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 $1 - \alpha$.

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 $\boldsymbol{\theta}$.

Teorema 1. 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 $\boldsymbol{\theta}$.

Intervalos para la media μ con varianza conocida

$X_1, \ldots, X_n$ i.i.d. $X_i \sim N(\mu, \sigma_0^2)$, σ_0^2 **conocida.**

Buscando un pivote....

Intervalos para la media μ con varianza conocida

$X_1, \ldots, X_n$ i.i.d. $X_i \sim N(\mu, \sigma_0^2)$, σ_0^2 **conocida.**

Buscando un pivote....

$$U = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma_0} \sim N(0,1)$$

Intervalos para la media μ con varianza conocida

$X_1, \ldots, X_n$ i.i.d. $X_i \sim N(\mu, \sigma_0^2)$, σ_0^2 **conocida.**

Buscando un pivote....

$$U = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma_0} \sim N(0,1)$$

Luego,

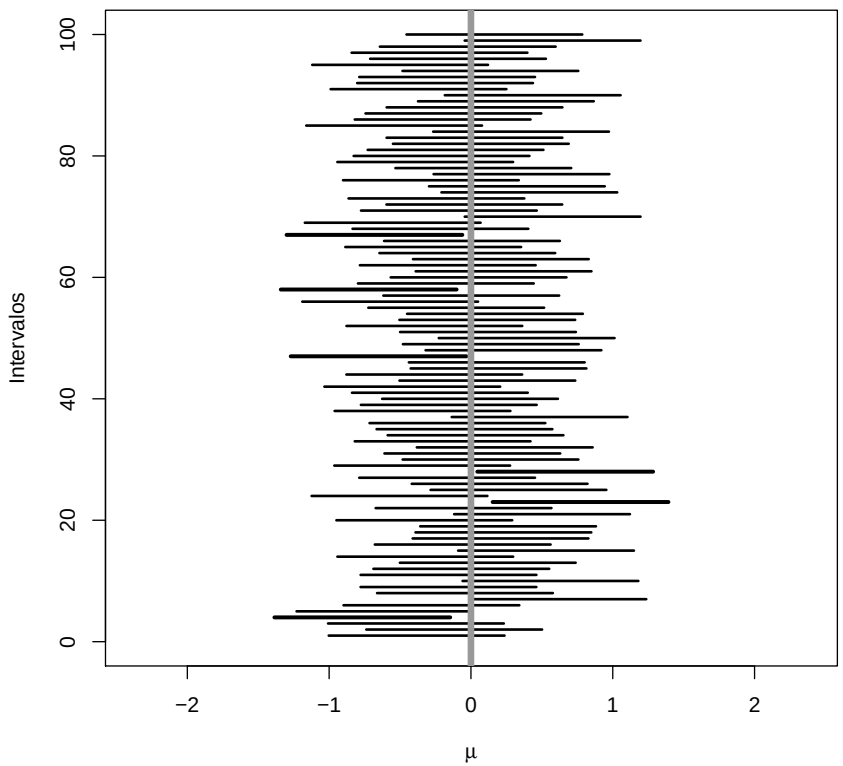
$$\left[\overline{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}, \overline{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right]$$

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

Intervalos para la media μ con varianza conocida

Para $1 \leq j \leq N = 100$,

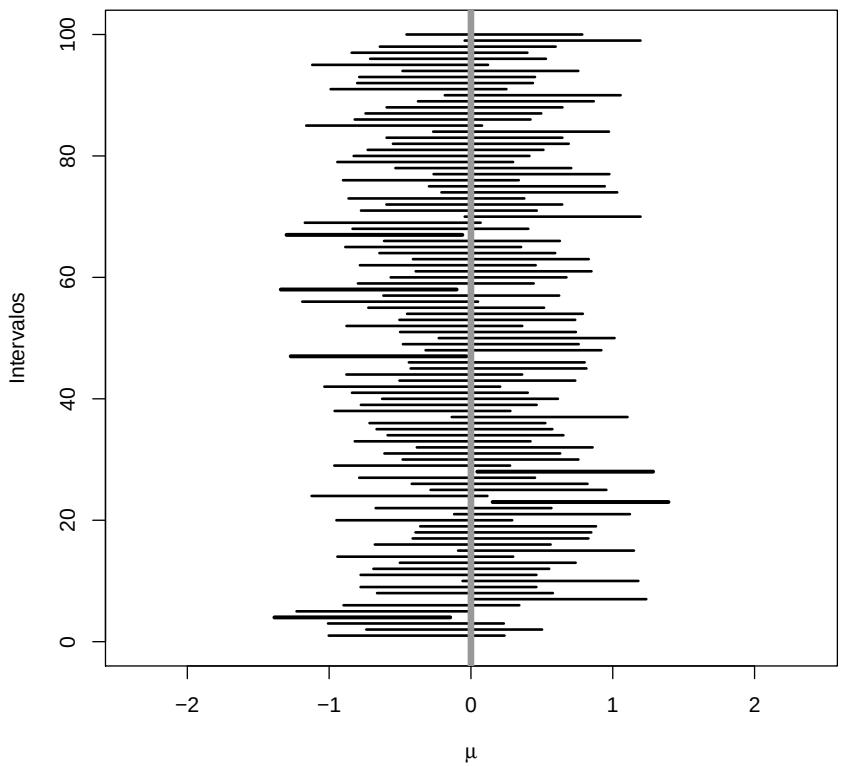
- Generamos X_1, X_n, $n = 10$, i.i.d. $X_i \sim N(\mu, 1)$
- Graficamos en cada caso los IC para μ de nivel 0.95

