Essential Machine Learning with Python

Luis Chacón

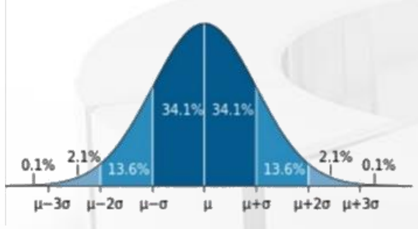
https://www.facebook.com/qslearningperu/?ref=page_internal

QS Learning
+51 937 012 707 / +51 915 111 457
quants.admission@gmail.com

@by_LChacon

Conceptos básicos

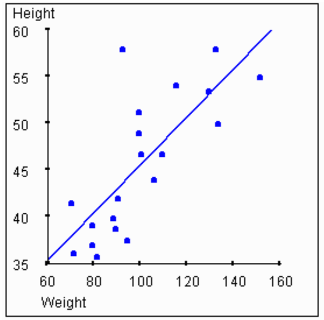
Descriptivos



Teorema Central de Límite

Muestreo

Contraste de hipótesis



Correlación

Medidas básicas de la estadística descriptiva

$$X = \{x_1, x_2, \cdots, x_n\}$$
$$|X| = n$$

Medidas de Centralización

media aritmética $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

mediana $P(X \leq m) = 0.5$

moda $p(X = M) \geq p(x = x_i)$
$\forall 1 \leq i \leq n$

percentiles $P(X \leq x_p) = p$
$p \in [0, 1]$

Medidas de Dispersión

varianza $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$

desviación típica $s = +\sqrt{s^2}$

coeficiente de variación $C_V = \frac{s}{\overline{x}} \cdot 100$

Momento de orden r respecto de la media

$$m_r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^r}{n}$$

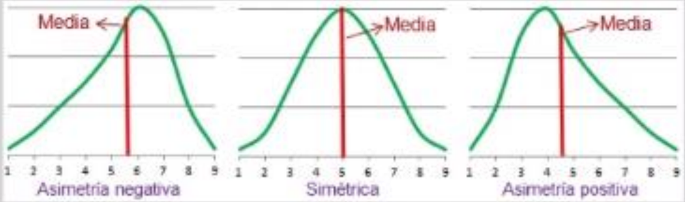
Medidas básicas de la estadística descriptiva

$$X = \{x_1, x_2, \cdots, x_n\}$$

$$|X| = n$$

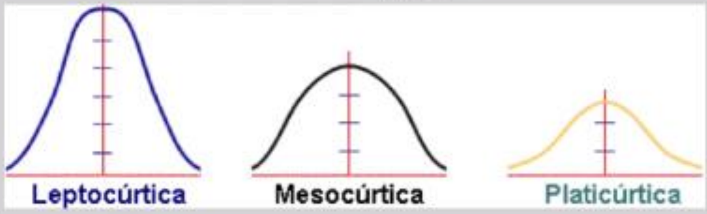
Medidas de Asimetría

asimetría de Fisher $CA_F = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n \cdot s^3}$



