

# TEST DE HIPÓTESIS E INTERVALO DE CONFIANZA. DOS MUESTRAS

IRONHACK

# Intro

En esta sección continuaremos explorando los contrastes de hipótesis y aprenderemos a realizar pruebas con dos muestras de datos.

Es común que los investigadores realicen estudios para comparar dos grupos y comprobar si se comportan de manera diferente. Un ejemplo es aplicar un tratamiento a un grupo mientras que otro se deja sin tratar. Luego se comparan los resultados para ver si los dos grupos difieren.



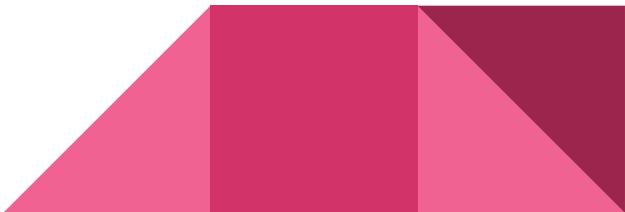
# Intro

Los contrastes de hipótesis son herramientas imprescindibles en estos problemas, en los que necesitamos utilizar la estadística para fundamentar la toma de decisiones. En esta lección aprenderemos sobre las diferentes maneras en que podemos comparar dos muestras para ver si difieren significativamente.



# Intro

Los pasos a seguir para el testeo de hipótesis de dos muestras es el mismo que con una muestra. Recordemos:

1. Identificar parámetro
  2. Elegir nivel de significación alfa
  3. Definir hipótesis en función del objetivo del test
  4. Obtener el valor del estadístico del test (z o t)
  5. Definir la región de rechazo con los valores críticos o calcular el pvalue
  6. Utilizar el pvalue del test o la región de rechazo para obtener una conclusión del test
- 

# Intro

En este caso, vamos a encontrarnos con muchos tipos de test dependiendo de:

- **El parámetro a contrastar** → diferencia entre medias poblacionales o proporciones poblacionales
- **El tamaño de las muestras** → si ambas son grandes, distribución z; de lo contrario, distribución t.
- **La relación entre las muestras** → dependientes o independientes



# Intro

- Diferencia entre **dos medias poblacionales**, muestras **independientes** con tamaño muestral **grande** en ambas.
- Diferencia entre **dos medias poblacionales**, muestras **independientes sin tamaño muestral grande** en ambas.
- Diferencia entre **dos medias poblacionales**, muestras **dependientes** con tamaño muestral **grande** en ambas.
- Diferencia entre **dos medias poblacionales**, muestras **dependientes sin tamaño muestral grande** en ambas.
- Diferencia entre **dos proporciones poblacionales**, muestras **independientes** con tamaño muestral **grande** en ambas.

Parameter	Key Words or Phrases	Type of Data
$\mu_1 - \mu_2$	Mean difference; differences in averages	Quantitative
$p_1 - p_2$	Differences between proportions, percentages, fractions, or rates; compare proportions	Qualitative



# Comparación de dos medias poblacionales. Muestras dependientes



# Muestras dependientes

Cuando nos enfocamos en un contraste de hipótesis de dos muestras, primero debemos determinar qué tipo de datos tenemos. El primer tipo de contraste de hipótesis de 2 muestras que vamos a realizar es el de dos muestras relacionadas. Esto significa que los datos de los dos muestreos son dependientes.

Por ejemplo, en un ensayo clínico de medicamentos, podemos administrar un medicamento para la presión arterial a un grupo de personas y examinar su presión arterial antes y después del tratamiento. Luego trataremos el antes y el después como dos muestras y las compararemos.



# Ejemplo 1

Para hacer inferencias a partir de dos muestras, no siempre tienen que ser independientes. Podemos estudiar los valores de la variable cuantitativa en la misma muestra pero en dos momentos diferentes o en muestras diferentes que estén emparejadas (no independientes).

Tenemos primero una muestra de trabajadores aprendiendo Excel y luego la misma muestra aprendiendo Word y tratamos de comparar el tiempo que tardan en aprender cada plataforma.



Block	WORD	EXCEL	d
Sam	Time to learn word	Time to learn exc	Time Word - Time exc
Mike	Time to learn word	Time to learn exc	Time Word - Time exc
Elena	Time to learn word	Time to learn exc	Time Word - Time exc
Jesús	Time to learn word	Time to learn exc	Time Word - Time exc
...	...	...	...

## Ejemplo 2

También puede darse el caso de que queramos utilizar un bloque para hacer la comparación más justa. Imagina que quieres testear si sacan mejores notas hombres o mujeres.

Imaginemos que sacamos una muestra en la que la mayoría de los hombres han estudiado periodismo, mientras que la mayoría de las mujeres de la muestra han estudiado física. ¿Es esto justo? No, la titulación sería otra fuente importante de variabilidad que habría que tener en cuenta. Por lo tanto, sólo se podrían comparar pares muy similares.

De este modo, si disponemos de esta información extra (carrera que ha estudiado cada uno), podemos utilizar esta metodología para obtener una prueba más rigurosa.