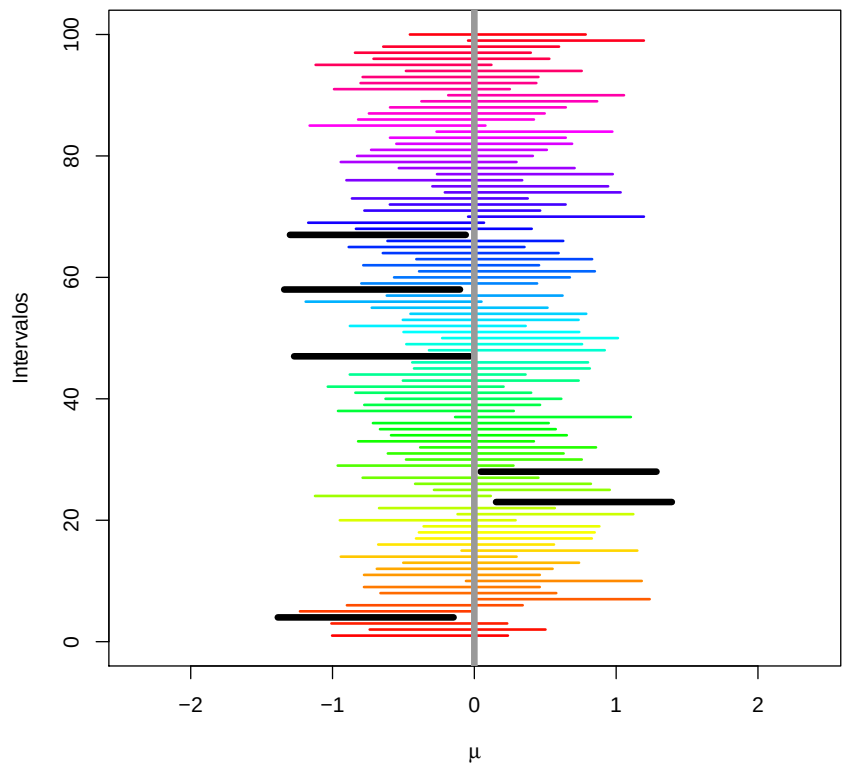


Intervalos para la media μ con varianza conocida

Para $1 \leq j \leq N = 100$,

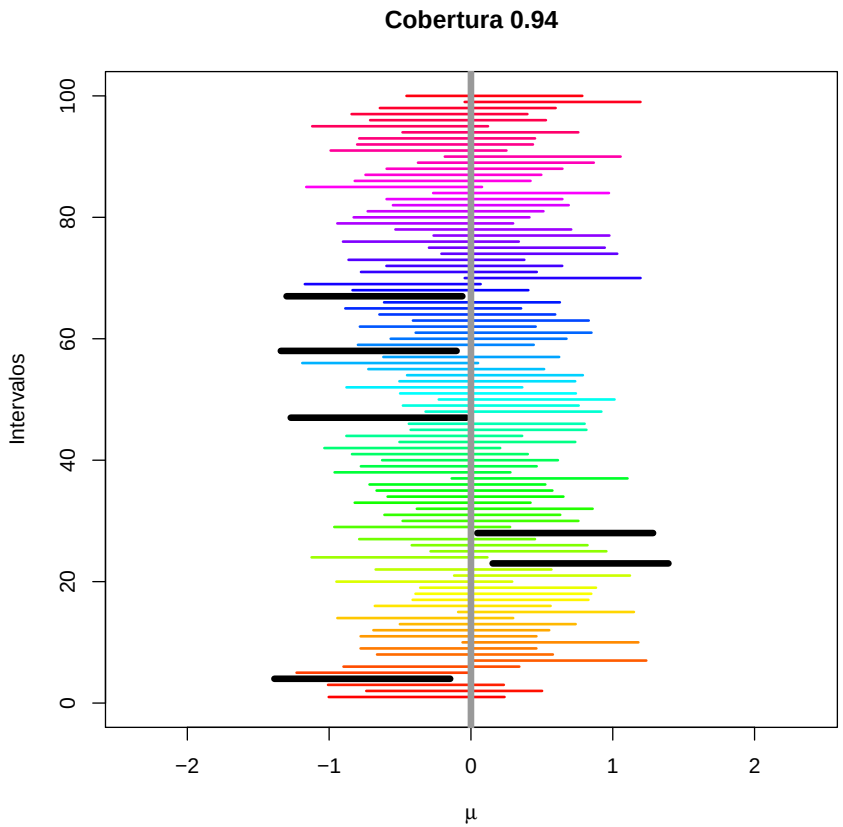
- Generamos X_1, X_n, $n = 10$, i.i.d. $X_i \sim N(\mu, 1)$
- Graficamos en cada caso los IC para μ de nivel 0.95

