ANALISIS DE DATOS

Pelli, Nahuel

ANÁLISIS ESTADÍSTICO BÁSICO

Lo que ya conocemos... pero con otro saborcito

ESTADÍSTICA BÁSICA

Un repaso y unas pruebitas

Análisis estadístico básico de los datos

Lo interesante de los datos no son los datos en sí sino la información que podamos 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:

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in A} x \cdot p_X(x) \\ \int_{\mathbb{D}} x \cdot f_X(x) dx \end{cases}$$

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

¿Cómo la estimamos a partir de los datos?

Consideremos $\bar{x} = (x_1, \ldots, x_n)$ una muestra tamaño *n*. La media muestral se calcula como:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_i x_i$$

Relación entre momentos y su estimación

Varianza

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

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

Esta se llama **momento de orden 2** porque depende de X^2

¿Cómo la estimamos a partir de los datos?

Consideremos $\bar{x} = (x_1, \ldots, x_n)$ una muestra tamaño *n*. La varianza muestral se calcula como:

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

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:

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

Este también se considera un momento de orden 2, pero es referente a dos VA

¿Cómo la estimamos a partir de los datos?

Consideremos dos vectores aleatorios X e Y de tamaño *n*. La **covarianza** muestral la podemos obtener como:

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

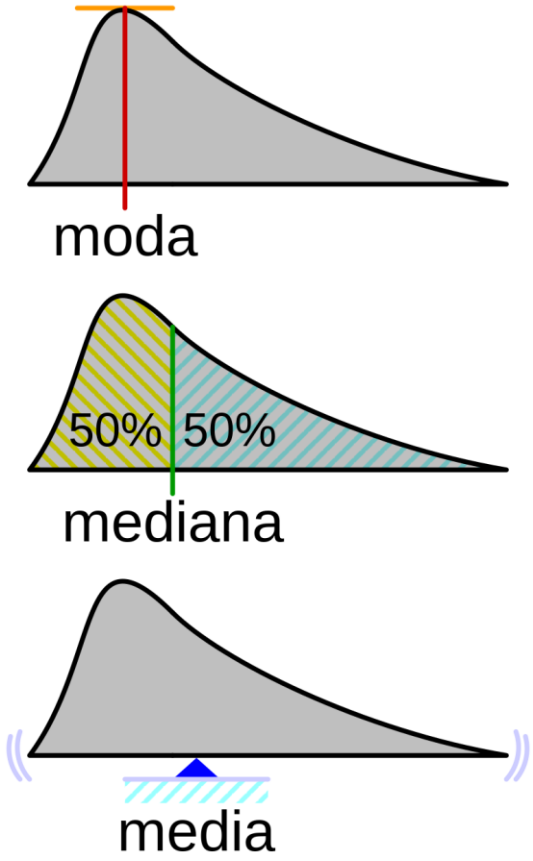
Otras medidas de tendencia central
Mediana

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

$$med(x) = \begin{cases} \frac{1}{2}(x_{(n/2)} + x_{(n/2)+1}) & si\ n\ es\ par \\ x_{(n//2+1)} & si\ n\ es\ impar \end{cases}$$



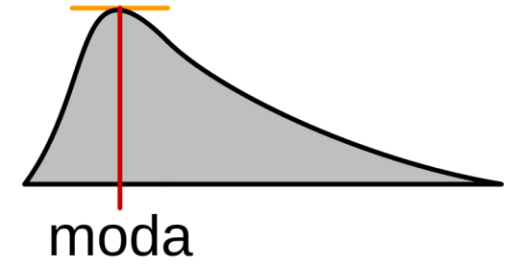
Otras medidas de tendencia central
Moda

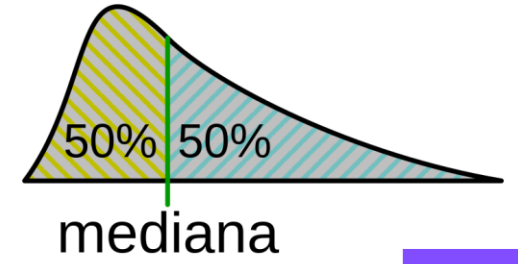
La moda se corresponde con el valor más probable de la distribución. (Recordar método de máxima verosimilitud)

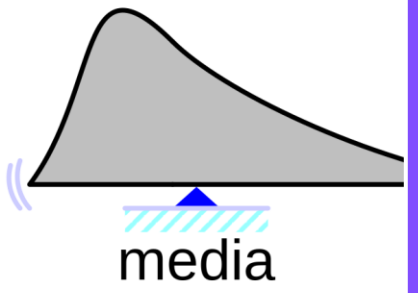
Observación: Puede haber más de una moda (distribuciones multimodales)

¿Cómo la estimamos a partir de los datos?

