Inferencia Estadística

Marisol García Peña

Departamento de Matemáticas
Pontificia Universidad Javeriana

Bogotá, 2022

Tamaño de muestra para la proporción

Se quiere estimar la proporción poblacional con error admisible $\pm E$.

$$p \pm Z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\pi(1-\pi)}{n}} \Longrightarrow p \pm E$$

- La proporción estrá entre 0 y 1 $\Longrightarrow$ E también debe estar entre 0 y 1.
- Igualando el error admisible (E) a la semiamplitud del IC.
- n se aproxima al mayor entero más próximo.

Tamaño de muestra para la proporción

$$E = Z\sqrt{\frac{\pi(1-\pi)}{n}} \Longrightarrow E^2 = Z^2\left[\frac{\pi(1-\pi)}{n}\right] \Longrightarrow n = \frac{Z^2\pi(1-\pi)}{E^2}$$

$$n = \left(\frac{Z}{E}\right)^2 \pi(1-\pi).$$

¿Cómo estimar π?

Método 1 Muestra piloto. Seleccionar una muestra, calcular p y reemplazar por π. Útil si se piensa que $\pi \neq 0.5$.

Método 2 Datos históricos.

Método 3 Asumir $\pi = 0.5 \Longrightarrow$ Puede obtenerse un tamaño de muestra mayor del necesario.

Ejemplo

Una agencia bancaria quiere estimar la proporción de retiros en dinero que superen los \$500.000 en un cajero electrónico de un centro comercial. Considerando un error de $\pm 2\,\%$ y coeficiente de confianza del $95\,\%$, ¿qué tamaño de muestra es necesario para estimar la proporción de retiros superiores a \$500.000?

- $Z = 1.960$, $E = 0.02$.
- Asumiendo $\pi = 0.5$.
- $n = \left(\frac{1.960}{0.02}\right)^2 0.50(1 - 0.50) = 2401$ retiros.

Intervalo de confianza para dos poblaciones

- Se muestrean dos poblaciones normales para estimar los parámetros "comparativamente".
- $X_{11}, X_{12}, \ldots, X_{1n1}$, muestra aleatoria de tamaño n_1, μ_1 y σ_1^2.
- $X_{21}, X_{22}, \ldots, X_{2n2}$, muestra aleatoria de tamaño n_2, μ_2 y σ_2^2.
- Si el intervalo de confianza para la diferencia de dos parámetros, NO incluye al 0 $\Longrightarrow$ Existe diferencia significativa entre ellos.
- Si el intervalo de confianza para el cociente de dos parámetros, NO incluye al 1 $\Longrightarrow$ Existe diferencia significativa entre ellos.

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$

- Sirve para comparar dos medias.
- Cuando las varianzas poblacionales son desconocidas (situación usual) $\Longrightarrow$ el IC depende de los supuestos que se hagan sobre ellas.
- Población normal y las varianzas son desconocidas pero pueden considerarse iguales.
- La diferencia de medias sigue una distribución t-Student con $(n_1 - 1) + (n_2 - 1)$ g.l.
- La varianza combinada es una media ponderada de las varianzas muestrales con pesos $(n_1 - 1)$ y $(n_2 - 1)$.

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$

- Cantidad pivotal, varianzas conocidas $\Longrightarrow$?.
- Cantidad pivotal, varianzas desconocidas $\Longrightarrow$?.

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$ (cont.)

Asumiendo varianzas conocidas

$$(\overline{X}_1 - \overline{X}_2) \pm Z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$ (cont.)

Ejemplo

Construya un intervalo de confianza del 94 % para la diferencia real entre las duraciones de dos marcas de bombillos, si una muestra de 40 bombillos tomada aleatoriamente de la primera marca proporcionó una duración media de 418 horas, y una muestra de 50 bombillos de otra marca, dió una duración media de 402 horas. Las desviaciones estándar de las dos poblaciones son 26 horas y 22 horas, respectivamente.

- $\overline{X}_1 = 418, \sigma_1 = 26, n_1 = 40, \overline{X}_2 = 402, \sigma_2 = 22, n_2 = 50, \alpha = 0.06$ y $Z_{1-\frac{\alpha}{2}} = Z_{0.97} = 1.88$.
- $(418 - 402) \pm 1.88\sqrt{\frac{26^2}{40} + \frac{22^2}{50}}$.
- IC $(94\,\%) \Longrightarrow 6.3075 \leq \mu_1 - \mu_2 \leq 25.6925$.

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$

Asumiendo varianzas desconocidas e iguales

$$(\overline{X}_1 - \overline{X}_2) \pm t_{(\frac{\alpha}{2},v)} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}};$$

$$v = (n_1 - 1) + (n_2 - 1)$$

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$ (cont.)

Ejemplo

Estudiantes de último año de marketing fueron asignados aleatoriamente a un equipo virtual, o a un equipo cara a cara. Los dos equipos recibieron la tarea de analizar ocho casos complejos de marketing.

Después de completar la tarea, respondieron una escala Likert de 1 a 5 la siguiente pregunta: "Comparado a otros equipos, los miembros trabajan bien juntos".

Estadística	Equipo virtual (G1)	Equipo c-c (G2)
Media muestral	$\overline{X}_1 = 2.48$	$\overline{X}_2 = 1.83$
Desviación estándar muestral	$S_1 = 0.76$	$S_2 = 0.82$
Tamaño de muestra	$n_1 = 44$	$n_2 = 42$

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$ (cont.)

