

TEST DE HIPÓTESIS E INTERVALO DE CONFIANZA. UNA MUESTRA

IRONHACK

Estadística inferencial

Cuando hacemos inferencia estadística es porque no tenemos los datos de toda la población (prácticamente siempre) y trabajamos con muestras que nos ayudan a estimar, con un error, estos parámetros poblacionales que no podemos llegar a conocer.

Es decir, cuando nuestro propósito es llegar a conocer ciertas características de la población a partir de la muestra de la que disponemos, estamos haciendo el proceso de inferencia.




Estadística inferencial

Dentro de la inferencia estadística, un contraste de hipótesis (también denominado test de hipótesis o prueba de significación) es un procedimiento para juzgar si una propiedad que se supone en una población estadística es compatible con lo observado en una muestra de dicha población.

Ejemplo. Contraste de hipótesis bilateral para la media, sabiendo varianza poblacional y con una sola muestra.

En una población para la cual la std es 29, contrasta la hipótesis de que $\mu=347$, con un nivel de significación del 1%, mediante una muestra de 200 individuos en la que se obtiene una media de 352.



Estadística inferencial

Llamamos **población** estadística, universo o colectivo al conjunto de referencia del que extraemos las observaciones, es decir, el conjunto de todas las posibles **unidades experimentales**.

Llamamos **muestra** a un subconjunto de elementos de la población que habitualmente utilizaremos para realizar un estudio estadístico. El número de elementos que componen la muestra es a lo que llamamos tamaño muestral y se suele representar por la letra minúscula **n**.



Estadística inferencial

Cuando queremos referirnos a las características que presentan estos conjuntos de datos y cómo medirlas, trabajamos con tres términos clave, **el estadístico, el parámetro y el estimador**:

- Un estadístico es una medida usada para describir alguna característica de una muestra y un parámetro es una medida usada para describir las mismas características pero de la población. Cuando el estadístico se calcula en una muestra con idea de hacer inferencia sobre la misma característica en la población, se le llama **estimador**.
- Lo que hacemos en contraste de hipótesis es **comparar un estadístico de la muestra** (que si podemos obtener) **con un hipotético parámetro de la población** (que solo podemos estimar).



Característica	Muestra (Estadístico)	Población (Parámetro)
Variable Cuantitativa		
Media	\bar{x}	μ
Desviación típica	s	σ
Varianza	s^2	σ^2
Variable Categórica		
Porcentaje	\hat{P}	P

Test de Hipótesis. Una muestra

Pasos test de hipótesis

1. Identificación del parámetro a estudiar

En primer lugar hemos de analizar el problema que se plantea y determinar qué parámetro poblacional queremos contrastar. Esto va a definir nuestra forma de trabajar de aquí en adelante.

En el caso de nuestro ejemplo el parámetro a estudiar es la media poblacional μ a la que le dan un valor de 347 (Este valor no coincide con la media de la muestra que hemos sacado).



Pasos test de hipótesis

2. Especificación de hipótesis nula y alternativa

En cualquier contraste de hipótesis tendremos 2 alternativas complementarias en las que se especificarán distintos valores de un parámetro poblacional y a la vista de los datos habremos de optar por una de ellas. Por ejemplo, si deseamos conocer si el valor de un parámetro μ puede ser igual a 25 o por el contrario es inadmisibile a la vista de los datos que disponemos, nuestras hipótesis serán: $\mu=25$ y $\mu\neq 25$.

Estas 2 hipótesis que hemos señalado no jugarán el mismo papel dentro de cualquier contraste de hipótesis, y por tanto cada una de ellas recibirá un nombre específico:

- Hipótesis nula, a la que habitualmente nos referimos como **H0**.
- Hipótesis alternativa, a la que habitualmente nos referimos como **HA** o **H1**.

Pasos test de hipótesis

A la hipótesis nula siempre se le concederá el beneficio de la duda e intentaremos encontrar en nuestra muestra evidencias en contra de ella. Así, al terminar el contraste habremos de optar por no rechazar H_0 (si no tenemos evidencia suficiente en su contra) o rechazarla (si los datos hacen que la descartemos), como en un juicio.

