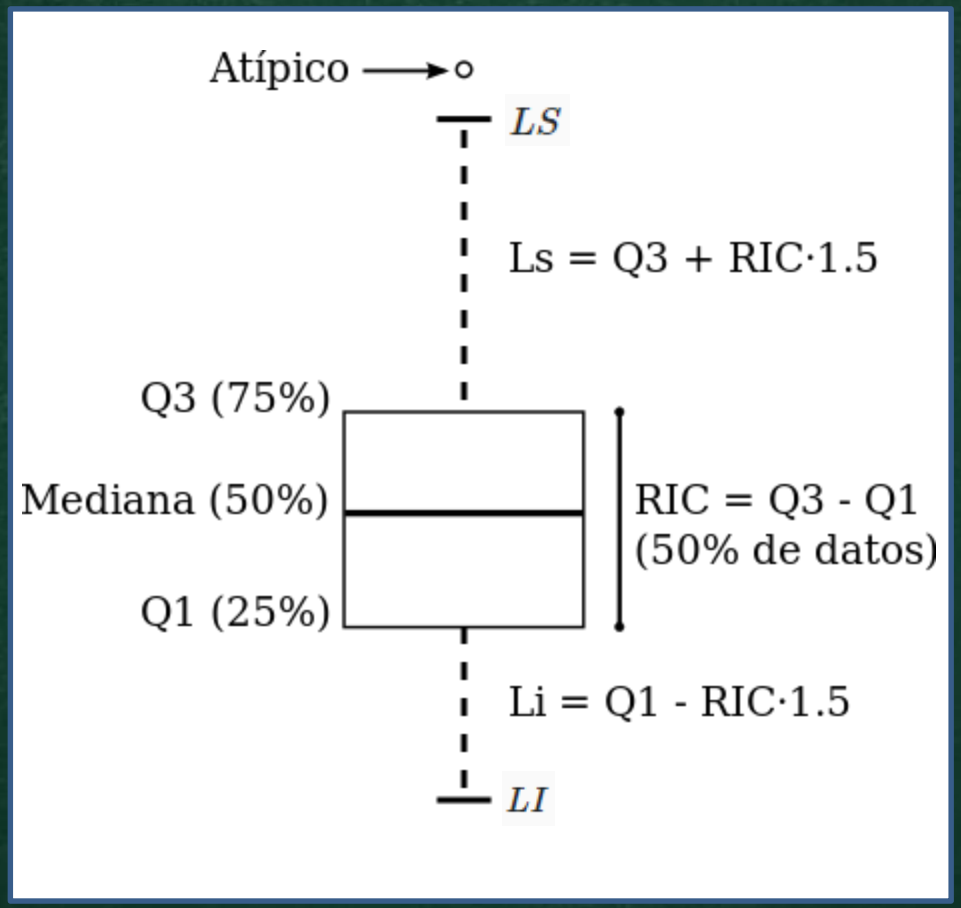



**As <0** As =0 As >0

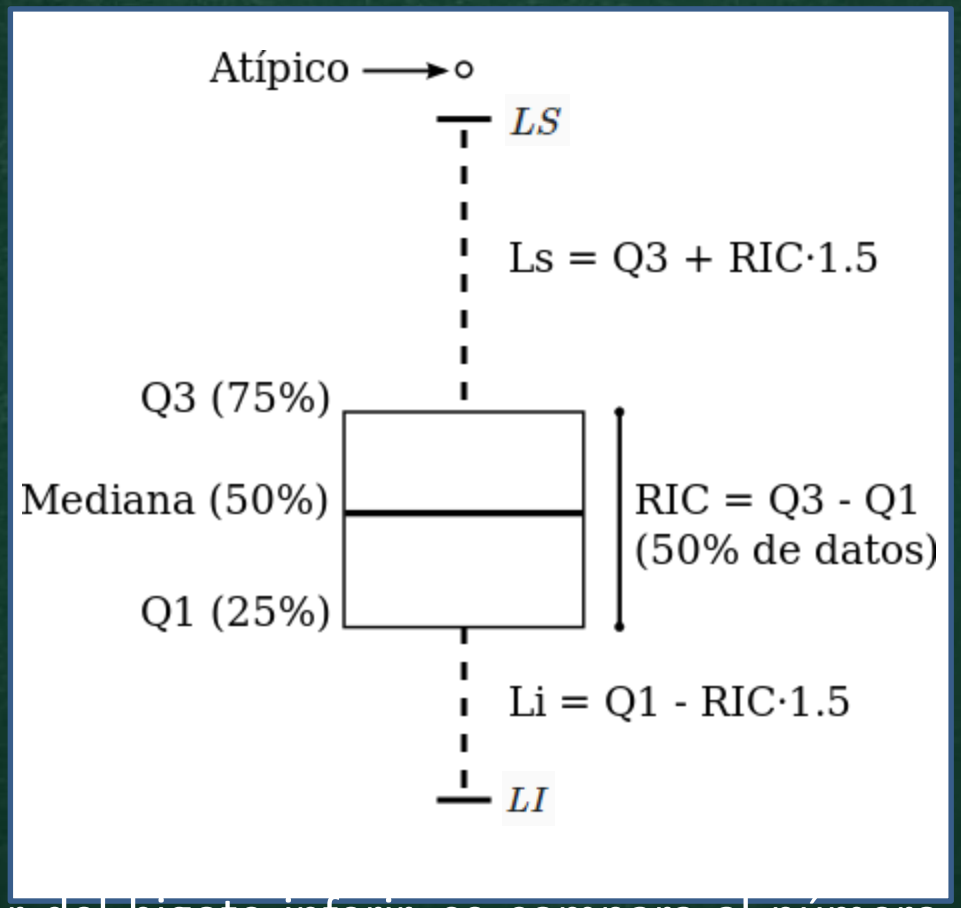
# Observaciones finales

- Comenzar por el estudio de la variabilidad de los datos, puede ahorrar pasos en el análisis.
- Si el CV es mayor que 30 %, ninguna medida resume los datos.
- Si existe poca variación en los datos, debemos analizar la forma. En ese caso, si los datos son simétricos, la media representa los mismos. Si son asimétricos, la medida que los representa es la mediana.

# Gráfico de caja y bigotes (Box-Plot)



# Gráfico de caja y bigotes (Box-Plot)

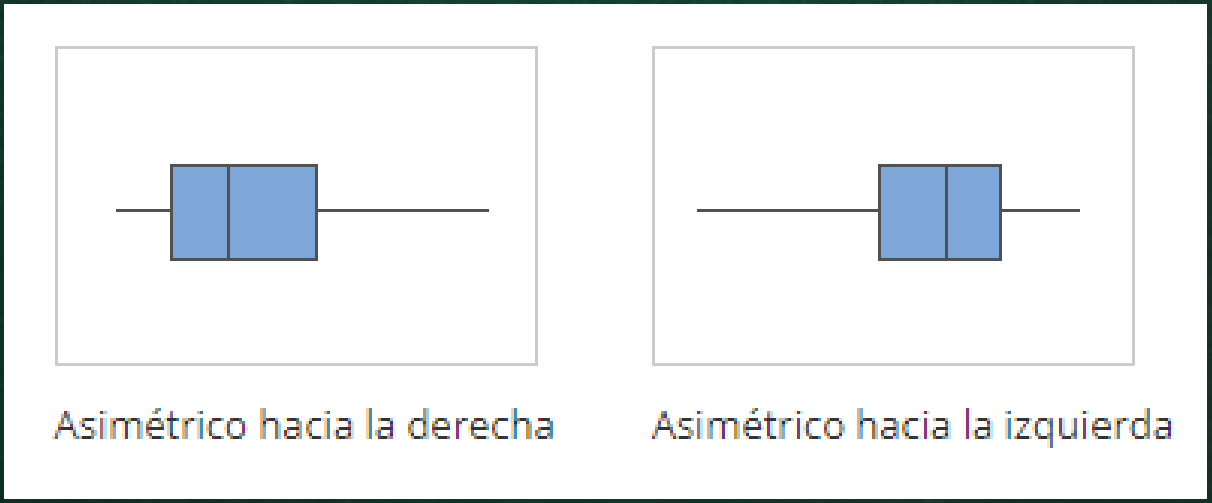



Para calcular el valor del bigote inferir, se compara al número Li con el valor más chico de todos los datos y el más grande entre ellos será el bigote inferior, LI. Para calcular el valor del bigote superior, se comparar al número Ls con el valor más grande de todos los datos y el más chico entre los dos números será el bigote superior LS

# Gráfico de caja y bigotes (Box-Plot) Interpretación.

## Datos asimétricos

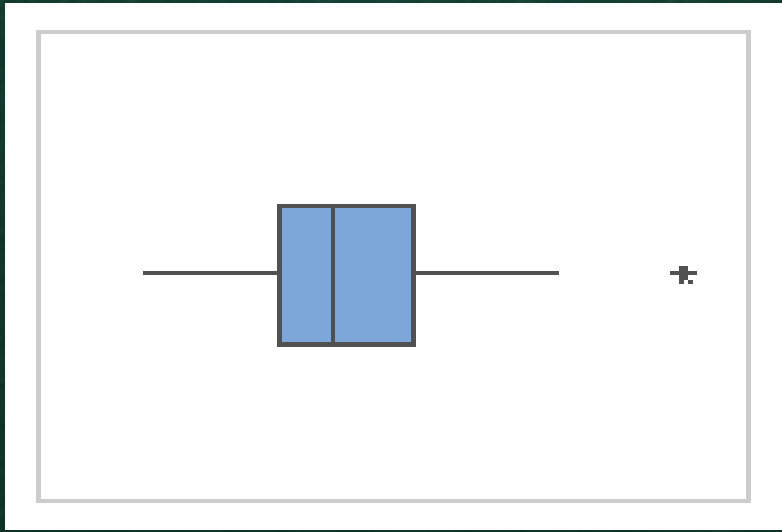
Cuando los datos son asimétricos, la mayoría de los datos se ubican en la parte superior o inferior de la gráfica.



# Gráfico de caja y bigotes (Box-Plot) Interpretación.

## Valores atípicos

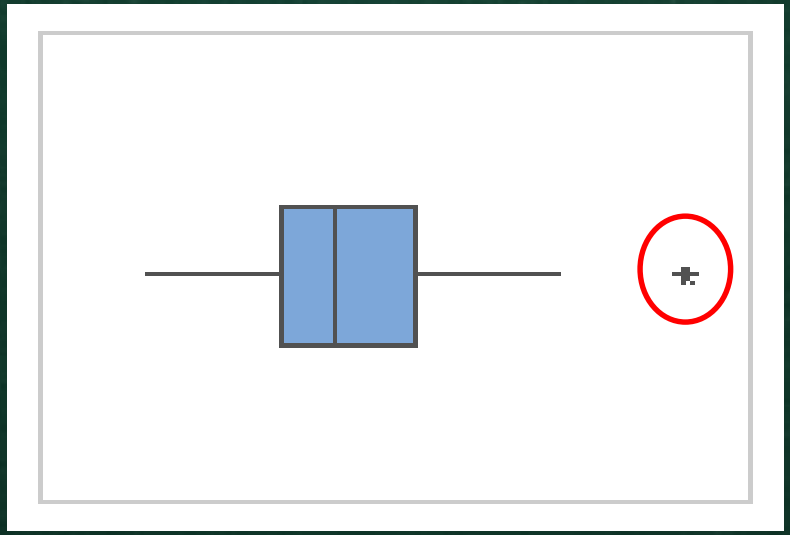
Los valores atípicos, que son valores de datos que están muy alejados de otros valores de datos, pueden afectar fuertemente sus resultados. Frecuentemente, es más fácil identificar los valores atípicos en una gráfica de caja..



# Gráfico de caja y bigotes (Box-Plot) Interpretación.

## Valores atípicos

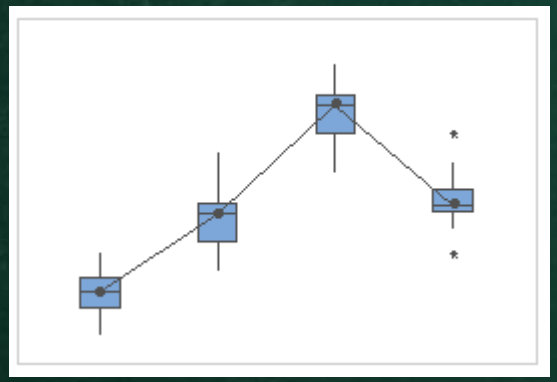
Los valores atípicos, que son valores de datos que están muy alejados de otros valores de datos, pueden afectar fuertemente sus resultados. Frecuentemente, es más fácil identificar los valores atípicos en una gráfica de caja..

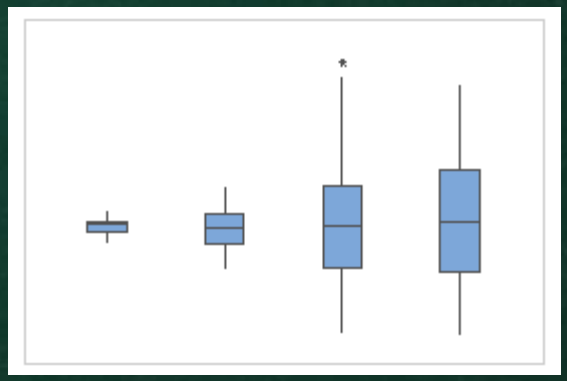


# Gráfico de caja y bigotes (Box-Plot) Interpretación.

## Evaluar y comparar los grupos

Permiten evaluar y comparar el centro y la dispersión de distintos grupos.





La mediana de los grupos son similares, pero algunos de los grupos presentan mayor variabilidad.

Universidad Nacional de Mar del Plata

Facultad de Ingeniería

# Análisis de Regresión y Correlación Lineal

2° Cuatrimestre 2021

# Análisis de Regresión y Correlación Lineal

**Introducción**: Hasta ahora hemos centrado nuestra atención principalmente en estudios que involucraban una única variable de respuesta numérica o en series de datos que contienen una única observación de cada individuo.

Por ejemplo:

- ✓ duración de cierto proceso medido en horas,
- ✓ longitud de una pieza,
- ✓ resistencia de cierto material a roturas,
- ✓ Número de hijos de una familia,
- ✓ entre otras…

# Análisis de Regresión y Correlación Lineal

Estudiar si existe relación o dependencia entre dos o más variables puede ser de interés en numerosas actividades.