Sea $\bar{x} = (x_1, \dots, x_n)$ 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.

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?

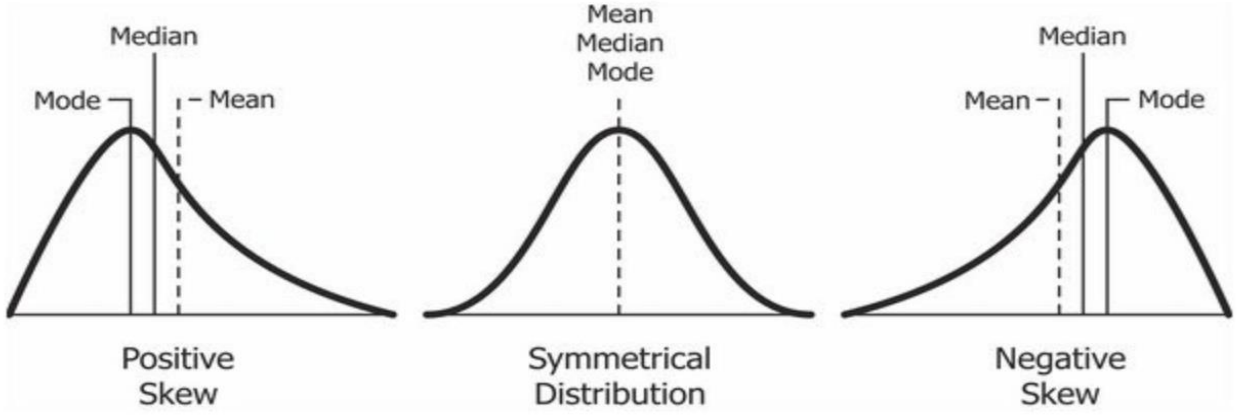
Consideremos $\bar{x} = (x_1, \ldots, x_n)$ una muestra tamaño *n*. La media muestral truncada a un p% el promedio sobre las muestras luego de descartar el p% superior y el p% inferior.

Momentos de orden mayor

Oblicuidad (*Skewness*)

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

$$\mu_3 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



¿Cómo la estimamos a partir de los datos?

Consideremos $\bar{x} = (x_1, \ldots, x_n)$ una muestra tamaño *n*. La varianza muestral se calcula como:

$$\widetilde{\mu_3} = \frac{n^{-1}\sum(x_i - \bar{x})^3}{s^3}$$

¿Qué ocurre con distribuciones simétricas?

Momentos de orden mayor

Curtósis

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:

$$\mu_4 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

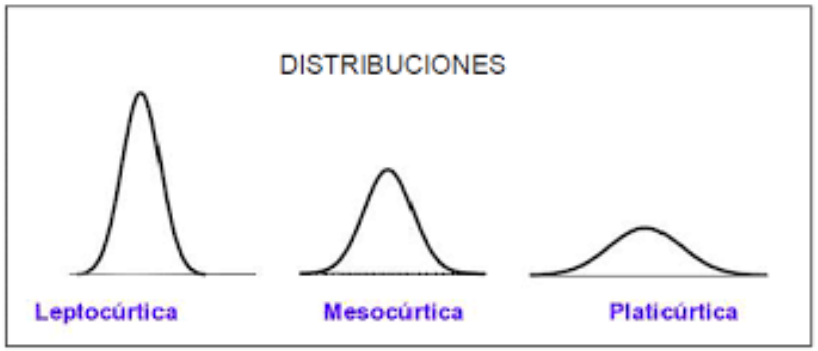
A su vez, suele definirse la **curtosis por exceso =** $\mu_4 - 3$, que se encuentra referenciada a la curtosis de la dist. normal.

¿Cómo la estimamos a partir de los datos?

Consideremos $\bar{x} = (x_1, \ldots, x_n)$ una muestra tamaño *n*.

$$\widetilde{\mu_4} = n^{-1}\frac{\sum(x_i - \bar{x})^4}{s^4}$$

¿Cómo se interpreta este estadístico?



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:

$$q_k = \min\{x: \mathbb{P}(X \leq x) \geq k/q\}$$

Donde $k = 1,2, \dots, q-1$

¿Cómo la estimamos a partir de los datos?

