Análisis Estadístico con R

Importancia de definir el problema

Objetivos / hipótesis de trabajo

¿Cuál es el problema detectado?

¿Está validado que sea un problema?

¿Por qué?

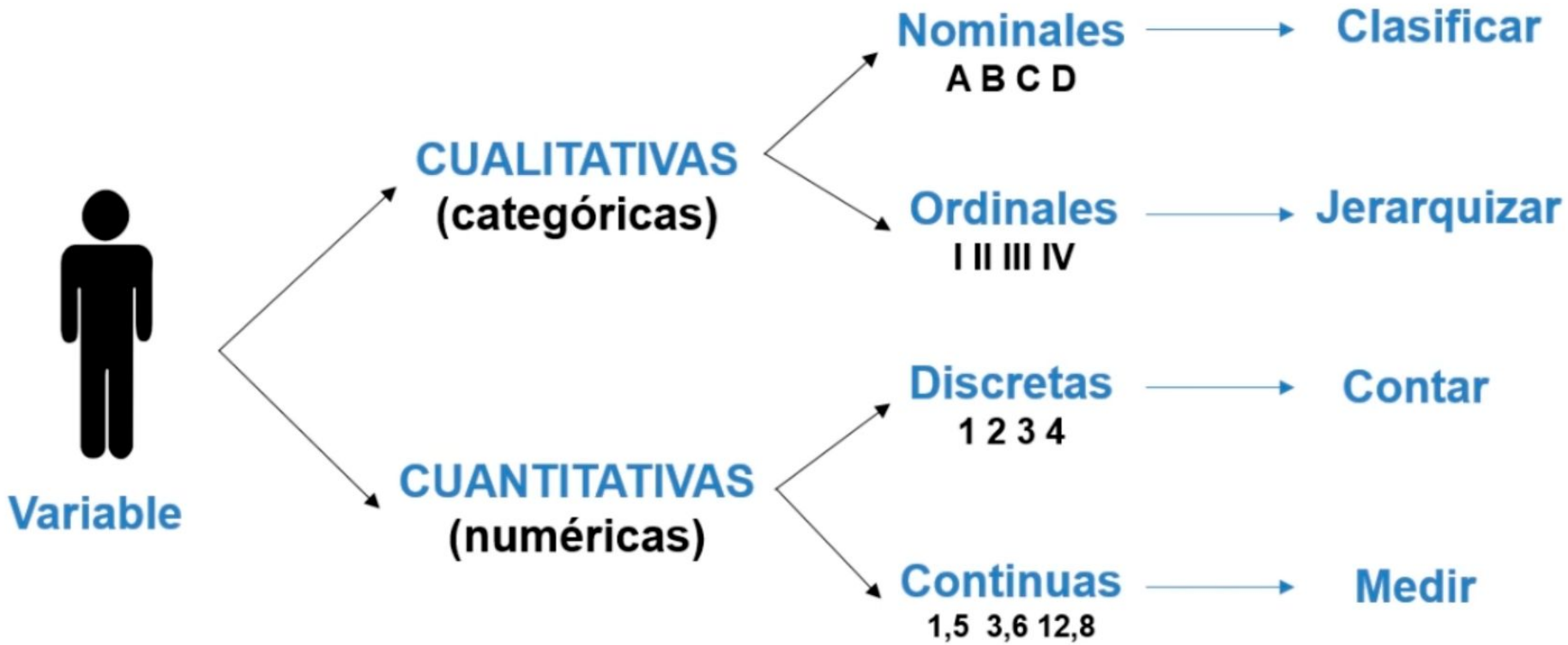
¿Para qué?

