Intervalos y regiones de confianza

Graciela Boente[1]

[1]Universidad de Buenos Aires and CONICET, Argentina

Regiones de confianza

Dado un vector $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, *una región de confianza* $\mathcal{S}(\mathbf{X})$ *para* $\boldsymbol{\theta}$ *con nivel de confianza* $1 - \alpha$ será una función que a cada $\mathbf{X}$ le hace corresponder un subconjunto $\mathcal{S}(\mathbf{X}) \subset \Theta$ de manera que

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{S}(\mathbf{X})) = 1 - \alpha, \qquad \forall \boldsymbol{\theta} \in \Theta$$

Es decir, $\mathcal{S}(\mathbf{X})$ cubre el valor verdadero del parámetro con probabilidad $1 - \alpha$.

Regiones de confianza

Dado un vector $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, *una región de confianza* $\mathcal{S}(\mathbf{X})$ *para* $\boldsymbol{\theta}$ *con nivel de confianza* $1-\alpha$ será una función que a cada $\mathbf{X}$ le hace corresponder un subconjunto $\mathcal{S}(\mathbf{X}) \subset \Theta$ de manera que

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in \mathcal{S}(\mathbf{X})) = 1-\alpha, \qquad \forall \boldsymbol{\theta} \in \Theta$$

Es decir, $\mathcal{S}(\mathbf{X})$ cubre el valor verdadero del parámetro con probabilidad $1-\alpha$.

Caso particular: Si $\boldsymbol{\theta} \in \mathbb{R}$ se dirá que $\mathcal{S}(\mathbf{X})$ es un intervalo de confianza

$$\mathcal{S}(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$$

La longitud de $\mathcal{S}(\mathbf{X})$ es

$$L = b(\mathbf{X}) - a(\mathbf{X})$$

Procedimientos generales para obtener RC

$\mathbf{X}$ un vector aleatorio cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Una función $G(\mathbf{X}, \boldsymbol{\theta})$ se llama un pivote si y sólo si la distribución de $G(\mathbf{X}, \boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$.

Procedimientos generales para obtener RC

$\mathbf{X}$ un vector aleatorio cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Una función $G(\mathbf{X}, \boldsymbol{\theta})$ se llama un pivote si y sólo si la distribución de $G(\mathbf{X}, \boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$.

Teorema Sea $\mathbf{X}$ un vector aleatorio cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Sea

- $U = G(\mathbf{X}, \boldsymbol{\theta})$ una variable aleatoria cuya distribución es independiente de $\boldsymbol{\theta}$.
- A y B tales que $\mathbb{P}(A \leq U \leq B) = 1 - \alpha$.

Luego, si $\mathcal{S}(\mathbf{X}) = \{\boldsymbol{\theta} : A \leq G(\mathbf{X}, \boldsymbol{\theta}) \leq B\}$,

$\mathcal{S}(\mathbf{X})$ es una región de confianza a nivel $(1 - \alpha)$ para $\boldsymbol{\theta}$.

Relación entre test y Regiones de confianza

- $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.
  Para cada $\boldsymbol{\theta}_0$ fijo sea $\phi_{\boldsymbol{\theta}_0}$, un test no aleatorizado de nivel α, para

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Relación entre test y Regiones de confianza

- $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.
  Para cada $\boldsymbol{\theta}_0$ fijo sea $\phi_{\boldsymbol{\theta}_0}$, un test no aleatorizado de nivel α, para
  $$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$
- $\mathcal{A}(\boldsymbol{\theta}_0)$ Región de aceptación de $\phi_{\boldsymbol{\theta}_0}$
- $\mathcal{S}(\mathbf{X}) = \{\boldsymbol{\theta} : \phi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\} = \{\boldsymbol{\theta} : \mathbf{X} \in \mathcal{A}(\boldsymbol{\theta})\}$
  $\Longrightarrow \mathcal{S}(\mathbf{X})$ es una región de confianza de nivel $1 - \alpha$ para $\boldsymbol{\theta}$

