

# Clase 2

Análisis estadístico básico



# Análisis estadístico básico de los datos

Lo interesante de los datos no son los datos en sí sino la información que podemos extraer de ellos.

En lo primero que podemos pensar es en los momentos de una v.a.

- Esperanza
- Varianza
- Covarianza
- Otros

En general podemos hablar de los momentos de orden  $n$  calculados como  $\mathbb{E}[X^n]$ .

# Relación entre momentos y su estimación

## Esperanza

Recordemos que la esperanza representaba el valor medio o esperado de la v.a. y la calculamos como:

$$E[X] = \sum_{x \in A} x p_X(x) \text{ (si } X \text{ es v.a.d)} \text{ o } E[X] = \int_{\mathbb{R}} x f_X(x) dx \text{ ( si } X \text{ es v.a.c.)}.$$

La esperanza se corresponde con lo que se conoce como **momento de orden 1**

**¿Cómo la estimamos a partir de los datos?**

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ . La **media muestral** se calcula como:

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

# Relación entre momentos y su estimación

## Varianza

Recordemos que la varianza representaba la dispersión alrededor del valor medio y la calculamos como:

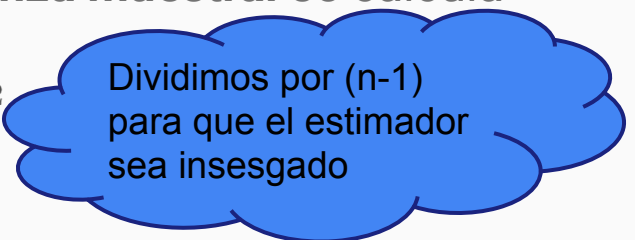
$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

La es un **momento de orden 2** porque depende de  $X^2$

**¿Cómo la estimamos a partir de los datos?**

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ . La **varianza muestral** se calcula como:

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



Dividimos por  $(n-1)$   
para que el estimador  
sea insesgado

# Relación entre momentos y su estimación

## Covarianza

La covarianza es un momento calculado entre dos v.a., y representa el grado de relación lineal entre las variables. La misma se calcula como

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

**¿Cómo la estimamos a partir de los datos?**

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ . La **varianza muestral** se calcula como:

$$\widehat{c_{X,Y}} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

# Otras medidas de tendencia central

## Mediana

La mediana se corresponde con el valor que acumula el 50% de la probabilidad y coincide con el cuantil 0.5

### ¿Cómo la estimamos a partir de los datos?

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ , y  $\underline{x}_{ord} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$  el vector ordenado de muestras.

- Si  $n\%2 = 0$ , la mediana se computa como  $med = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$
- Si  $n\%2 = 1$ , la mediana se computa como  $med = x_{(n//2+1)}$

# Otras medidas de tendencia central

## Media truncada (*Truncated or trimmed mean*)

Para el caso de variables con valores extremos, puede ser de utilidad analizar la media sobre la distribución truncada.

La media truncada a un  $p\%$  de  $X$ , es la media que se obtiene descartando el  $p\%$  superior y el  $p\%$  inferior de la distribución.

### ¿Cómo la estimamos a partir de los datos?

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ . La **media muestral truncada a un  $p\%$**  el promedio sobre las muestras luego de descartar el  $p\%$  superior y el  $p\%$  inferior.

# Otras medidas de tendencia central

## Moda

La moda se corresponde con el valor más probable de la distribución.  
(Recordar método de máxima verosimilitud)

Observación: Puede haber más de una moda (distribuciones multimodales)

### **¿Cómo la estimamos a partir de los datos?**

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ , la moda es el valor que más veces se repite.

Observación: Esta definición en general es válida para variables discretas, si la variable fuera continua no es tan directo el cálculo de moda ya que posiblemente no se observen valores repetidos.



# Otros momentos

## Oblicuidad (*Skewness*)

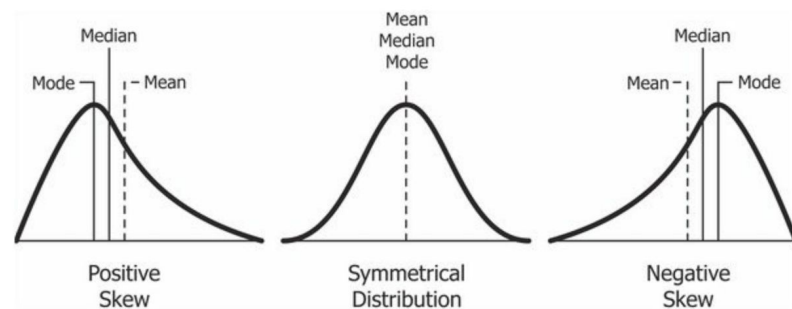
La oblicuidad es un momento de tercer orden estandarizado, y se utiliza como medida de asimetría respecto de la media

$$\tilde{\mu}_3 = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right]$$

**¿Cómo la estimamos a partir de los datos?**

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ , la oblicuidad se puede estimar como:

$$\tilde{\mu}_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$



# Otros momentos

## Curtosis

La curtosis es un momento de cuarto orden estandarizado, y se utiliza como medida para caracterizar las colas de la distribución. La medida estándar de curtosis es la propuesta por Pearson:

$$\tilde{\mu}_4 = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right]$$

A su vez, suele definirse la **curtosis por exceso** =  $\tilde{\mu}_4 - 3$ , que se encuentra referenciada a la curtosis de la dist. normal.