¿El peso de las personas está relacionado con la estatura? ¿El peso y la presión arterial se relacionan? ¿La demanda de un producto dependerá de los precios? ¿La presión de una masa de gas depende de su volumen y de su temperatura?

Estudiaremos, siempre que sea posible, si una de las variables puede expresarse matemáticamente en función de la otra. Es decir, si se puede expresar $Y = f(X)$.

Nos va a interesar estudiar la relación que existen entre ellas y de qué forma se asocian. Para esto analizaremos dos técnicas: la de ***regresión*** y la de ***correlación***.

## Tipos de relación entre variables

Dos variables pueden estar relacionadas por una **dependencia funcional o determinista**, por una **dependencia estadística** o pueden ser **independientes**.

**Tipos de relación entre variables**

Dos variables pueden estar relacionadas por una **dependencia funcional o determinista**, por una **dependencia estadística** o pueden ser **independientes**.

Tipos de relación

Determinista: Conocido el valor de X, el valor de Y queda perfectamente establecido. Son del tipo: Y = f (X)

Ejemplo: La relación existente entre la temperatura expresada en grados centígrados (X) y grados Fahrenheit (Y ) es:

$$y = 1{,}8x + 32$$

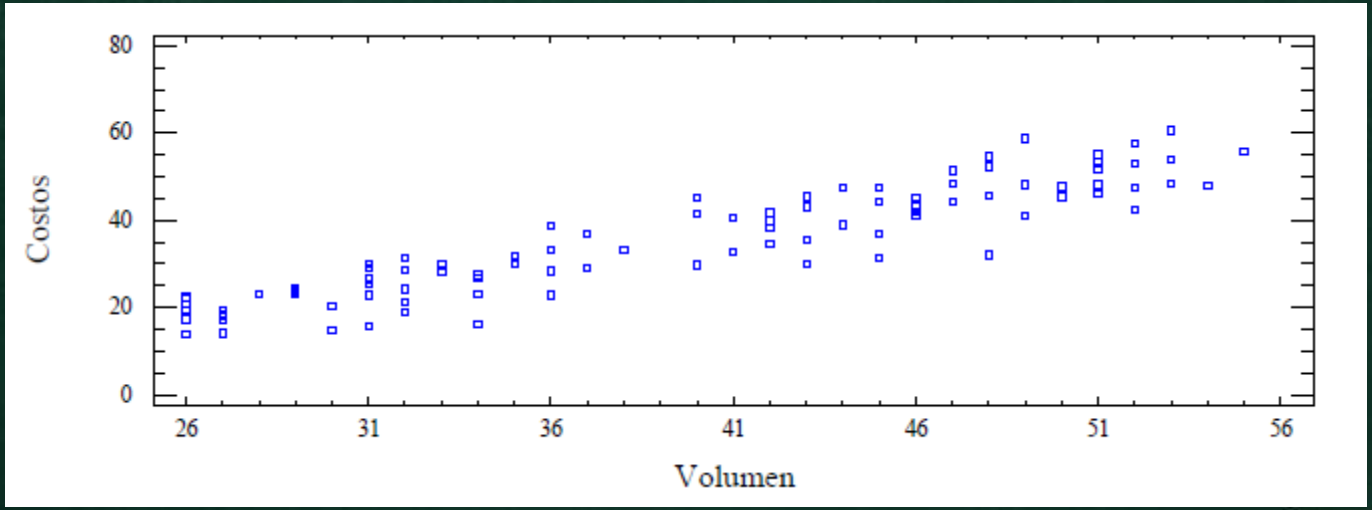
**Tipos de relación entre variables**

Dos variables pueden estar relacionadas por una **dependencia funcional o determinista**, por una **dependencia estadística** o pueden ser **independientes**.

Tipos de relación

**No** Determinista: Conocido el valor de X, el valor de Y no queda perfectamente establecido. Son del tipo: y = f (x) + *e* , donde *e (error)* es una perturbación desconocida.

Ejemplo: Se tiene una muestra del volumen de producción (X) y el costo total (Y) asociado a un producto en un grupo de empresas:



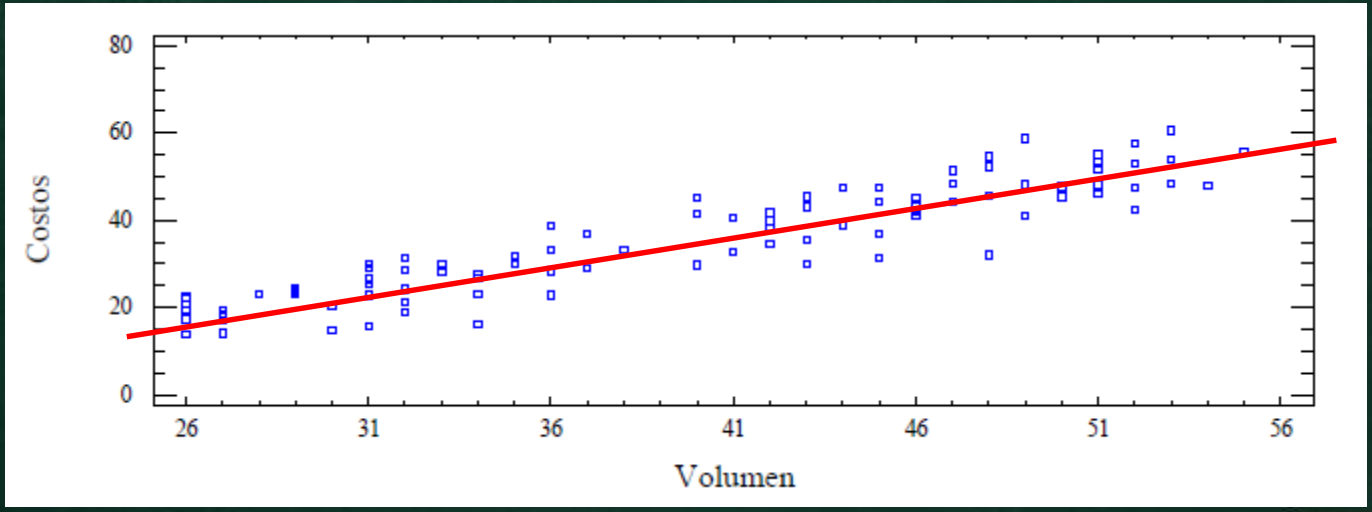
**Tipos de relación entre variables**

Dos variables pueden estar relacionadas por una **dependencia funcional o determinista**, por una **dependencia estadística** o pueden ser **independientes**.

Tipos de relación

**No** Determinista: Conocido el valor de X, el valor de Y no queda perfectamente establecido. Son del tipo: $y = f(x) + e$ , donde *e (error)* es una perturbación desconocida.

Ejemplo: Se tiene una muestra del volumen de producción (X) y el costo total (Y) asociado a un producto en un grupo de empresas:

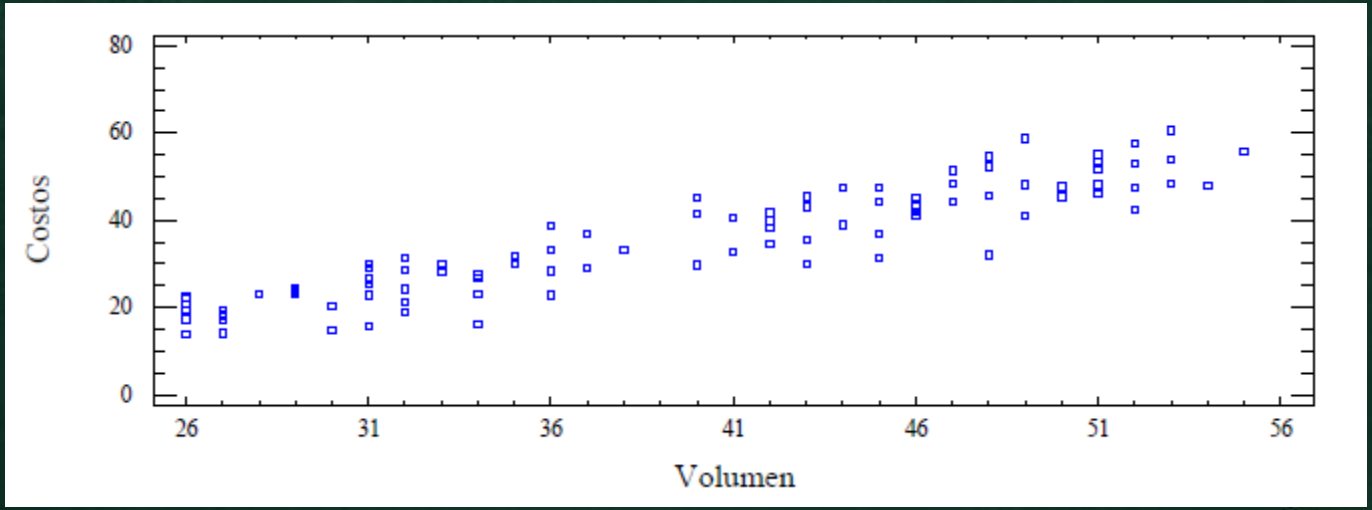


**Modelo de Regresión**

Es una técnica estadística que permite encontrar una ecuación (ecuación de regresión) que aproxime una variable en función de otras.
Permite describir cómo influye una variable X sobre otra variable Y.

X: Variable independiente o explicativa
Y: Variable dependiente o respuesta

El objetivo es obtener estimaciones "razonables" de Y para distintos valores de X a partir de una muestra de ***n*** pares de valores (x1, y1), . . . , (xn, yn).

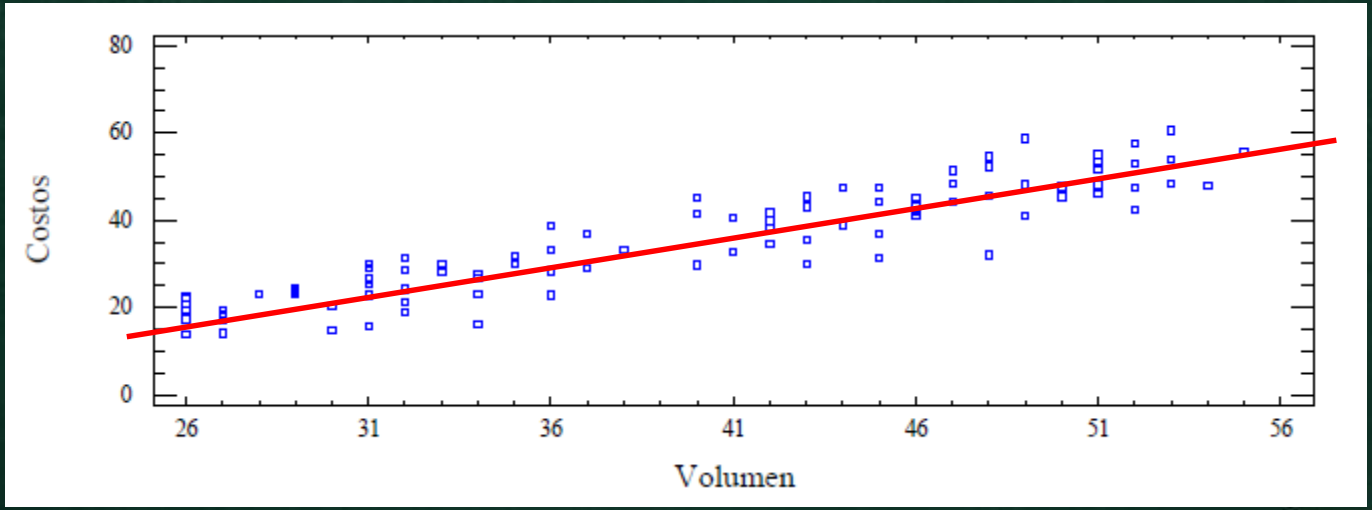


**Modelo de Regresión**

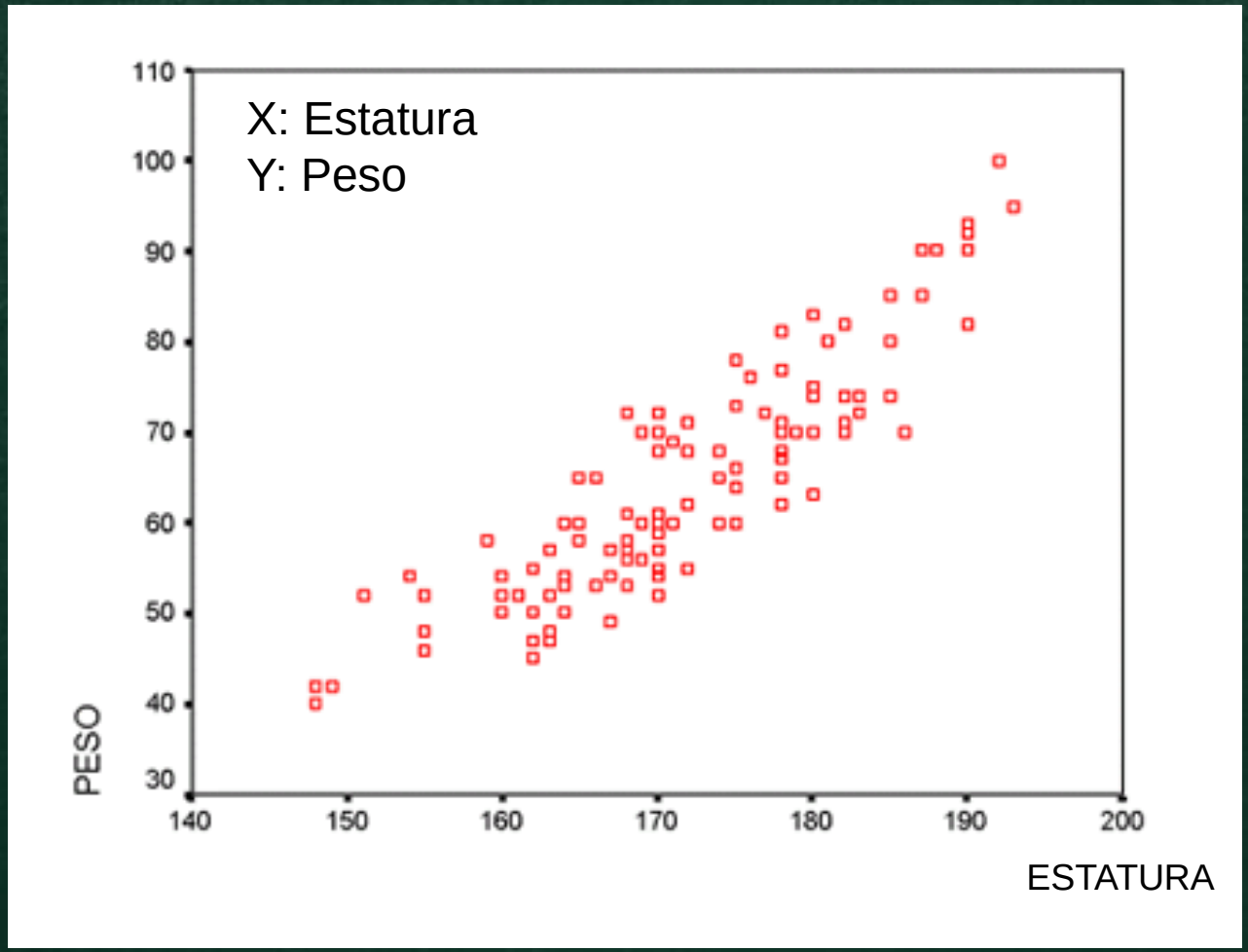
Es una técnica estadística que permite encontrar una ecuación (ecuación de regresión) que aproxime una variable en función de otras.
Permite describir cómo influye una variable X sobre otra variable Y.