Relación entre test y Regiones de confianza

- $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.
  Para cada $\boldsymbol{\theta}_0$ fijo sea $\phi_{\boldsymbol{\theta}_0}$, un test no aleatorizado de nivel α, para
  $$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$
- $\mathcal{A}(\boldsymbol{\theta}_0)$ Región de aceptación de $\phi_{\boldsymbol{\theta}_0}$
- $\mathcal{S}(\mathbf{X}) = \{\boldsymbol{\theta} : \phi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\} = \{\boldsymbol{\theta} : \mathbf{X} \in \mathcal{A}(\boldsymbol{\theta})\}$
  $\Longrightarrow \mathcal{S}(\mathbf{X})$ es una región de confianza de nivel $1 - \alpha$ para $\boldsymbol{\theta}$
- Recíprocamente, si $\mathcal{S}(\mathbf{X})$ es una región de confianza de nivel $1 - \alpha$ para $\boldsymbol{\theta}$, el test

$$\phi_{\boldsymbol{\theta}_0}(\mathbf{X}) = \begin{cases} 1 & \text{si } \boldsymbol{\theta}_0 \notin \mathcal{S}(\mathbf{X}) \\ 0 & \text{si } \boldsymbol{\theta}_0 \in \mathcal{S}(\mathbf{X}) \end{cases}.$$

es un test de nivel de α para testear
$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Intervalo de longitud prefijada para μ en una $N(\mu, \sigma^2)$

- Tomemos una muestra inicial $X_1, \ldots, X_n$.
- Estimamos σ^2 por

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

con

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Sea m tal que

$$\frac{2 s_n t_{\frac{\alpha}{2}, n-1}}{\sqrt{n+m}} \leq L$$

(m es una variable aleatoria)

Intervalo de longitud prefijada para μ en una $N(\mu, \sigma^2)$

- Sea $X_{n+1}, \ldots, X_{n+m}$ una muestra complementaria y

$$\overline{X}_{n+m} = \frac{1}{n+m} \sum_{i=1}^{n+m} X_i$$

El intervalo de confianza de nivel $1-\alpha$ con longitud menor o igual a L es

$$[\overline{X}_{n+m} - t_{\frac{\alpha}{2},n-1} \frac{s_n}{\sqrt{n+m}}, \overline{X}_{n+m} + t_{\frac{\alpha}{2},n-1} \frac{s_n}{\sqrt{n+m}}]$$

Intervalo de longitud prefijada para μ en una $N(\mu, \sigma^2)$

Sean $X_1, \ldots X_n$ variables aleatorias independientes con distribución $\mathrm{N}(\mu, \sigma^2)$ y sea m como antes.

(i) $W = (n-1)s_n^2/\sigma^2 \sim \chi^2_{n-1}$

(ii) $V = \sqrt{m+n}(\overline{X}_{m+n} - \mu)/\sigma \sim N(0,1)$

(iii) V y W son independientes

(iv) $\sqrt{m+n}(\overline{X}_{m+n} - \mu)/s_n \sim \mathcal{T}_{n-1}$

Por lo tanto,

$$[\overline{X}_{n+m} - t_{\frac{\alpha}{2},n-1}\frac{s_n}{\sqrt{n+m}}, \overline{X}_{n+m} + t_{\frac{\alpha}{2},n-1}\frac{s_n}{\sqrt{n+m}}]$$

es un intervalo de confianza para μ de nivel $1-\alpha$ con longitud menor o igual a L.

Intervalo de confianza para diferencia de medias

Sean $X_1, \cdots, X_{n_1}$ $X_i \sim \mathrm{N}(\mu_1, \sigma^2)$ y $Y_1, \cdots, Y_{n_1}$ $Y_i \sim \mathrm{N}(\mu_2, \sigma^2)$, independientes Sean

$$
\begin{aligned}
V &= \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2}{\sigma^2} \\
W &= \frac{\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{\sigma^2} \\
s^2 &= \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2\right) \\
T &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{s}\right)
\end{aligned}
$$

Intervalo de confianza para diferencia de medias

(i)
$$U = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma} \right) \sim N(0,1)$$

(ii) $V \sim \chi^2_{n_1-1}$, $W \sim \chi^2_{n_2-1}$

(iii) U, V y W son independientes

(iv) $Z = V + W \sim \chi^2_{n_1+n_2-2}$

(v) $T \sim \mathcal{T}_{n_1+n_2-2}$

Por lo tanto,

$$[\overline{X} - \overline{Y} - s \frac{n_1 + n_2}{n_1\, n_2} t_{\frac{\alpha}{2}, n_1+n_2-2}, \overline{X} - \overline{Y} + s \frac{n_1 + n_2}{n_1\, n_2} t_{\frac{\alpha}{2}, n_1+n_2-2}]$$

es un intervalo de confianza para $\mu_1 - \mu_2$ de nivel $1 - \alpha$.

