Intervalos de confianza

Kevin Piterman

30 de Junio, 2020

Resumen.
Sean $X_1, \dots, X_n$ variables aleatorias independientes y sea θ un parámetro de interés en la distribución de estas variables. Sea $\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$ un estimador de θ

- **Sesgo:** $b(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta$.
- **Insesgado:** si $b(\hat{\theta}_n) = 0$.
- **Asintóticamente insesgado:** si $\lim_n b(\hat{\theta}_n) = 0$.
- **Error cuadrático medio (ECM):** $\mathrm{ECM}_\theta(\hat{\theta}_n) = \mathbb{E}((\hat{\theta}_n - \theta)^2) = V_\theta(\hat{\theta}_n) + b(\hat{\theta}_n)^2$.
- **Criterio de estimadores:** elegir el de menor ECM.
- **Consistencia:** si $\hat{\theta}_n \xrightarrow{P} \theta$.

Convergencia en probabilidad. $X_n \xrightarrow{P} X$ si para todo $\varepsilon > 0$, $\lim_n P(|X_n - X| > \varepsilon) = 0$.

Convergencia en distribución. $X_n \xrightarrow{D} X$ si para todo t punto de continuidad de F_X, $\lim_n F_{X_n}(t) = F_X(t)$.

¿Cómo podemos probar *convergencia en probabilidad o distribución*?

Teorema. (Criterio de consistencia) Si $\lim_n \mathbb{E}_\theta(\hat{\theta}_n) = \theta$ y $\lim_n V_\theta(\hat{\theta}_n) = 0$ entonces $\hat{\theta}_n$ es consistente. Equivalentemente,

$$\text{Si } \lim_n \mathrm{ECM}(\hat{\theta}_n) = 0 \text{ entonces } \hat{\theta}_n \xrightarrow{P} \theta \text{ (es consistente).}$$

Propiedades. (Herramientas para probar convergencia) Sean $(X_i)_{i \geq 1}$ variables aleatorias.

1. Si $X_n \xrightarrow{P} X$ entonces $X_n \xrightarrow{D} X$.
2. (Vuelta) Si $X = c$ es constante y $X_n \xrightarrow{D} c$ entonces $X_n \xrightarrow{P} c$.
3. Si g es continua y $X_n \xrightarrow{P} X$ entonces $g(X_n) \xrightarrow{P} g(X)$.
4. (Álgebra de límites) Si $X_n \xrightarrow{P} X$ y $Y_n \xrightarrow{P} Y$ entonces $\frac{X_n}{Y_n} \xrightarrow{P} \frac{X}{Y}$, $X_n Y_n \xrightarrow{P} XY$ y $aX_b + bY_n \xrightarrow{P} aX + bY$.
5. Si g es continua y $X_n \xrightarrow{D} X$ entonces $g(X_n) \xrightarrow{D} g(X)$.
6. Si $X_n \xrightarrow{D} X$ y $Y_n \xrightarrow{D} c$ (constante) entonces $X_n Y_n \xrightarrow{D} Xc$.

Si además son iid con media μ y varianza σ^2 finitas:

7. (LGN) $\bar{X} \xrightarrow{P} \mu$.
8. (TCL) $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \xrightarrow{D} N(0,1)$. Equivalente: $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$ para n grande.

Intervalos de confianza: Sean $X_1, \ldots, X_n$ iid con distribución F_θ. Sea $\hat{\theta}_n$ un estimador de θ. Buscamos un intervalo aleatorio $(a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)) = I(X_1, \ldots, X_n)$, que depende de un parámetro $0 < \alpha < 1$, de manera que

$$P(\theta \in I(X_1, \ldots, X_n)) = 1 - \alpha.$$

Preferentemente α es chico y la longitud del intervalo también.

- Decimos que $I(X_1, \ldots, X_n)$ es el intervalo de confianza de nivel $1 - \alpha$ para θ.
- **Antes de observar:** $I(X_1, \ldots, X_n)$ contiene a θ con probabilidad $1 - \alpha$.
- **Después de observar:** $I(x_1, \ldots, x_n)$ contiene a θ con una "confianza" de $1 - \alpha$.
  Por ejemplo, si $I(x_1, \ldots, x_n) = (20, 30)$, $\alpha = 0.10$, entonces el intervalo $(20, 30)$ contiene a θ con una confianza de 0.90.