X: Variable independiente o explicativa
Y: Variable dependiente o respuesta

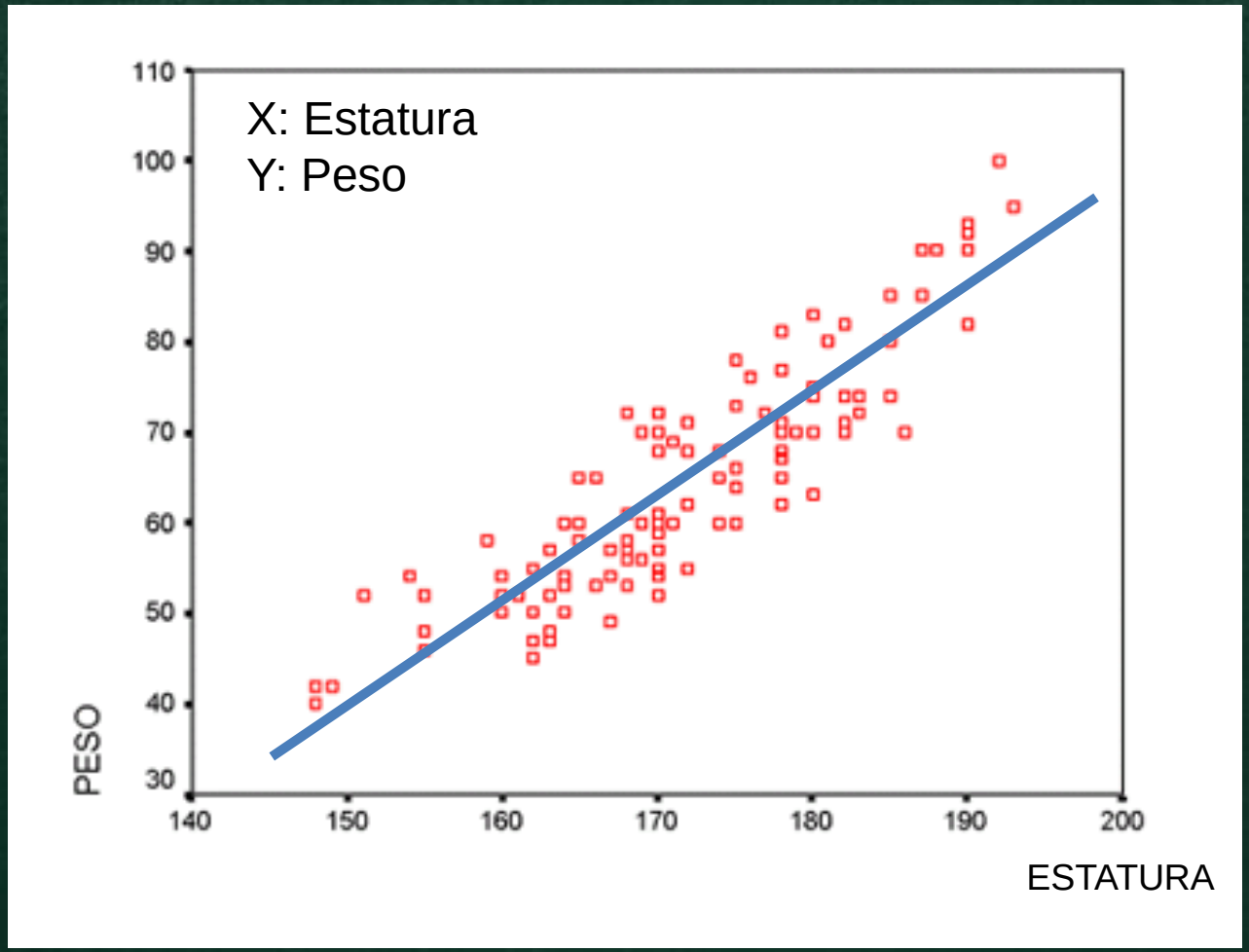
El objetivo es obtener estimaciones "razonables" de Y para distintos valores de X a partir de una muestra de ***n*** pares de valores (x1, y1), . . . , (xn, yn).



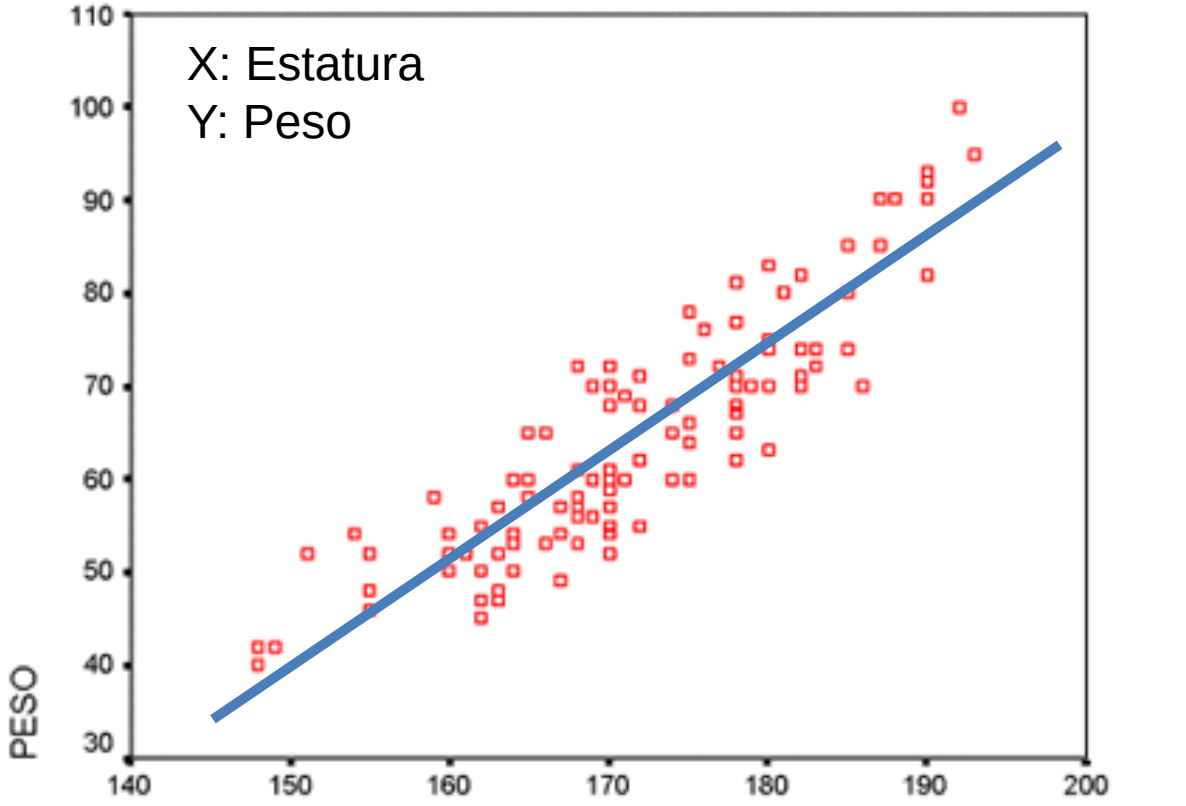
Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).



Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).

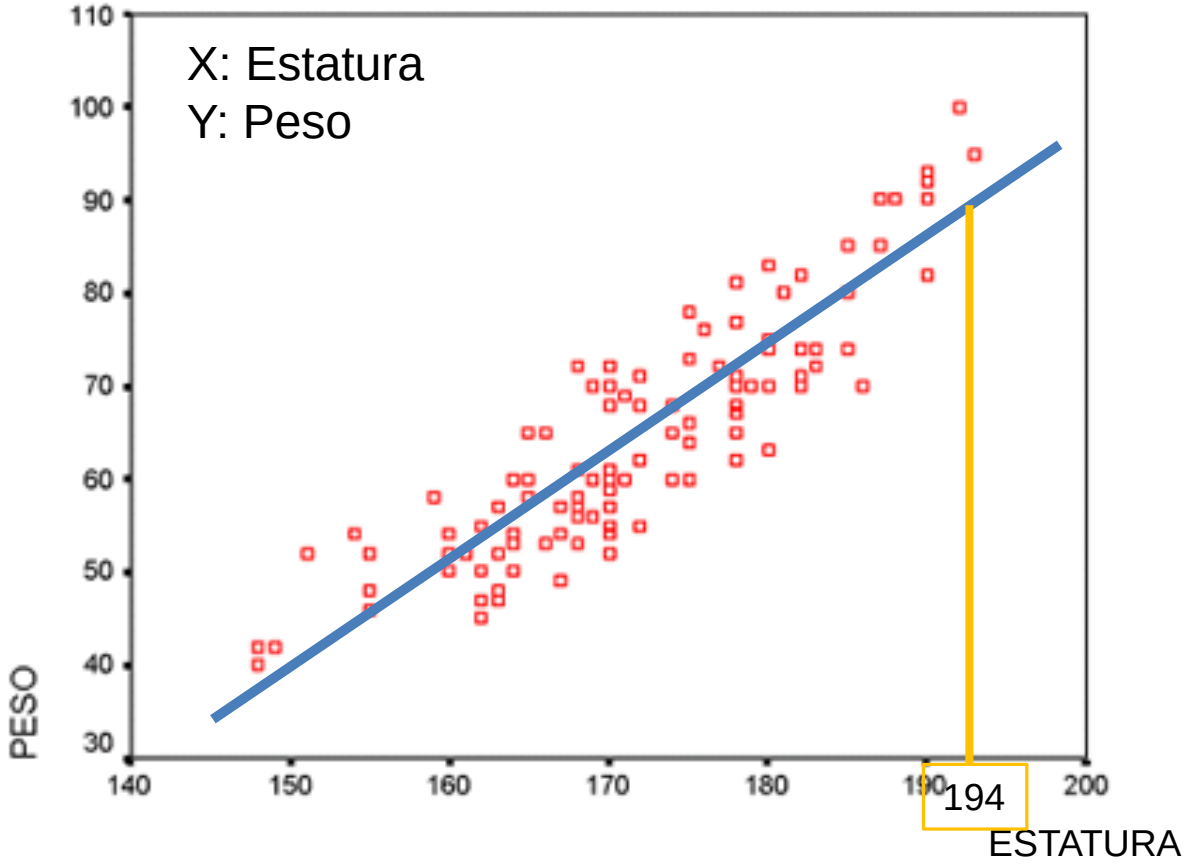


Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).




Podríamos predecir el peso de una persona con una estatura de 194 cm. Existe un componente aleatorio por lo que las predicciones tienen asociado un error de predicción.

Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).




Podríamos predecir el peso de una persona con una estatura de 194 cm. Existe un componente aleatorio por lo que las predicciones tienen asociado un error de predicción.

Resumiendo:

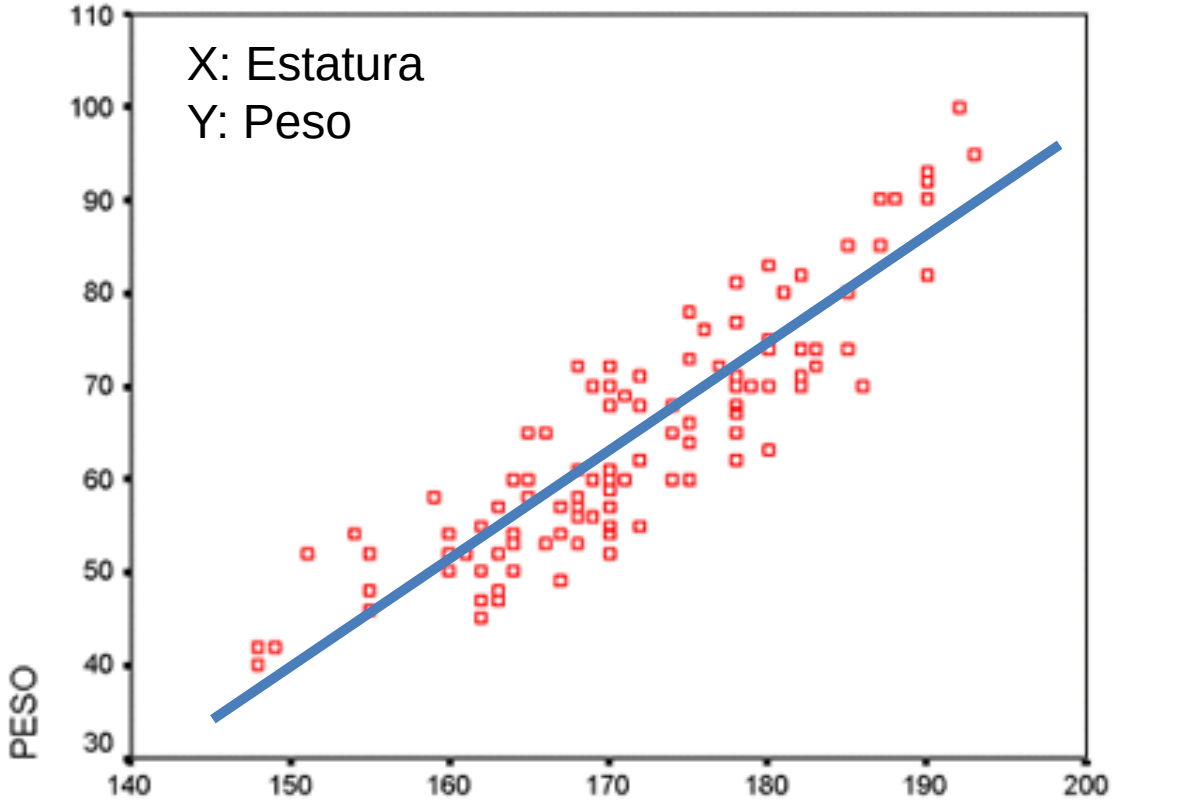
Si se trata de predecir o explicar el comportamiento de una variable **Y**, a la que se denomina **dependiente**, en función de otra variable **X denominada independiente**, **Y =f(X)**, estamos frente a un problema de análisis de regresión simple; pero si deseamos investigar el grado de asociación entre las variables X e Y estamos frente a un problema de análisis de correlación.