Sea $\bar{x} = (x_1, \dots, x_n)$ una muestra de tamaño n, y $\overline{x_{ord}} = (x_{(1)}, \dots, x_{(n)})$ el vector ordenado de muestras. Para estimar q_k, primero se debe calcular el índice h a partir de x_{ord}. Si h no fuese entero se trunca usando floor o ceil

En Python → $h = (n+1)p/q$

Cuantiles
Casos particulares

- Percentiles: q=100 $\rightarrow$ $q_{0.01}$,$q_{0.02}$,... , $q_{0.99}$
- 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:

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

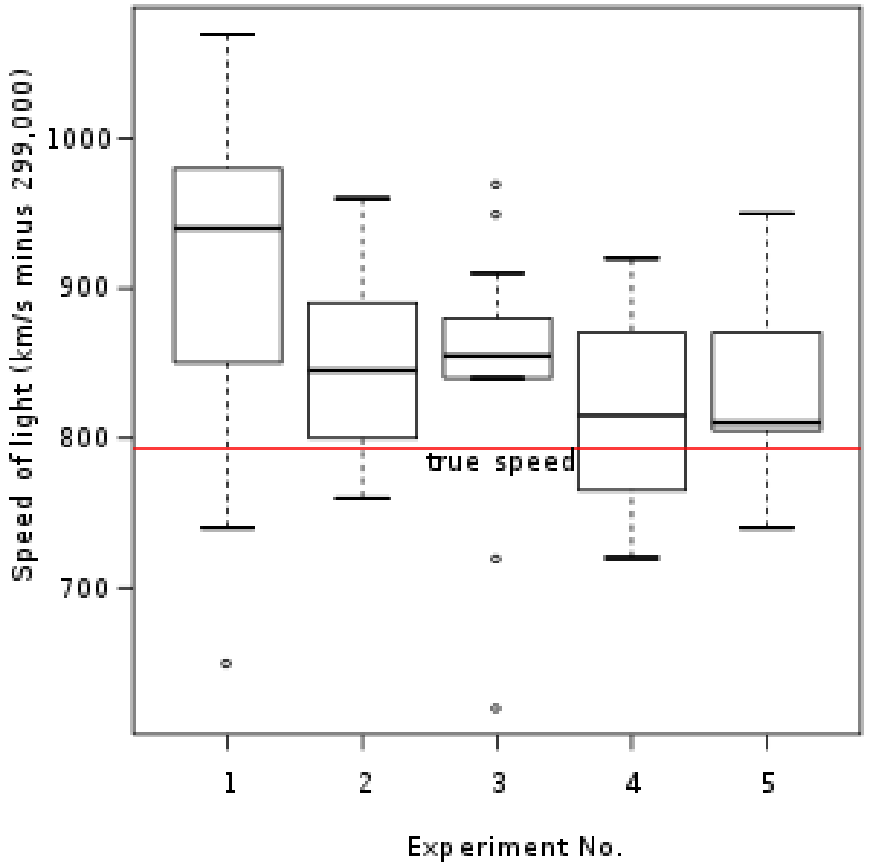
Es una medida de los dispersión de los datos y se corresponde con un estimador truncado, ya que descarta las colas superior e inferior de los datos.

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*IQR y desde Q3 hasta Q3+1.5*IQR



