PRUEBA DE HIPÓTESIS PARA DIFERENCIA DE MEDIAS

Introducción

Como hemos visto hasta ahora ya sabemos cómo hacer inferencia sobre bases de datos para medias con valores conocidos y desconocidos de desviación estándar. Sin embargo, sabemos que en muchos casos vamos a querer saber y entender que sucede con dos variables iguales, pero que corresponden a diferentes muestras, por lo que debemos ampliar nuevamente nuestro estudio para poder incluir dentro de nuestro análisis como proceder a inferencia sobre medias para distintas muestras.

Procedimiento

1.- Lo primero que debemos hacer es plantear como hasta ahora la hipótesis nula y la alternativa

2.- Calculamos el estadístico respectivo (Z para muestras grandes, t para muestras pequeñas)

3.- Se formula la regla de decisión

4.- Inferimos y concluimos en base a nuestros resultados

Vamos a hacer una diferencia con lo estudiado hasta ahora, en este caso asumiremos que las desviaciones estándar solo son muestrales, y haremos la diferencia solo de si las muestras en cuestión son lo suficientemente grande para ocupar o no el teorema central del límite.

Para muestras grandes:

Si las muestras tienen los suficientes datos, por el teorema central del límite podemos asumir que la distribución de las muestras es normal y trabajamos con el siguiente estadístico:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

Donde los valores de X, son las medias muestrales, los valores de mu corresponden a las diferencias de las medias de la hipótesis nula, y el valor de S, es la desviación estándar de ambas muestras, es decir:

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Nuestros valores críticos como estamos asumiendo una distribución normal serán exactamente los mismo de antes, por ejemplo, si nuestro nivel de significancia es al 5%, los niveles de Tabla serán -2,57 +2,57 dependiendo de si nuestra hipótesis alternativa tiene una cola, a la derecha o a la izquierda o -1,96 y +1,96 (recordemos que en este caso el valor de α es α/2) si tiene dos colas.

Ejemplo

Un club de golf quiere saber si existe una diferencia o no entre el tiempo de juego que demoran las mujeres al que demoran los hombres en terminar una partida a un 5% nivel de significancia, para ellos toma muestras tanto del tiempo que demoran hombres como mujeres obteniendo:

Hombres: Promedio=3,5 horas/ s=0,9 horas /n=50

Mujeres: Promedio=4,9 horas/ s=1,5 horas /n=45

Planteamos las hipótesis:

$$H_0: \mu_1 = \mu_2 \qquad \mu_1 \leq \mu_2 \qquad \mu_1 \geq \mu_2$$

$$H_1: \mu_1 \neq \mu_2 \qquad \mu_1 > \mu_2 \qquad \mu_1 < \mu_2$$

Principalmente para este caso nos importa la primera de las hipótesis alternativas pero veremos como hacerlo en todos los casos posibles:

El estadístico sera:

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{0{,}9^2}{50} + \frac{1{,}5^2}{45}} = 0{,}257$$

$$Z = \frac{(\bar{X}_1-\bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1-\bar{X}_2}} = \frac{3{,}5 - 4{,}9 - 0}{0{,}257} = -5{,}45$$

Si tomamos la primera hipótesis alternativa, es decir que queremos que ver si existe una diferencia entre las medias solamente, nuestro criterio será el siguiente: **Rechazar si el Z calculado es mayor que el Z de tabla positivo, o si es menor al Z de tabla negativo.** Dicho de otra forma, **aceptar la nula si $-\overline{Z} \leq Z \leq \overline{Z}$.**

En este caso, el Z calculado es igual a -5,45 por lo tanto esta fuera del intervalo dado por los valores de tabla, -1,96,+1,96, por lo que rechazamos la hipótesis nula, y tenemos evidencia que nos permite saber que existe una diferencia entre las medias.

Si lo analizamos desde el punto de vista de la segunda hipótesis alternativa, nos dice que la media de la primera muestra es mayor a la segunda, o dicho de otra forma que la diferencia de medias es positiva, por lo tanto estamos trabajando en la parte positiva de nuestra distribución, es decir, que el valor critico que nos importa es el valor critico positivo. De ahí que nuestro criterio de decisión será: **Rechazaremos la hipótesis nula si Z calculado es mayor al Z de tabla.**

Nuestro Z calculado es -5,45 mientras que nuestro Z de tabla en este caso es 2.57, por lo que no podemos rechazar la nula, pero ojo, aquí como la hipótesis nula es que en promedio los hombres se demoran menos o igual que las mujeres. por lo que también podríamos inferir que los hombres no se demoran más, si no que es igual o menos que las mujeres.

