Certificación
Universitaria

DATA SCIENCE

E

Análisis de Datos I

Sol Represa

CC

UNC FCEFyN mundosE PEOPLE & BUSINESS SCHOOL

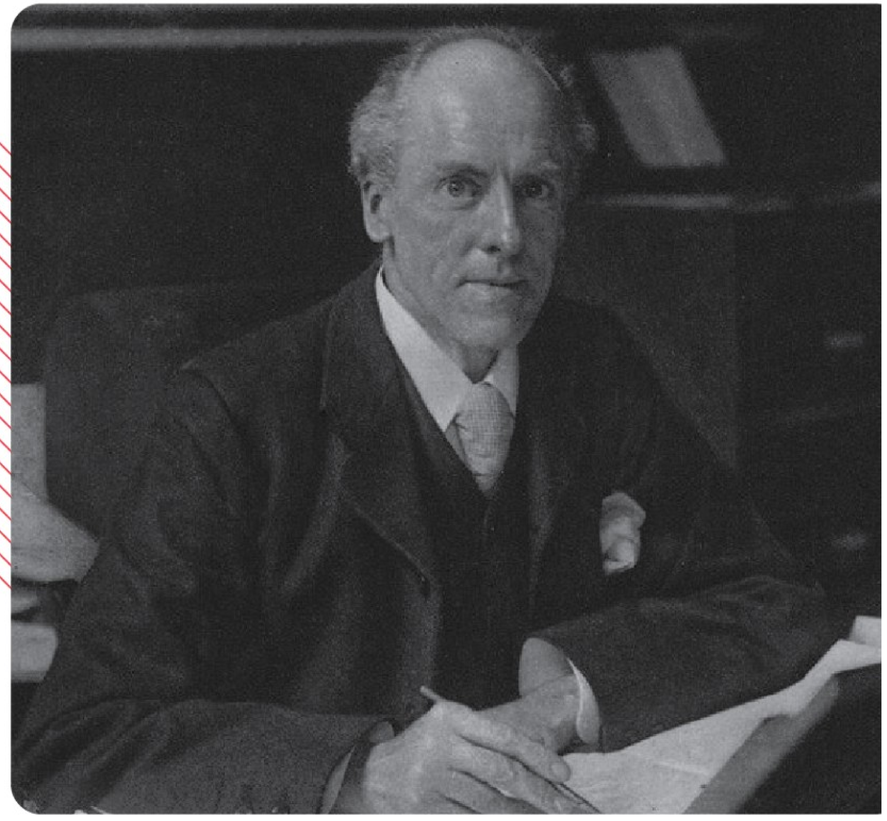
PROFESORA
Sol Represa

Actualmente

- Líder en Ciencias de Datos - NTT Data
- Docente universitaria UNLP
- Consultora en Ciencia de Datos

Statistics is the grammar of science.

Karl Pearson



¿Para qué sirve la estadística?

¿Para qué sirve la estadística?

Para **descubrir patrones** → **Estadística descriptiva o deductiva**

Para **confirmar una hipótesis** → **Estadística inferencial o inductiva**

Análisis exploratorio

(o el arte de hacer
confesar a los datos)

Análisis exploratorio

/ Tiene como **objetivo** identificar:

- patrones subyacentes
- relaciones entre las variables
- el modelo teórico más adecuado para representar la población de la cual proceden los datos muestrales.

Qué veremos en este encuentro?