¿Cómo calculamos intervalos de confianza?

Idealmente: usar los estimadores que conocemos de θ para poder hallar las variables aleatorias $a(X_1, \ldots, X_n)$ y $b(X_1, \ldots, X_n)$.

Receta: método del pivote.

1. Encontrar una función pivote $T(X_1, \ldots, X_n, \theta)$ cuya distribución no dependa de θ (aunque la expresión de la función sí pueda depender de θ).
2. Plantear
   $$P(a < T(X_1, \ldots, X_n, \theta) < b) = 1 - \alpha.$$
   Como la distribución de T no depende de θ, podemos hallar a y b explícitamente.
3. Tratar de despejar θ de adentro de T para que nos quede algo del estilo
   $$P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) = 1 - \alpha.$$

Intervalos para la Normal $N(\mu, \sigma^2)$.

	$\sigma = \sigma_0$ conocido	σ desconocido
μ	$\left[\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]$	$\left[\overline{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right]$
	$\mu = \mu_0$ conocido	μ desconocido
σ	$\left[\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\chi^2_{n,1-\alpha/2}}\right]$	$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$

Propiedades de distribuciones conocidas. Sean $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independientes.

1. $S^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$.
2. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ y $\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \sim N(0, 1)$.
3. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.
4. $\bar{X}$ y S^2 son independientes.
5. $\sqrt{n}\left(\frac{\bar{X} - \mu}{S}\right) \sim t_{n-1}$.
6. $Z \sim N(0, 1)$ entonces $\sqrt{Z} \sim \chi^2_1$.

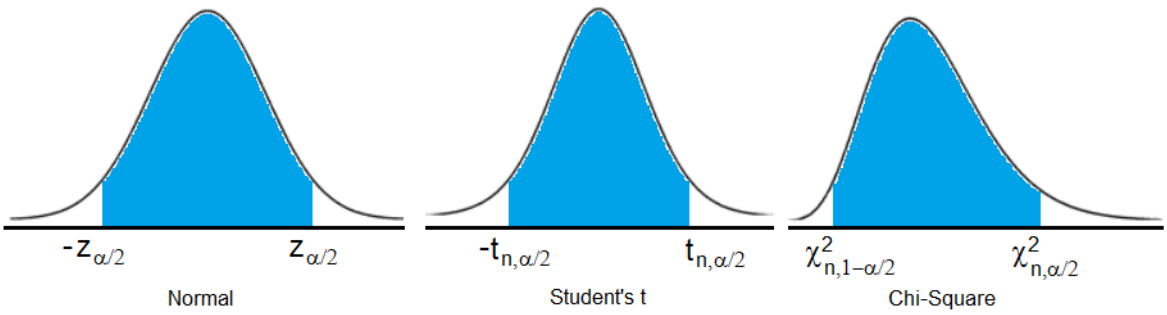

Figure 1: Cuantiles para las distribuciones conocidas. Las zonas en azul concentran una probabilidad de $1-\alpha$.

Ejemplo 1. Se supone que la longitud de cierto tipo de eje tiene distribución normal con desvío $\sigma = 0.05$ mm. Se toma una muestra de 20 ejes y se sabe que la longitud media (observada) es de 52.3 mm.

(a) Hallar un intervalo de confianza para la longitud media de nivel 0.99.

(b) ¿Qué tamaño deberá tener la muestra para que la longitud media del intervalo sea a lo sumo $\frac{\sigma}{10}$?

Solución. Tenemos una muestra aleatoria (independiente) $X_1, \ldots, X_n$, con $n = 20$, que sigue una distribución normal $N(\mu, \sigma^2)$. Además, $\sigma = 0.05$. Nos están pidiendo un intervalo de confianza para la media de una normal con desvío conocido, de nivel $1-\alpha = 0.99$ (o sea que $\alpha = 0.01$).