Intervalo de confianza para diferencia de medias, Muestras apareadas

$(X_1, Y_1), \ldots, (X_n, Y_n)$ independientes tales que

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix} \right).$$

Queremos un IC para $\lambda = \mu_1 - \mu_2$.

(i) $Z_i = X_i - Y_i$ $Z_i \sim \mathrm{N}(\lambda, \sigma_Z^2)$, con

$$\sigma_Z^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2,$$

(ii) Sea $s_Z^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$, entonces

$$\left[\overline{Z} - t_{n-1,\frac{\alpha}{2}}\frac{s_Z}{\sqrt{n}}, \overline{Z} + t_{n-1,\frac{\alpha}{2}}\frac{s_Z}{\sqrt{n}}\right]$$

es un intervalo de confianza para $\mu_1 - \mu_2$ de nivel $1 - \alpha$.

Regiones de confianza con nivel asintótico $(1-\alpha)$

Sea $X_1, X_2, \ldots, X_n$ m.a. $X_i \sim F(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Se dice que $S_n(X_1, \ldots, X_n)$ es una sucesión de regiones de confianza con nivel asintótico $1-\alpha$ si:

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S_n(X_1, \ldots, X_n)) = 1-\alpha \quad \forall\, \theta \in \Theta\ .$$

Procedimiento para obtener RC con nivel asintótico

Teorema Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $F(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Supongamos que

- $\forall n$, $\exists$ v.a. $U_n = G_n(X_1, \ldots, X_n, \boldsymbol{\theta})$ tales que $U_n \xrightarrow{D} U$, donde U es una variable aleatoria con distribución independiente de $\boldsymbol{\theta}$
- A y B puntos de continuidad de F_U tales que $\mathbb{P}(A \leq U \leq B) = 1 - \alpha$.

Luego, si

$$\mathcal{S}_n(X_1, \ldots, X_n) = \{\boldsymbol{\theta} : A \leq G_n(X_1, \ldots, X_n, \boldsymbol{\theta}) \leq B\}$$

$\mathcal{S}_n(\mathbf{X})$ es una sucesión de RC con nivel asintótico $(1 - \alpha)$.

Optimalidad de Regiones de confianza

Sea $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

Definición Sean $\{\mathcal{S}_1(\mathbf{X})\}$ y $\{\mathcal{S}_2(\mathbf{X})\}$ dos regiones de confianza de nivel $1-\alpha$. Diremos que $\{\mathcal{S}_1(\mathbf{X})\}$ es mejor que $\{\mathcal{S}_2(\mathbf{X})\}$ en θ_1 si

$$P_{\theta_1}\left(\theta \in \mathcal{S}_1(\mathbf{X})\right) < P_{\theta_1}\left(\theta \in \mathcal{S}_2(\mathbf{X})\right) \quad \theta \neq \theta_1$$

Optimalidad de Regiones de confianza

Sea $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

Definición Sean $\{\mathcal{S}_1(\mathbf{X})\}$ y $\{\mathcal{S}_2(\mathbf{X})\}$ dos regiones de confianza de nivel $1-\alpha$. Diremos que $\{\mathcal{S}_1(\mathbf{X})\}$ es mejor que $\{\mathcal{S}_2(\mathbf{X})\}$ en θ_1 si

$$P_{\theta_1}\left(\theta \in \mathcal{S}_1(\mathbf{X})\right) < P_{\theta_1}\left(\theta \in \mathcal{S}_2(\mathbf{X})\right) \quad \theta \neq \theta_1$$

Teorema Sea $\mathcal{A}(\boldsymbol{\theta}_0)$ Región de aceptación de $\phi_{\boldsymbol{\theta}_0}$, test UMP de nivel α para

