Resumen del capítulo: Verificación de hipótesis

Muestreo aleatorio y medias muestrales

La lógica de la prueba de hipótesis estadística es un poco diferente a la de la teoría de la probabilidad, donde se asumen condiciones ideales. En primer lugar, cuando probamos las hipótesis, evaluamos un gran dataset o **población estadística**, en función de las **muestras**.

No es necesario analizar el dataset completo. Todo lo que necesitas es una porción pequeña pero **representativa** de los datos que refleje las características de la población en su conjunto. La forma más fácil de asegurar la representatividad es tomar una **muestra aleatoria**. Utilizando elementos seleccionados aleatoriamente, podemos sacar conclusiones sobre la población en su conjunto.

Algunos datasets pueden tener varias partes de tamaños desiguales que difieren mucho con respecto al parámetro que estás estudiando. En estos casos, es una buena idea tomar muestras aleatorias proporcionales de cada parte y luego combinarlas. El resultado es una **muestra estratificada** que es más representativa que una muestra aleatoria normal. Lo llamamos "estratificado" porque dividimos la población en estratos o grupos que tienen algo en común. Estos estratos luego se usan para producir muestras aleatorias.

Con una muestra, puedes sacar conclusiones sobre la población, o sea sobre sus parámetros estadísticos, para ser precisos. Por lo general, es suficiente estimar los valores medios y la varianza para sacar una conclusión sobre **igualdad o desigualdad** con respecto a los valores medios de las poblaciones.

¿Qué podemos aprender acerca de la media y la varianza de una población en función de la media y la varianza que calculamos para una muestra (también denominadas **media muestral** y **varianza muestral**)? Casi todo, siempre y cuando nuestra muestra sea suficientemente grande.

Esta es una manera de establecer el teorema del límite central: si hay suficientes observaciones en una muestra, la **distribución muestral** de la media muestral de cualquier población estadística se distribuye normalmente alrededor de la media de esta población. "Cualquier población estadística" significa que la población estadística puede tener cualquier distribución. Los valores medios de las muestras

seguirán distribuidos normalmente alrededor de la media de toda la población estadística.

La medida del grado en que la media muestral se desvía de la media de la población se llama **error estándar** y se calcula mediante la fórmula:

$$E.S.E. = \frac{S}{\sqrt{n}}$$

E.S.E. significa error estándar estimado (estimated standard error). Es "estimado" porque solo tenemos una muestra. No sabemos el error exacto, solo lo estimamos en función de los datos que tenemos.

S es la desviación estándar estimada de la población.

n es el tamaño de la muestra. Dado que la raíz cuadrada de *n* está en el denominador, el error estándar disminuye a medida que aumenta el tamaño de la muestra.

Formular hipótesis

Ningún dato obtenido experimentalmente confirmará ninguna hipótesis. Esta es nuestra limitación fundamental. Los datos solo pueden contradecir la hipótesis o, por el contrario, mostrar que los resultados son extremadamente improbables (suponiendo que la hipótesis sea verdadera). Pero en ambos casos no tenemos motivos para afirmar que la hipótesis ha sido *probada*.

