ESTIMACIÓN POR INTERVALO

Ya estudiamos cómo estimar intervalos de confianza cuando la desviación estándar poblacional es conocida. En ese caso usamos una estimación de la desviación estándar poblacional tomando como referencia datos históricos disponibles o algún otro valor que sirviera de referencia para este parámetro. Ahora, si lo pensamos bien, esta forma de proceder es netamente teórica, porque sería muy extraño conocer la desviación estándar de una población y no su media. La principal razón para estudiar inferencia por intervalo de una media con una desviación estándar poblacional conocida es facilitar el aprendizaje de esa misma estimación cuando la desviación estándar poblacional es desconocida.

Estimación por intervalos de la media poblacional con desviación estándar desconocida

Generalmente el cálculo del intervalo de confianza de la media poblacional se realiza sin conocer la desviación estándar de la población. En esta clase veremos cómo resolver este problema utilizando por un lado la desviación estándar muestral s, y por otro lado la misma muestra para obtener $\mu\ y\ \sigma$.

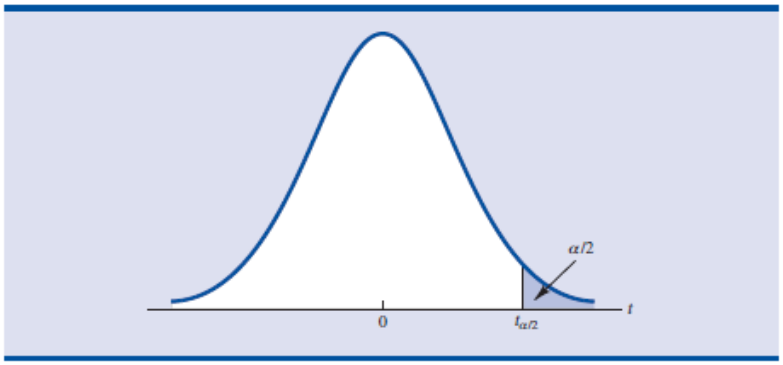
En este procedimiento ya no utilizaremos la tabla de distribución normal estándar, sino una nueva distribución de probabilidad llamada distribución t o distribución $t - student$. Este nombre proviene del seudónimo ("Student") que ocupaba Sealy Gosset en sus publicaciones, donde se describió por primera vez esta distribución. La distribución t matemáticamente parte de una distribución normal, pero se ha demostrado que sirve para muchos casos en que la población se desvía significativamente de una población normal.

Distribución t

Esta distribución es un conjunto de distribuciones similares, donde cada distribución está sujeta al número de grados de libertad que tenga. Cada distribución t depende del valor de los grados de libertad, de tal forma que mientras mayor sea este valor, más se acerca a una distribución normal y menor es su varianza. Otra característica importante de señalar es que la media de toda la distribución t es 0.

Al igual que con la tabla de probabilidades normal estándar, el valor t también tiene un sub índice que denota el área en la cola superior de la distribución de probabilidad $t - student$. En la siguiente figura se puede observar la distribución t con la probabilidad $\alpha/2$ en la cola superior ($t_{\alpha/2}$).

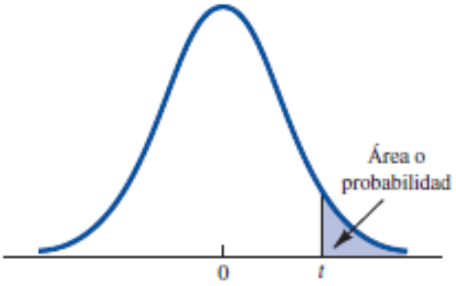
FIGURA 8.5 ÁREA DE DISTRIBUCIÓN t CON UN ÁREA O PROBABILIDAD $\alpha/2$ EN LA COLA SUPERIOR



Anderson, Sweeney y Williams (2008)

Sin embargo, la tabla de distribución t tiene ciertas diferencias con la tabla de distribución normal estándar. Para el caso de t hay dos parámetros que necesitamos conocer para obtener una probabilidad: los grados de libertad y el área de la cola superior. La siguiente figura nos muestra una sección de la tabla, que incluye grados de libertad de 1 a 9. Si bien los grados de libertad van de 1 a infinito, para grados de libertad mayores a 100 generalmente se asumen que son infinitos y se ocupa el valor Z normal estándar como una buena aproximación al valor t.