Resumiendo:

Si se trata de predecir o explicar el comportamiento de una variable **Y**, a la que se denomina **dependiente**, en función de otra variable **X denominada independiente**, **Y =f(X)**, estamos frente a un problema de análisis de regresión simple; pero si deseamos investigar el grado de asociación entre las variables X e Y estamos frente a un problema de análisis de correlación.

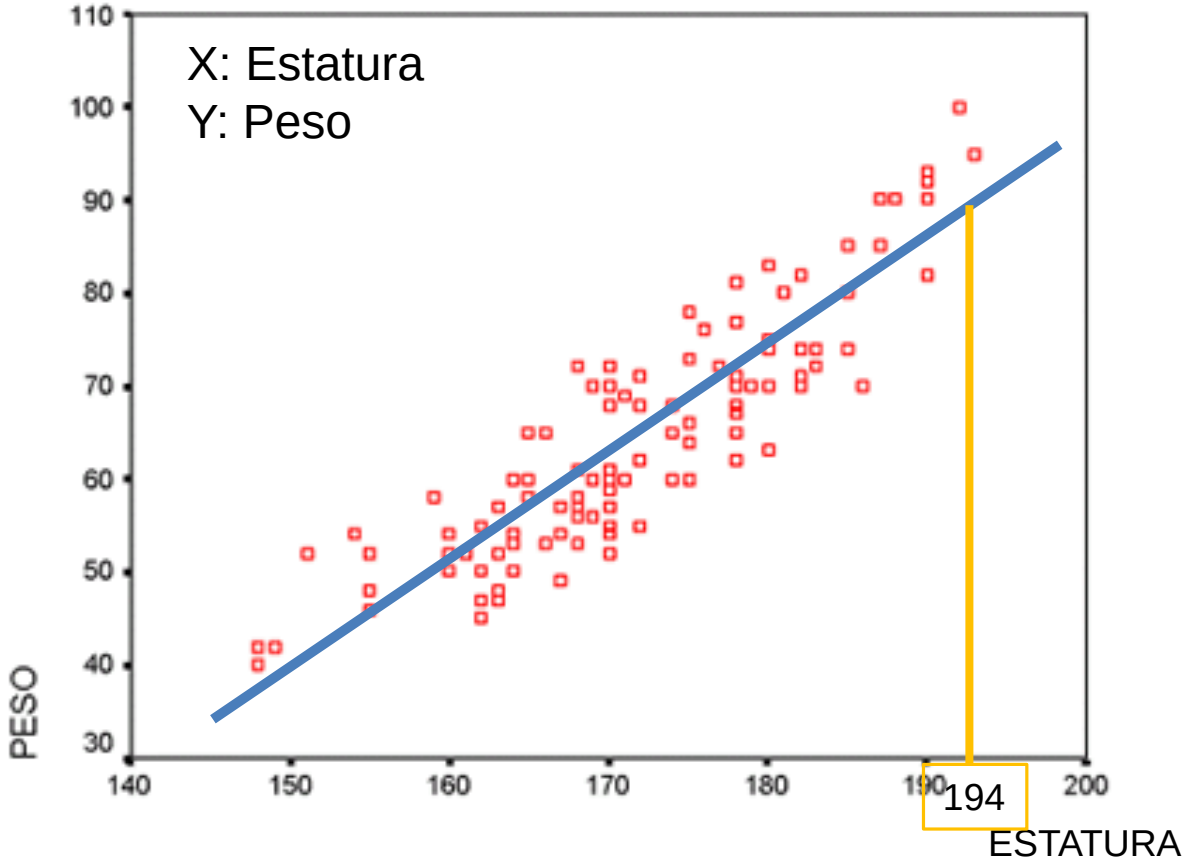
El objetivo es analizar la relación existente entre dos variables, X e Y, de forma que podamos predecir o aproximar el valor de la variable Y a partir del valor de la variable X.

Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).




Podríamos predecir el peso de una persona con una estatura de 194 cm. Existe un componente aleatorio por lo que las predicciones tienen asociado un error de predicción.

Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).




Podríamos predecir el peso de una persona con una estatura de 194 cm. Existe un componente aleatorio por lo que las predicciones tienen asociado un error de predicción.

Resumiendo:

Si se trata de predecir o explicar el comportamiento de una variable **Y**, a la que se denomina **dependiente**, en función de otra variable **X denominada independiente**, **Y =f(X)**, estamos frente a un problema de análisis de regresión simple; pero si deseamos investigar el grado de asociación entre las variables X e Y estamos frente a un problema de análisis de correlación.

El objetivo es analizar la relación existente entre dos variables, X e Y, de forma que podamos predecir o aproximar el valor de la variable Y a partir del valor de la variable X.

**Observación**: En un problema de regresión el papel de las dos variables no es simétrico.

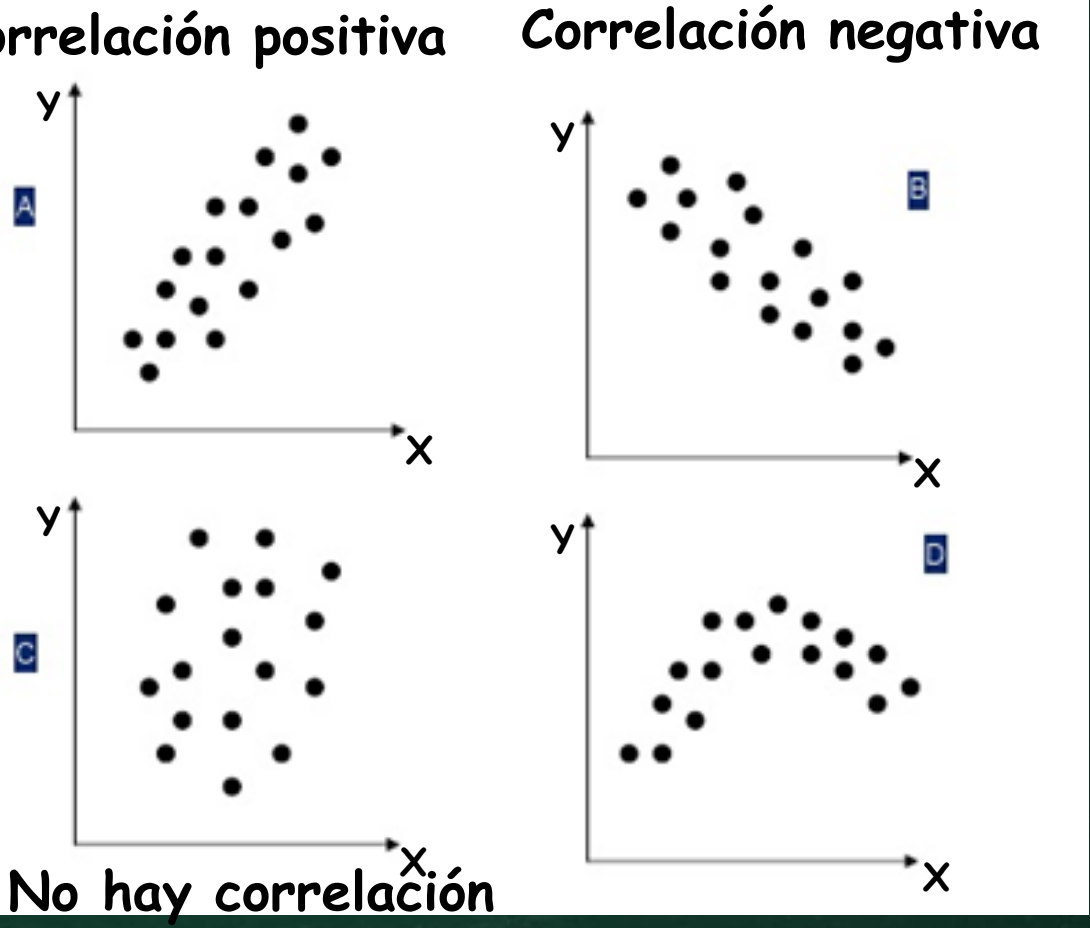
X: Variable independiente o explicativa
Y: Variable dependiente o respuesta

## Análisis de regresión entre dos variables X e Y

Consideremos el problema de tratar de hallar la relación funcional existente entre dos variables **X** e **Y**. Supongamos que en ***n*** experimentos las variables asumieron pares de valores {(xi,yi):i=1,…,*n*}, podemos inicialmente observar su comportamiento graficando dichos pares de valores sobre un sistema de coordenadas. Dicho gráfico, llamado ***diagrama de dispersión*** a menudo permite discernir si existe alguna tendencia hacia algún tipo de interrelación entre ambas variables, y, si es posible, la naturaleza de dicho tipo de interrelación.

## Diagrama de Dispersión

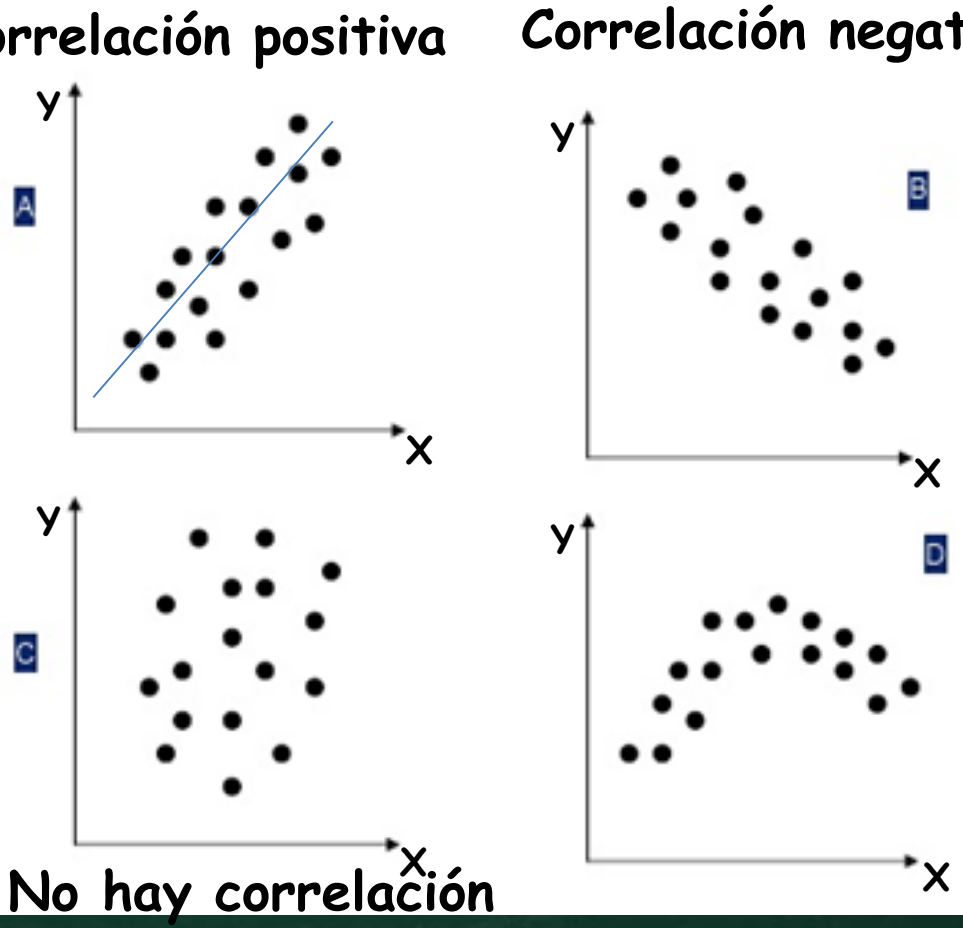
| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $x_3$ | $y_3$ |
| ⋮ | ⋮ |
| $x_n$ | $y_n$ |




**Observación**: Solo nos ocuparemos del caso lineal en esta unidad.

## Diagrama de Dispersión

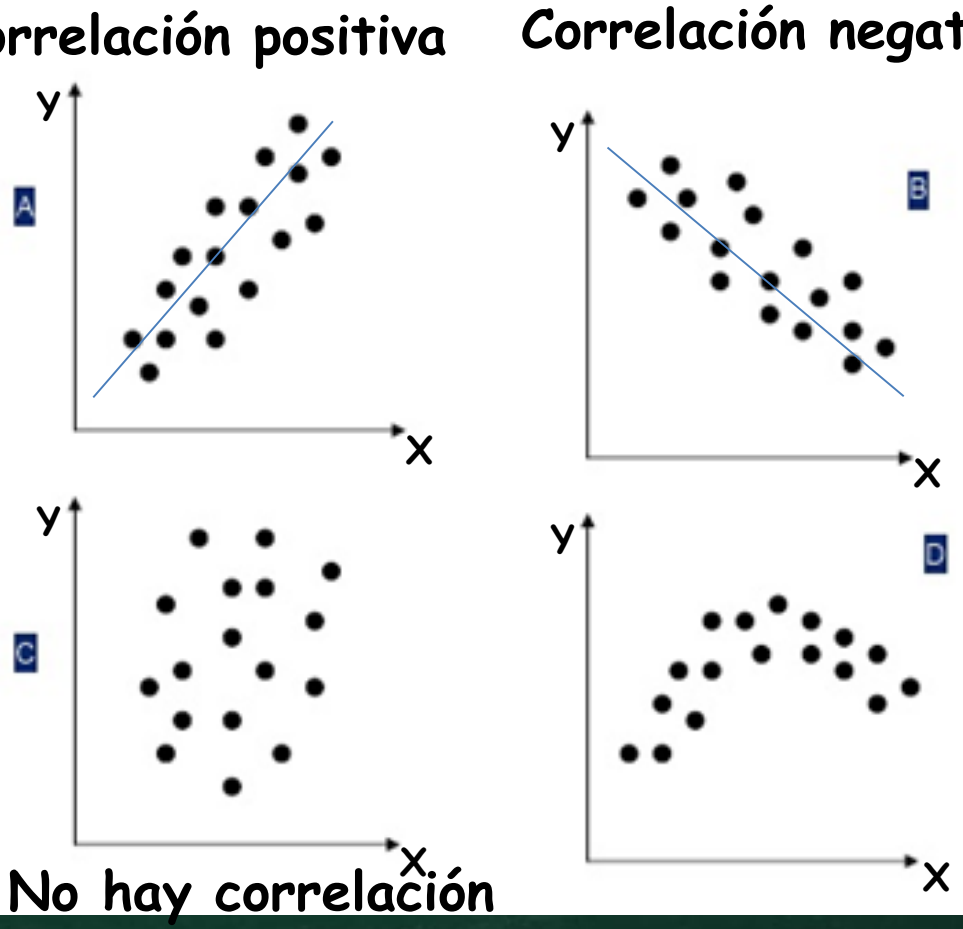
| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $x_3$ | $y_3$ |
| ⋮ | ⋮ |
| $x_n$ | $y_n$ |




