PRUEBA DE HIPÓTESIS PARA LA MEDIA CON VALORES DESCONOCIDOS DE σ

Introducción

Hemos visto que no es difícil generar pruebas ni hacer inferencia de cuando podemos asumir que la muestra se distribuye normal y además conocemos de ante mano el valor de la desviación estándar de la población, pero ¿Qué pasa si nos enfrentamos a datos donde no conocemos dicha desviación? ¿Existe una manera de hacer inferencia en dicho caso?

En este caso mostraremos la forma de cómo hacer inferencia sobre la media de la población, pero esta vez, a diferencia de los ejemplos anteriores, con una desviación estándar de la población desconocida. Esto porque no es simple encontrar datos en los cuales se pueda conocer la desviación estándar de la población, por lo que es gran ayuda generar este cambio en la forma de hacer la prueba.

Anteriormente hemos visto que nos bastaba con poder mostrar que la muestra se distribuía normal, o tener el suficiente número de datos para mediante el teorema central del límite, asumir que podíamos decir que la distribución de la muestra era una normal estándar. En este caso seguiremos asumiendo que la muestra se distribuye normal, pero como de la muestra tendremos que extraer los datos tanto para la media como para la desviación estándar, la distribución que sigue nuestro estadístico es la distribución t, al igual que en el cálculo del intervalo de confianza con desviación estándar poblacional desconocida. De igual manera tenemos que modificar nuestros valores del estadístico de prueba y los valores críticos que queramos calcular.

Como sabemos en este caso tanto el valor de la media(μ) como el valor de la desviación estándar(σ) deberán ser calculadas sobre los datos de la muestra, sabemos por la estimación por intervalos que la desviación estándar que buscamos tiene una distribución t, y por lo tanto el estadístico de prueba de este caso también se distribuye bajo la distribución t, con n-1 grados de libertad.

En este caso el estadístico de prueba se mide:

$$t = \frac{\bar{x} - H_0}{s/\sqrt{n}}$$

Prueba con una cola

Para ejemplificar esta situación supongamos que una empresa quiere catalogar a distintos aeropuertos, según sus servicios, encuestando a sus clientes de clase ejecutiva, pidiéndole que le pongan nota al aeropuerto de 0 a 10, y donde se categorizara a los aeropuertos con nota igual 0 superior a 7 como aeropuertos de servicios superior.

Se sabe que en el aeropuerto de Londres se encuestaron a 60 pasajeros de clase ejecutiva y que la media obtenida de esta fue de $\bar{x} = 7{,}25$ y la desviación estándar de estos datos es s=1.052. ¿De acuerdo con esto podemos catalogar a un nivel de significancia del 5% al aeropuerto de Londres como uno de servicios superior?

Como siempre lo primero que debemos hacer es identificar la hipótesis nula de la alternativa:

$$H_0: \mu \leq 7$$

$$H_0: \mu > 7$$

Donde rechazar la hipótesis nula en este caso nos llevaría a inferir que dicho aeropuerto si puede ser catalogado como uno de servicio superior.

Debemos como en los otros casos calcular el valor de prueba para este caso con la formula mostrada anteriormente con distribución t:

$$t = \frac{\bar{x} - H_0}{s/\sqrt{n}} = \frac{7{,}25 - 7}{1.052/\sqrt{60}} = 1.84$$

Este es el valor de prueba que nosotros calculamos. Ahora bien debemos encontrar el valor crítico con el cual debemos comparar nuestro valor de prueba, para poder ver las condiciones bajo las cuales rechazaremos o no la hipótesis nula.

En este caso, la hipótesis alternativa se encuentra a la derecha de la hipótesis nula, por lo tanto debemos fijarnos en los valores positivos de la tabla de distribución t. Debemos encontrar el valor $\bar{t}$, asociado a (n-1) grados de libertad (59), y como el nivel de significancia es de 5% (α=0,05), y estamos trabajando en la parte superior de la distribución es el valor 1-α=0.95. En este caso dicho valor es:

$$\bar{t} = t_{(59{,}0{,}95)} = 1.671$$

Como lo analizamos en el caso donde la desviación estándar era conocida en la población, lo que buscamos nosotros dado el nivel de significancia, que representa como siempre el máximo de tolerancia de la probabilidad de cometer el error tipo I, por lo que nosotros queremos ver si dados los datos estamos dentro de la región aceptada dentro de la tolerancia. Dicho de otra manera, dado que estamos trabajando en el lado derecho de la distribución, lo que nos interesa a nosotros es saber si nuestro valor de prueba está dentro del nivel aceptado o no, por lo tanto la regla de decisión en este caso seria, Rechazar la hipótesis nula si $t \geq \bar{t}$.