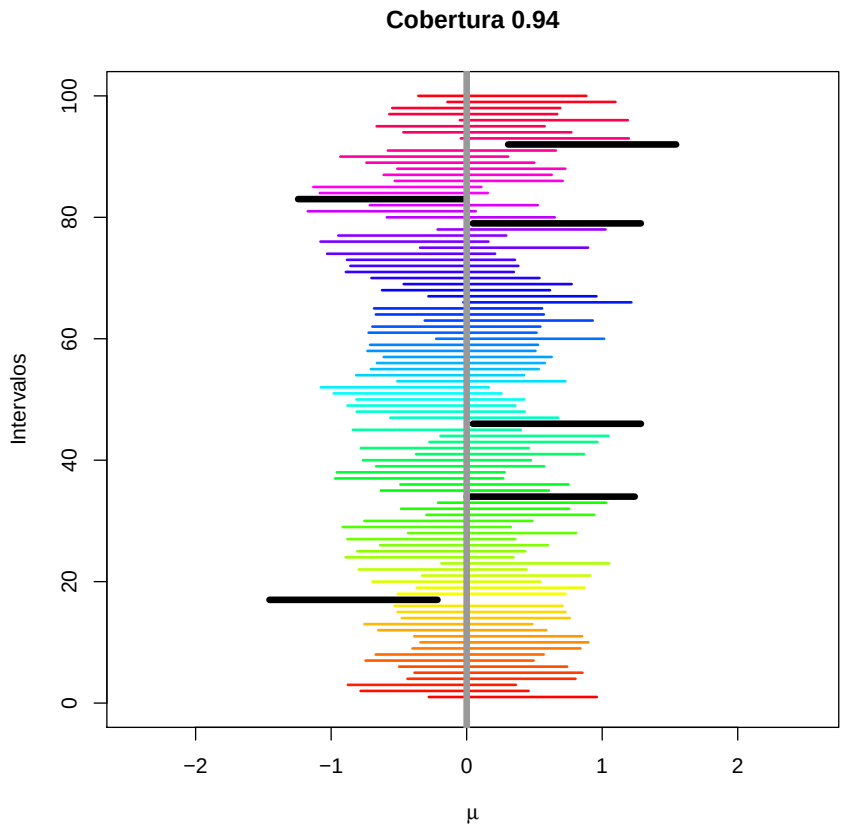




Intervalos para la media μ con varianza conocida

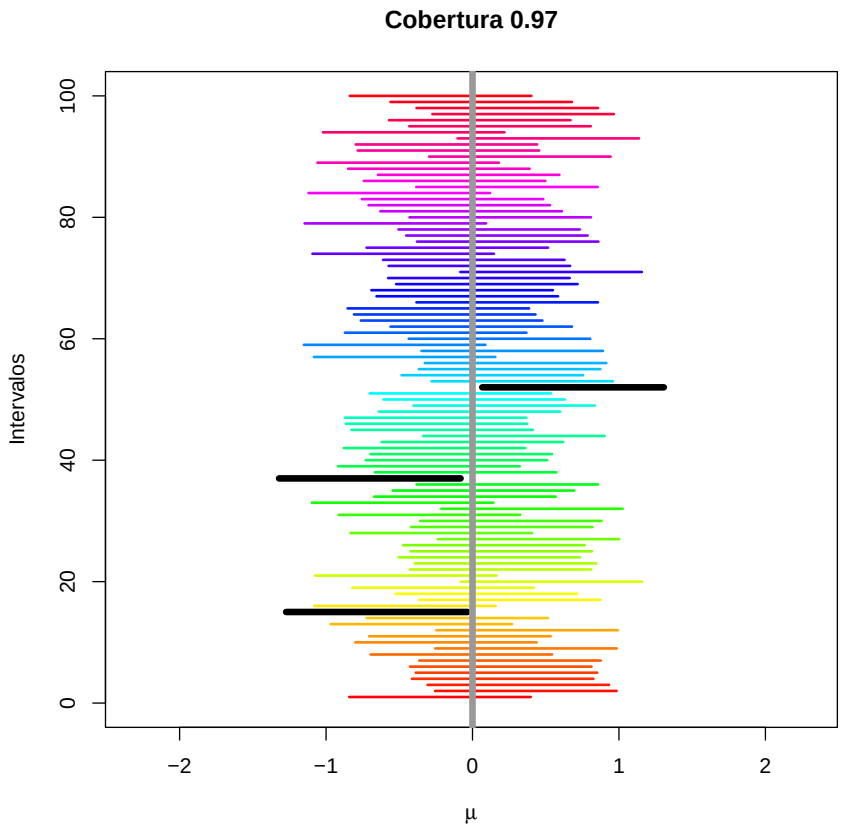
Lo hacemos con varias semillas...