Algunos conceptos importantes → Técnicas de descripción numérica

Variable

/ Cuantitativa

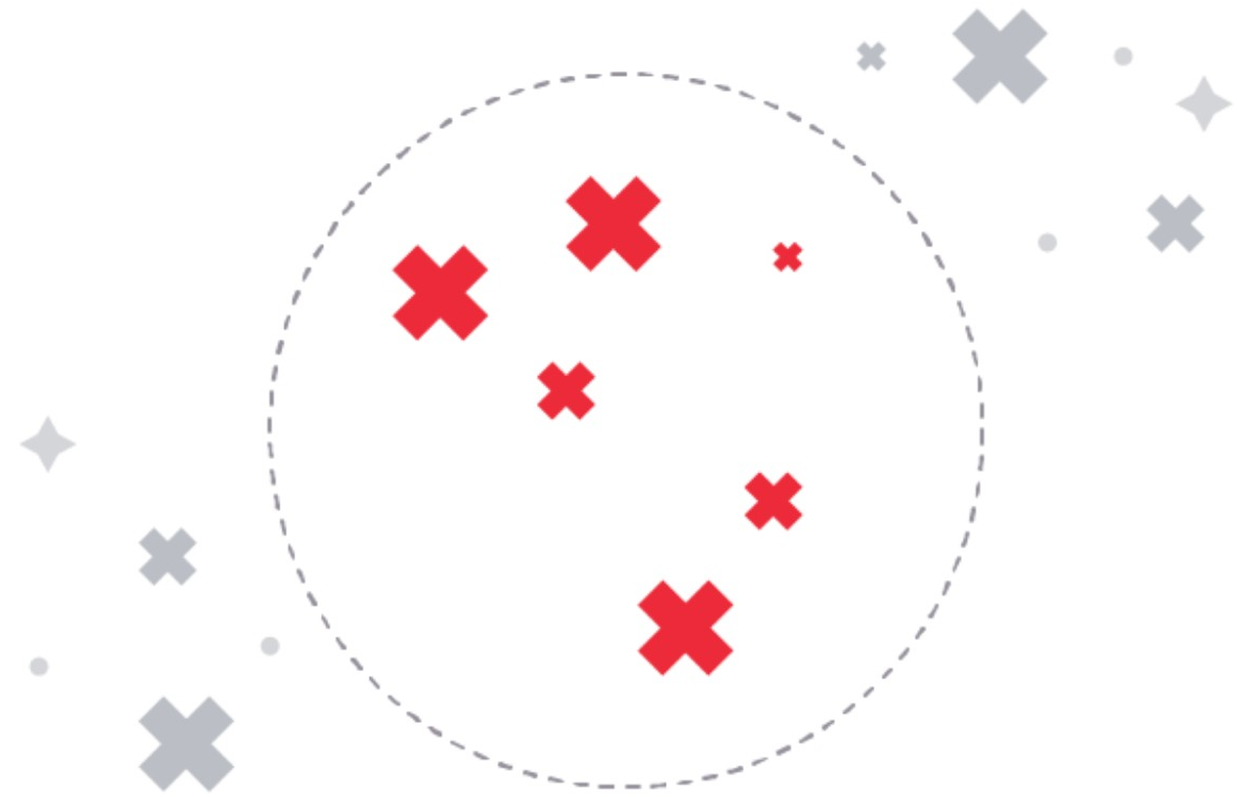
- Discreta
- Continua

/ Cualitativa

- Nominal
- Ordinal

Muestra

Subconjunto de una **población**.