Aportes potenciales

Metodología

Duración y etapas del proyecto

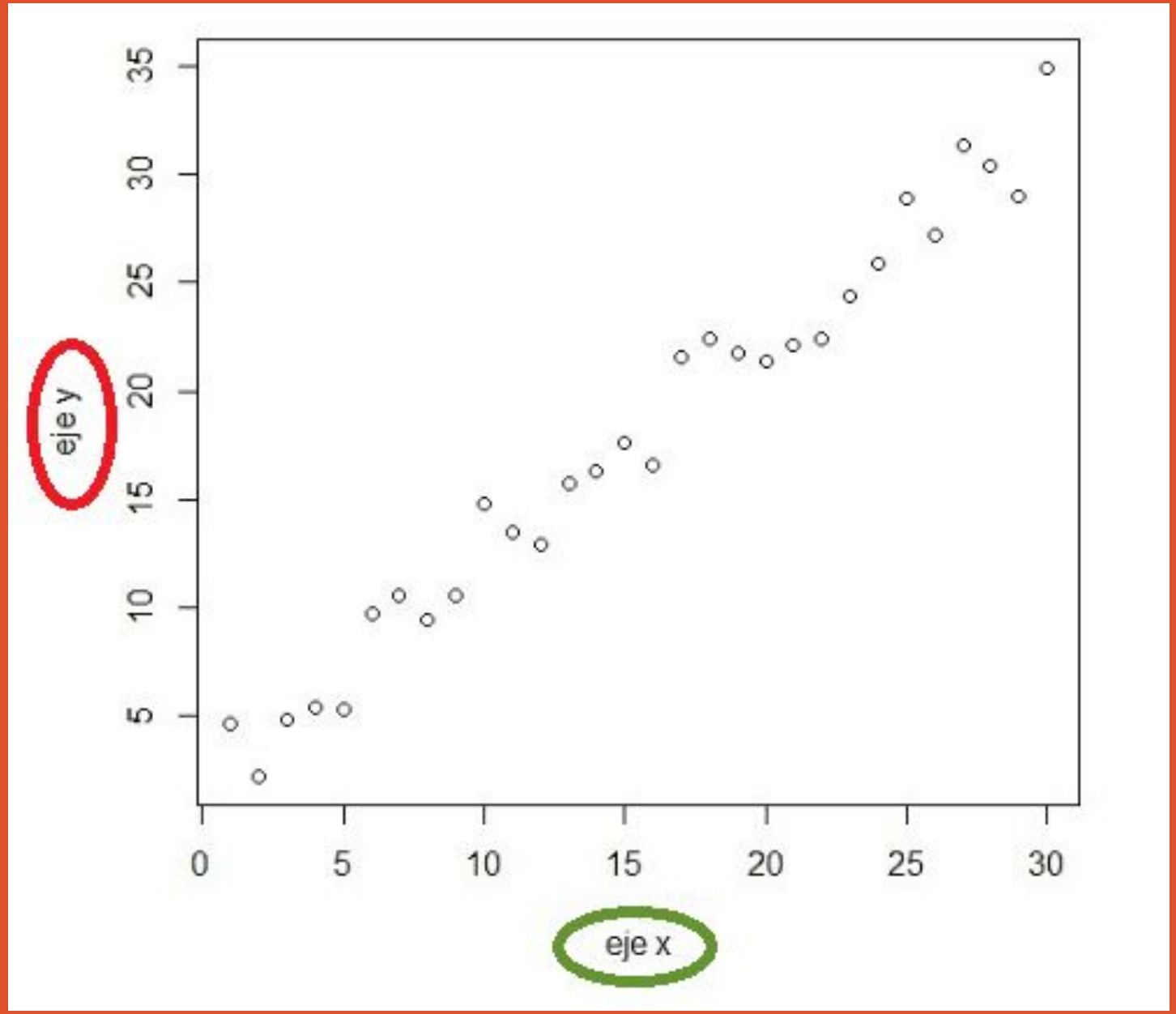
Presupuesto

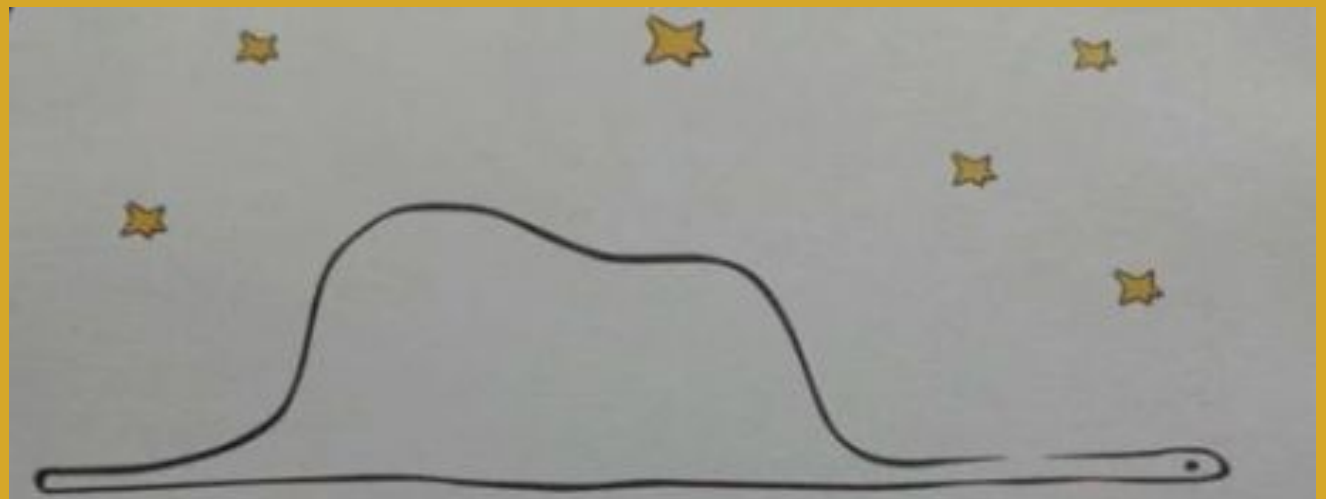


$$y = f(x)$$

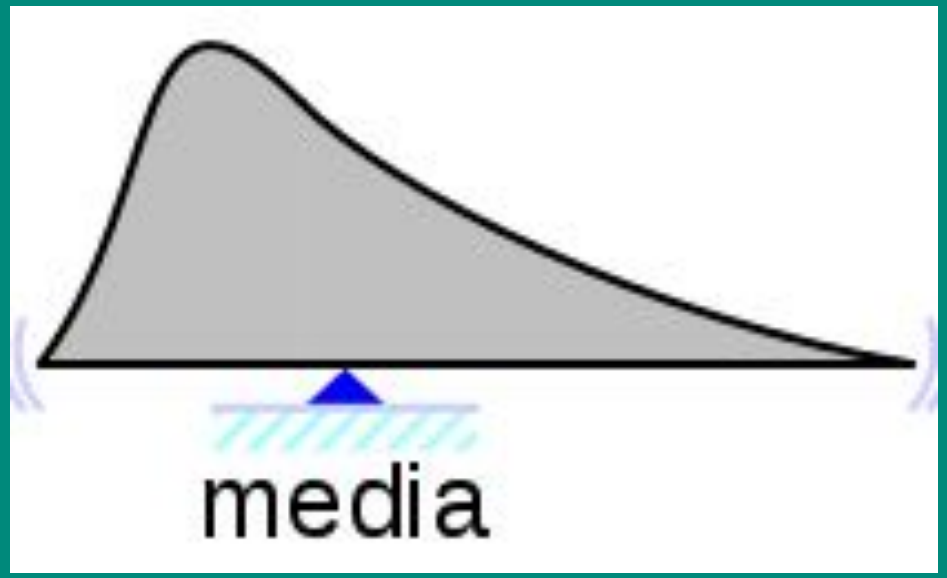
Variable dependiente

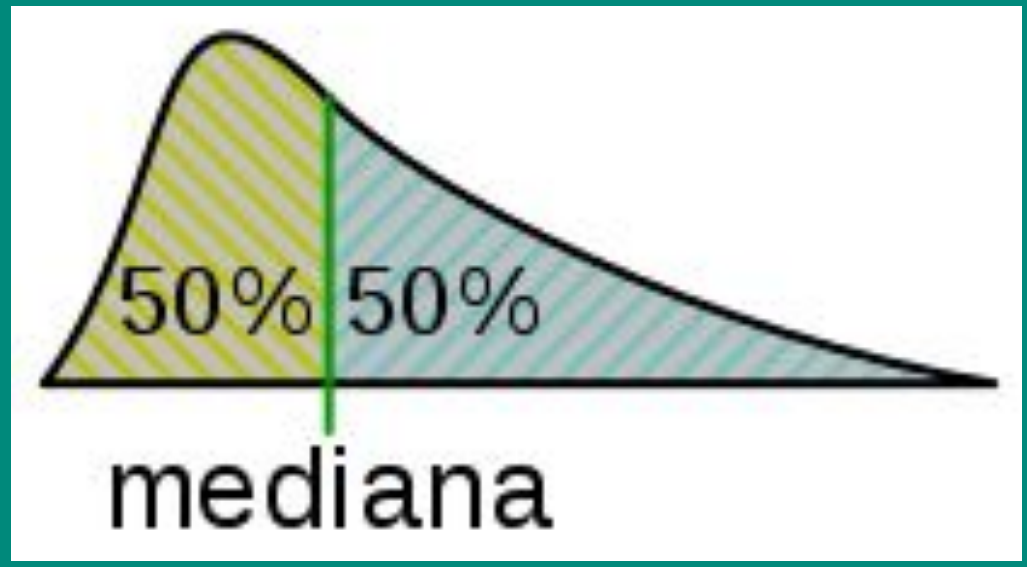
Variable independiente

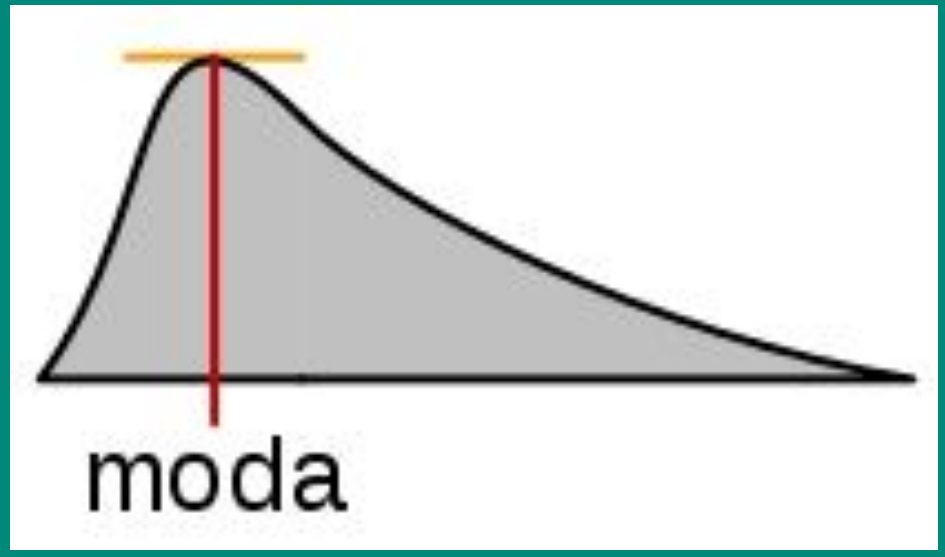




Este es mi secreto. Es muy simple: solo se ve bien con el corazón.
Lo esencial es invisible a los ojos.


media


50%
50%
mediana


moda

Media

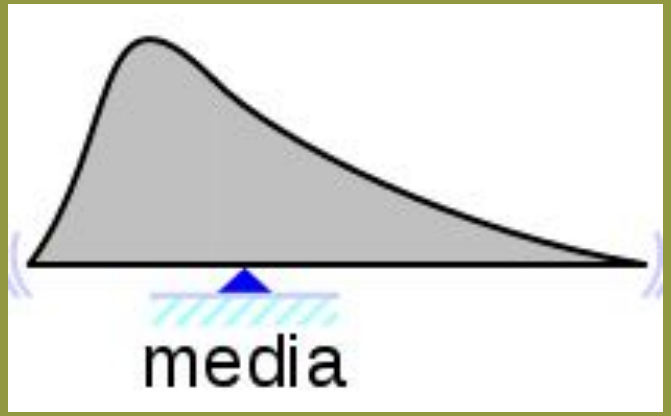
Es el valor característico de la serie de datos resultado de la suma de todas las observaciones dividido por el número total de datos

Ventajas:

- Es muy usado y fácil de comprender
- Es útil como medida de comparación entre datos.

Desventajas:

- se ve fuertemente afectado por los valores extremos
- No es recomendable emplearla en distribuciones muy asimétricas