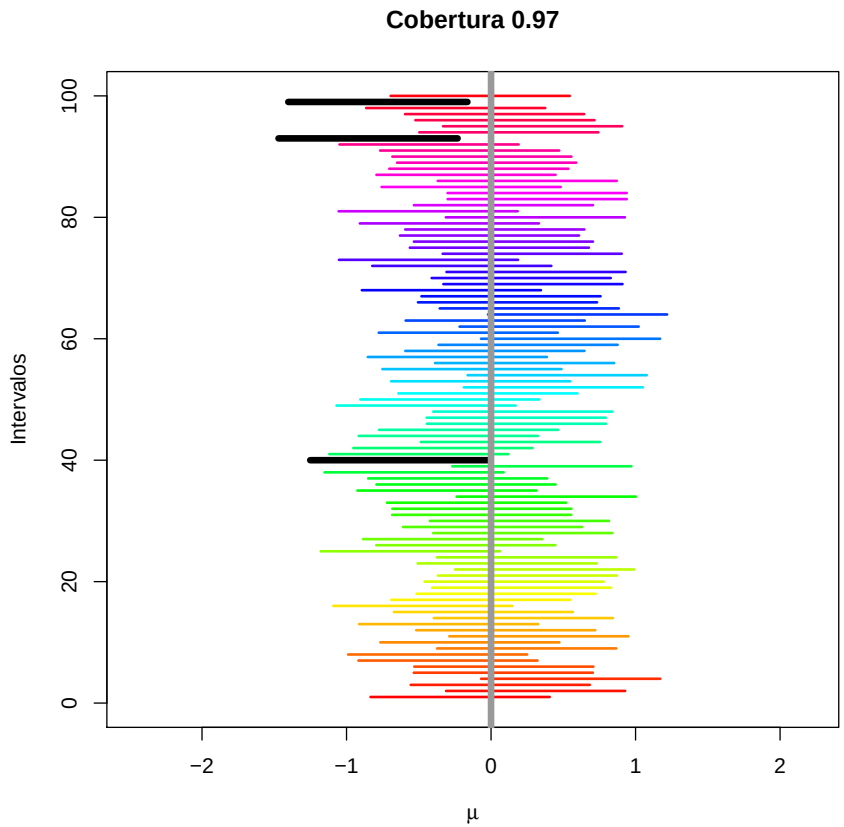




Intervalos para la media μ con varianza conocida

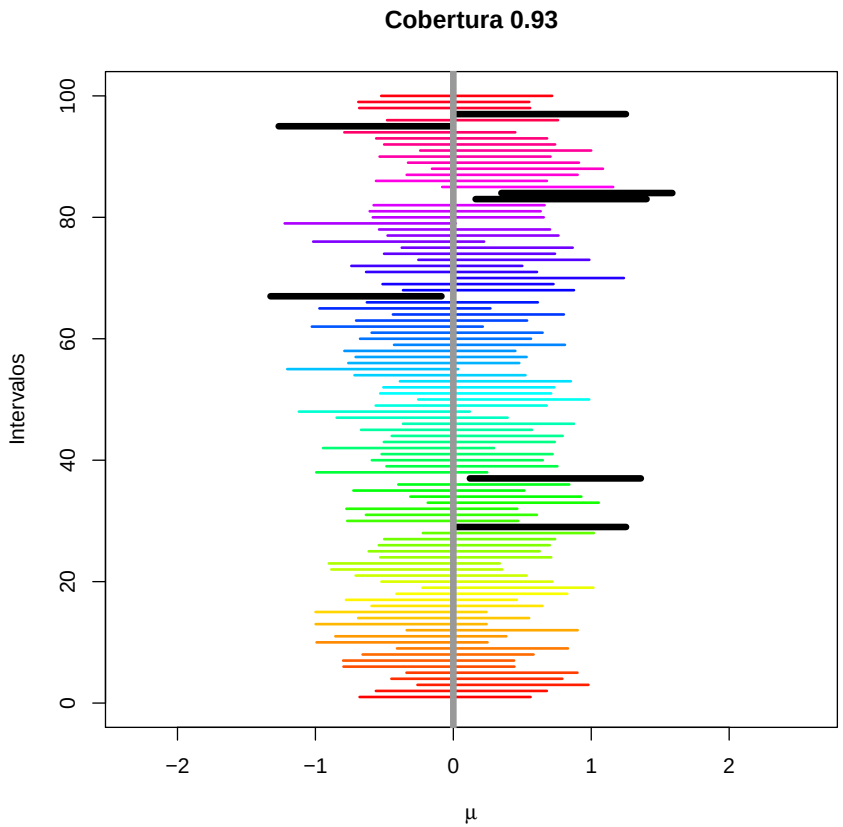
Lo hacemos con varias semillas...

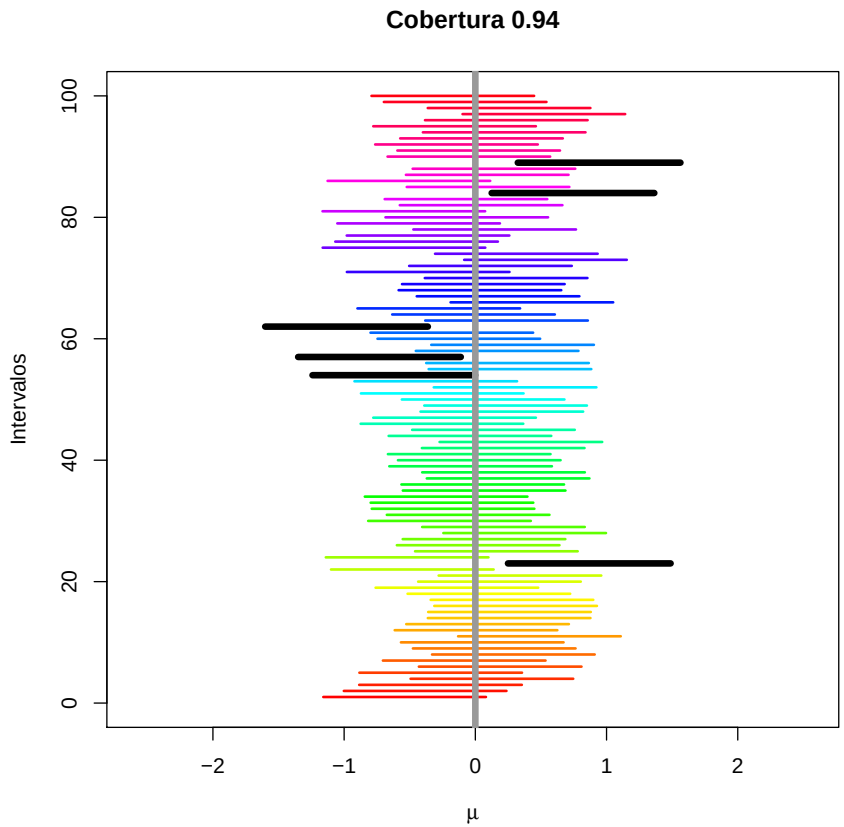




Intervalos para la media μ con varianza conocida

Lo hacemos con varias semillas...