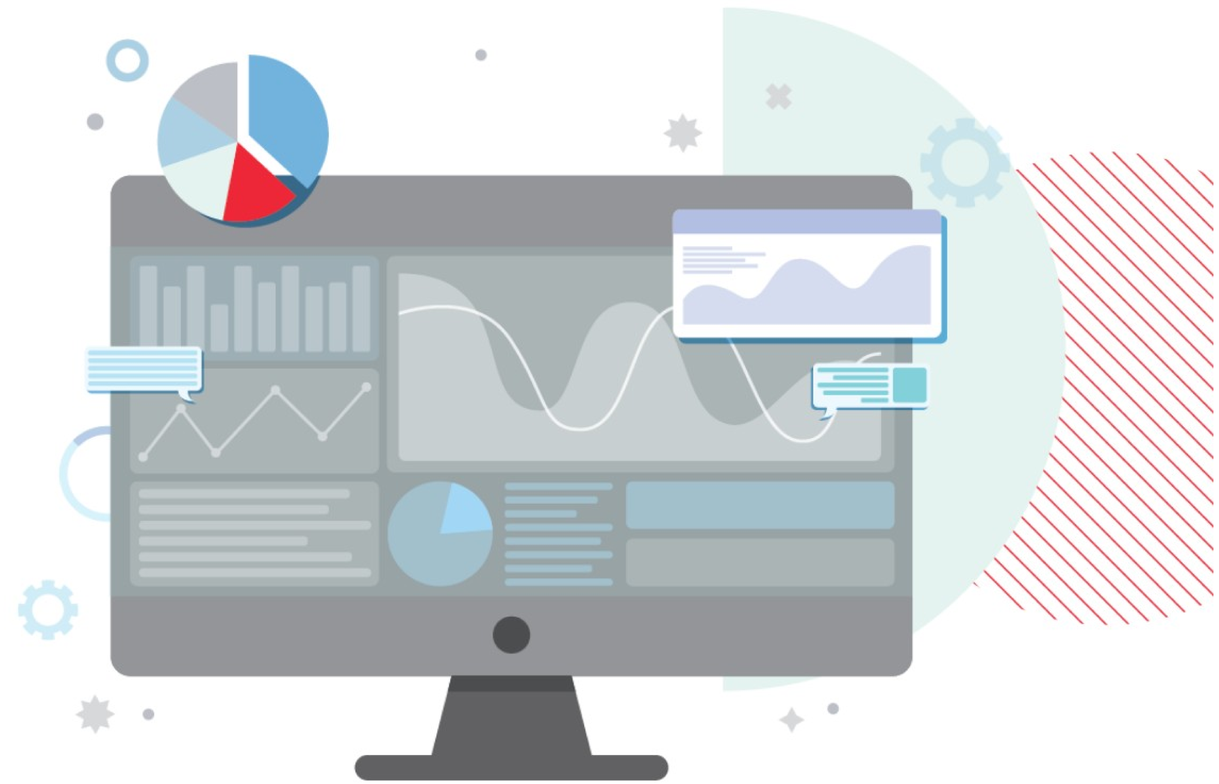


Inferencia estadística

Proceso mediante el cual basándonos en una muestra sacamos conclusiones acerca de la población.

El **objetivo final del análisis estadístico** no es el estudio de la muestra, sino el **estudio de la población**.