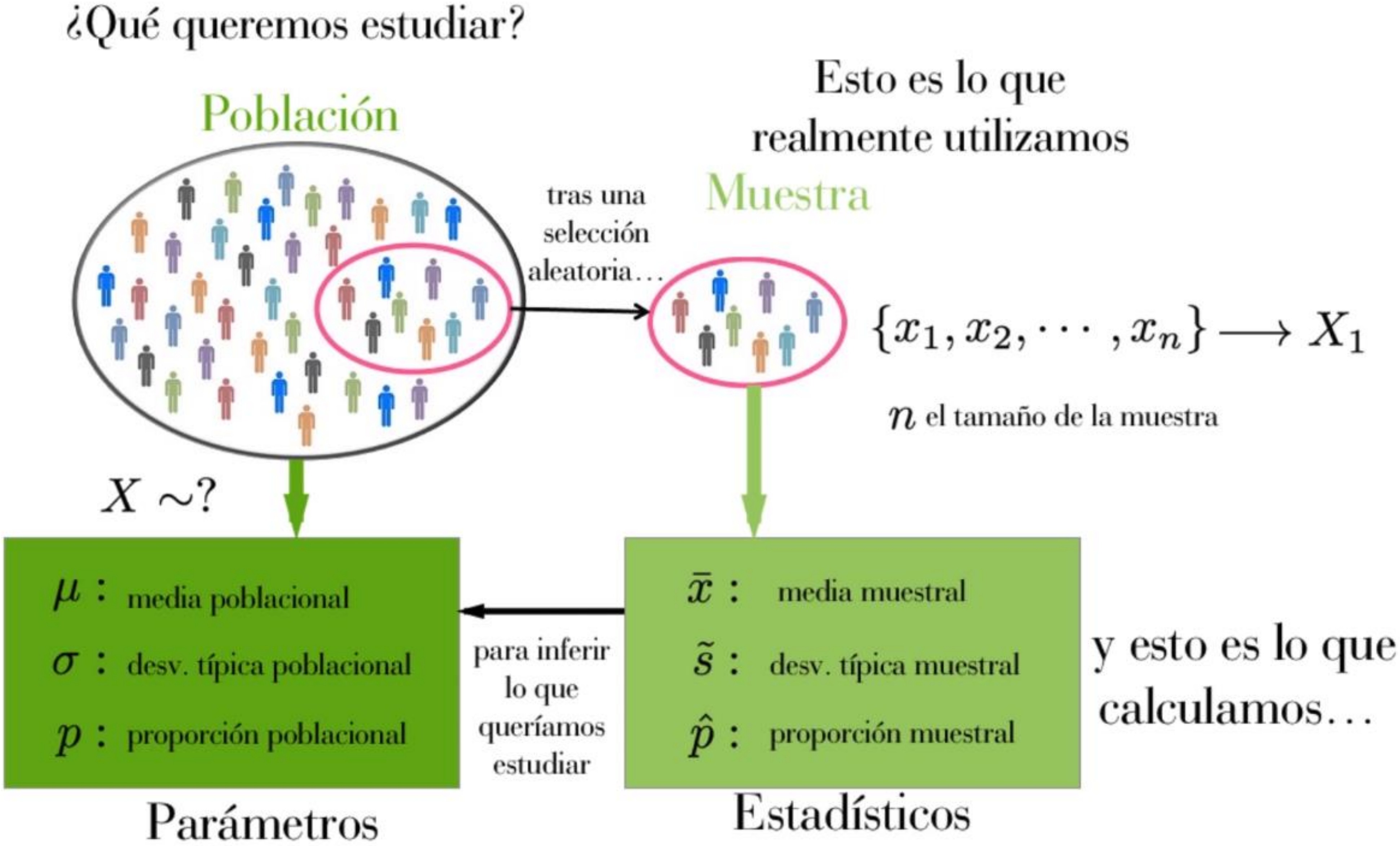
- 0 +

curtosis $c = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n \cdot s^4} - 3$