**Observación**: Solo nos ocuparemos del caso lineal en esta unidad.

## Diagrama de Dispersión

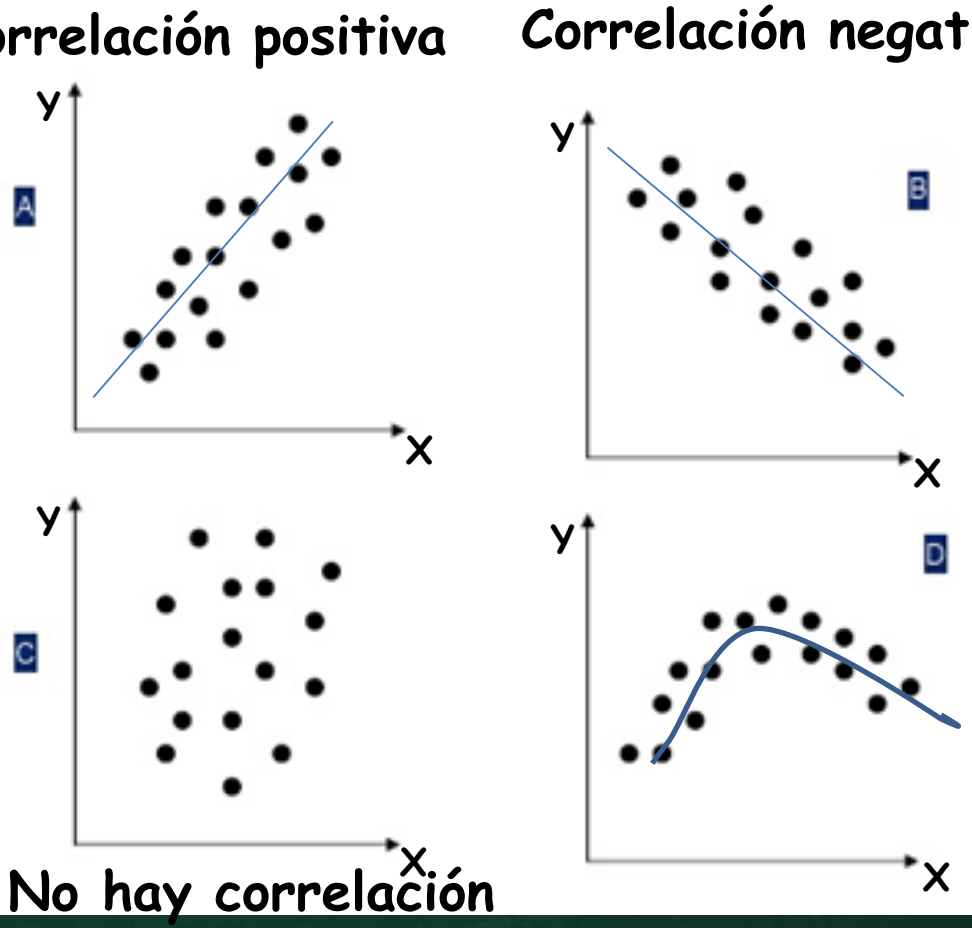
| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $x_3$ | $y_3$ |
| ⋮ | ⋮ |
| $x_n$ | $y_n$ |




**Observación**: Solo nos ocuparemos del caso lineal en esta unidad.

## Diagrama de Dispersión

| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $x_3$ | $y_3$ |
| ⋮ | ⋮ |
| $x_n$ | $y_n$ |




**Observación**: Solo nos ocuparemos del caso lineal en esta unidad.

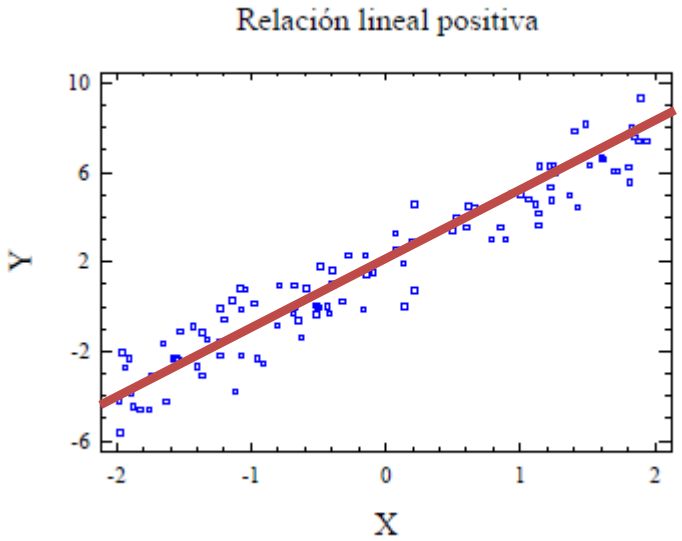
## Análisis de regresión entre dos variables X e Y

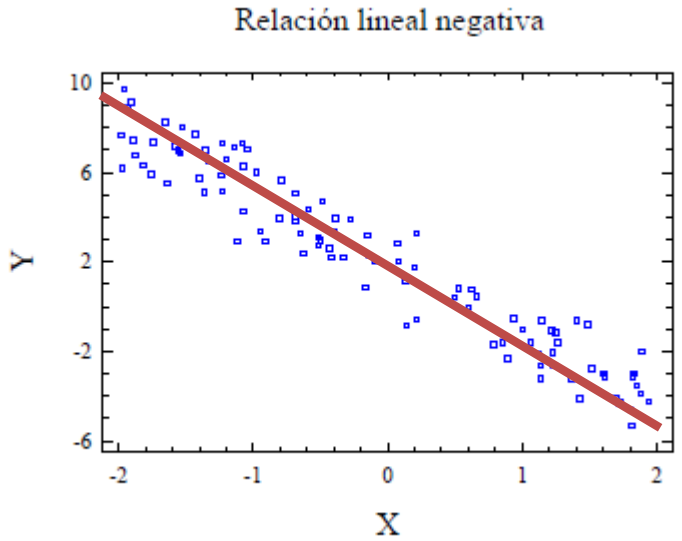
Lineal: Cuando la función f (x) es lineal, es decir de la forma:

$$f(x) = ax + b$$

Si a > 0 hay relación lineal positiva.

Si a < 0 hay relación lineal negativa.

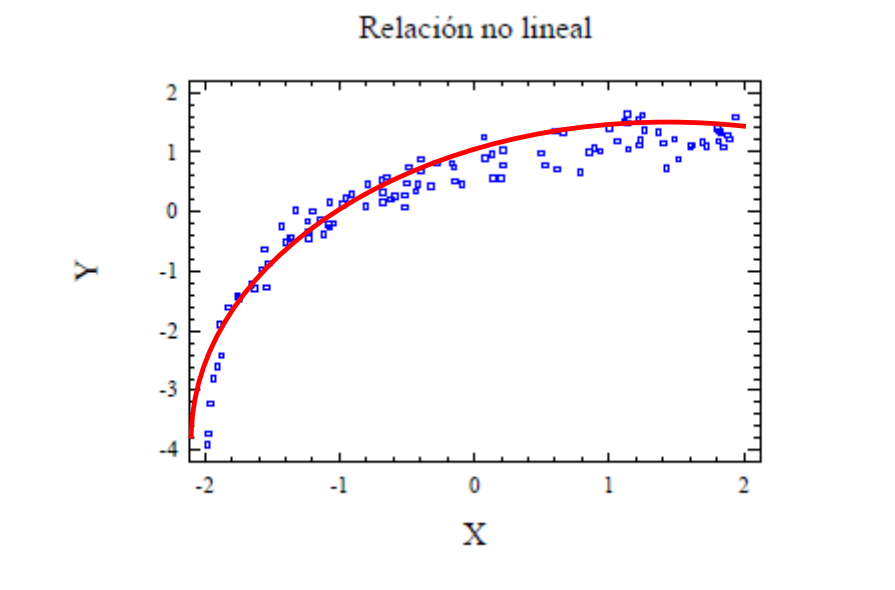


## Análisis de regresión entre dos variables X e Y

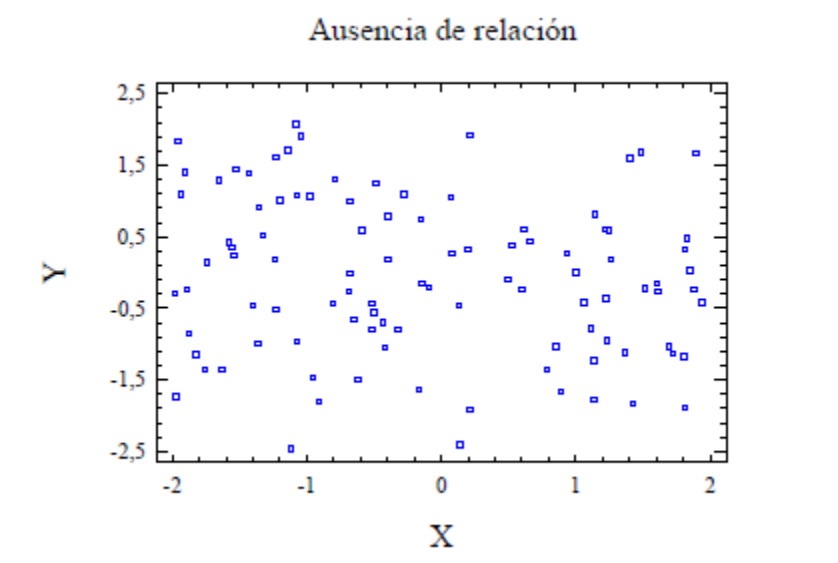
No lineal: Cuando la función f (x) no es lineal. Por ejemplo,

$$f(x) = \log(x)$$



## Análisis de regresión entre dos variables X e Y

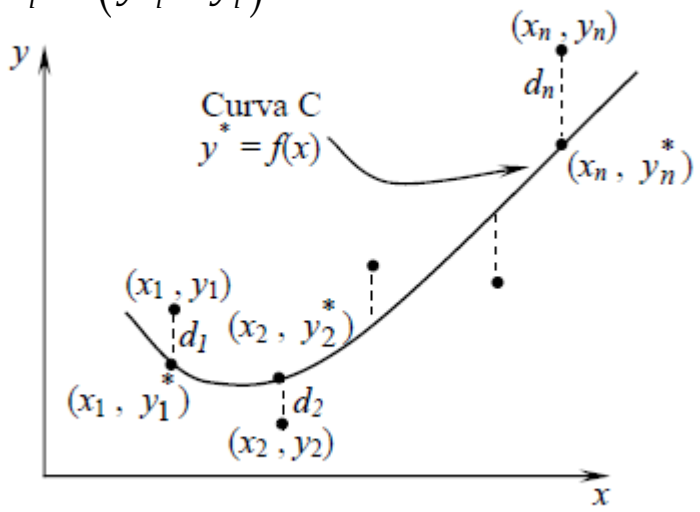
Ausencia de relación:



Ajuste de una función de regresión: **Método de mínimos cuadrados**

Ajustar una función de regresión significa encontrar, la función que exprese con mayor precisión la relación entre las variables X e Y. Gráficamente será aquella función cuya gráfica mejor se adecue a la nube de puntos. En este sentido, es recomendable como primer paso construir el diagrama de dispersión o diagrama de nube de puntos para, luego de analizar su forma, decidir por el tipo de función matemática (modelo) o la ecuación de regresión que exprese la relación entre las variables X e Y. Luego, se estiman los parámetros del modelo, para lo cual existen varios métodos, siendo el más usado el ***método de mínimos cuadrados***.

$$d_i^2 = \left(y^*_i - y_i\right)^2$$

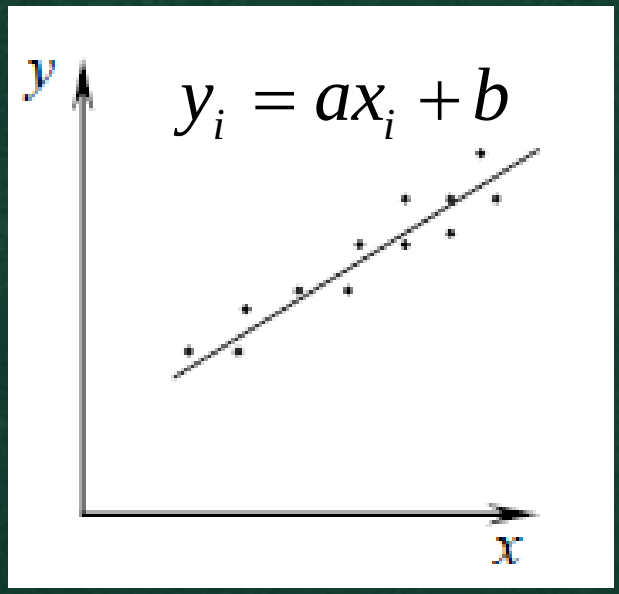



$$D = d_1^2 + d_2^2 + \ldots + d_{n-1}^2 + d_n^2$$

El problema queda ahora reducido a encontrar los coeficientes de un tipo de curva que hagan mínimo el valor *D*. Una vez determinados estos valores, a la curva correspondiente se la llamará ***curva de regresión de Y sobre X***.