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \in \mathcal{R}_{\boldsymbol{\theta}_0} \quad \theta_0 \notin \mathcal{R}_{\theta_0}.$$

$\mathcal{S}(\mathbf{X}) = \{\boldsymbol{\theta} : \phi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\} = \{\boldsymbol{\theta} : \mathbf{X} \in \mathcal{A}(\boldsymbol{\theta})\}$

$\Longrightarrow \forall \theta_1 \in \mathcal{R}_{\boldsymbol{\theta}_0}$, $\mathcal{S}(\mathbf{X})$ minimiza

$$P_{\theta_1}\left(\theta \in \mathcal{S}_1(\mathbf{X})\right) \quad \theta \neq \theta_1$$

entre todas las regiones $\mathcal{S}_1(\mathbf{X})$ de nivel $(1-\alpha)$ para $\boldsymbol{\theta}$.

Optimalidad de Regiones de confianza

Sea $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

Definición Una familia $\{\mathcal{S}(\mathbf{X})\}$ de regiones de confianza de nivel $1-\alpha$ se dice UMA (*uniformemente más exacta*) si dada $\{\mathcal{S}^{\star}(\mathbf{X})\}$ de nivel $1-\alpha$ se tiene

$$P_{\theta_1}\left(\theta \in \mathcal{S}(\mathbf{X})\right) \leq P_{\theta_1}\left(\theta \in \mathcal{S}^{\star}(\mathbf{X})\right) \quad \theta, \theta_1$$

Optimalidad de Regiones de confianza

Sea $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

Definición Una familia $\{\mathcal{S}(\mathbf{X})\}$ de regiones de confianza de nivel $1-\alpha$ se dice UMA (*uniformemente más exacta*) si dada $\{\mathcal{S}^{\star}(\mathbf{X})\}$ de nivel $1-\alpha$ se tiene

$$P_{\theta_1}\left(\theta \in \mathcal{S}(\mathbf{X})\right) \leq P_{\theta_1}\left(\theta \in \mathcal{S}^{\star}(\mathbf{X})\right) \quad \theta, \theta_1$$

Definición Sea $\mathbf{X} \sim F(\mathbf{x}, \boldsymbol{\theta}, \mu)$, $\boldsymbol{\theta} \in \Theta$. Una familia $\{\mathcal{S}(\mathbf{X})\}$ de regiones de confianza de nivel $1-\alpha$ se dice (*insesgada*) si

$$P_{\theta,\mu}\left(\theta \in \mathcal{S}(\mathbf{X})\right) = 1-\alpha \quad \forall \theta, \mu$$

$$P_{\theta_1,\mu}\left(\theta \in \mathcal{S}(\mathbf{X})\right) \leq 1-\alpha \quad \forall \theta \neq \theta_1, \forall \mu$$

Optimalidad de Regiones de confianza

Definición Una familia $\{\mathcal{S}(\mathbf{X})\}$ de regiones de confianza de nivel $1-\alpha$ se dice IUMA (*uniformemente más exacta entre las insesgadas*) si

- $\{\mathcal{S}(\mathbf{X})\}$ es insesgada
- $\{\mathcal{S}(\mathbf{X})\}$ tiene nivel $1-\alpha$
- dada $\{\mathcal{S}^\star(\mathbf{X})\}$ insesgada de nivel $1-\alpha$ se tiene

$$P_{\theta_1}\left(\theta \in \mathcal{S}(\mathbf{X})\right) \leq P_{\theta_1}\left(\theta \in \mathcal{S}^\star(\mathbf{X})\right) \quad \theta, \theta_1$$

Optimalidad de Regiones de confianza

Definición Una familia $\{\mathcal{S}(\mathbf{X})\}$ de regiones de confianza de nivel $1-\alpha$ se dice IUMA (*uniformemente más exacta entre las insesgadas*) si

- $\{\mathcal{S}(\mathbf{X})\}$ es insesgada
- $\{\mathcal{S}(\mathbf{X})\}$ tiene nivel $1-\alpha$
- dada $\{\mathcal{S}^\star(\mathbf{X})\}$ insesgada de nivel $1-\alpha$ se tiene

$$P_{\theta_1}\left(\theta \in \mathcal{S}(\mathbf{X})\right) \leq P_{\theta_1}\left(\theta \in \mathcal{S}^\star(\mathbf{X})\right) \quad \theta, \theta_1$$