+ 0 -

Muestreo aleatorio



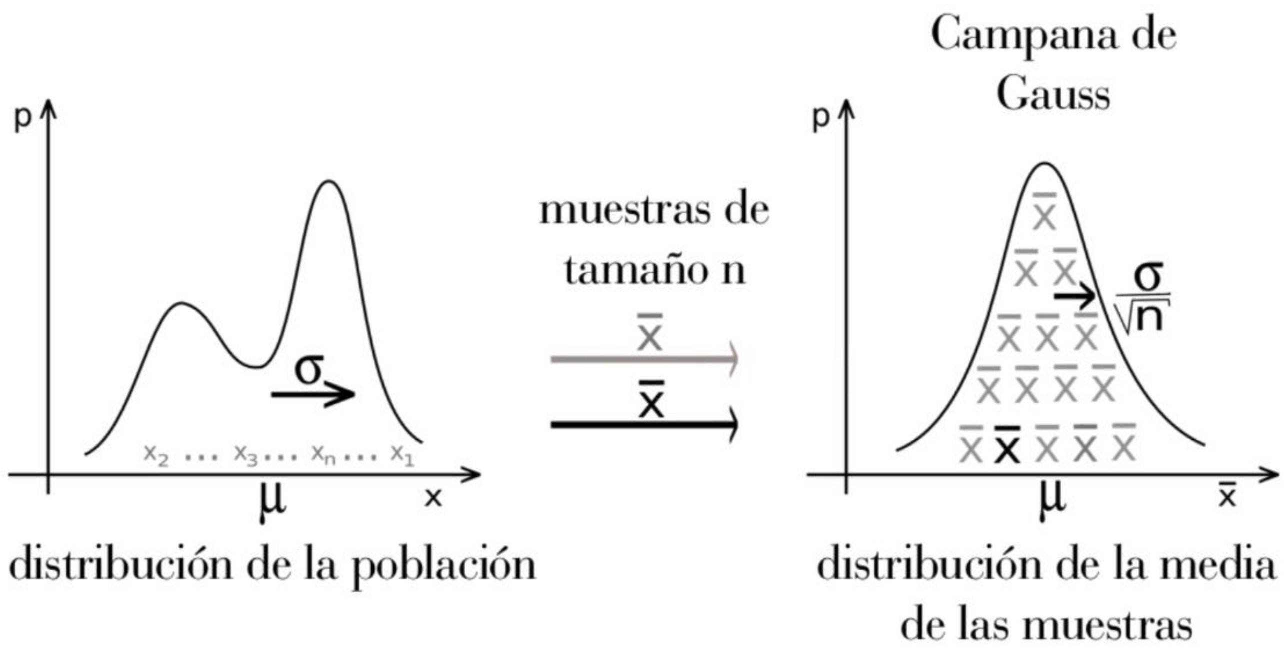
Teorema central de límite

Enunciado

- Si x_1, x_2, … x_n es una muestra aleatoria de tamaño **n** tomada de una población con media **μ** y varianza $\boldsymbol{\sigma^2}$, entonces el límite de la distribución

$$z_n = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

- cuando $n \to \infty$, es la distribución normal.



Estimamos

Contraste de hipótesis

El contraste es una afirmación respecto a alguna característica de una población.

Contrastar una hipótesis es comparar las predicciones con la realidad que observamos.

Si dentro del margen de error que nos permitimos admitir, hay coincidencia, aceptaremos la hipótesis y en caso contrario la rechazaremos

Contraste bilateral

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

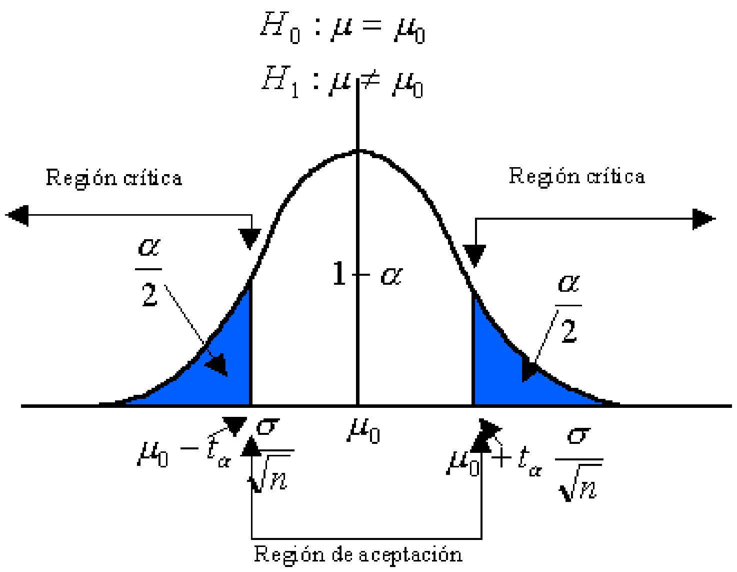
Contrastes unilaterales

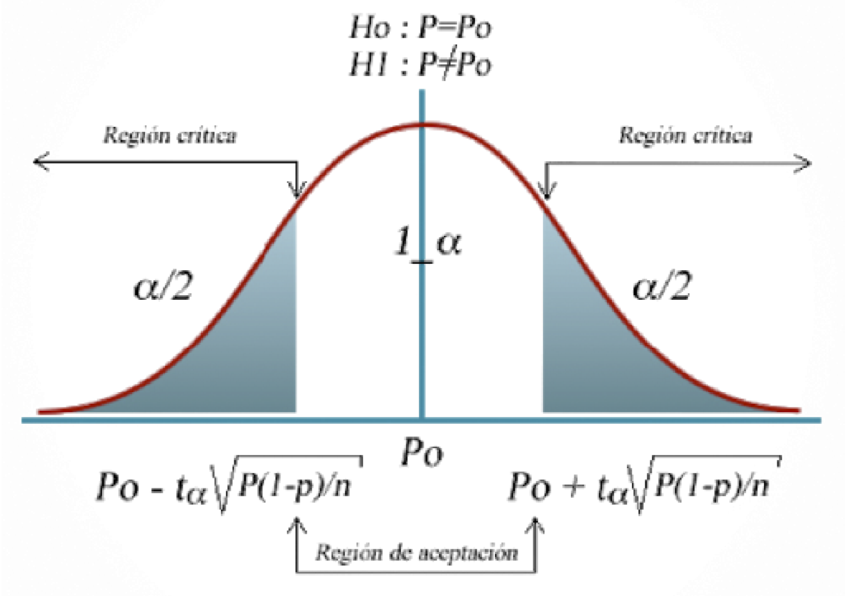
$$\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

