Análisis estadístico de datos simulados
Estimadores por intervalos

Georgina Flesia

FaMAF

5 de mayo, 2016

Estimador por intervalos

Un estimador por intervalo de un parámetro es un intervalo aleatorio con una probabilidad de cobertura para el parámetro.

Sean $X_1, \ldots, X_n$ independientes e idénticamente distribuídos $F(\theta)$. Quiero encontrar $L(X_1, \ldots, X_n), R(X_1, \ldots, X_n)$ tal que

$$P\left(L(X_1, \ldots, X_n) \leq \theta \leq R(X_1, \ldots, X_n)\right) = 1 - \alpha$$

La confianza que se da al intervalo es la probabilidad de que el intervalo contenga al parámetro, usualmente $1 - \alpha$.

Estimador por intervalos

Estimador por intervalo de la media poblacional

Sean $X_1, \ldots, X_n$ i.i.d con media μ.

Quiero encontrar $L(X_1, \ldots, X_n), R(X_1, \ldots, X_n)$ tal que

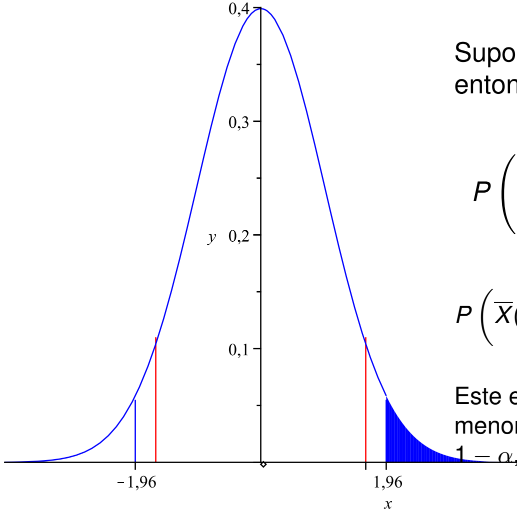
$$P\left(L(X_1, \ldots, X_n) \leq \mu \leq R(X_1, \ldots, X_n)\right) = 1 - \alpha$$

Sabemos que

- $\overline{X}(n)$ es un estimador puntual de la media basado en $X_1, \ldots, X_n$.
- Si la población es normal con media μ y d.s. σ,

$$\frac{\overline{X}(n) - \mu}{\sigma/\sqrt{n}} \sim Z = N(0, 1)$$

Ejemplo



Supongamos $1-\alpha = 0.95$, entonces

$$P\left(\frac{|\overline{X}(n)-\mu|}{\sigma\sqrt{n}} \leq 1.96\right) = 0.95.$$

$$P\left(\overline{X}(n) - 1.96\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X}(n) + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Este es un intervalo posible, el de menor ancho con probabilidad fija $1-\alpha$, y es simétrico.

Estimador por intervalos

- El intervalo aleatorio con extremos

$$\overline{X}(n) - 1.96\,\sigma/\sqrt{n} \qquad \text{y} \qquad \overline{X}(n) + 1.96\,\sigma/\sqrt{n}$$

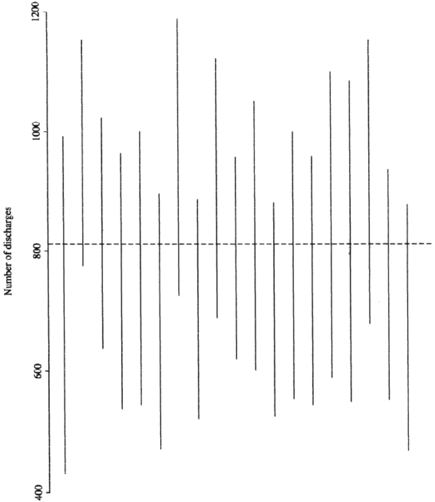
  se dice que es un estimador por intervalo, con un 95% de confianza para la media μ.

- Si $\overline{x}$ es un valor observado de $\overline{X}(n)$, el intervalo con extremos

$$\overline{x} - 1.96\,\sigma/\sqrt{n} \qquad \text{y} \qquad \overline{x} + 1.96\,\sigma/\sqrt{n}$$

  es el valor estimado del estimador por intervalo de μ, con un 95% de confianza.

Estimador por intervalos: Significado



- $(\overline{X} - \frac{1.96\sigma}{\sqrt{n}}, \overline{X} + \frac{1.96\sigma}{\sqrt{n}})$.
- $z_{0.025} = 1.96$.
- El 95% de los intervalos cubren la media.

Estimador por intervalo de la media poblacional

- $\overline{X}(n)$ es un estimador puntual de la media.
- Si la población es normal con media θ y d.s. σ,

$$\frac{\overline{X}(n) - \theta}{\sigma/\sqrt{n}} \sim Z = N(0, 1)$$

- $P(Z > z_\alpha) = \alpha$, para $0 < \alpha < 1$.
- Si el nivel de confianza deseado es $1 - \alpha$,

$$P\left(\frac{|\overline{X}(n) - \mu|}{\sigma\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

$$P\left(\overline{X}(n) - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X}(n) + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Estimador por intervalos

- El intervalo aleatorio con extremos

$$\overline{X}(n) - z_{\alpha/2}\,\sigma/\sqrt{n} \qquad \text{y} \qquad \overline{X}(n) + z_{\alpha/2}\,\sigma/\sqrt{n}$$

  se dice que es un estimador por intervalo, con un $100(1-\alpha)\%$ de confianza para la media μ.