Si lo vemos desde el punto de vista de la tercera hipótesis alternativa, esta nos dice que la diferencia de medias nos entrega un valor negativo, por lo que estamos hablando del lado negativo de la distribución y por lo tanto el valor de

tabla que nos interesa es el -2,57. Nuestro criterio de decisión será: **Rechazar la hipótesis nula si el valor Z calculado es menor que el valor Z de tabla.**

En este caso el valor Z calculado es efectivamente menor que el de tabla, por lo que podemos rechazar la nula, que en este caso nos dice que en promedio las mujeres efectivamente demoran más en terminar el circuito de juego.

Dato adicional hasta aquí hemos visto solo diferencias asumiendo que esta es cero, si esto no es asi, es decir si nuestra hipótesis nula gira en torno a un numero dado, por ejemplo: H_0: $\mu_1 - \mu_2 = K$, lo único que cambia del estadístico es que ahora:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - K}{s_{\bar{X}_1 - \bar{X}_2}}$$

El resto se mantiene igual.

Para muestras pequeñas

En este caso debido a que las muestras son pequeñas no podemos asumir que se distribuyen normalmente, por lo que tenemos que ocupar una distribución que se acomode de mejor manera a esta última relación. Dicha distribución es la t.

Nuestro estadístico de prueba seria entonces:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Donde la diferencia estará ahora en el valor critico de Tabla t, que dependerá no solo del nivel de significancia sino también de los grados de libertad, los que se calcularan como los calculábamos para los intervalos de confianza, es decir:

$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Ejemplo

Una empresa está instalando un nuevo software para mejorar el tiempo que demoran los procesos productivos en su planta, para eso toma una muestra de 12 distintos procesos hechos tanto con el antiguo como con el nuevo software. Sus datos fueron los siguientes.

Antiguo: Promedio 325 horas S=40 n=12

Nuevo: Promedio 286 horas S=44 n=12

La empresa quiere saber si esto es efectivo a un nivel de significancia del 5%.

Como es una investigación lo mejor es poner lo que queremos probar como hipótesis alternativa, y asumir que por ahora no tenemos prueba de ello, por lo tanto:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

El estadístico t nuestro sera:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(325 - 286) - 0}{\sqrt{\frac{40^2}{12} + \frac{44^2}{12}}} = 2{,}27$$

Calculamos los grados de libertad:

$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2} = \frac{\left(\frac{40^2}{12} + \frac{44^2}{12}\right)^2}{\frac{1}{11}\left(\frac{40^2}{12}\right)^2 + \frac{1}{11}\left(\frac{44^2}{12}\right)^2} = 21{,}8$$

Según los grados de libertad y el nivel de significancia, el valor de Tabla será t=1,721

Tal como lo hacíamos antes, nosotros **podremos rechazar la hipótesis nula si el valor del estadístico calculado es mayor al de Tabla para el lado derecho de la distribución** (hipótesis alternativa en la cola superior). Así como nuestro valor calculado es 2,27 y el estadístico de tabla es 1,721 podemos rechazar la hipótesis nula, y por lo tanto inferir que el tiempo en que demoran los procesos productivos con el nuevo software es menor.

De la misma manera que lo hicimos con el ejercicio anterior para muestras grandes extraemos las formas para muestras pequeñas, lo único que cambia es que habrá que calcular los grados de libertad para encontrar los valores críticos, y que la distribución será la t-student.

STATA

Diferencia de media de ingreso por sexo. Y diferencia de media de ingreso por sexo con condicionante.

```
. ttest ytotaj, by(sexo)


Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
```

```
  hombre |     2396    806409.8    26517.91     1298024    754409.3    858410.2
   mujer |     2457    454218.6       13742    681165.2    427271.5    481165.7
---------+--------------------------------------------------------------------
combined |     4853    628100.8    15038.53     1047636    598618.4    657583.1
---------+--------------------------------------------------------------------
    diff |            352191.2    29654.49                294054.9    410327.4
------------------------------------------------------------------------------
    diff = mean(hombre) - mean(mujer)                             t =  11.8765
Ho: diff = 0                                     degrees of freedom =     4851

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

. ttest ytotaj if ytotaj>=200000, by(sexo)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  hombre |    1872     1004078    32494.42     1405923    940349.3     1067808
   mujer |    1504    683754.2    20303.34    787392.6    643928.3    723580.1
---------+--------------------------------------------------------------------
combined |    3376    861374.8    20343.87     1182047    821487.2    901262.3
---------+--------------------------------------------------------------------
    diff |            320324.3    40564.58                240790.6    399857.9
------------------------------------------------------------------------------
```

```
    diff = mean(hombre) - mean(mujer)                             t =   7.8967
Ho: diff = 0                                     degrees of freedom =     3374

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Ayuda Trabajo

sum ytotaj,d

gen g1=.

replace g1=1 if ytotaj<=303585

replace g1=2 if ytotaj>303585

ttest edad, by(g1)