2.5 INTERVALOS DE CONFIANZA PARA DIFERENCIA DE MEDIAS Y PROPORCIONES[1]

Si en lugar de tratarse de algún parámetro desconocido de una población se trata de dos poblaciones en referencia a una misma variable y con el objeto de considerar las diferencias que pudieran existir entre ambas, se desea comparar sus parámetros, surge el problema de estimar funciones en que intervienen parámetros de las dos poblaciones.

2.5.1 Diferencia de medias, muestras independientes

Se trata de estimar la diferencia existente entre μ_1 y μ_2, medias de dos poblaciones, a partir de dos muestras independientes, esto es tomando una muestra de cada población; se plantea:

a) El estimador de la diferencia de medias poblacionales, es la diferencia entre las medias muestrales $\bar{X}_1 - \bar{X}_2$

b) La esperanza de la diferencia, es igual a la diferencia de las esperanzas, por lo tanto la esperanza del estimador, es $\mu_1 - \mu_2$; la varianza del estimador por ser las muestras independientes, es igual a la suma de las varianzas, luego:

$$\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Se establecen algunas alternativas para determnar el estadístico adecuado:

2.5.1.1 Varianzas poblacionales conocidas

Siendo las poblaciones normales o las muestras suficientemente grandes como para aplicar Teorema Central del Límite, entonces el estadístico tiene distribución normal:

$$\frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Despejando el parámetro a estimar, resulta:

[1] Sección extraída de la Guía de Estadística II del Ciclo Básico a Distancia de Blanch, eta l 2004.

$$P((\bar{X}_1-\bar{X}_2)-t_{n_1+n_2-2}\sigma_{\bar{X}_1-\bar{X}_2}<(\mu_1-\mu_2)<(\bar{X}_1-\bar{X}_2)+z_{1-\frac{\alpha}{2}}\sigma_{\bar{X}_1-\bar{X}_2})=1-\alpha$$

Que es la fórmula adecuada para estimar un intervalo de confianza para la diferencia de medias, cuando se conocen las varianzas poblacionales.

2.5.1.2 Varianzas poblacionales desconocidas

En este caso es necesario utilizar un estadístico adecuado que no contenga las varianzas poblacionales. Como en el caso de la estimación por intervalos de la media, cuando se desconoce la varianza, el estadístico está asociado a la distribución t. Pero ahora es necesario además del supuesto de distribución normal, otro supuesto de igualdad de varianzas poblacionales.

Por un lado conocemos el estadístico

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Y por otro, el estadístico $\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi_{n_1+n_2-2}$ posee distribución chi cuadrado con n_1+n_2-2 grados de libertad por ser una suma de dos estadísticos con distribuciones chi cuadrado.

Con ambas expresiones se compone un estadístico con distribución T de Student:

$$\frac{\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2}}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$$

Simplificando las σ^2, resulta:

$$\frac{\left(\bar{X}_1-\bar{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\,\frac{n_1+n_2}{n_1.n_2}}} \sim t_{n_1+n_2-2}$$

En el denominador se observa la raíz cuadrada de la media ponderada de ambas varianzas muestrales, lo que se suele denominar como varianzas combinadas o pooled.

Es necesario probar el supuesto de igualdad de varianzas, lo que veremos en la sección siguiente. A partir de esa prueba, si se concluye que no hay evidencias para sospechar que las varianzas son diferentes, entonces aplicamos el estadístico que acabamos de presentar; si las hay, esto es si las varianzas no pueden considerarse iguales, existe un estadístico desarrollado por Satterthwaite, en el cual se consideran las varianzas de ambas muestras separadas, ya que no es correcto combinarlas; la distribución es también t de student pero es necesario recalcular los grados de libertad, los cuales resultan menores que los calculados en el estadístico anterior. Dicho cálculo es:

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

El estadístico es entonces:

$$\frac{\left(\bar{X}_1-\bar{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_v$$

Si las varianzas poblaciones son iguales, tenemos:

$$P((\bar{X}_1-\bar{X}_2)-t_{(n_1+n_2-2),(1-\frac{\alpha}{2})}S_{\bar{X}_1-\bar{X}_2}<\left(\mu_1-\mu_2\right)<(\bar{X}_1-\bar{X}_2)+t_{(n_1+n_2-2),(1-\frac{\alpha}{2})}S_{\bar{X}_1-\bar{X}_2})=1-\alpha$$

Si las varianzas poblacionales no son iguales, tenemos:

$$P((\bar{X}_1-\bar{X}_2)-t_{v,(1-\frac{\alpha}{2})}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}<\left(\mu_1-\mu_2\right)<(\bar{X}_1-\bar{X}_2)+t_{v,(1-\frac{\alpha}{2})}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}})=1-\alpha$$

En ambos casos, si los grados de libertad son lo suficientemente grande, los estadísticos tienden a la distribución normal.

2.5.2 DIFERENCIA DE MEDIAS, MUESTRAS DEPENDIENTES

Cuando se quiere construir un intervalo de confianza para la diferencia de medias poblacionales, pero se trata de muestras dependientes, no puede utilizarse ninguno de los estadísticos planteados, ya que todos se basan en las propiedades de la varianza para variables independientes.