- Si  $\tilde{\mu}_4 - 3 > 0$  se trata de una distribución de colas más pesadas, por ejemplo de t-Student
- Si  $\tilde{\mu}_4 - 3 < 0$  se trata de una distribución con colas más delgadas, por ejemplo la distribución uniforme.

# Otros momentos

## Curtosis

¿Cómo la estimamos a partir de los datos?

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ , la curtosis se puede estimar como:

$$\tilde{\mu}_4 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$$

# Cuantiles

Los cuantiles son los puntos que dividen el rango de la función de distribución en segmentos de igual probabilidad. De esta forma, los  $q$ -cuantiles se definen como:

$$q_k = \min\{x : \mathbb{P}(X \leq x) \geq k/q\}, \quad k = 1, 2, \dots, q - 1$$

## ¿Cómo la estimamos a partir de los datos?

Sea  $\underline{x} = (x_1, \dots, x_n)$  una muestra de tamaño  $n$ , y  $\underline{x}_{ord} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$  el vector ordenado de muestras. Para estimar  $q_k$ , primero se debe calcular el índice  $h$  a partir de  $\frac{x_{ord}}{n}$ . Si el valor de  $h$  no fuera entero se puede interpolar entre  $\text{floor}(h)$  y  $\text{ceil}(h)$ .

Cómo se calcula  $h$  depende del software empleado. Python usa  $h=(n+1)p/q$

# Cuantiles

## Casos particulares

- Percentiles:  $q=100 \rightarrow q_{0.01}, q_{0.02}, \dots, q_{0.99}$
- Cuartiles:  $q=4 \rightarrow q_{0.25}, q_{0.5}, q_{0.75}$ . Se suele llamar primer, segundo y tercer cuartil (Q1, Q2, Q3) respectivamente.

Ejemplos:

- $\underline{x} = [1, 2, 3, 3, 4, 5, 5, 5, 6]$
- $\underline{x} = [1, 2, 3, 3, 4, 5, 5, 6]$

# Rango intercuartil (IQR)

El **rango intercuartil (IQR)** es la diferencia entre el tercer y el primer cuartil:

$$\text{IQR} = q_{0.75} - q_{0.25} = Q3 - Q1$$

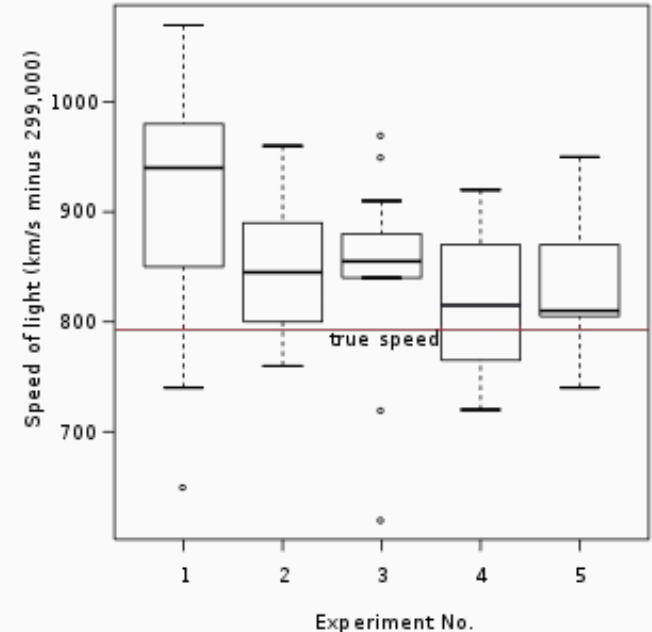
Es una medida de la dispersión de los datos y se corresponde con un estimador truncado, ya que descarta las colas superior e inferior de los datos.

Sin hacer mucho spoiler... sirve para hacer análisis de datos extremos.

# Box-Plot

El Box Plot o gráfico de cajas y bigotes es una forma sintética de representar los datos basándose en el IQR. Los pasos para graficar el Box plot son:

1. Identificar Q1, Q2, Q3
2. Graficar una caja entre Q1 y Q3
3. Graficar los “bigotes” desde Q1 hasta  $Q1 - 1.5 \times IQR$  y desde Q3 hasta  $Q3 + 1.5 \times IQR$



# BONUS - Repaso IC

[Api](#) para entender mejor los intervalos de confianza