Podemos hablar de un contraste **unilateral o bilateral** en función de cómo se plantean las hipótesis:

Unilateral	Bilateral
$H_0: \mu \leq k$	$H_0: \mu = k$
$H_1: \mu > k$	$H_1: \mu \neq k$
$H_0: \mu \geq k$	
$H_1: \mu < k$	

Pasos test de hipótesis

¿Cómo se enunciarían las hipótesis del test de nuestro ejemplo?

En una población para la cual la std es 29, contrasta la hipótesis de que $\mu=347$, con un nivel de significación del 1%, mediante una muestra de 200 individuos en la que se obtiene una media de 352.



Pasos test de hipótesis

3. Fijar un valor para el nivel de significación (α)

La interpretación de este parámetro sería: Máxima probabilidad de equivocarnos que estamos dispuestos a asumir en caso de que rechazemos la hipótesis nula.

En la práctica totalidad de estudios estadísticos el valor que se suele elegir para α es 0.05, aunque también suelen tomarse $\alpha = 0.01$ o $\alpha = 0.10$ dependiendo de si queremos asumir menos o más riesgo de equivocarnos, respectivamente, en caso de rechazar la hipótesis nula.



Pasos test de hipótesis

Los tipos de error que podemos cometer quedarían reflejados en esta tabla

		Naturaleza de H_0	
		Verdadera	Falsa
Decisión	Rechazar H_0	Error de tipo I $P = \alpha$	Decisión correcta $P = 1 - \beta$
	No rechazar H_0	Decisión correcta $P = 1 - \alpha$	Error de tipo II $P = \beta$

Alfa' es el nivel de significación.

'1 - Beta' es la potencia de contraste

'1 - Alfa' es el nivel de confianza

Pasos test de hipótesis

La única forma de disminuir alfa y beta simultáneamente es aumentando el tamaño de nuestra muestra, cosa que no suele ser posible. Por esto, tendremos que decidir si disminuimos alfa o la aumentamos en función del problema con el que nos encontremos.

- En el caso de un juicio buscaremos reducir alfa por la presunción de inocencia. Preferimos cometer Error 2.
- En el caso de contrastar efectos de un medicamento cuando nuestra H_0 es que no son dañinos. Preferimos cometer Error 1.

Muchas veces te dan un nivel de significación deseado, como en nuestro ejemplo, donde $\alpha = 0.01$

Pasos test de hipótesis

4. Obtener el valor del estadístico de contraste para la muestra elegida

En el caso de los contrastes de hipótesis sobre la media podemos utilizar los valores de los estadísticos de contraste para decidir qué hacemos con la hipótesis nula.

Tenemos que tipificar nuestra distribución. Si tenemos una distribución normal $N(\mu, \sigma)$, llamamos tipificar la variable al proceso de convertirla en una Normal Estándar $N(0,1)$, lo cual nos permitirá poder consultarla en las tablas.

Si $X \rightarrow N(\mu, \sigma)$, entonces: $Z = (X - \mu) / \sigma \rightarrow N(0,1)$

Es decir, los valores X de la distribución original pasan a ser los valores Z de la distribución estandarizada $N(0,1)$.

	Population Standard Deviation is known	Population Standard Deviation is not known
Sample Size is more than 30	Z-Distribution	Z-Distribution
Sample Size is less than 30	Z-Distribution	T-Distribution

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

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \text{ where we have } n - 1 \text{ degrees of freedom}$$

*#En nuestro ejemplo conocemos la std y la muestra es mayor que 30 por lo que usaremos la prueba z
#hallamos el resultado de nuestro estadístico z*

```
import numpy as np
```

```
media_muestra = 352
```

```
Mu = 347
```

```
std_pob = 29
```

```
n = 200
```

```
z = (media_muestra - Mu) / (std_pob / np.sqrt(n))
```

```
z
```

2.4382992454708536

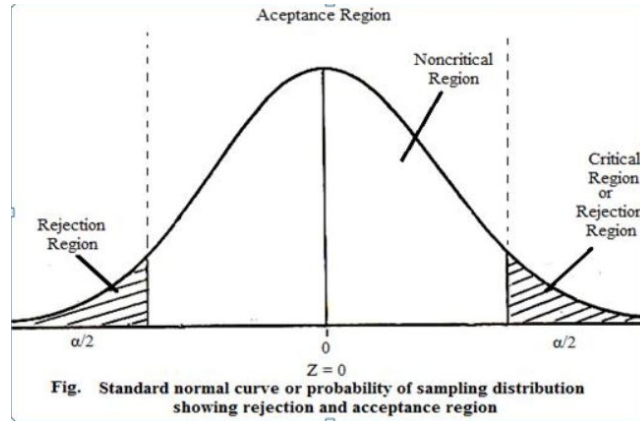
Habiendo calculado nuestro estadístico z o t tenemos que comprobar si está en la región de aceptación o en la de rechazo de la H_0 .



