DIFERENCIA ENTRE MEDIAS PARA MUESTRAS INDEPENDIENTES

Cuando hacemos estudios, hay veces que solamente queremos identificar los intervalos en que se encuentra un estimador puntual. Sin embargo, también es frecuente el querer encontrar diferencias entre dos grupos. Por ejemplo, podríamos estar interesados en la diferencia de ingreso promedio entre dos muestras, la capacidad promedio de engordar que tienen dos alimentos para aves, o la tensión que resisten en promedio dos tipos de materiales distintos para construir alas de avión. Para identificar este tipo de diferencias entre medias, utilizaremos estimadores puntuales observados en dos muestras y veremos los intervalos de confianza en que se encuentran estas diferencias. Los supuestos serán que estas muestras son de tipo aleatorio simple e independientes entre sí.

Nos enfrentaremos a una situación en donde tenemos dos poblaciones distintas (1 y 2), cada una con su respectiva media poblacional (μ_1 y μ_2). Al igual que en clases anteriores, utilizaremos las medias muestrales ($\bar{x}_1$ y $\bar{x}_2$) para obtener el intervalo de las medias poblacionales. Pero esta vez el foco será hacer una inferencia acerca de la diferencia entre ambas medias poblacionales ($\mu_1 - \mu_2$). Las dos muestras aleatorias simples e independientes que utilizaremos para la inferencia pueden tener distinta cantidad de observaciones (n_1 y n_2). Como pueden anticipar, el procedimiento de inferencia será distinto dependiendo de si conocemos o no la desviación estándar de las poblaciones (σ_1 y σ_2). En esta clase estudiaremos ambos casos con un ejemplo concreto de una empresa de zapatos con dos sucursales en distintas ciudades.

Estimación por Intervalo de la Diferencia entre Medias Poblacionales con Desviación Estándar Conocida

Primero estudiaremos la estimación por intervalo de la diferencia entre medias poblacionales con desviación estándar conocida. Supongamos que la dueña de la empresa de zapatos está preocupada por la diferencia de ventas que tiene entre las dos sucursales ubicadas en distintas ciudades. Ella cree que esta diferencia en las ventas se explica por diferencias en la edad de los clientes. Por lo tanto los contrata a ustedes para hacer una investigación acerca de la diferencia de edad de los clientes de ambas sucursales. Al comenzar su trabajo ella les entrega las desviaciones estándar de las edades de ambas poblaciones: $\sigma_1 = 9$ y $\sigma_2 = 10$. Ustedes, ya habiendo tomado varias clases de inferencia estadística, saben que pueden ocupar una distribución normal para el cálculo.

Lo primero que tienen que hacer es definir las poblaciones de clientes de ambas sucursales por separado, a las que llamaremos sucursal 1 y sucursal 2. En la estimación de la diferencia de las edades promedio de los clientes, $\mu_1 - \mu_2$, tendremos que $\mu_1$es la media poblacional de la sucursal 1 y μ_2 la media poblacional de la sucursal 2. Para estimar $\mu_1 - \mu_2$, tomaremos una muestra aleatoria simple de cada población, definidas como n_1 y n_2 y calcularemos las medias muestrales de cada muestra: $\bar{x}_1$ siendo la media muestral de la edad de los clientes de la sucursal 1 y $\bar{x}_2$ siendo el homólogo.

Para el cálculo de intervalo de confianza para la diferencia de medias, el estimador puntual será el siguiente:

$$\bar{x}_1 - \bar{x}_2$$

Al igual que con otros estimadores puntuales, tendremos un error estándar que muestra la variación de la distribución muestral del estimador. Para dos muestras aleatorias simples independientes, el error de nuestro estimador puntual será el siguiente:

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

Como vimos en clases anteriores en las estimaciones por intervalos tenemos nuestro estimador puntual $\pm$ un margen de error. Para la estimación de la diferencia entre dos medias poblacionales tenemos el siguiente margen de error:

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

Por ende nuestra estimación por intervalo de la diferencia entre dos medias poblacionales con desviación estándar conocida quedará de la siguiente manera:

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

El coeficiente de confianza seguirá siendo $(1-\alpha)$.