$$\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$$

H_0: hipótesis nula
H_1: hipótesis alternativa

¿Qué distribución sigue?
¿Estadístico de Contraste?





Advanced Analytics @cademy

Nos preguntamos si es cierto que la población tiene una media

$$\mu = \mu_0$$

Podríamos usar el TCL

¿Pero que pasa con σ?

$$X \sim N(\mu, \sigma)$$

$\{x_1, x_2, \cdots, x_n\}$ m.a.s.

$$\mu_{\bar{X}} \longrightarrow \mu$$

$$\sigma_{\bar{X}} \longrightarrow \frac{\sigma}{\sqrt{n}}$$

<u>Caso 1:</u> σ conocida $X \sim N(\mu_0, \sigma)$

Podemos aplicar el TCL directamente

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

z-test

<u>Caso 2:</u> σ desconocida $X \sim N(\mu_0, ?)$

Estimamos primero la desviación típica

$$S = \frac{\sum (X_i - \mu_{\bar{X}})^2}{n-1} \rightarrow \sigma$$

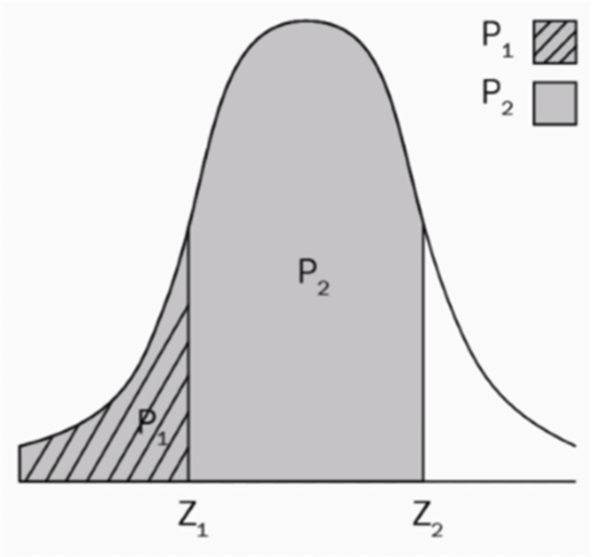
y los datos se distribuyen según la distribución t

