Intervalos de confianza - continuación

Clase práctica 17/11/20

Daniela Parada

Departamento de Matemática
Universidad de Buenos Aires

Probabilidad y Estadística (C)

Plan de trabajo de hoy

Pendientes de la clase pasada

- Ejercicio 5 de la guía 8
- Ejercicio extra

IC asintóticos

- Motivación y ejemplos
- Ejercicios

IC para la varianza con μ conocido

Queremos obtener un IC para σ^2 bajo el supuesto de distribución $N(\mu, \sigma^2)$ con μ conocido.

Tomamos la sugerencia. Vamos a probar que $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi_n^2$.

Hagamos un cambio de variable de modo de poder estudiar la distribución de $Y = X^2$ con $X \sim N(0,1)$.

$$
\begin{aligned}
F_Y(y) &= P(Y \leq y) = P\left(X^2 \leq y\right) = P\left(|X| \leq y\right) \\
&= P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) = F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right) \\
f_Y(y) &= f_X\left(\sqrt{y}\right)\left(\frac{1}{2\sqrt{y}}\right) I_{(0,\infty)}(y) - f_X\left(-\sqrt{y}\right)\left(-\frac{1}{2\sqrt{y}}\right) I_{(0,\infty)}(y) \\
&= \frac{1}{2\sqrt{2\pi y}} e^{-y/2} I_{(0,\infty)}(y) + \frac{1}{2\sqrt{2\pi y}} e^{-y/2} I_{(0,\infty)}(y) \\
&= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} I_{(0,\infty)}(y)
\end{aligned}
$$

De la densidad de Y anterior, vemos que se distribuye como una $\Gamma(1/2, 1/2)$. A esta distribución se la conoce como chi-cuadrado con 1 grado de libertad: $Y \sim \chi_1^2$.

Por otro lado, como las X_i son iid $N(\mu, \sigma^2)$, entonces

- $\frac{X_i - \mu}{\sigma} \sim N(0, 1)$
- $\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_1^2$
- $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2$

Este último punto se deduce del hecho de que la distribución χ_1^2 es equivalente a una distribución $\Gamma(1/2, 1/2)$. Entonces, la suma de n variables aleatorias independientes (lo heredan de la m.a.) e idénticamente distribuídas como $\Gamma(1/2, 1/2)$ se distribuye como una $\Gamma(n/2, 1/2)$ o, lo que es lo mismo, como una χ_n^2.

Probamos que $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi_n^2$. Con esto, podemos construir un IC de nivel deseado para σ^2.

$$
\begin{aligned}
1-\alpha &= P\left(a \leq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \leq b\right) \\
&= P\left(1/a \geq \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\mu)^2} \geq 1/b\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{a} \geq \sigma^2 \geq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{b}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{b} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{a}\right)
\end{aligned}
$$

Donde a y b son los percentiles de la distribución que dejan área igual a $\alpha/2$ en cada extremo. En particular y con la notación que estamos usando acá: $a = \chi^2_{n,1-\alpha/2}$ y $b = \chi^2_{n,\alpha/2}$.

Conclusión: un intervalo de confianza exacto de nivel (1-α) para σ^2 bajo el supuesto de distribución $N(\mu, \sigma^2)$ con μ conocido es

$$\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n,\alpha/2}} ; \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\chi^2_{n,1-\alpha/2}}\right]$$

¿Y si μ es desconocido? Necesito un nuevo pivot (ejercicio 6 de la guía).

Ejercicio

Sean $X_1, \ldots, X_n$ variables aleatorias i.i.d con función de densidad dada por

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

Verificar que $Y_i = -4 \log \frac{X_i}{\theta}$ *tiene distribución* $Exp(1/2)$

Hallar la distribución de $\sum_{i=1}^{n} Y_i$.

Hallar un intervalo de confianza de nivel $1 - \alpha$ *para* θ.

Hallar un intervalo de confianza de nivel 0.90 para θ, *si* $\prod_{i=1}^{n} X_i = 10$.

Resolución

Buscamos la distribución de Y.