Pasos test de hipótesis

5. Determinar la región de aceptación y la región de rechazo

En este paso utilizamos el nivel de significación para estimar los valores de nuestra distribución de contraste. Si el estadístico está dentro de la región de aceptación, no podemos rechazar la H_0 . Si el estadístico está en la región de rechazo, rechazamos H_0 .



Pasos test de hipótesis

En la imagen vemos cómo pintar estas regiones en un contraste bilateral. Para calcular el valor de $z_{\alpha/2}$ y $-z_{\alpha/2}$ hemos de calcular la probabilidad acumulada hasta $z_{\alpha/2}$ y consultar el valor del punto crítico en una **tabla de distribución normal** (en [python](#))

En nuestro ejemplo nuestra distribución pintada también sería bilateral y nuestro $\alpha = 0.01$, así que la distribución acumulada hasta nuestro $z_{\alpha/2}$ sería: $1 - \alpha/2 = 0.995$.

Si consultamos el valor 0.995 en la tabla vemos que: $z_{\alpha/2} = 2.575$ y, por lo tanto, $-z_{\alpha/2} = -2.575$

Obtener punto crítico distribución t en [Python](#)



Pasos test de hipótesis

6. ¿Rechazamos la hipótesis nula?

Si hemos calculado nuestra región de aceptación y de rechazo y nuestro estadístico solo tenemos que ver en qué región se encuentra este para determinar qué hacemos con la H_0 .

En el caso del ejemplo hemos determinado que nuestra región de aceptación es $(-2.575, 2.575)$ y el valor de nuestro estadístico es $z = 2.438$, por lo que **no podemos rechazar la hipótesis nula** y nos arriesgamos a cometer un error del tipo 2 pero tenemos un nivel de confianza muy alto, seguramente no fallemos.



Intervalo de confianza. Una muestra

Intervalo de confianza

Otra forma rechazar o no la hipótesis en un contraste de hipótesis para la media es calculando el intervalo de confianza y ver si la media muestral está dentro; sin tipificar la distribución ni estadísticos de contraste.

$$\text{C.I.:}(\text{media_muestral} - z_{\alpha/2} * \sigma/\sqrt{n}, \text{media_muestral} + z_{\alpha/2} * \sigma/\sqrt{n})$$



Intervalo de confianza

#En el caso del ejemplo, si calculamos el intervalo de confianza (C.I.).

```
ci =(352 - 2.575 * 29/(np.sqrt(200)), 352 + 2.575 * 29/(np.sqrt(200)))  
ci
```

```
(346.7196801114895, 357.2803198885105)
```

Dado este intervalo de confianza, podemos ver que el valor de la hipótesis nula (347) estaría entre el rango de valores del intervalo, por lo que sería un posible valor de la media poblacional y, por tanto, no podríamos rechazar la hipótesis nula.



pvalue (más común)

Pvalues

También podemos calcular el p value que se define como la probabilidad correspondiente al estadístico de ser posible bajo la H_0 . Si cumple con la condición de **ser menor al nivel de significancia (alfa)** impuesto, entonces la H_0 será rechazada. Para calcularlo cogemos el valor de la tabla de la distribución correspondiente al estadístico de contraste y lo restamos a 1 (unilateral). En el caso bilateral = $(1 - \text{valor de la tabla}) * 2$

También podemos utilizar una [calculadora de p values](#)

O Python: [z pvalue](#), [t pvalue](#)





Vuestro turno

House Prices Dataset

¿Es el precio medio de venta de una casa en la zona estudiada mayor de \$180k?

- Contrasta esta hipótesis a través de intervalos de confianza, pvalue del test y región de rechazo.

El 10% de las casas vendidas son nuevas (Extra)

- Parámetro: proporción.
 - Busca por internet y/o pregúntame si te atascas.
 - Contrasta esta hipótesis a través de la técnica que prefieras.
- 