Intervalos para la media μ y la varianza de una normal

Buscando un pivote....

Intervalos para la media μ y la varianza de una normal

Buscando un pivote....

Proposición. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes donde $X_i \sim N(\mu, \sigma^2)$. Luego*

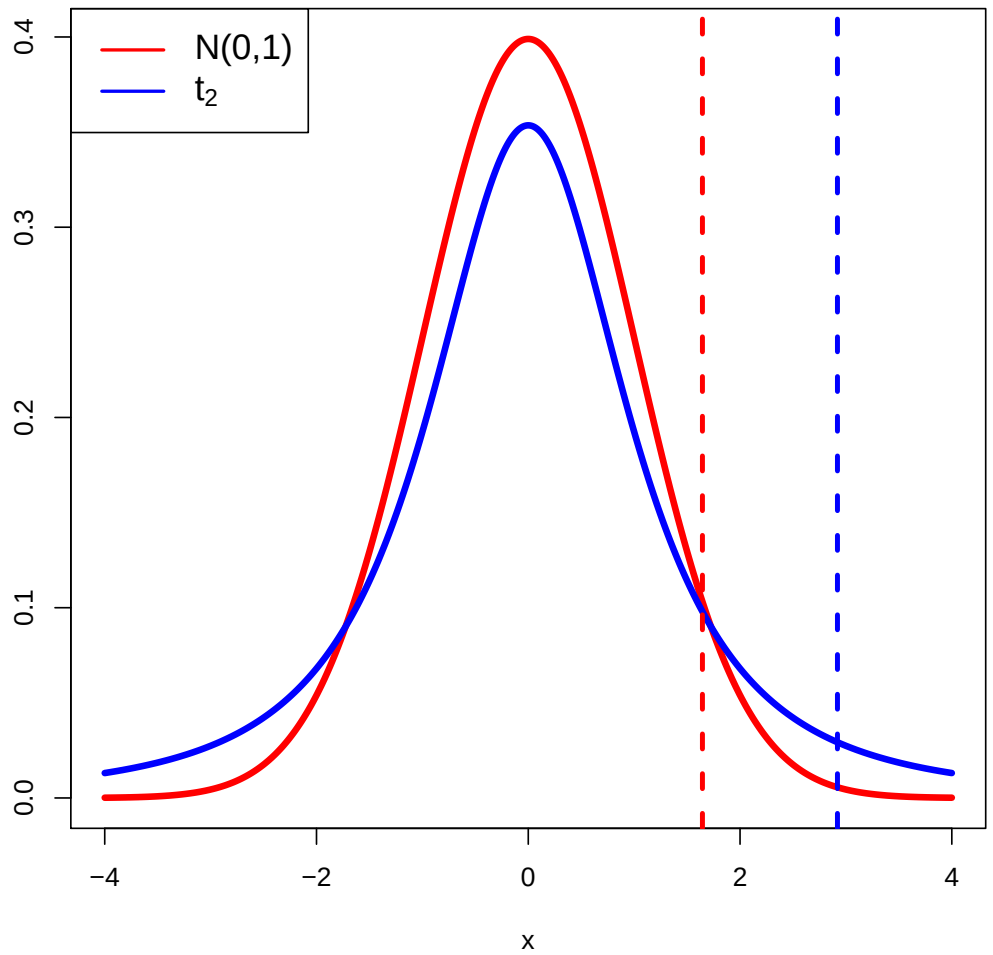
a) $U = \sqrt{n}\dfrac{\overline{X}_n - \mu}{\sigma} \sim N(0,1)$, *con* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

b) $s_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ *es independiente de* $\overline{X}_n$.

c) $V = \dfrac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

d) $W = \sqrt{n}\dfrac{(\overline{X}_n - \mu)}{s_n} \sim \mathcal{T}_{n-1}$.

Normal vs. $\mathcal{T}_2$: mirando percentiles...



Las líneas verticales corresponden al percentil 0.95

Intervalo para la media μ de una normal con varianza desconocida

Sea T una v.a. tal que $T \sim \mathcal{T}_m$. Dado $0 < \eta < 1$ llamamos $t_{m,\eta}$ al punto que satisface:

$$\mathbb{P}(V > t_{m,\eta}) = \eta$$

Intervalo para la media μ de una normal con varianza desconocida

Sea T una v.a. tal que $T \sim \mathcal{T}_m$. Dado $0 < \eta < 1$ llamamos $t_{m,\eta}$ al punto que satisface:

$$\mathbb{P}(V > t_{m,\eta}) = \eta$$

Teorema 2. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes donde $X_i \sim N(\mu, \sigma^2)$ con μ y σ desconocidas. Luego, si*

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

tenemos que

$$\left[\overline{X}_n - t_{n-1,\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}, \overline{X}_n + t_{n-1,\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right]$$

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

Intervalo para la varianza σ^2 de una normal

Sea V una v.a. tal que $V \sim \chi^2_m$. Dado $0 < \eta < 1$ llamamos $\chi^2_{m,\eta}$ al punto que satisface:

$$\mathbb{P}(V > \chi^2_{m,\eta}) = \eta$$

Intervalo para la varianza σ^2 de una normal

Sea V una v.a. tal que $V \sim \chi^2_m$. Dado $0 < \eta < 1$ llamamos $\chi^2_{m,\eta}$ al punto que satisface:

$$\mathbb{P}(V > \chi^2_{m,\eta}) = \eta$$

Teorema 3. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes donde $X_i \sim N(\mu, \sigma^2)$ con σ desconocida. Luego, sean β y γ son dos números positivos tales que $\beta + \gamma = \alpha$.*

a) *[Ejercicio] Si μ es conocido un intervalo de confianza de nivel $1 - \alpha$ para σ^2 es*

