Probabilidad y Estadística I
Semana 11
Intervalos de confianza, simulación y tamaño muestral

Profesor: Nicolás López MSc

Universidad del Rosario

Universidad del Rosario

Contenido

Universidad del Rosario

Introducción

El objetivo de la estadística es hace inferencias de la población a través de una muestra de la misma. Como las poblaciones están caracterizadas por **parámetros**, el objetivo resulta ser encontrar estimaciones de dichos parámetros.

1. (p) Proporción de cestas logradas por un nuevo jugador.
2. (μ) Tiempo medio de espera en la fila del contact center.
3. (σ) Desviación estándar en el error de medición en la capacidad pulmonar mediante un nuevo instrumento de medida.

Hay diferentes **parámetros objetivo** de interés dependiendo el problema.

Introducción

Estos parámetros se pueden estimar de manera **puntual** o por **intervalo** (o ambos). Un **estimador** es una fórmula que indica como estimar el parámetro con la información muestral.

1. La estimación puntual requiere una fórmula.
2. **La estimación por intervalo requiere dos fórmulas**.

Introducción

- Así como el estimador puntual $\hat{\theta}$ es aleatorio (varía de muestra a muestra), el estimador por intervalo también lo hará: su longitud y ubicación es aleatoria.
- Así como con el estimador puntual $\hat{\theta}$ se tiene que $\hat{\theta} - \theta$ es aleatorio, y no se sabe que tan cerca o lejos se encuentra su realización de θ, el estimador por intervalo también tiene incertidumbre de contener, o no, a θ.

Introducción

Buscamos entonces **intervalos de confianza**, intervalo definido por los límites inferior $\hat{\theta}_L$ y superior $\hat{\theta}_U$ de confianza (ambos aleatorios), que sean

- Precisos: lo más angostos posibles

$$\hat{\theta}_U - \hat{\theta}_L$$

- Confiables: altamente probables de contener θ.

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_L) = 1 - \alpha$$

Con $1 - \alpha$ el nivel de confianza (usualmente igual a 0.95, es decir, $\alpha = 0{,}05$).

Universidad del Rosario

Introducción

Podemos tener intervalos de confianza a una o dos colas:

- A una cola, a la derecha $[\hat{\theta}_L, +\infty)$

$$P(\hat{\theta}_L \leq \theta) = 1 - \alpha$$

- A una cola, a la izquierda $(-\infty, \hat{\theta}_U]$

$$P(\hat{\theta}_U \geq \theta) = 1 - \alpha$$

- A dos colas $[\hat{\theta}_L, \hat{\theta}_U]$

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

Contenido

Universidad del Rosario

Intervalos de confianza para muestras grandes

Método pivote

Un método para encontrar intervalos de confianza se basa en el principio de cambio de escala y translación para una probabilidad dada $1-\alpha$ de una variable aleatoria. Sea Y v.a, entonces

$$P(a \leq Y \leq b) = 1 - \alpha$$

Traslación

$$P(a + d \leq Y + d \leq b + d) = 1 - \alpha$$

Escalamiento

$$P(ca \leq cY \leq cb) = 1 - \alpha$$

Se mantiene la misma probabilidad. Usar estas operaciones y despejar en la desigualdad el parámetro de interés θ de la forma

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_L) = 1 - \alpha$$

es denominado el método pivote en la construcción de intervalos de confianza a dos colas.

Universidad del Rosario

Intervalos de confianza para muestras grandes

Recordando los estimadores puntuales usuales revisados anteriormente,

Figura 1: Estimadores insesgados usuales. Tomado de [2]

Target Parameter θ	Sample Size(s)	Point Estimator $\hat{\theta}$	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^{*\dagger}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}^{\dagger}$

*σ_1^2 and σ_2^2 are the variances of populations 1 and 2, respectively.

†The two samples are assumed to be independent.

Universidad del Rosario

Intervalos de confianza para muestras grandes

Tenemos que

- Los 4 estimadores mencionados no asumen supuestos distribucionales sobre la muestra aleatoria.
- Los 4 estimadores se aproximan a la distribución normal por el TLC para muestras grandes.

Intervalos de confianza para muestras grandes

Esto significa que, si el parámetro objetivo θ es alguno de los 4 listados, para una muestra grande se tiene que

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

Se distribuye de manera aproximadamente normal estándar, y sobre esta podremos establecer intervalos de confianza de manera aproximada.

Intervalos de confianza para muestras grandes

En general, si un estadístico arbitrario $\hat{\theta} \sim N(\theta, \sigma^2_{\hat{\theta}})$, un intervalo de confianza a dos colas para θ a una confianza $1 - \alpha$ parte de:

$$P(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha$$

Con el método pivote se obtiene que un intervalo de confianza a dos colas $[\hat{\theta}_L, \hat{\theta}_U]$ para θ es igual a

$$P(\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{1-\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha$$

Que usualmente se escribe de la forma

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$$

Universidad del Rosario

Intervalos de confianza para muestras grandes

Usando el mismo principio para encontrar los intervalos a una cola

- A la derecha $[\hat{\theta}_L, +\infty)$

$$P(\hat{\theta}_L \leq \theta) = P(\hat{\theta} - z_\alpha \sigma_{\hat{\theta}} \leq \theta) = 1 - \alpha$$

- A la izquierda $(-\infty, \hat{\theta}_U]$

$$P(\hat{\theta}_U \geq \theta) = P(\hat{\theta} + z_\alpha \sigma_{\hat{\theta}} \geq \theta) = 1 - \alpha$$

Si para un intervalo a dos colas usara el $\hat{\theta}_L$ y $\hat{\theta}_U$ de los intervalos a una cola, cada uno con una confianza de $(1 - \alpha)$, el intervalo a dos colas resultante tiene una confianza de $(1 - 2\alpha)$