El intervalo de confianza es para la media de una normal con desvío conocido. Por la tabla anterior, este intervalo es de la forma:

$$I(X_1, \ldots, X_n) = \left[\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right].$$

Reemplazamos los datos: $z_{\alpha/2} = 2.57589$, $\overline{x} = 52.3$, $\sigma_0 = \sigma = 0.05$ y $\sqrt{n} = \sqrt{20}$,

$$I = \left(52.3 - 2.57589\frac{0.05}{\sqrt{20}}, 52.3 + 2.57589\frac{0.05}{\sqrt{20}}\right) = (52.2712, 52.3288)\,.$$

Para el segundo ítem, lo que nos están pidiendo es que la esperanza de la longitud del intervalo aleatorio que encontramos en el ítem anterior sea a lo sumo $\frac{\sigma}{10}$.

La longitud del intervalo es:

$$L = L(X_1, \ldots, X_n) = \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} - \left(\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 2z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}.$$

Notemos que la longitud del intervalo no depende de los valores de la muestra (o sea es constante). Sí depende del tamaño de la muestra.

Como el intervalo es constante (no aleatorio), su esperanza es él mismo, y lo que debemos hacer es calcular n de manera que tengamos la siguiente desigualdad:

$$2z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} = \mathbb{E}(L) \leq \frac{\sigma}{10} = \frac{0.05}{10}$$

$$2z_{\alpha/2}\frac{0.05}{\sqrt{n}} \leq \frac{0.05}{10}$$

Cancelamos:

$$2z_{\alpha/2}\frac{1}{\sqrt{n}} \leq \frac{1}{10}$$

Pasamos $\sqrt{n}$ multiplicando hacia la derecha, y el 10 multiplicando hacia la izquierda, de manera que, si luego elevamos al cuadrado,

$$(20z_{\alpha/2})^2 \leq n.$$

O sea que $n \geq 400z^2_{\alpha/2} = 400(2.58)^2 = 2654.084$.

Ejemplo 2. Sean $X_1, \ldots, X_n$ variables aleatorias iid con función de densidad dada por

$$f_X(x, \theta) = \frac{2x}{\theta^2} I_{(0,\theta)}(x), \quad \theta > 0.$$

(a) Verificar que $Y_i = -4 \log \frac{X_i}{\theta}$ tiene distribución $\mathcal{E}(1/2)$.

(b) Hallar la distribución de $\sum_{i=1}^n Y_i$ (Sugerencia: recordar que $\Gamma(\frac{k}{2}, \frac{1}{2}) = \chi^2_k$).

(c) Hallar un intervalo de confianza de nivel $1 - \alpha$ para θ.

(d) Hallar un intervalo de confianza de nivel 0.90 para θ, si $\prod_{i=1}^n x_i = 10$ con $n = 20$.

Solución.

Veamos que $Y_i \sim \mathcal{E}(1/2)$. Notemos que $Y_i = g(X_i)$, donde $g(x) = -4 \log \frac{x}{\theta}$. Luego $g^{-1}(y) = \theta e^{-\frac{y}{4}}$. Buscamos su derivada:

$$(g^{-1}(y))' = -\frac{1}{4} \theta e^{-\frac{y}{4}}.$$

Por el teorema del cambio de variable

$$f_Y(y) = f_X(g^{-1}(y)) |(g^{-1}(y))'| = \frac{2}{\theta^2} \theta e^{-\frac{y}{4}} I_{(0,\theta)}(\theta e^{-\frac{y}{4}}) \frac{1}{4} \theta e^{-\frac{y}{4}}.$$

Para calcular la indicadora, planteamos $0 < \theta e^{-\frac{y}{4}} < \theta$. Despejando y de esta inecuación, obtenemos que $y > 0$. Por otro, lado, cancelamos θ en la ecuación anterior y las exponencial se juntan de manera de que la densidad de Y nos queda:

$$f_Y(y) = \frac{1}{2} e^{-\frac{y}{2}} I_{(0,+\infty)}(y).$$

Por lo tanto, $Y_i = g(X_i)$ es exponencial de parámetro $1/2$ pues su densidad es la de tal distribución.

Como las X_i son independientes, las Y_i son independientes. Además, $Y_i \sim \mathcal{E}(1/2)$, por lo que $\sum_i Y_i$ es la suma de n exponenciales independientes de parámetro $\frac{1}{2}$. Luego su distribución es $\Gamma(n, \frac{1}{2})$, la cual es una distribución χ^2_{2n}.