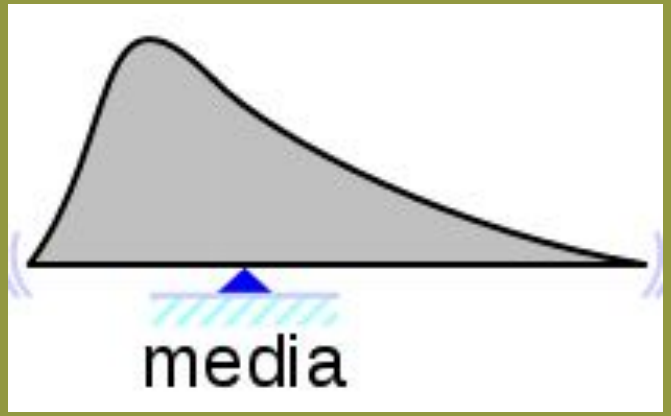
Media

1	1	1	3	3	4	4	6	6

3.22

Ejemplo:

$$\bar{x} = \frac{1+1+1+3+3+4+4+6+6}{9} = 3.22$$



Media

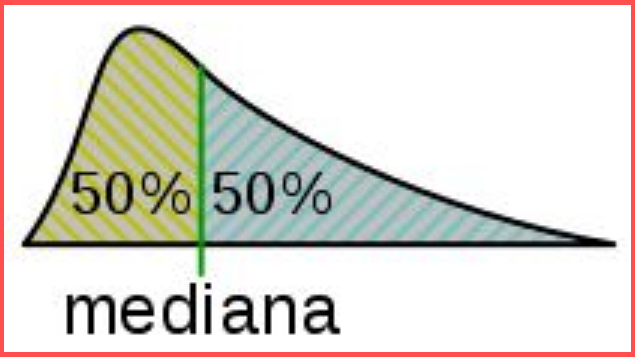
Ventajas:

- Es muy usado y fácil de comprender
- Es útil como medida de comparación entre datos.

Desventajas:

- se ve fuertemente afectado por los valores extremos
- No es recomendable emplearla en distribuciones muy asimétricas

Mediana

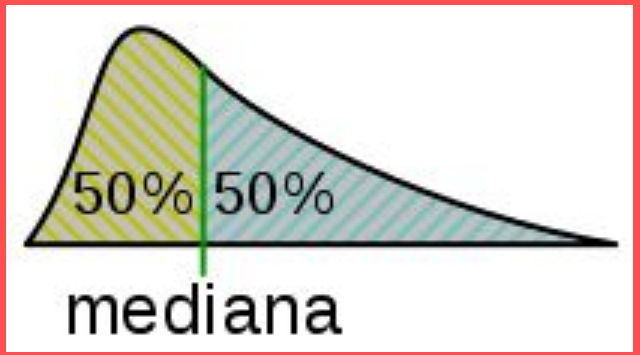


Es el número de la mitad en un conjunto de números ordenado de menor a mayor.

Mediana

1	1	1	3	3	4	4	6	6
1	2	3	4	5	6	7	8	9

3



Ejemplo:

$\tilde{x}$ = 0.5*(9+1) = 5

nos da la posición que tenemos que mirar en la tabla

Mediana

Ventajas:

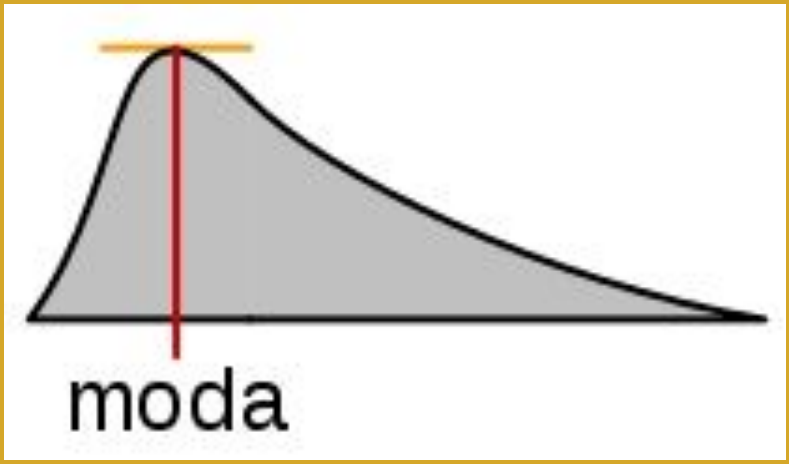
- No se ve afectado por valores extremos
- Es fácil de comprender
- Es la medida de tendencia central más representativa en el caso de variables que solo admiten la escala ordinal.