Distribución muestral



Funciones de distribución poblacional

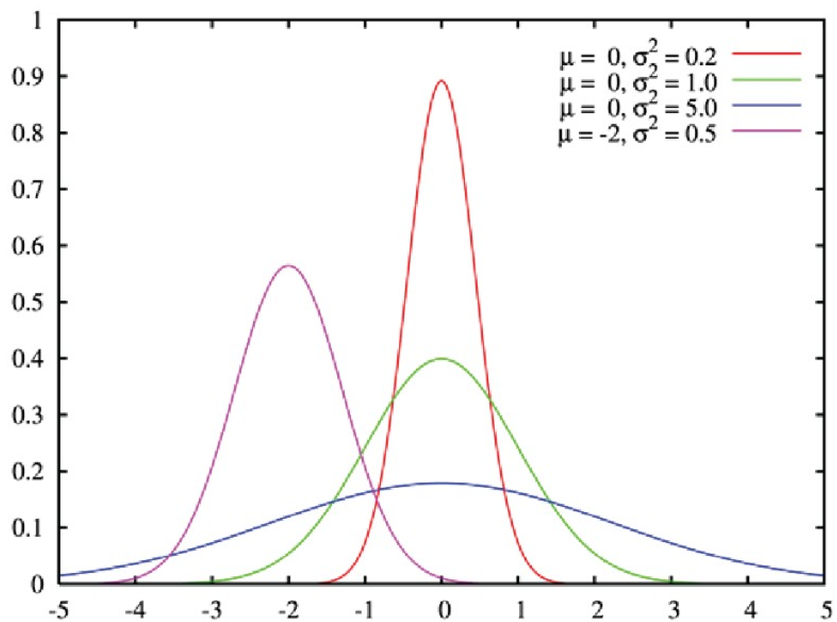
/ Variable contínua

- Normal
- Lognormal
- Gamma

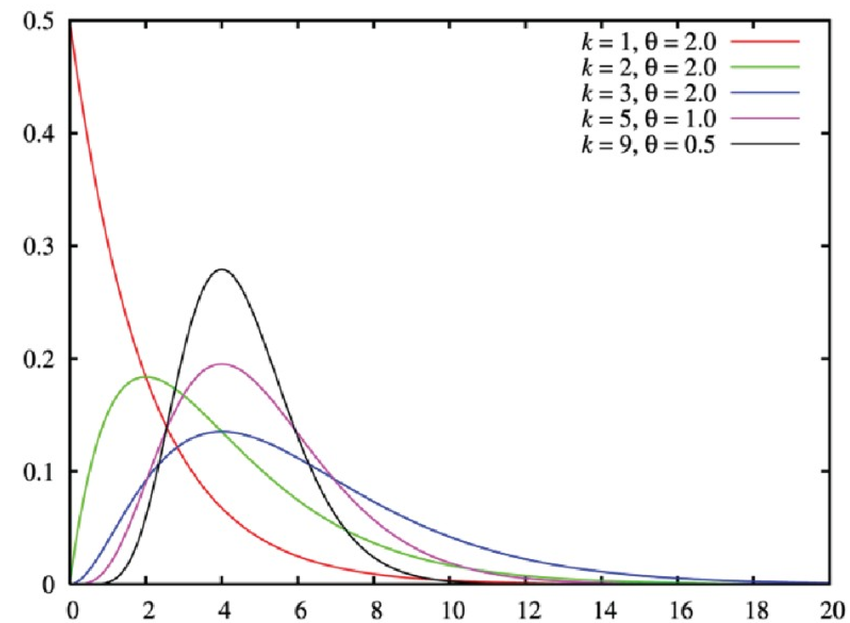
/ Variable discreta

- Poisson
- Binomial
Negativa

Funciones de distribución continuas

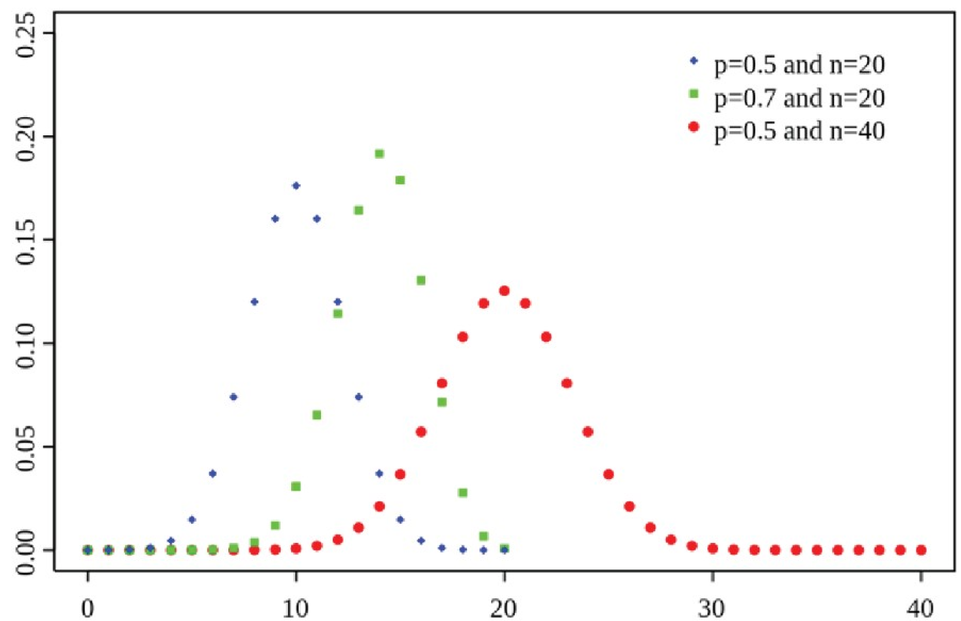