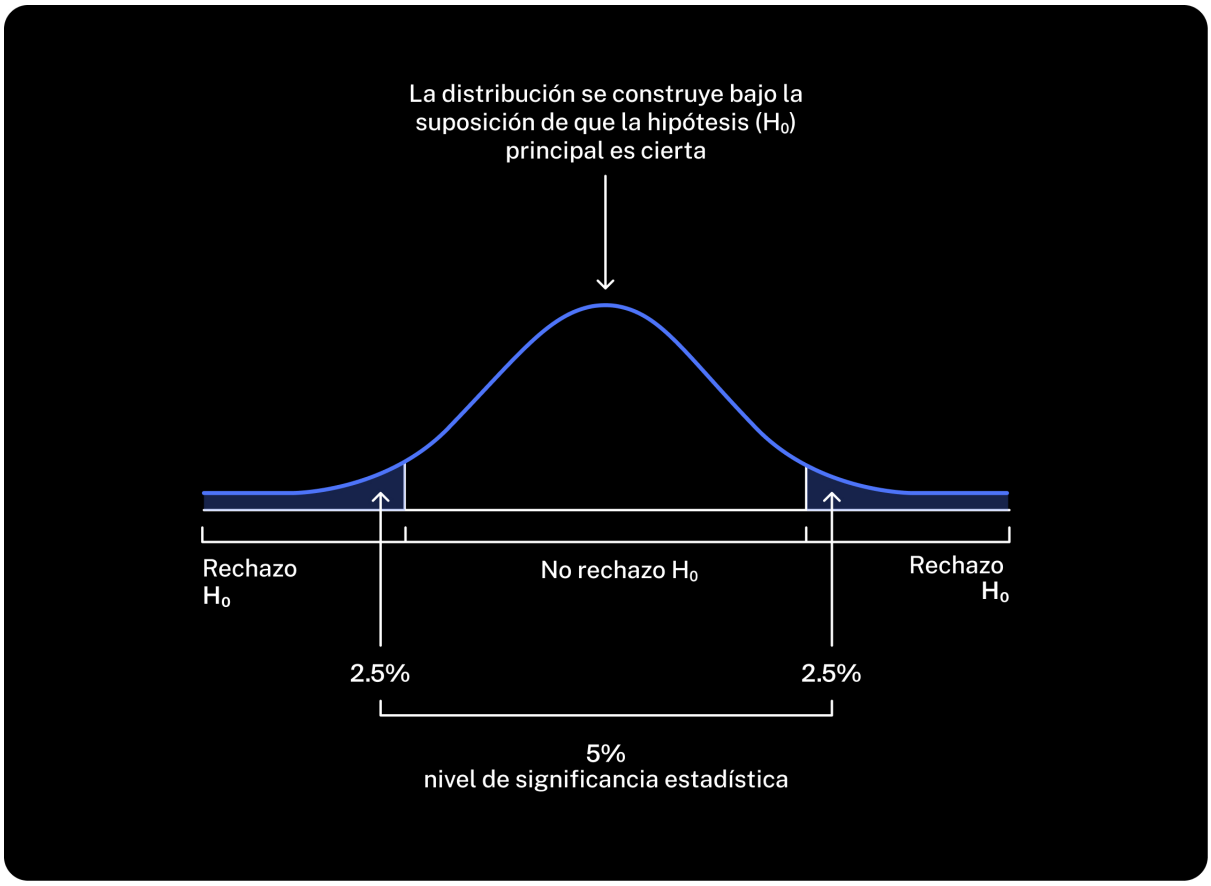
Si los datos no contradicen la hipótesis, simplemente *no la rechazamos*. Pero si, suponiendo que la hipótesis sea verdadera, es muy poco probable que obtengamos esos datos, tenemos una razón para rechazar la hipótesis.

Las hipótesis típicas pertenecen a las medias de poblaciones estadísticas y se ven así:

- La media de una población es igual a un valor determinado.
- Las medias de dos poblaciones son iguales entre sí.

La prueba de una hipótesis estadística siempre comienza con un enunciado de la hipótesis. En primer lugar, establecemos la **hipótesis nula**, **H_0**. Por ejemplo, "La media de la población en cuestión es igual a *A*", donde *A* es un número. La **hipótesis alternativa**, **H_1**, se basa en H_0. Para este H_0, H_1 sería "La media de la población no es igual a *A*". H_0 siempre se indica con un signo igual.

Construimos una distribución basada en la suposición de que H_0 es verdadera. En este caso, se trataría de una distribución normal alrededor del parámetro de interés, o sea la media. La varianza o su raíz cuadrada, la desviación estándar (una medida de la dispersión de la distribución), se estima en función de la muestra.



Para la distribución normal, la probabilidad de estar comprendido en algún intervalo es igual al área bajo la curva de ese intervalo. Habrá valores dentro de un cierto rango de la media que es muy probable que se obtengan aleatoriamente.