Supongamos que de la muestra n_1 se obtuvo una media de $\bar{x}_1 = 40\ años$ y un tamaño muestral de $n_1 = 36$, mientras que de la muestra n_2 se obtuvo una media de $\bar{x}_2 = 35\ años$ y un tamaño muestral de $n_2 = 49$.

Reemplazando estos valores en la fórmula del intervalos de confianza para la diferencia de medias, obtenemos lo siguiente:

$$40 - 35 \pm z_{\alpha/2}\sqrt{\frac{9_1^2}{36_1} + \frac{10_2^2}{49_2}}$$

Aquí el estimador puntual de la diferencia entre las medias poblacionales es 5 (40-35). ¿Qué significa esto? Que en promedio los clientes de la sucursal 1 son 5 años mayores que los clientes de la sucursal 2.

Para obtener el intervalo de confianza necesitamos el margen de error, que en este caso lo calcularemos con un 95% de confianza, siendo $z_{\alpha/2} = z_{0.025} = 1.96$.

$$5 \pm 1.96 * \sqrt{\frac{9_1^2}{36_1} + \frac{10_2^2}{49_2}}$$

$$5 \pm 4.06$$

Dado que el margen de error para el estimador puntual es de 4,06 años, podemos concluir que el intervalo de confianza del 95% para la diferencia de medias de la población es de $[0{,}94\ ,\ 9.06]$. En otras palabras, con un 95% de seguridad la diferencia de medias de la población se encuentra en ese intervalo.

Estimación por Intervalo de la Diferencia entre Medias Poblacionales con Desviación Estándar Desconocida

Imagínense que ustedes le entregan los resultados anteriores a la dueña de la empresa de zapatos y que ella se da cuenta que las desviaciones estándar poblacionales que les entregó eran sus números del Kino. Como de todas formas está contenta con el trabajo que ustedes hicieron, les ofrece pagarles el doble si repiten el estudio, pero esta vez sin información acerca de la desviación estándar poblacional. Ustedes nuevamente se sienten confiados para enfrentar el problema, ya que en sus clases de inferencia estadística aprendieron que en este caso tienen que usar una distribución $t - student$. También saben que cuando no conocemos las desviaciones estándar poblacionales es necesario usar las desviaciones estándar muestrales (s_1 y s_2) y $t_{\alpha/2}$ en vez de $z_{\alpha/2}$.

Según lo anterior, la estimación por intervalo para la diferencia entre dos medias poblacionales con desviación estándar poblacional desconocida tiene la siguiente forma:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

El coeficiente de confianza seguirá siendo $(1-\alpha)$. Algo que cambia y se vuelve algo más engorroso en este tipo de estimación es el cálculo de grados de libertad. A diferencia de cuando calculamos grados de libertad para una muestra $(n-1)$, para dos muestras ocuparemos la siguiente fórmula:

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

En la fórmula anterior s_1^2 y s_2^2 son las varianzas muestrales, mientras que n_1y n_2 los tamaños muestrales.

Supongamos que obtuvimos las siguientes desviaciones estándar para cada muestra: $s_1 = 8$ y $s_2 = 6$. Entonces podemos reemplazar estos valores y otros que ya sabemos en la fórmula del intervalo de confianza:

$$40 - 35 \pm t_{\alpha/2}\sqrt{\frac{8_1^2}{36_1} + \frac{6_2^2}{49_2}}$$

El cálculo del error estándar nos dará aproximadamente 1,59:

$$\sqrt{\frac{8_1^2}{36_1} + \frac{6_2^2}{49_2}} \approx 1{,}59$$

$$5 \pm t_{\alpha/2} * 1{,}59$$

Y el cálculo de los nuestros grados de libertad, utilizando la fórmula anterior, nos dará aproximadamente 62,2:

$$gl = \frac{\left(\frac{64}{36} + \frac{36}{49}\right)^2}{\frac{1}{36-1}\left(\frac{64}{36}\right)^2 + \frac{1}{49-1}\left(\frac{36}{49}\right)^2} \approx 62{,}2$$