Tratemos de encontrar ahora el intervalo de confianza de nivel $1 - \alpha$ para θ. Podemos utilizar como función pivote a $T = \sum_i Y_i$, ya que tiene distribución χ^2_{2n} que no depende del parámetro desconocido θ (ni de ningún otro parámetro desconocido). Con esta T, podemos elegir a, b constantes de manera que

$$P(a < T < b) = 1 - \alpha. \tag{0.1}$$

Como vimos en la Figura 1, podemos elegir $a = \chi^2_{2n,1-\alpha/2}$ y $b = \chi^2_{2n,\alpha/2}$. Notemos que

$$T = T(X_1, \ldots, X_n, \theta) = \sum_i Y_i = \sum_i -4 \log \frac{X_i}{\theta}.$$

Reemplazamos la expresión de T en la Ecuación 0.1 para ver si podemos despejar θ.

$$\begin{aligned}
1 - \alpha &= P(a < T < b) \\
&= P\left(a < \sum_i -4 \log \frac{X_i}{\theta} < b \right) \\
&= P\left(a < \sum_i -4(\log X_i - \log \theta) < b \right)
\end{aligned}$$

$$= P\left(a < \sum_i -4(\log X_i) - \sum_i (-4)\log\theta < b\right)$$
$$= P\left(a < Z_n + 4n\log\theta < b\right) \qquad Z_n := \sum_i -4(\log X_i)$$
$$= P\left(e^{(a-Z_n)/4n} < \theta < e^{(b-Z_n)/4n}\right)$$

Notemos que

$$e^{-\frac{Z_n}{4n}} = e^{-\frac{-4\sum_i \log X_i}{4n}} = \prod_i e^{\frac{1}{n}\log X_i} = \prod_i X_i^{1/n}.$$

Por lo tanto, el intervalo de confianza de nivel $1-\alpha$ para θ es

$$I(X_1,\ldots,X_n) = \left[e^{a/4n}\prod_i X_i^{1/n}, e^{b/4n}\prod_i X_i^{1/n}\right].$$

Para encontrar el intervalo de nivel 0.90 para θ sabiendo que $\prod_i x_i = 10$, reemplazamos en la fórmula del intervalo que hallamos en el inciso anterior estos valores. Recordar que $n = 20$ y $\alpha = 0.10$.

$$I = I(x_1,\ldots,x_{20}) = \left[e^{a/80}10^{1/20}, e^{b/80}10^{1/20}\right].$$

Además, $a = \chi^2_{40,0.95} = 26.5093$ y $b = \chi^2_{40,0.05} = 55.75848$. Por lo tanto,

$$I = \left[e^{26.5093/80}10^{1/20}, e^{55.75848/80}10^{1/20}\right] = [1.56283, 2.25266].$$

Ejemplo 3. Sean $x_1, \ldots, x_{15}$ números generados aleatoriamente de una distribución $N(\mu, \sigma^2)$. Se sabe que $\sum_{i=1}^{15} \frac{x_i}{15} = 0.706$ y que $\sum_{i=1}^{15}(x_i - 0.706)^2 = 241.9234$.

(a) Hallar un intervalo de confianza para μ de nivel 0.95.

(b) Ídem para σ^2.

Solución. El intervalo para μ con desvío estándar desconocido en una normal es

$$\left[\bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right].$$

En este caso, tenemos que $\bar{x} = 0.706$, $\alpha = 0.05$, $n = 15$, $s^2 = \frac{241.9234}{14}$. Además, $t_{n-1,\alpha/2} = t_{14,0.025} = 2.144787$. Luego el intervalo buscado es:

$$I = \left[\bar{x} - t_{14,0.025}\frac{s}{\sqrt{15}}, \bar{x} - t_{14,0.025}\frac{s}{\sqrt{15}}\right] = [-1.59604, 3.00804].$$

Para σ^2, planteamos el intervalo correspondiente:

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

En este caso tenemos que $\chi^2_{n-1,1-\alpha/2} = 5.628726$ y $\chi^2_{n-1,\alpha/2} = 26.11895$. Luego el intervalo es

$$I = \left[\frac{241.9234}{26.11895}, \frac{241.9234}{5.628726}\right] = [9.262371, 42.98013].$$

Nota: estábamos resolviendo los ejercicios con los cuantiles de lower tail para la distribución chi cuadrado, pero los intervalos estaban calculados teniendo en cuenta los cuantiles del upper tail. Corregí estos problemas junto con la Figura 1.