ALGUNOS VALORES DE LA TABLA DE LA DISTRIBUCIÓN t



Grados de libertad	Área en la cola superior					
	0.20	0.10	0.05	0.025	0.01	0.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250

Obtención de los intervalos de confianza

Cuando conocemos la desviación estándar poblacional, ocupamos la siguiente fórmula para la estimación por intervalo de la media poblacional:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Cuando la desviación estándar poblacional es desconocida, se utiliza la desviación estándar muestral (s) en reemplazo de sigma (σ) y se reemplaza el valor $z_{\frac{\alpha}{2}}$ por $t_{\frac{\alpha}{2}}$ de la distribución t. Por lo tanto, la estimación por intervalo de la media poblacional cuando se desconoce la desviación estándar poblacional utiliza la siguiente fórmula:

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Aquí, la expresión formal de los parámetros que utilizamos para la inferencia por intervalo es la siguiente:

- Margen de error: $t_{\alpha/2}\frac{s}{\sqrt{n}}$
- Coeficiente de confianza: $(1-\alpha)$
- Intervalo de confianza: $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
- Grados de libertad: $n-1$

Necesitamos los grados de libertad porque utilizamos la desviación estándar muestral como estimación de la desviación estándar poblacional. Los grados de libertad se refieren al número de valores independientes en el cálculo de $\sum(x_i - \bar{x})$, ya que en cualquier conjunto de datos $\sum(x_i - \bar{x}) = 0$, por lo tanto $n-1$ de esos valores son independientes, mientras que el último se puede determinar conociendo los otros. Como vimos anteriormente, la fórmula de la desviación estándar muestral es:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Ejemplo

Chile y Brasil se han enfrentado en 71 oportunidades por partidos de campeonato, ya sean considerados oficiales por la FIFA o no, recibiendo un total de 161 goles en contra y 61 goles a favor. Obtengamos la estimación por intervalo de confianza de la media poblacional de goles que Chile recibe por partido al 95% de confianza. Asumamos en este ejercicio que no podemos obtener la desviación estándar poblacional, pero que tenemos acceso a una muestra aleatoria de tamaño 10, cuyos resultados fueron los siguientes:

Ch 1-1 Br	Ch 1-6 Br
Ch 0-1 Br	Ch 2-4 Br
Ch 2-2 Br	Ch 1-1 Br
Ch 4-1 Br	Ch 0-3 Br
Ch 1-1 Br	Ch 1-2 Br

Chile recibe 2,2 goles por partido con una desviación estándar muestral de 1,7 aproximadamente. Ocupando la tabla de la distribución t presentada arriba encontramos que el valor de $t_{\frac{\alpha}{2}}$ es 2,262. Reemplazando por la fórmula correspondiente, obtenemos lo siguiente:

$$2{,}2 \pm 2{,}262 * \frac{1{,}7}{\sqrt{10}}$$

$$2{,}2 \pm 1{,}2064$$

Esto quiere decir que cuando Chile juega ante Brasil recibe en promedio 2,2 goles por partido, con un margen de error de 1,2064. Por lo tanto estimamos un intervalo de confianza para la media que va desde 0,99 a 3,41 goles por partido.

Consideraciones generales

Al enfrentarnos a un problema de estimación por intervalo de la media poblacional debemos seguir el siguiente procedimiento: Primero preguntarnos si se conoce o no la desviación estándar poblacional σ. Si la desviación estándar

poblacional es conocida, utilizamos la formula vista en clases anteriores con la tabla de probabilidad normal estándar. Si la desviación estándar poblacional no es conocida, utilizaremos la desviación estándar de la muestra (para estimar σ) con la tabla de distribución $t - student$.

Este último método de estimación lo podemos utilizar con muestras de distinto tamaño. El intervalo de confianza que estimemos será exacto cuando la población tenga una distribución normal. Si la población no se distribuye normal, los intervalos de confianza serán aproximados. La calidad de esta aproximación dependerá de la calidad de los datos, es decir, del tamaño de la muestra y de la distribución de la población. Como hemos visto a lo largo del curso, una muestra de tamaño mayor o igual a 30 puede ser suficiente, asumiendo que no tenemos observaciones atípicas o que la distribución sea muy sesgada. En caso que esto no ocurra, es recomendable aumentar el tamaño de nuestra muestra a 50 o más. Es recomendable trabajar con muestras pequeñas exclusivamente si podemos asumir normalidad en la distribución de la población.