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,\beta}}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,1-\gamma}} \right).$$

b) *Si μ es desconocido un intervalo de confianza de nivel $1 - \alpha$ para σ^2 es*

$$\left(\frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\chi^2_{n-1,\beta}}, \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{\chi^2_{n-1,1-\gamma}} \right).$$

Mundo Normal

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Buscamos intervalo de confianza para μ.
- σ conocido:

  IC nivel $1-\alpha$ para μ

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

- Longitud: $\rightarrow \ell = 2\, z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Mundo Normal

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Buscamos intervalo de confianza para μ.
- σ conocido:

  IC nivel $1-\alpha$ para μ

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

- Longitud: $\rightarrow \ell = 2\, z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \ell_0$

Mundo Normal

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Buscamos intervalo de confianza para μ.
- σ conocido:

  IC nivel $1-\alpha$ para μ

$$\left(\overline{X}_n - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad , \quad \overline{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

- Longitud: $\rightarrow \ell = 2\, z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \ell_0$

- Si queremos $\ell \leq \ell_0 \Longrightarrow$

$$\frac{4 z_{\frac{\alpha}{2}}^2 \sigma^2}{\ell_0^2} \leq n$$

Mundo Normal

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Buscamos intervalo de confianza para μ.
- σ desconocido .....llegó la $\mathcal{T}$...
- IC nivel $1-\alpha$ para μ

$$\left(\overline{X}_n - t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \quad , \quad \overline{X}_n + t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \right)$$

- Longitud: $\rightarrow L = 2\, t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}$

Mundo Normal

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. Buscamos intervalo de confianza para μ.
- σ desconocido .....llegó la $\mathcal{T}$...
- IC nivel $1-\alpha$ para μ

$$\left(\overline{X}_n - t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \quad , \quad \overline{X}_n + t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \right)$$

- Longitud: $\rightarrow L = 2\, t_{n-1,\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}$

- L es una v.a. Si queremos que $L \leq \ell_0$ ¿Cómo hacemos?

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

$$\frac{2 s_n t_{\frac{\alpha}{2}, n-1}}{\sqrt{n+m}} \leq \ell_0$$

- m es una variable aleatoria que depende sólo de s_n pues

$$n + m \geq \frac{4 s_n^2 t^2_{\frac{\alpha}{2}, n-1}}{\ell_0^2}$$

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

- Sea $X_{n+1}, \ldots, X_{n+m}$ una muestra complementaria, independiente de la anterior, 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_0 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 $N(\mu, \sigma^2)$

Teorema 3. *Sean $X_1, \cdots, X_n, \cdots$ variables aleatorias independientes con distribución $N(\mu, \sigma^2)$ 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_0.

Intervalo de confianza para diferencia de medias

Sean $X_1, \cdots, X_{n_1}$ $X_i \sim \mathrm{N}(\mu_1, \sigma^2)$ y $Y_1, \cdots, Y_{n_2}$ $Y_i \sim \mathrm{N}(\mu_2, \sigma^2)$, 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

Teorema 4.

(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,

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

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 $(1 - \alpha)$.

Ejemplos

- Consideremos el caso en que $X_1, \ldots, X_n$ son una muestra aleatoria $\mathbb{E}(X_i) = \mu$ y $\mathbb{V}(X_i) = \sigma^2$, ambas desconocidas.

- El intervalo

$$\left[\overline{X}_n - z_{\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}, \overline{X}_n + z_{\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}\right]$$

tiene nivel asintótico $1 - \alpha$.

Ejemplos

Consideremos el caso en que $X_1, \ldots, X_n$ son una muestra aleatoria donde $X_i \sim Bi(1, p)$

Vamos a deducir dos intervalos de confianza asintóticos diferentes para p.

- El intervalo

$$[\widehat{p}_{1,n}, \widehat{p}_{2,n}]$$

donde $\widehat{p}_{1,n} \leq \widehat{p}_{2,n}$ son las raíces del polinomio en p

$$n\overline{X}_n^2 - p(2n\overline{X}_n + z^2_{\frac{\alpha}{2}}) + p^2(z^2_{\frac{\alpha}{2}} + n)$$

tiene nivel de confianza asintótico $1 - \alpha$.

Ejemplos

Consideremos el caso en que $X_1, \ldots, X_n$ son una muestra aleatoria donde $X_i \sim Bi(1, p)$

Vamos a deducir dos intervalos de confianza asintóticos diferentes para p.

- El intervalo

$$[\widehat{p}_{1,n}, \widehat{p}_{2,n}]$$

donde $\widehat{p}_{1,n} \leq \widehat{p}_{2,n}$ son las raíces del polinomio en p

$$n\overline{X}_n^2 - p(2n\overline{X}_n + z^2_{\frac{\alpha}{2}}) + p^2(z^2_{\frac{\alpha}{2}} + n)$$

tiene nivel de confianza asintótico $1 - \alpha$.

- El intervalo

$$\left[\overline{X}_n - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}_n(1-\overline{X}_n)}{n}}, \bar{X}_n + z_{\frac{\alpha}{2}}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

también tiene nivel de confianza asintótico $1 - \alpha$.