- Si $\overline{x}$ es un valor observado de $\overline{X}(n)$, el intervalo con extremos

$$\overline{x} - z_{\alpha/2}\,\sigma/\sqrt{n} \qquad \text{y} \qquad \overline{x} + z_{\alpha/2}\,\sigma/\sqrt{n}$$

  es el valor estimado del estimador por intervalo de μ, con un $100(1-\alpha)\%$ de confianza.

Estimador por intervalos

- Si la varianza σ^2 es desconocida, utilizamos el estimador $S^2(n)$.
- Para determinar un intervalo de confianza, es necesario conocer la distribución del estadístico:

$$\sqrt{n}\,\frac{\overline{X}(n)-\theta}{S(n)}$$

<u>Distribuciones derivadas de la normal</u>

- χ^2 de Pearson con k grados de libertad: si $Z_1, Z_2, \ldots, Z_k$ son v.a. N(0,1), independientes:

$$\chi^2_k = Z_1^2 + \cdots + Z_k^2$$

- T_k de Student, con k grados de libertad: (W. S. Gosset)

$$T_k = \frac{Z}{\sqrt{\frac{\chi^2_k}{k}}}$$

Estimador por intervalos

Distribuciones derivadas de la normal

- Si $X_1, X_2, \ldots, X_k$ son v.a. $N(\mu, \sigma^2)$, independientes: el estadístico S^2 tiene una distribución T_{n-1}:

$$\sqrt{n}\frac{\overline{X}(n) - \mu}{S(n)} \sim T_{n-1}$$

- Sea t_α tal que $\quad P(|T_{n-1}| > t_\alpha) = 1 - \alpha.$

$$P\left(\overline{X}(n) - t_{\alpha/2}\frac{S(n)}{\sqrt{n}} \leq \mu \leq \overline{X}(n) + t_{\alpha/2}\frac{S(n)}{\sqrt{n}}\right) = 1 - \alpha.$$

Estimador por intervalos

- El intervalo aleatorio con extremos

$$\overline{X}(n) - t_{\alpha/2}\, S(n)/\sqrt{n} \qquad \text{y} \qquad \overline{X}(n) + t_{\alpha/2}\, S(n)/\sqrt{n}$$

  se dice que es un estimador por intervalo, con un $100(1-\alpha)\%$ de confianza para la media μ con σ desconocido.

- Si $\overline{x}$ es un valor observado de $\overline{X}(n)$, el intervalo con extremos

$$\overline{x} - t_{\alpha/2}\, s/\sqrt{n} \qquad \text{y} \qquad \overline{x} + t_{\alpha/2}\, s/\sqrt{n}$$

  es el valor estimado del estimador por intervalo de μ, con un $100(1-\alpha)\%$ de confianza, con σ desconocido.
- Para $n > 120$, puede usarse la distribución normal, es decir, $t_\alpha \approx z_\alpha$.

Intervalos de confianza para proporciones

- $X_1, X_2, \ldots, X_n$: Bernoulli, independientes, con probabilidad p de éxito.
- Para n suficientemente grande tal que np y $n(1-p)$ es mayor que 5,

$$X_1 + \cdots + X_n = Bi(n,p) \sim N(np, np(1-p)).$$

- Si p es desconocido, podemos estimar p con la media muestral:

$$\hat{p} = \overline{X}(n) \quad \text{y} \quad \mathrm{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}.$$

- Intervalos de confianza del $100(1-\alpha)\%$:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Longitud del intervalo de confianza

- Estimación de la media: $s(n)$: valor observado de la varianza muestral.

$$2\,\frac{z_{\alpha/2}\sigma}{\sqrt{n}} \quad \text{o} \quad 2\,\frac{z_{\alpha/2}s(n)}{\sqrt{n}}.$$

- Estimación de la proporción:

$$2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- La longitud del intervalo de confianza al $100(1-\alpha)\%$ depende del tamaño de la muestra.

Cuando parar una simulación para estimar la media

- Definir α y d, para el nivel de confianza y el del error.
- Generar al menos 30 datos.
- Continuar generado datos hasta que k, el número de datos generados produzca

$$2\,\frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq d$$

- si σ es desconocido $S(k)$ debe ser calculado a cada paso

$$2\,\frac{z_{\alpha/2}s(k)}{\sqrt{n}} \leq d$$

Desea determinarse mediante Monte Carlo el valor de la integral

$$I = \int_A^B g(x)\, dx$$

a) Indicar cómo se obtiene mediante simulación el valor de la integral.
b) Escribir la expresión genérica correspondiente a un intervalo de confianza del 95% para el valor de la integral. Detalle cómo se obtiene cada uno de los componentes de la expresión.
c) Obtener mediante simulación en computadora el intervalo de confianza del 95% cuyo semi-ancho sea igual a 0.005. Indicar cuál es el número de simulaciones N_s necesarias para lograr la condición pedida y completar con los valores obtenidos (usando **5** decimales) la siguiente tabla:

N^o de sim.	I	S
100		
1 000		
3 000		
5 000		
$N_s =$		