## Análisis de regresión lineal simple

Es frecuente suponer que existe entre las variables observadas una relación aproximadamente lineal:

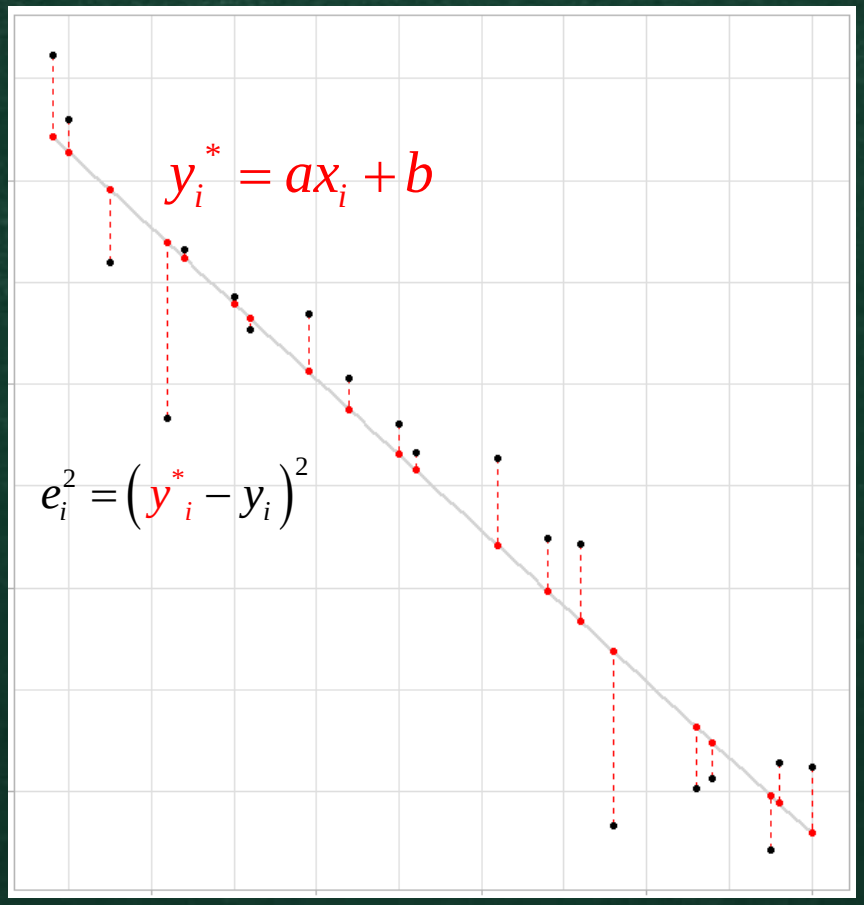



La recta ***y=ax+b*** es una recta de regresión. El parámetro ***a*** es la pendiente de la recta e indica cómo cambia la variable respuesta o dependiente cuando el incremento de x es una unidad. El parámetro ***b*** es el término independiente de la recta e indica el valor de Y cuando X = 0.

**Problema estadístico**: Estimar los parámetros a y b a partir de los datos, de una muestra.

## Determinación de las rectas de regresión por el método e mínimos cuadrados

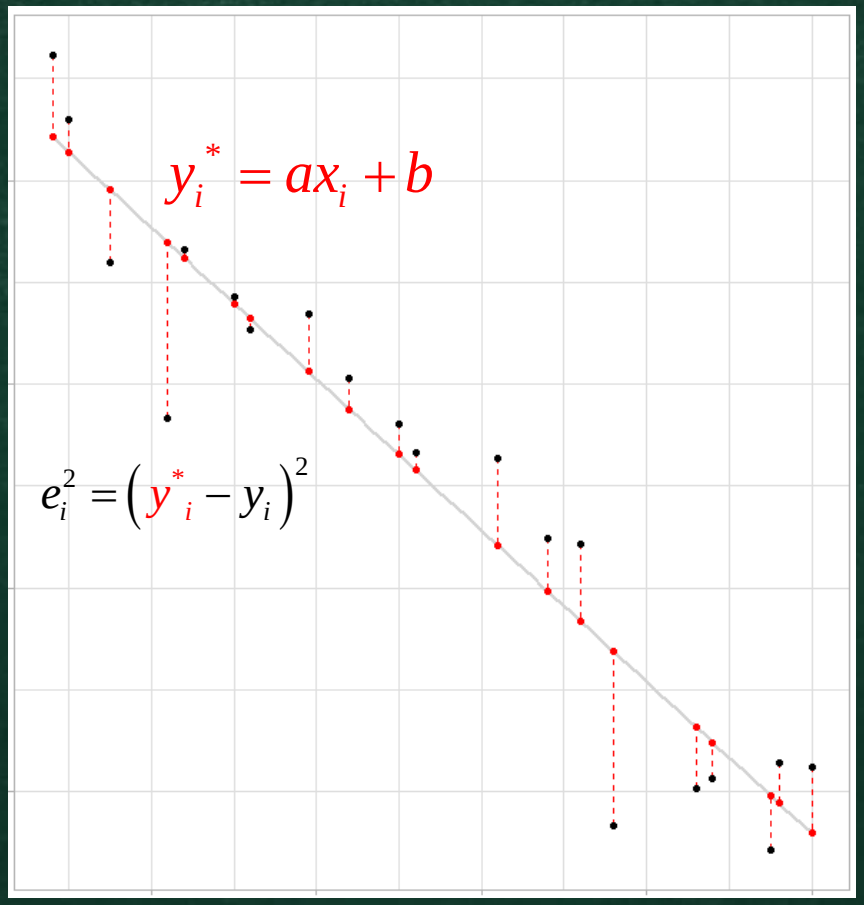
Gauss propuso en 1809 el método de mínimos cuadrados para obtener los valores a y b que mejor se ajustan a los datos:




El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, minimizar la suma de los residuos al cuadrado.

## Determinación de las rectas de regresión por el método e mínimos cuadrados

Gauss propuso en 1809 el método de mínimos cuadrados para obtener los valores a y b que mejor se ajustan a los datos:

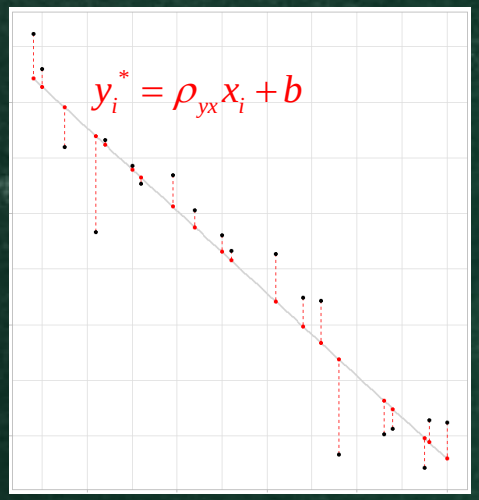



$$\sum_i e_i^2 = \sum_i \left(y^*_i - y_i\right)^2$$

minimizar

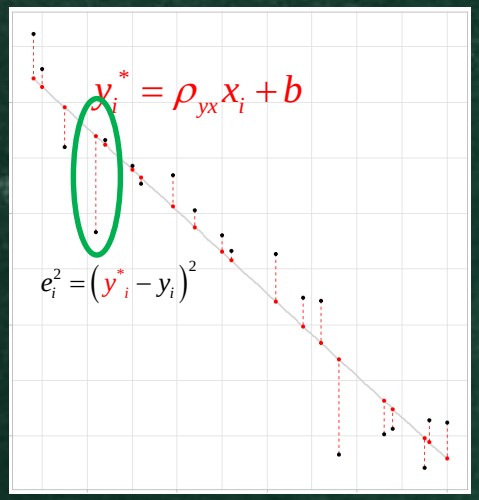
El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, minimizar la suma de los residuos al cuadrado.

## Determinación de las rectas de regresión por el método e mínimos cuadrados




$$y^* = ax + b \qquad a = \rho_{yx}$$

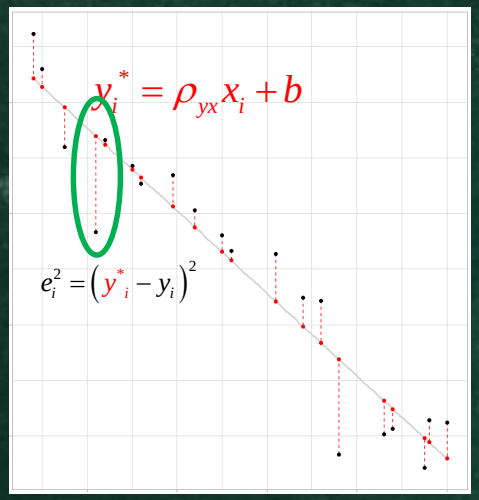
# Determinación de las rectas de regresión por el método e mínimos cuadrados




$$y^* = ax + b$$

$$e_i^2 = \left(y^*_i - y_i\right)^2$$

# Determinación de las rectas de regresión por el método e mínimos cuadrados



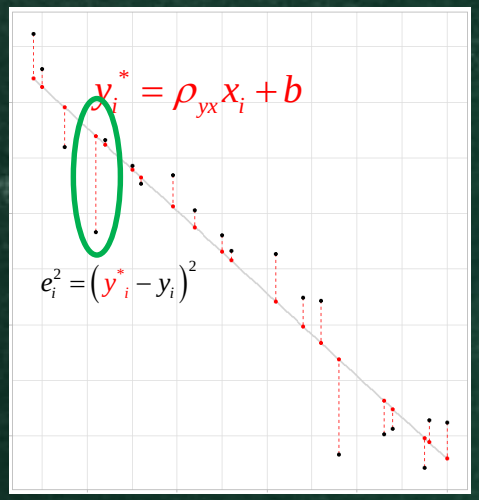

$$y^* = ax + b$$

$$e_i^2 = \left(y^*_i - y_i\right)^2$$

$$D = \sum_i e_i^2 = \sum_i \left(y^*_i - y_i\right)^2$$

$$D = \sum_i \left(y^*_i - y_i\right)^2 = \sum_i \left(\rho_{yx} x_i + b - y_i\right)^2$$

# Determinación de las rectas de regresión por el método e mínimos cuadrados




$$y^* = ax + b$$

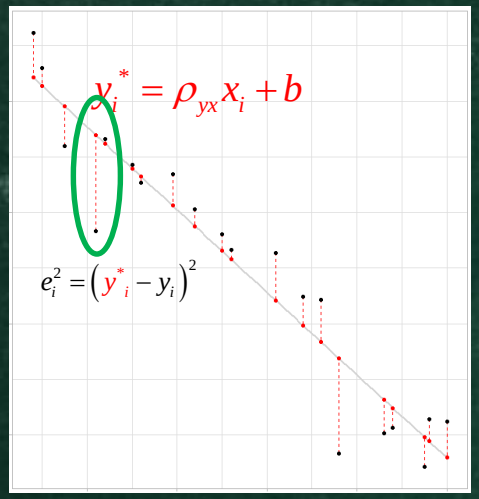
$$e_i^2 = \left(y^*_i - y_i\right)^2$$

$$D = \sum_i e_i^2 = \sum_i \left(y^*_i - y_i\right)^2$$

$$D = \sum_i \left(y^*_i - y_i\right)^2 = \sum_i \left(\rho_{yx} x_i - b - y_i\right)^2$$

$$\begin{cases} \dfrac{\partial D}{\partial \rho} = 0 \\ \dfrac{\partial D}{\partial b} = 0 \end{cases}$$

# Determinación de las rectas de regresión por el método e mínimos cuadrados




$$y^* = \rho_{yx} x + b$$

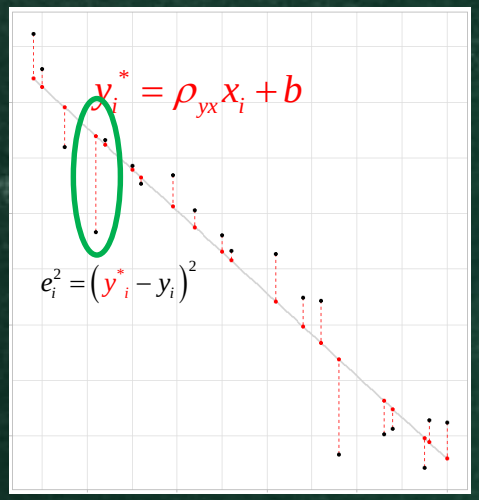
$$e_i^2 = \left(y^*_i - y_i\right)^2$$

$$D = \sum_i e_i^2 = \sum_i \left(y^*_i - y_i\right)^2$$

$$D = \sum_i \left(y^*_i - y_i\right)^2 = \sum_i \left(\rho_{yx} x_i + b - y_i\right)^2$$

$$\begin{cases} \dfrac{\partial D}{\partial \rho} = 0 \\ \dfrac{\partial D}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \dfrac{\partial D}{\partial \rho} = 2\sum_{i=1}^{n} \left(\rho_{yx} x_i - b - y_i\right) x_i = 0 \\ \dfrac{\partial D}{\partial b} = 2\sum_{i=1}^{n} \left(\rho_{yx} x_i - b - y_i\right) = 0 \end{cases} \Rightarrow \begin{cases} \sum x_i y_i = \rho_{yx} \sum x_i + \rho_{yx} \sum x_i^2 \\ \sum y_i = bn + \rho_{yx} \sum x_i \end{cases}$$

# Determinación de las rectas de regresión por el método e mínimos cuadrados




$$y^* = ax + b$$

$$e_i^2 = \left(y^*_i - y_i\right)^2$$

$$D = \sum_i e_i^2 = \sum_i \left(y^*_i - y_i\right)^2$$

$$D = \sum_i \left(y^*_i - y_i\right)^2 = \sum_i \left(\rho_{yx} x_i + b - y_i\right)^2$$

$$\begin{cases} \dfrac{\partial D}{\partial \rho} = 0 \\ \dfrac{\partial D}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \dfrac{\partial D}{\partial \rho} = 2\sum_{i=1}^{n}\left(\rho_{yx} x_i - b - y_i\right)x_i = 0 \\ \dfrac{\partial D}{\partial b} = 2\sum_{i=1}^{n}\left(\rho_{yx} x_i - b - y_i\right) = 0 \end{cases} \Rightarrow \begin{cases} \sum x_i y_i = \rho_{yx}\sum x_i + \rho_{yx}\sum x_i^2 \\ \sum y_i = bn + \rho_{yx}\sum x_i \end{cases}$$

Resolviendo el sistema obtenemos:

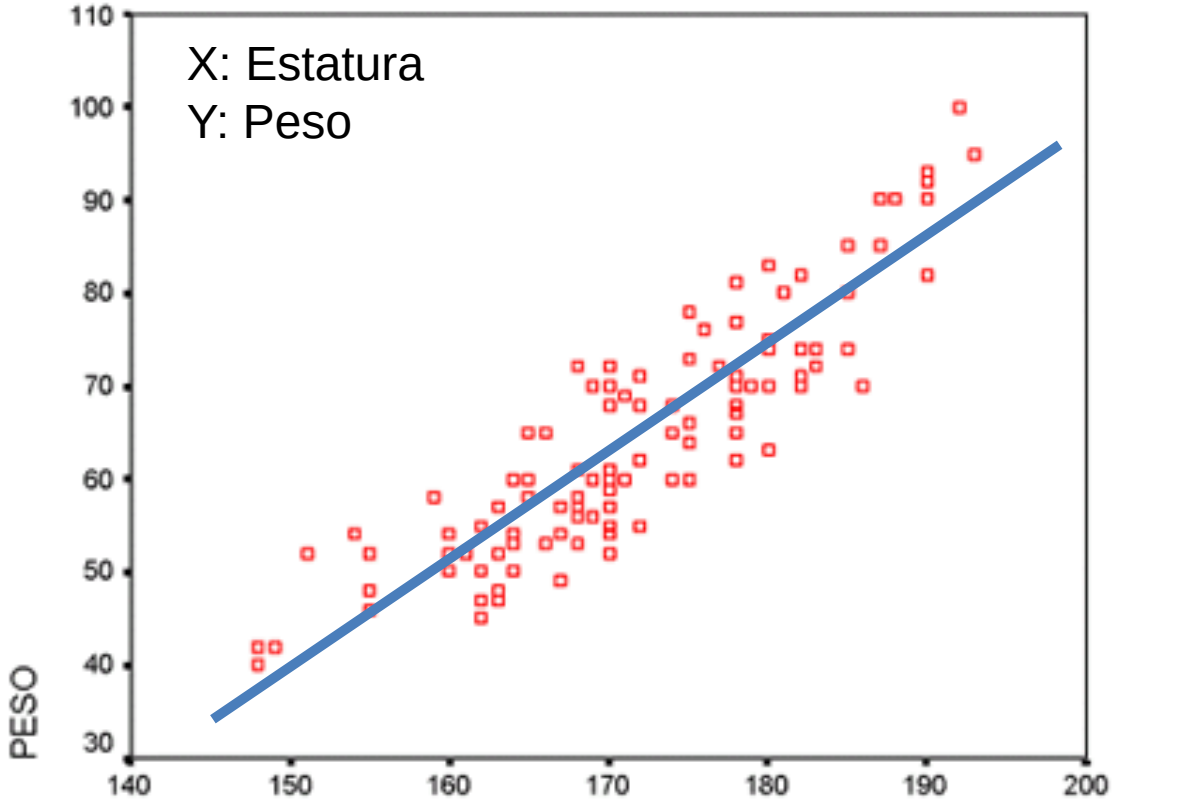
$$\bar{y}_x = \rho_{yx} x + b$$

$$\rho_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad b = \frac{\sum_{i=1}^{n} y_i}{n} - \rho_{yx}\frac{\sum_{i=1}^{n} x_i}{n}$$

**INFERENCIA EN REGRESION**

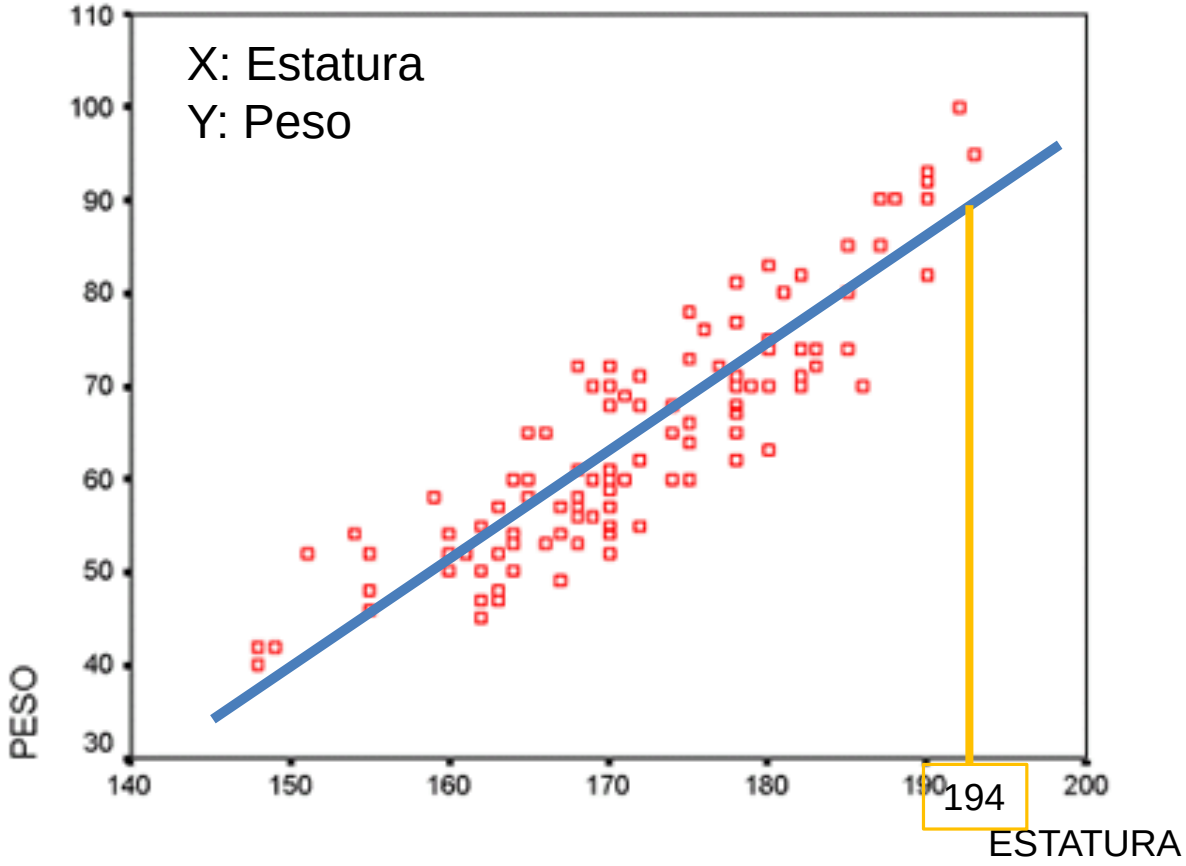
La recta de regresión nos permite, basándonos en los datos de la muestra, estimar un valor de la variable Y, correspondiente a un valor dado xi de la variable X. Para ello es suficiente reemplazar el valor de xi en la recta de regresión y encontrar el correspondiente valor estimado.

Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).




Podríamos predecir el peso de una persona con una estatura de 194 cm. Existe un componente aleatorio por lo que las predicciones tienen asociado un error de predicción.

Ejemplo: Se tiene una muestra de las variables Peso y Estatura, es decir ***n*** pares de valores (x1, y1), . . . , (xn, yn).




Podríamos predecir el peso de una persona con una estatura de 194 cm. Existe un componente aleatorio por lo que las predicciones tienen asociado un error de predicción.

**Ejemplo**: Para ilustrar el método de mínimos cuadrados lo aplicamos a los siguientes datos apareados, que representan el tiempo de calentamiento y la superficie de óxido generado de cierta pieza.

X: representa el tiempo de recalentamiento
Y: los espesores de óxido de cierta pieza

| **X (min)** | 20 | 30 | 40 | 60 | 70 | 90 | 100 | 120 | 150 | 180 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Y (Ang)** | 3,5 | 7,4 | 7,1 | 15,6 | 11,1 | 14,9 | 23,5 | 27,1 | 22,1 | 32,9 |

$$\sum x_i y_i = 18469 \qquad \sum x_i = 860 \qquad \sum y_i = 165,2$$

$$\sum x_i^2 = 98800$$

$$\rho_{yx} = 0,17 \quad b = 1,76 \quad \bar{y}_x = 0,17x + 1,76$$

Ejemplo: Determinar la recta de regresión lineal

$$\bar{y}_x = \rho_{yx} x + b$$

$$\rho_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad b = \frac{\sum_{i=1}^{n} y_i}{n} - \rho_{yx} \frac{\sum_{i=1}^{n} x_i}{n}$$

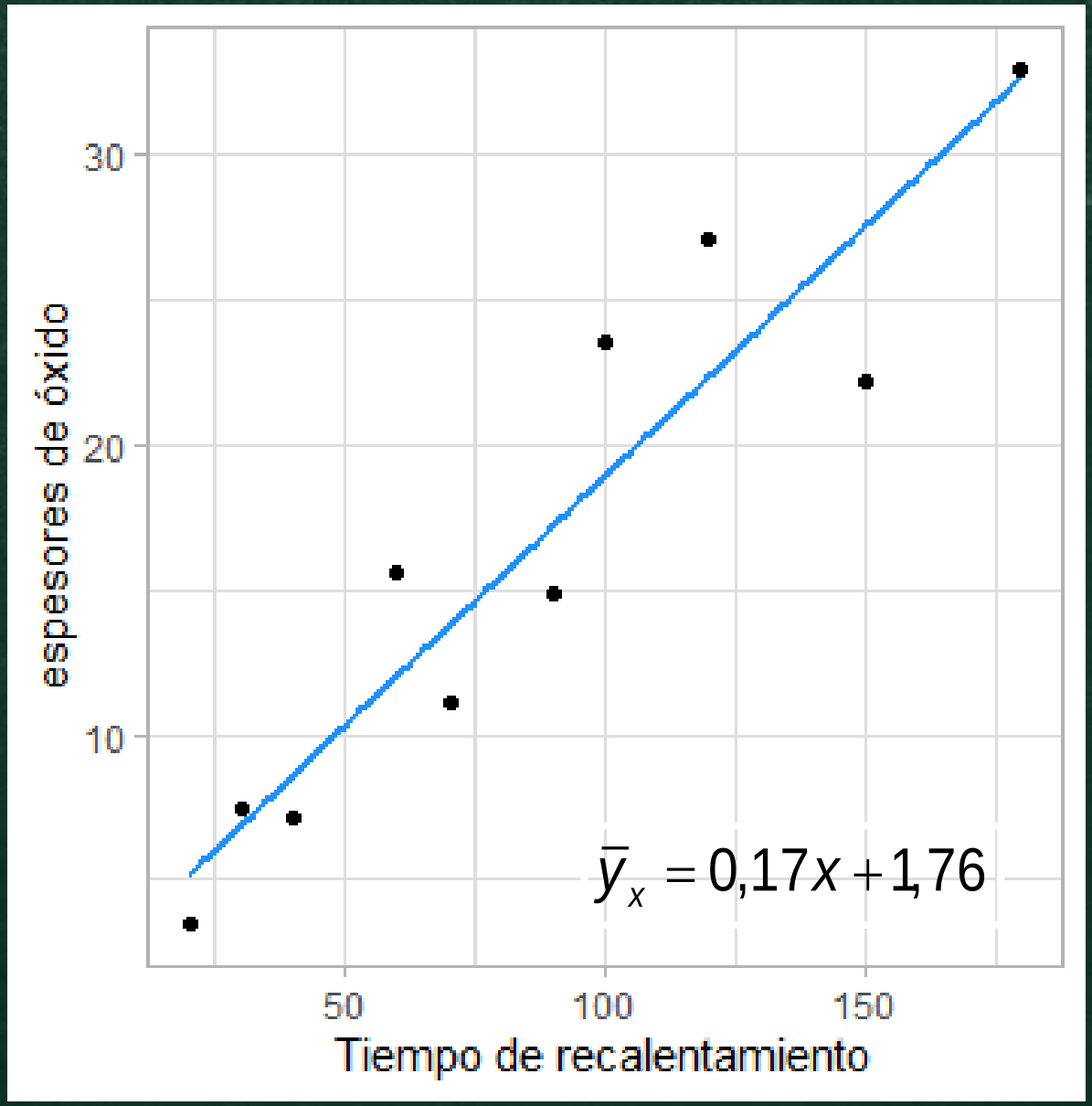
X: representa el tiempo de recalentamiento
Y: los espesores de óxido de cierta pieza

| **X (min)** | 20 | 30 | 40 | 60 | 70 | 90 | 100 | 120 | 150 | 180 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Y (Ang)** | 3,5 | 7,4 | 7,1 | 15,6 | 11,1 | 14,9 | 23,5 | 27,1 | 22,1 | 32,9 |

$$\sum x_i y_i = 18469 \qquad \sum x_i = 860 \qquad \sum y_i = 165,2$$

$$\sum x_i^2 = 98800$$

$$\rho_{yx} = 0,17 \quad b = 1,76 \quad \bar{y}_x = 0,17x + 1,76$$


espesores de óxido
30
20
10
50
100
150
Tiempo de recalentamiento
$\bar{y}_x = 0{,}17x + 1{,}76$

## EL COEFICIENTE DE CORRELACIÓN LINEAL PEARSON

Algunas veces es deseable tener un indicador del grado de intensidad o fuerza de la relación lineal entre dos variables X e Y que sea independiente de sus respectivas escalas de medición. A este indicador se le denomina ***coeficiente de correlación lineal entre X e Y***. El estadígrafo comúnmente utilizado se llama coeficiente de correlación del producto momento de Pearson.

Definición. Sea (X, Y) dos variables, para ***n*** pares de valores (x1, y1), . . . , (xn, yn), definimos ***r*** el **coeficiente de correlación muestral entre X e Y** como sigue:

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}\sqrt{\sum \left(y_i - \bar{y}\right)^2}}$$

**donde**

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

$$r = \rho_{yx}\frac{S_x}{S_y} = \rho_{xy}\frac{S_y}{S_x}$$

Notar que si $S_x = S_y$

$$r = \rho_{yx}$$

## INTERPRETACIÓN

**El coeficiente de correlación lineal de *Pearson* (r):**

-Está acotado entre -1 y 1.