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

t-test

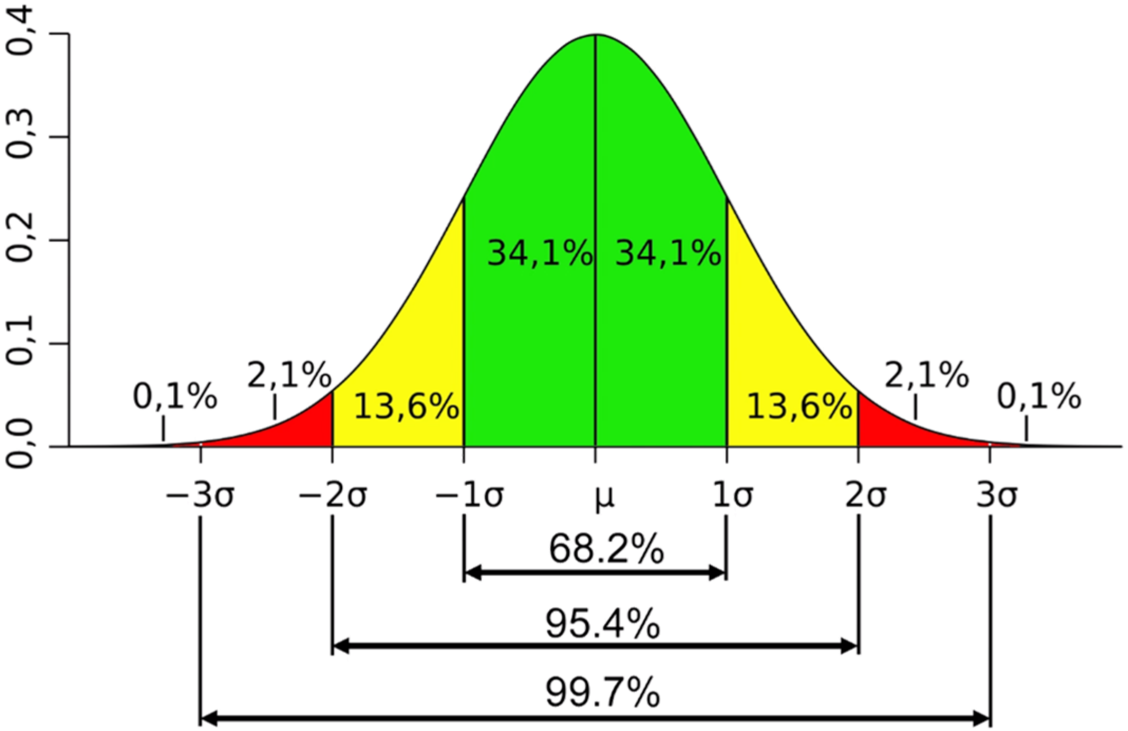
Contraste de hipótesis

Si H_0 es cierta, hemos modelado X como una distribución normal o t de Student



$$P(X < Z_1) = p_1$$
$$P(X < Z_2) = p_2$$

$$P(X > Z_1) = 1 - p_1$$
$$P(X > Z_2) = 1 - p_2$$



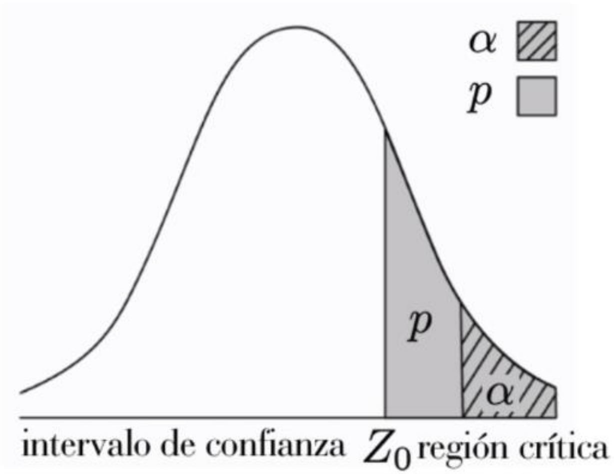
- **El área bajo cualquier distribución siempre es 1**
- **Podemos realizar estimaciones en intervalos**

- **Niveles de confianza**
- **Probabilidad con la que estamos seguros que el valor de la VA va a caer en el intervalo de confianza**