Desventajas:

- Hay que ordenar los datos antes de determinarla
- No pondera cada valor por el número de veces que se ha repetido.

Moda

Es el valor con mayor frecuencia en una de las distribuciones de datos



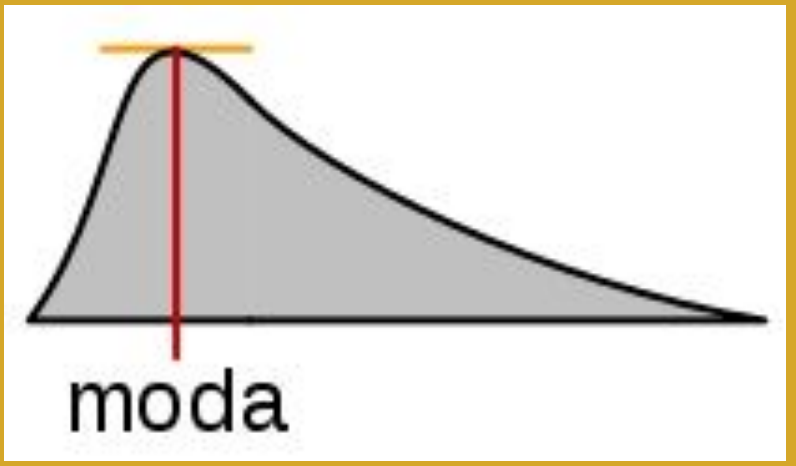
Moda

1	1	1	3	3	4	4	6	6

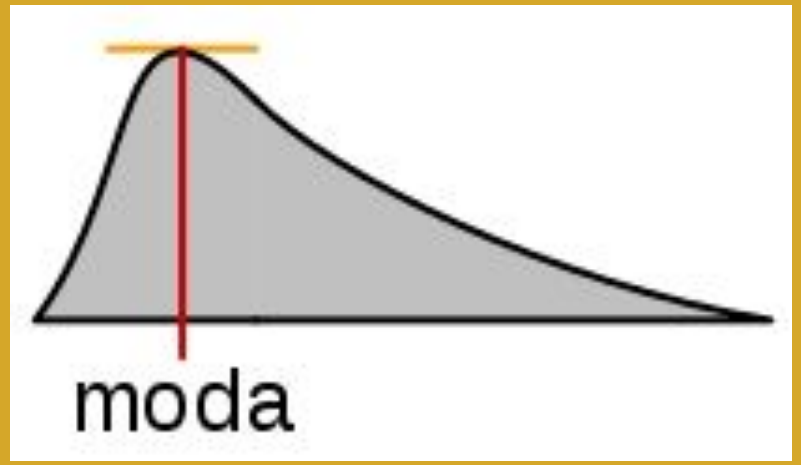
1

Ejemplo:

Valor	Cant de veces que aparece
1	3
3	2
4	2
6	2



Moda



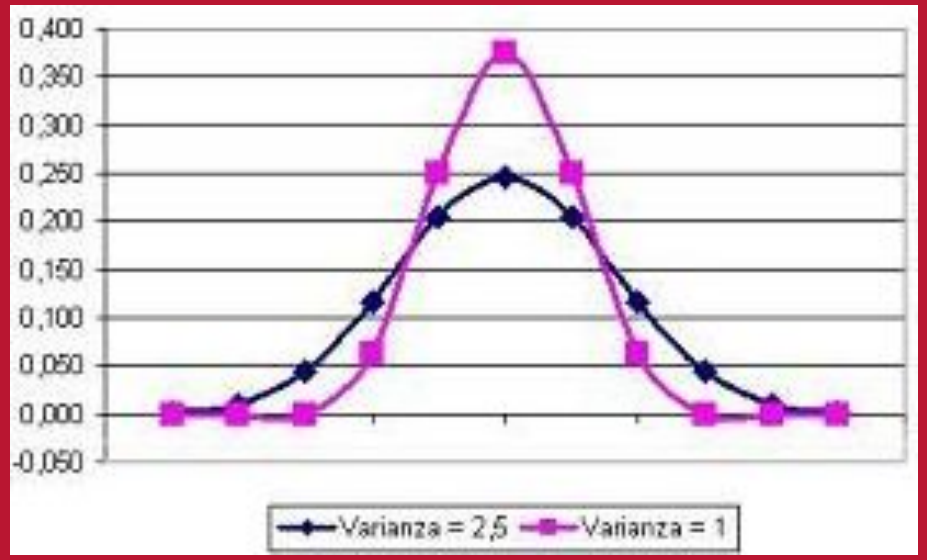
Ventajas:

- No requiere cálculos.
- Fácil de interpretar.
- No se ve influenciada por valores extremos.

Desventajas:

- No siempre existe, si los datos no se repiten.
- Muchas veces no existe moda (distribución amodal).
- No tiene un uso tan frecuente como la media.

Varianza



Mide la variabilidad de los datos alrededor de la media.

Es decir, es la distancia de c/ valor de la media

Varianza

1	1	1	3	3	4	4	6	6

S^2 = 3.94

Ejemplo:

$$s^2 = \frac{(1-3.22)^2+ (1-3.22)^2 + (1-3.22)^2 + (3-3.22)^2 + (3- 3.22)^2 + (3-3.22)^2 + (4-3.22)^2 + (4-3.22) + (6-3.22)^2 + (6-3.22)^2}{9 - 1} = 3.94$$

Desvío Estándar

Indica qué tan dispersos están los datos con respecto a la media. Mientras mayor sea la desviación estándar, mayor será la dispersión de los datos.