Como vemos nuestro t = 1.84, es efectivamente mayor que el $\bar{t} = 1.671$, por lo que podemos rechazar la hipótesis nula, y efectivamente tenemos la evidencia necesaria para poder identificar al aeropuerto de Londres como uno de servicio superior.

Prueba con dos colas

Este caso es muy similar al del caso con dos colas para un valor de σ conocido de la población. Lo único que cambia con respecto ese caso es que tanto nuestra prueba estadística como nuestro valor crítico serán tomados bajo la distribución t, debido a que σ es un dato no conocido de la población.

Al igual que el caso anterior debemos tener en cuenta que la hipótesis alternativa ahora son regiones que están a ambos lados de la media, es decir, son las colas que están a la derecha y a la izquierda de la distribución con un nivel de significancia α/2, dada la condición.

El cálculo del estadístico de prueba es exactamente igual al anterior, es decir:

$$t = \frac{\bar{x} - H_0}{s/\sqrt{n}}$$

Ahora son los criterios de la prueba los que cambian con respecto al caso de una cola. Ya sabemos que los valores críticos que debemos encontrar dependerán del nivel de significancia, y os grados de libertad, estos serán entonces

$$\bar{t} = \bar{t}_{(n-1,\alpha/2)} \;\; o \;\; \bar{t} = \bar{t}_{(n-1,1-\alpha/2)}$$

Dependiendo de si el valor $\bar{x}$ que calculemos en la muestra sea mayor o menor que el de la hipótesis alternativa. Si $\bar{x} < \mu_0$, usaremos el criterio de rechazo $\bar{t} = \bar{t}_{(n-1,\alpha/2)}$, mientras que si $\bar{x} > \mu_0$, usaremos $\bar{t} = \bar{t}_{(n-1,1-\alpha/2)}$

Como siempre nuestra regla de la prueba dependerá de cómo es nuestro t estimado en comparación con el $\bar{t}$, donde rechazaremos la hipótesis alternativa si:

$$t \leq \bar{t} \; o \; t \geq \bar{t}$$

Obviamente el criterio que ocuparemos dependerá como dijimos anteriormente de si el valor de la hipótesis nula es mayor o menor que el calculado con los datos.

Otra forma de verlo es que no podemos rechazar la hipótesis nula si $\bar{t} \leq t \leq \bar{t}$

Finalmente podemos resumir estos casos de prueba de hipótesis sobre la media de la población, pero con la desviación estándar desconocida en:

Hipótesis nula	*Valor de la estadística de prueba bajo H_0*
H_0: $\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Hipótesis alternativa	*Criterios de rechazo*
H_1: $\mu \neq \mu_0$	Rechazar H_0 cuando $t \leqslant t_{\alpha/2,n-1}$ o cuando $t \geqslant t_{1-\alpha/2,n-1}$
H_1: $\mu > \mu_0$	Rechazar H_0 cuando $t \geqslant t_{1-\alpha,n-1}$
H_1: $\mu < \mu_0$	Rechazar H_0 cuando $t \leqslant t_{\alpha,n-1}$

Stata

Vamos a seguir ocupando la base de datos CASEN. Fijémonos en varios ejemplos

ttest ytotaj=500000 if edad<30
ttest ytotaj=500000 if ecivil==1
ttest ytotaj=200000 if edad<20

Para cada uno de estos resultados, ¿qué podemos inferir de la hipótesis alternativa con dos colas?

STATA para el ttest siempre ocupa una distribución t-student, ya que al considerar la base de datos ingresada como una muestra de la población, se trabaja con la desviación estándar muestras, es decir, como si la desviación estándar poblacional es desconocida.

```
. ttest ytotaj=500000 if edad<30


One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ytotaj |    1269    353411.3    13823.47      492434    326291.9    380530.7
------------------------------------------------------------------------------
    mean = mean(ytotaj)                                           t = -10.6043
Ho: mean = 500000                                degrees of freedom =     1268

  Ha: mean < 500000           Ha: mean != 500000           Ha: mean > 500000
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000         Pr(T > t) = 1.0000



. ttest ytotaj=500000 if ecivil==1

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ytotaj |    1874      879319    31586.56     1367374    817370.4    941267.5
------------------------------------------------------------------------------
    mean = mean(ytotaj)                                           t =  12.0089
Ho: mean = 500000                                degrees of freedom =     1873

  Ha: mean < 500000           Ha: mean != 500000           Ha: mean > 500000
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000         Pr(T > t) = 0.0000

.
. ttest ytotaj=200000 if edad<20

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  ytotaj |     341    63721.97    5866.154    108325.5    52183.45     75260.5
------------------------------------------------------------------------------
    mean = mean(ytotaj)                                           t = -23.2312
Ho: mean = 200000                                degrees of freedom =      340

  Ha: mean < 200000           Ha: mean != 200000           Ha: mean > 200000
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000         Pr(T > t) = 1.0000
```