¿Cuándo se trata de muestras dependientes? Un caso frecuente, es aquel en que ambas muestras corresponden a las mismas observaciones en distintos momentos del tiempo: por ejemplo, se trata de construir un intervalo para la diferencia de medias de tensión arterial antes y después de un tratamiento, para una muestra de pacientes hipertensos (hay una diferencia de medias "antes" y "después", pero se tomo una sola muestra y se realizaron a los mismos individuos dos mediciones, o se trata de analizar la diferencia entre el rendimiento promedio de un grupo de operarios trabajando en el turno matutino y el mismo grupo trabajando en turno vespertino; o las diferencias entre los montos promedio de ventas de una muestra de vendedores antes o después de haber realizado un curso de capacitación en ventas. En estos casos, las diferencias no constituyen observaciones de muestras independientes puesto que se trata de la misma muestra observada en dos oportunidades; el valor de la variable no depende solo del momento en que se mide, sino también de la observación de que se trata.

Para estas situaciones, es conveniente, en lugar de trabajar con las variables X_1 y X_2, definir una nueva variable que sea igual a la diferencia entre ambas:

$D = X_1 - X_2$, siendo cada observación $d_i = x_{1i} - x_{2i}$.

De esta manera se calculan la media y varianza muestrales de la nueva variable D y se utiliza un estadístico similar al necesario para construir un intervalo de confianza para la media de una variable (en estecaso, la variable D). Como se trabaja siempre con la varianza muestral, siendo la varianza poblacional desconocida, corresponde utilizar el estadístico con distribución t de Student y se necesita el supuesto de poblaciones normales.

Media muestral: $\bar{d} = \dfrac{\sum d_i}{n}$

Media poblacional: Δ

Varianza estimada muestral: $S_d^2 = \dfrac{\sum d_i^2 - n\bar{d}^2}{n-1}$

El estadístico adecuado es entonces: $\dfrac{\bar{d} - \Delta}{S_d / \sqrt{n}} \sim t_{n-1}$

Y el intervalo de confianza para la diferencia poblacional Δ resulta:

$$P(\bar{d} - t_{n-1,(1-\frac{\alpha}{2})}\sqrt{\frac{S_d^2}{n}} < (\Delta) < \bar{d} + t_{n-1,(1-\frac{\alpha}{2})}\sqrt{\frac{S_d^2}{n}}) = 1-\alpha$$

Si se obtiene el intervalo de confianza para la diferencia de medias de dos poblaciones, llamando a y b a los limites inferior y superior del mismo, puede decirse con una confianza del 1-α que:

$$a < \mu_1 - \mu_2 < b$$

Esto puede leerse como que, con una elevada confianza, puede esperarse que la diferencia de medias este comprendida entre a y b. Si a y b son ambos positivos, significa que es muy probable que μ_1 sea mayor que μ_2; si, por el contrario, ambos límites son negativos, ello implica que es muy probable que μ_1 sea menor que μ_2 (por eso la diferencia negativa).

En cambio si a es negativo y b positivo, ello sugiere que la diferencia entre las medias poblacionales puede ser tanto negativa como positiva, no pudiendo afirmar nada respecto a cuál de las medias es mayor.

Es por este razonamiento que suele afirmarse, en cualquier a de los dos primeros casos (ambos extremos de igual signo), que la diferencia observada entre las medias muestrales es significativa, en el sentido que indica con alta probabilidad que una de las medias poblacionales es mayor que la otra. Por el contrario, si los signos son diferentes, se afirma que la diferencia entre las medias muestrales no es significativa, en el sentido que no puede afirmarse, a partir de los resultados muestrales, que alguna de las medias poblacionales supere a la otra.

2.5.3 Diferencia de proporciones. Muestras independientes

Se trata de obtener el estadístico adecuado para construir intervalos de confianza para una diferencia de proporciones, estamos hablando por lo tanto de dos poblaciones dicotómicas. Sólo

consideramos el caso de muestras independientes y suficientemente grandes como para usar aproximación normal (esto es np y nq mayores que 5 para cada una de las muestras).

Los estimadores de cada una de las proporciones poblacionales son las proporciones muestrales, y por propiedades de esperanza y varianza de una diferencia de variables aleatorias independientes, el estadístico es:

$$\frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \; N(0,1)$$

Las proporciones estimadas en el denominador, utilizadas como estimadores de los errores estándar de la proporción en cada muestra, se aplican por desconocimiento de las poblaciones, produciéndose por este motivo un error adicional que, en general, no es importante.

Simbolizando con $\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$ el denominador de la expresión anterior, el intervalo resulta:

$$P((\hat{p}_1 - \hat{p}_2) - z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\hat{p}_1-\hat{p}_2}) < (P_1 - P_2) < (\hat{p}_1 - \hat{p}_2) + z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = 1-\alpha$$

El mismo razonamiento expresado en el punto anterior respecto a las diferencias significativas se aplica en este caso.