Desvío Estándar

1	1	1	3	3	4	4	6	6

S = 1.98

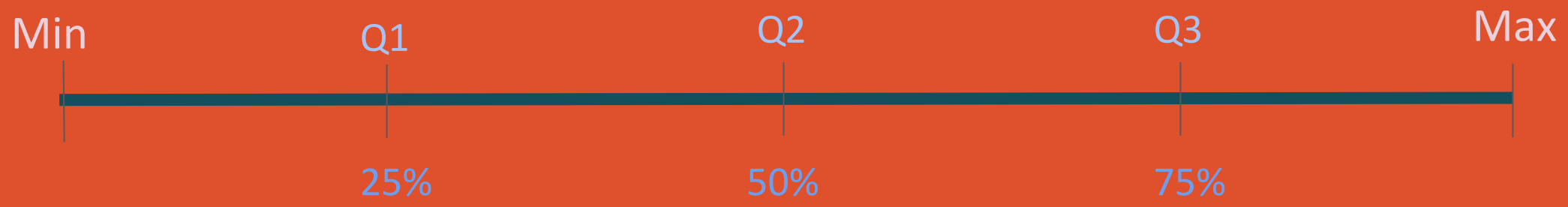
Ejemplo:

$$S = \sqrt{\frac{(1-3.22)^2 + (1-3.22)^2 + (3-3.22)^2 + (3- 3.22)^2 + (3-3.22)^2 + (4-3.22)^2 + (4-3.22) + (6-3.22)^2 + (6-3.22)^2}{9 - 1}} = 1.98$$

Cuartiles

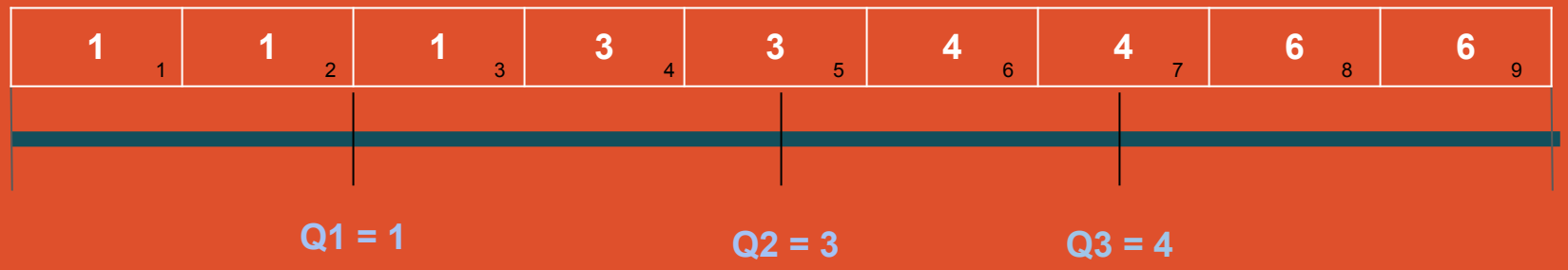
Son valores que dividen una muestra en partes iguales

Cuartiles



Son valores que dividen una muestra en partes iguales

Cuartiles



Ejemplo:

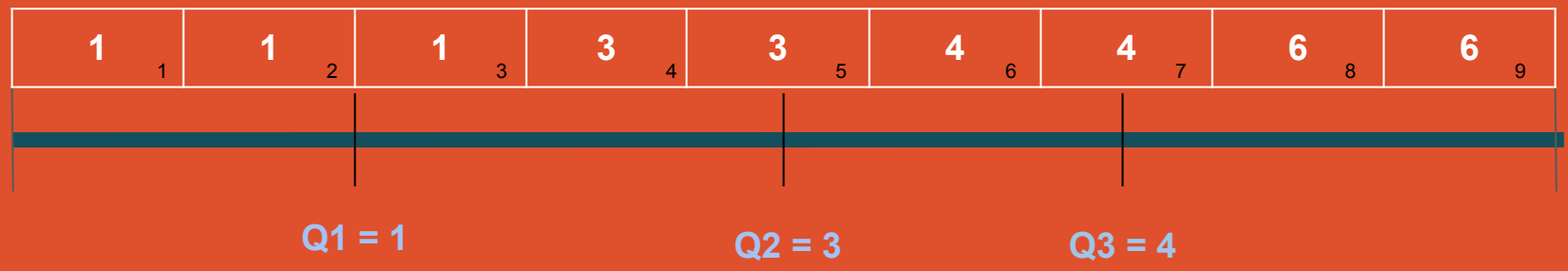
	Cálculo	Posición en la tabla	Valor de la tabla
Q1	0.25*(9+1)	2.5	1
Q2	0.50*(9+1)	5	3
Q3	0.75*(9+1)	7.5	4
	0.95*(9+1)	9.5	6

Rango Intercuartil

Es la diferencia entre el tercer y el primer cuartil

IQR = Q3-Q1

Rango Intercuartil



Ejemplo:

IQR = 4 - 1 = 3

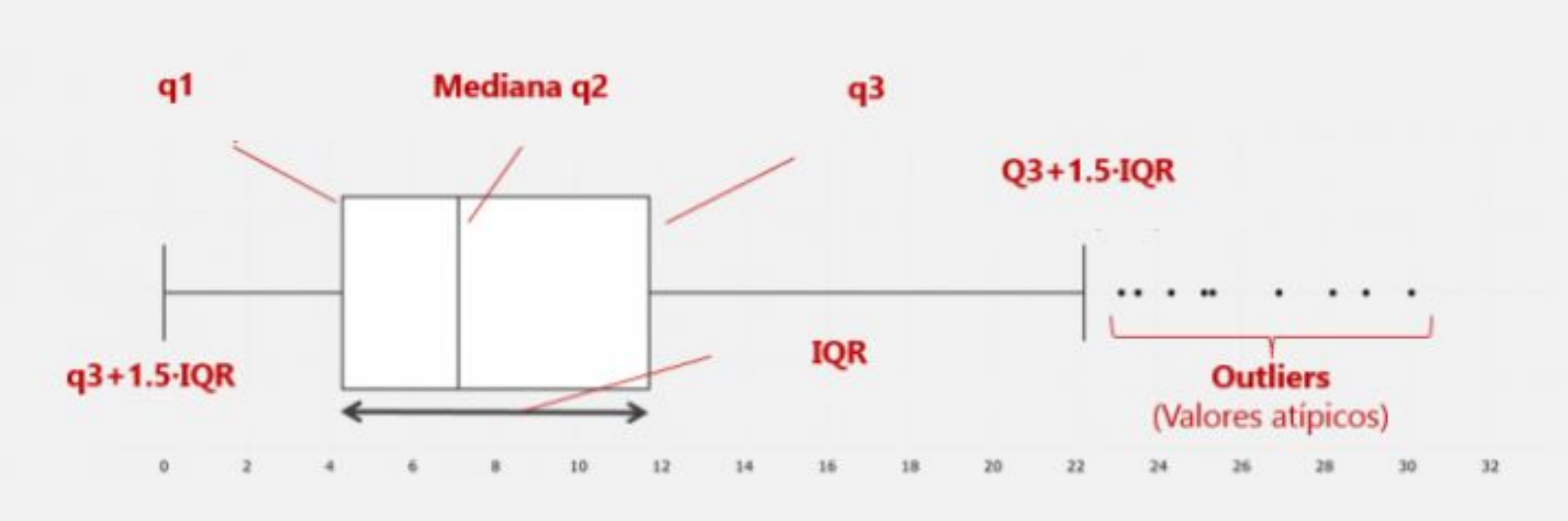
IQR = Q3-Q1

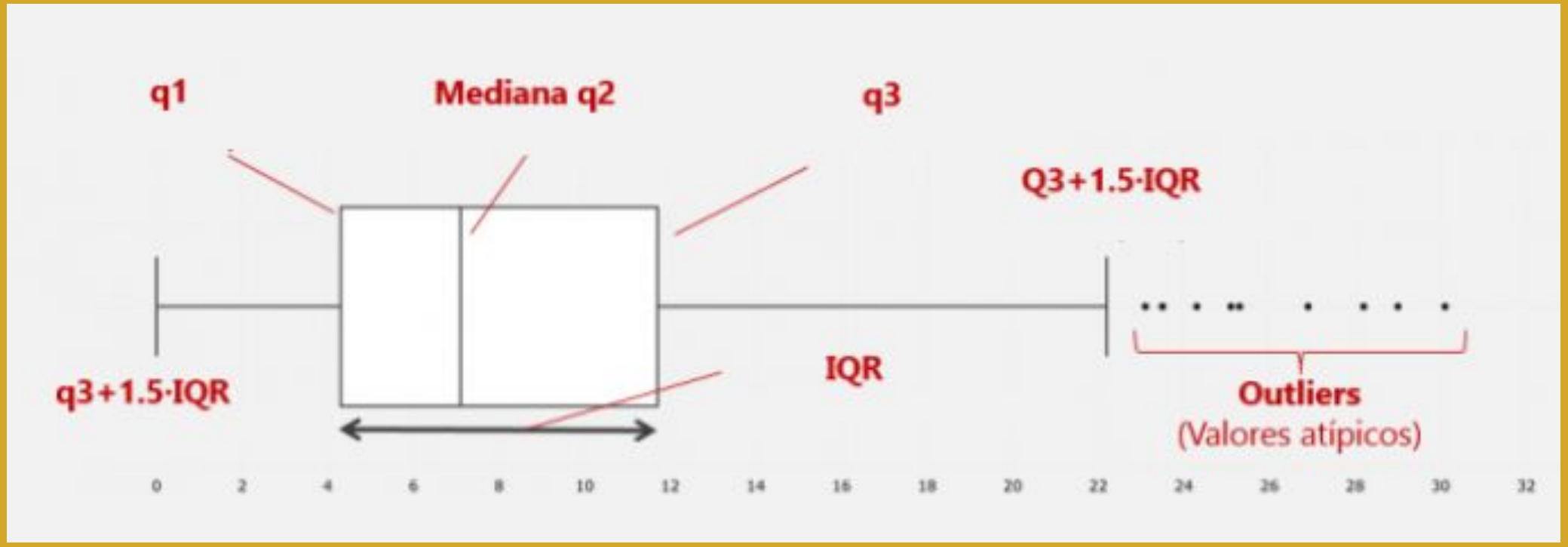
Boxplot

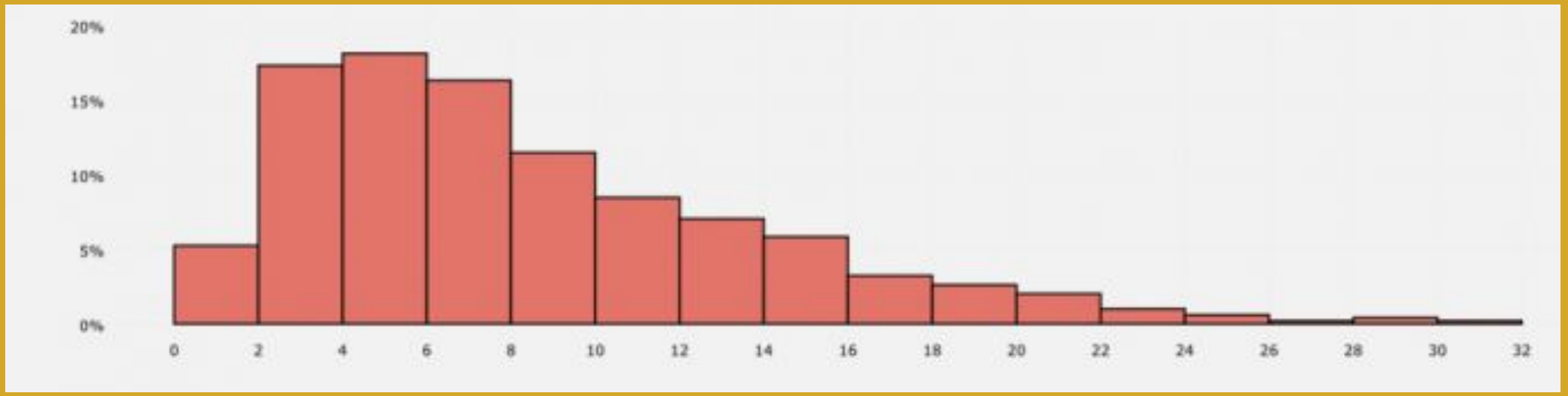
Es un tipo de gráfico que muestra un resumen de una gran cantidad de datos en cinco medidas descriptivas, además de intuir su morfología y simetría.