Teorema Sea $\mathcal{A}(\boldsymbol{\theta}_0)$ Región de aceptación de $\phi_{\boldsymbol{\theta}_0}$, test IUMP de nivel α para

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

$\mathcal{S}(\mathbf{X}) = \{\boldsymbol{\theta} : \phi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\} = \{\boldsymbol{\theta} : \mathbf{X} \in \mathcal{A}(\boldsymbol{\theta})\}$

$\Longrightarrow \mathcal{S}(\mathbf{X})$ es IUMA de nivel $1-\alpha$.

Bootstrap percentil

- Sean $X_1, \ldots, X_n$ i.i.d. $X_i \sim F$. Nos interesa conocer $\theta = T(F)$.

- Sea F_n la empírica asociada a $X_1, \ldots, X_n$ y $\widehat{\theta}_n = T(F_n)$ un estimador de θ.

- Sea $X_1^\star, \ldots, X_n^\star$ una muestra bootstrap obtenida a partir de F_n, o sea, X_i son i.i.d., $X_i^\star \sim F_n$.

Bootstrap percentil

- Sean $X_1, \ldots, X_n$ i.i.d. $X_i \sim F$. Nos interesa conocer $\theta = T(F)$.
- Sea F_n la empírica asociada a $X_1, \ldots, X_n$ y $\widehat{\theta}_n = T(F_n)$ un estimador de θ.
- Sea $X_1^\star, \ldots, X_n^\star$ una muestra bootstrap obtenida a partir de F_n, o sea, X_i son i.i.d., $X_i^\star \sim F_n$.
- Sea $F_n^\star$ la empírica asociada a $X_1^\star, \ldots, X_n^\star$.
- $\widehat{\theta}_n^\star = T(F_n^\star)$, la réplica bootstrap de $\widehat{\theta}_n$.

Bootstrap percentil

Sea

$$K_B(x) = \mathbb{P}_\star\left(\widehat{\theta}_n^\star \leq x\right)$$

donde $\mathbb{P}_\star$ indica la distribución de $\mathbf{X}^\star = (X_1^\star, \ldots, X_n^\star)$ condicional a $\mathbf{X} = (X_1, \ldots, X_n)$.

Definición. Sean

$$\underline{\theta}_{\mathrm{BP},\alpha} = K_B^{-1}(\alpha) \qquad \overline{\theta}_{\mathrm{BP},\alpha} = K_B^{-1}(1-\alpha) = \underline{\theta}_{\mathrm{BP},1-\alpha}$$

entonces el método percentil consiste en tomar

$$IC_{\mathrm{BP}}(1-2\alpha) = [\underline{\theta}_{\mathrm{BP},\alpha}, \overline{\theta}_{\mathrm{BP},\alpha}]$$

como intervalo de confianza para θ de nivel aproximado $1-2\alpha$.

Este método es correcto si

$$K_B(\widehat{\theta}_n) = \mathbb{P}_\star\left(\widehat{\theta}_n^\star \leq \widehat{\theta}_n\right) = \frac{1}{2}\,. \qquad (1)$$

Bootstrap percentil

Definamos

- $G_n(u) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq u)$
- $\widehat{G}_B(u) = \mathbb{P}_\star(\sqrt{n}(\widehat{\theta}^\star_n - \widehat{\theta}_n) \leq u)$
- Dadas dos distribuciones F y G, $\rho(F, G) = \sup_u |F(u) - G(u)|$

Bootstrap percentil

Teorema. Supongamos que

existe $h_n : \mathbb{R} \to \mathbb{R}$ monótona tal que

$$\Psi(x) = \mathbb{P}_F \left(h_n(\widehat{\theta}_n) - h_n(\theta) \leq x \right) \tag{2}$$

es una función de distribución continua, estrictamente creciente y simétrica alrededor de 0 para toda F (incluyendo $F = F_n$) y donde $\theta = T(F)$ y $\widehat{\theta}_n = T(F_n)$ siendo F_n la empírica asociada a $X_1, \ldots, X_n$ cuando $X_i \sim F$.