Usando la distribución asintótica de los EMV

$X_1, \ldots, X_n$ i.i.d. donde X_i tienen función de densidad o de probabilidad puntual $f(x, \theta)$.

Bajo condiciones de regularidad

- si $\widehat{\theta}_n = \widehat{\theta}_n^{\mathrm{MV}}$, entonces $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, \frac{1}{I_1(\theta)})$
- $$\sqrt{n}\sqrt{I_1(\theta)}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, 1)$$
- La región

$$S(\mathbf{X}) = \{\theta : -z_{\frac{\alpha}{2}} \leq \sqrt{n}\sqrt{I_1(\theta)}(\widehat{\theta}_n - \theta) \leq z_{\frac{\alpha}{2}}\}$$

no tiene porqué ser un intervalo y puede ser difícil de calcular.

Usando la distribución asintótica de los EMV

- Si $I_1(\theta)$ es una función continua de θ, como $\widehat{\theta}_n \xrightarrow{p} \theta$, bajo condiciones de regularidad obtendremos que $I_1(\widehat{\theta}_n) \xrightarrow{p} I_1(\theta)$, luego

$$\sqrt{n}\sqrt{I_1(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0,1)$$

- Entonces, un intervalo de nivel asintótico $1-\alpha$ será

$$\left[\widehat{\theta}_n - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n\, I_1(\widehat{\theta}_n)}}, \widehat{\theta}_n + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n\, I_1(\widehat{\theta}_n)}}\right]$$

$X_1, \ldots, X_n$ i.i.d. $X_i \sim F$

En muchos casos, el parámetro de interés puede escribirse como

$$\theta = T(F)$$

y el estimador (plug-in) resulta

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

donde $\widehat{F}_n$ es la empírica.

Intervalo de confianza para la media

- $\mu := T(F) = \mathbb{E}_F(X)$. Estimador plug-in:

$$\widehat{\mu}_n = T(\widehat{F}_n) = \mathbb{E}_{\widehat{F}_n}(X) = \overline{X}_n$$

- Distribución de $\widehat{\mu}_n$: asintóticamente normal

$$\frac{\widehat{\mu}_n - \mu}{\text{se}(\widehat{\mu}_n)} \approx N(0,1) \qquad n \text{ grande}$$

- Desvío del Estimador:

$$\text{se}(\widehat{\mu}_n) = \sqrt{\mathbb{V}_F(\widehat{\mu}_n)} = \sqrt{\frac{\sigma^2}{n}} = \text{se}$$

se estima con $\widehat{\text{se}} = \sqrt{\frac{\widehat{\sigma}^2}{n}}$ o con $\widehat{\text{se}} = \sqrt{\frac{S^2}{n}}$

Intervalo de confianza $\widehat{\mu}_n \pm z_{\frac{\alpha}{2}} \widehat{\text{se}}$

Intervalo de confianza para la mediana

- $\theta := T(F) = F^{-1}(\frac{1}{2})$. Estimador plug-in:

$$\widehat{\theta}_n = T(\widehat{F}_n) = \underset{1 \leq i \leq n}{\text{mediana}}(X_i)$$

- Distribución de $\widehat{\theta}_n$: asintóticamente normal

$$\frac{\widehat{\theta}_n - \theta}{\text{se}(\widehat{\theta}_n)} \approx N(0,1) \qquad\qquad n \text{ grande}$$

- Desvío del Estimador:

$$\text{se}(\widehat{\theta}_n) = \sqrt{\mathbb{V}_F(\widehat{\theta}_n)} = ???$$

- $\widehat{\text{se}} = ??$
- Bootstrap! $\widehat{\text{se}}_{\text{BOOT}}$

Intervalo de confianza $\widehat{\theta}_n \pm z_{\frac{\alpha}{2}} \, \widehat{\text{se}}_{\text{BOOT}}$

POBLACION $\leftrightarrow F$	MUESTRA $X_1, \ldots X_n$ i.i.d. $X_i \sim F$
Parámetro: Valor asociado de F	Estimador: estadístico para estimar θ
$\theta = \theta(F)$	$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$
θ: valor poblacional	$\widehat{\theta}_n$: NUEVA VARIABLE ALEATORIA

POBLACION $\leftrightarrow F$	MUESTRA $X_1, \ldots X_n$ i.i.d. $X_i \sim F$
Parámetro: Valor asociado de F	Estimador: estadístico para estimar θ
$\theta = \theta(F)$	$\widehat{\theta}_n = \widehat{\theta}_n(X_1, \ldots, X_n)$
θ: valor poblacional	$\widehat{\theta}_n$: NUEVA VARIABLE ALEATORIA

$$\widehat{\theta}_n = T(\widehat{F}_n)$$

x	X_1	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	X_n
$p(x)$	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$

Aproximación de la distribución de un estimador (si pudiéramos...)

$$\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(N_{\mathrm{B}})} ,$$

donde

$$\begin{array}{lll} X_1^{(1)}, \ldots, X_n^{(1)} & \text{i.i.d.} & X_i^{(1)} \sim F, \\ X_1^{(2)}, \ldots, X_n^{(2)} & \text{i.i.d.} & X_i^{(2)} \sim F \\ & \vdots & \\ X_1^{(N_{\mathrm{B}})}, \ldots, X_n^{(N_{\mathrm{B}})} & \text{i.i.d.} & X_i^{(N_{\mathrm{B}})} \sim F \end{array}$$

Aproximación de la distribución de un estimador

- $X_1 \ldots, X_n \quad X_i \sim F \qquad \widehat{\theta}_n = T(\widehat{F}_n)$