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) = P\left(-4\log\frac{X}{\theta} \leq y\right) = P\left(\log\frac{X}{\theta} \geq \frac{-y}{4}\right) \\
&= P\left(\frac{X}{\theta} \geq e^{-y/4}\right) = P\left(X \geq \theta e^{-y/4}\right) \\
&= 1 - P\left(X \leq \theta e^{-y/4}\right) = 1 - F_X\left(\theta e^{-y/4}\right) \\
f_Y(y) &= -f_X\left(\theta e^{-y/4}\right) \cdot \left(-\frac{\theta}{4}\right) e^{-y/4} \cdot I_{(0,\theta)}\left(\theta e^{-y/4}\right) \\
&= -\frac{2\theta e^{-y/4}}{\theta^2} \cdot \left(-\frac{\theta}{4}\right) e^{-y/4} \cdot I_{(0,\infty)}(y) \\
&= \frac{1}{2} e^{-y/2} \cdot I_{(0,\infty)}(y)
\end{aligned}$$

Vemos que $Y \sim Exp(1/2)$.

Buscamos la distribución de $\sum_{i=1}^{n} Y_i$.

Empecemos por ver la suma (p. 113 apunte). Sea $Z = X + Y$ con X e Y variables aleatorias continuas e independientes. Entonces, para cada $z \in \mathbb{R}$, $A_z = \left\{(x, y) \in \mathbb{R}^2 : x + y \leq z\right\}$, la función de distribución de Z es:

$$
\begin{aligned}
F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) &= \iint_{A_z} f_{X,Y}(x, y) dydx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-x} f_{X,Y}(x, y) dy \right) dx
\end{aligned}
$$

La densidad de Z se puede obtener derivando respecto de z la función de distribución anterior:

$$
f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x) dx
$$

Bajo el supuesto de X e Y variables aleatorias independientes, su densidad conjunta es $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Por lo anterior, entonces, la función de densidad de Z es

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)dx = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

Con X e Y iid exponencial de parámetro $\lambda > 0$, las densidades marginales $f_X(x)$ y $f_Y(y)$ están definidas en $[0,\infty)$, entonces:

$$\begin{aligned} f_Z(z) &= \int_0^z f_X(x)f_Y(z-x)dx = \int_0^z \lambda e^{-\lambda x}\lambda e^{-\lambda(z-x)}dx \\ &= \lambda^2 \int_0^z e^{-\lambda z}dx = \lambda^2 z e^{-\lambda z} \end{aligned}$$

Conclusión: coincide con la densidad de la distribución Gamma de parámetros (2, λ). Es decir: $Z = X + Y \sim \Gamma(2,\lambda)$.

Sabemos que $Y_i \overset{\text{iid}}{\sim} Exp(1/2)$. Luego, por lo que vimos recién, es posible extender la idea a n finitos sumandos:

$$Y_i \overset{\text{iid}}{\sim} \Gamma(n, 1/2)$$

Para agendar: la exponencial es un caso particular de la Gamma. En particular, $Exp(\lambda)$ se distribuye como $\Gamma(1, \lambda)$. Y la suma de n exponenciales independientes e idénticamente distribuidas como $Exp(\lambda)$ se distribuye como $\Gamma(n, \lambda)$. A su vez, la Chi es también un caso especial de la Gamma. En particular, una distribución χ^2_k es equivalente a una $\Gamma(k/2, 1/2)$.

Yapa: $Y_i \overset{\text{iid}}{\sim} \Gamma(n, 1/2) \equiv \chi^2_{2n}$.

¿Para qué hicimos todo esto? 

Queremos un IC de nivel $(1-\alpha)$ para θ. Pero θ estaba en el soporte de X...

Con la transformación: ¡nos fabricamos un pivot! Es decir, una función de la muestra aleatoria $X_1, ..., X_n$ cuya distribución no depende de θ y es conocida. ¡Listo!