Bootstrap percentil

Teorema. Supongamos que

existe $h_n : \mathbb{R} \to \mathbb{R}$ monótona tal que

$$\Psi(x) = \mathbb{P}_F \left(h_n(\widehat{\theta}_n) - h_n(\theta) \leq x \right) \tag{2}$$

es una función de distribución continua, estrictamente creciente y simétrica alrededor de 0 para toda F (incluyendo $F = F_n$) y donde $\theta = T(F)$ y $\widehat{\theta}_n = T(F_n)$ siendo F_n la empírica asociada a $X_1, \ldots, X_n$ cuando $X_i \sim F$.

entonces, el intervalo tiene nivel exacto $1 - \alpha$.

Bootstrap percentil

Teorema. Supongamos que

a) existe $h_n : \mathbb{R} \to \mathbb{R}$ monótona tal que

$$\mathbb{P}_F \left(h_n(\widehat{\theta}_n) - h_n(\theta) \leq x \right) = \Psi(x) + o(1)$$

es una función de distribución continua, estrictamente creciente y simétrica alrededor de 0 para toda F (incluyendo $F = F_n$) y donde $\theta = T(F)$ y $\widehat{\theta}_n = T(F_n)$ siendo F_n la empírica asociada a $X_1, \ldots, X_n$ cuando $X_i \sim F$.

Bootstrap percentil

Teorema. Supongamos que

a) existe $h_n : \mathbb{R} \to \mathbb{R}$ monótona tal que

$$\mathbb{P}_F\left(h_n(\widehat{\theta}_n) - h_n(\theta) \leq x\right) = \Psi(x) + o(1)$$

es una función de distribución continua, estrictamente creciente y simétrica alrededor de 0 para toda F (incluyendo $F = F_n$) y donde $\theta = T(F)$ y $\widehat{\theta}_n = T(F_n)$ siendo F_n la empírica asociada a $X_1, \ldots, X_n$ cuando $X_i \sim F$.

b) Existe una distribución continua, estrictamente creciente y simétrica G tal que $\rho(G_n, G) \to 0$

c) $\rho(\widehat{G}_B, G_n) \xrightarrow{p} 0$.

Bootstrap percentil

Entonces, el intervalo

$$IC_{\mathrm{BP}}(1-2\alpha) = [\underline{\theta}_{\mathrm{BP},\alpha}, \overline{\theta}_{\mathrm{BP},\alpha}]$$

tiene nivel de confianza asintótico $1-2\alpha$, es decir,

$$\lim_{n\to\infty} \mathbb{P}_\theta \left(\underline{\theta}_{\mathrm{BP},\alpha} \leq \theta \leq \overline{\theta}_{\mathrm{BP},\alpha}\right) = 1-2\alpha \qquad \forall \theta \in \Theta$$

Más aún, si $\psi_\alpha = \Psi^{-1}(\alpha) = -\Psi^{-1}(1-\alpha)$

$$\underline{\theta}_{\mathrm{BP},\alpha} = h_n^{-1}\left(h_n(\widehat{\theta}_n) + \psi_\alpha\right)$$

Si $\Psi = \Phi$, h_n se llama la transformación normalizadora y estabilizadora de varianza.

La condición $\rho(\widehat{G}_B, G_n) \xrightarrow{p} 0$ se cumple si, por ejemplo, el funcional T es diferenciable Hadamard en F_θ

Bootstrap percentil corregido por sesgo

Observemos que

- si se cumple (1), $K_B(\widehat{\theta}_n) = 1/2$, luego
  $z_0 = \Phi^{-1}\left(K_B(\widehat{\theta}_n)\right) = 0$

- si se cumple (2),

$$\begin{aligned} \frac{1}{2} &= \Psi(0) = \mathbb{P}_\star\left(h_n(\widehat{\theta}_n^\star) - h_n(\widehat{\theta}_n) \leq 0\right) \\ &= \mathbb{P}_\star\left(h_n(\widehat{\theta}_n^\star) \leq h_n(\widehat{\theta}_n)\right) \\ &= \mathbb{P}_\star\left(\widehat{\theta}_n^\star \leq \widehat{\theta}_n\right) = K_B(\widehat{\theta}_n) \end{aligned}$$