¿Cómo determinamos si rechazamos o no la hipótesis nula? Se especifica un valor crítico para la muestra del **nivel de significación** de la prueba de hipótesis. El nivel de significación es la probabilidad total de que un valor medido empíricamente se encuentre lejos de la media. Digamos que el valor observado en la muestra se encuentra dentro de este rango. Si asumimos que la hipótesis nula es correcta, la probabilidad de que tal evento ocurra se considera demasiado baja (con respecto al nivel de significación). Por lo tanto, tenemos motivos para rechazar la hipótesis nula. Cuando el valor se encuentra en el rango de "No rechazar H_0", no hay motivos para

rechazar la hipótesis nula. Llegamos a la conclusión de que los datos obtenidos empíricamente no refutan la hipótesis nula.

Hay un método en Python que devuelve la **estadística de diferencia** entre la media y el valor para comparar. La más importante es la significación estadística entre ellos, representada por el **valor *p***.

La **estadística de diferencia** es el número de desviaciones estándar entre los valores comparados, si ambas distribuciones se convierten en una distribución normal estándar con media 0 y desviación estándar 1. Sin embargo, este valor no te brinda suficiente información para llegar a una conclusión sobre tu hipótesis nula.

En cambio, usa el **valor *p*** para decidir si rechazar la hipótesis nula. Representa la probabilidad de obtener el resultado observado o un resultado más alejado de lo esperado, asumiendo que la hipótesis nula sea correcta. Los valores de umbral convencionales son 5% y 1%. En última instancia, la decisión sobre qué umbral considerar suficiente depende del analista.

Para probar la hipótesis de que la media de una población estadística es igual a algún valor, puedes usar el método `scipy.stats.ttest_1samp()`. Los parámetros del método son `array` (la matriz que contiene la muestra) y `popmean` (la media propuesta que usamos para la prueba). El método devuelve la estadística de la diferencia entre `popmean` y la media muestral de la matriz, así como el nivel de significación:

```
from scipy import stats as st

interested_value = 120

results = st.ttest_1samp(
    array,
    interested_value)

print('p-value: ', results.pvalue)
```

Hipótesis sobre la igualdad de las medias de dos poblaciones

A veces necesitas comparar las medias de dos poblaciones estadísticas diferentes. Para probar tu hipótesis de que las medias de dos poblaciones estadísticas son iguales según las muestras tomadas de ellas, aplica el método `scipy.stats.ttest_ind()`. El método toma los parámetros siguientes:

- `matriz1` y `matriz2` son matrices que contienen las muestras.
- `equal_var` es un parámetro opcional que especifica si las varianzas de las poblaciones deben considerarse iguales o no.

Si hay motivos para creer que las muestras se tomaron de poblaciones con parámetros similares, configura `equal_var = True` y la varianza de cada muestra se estimará a partir del dataset *combinado* de las dos muestras y no a partir de los valores de cada muestra *por separado*. Esto nos proporciona resultados más precisos. Sin embargo, lo hacemos solamente si las varianzas de las poblaciones estadísticas de las que se toman las muestras son aproximadamente iguales. De lo contrario, tenemos que configurar `equal_var = False`; de forma predeterminada, es Hipótesis sobre la igualdad de las medias de muestras pareadas `equal_var = True`.

```
from scipy import stats as st

sample_1 = [...]
sample_2 = [...]

results = st.ttest_ind(
    sample_1,
    sample_2)

print('p-value: ', results.pvalue)
```

Hipótesis sobre la igualdad de las medias de muestras pareadas

Cuando trabajamos con una población estadística, es útil saber si los cambios tienen un efecto en la media de la población. Una **muestra pareada** significa que estamos midiendo una variable de la misma entidad. En Python, para probar la hipótesis de que las medias de dos poblaciones estadísticas son iguales para muestras dependientes (pareadas) usamos la función `scipy.stats.ttest_rel()`:

```
from scipy import stats as st

before = [...]
after = [...]

results = st.ttest_rel(
    before,
    after)

print('p-value: ', results.pvalue)
```