Normal

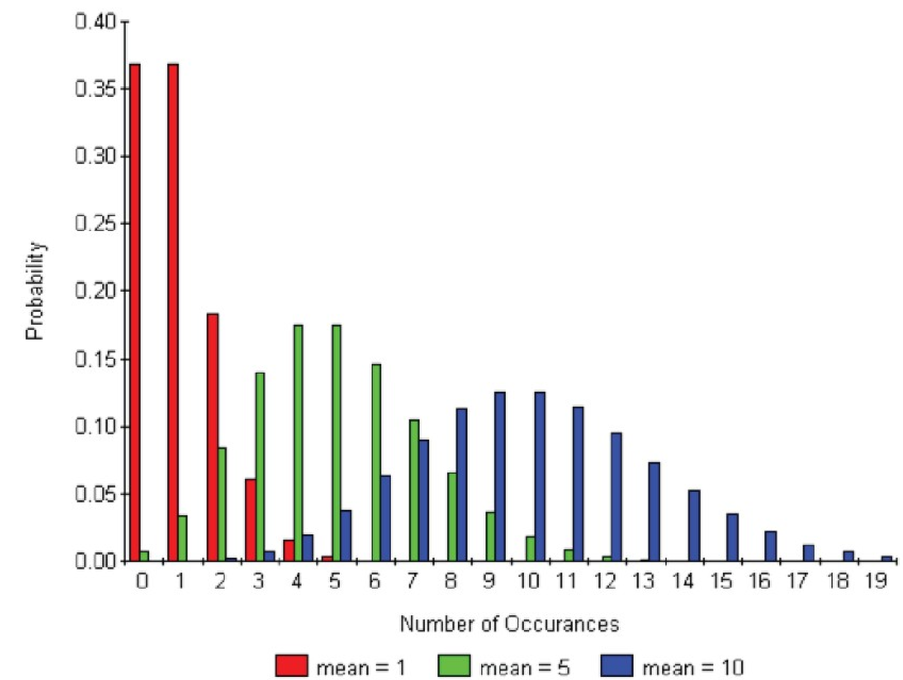


Gamma

Funciones de distribución discreta



Binomial negativa



Poisson

Parámetros estadísticos descriptivos

1

Posición

¿Qué valores toma nuestra variable?

2

Dispersión

¿Qué tan juntos o dispersos están los valores que toma la variable?