Los resultados para el cálculo de grados de libertad generalmente no son números enteros y tendremos que redondearlos, pero no de cualquier forma. En este cálculo de grados de libertad SIEMPRE debemos redondear hacia abajo. Si el resultado hubiera sido 62,8, el resultado de los grados de libertad seguiría siendo 62. Así se tiene una estimación por intervalo más prudente al tener un valor t mayor. En este caso, el valor t para un nivel de confianza de 95% es de $t_{0.025} = 1{,}999$. Por lo tanto, podemos reemplazar este número en la fórmula del intervalo de confianza:

$$5 \pm t_{\alpha/2} * 1{,}59$$

$$5 \pm 1{,}999 * 1{,}59$$

En este caso estimamos un margen de error es de 3,17 años, obteniendo un intervalo de confianza del 95% para la diferencia entre las medias de la población de $[1{,}83\ , 8{,}17]$. Es decir, con un 95% de seguridad la diferencia de medias de la población se encuentra en este intervalo.

STATA

Haremos una simulación de intervalo de confianza para la diferencia entre dos áreas dentro de una empresa: marketing y recursos humanos.

```
. version 13

. set seed 10

. set obs 60
obs was 0, now 60
```

Recordemos que Stata siempre asume distribución t-Student.

Generar Muestra

Primero generaremos una variable dicotómica que tenga valor 1 para indicar el área de marketing y 0 el área de recursos humanos.

```
. gen byte area = round(runiform(),1)
```

En el comando anterior, runiform() entrega valores aleatorios con distribución uniforme entre 0 y 1, mientras que round(x,y) aproxima x al decimal y.

```
. label def area_l 0 RRHH 1 Marketing

. label val area area_l
```

. tab area

```
       area |      Freq.     Percent        Cum.
------------+-----------------------------------
       RRHH |         25       41.67       41.67
  Marketing |         35       58.33      100.00
------------+-----------------------------------
      Total |         60      100.00
```

Luego generaremos una variable continua para el ingreso de cada persona en cada área.

```
. gen ingreso = rnormal(500,300) if area == 0
(35 missing values generated)

. replace ingreso = rnormal(450,300) if area == 1
(35 real changes made)
```

En los comandos anteriores, rnormal(x,y) general valores a partir de una normal con media x y desviación estándar y.

Calcular Intervalos de Confianza para Varianzas Iguales

Primero observemos la estadística descriptiva de los ingresos de cada área:

```
. tabstat ingreso, by(area) s(N mean sd)
     area |         N      mean        sd
----------+------------------------------
     RRHH |        25  536.9574  341.9486
Marketing |        35  482.7045  337.8281
----------+------------------------------
    Total |        60  505.3099  337.7283
--------------------------------------
```

Fíjense que las desviaciones estándar son casi iguales.

Obtenemos los datos necesarios de una muestra y los guardamos como escalares:

```
. sum ingreso if area == 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     ingreso |         25    536.9574    341.9486  -450.0545    1173.13

. scalar X_1 = `r(mean)'
```

```
. scalar S2_1 = `r(Var)'

. scalar n_1 = `r(N)'
```

Repetimos el proceso para la otra muestra:

```
. sum ingreso if area == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     ingreso |         35    482.7045    337.8281  -177.8142   1381.874

. scalar X_2 = `r(mean)'

. scalar S2_2 = `r(Var)'

. scalar n_2 = `r(N)'
```

Cuando son varianzas iguales, creamos una varianza conjunta:

```
. scalar S2_c = (S2_1*(n_1 - 1) + S2_2*(n_2 - 1))/(n_1+ n_2 - 2)
```

Definimos el valor t (alfa = 0.05):

```
. scalar alfa = 0.05

. scalar t = invttail(n_1 + n_2 - 2, alfa/2)
```

Calculamos diferencia de los promedios:

```
. scalar dif = X_1 - X_2
```

Calculamos la cota inferior (sumando la misma varianza conjunta S2_c):

```
. scalar lim_inf = dif - t*(S2_c/n_1 + S2_c/n_2)^0.5
```