QQ-plot

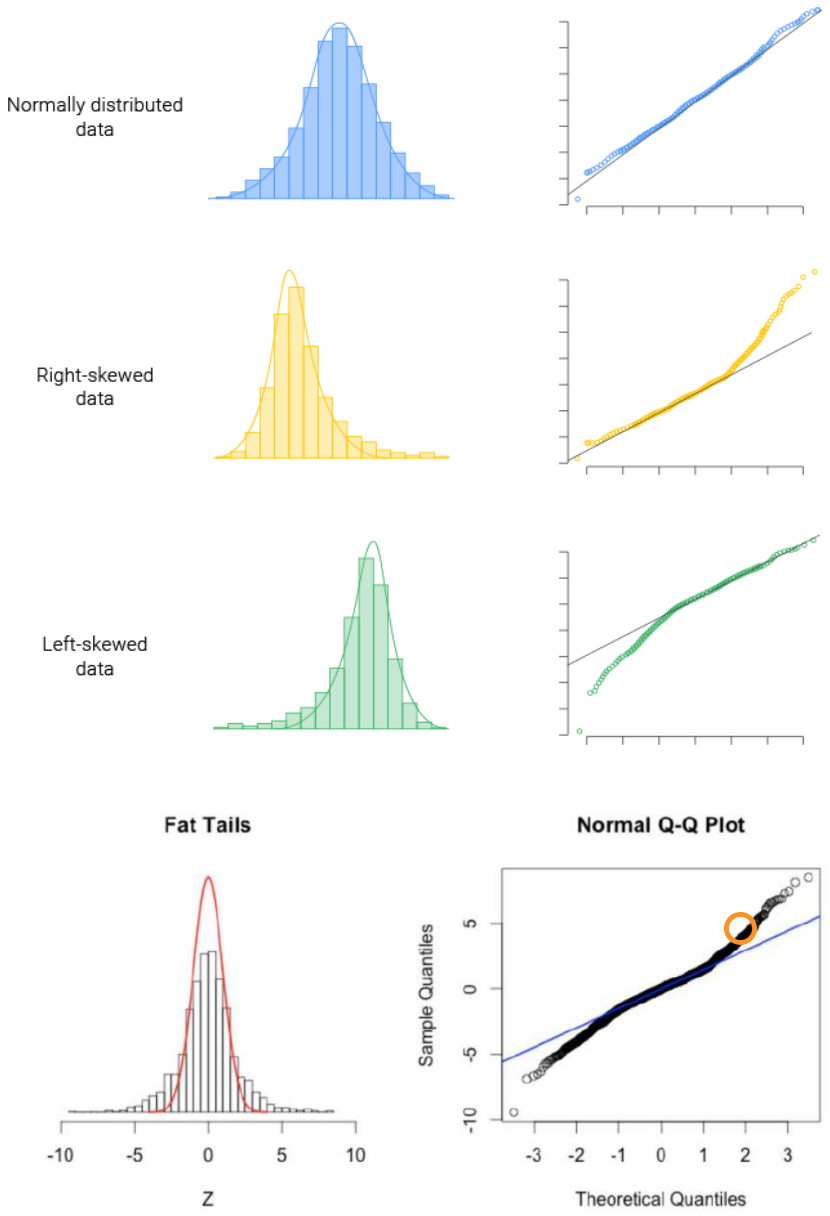
El grafico Q-Q plot es una manera de representar gráficamente los momentos de orden mayor que vimos.

Este nos sirve también para comparar dos distribuciones respecto a sus cuantiles.

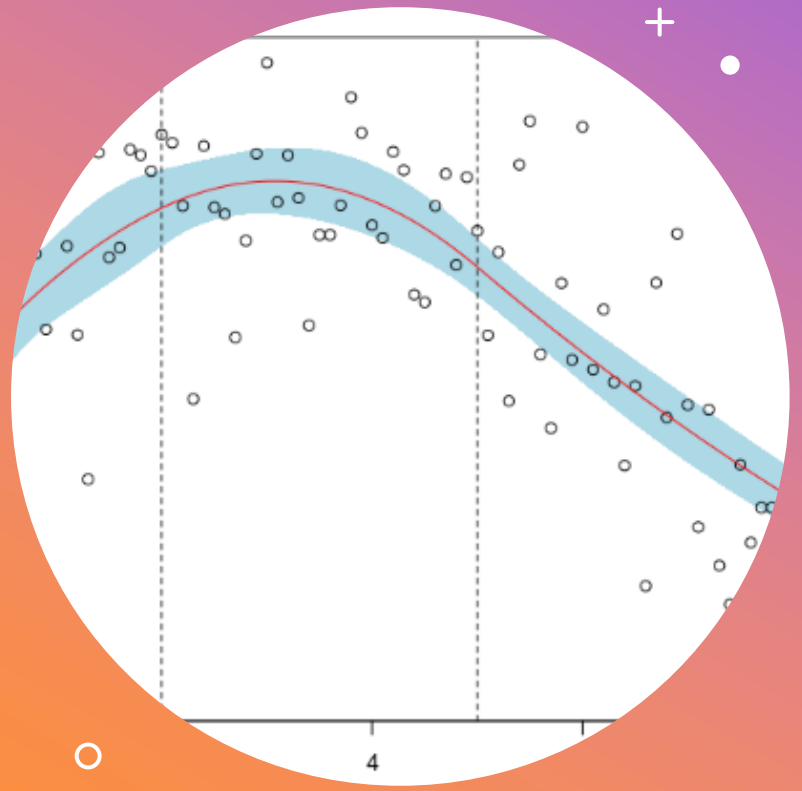
Obs: si lo comparamos contra una distribución normal podemos valernos de este gráfico para aseverar el cumplimiento de la normalidad en nuestro Sistema.

Construccion:

- Construir el vector de percentiles q_k a analizar (por ejemplo $v_q = (q_{0.1}, q_{0.2}, \ \dots\ , q_{0.9}, q_{1.0})$
- Calculamos los percentiles v_q para las VA's X, Y
- Realizamos un gráfico de puntos para v_{q_X} vs v_{q_Y}
- Trazamos una recta X=Y para poder comparar



BONUS - REPASO IC



Api para entender mejor los intervalos de confianza

DUDAS?

ENCUESTA