3

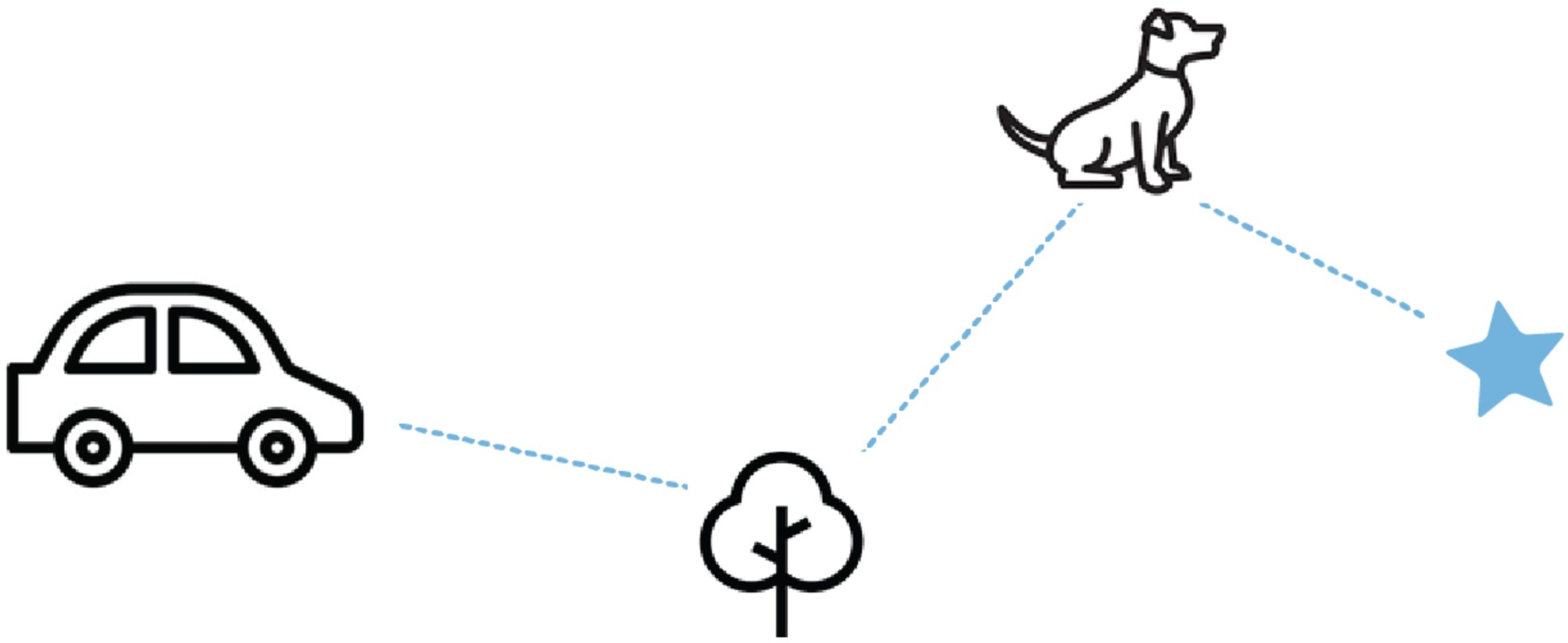
Forma

¿Cómo se distribuyen esos valores entre los sujetos?

1

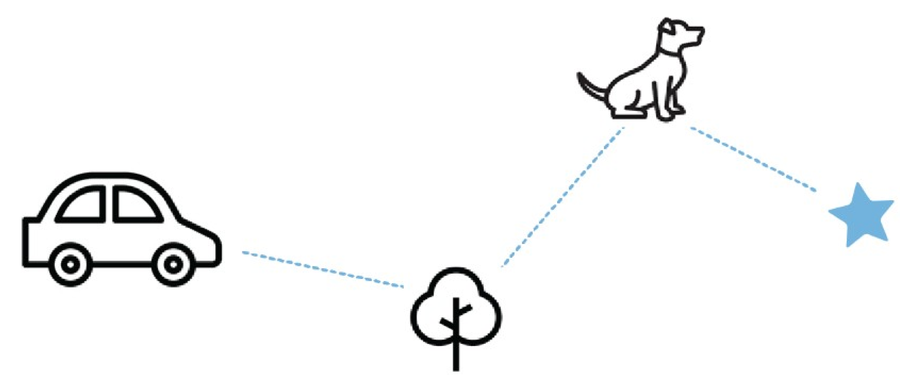
Parámetros de posición

¿Qué valores toma nuestra variable?