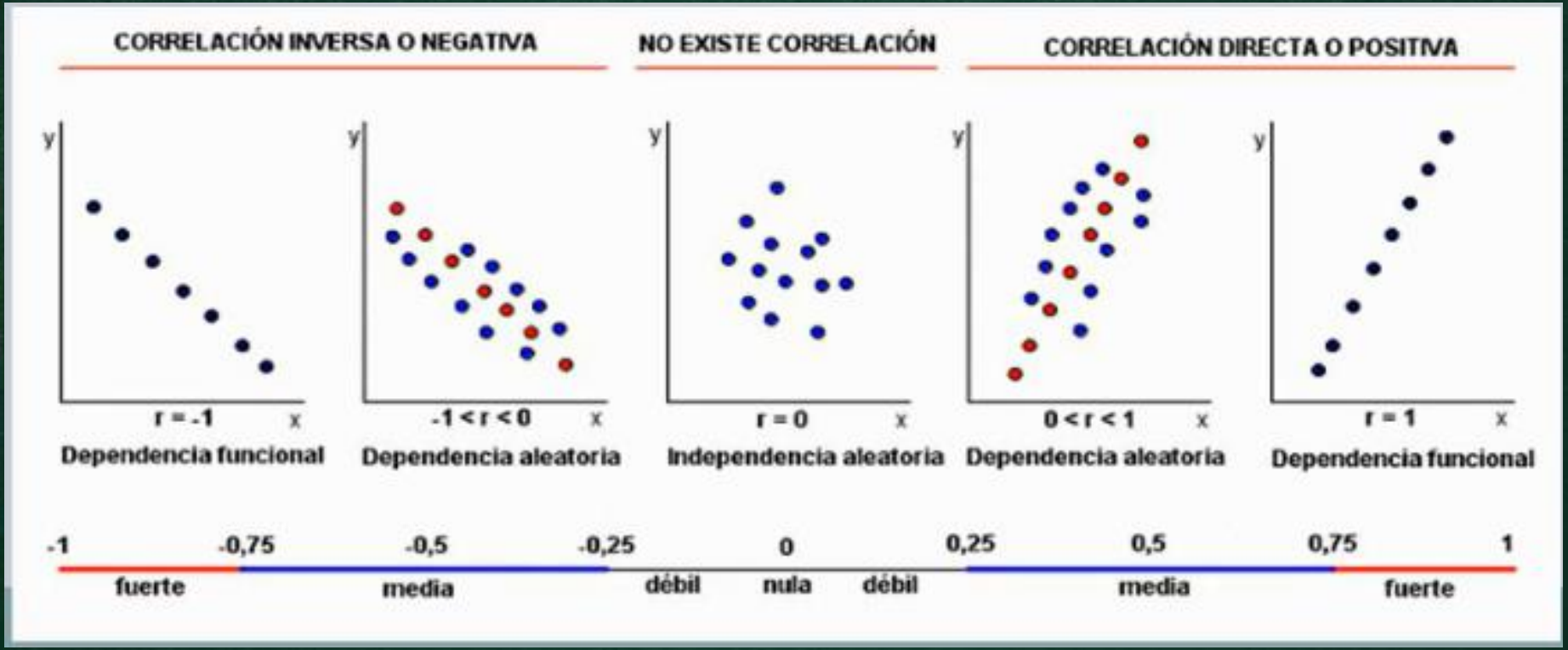
-Un valor positivo se interpreta como indicador de una relación directa: A medida que aumentan los valores de una variable aumentan los valores de la otra.

-Un valor negativo se interpreta como indicador de una relación inversa : A medida que aumentan los valores de una variable disminuyen los valores de la otra.

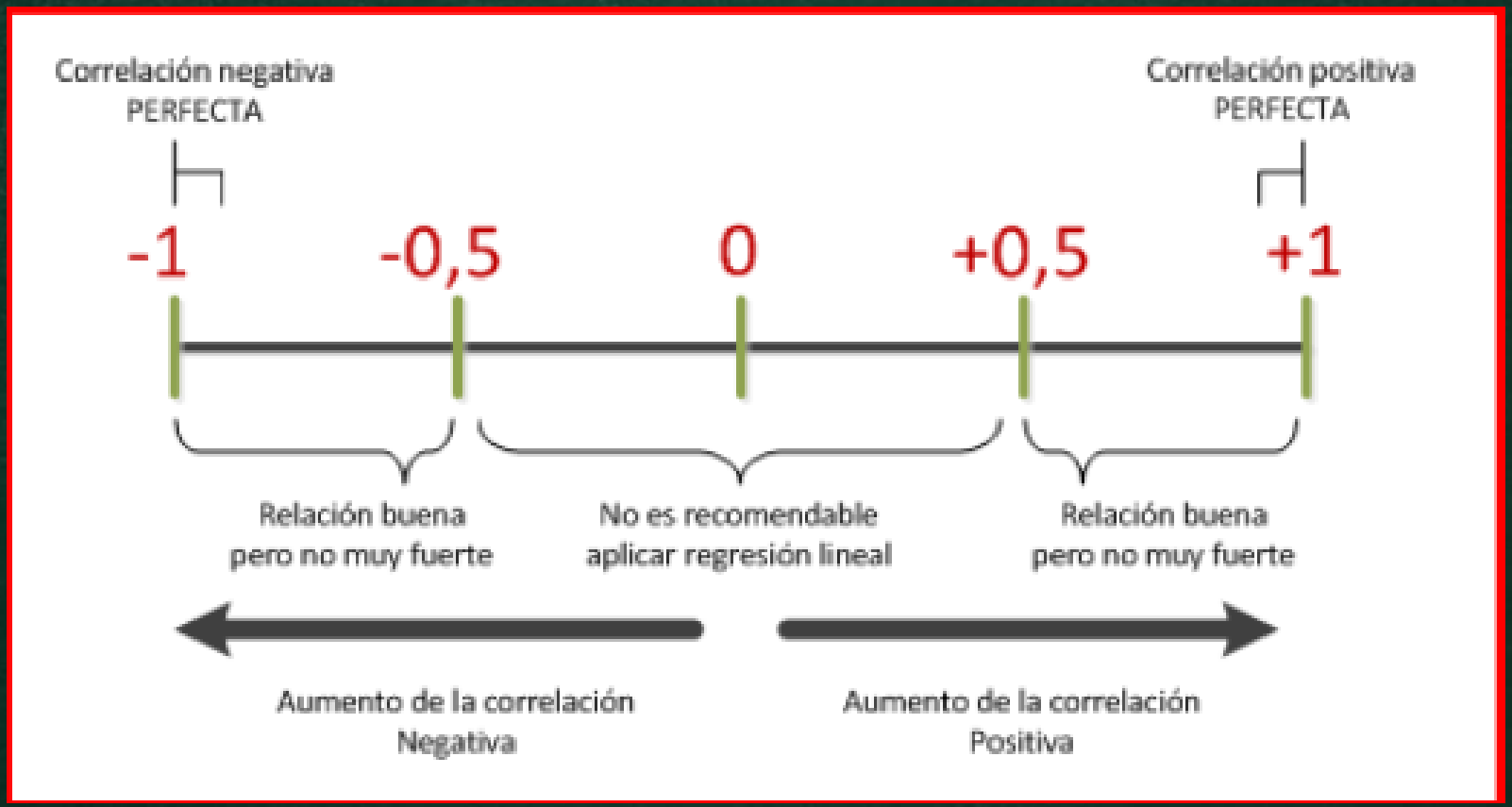
-El valor absoluto se interpreta como el grado de relación lineal existente entre las variables, que será mayor cuanto más cercano sea a 1.

-Si el valor del coeficiente de correlación muestral, en valor absoluto, es mayor de 0,90 se considera buena la estimación que se realiza con la recta de regresión.

# INTERPRETACIÓN



# INTERPRETACIÓN



# Ejemplo de Aplicación.

Un Ingeniero Agrónomo ha analizado la relación entre la cantidad de agua aplicada mediante riego artificial por aspersión en $m^3$ y las toneladas de granos cosechadas en toneladas por hectárea obteniendo la siguiente tabla:

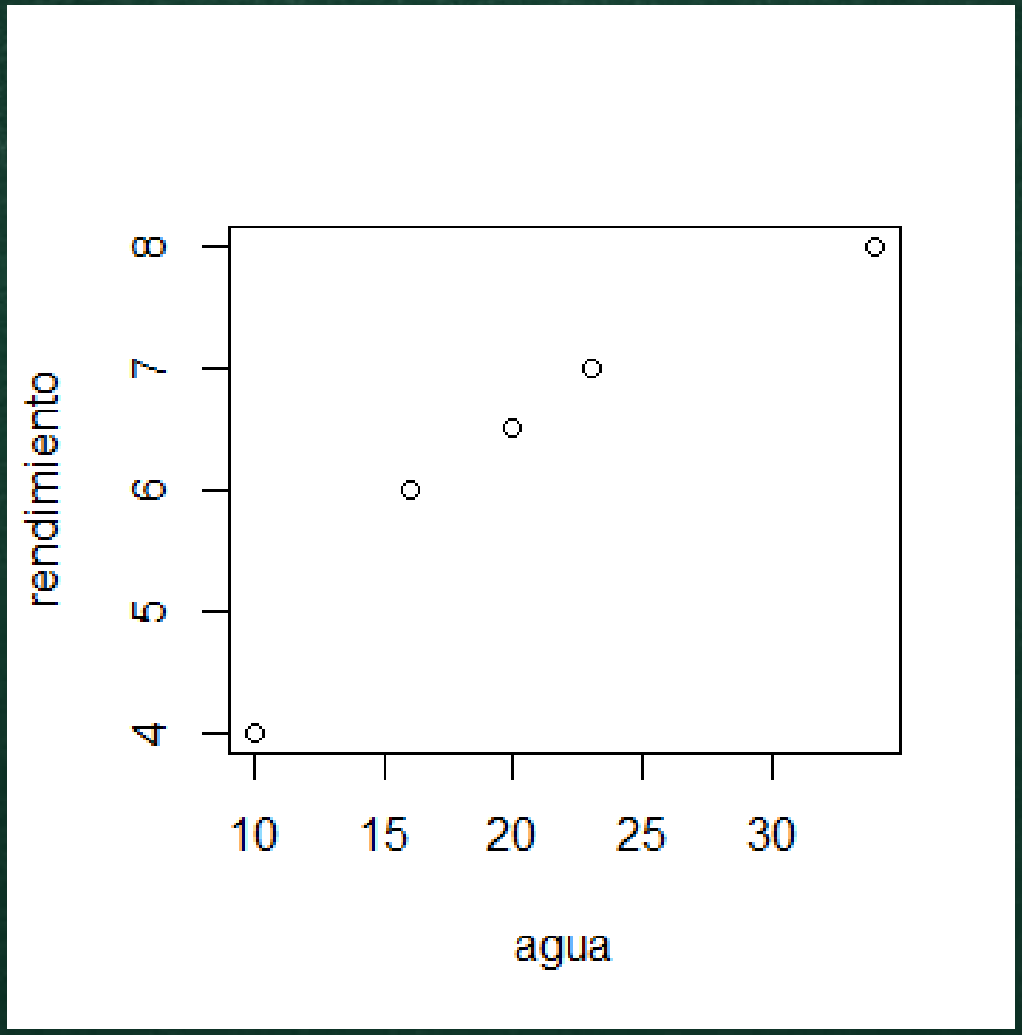
| **X: cantidad de agua en $m^3$** | **20** | **16** | **34** | **10** | **23** |
|---|---|---|---|---|---|
| **Y: toneladas por hectárea** | 6,5 | 6 | 8 | 4 | 7 |

a) Calcular la recta de regresión lineal de las toneladas de granos por hectárea en función del agua en $m^3$

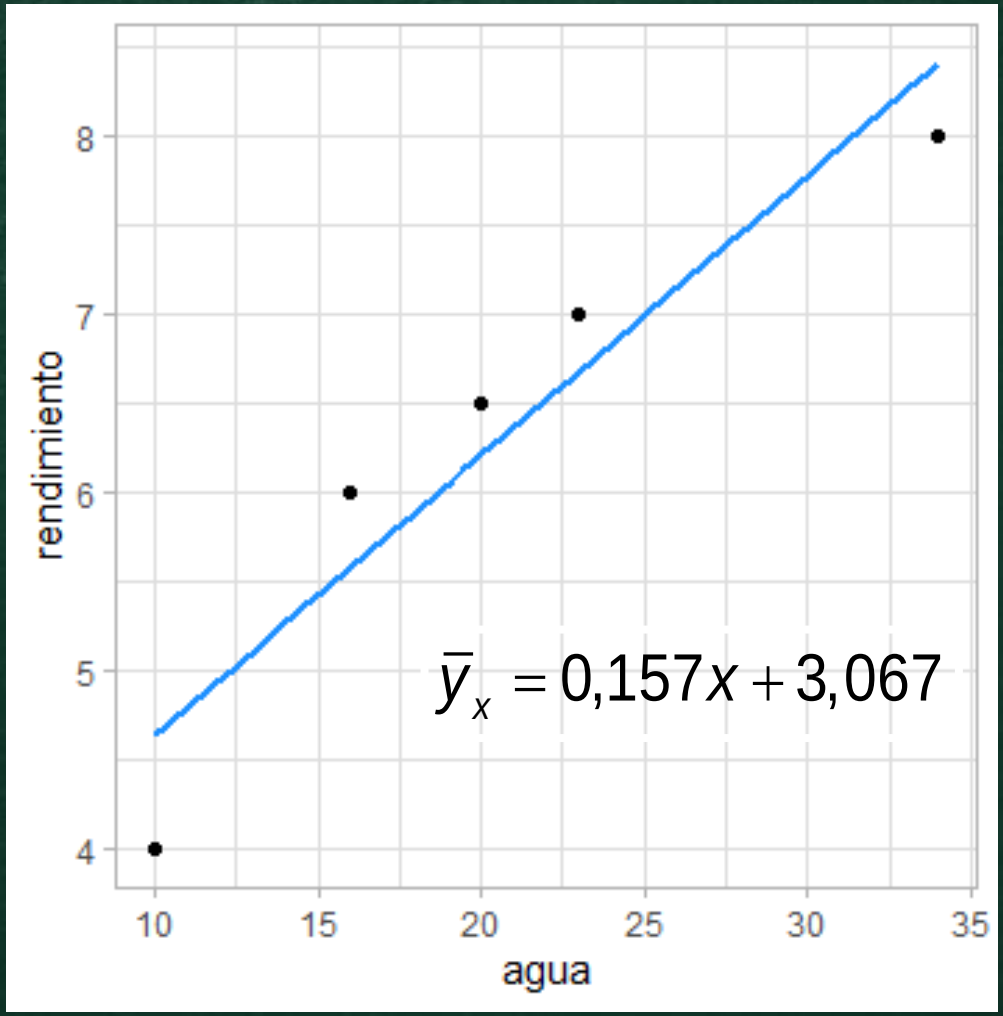
b) Calcular el coeficiente de correlación e interpretar su valor

c) Calcular la cantidad estimada de toneladas por hectárea si se usan 28 $m^3$ de agua

## Diagrama de dispersión



## Diagrama de dispersión



# Clase Práctica 10 hs….