Advanced Analytics @cademy

Contraste de hipótesis

El p-valor y el nivel de significación



Z_0 el estadístico del contraste

$p - valor = P(X > Z_0)$

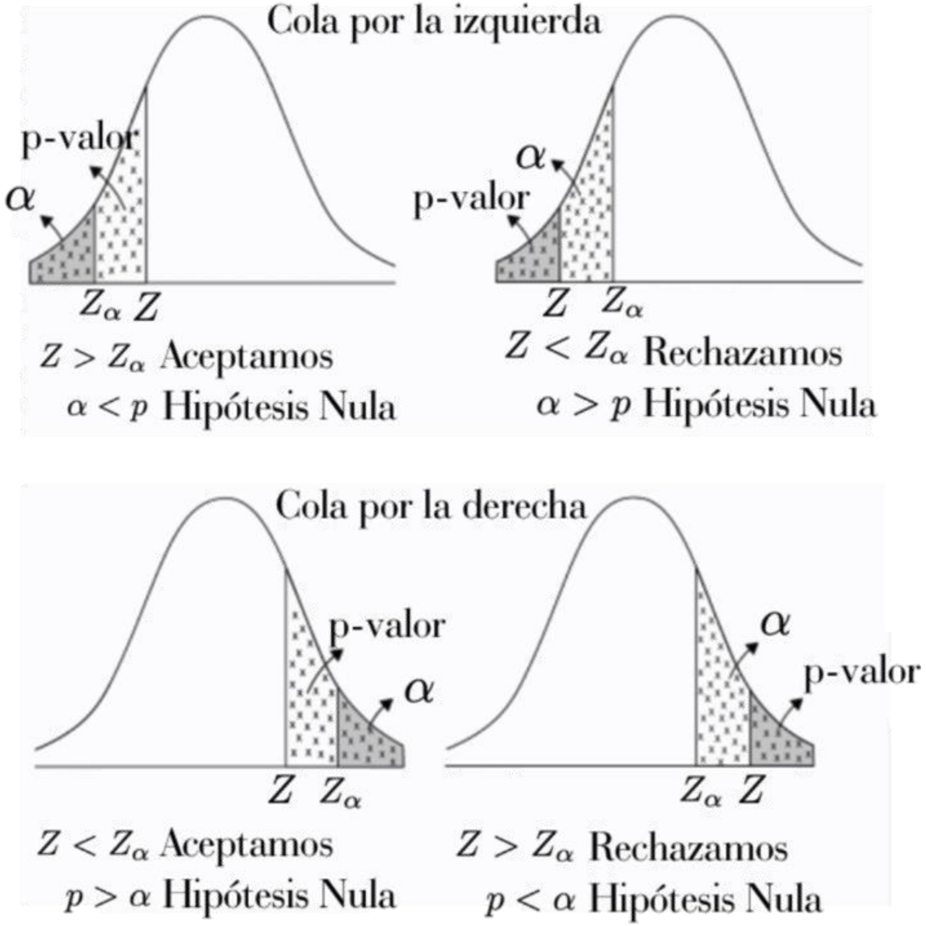
α el nivel de significación

$p - valor > \alpha \Rightarrow$

Mi estudio me da razones para aceptar la hipótesis nula y rechazar la hipótesis alternativa

$p - valor < \alpha \Rightarrow$

Tenemos evidencias para poder rechazar la hipótesis nula y aceptar como válida la alternativa



Dos formas de concluir si aceptamos o bien rechazamos la hipótesis nula
1. Comparando estadísticos
2. Comparando el p-valor con el nivel de significación

Contraste de hipótesis

Procedimiento

1. Definir hipótesis nula (μ_0) y alternativa uni o bilateral
2. Tomar una muestra aleatoria de tamaño n y calcular el valor del estimador (promedio, proporción…)
3. Calcular el estadístico de contraste Z-valor o t-valor,
4. Calcular el p-valor asociado,
5. Comparar p-valor y nivel de significación y decidir.

El ejemplo de Just Eat

El pizzero de Just-Eat afirma que trae la comida en un tiempo promedio inferior a 20 minutos con una desviación típica de 3.

$$\begin{cases} H_0 : \mu \leq 20 \\ H_1 : \mu > 20 \end{cases}$$

$$\sigma = 3$$

Como sospechamos que es falso, tomamos 64 de las entregas de la última semana y obtenemos una media de 21.2 minutos.

$$\bar{X} = 21.4, n = 64$$

¿Podemos aceptar su afirmación a un nivel de confianza del 95%? $\alpha = 0.05$

$$Z = \frac{21.2 - 20}{\frac{3}{\sqrt{64}}} = 3.2$$

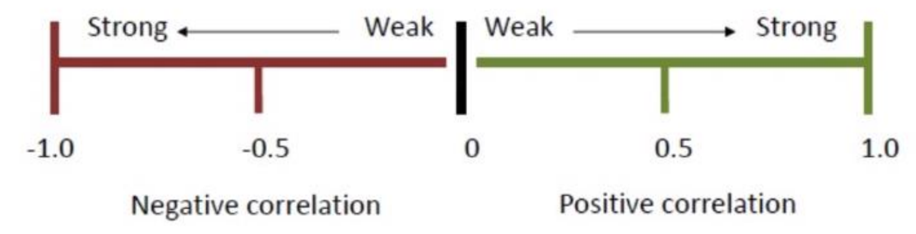
$$p = P(Z > 3.2) = 1 - P(Z < 3.2) = 1 - 0.999 = 0.001 < 0.05 = \alpha$$