Intervalos de confianza para muestras grandes

Algunos comentarios

1. Dado un intervalo de confianza al $(1-\alpha)$, el parámetro poblacional **se encuentra o no en este**, para cualquier α.
2. Dado un intervalo de confianza al $(1-\alpha)$, se tiene una confianza del $100(1-\alpha)\,\%$ que el parámetro poblacional se encuentre en este. De repetir el experimento m veces y obtener m de intervalos diferentes, $m(1-\alpha)$ de estos lo contienen y $m\alpha$ no.
3. Con estos resultados puede encontrar los tres tipos de intervalos de confianza al $100(1-\alpha)\,\%$ para cualquiera de los 4 parámetros descritos, siempre que la muestra sea grande. Note, sin embargo, que necesita $\sigma_{\hat{\theta}}$ en el cálculo. Use la estimación muestral para obtener un intervalo aproximado.

Simulación en R

Ahora se hará una simulación de inbtervalos de confianza desde R.

Contenido

Universidad del Rosario

Intervalos de confianza para muestras grandes

Suponga que para un total de $n = 64$ estudiantes seleccionados aleatoriamente, el tiempo promedio de estancia en la cafetería a la hora de almuerzo es de 33 minutos, con una desviación estándar de 16 minutos. Encuentre una estimación por intervalo del tiempo promedio poblacional μ de estancia en la cafetería a la hora de almuerzo a una confianza del 90 %.

Intervalos de confianza para muestras grandes

Tenemos que $\theta = \mu$, $\hat{\theta} = \bar{y} = 33$ y $s = 16$ para $n = 64$, y además $\alpha = 0{,}1$

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}} = \bar{y} \pm z_{0{,}05}\frac{\sigma}{\sqrt{n}} \approx \bar{y} \pm z_{0{,}05}\frac{s}{\sqrt{n}} = 33 \pm 1{,}645\frac{16}{8}$$

Obteniendo un intervalo de confianza al 90 % para el tiempo promedio poblacional μ de estancia en la cafetería a la hora de almuerzo como

$$[29{,}71, 36{,}29]$$

Universidad del Rosario

Intervalos de confianza para muestras grandes

Notas

1. Recuerde que si repitiéramos el experimento 100 veces y en cada repetición obtuviéramos un nuevo intervalo, esperamos que 90 de estos contengan el parámetro poblacional de interés μ.
2. Aunque no sabemos si μ está en el intervalo [29,71, 36,29], el proceso que lo genera produce intervalos que capturan μ un 90 % de las veces.

Intervalos de confianza para muestras grandes

Suponga que de un total de 50 estudiantes seleccionados de manera aleatoria de la facultad de ciencias, 12 están de acuerdo con las políticas del rector actual. Para 60 estudiantes seleccionados de las demás facultades, 12 están de acuerdo. Estime la diferencia real entre las proporciones de estudiantes de acuerdo con el rector actual. Existe diferencia de opinión significativa entre ambos grupos de estudiantes? Realice un intervalo de confianza al 98 % para corroborarlo.

Intervalos de confianza para muestras grandes

Tenemos que $\theta = p_1 - p_2$, $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = 0{,}04$, $n_1 = 50$, $n_2 = 60$ y además $\alpha = 0{,}02$

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}} = (\hat{p}_1 - \hat{p}_2) \pm z_{0{,}01}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

De manera aproximada se tiene

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}} \approx (\hat{p}_1 - \hat{p}_2) \pm z_{0{,}01}\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Igual a

$$0{,}04 \pm 2{,}33\sqrt{\frac{(0{,}24)(0{,}76)}{50} + \frac{(0{,}20)(0{,}80)}{60}} = 0{,}04 \pm 0{,}1851$$

Obteniendo un intervalo de confianza al 98 % para la diferencia entre las proporciones $p_1 - p_2$ de estudiantes de acuerdo con el rector actual.

$$[-0{,}1451, 0{,}2251]$$

Note que el intervalo que incluye al cero.

Universidad del Rosario

Contenido

Universidad del Rosario

Tamaño muestral

Bajo muestreo aleatorio de una población determinada, se pregunta por el tamaño requerido de la muestra para obtener conclusiones confiables del experimento a realizar. Un tamaño pequeño de muestra no permite usar el TLC, pero esto no acota el tamaño poblacional.

Tamaño muestral

Para determinar n, se fija un margen de error admisible E y una confianza determinada $1-\alpha$ en la estimación. Se tiene entonces:

$$z_{\alpha/2}\sigma_{\hat{\theta}} < E$$

De la inecuación resultante se despeja n y se toma la cota inferior de dicho resultado. Suponga que quiere determinar n para estimar μ:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < E \longrightarrow n > \left(\frac{z_{\alpha/2}}{E}\right)^2 \sigma^2 \longrightarrow n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \sigma^2$$

Suponga ahora que quiere determinar n para estimar p:

$$z_{\alpha/2}\sqrt{\frac{pq}{n}} < E \longrightarrow n > \left(\frac{z_{\alpha/2}}{E}\right)^2 pq \longrightarrow n = \left(\frac{z_{\alpha/2}}{E}\right)^2 pq$$

Tamaño muestral

Note que en la estimación del tamaño muestral:

1. Para μ se desconoce σ. Puede aproximarse mediante s de un experimento previo o sabiendo el rango de valores de la variable R, ya que $R \approx 4s$.
2. Para p se desconoce p naturalmente. A menos que se tenga información auxiliar que soporte lo contrario, el valor conservador usado para p es 0.5.
3. Para $\mu_1 - \mu_2$ y $p_1 - p_2$ se sigue un proceso similar al descrito anteriormente y se tienen las mismas consideraciones.