  Por lo tanto, $z_0 = \Psi^{-1}\left(K_B(\widehat{\theta}_n)\right) = 0$

Luego, una manera de medir el sesgo es través de

$$z_0 = \Psi^{-1}\left(K_B(\widehat{\theta}_n)\right)$$

Bootstrap percentil corregido por sesgo

Supongamos que existe $h_n : \mathbb{R} \to \mathbb{R}$ monótona tal que

$$\mathbb{P}_F \left(h_n(\widehat{\theta}_n) - h_n(\theta) + z_0 \leq x \right) = \Psi(x) + o(1) \qquad (3)$$

es una función de distribución continua, estrictamente creciente y simétrica alrededor de 0 para toda F (incluyendo $F = F_n$) y donde

- $\theta = T(F)$
- $\widehat{\theta}_n = T(F_n)$ siendo F_n la empírica asociada a $X_1, \ldots, X_n$ cuando $X_i \sim F$ y
- $z_0 = \Psi^{-1}\left(K_B(\widehat{\theta}_n)\right)$.

Sea $\psi_\alpha = \Psi^{-1}(\alpha)$ y definamos

$$\begin{aligned} \underline{\theta}_{\mathrm{BC},\alpha} &= h_n^{-1}\left(h_n(\widehat{\theta}_n) + z_0 + \psi_\alpha\right) = K_B^{-1}\left(\Psi(\psi_\alpha + 2z_0)\right) \\ \overline{\theta}_{\mathrm{BC},\alpha} &= h_n^{-1}\left(h_n(\widehat{\theta}_n) + z_0 + \psi_{1-\alpha}\right) = K_B^{-1}\left(\Psi(\psi_\alpha + 2z_0)\right) \end{aligned}$$

Bootstrap percentil corregido por sesgo

Si Ψ es una función conocida como, por ejemplo, $\Psi = \Phi$, el intervalo

$$IC_{\mathrm{BC}}(1-2\alpha) = [\underline{\theta}_{\mathrm{BC},\alpha}, \overline{\theta}_{\mathrm{BC},\alpha}]$$

se llama intervalo bootstrap con corrección por sesgo y tiene nivel de confianza asintótico $1-2\alpha$, es decir,

$$\mathbb{P}_\theta\left(\underline{\theta}_{\mathrm{BC},\alpha} \leq \theta \leq \overline{\theta}_{\mathrm{BC},\alpha}\right) \to 1-2\alpha \qquad \forall \theta \in \Theta$$

si

- se cumple (3)
- Existe una distribución continua, estrictamente creciente y simétrica G tal que $\rho(G_n, G) \to 0$
- $\rho(\widehat{G}_B, G_n) \xrightarrow{p} 0$.

Bootstrap hibrido

- $G_n(u) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq u)$
- $\widehat{G}_B(u) = \mathbb{P}_\star(\sqrt{n}(\widehat{\theta}^\star_n - \widehat{\theta}_n) \leq u) = K_B\left(\widehat{\theta}_n + u\, n^{-1/2}\right)$

Luego,

$$\mathbb{P}\left(\widehat{\theta}_n - n^{-1/2} G_n^{-1}(1-\alpha) \leq \theta \leq \widehat{\theta}_n - n^{-1/2} G_n^{-1}(\alpha)\right) = 1 - 2\alpha$$

El problema es que $G_n^{-1}(1-\alpha)$ y $G_n^{-1}(\alpha)$ no se conocen en general.

Por eso los aproximaremos por $\widehat{G}_B^{-1}(1-\alpha)$ y $\widehat{G}_B^{-1}(\alpha)$, respectivamente ya que si $\rho(\widehat{G}_B, G_n) \xrightarrow{p} 0$, se cumple que $\widehat{G}_B^{-1}(u) - G_n^{-1}(u) \xrightarrow{p} 0$.