Calculamos la cota superior (sumando la misma varianza conjunta S2_c):

```
. scalar lim_sup = dif + t*(S2_c/n_1 + S2_c/n_2)^0.5
```

Resumimos lo encontrado:

```
. dis in red  _newline(1) "La diferencia entre " X_1 " y " X_2 " es " dif " y se
encuentra en el intervalo entre "lim_inf " y " lim_sup " con un un nivel de
confianza de " 1 - alfa _newline(1)

La diferencia entre 536.95735 y 482.70453 es 54.252818 y se encuentra en el
intervalo entre -123.72437 y 232.23001 con un un nivel de confianza de .95
```

La gracia de la respuesta anterior es que va a ir cambiando cada vez que cambien los valores. Por ejemplo, pueden recalcularla cambiando el alfa a 0.01 o 0.1.

Calcular Intervalos de Confianza para Varianzas Distintas

Ahora calcularemos la diferencia en el promedio de respuestas correctas a una prueba que respondieron ambas áreas. La principal diferencia será que tendremos varianzas distintas.

Lo primero es generar una variable que mida la cantidad de respuestas buenas que tuvo cada individuo en la prueba. Lo modelaremos como una distribución Poisson.

```
. gen prueba = rpoisson(55) if area == 0
(35 missing values generated)

. replace prueba = rpoisson(95) if area == 1
(35 real changes made)
```

Luego limpiamos los escalares:

```
. scalar drop _all
```

A continuación observemos la estadística descriptiva del rendimiento en la prueba de cada área:

```
. tabstat prueba, by(area) s(N mean sd)

     area |         N      mean        sd
----------+------------------------------
     RRHH |        25     53.12  7.178208
Marketing |        35      95.2    11.378
----------+------------------------------
    Total |        60  77.66667  23.09205
--------------------------------------
```

Las desviaciones, para la escala que manejan son diferentes (una es 50% más grande que la otra). Por lo tanto no podemos calcular una varianza conjunta y tenemos que calcular de forma especial los grados de libertad.

Obtenemos los datos (con varianzas distintas para cada muestra):

```
. sum prueba if area == 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      prueba |        25       53.12    7.178208         42         67

. scalar X_1 = `r(mean)'

. scalar S2_1 = `r(Var)'
```

```
. scalar n_1 = `r(N)'

. summ prueba if area == 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      prueba |        35        95.2      11.378         73        114

. scalar X_2 = `r(mean)'

. scalar S2_2 = `r(Var)'

. scalar n_2 = `r(N)'
```

Calculamos los grados de libertad:

```
. scalar df = floor((S2_1/n_1 + S2_2/n_2)^2/((S2_1/n_1)^2/(n_1-1) +
(S2_2/n_2)^2/(n_2-1)))

. dis as text "Hay " as result df as text " grados de libertad y no " as result
n_1 + n_2 - 2 as text " como en el caso anterior."
Hay 57 grados de libertad y no 58 como en el caso anterior.
```

Calculamos el valor t:

```
. scalar alfa = 0.05

. scalar t = invttail(df, alfa/2)
```

Calculamos la diferencia de las medias:

```
. scalar dif = X_1 - X_2
```

Calculamos la cota inferior (sumando varianzas distintas para cada muestra S2_1 y S2_2):

```
. scalar lim_inf = dif - t*(S2_1/n_1 + S2_2/n_2)^0.5
```

Calculamos la cota superior (sumando varianzas distintas para cada muestra S2_1 y S2_2):

```
. scalar lim_sup = dif + t*(S2_1/n_1 + S2_2/n_2)^0.5
```

Resumimos lo encontrado:

```
. dis in red _newline(1) "La diferencia entre " X_1 " y " X_2 " es " dif " y se
encuentra en el intervalo entre "lim_inf " y " lim_sup " con un un nivel de
confianza de " 1 - alfa _newline(1)
```

La diferencia entre 53.12 y 95.2 es -42.08 y la diferencia de las medias poblacionales se encuentra en el intervalo entre -46.885871 y -37.274129 con un nivel de confianza de .95