STATA

La clase pasada aprendimos a utilizar STATA para obtener una estimación por intervalo de confianza de la media poblacional. Ahora les pediré a ustedes que trabajen en pareja y apliquen lo aprendido a otros dos ejercicios:

1. Utilicen el comando <cii> para elaborar intervalos de confianza inmediatos para variables distribuidas normalmente. Asuman los siguientes datos sobre la variable: $\bar{x} = 100$; $\sigma = 49$; $n = 120$. Asuman también que la muestra de donde se obtuvieron los valores está distribuida de forma normal, que se trata de una muestra grande y que se usó una tabla de distribución normal estándar para obtener los valores. A continuación repitan el ejercicio cambiando el nivel de confianza a 90. Luego interpreten y expliquen todos los valores expresados en la tabla de intervalos de confianza inmediatos y las diferencias que se producen al cambiar el nivel de confianza. ¿Por qué cambian los resultados? ¿Cómo se interpreta que los intervalos de confianza sean distintos?

2. Abran la base de datos casen2011_ie, y asumiendo que el total de la base de datos es una población, generen seis muestras aleatorias de la base de datos y estimen los intervalos de confianza de la variable *ytotaj* al 99% de confiabilidad para todas las muestras. Expliquen por qué se diferencian y compare con la media poblacional.

NOTAS PROFESOR

Ejercicio 1

En la primera pregunta simplemente hay que utilizar el comando <cii> con los datos que nos entrega el enunciado. La única dificultad es que los datos no estaban en orden.

```
. cii 120 100 49

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        120         100    4.473068        91.14288    108.8571
```

Por defecto el nivel de confianza utilizado es del 95%. Para agregarle un nivel de confianza menor se requiere especificar una opción.

```
. cii 120 100 49, level(90)

    Variable |        Obs        Mean    Std. Err.       [90% Conf. Interval]
-------------+---------------------------------------------------------------
             |        120         100    4.473068        92.58473    107.4153
```

Comandos para generar lo pedido en la parte 2, realizar usando el programa y mostrar los resultados. Los resultados pueden diferir a los entregados acá.

- Cambios inmediatos: Al disminuir el nivel de confianza, los intervalos se hacen más pequeños. Esto es así porque disminuye la probabilidad de que la media se encuentre dentro de la distribución de 90%.

- Interpretación: A un menor nivel de confianza los intervalos se hacen más pequeños. Esto disminuye nuestra confiabilidad de que la media poblacional se encuentre en el intervalo arrojado. Si antes teníamos una seguridad de un 95% de que la media poblacional se encontrara dentro el intervalo, ahora tenemos un 90% de seguridad de que la media poblacional se encuentra dentro del intervalo de confianza.

Ejercicio 2

```
. set seed 1234

. generate random=runiform()

. sort random

. gen group=ceil(6*_n/_N)

. ci ytotaj if group==1, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
      ytotaj |        808    609675.7     38733.5        509668.3    709683.1
```

```
. ci ytotaj if group==2, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
      ytotaj |        809    676581.4    38958.18        575994.2    777168.6

. ci ytotaj if group==3, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
      ytotaj |        809    564616.5    27175.43        494451.5    634781.5

. ci ytotaj if group==4, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
      ytotaj |        809    608267.6    38061.61        509995.3    706539.9

. ci ytotaj if group==5, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
      ytotaj |        809    611613.8       31576        530086.9    693140.8


. ci ytotaj if group==6, level(99)

    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
      ytotaj |        809    697826.8    43965.84        584310.2    811343.4

. sum ytotaj

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      ytotaj |      4853    628100.8     1047636       1342   1.96e+07
```

Podemos ver que la media poblacional del ingreso es de $628101 pesos. Si comparamos los intervalos de confianza generados en los seis grupos, vemos que la media poblacional se encuentra dentro de cada intervalo. Esto nos dice que con un 99% de seguridad la media poblacional se encontrará en el intervalo de confianza, o que el 99% de los intervalos de confianza que se generen a partir de distintas muestras contendrá la media poblacional.