- Varianzas poblacionales desconocidas, asumidas iguales.
- Con un nivel de confianza del 90 %, t con $v = 44 + 42 - 2 = 84 \Longrightarrow t = 1.663$.
- $(2.48 - 1.83) \pm 0.2834 \Longrightarrow$ IC $(90\,\%)$ $0.366 \leq \mu_1 - \mu_2 \leq 0.934$

Conclusión

Puede afirmarse con el 90 % de confianza que existe una diferencia significativa entre las medias.

El intervalo no contiene al cero.

Asumiendo varianzas desconocidas y distintas

$$(\overline{X}_1 - \overline{X}_2) \pm t_{(\frac{\alpha}{2}, v)} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}};$$

$$v = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{\left[\frac{S_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2}\right]^2}{n_2 - 1}}, \text{Fórmula de Welch}$$

Alternativa a la Fórmula de Welch $\Longrightarrow v^* = \text{mín}(n_1 - 1, n_2 - 1)$.

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$ (cont.)

Ejemplo

Cierto metal se produce, por lo común, mediante un proceso estándar. Se desarrolla un nuevo proceso en el que se añade una aleación a la producción del metal. Los fabricantes se encuentran interesados en estimar la verdadera diferencia entre las tensiones medias de ruptura de los metales producidos por los dos procesos. Para cada metal se seleccionan 12 ejemplares y cada uno de éstos se somete a tensión hasta que se rompe.

Si se supone que el muestreo se llevó a cabo sobre dos distribuciones normales e independientes, obtener los intervalos de confianza estimados del 95 % y 99 % para la diferencia entre las medias de los dos procesos.

Intervalo de confianza para la diferencia de dos medias $(\mu_1 - \mu_2)$ (cont.)

Tensiones de ruptura de los ejemplares, en kilogramos por centimetro cuadrado.

Proceso estándar	446	401	476	421	459	438
	481	411	456	427	459	445
Proceso nuevo	462	448	435	465	429	472
	453	459	427	468	452	447

- $\overline{X}_1 = 443.333$, $S_1 = 24.824$, $n_1 = n_2 = 12$, $\overline{X}_2 = 451.417$, $S_2 = 14.939$.
- $v = 18$, $t_{0.975} = 2.10$, $t_{0.995} = 2.88$
- IC 95 % $\Longrightarrow -25.648 \leq \mu_1 - \mu_2 \leq 9.480$
- IC 99 % $\Longrightarrow -32.171 \leq \mu_1 - \mu_2 \leq 16.003$

Intervalo de confianza para la diferencia de dos proporciones $(\pi_1 - \pi_2)$

$$(p_1 - p_2) \pm Z_{(1-\frac{\alpha}{2})}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Muestras lo suficientemente grandes para asumir normalidad.

Intervalo de confianza para la diferencia de dos proporciones $(\pi_1 - \pi_2)$ (cont.)

Ejemplo

Se cree que la osteoporosis está relacionada con el género. Para esto se elige: Una muestra de 100 hombres de más de 50 años y una muestra de 200 mujeres en las mismas condiciones. Se obtiene que 10 hombres y 40 mujeres tienen algún grado de osteoporosis. ¿Qué se puede concluir con una confianza del 95 %?

- p_1: proporción de osteoporosis en mujeres mayores de 50 años.
  p_2: proporción de osteoporosis en hombres mayores de 50 años.
- $p_1 = \frac{40}{200} = 0.2,\ p_2 = \frac{10}{100} = 0.1,\ Z = 1.96.$
- $(0.2 - 0.1) \pm 1.96\sqrt{\frac{0.2(0.8)}{200} + \frac{0.1(0.9)}{100}} \Longrightarrow 0.1 \pm 0.08$
- IC $(95\,\%) \Longrightarrow 0.02 \leq \pi_1 - \pi_2 \leq 0.18.$

Intervalo de confianza para el cociente de dos varianzas $\frac{\sigma_1^2}{\sigma_2^2}$

- Se quiere comparar la precisión de un instrumento de medición con la de otro.
- Estabilidad de un proceso de manufactura con la de otro.
- Forma en que varía el procedimiento para calificar de un profesor universitario con la de otro.
- Se usa la distribución F.
- Se asume que las dos poblaciones siguen una distribución normal.

Intervalo de confianza para el cociente de dos varianzas $\frac{\sigma_1^2}{\sigma_2^2}$ (cont.)

$$\frac{\frac{S_1^2}{S_2^2}}{F_{1-\frac{\alpha}{2};n_1-1,n_2-1}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{\frac{S_1^2}{S_2^2}}{F_{\frac{\alpha}{2};n_1-1,n_2-1}}$$

- $F_{\frac{\alpha}{2};n_1-1,n_2-1}$ y $F_{1-\frac{\alpha}{2};n_1-1,n_2-1}$, percentiles de la distribución F.
- $F_{n_1-1;n_2-1;1-\frac{\alpha}{2}} = \frac{1}{F_{n_2-1;n_1-1;\frac{\alpha}{2}}}$.
- Se asume poblaciones normales.
- Para la construcción del IC, la mayor varianza debe estar en el numerador.
- `inv.f.cd(prob=`$\frac{\alpha}{2}$ `o` $1-\frac{\alpha}{2}$`;g.l.1;g.l.2)`