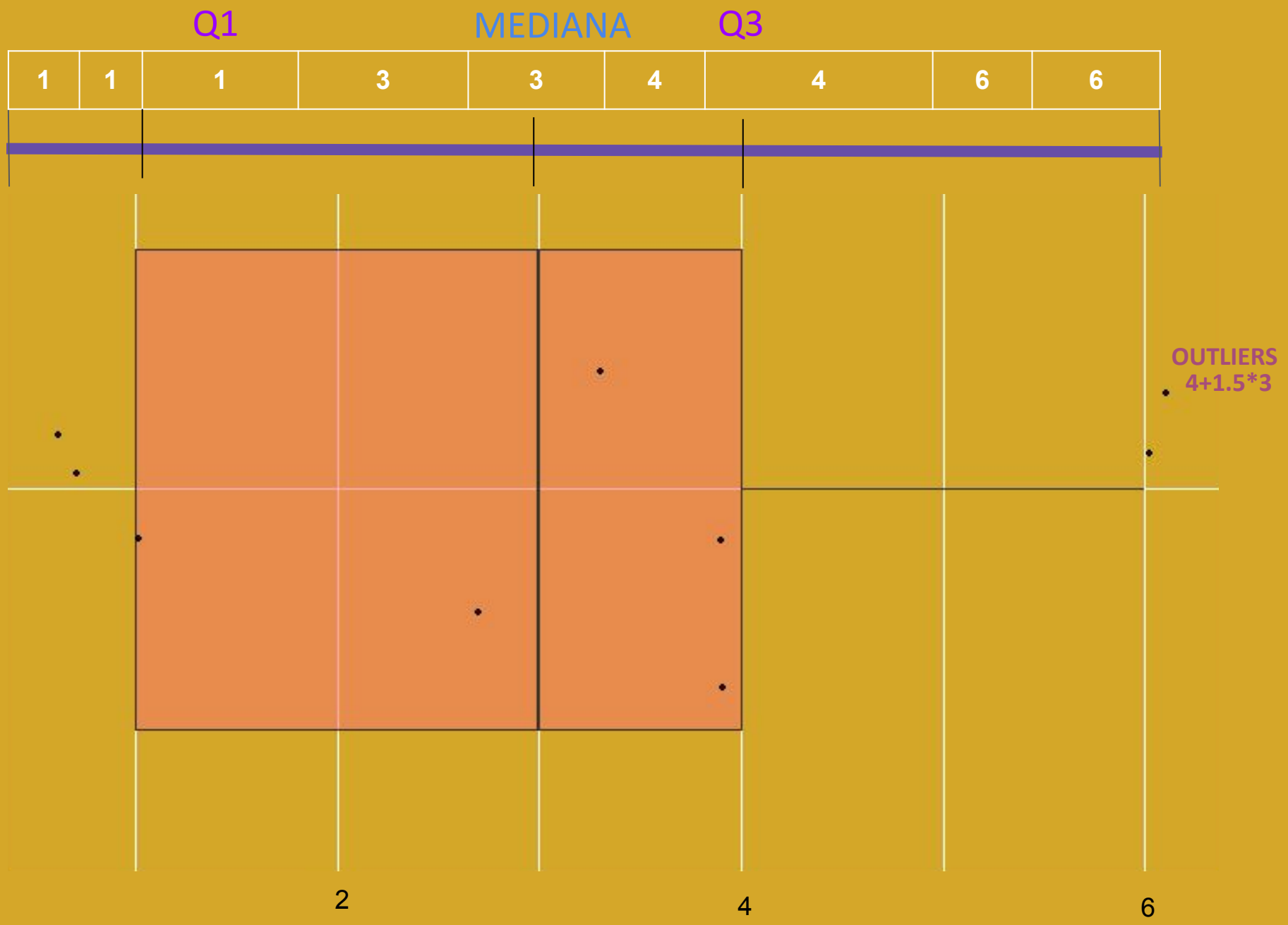
Nos permite identificar valores atípicos y comparar distribuciones.

Boxplot

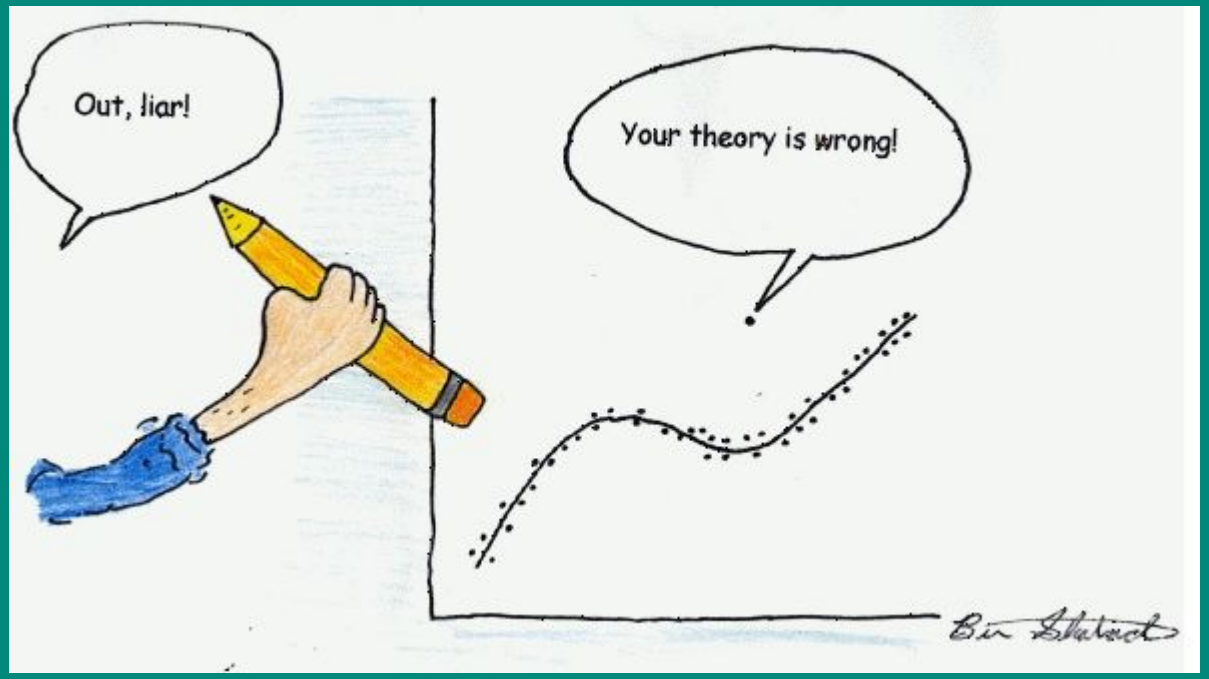

q1
Mediana q2
q3
Q3+1.5·IQR
q3+1.5·IQR
IQR
Outliers
(Valores atípicos)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32


q1
Mediana q2
q3
Q3+1.5·IQR
q3+1.5·IQR
IQR
Outliers
(Valores atípicos)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32