Block	Male	Female	d
Physics	Grademale1	Gradefemale1	Grademale1 - Gradefemale 1
Journalism	Grademale2	Gradefemale2	Grademale2 - Gradefemale2
Law	Grademale3	Gradefemale3	Grademale3 - Gradefemale3
Biology	Grademale4	Gradefemale4	Grademale4 - Gradefemale4
...	...	...	...

# Definir hipótesis

Dado que podemos comparar los datos entre las muestras, tomamos la diferencia entre las dos muestras en cada fila y luego volvemos a realizar el contraste de hipótesis sobre una muestra. Nuestro contraste de hipótesis comprobará si la media de las diferencias es significativamente diferente de cero (también podríamos probar que la media es mayor o menor que cero utilizando un contraste de hipótesis unilateral). Para un contraste de hipótesis bilateral, las hipótesis nula y alternativa son las siguientes:

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$



## **Large Sample**

Test statistic: 
$$z = \frac{\bar{d} - D_0}{\sigma_d / \sqrt{n_d}} \approx \frac{\bar{d} - D_0}{s_d / \sqrt{n_d}}$$

$d$  es la  
muestra/columna de  
diferencias

$D_0$  es el valor de la  
diferencia fijado en la  
hipótesis nula

## **Small Sample**

Test statistic: 
$$t = \frac{\bar{d} - D_0}{s_d / \sqrt{n_d}}$$



# Test

De esta forma, ya que solo tendríamos un grupo (el grupo o columna de las diferencias), simplemente seguiríamos las normas que vimos en test de hipótesis para una muestra en caso de que queramos sacar la columna de diferencias o hacerlo manualmente.

En Python usaremos la función de scipy llamada ttest\_rel en caso de que las muestras no sean grandes.

Y, si son grandes ambas, podemos utilizar la función de ztest dando solo el valor de una muestra que será la diferencia entre las dos que tenemos ( $x1 = \text{sample1} - \text{sample2}$ )





# Test

En nuestro ejemplo veremos un estudio de presión arterial con 100 participantes. A todos nuestros participantes se les midió la presión arterial antes del inicio del estudio y un mes después del inicio del mismo. Compararemos la presión arterial sistólica de los participantes antes y después.

Utiliza el dataset de blood pressure para hacer este test y explica tu conclusión.



# Comparación de dos medias poblacionales. Muestras independientes

# Muestras independientes

El segundo tipo de contraste de hipótesis es sobre dos muestras independientes. En este caso, tenemos dos grupos en los que no podemos emparejar las filas entre sí. Por ejemplo, comparamos el efecto de un determinado medicamento en una muestra de hombres y una muestra de mujeres. Luego realizamos un contraste de hipótesis para ver si hay una diferencia significativa en la forma en que la medicación afecta a los grupos.



# Muestras independientes

Otro ejemplo es un test A/B en un sitio web. Podemos implementar una serie de cambios en la interfaz de usuario de un sitio web de comercio electrónico.

Publicaremos la versión A para una muestra de clientes y la versión B para otra muestra. A continuación, comprobaremos si hay una diferencia en los ingresos entre las diferentes muestras.



# Muestras independientes

Al examinar dos muestras independientes, debemos comprobar que se cumplen algunos supuestos. El primer supuesto es la independencia. Un ejemplo de lo que podría causar una dependencia entre dos grupos es si tuviéramos un estudio sobre el impacto de la nutrición en la salud y tuviéramos un marido en un grupo y una mujer en el otro. Aunque no son la misma persona, lo más probable es que vivan en el mismo hogar. Por lo tanto, hay algunas cosas que hacen que pueden ser similares como hábitos de sueño o hábitos de desplazamiento.

Como investigadores, cuando esto sucede, no podemos estar seguros de si la intervención en nuestro estudio fue la causa principal de la diferencia (o similitud) entre los sujetos.



# Muestras independientes

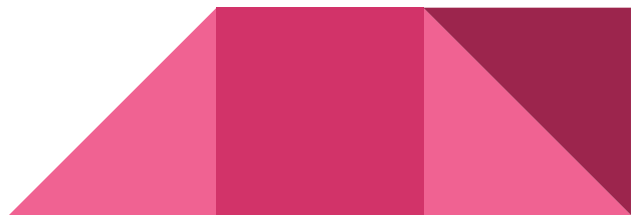
En un contraste de 2 muestras independientes, nuestro contraste de hipótesis (para un contraste bilateral) es una comparación de las medias de los dos grupos. También debemos asumir que las muestras fueron tomadas al azar de una población normalmente distribuida.

Hypothesis	Research Questions		
	No Difference Any Difference	Pop 1 $\geq$ Pop 2 Pop 1 < Pop 2	Pop 1 $\leq$ Pop 2 Pop 1 > Pop 2
$H_0$	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 \leq 0$
$H_a$	$\mu_1 - \mu_2 \neq 0$	$\mu_1 - \mu_2 < 0$	$\mu_1 - \mu_2 > 0$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sigma_{(\bar{x}_1 - \bar{x}_2)}}$$

$$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



# Test

Si ambas muestras son grandes ztest

Si no, ttest

En Scipy también hay funciones para hacerlo, pero estas me parecieron mejores.