Quiero contruir $\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(N_B)}$, por simplicidad, $\widehat{\theta}_{(1)}, \ldots, \widehat{\theta}_{(N_B)}$,

Aproximación de la distribución de un estimador

- $X_1 \ldots, X_n \quad X_i \sim F \qquad \widehat{\theta}_n = T(\widehat{F}_n)$

Quiero contruir $\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(N_B)}$, por simplicidad, $\widehat{\theta}_{(1)}, \ldots, \widehat{\theta}_{(N_B)}$,

$x_1 \ldots, x_n$: datos originales, realizaciones de $X_i \sim F \longrightarrow \widehat{F}_n$

$\widehat{F}_n$: distribución construida con los datos originales que aproxima a F

Podemos ahora tomar muestras con distribución $\widehat{F}_n$ para aproximar la distribución de $\widehat{\theta}_n$

$$X_1^* \ldots, X_n^* \quad X_i^* \sim \widehat{F}_n \longrightarrow \widehat{\theta}^*$$

Aproximación de la distribución de un estimador

- $X_1 \ldots, X_n \quad X_i \sim F \qquad \widehat{\theta}_n = T(\widehat{F}_n)$

Quiero contruir $\widehat{\theta}_n^{(1)}, \ldots, \widehat{\theta}_n^{(N_{\mathrm{B}})}$, por simplicidad, $\widehat{\theta}_{(1)}, \ldots, \widehat{\theta}_{(N_{\mathrm{B}})}$,

$x_1 \ldots, x_n$: datos originales, realizaciones de $X_i \sim F \longrightarrow \widehat{F}_n$

$\widehat{F}_n$: distribución construida con los datos originales que aproxima a F

Podemos ahora tomar muestras con distribución $\widehat{F}_n$ para aproximar la distribución de $\widehat{\theta}_n$

$$X_1^* \ldots, X_n^* \quad X_i^* \sim \widehat{F}_n \longrightarrow \widehat{\theta}^*$$

Si tomo N_{B} muestras obtenemos

$$\widehat{\theta}_{(1)}^*, \ldots, \widehat{\theta}_{(N_{\mathrm{B}})}^* , \quad X_i^* \sim \widehat{F}_n$$

Bootstrap basado en la distribución asintótica

Supongamos que

- $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, \sigma_F^2)$

Si conocemos σ_F podemos construir un intervalo de confianza asintótico como

$$\widehat{\theta}_n \pm z_{\frac{\alpha}{2}} \frac{\sigma_F}{\sqrt{n}} = \widehat{\theta}_n \pm z_{\frac{\alpha}{2}} \text{se}_{\widehat{\theta}}$$

Si σ_F no es conocida o no se puede estimar fácilmente:

Necesitamos aproximar σ_F

Bootstrap basado en la distribución asintótica

Consideramos N_{B} submuestras obtenemos

$$\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{N_{\mathrm{B}}}\ , \quad X^*_i \sim \widehat{F}_n$$

Definimos

$$\widehat{\mathrm{se}}_{\mathrm{BOOT}} = \widehat{\sigma}_F = \left\{ \frac{1}{N_{\mathrm{B}}} \sum_{j=1}^{N_{\mathrm{B}}} \left(\widehat{\theta}^*_j - \frac{1}{N_{\mathrm{B}}} \sum_{\ell=1}^{N_{\mathrm{B}}} \widehat{\theta}^*_\ell \right)^2 \right\}^{\frac{1}{2}}$$

Intervalo de confianza $\widehat{\theta}_n \pm z_{\frac{\alpha}{2}}\ \widehat{\mathrm{se}}_{\mathrm{BOOT}}$

Bootstrap percentil

Consideramos la muestra boostrap

$$\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{N_{\mathrm{B}}} , \quad X^*_i \sim \widehat{F}_n$$

Definimos

- $\widehat{\theta}^*_{(\gamma)}$: percentil γ de la muestra $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{N_{\mathrm{B}}}$

El intervalo de confianza bootstrap percentil de nivel $1-\alpha$ se define como

$$\left(\widehat{\theta}^*_{\left(\frac{\alpha}{2}\right)} , \widehat{\theta}^*_{\left(1-\frac{\alpha}{2}\right)} \right)$$

Bootstrap percentil

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

Definición. Sean

$$\underline{\theta}_{\mathrm{BP},\frac{\alpha}{2}} = K_B^{-1}\left(\frac{\alpha}{2}\right) \qquad \overline{\theta}_{\mathrm{BP},\frac{\alpha}{2}} = K_B^{-1}\left(1-\frac{\alpha}{2}\right) = \underline{\theta}_{\mathrm{BP},1-\frac{\alpha}{2}}$$

entonces el método percentil consiste en tomar

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

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

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

a) 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 $F = \widehat{F}_n$) y donde $\theta = T(F)$ y $\widehat{\theta}_n = T(\widehat{F}_n)$ siendo $\widehat{F}_n$ 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

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

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

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

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

$$\underline{\theta}_{\mathrm{BP},\frac{\alpha}{2}} = h_n^{-1}\left(h_n(\widehat{\theta}_n) + \psi_{\frac{\alpha}{2}} \right)$$

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-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 $1-\alpha$ para θ como

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

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

Se define el intervalo bootstrap$-t$ como

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

donde

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

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

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

- Davison, A. & Hinkley, D. (1997). *Bootstrap methods and their application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge.
- Efron, B. (1982). *The jacknife, the bootstrap and other resampling plans*, Society for Industrial and Applied Mathematics, Philadelphia.