(0, 0)
(20, 0)

Parámetros de posición



A

/ De posición relativa

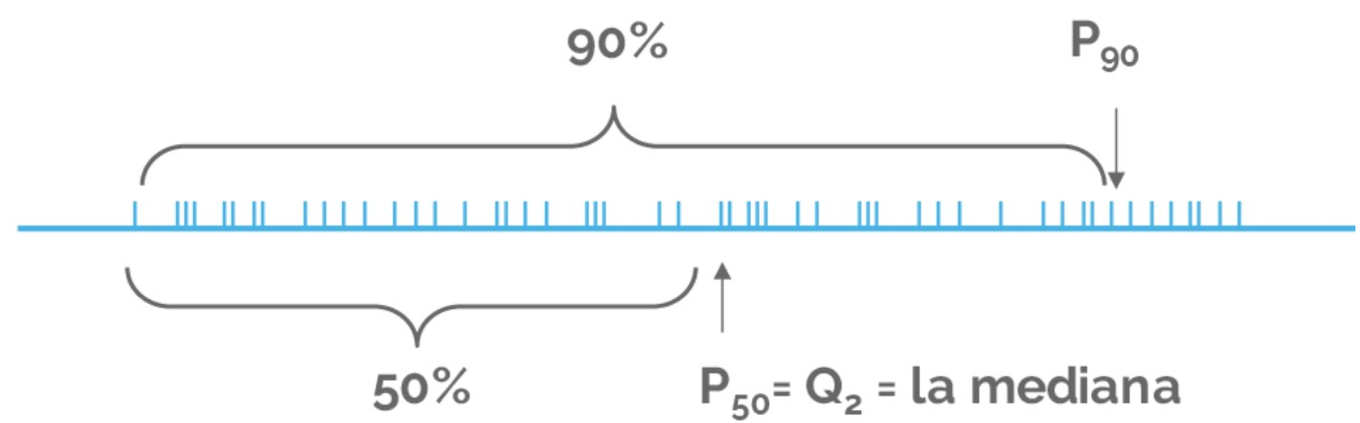
Su construcción hace referencia al total de los datos muestrales.

B

/ De tendencia central

Nos dan una idea en torno a qué valor se encuentran concentrados nuestros datos.

A Parámetros de posición relativa



/ Percentiles

Ordenados los datos de menor a mayor, el percentil p representa el valor de la variable por debajo del cual se encuentra el p% de las observaciones.

/ Cuartiles

Son los percentiles más importantes.

B Parámetros de posición de tendencia central

mE

Resumen nuestro conjunto de datos en un solo valor que responde a dónde están concentrados nuestros datos.

/ **Ej: Media o promedio, mediana, moda.**

1

La media o promedio

Es el valor que se obtiene al dividir la suma de un conjunto de números por su cantidad.

2

La mediana

Es el valor que se encuentra a la mitad de los otros valores. Es decir, al ordenar los número de menor a mayor, éste se encuentra justamente en el medio.

3

La moda

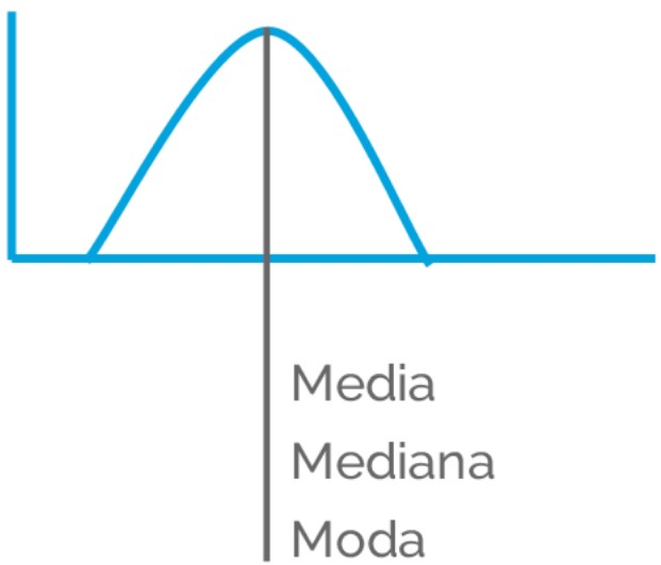
Es el valor que más se repite.