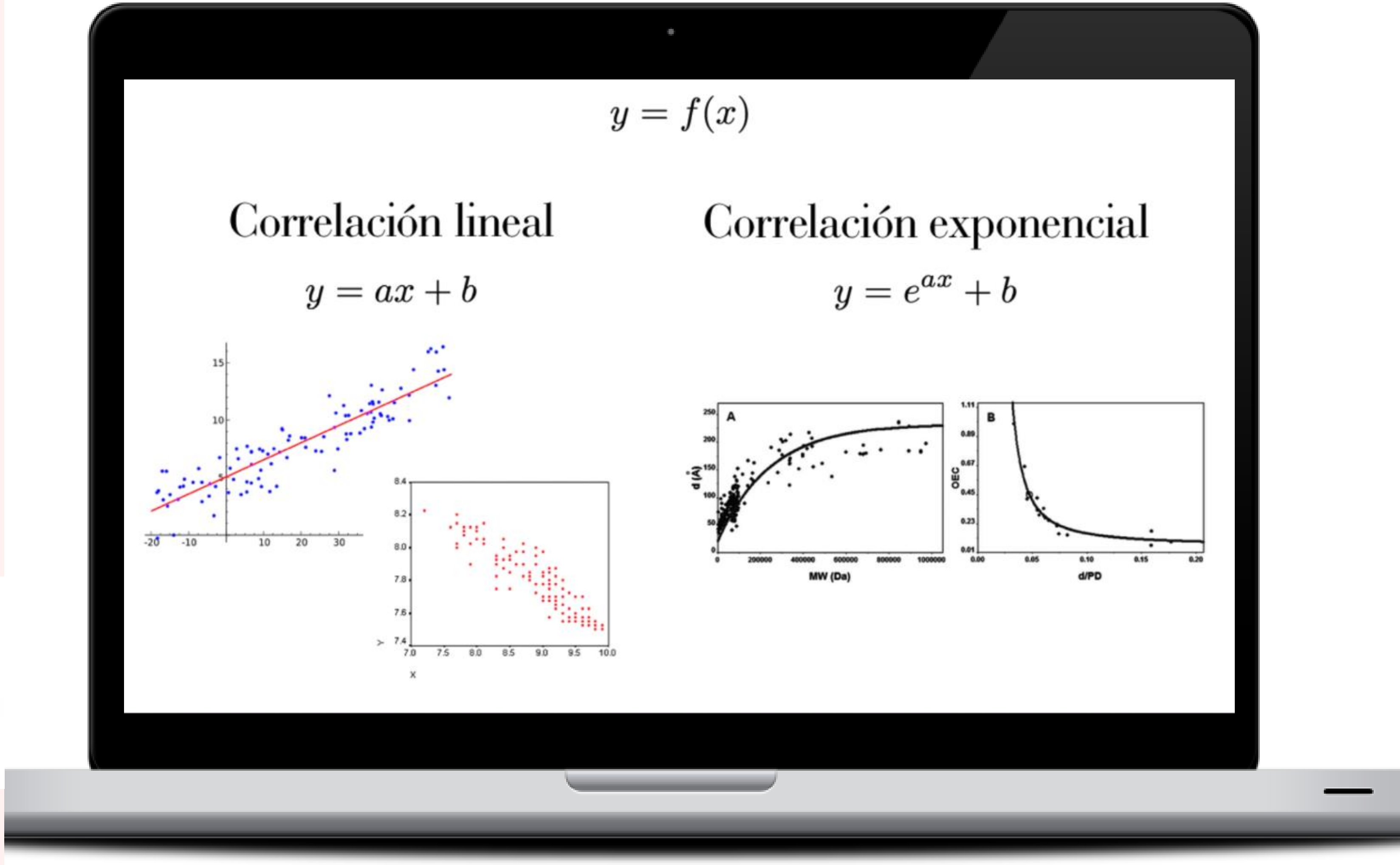
Correlación

La **correlación** alude a la proporcionalidad y la relación lineal que existe entre distintas variables.

Coeficiente de Correlación de Pearson

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$





Gracias!!!

Los esperamos en el siguiente módulo

QS Learning

+51 937 012 707 / +51 915 111 457

quants.admission@gmail.com