Bootstrap hibrido

Usando que $\widehat{G}_B^{-1}(u) - G_n^{-1}(u) \xrightarrow{p} 0$ para todo $u \in [0,1]$, podemos definir el intervalo bootstrap hibrido como

$$IC_{\mathrm{HB}}(1-2\alpha) = [\underline{\theta}_{\mathrm{HB},\alpha}, \overline{\theta}_{\mathrm{HB},\alpha}]$$

donde

$$\underline{\theta}_{\mathrm{HB},\alpha} = \widehat{\theta}_n - n^{-1/2}\widehat{G}_B^{-1}(1-\alpha) \qquad \overline{\theta}_{\mathrm{HB},\alpha} = \widehat{\theta}_n - n^{-1/2}\widehat{G}_B^{-1}(\alpha)$$

El intervalo bootstrap hibrido tiene nivel de confianza asintótico $1-2\alpha$ si $\rho(\widehat{G}_B, G_n) \xrightarrow{p} 0$.

Bootstrap-t

Supongamos que

- $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, \sigma_F^2)$, o sea, $G_n \xrightarrow{w} G$ donde $G = \Phi(\cdot/\sigma_F)$.
- $\widehat{\sigma}_F \xrightarrow{p} \sigma_F$

Definamos el estadístico *studentizado*

$$t(\mathbf{X}, \theta) = \sqrt{n}\frac{\widehat{\theta}_n - \theta}{\widehat{\sigma}_F}$$

Sean

- $G_{n,t}(u) = \mathbb{P}(t(\mathbf{X}, \theta) \leq u)$
- $\widehat{G}_{B,t}(u) = \mathbb{P}_\star\left(t(\mathbf{X}^\star, \widehat{\theta}) \leq u\right)$

Bootstrap-t

Si $G_{n,t}$ fuera conocida podríamos usar el método del pivote para dar un intervalo de confianza de nivel $1-2\alpha$ para θ como

$$[\widehat{\theta}_n - n^{-1/2}\,\widehat{\sigma}_F\,G_{n,t}^{-1}(1-\alpha), \widehat{\theta}_n - n^{-1/2}\,\widehat{\sigma}_F\,G_{n,t}^{-1}(\alpha)]$$

Como no conozco $G_{n,t}$, aproximo $G_{n,t}^{-1}(1-\alpha)$ y $G_{n,t}^{-1}(\alpha)$ por $\widehat{G}_{B,t}^{-1}(1-\alpha)$ y $\widehat{G}_{B,t}^{-1}(\alpha)$, respectivamente.

Se define el intervalo bootstrap$-t$ como

$$IC_{\mathrm{BT}}(1-2\alpha) = [\underline{\theta}_{\mathrm{BT},\alpha}, \overline{\theta}_{\mathrm{BT},\alpha}]$$

donde

$$\underline{\theta}_{\mathrm{BT},\alpha} = \widehat{\theta}_n - n^{-1/2}\,\widehat{\sigma}_F\,\widehat{G}_{B,t}^{-1}(1-\alpha) \qquad \overline{\theta}_{\mathrm{BT},\alpha} = \widehat{\theta}_n - n^{-1/2}\,\widehat{\sigma}_F\,\widehat{G}_{B,t}^{-1}(\alpha)$$

Bootstrap-t

- El intervalo bootstrap$-t$ tiene nivel de confianza asintótico $1-2\alpha$ si $\rho(\widehat{G}_{B,t}, G_{n,t}) \xrightarrow{p} 0$.

- El intervalo $IC_{\mathrm{BT}}(1-2\alpha)$ es más preciso en general que los intervalos $IC_{\mathrm{BP}}(1-2\alpha)$, $IC_{\mathrm{BC}}(1-2\alpha)$ o $IC_{\mathrm{HB}}(1-2\alpha)$ pero requiere un estimador consistente de la varianza asintótica.

- Una razón por la cual $IC_{\mathrm{BT}}(1-2\alpha)$ es mejor que $IC_{\mathrm{HB}}(1-2\alpha)$ es que $t(\mathbf{X},\theta)\sqrt{n}(\widehat{\theta}_n-\theta)/\widehat{\sigma}_F$ es más pivotal que $\sqrt{n}(\widehat{\theta}_n-\theta)$ en el sentido que la distribución de $t(\mathbf{X},\theta)$ es menos dependiente de F que la de $\sqrt{n}(\widehat{\theta}_n-\theta)$.