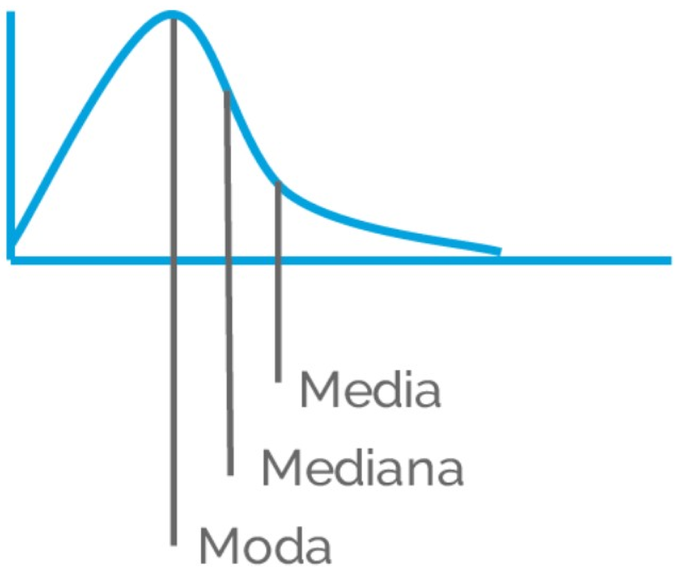
/ **Ejemplo:**
Tenemos los siguientes valores:
[1, 2, 3, 3, 3]
¿Cuánto vale la media, la mediana y la moda?

B Parámetros de posición de tendencia central

/ ¿Cuál elegir?

¿Distribución simétrica o sesgada?





2

Parámetros de dispersión

¿Qué tan juntos o dispersos están los valores que toma la variable?

Parámetros de dispersión

- Rango = máximo - mínimo
- Varianza (s2)
- Desviación estándar o típica (s)

Parámetros de dispersión

sumatoria

la media

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

la cantidad de datos

/ Ejemplo:
Tenemos los siguientes valores:
[1, 2, 3, 3, 3]

¿Cuánto vale la varianza?

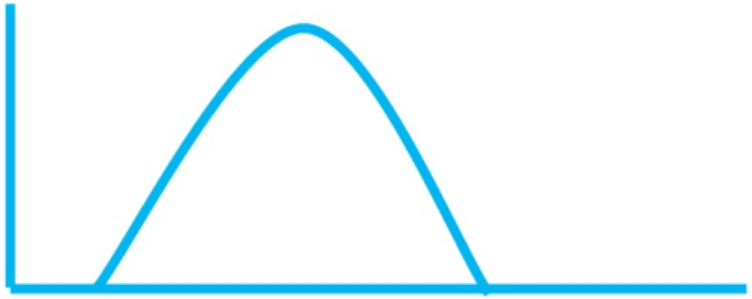
3

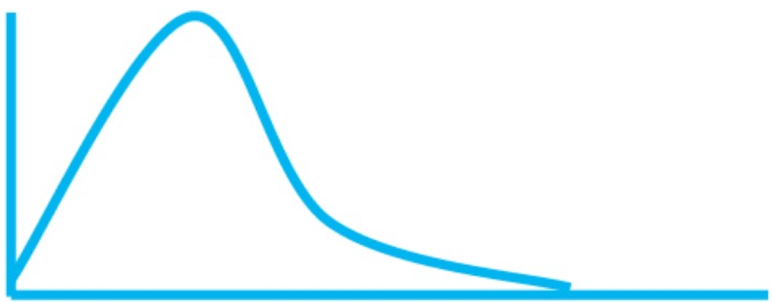
Parámetros de forma

¿Cómo se distribuyen esos valores entre los sujetos?

Parámetros de forma

- Coeficientes de asimetría
- Apuntamiento o curtosis





Resulta más sencillo referirse a su parecido con una función de distribución ¿es **uniforme**? ¿se parece a una **distribución normal**? ¿es **asimétrica**?

MUCHAS GRACIAS!

UNC

FCEFyN

mundosE
PEOPLE & BUSINESS SCHOOL