20%
15%
10%
5%
0%
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32


Q1
MEDIANA
Q3
1
1
1
3
3
4
4
6
6
OUTLIERS
4+1.5*3
2
4
6

OUTLIERS

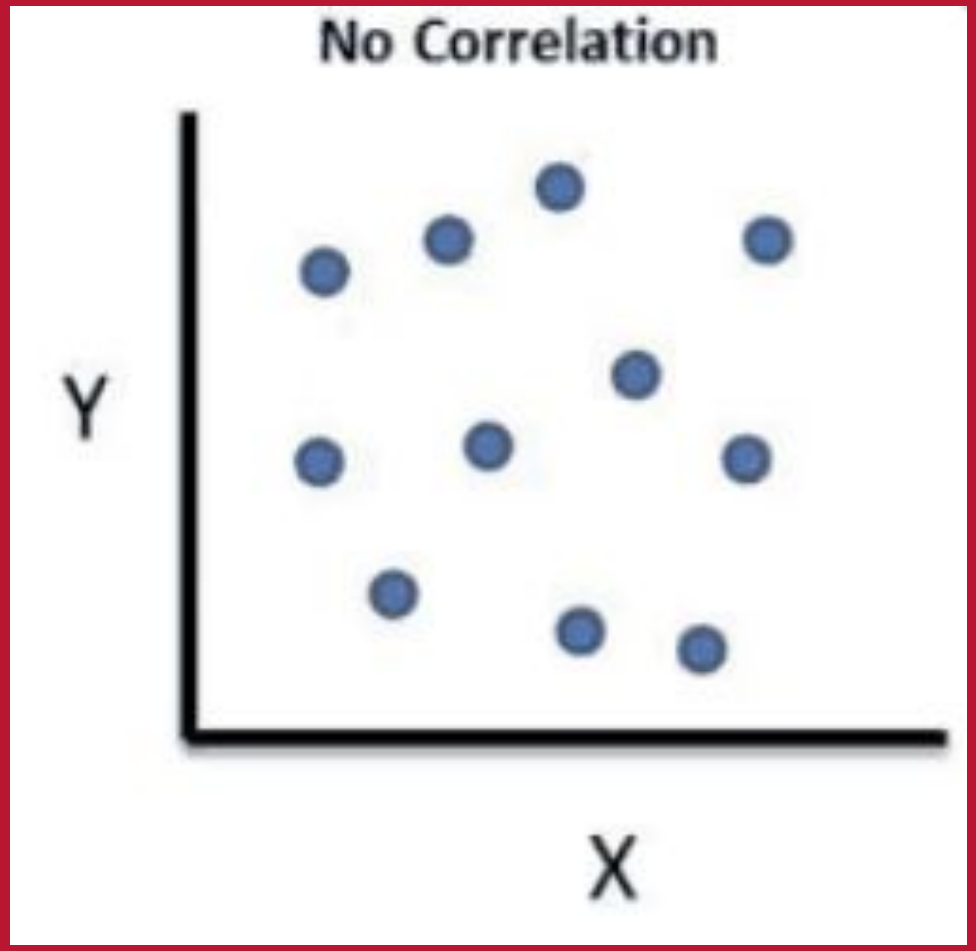


OUTLIERS

Son observaciones numéricamente distante del resto de los datos

Métodos para lidiar con outliers:

- Media podada
- Eliminarlos a partir de un umbral de interés
- Reemplazar por la media los que están debajo del umbral y por la mediana los que están por encima del umbral

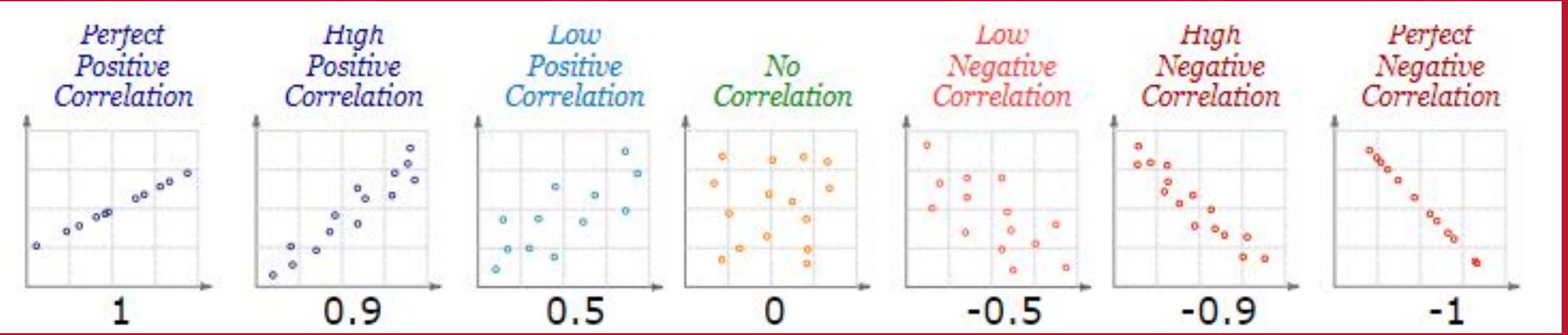


https://www.tylervigen.com/spurious-correlations

Correlación

Indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas

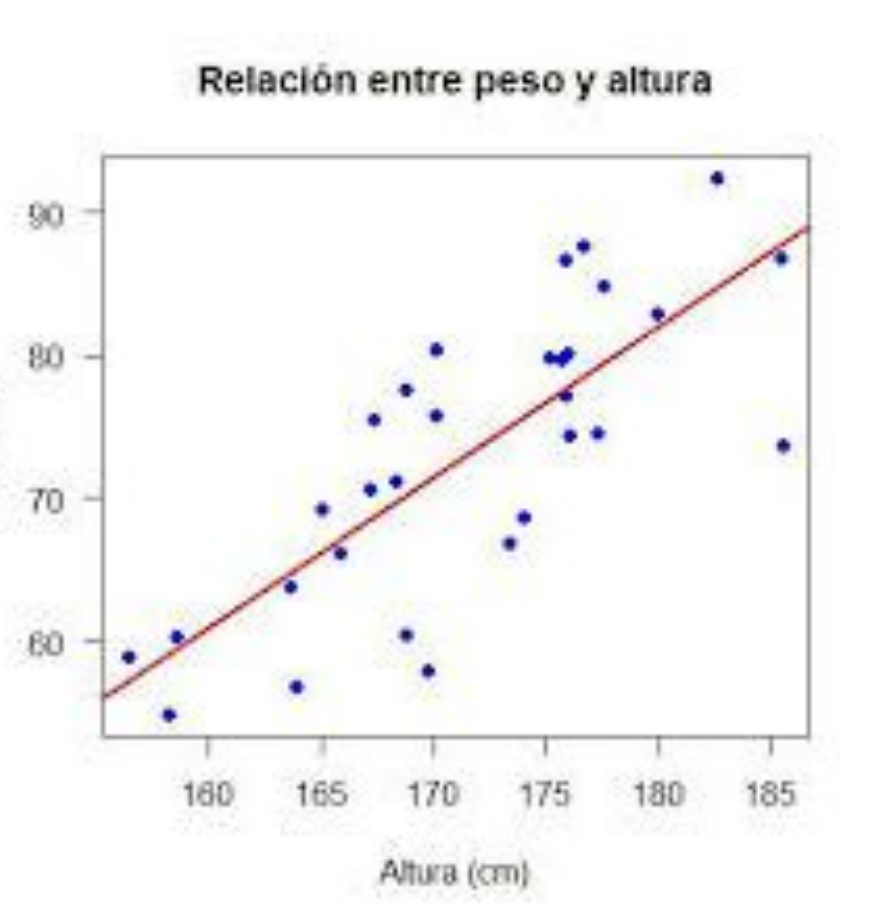
Correlación



Regresión Lineal

*Se utiliza para **predecir** el valor de y (variable objetivo, dependiente) dados los valores de x (denominadas variables explicativas, independientes o regresores)*

Regresión Lineal



Ahora les toca a uds...

¿Qué análisis deberían hacer para la validación de datos?