# Test

Test A/B en un sitio web. Podemos implementar una serie de cambios en la interfaz de usuario de un sitio web de comercio electrónico. Publicaremos la versión A para una muestra de clientes y la versión B para otra muestra. A continuación, comprobaremos si hay una diferencia en los ingresos entre las diferentes muestras.

Utiliza los datos del dataset de ab test para hacer el test de hipótesis y explica tu conclusión.





# Comparación de dos proporciones poblacionales. Muestras independientes

# Trabajando con proporciones

Eres director de personal de una empresa y quieres comprobar la percepción que tienen tus empleados de dos métodos de evaluación del rendimiento. Por percepción entendemos si los empleados consideran que estos métodos son justos o no.

¿Consideran que un método es más justo que el otro?

**Una variable cualitativa, dos grupos**



# Trabajando con proporciones

Supongamos que un candidato presidencial quiere comparar las preferencias de los votantes registrados en el noreste de Estados Unidos con los del sureste. Esta comparación ayudaría a determinar dónde concentrar los esfuerzos de la campaña. El candidato contrata a un encuestador profesional para que elija al azar 1000 votantes registrados en el noreste y 1000 en el sureste y entreviste a cada uno de ellos para conocer su preferencia de voto.

El objetivo es utilizar esta información de la muestra para hacer una inferencia sobre la diferencia entre la proporción de todos los votantes registrados en el noreste y la proporción de todos los votantes registrados en el sureste que planean votar por el candidato presidencial.

**Una variable cualitativa, dos grupos**



# Trabajando con proporciones

La comparación de dos proporciones, al igual que la comparación de dos medias, es habitual. Si dos proporciones estimadas son diferentes, puede deberse a una diferencia en las poblaciones o al azar.

Una prueba de hipótesis puede ayudar a determinar si una diferencia en las proporciones estimadas refleja una diferencia en las proporciones de la población.

Solo trabajaremos con proporciones en el caso de muestras independientes y grandes ( $n_1\hat{p}_1 \geq 15$ ,  $n_1\hat{q}_1 \geq 15$  y  $n_2\hat{p}_2 \geq 15$ ,  $n_2\hat{q}_2 \geq 15$ )

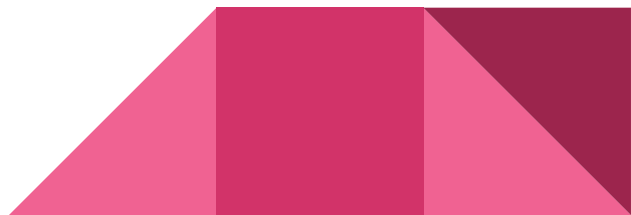


# Trabajando con proporciones

Hypothesis	Research Questions		
	No Difference Any Difference	Pop 1 $\geq$ Pop 2 Pop 1 $<$ Pop 2	Pop 1 $\leq$ Pop 2 Pop 1 $>$ Pop 2
$H_0$	$p_1 - p_2 = 0$	$p_1 - p_2 \geq 0$	$p_1 - p_2 \leq 0$
$H_a$	$p_1 - p_2 \neq 0$	$p_1 - p_2 < 0$	$p_1 - p_2 > 0$

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \approx \sqrt{\hat{p} \hat{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$



# Test

Para hacer un test de proporciones en python, puedes utilizar [esta función](#) de la librería statsmodels.





# Intervalos de confianza

# En Python

Función para estimar el valor de la diferencia entre dos proporciones poblacionales independientes.

Función para estimar el valor de la diferencia entre dos medias poblacionales independientes (muestras grandes).

Para muestras relacionadas, podemos sacar el intervalo de confianza de la columna de diferencias (una muestra).

Para muestras pequeñas independientes, sigue la fórmula.

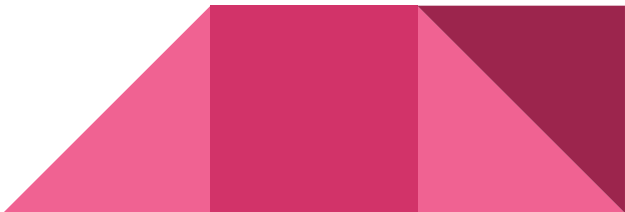


# Muestras Grandes. Comparación de medias

$\sigma_1^2, \sigma_2^2$  known:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{(\bar{x}_1 - \bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$\sigma_1^2, \sigma_2^2$  unknown:


$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sigma_{(\bar{x}_1 - \bar{x}_2)} = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$


# Muestras Pequeñas. Comparación de medias

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{s_p^2 \cdot \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where  $s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$

and  $t_{\alpha/2}$  is based on  $(n_1 + n_2 - 2)$  degrees of freedom.



## Muestras Grandes. Comparación de proporciones

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sigma_{(\hat{p}_1 - \hat{p}_2)} &= (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \\ &\approx (\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}\end{aligned}$$





# Vuestro turno

# House Prices Dataset

- Estima la diferencia de precio media entre una casa nueva y una de segunda mano.
- ¿Que la casa tenga aire acondicionado central influye en el precio significativamente?