Con a y b los percentiles correspondientes de la Gamma, tenemos:

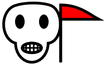
$$\begin{aligned}
1-\alpha &= P\left(a \leq \sum_{i=1}^{n} Y_i \leq b\right) \\
&= P\left(a \leq \sum_{i=1}^{n}\left(-4\log\frac{X_i}{\theta}\right) \leq b\right) \\
&= P\left(a \leq -4\log\prod_{i=1}^{n}\left(\frac{X_i}{\theta}\right) \leq b\right)
\end{aligned}$$

$$
\begin{aligned}
1-\alpha &= P\left(a \le -4\log\prod_{i=1}^{n}\left(\frac{X_i}{\theta}\right) \le b\right) \\
&= P\left(-a/4 \ge \log\prod_{i=1}^{n}\left(\frac{X_i}{\theta}\right) \ge -b/4\right) \\
&= P\left(e^{-a/4} \ge \prod_{i=1}^{n}\left(\frac{X_i}{\theta}\right) \ge e^{-b/4}\right) \\
&= P\left(e^{-a/4} \ge \frac{\prod_{i=1}^{n} X_i}{\theta^n} \ge e^{-b/4}\right) \\
&= P\left(e^{a/4} \le \frac{\theta^n}{\prod_{i=1}^{n} X_i} \le e^{b/4}\right) \\
&= P\left(e^{a/4}\prod_{i=1}^{n} X_i \le \theta^n \le e^{b/4}\prod_{i=1}^{n} X_i\right)
\end{aligned}
$$

$$1 - \alpha = P\left(\sqrt[n]{e^{a/4}\prod_{i=1}^{n} X_i} \leq \theta \leq \sqrt[n]{e^{b/4}\prod_{i=1}^{n} X_i}\right)$$

Con lo cual, el IC de nivel 0.90 para θ, si $\prod_{i=1}^{n} X_i = 10$ es algo que se obtiene de forma directa. Ojo, sin conocer n no es posible obtener a y b ya que estos percentiles son los de la distribución $\Gamma(n, 1/2)$ que dejan área $\alpha/2$ en las colas (su obtención depende de n).

¡Terminamos con los exactos! Pero... ¿y los asintóticos?



Motivación

Mismo espíritu que en los anteriores... solo que ahora no vamos a usar la distribución exacta del pivot sino la asintótica, usualmente dada por la convergencia en distribución de TCL. ¿Por qué no usamos la exacta?

- Porque no la conocemos o es difícil hallarla, o
- porque no logramos un pivot que no dependa del parámetro, o
- porque no sabemos la distribución subyacente de los datos.

Definición

Ver teórica y/o apunte (p. 188)

Algunos IC asintóticos

Supuestos: $X_1, \cdots, X_n$ una muestra aleatoria de una distribución desconocida con $E(X_1) = \mu$ y $V(X_1) = \sigma^2 < \infty$.

Ejemplo

IC para μ con σ^2 conocido. Por TCL:

$$\sqrt{n}\frac{\bar{X} - \mu}{\sigma} \xrightarrow{\mathcal{D}} N(0, 1).$$

Esto puede ser un pivote para la construcción del IC.

$$P\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \leq z_{\alpha/2}\right) \longrightarrow 1 - \alpha$$

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Algunos IC asintóticos

Si σ^2 no es conocido, usamos su estimador S^2.

Ejemplo

IC para μ con σ^2 desconocido. Por TCL, consistencia de S^2 y Slutsky:

$$\left.\begin{array}{c} \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \xrightarrow{\mathcal{D}} N(0,1) \\ \frac{\sigma}{S} \xrightarrow{p} 1 \end{array}\right\} \quad \Rightarrow \quad \sqrt{n}\frac{\bar{X}-\mu}{S} \xrightarrow{\mathcal{D}} N(0,1)$$

Esto puede ser un pivote para la construcción del IC.

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

Ejercicio 9 de la guía 8

Resolución

Sea $X \sim Be(p)$ con p la probabilidad de que el ciudadano se oponga a la propuesta política: codifico binario. Considero una muestra aleatoria (¿encuesta?) $X_1, \cdots, X_n$ de esa distribución y quiero hallar un IC asintótico de nivel 0.90 para p. Entonces, por TCL:

$$\sqrt{n}\frac{\bar{X} - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{D}} N(0,1).$$

Podría armar un pivot con eso pero ya no sería trivial despejar p. Una alternativa es usar su estimador ya que por LGN: $\bar{X} \xrightarrow{p} p$.

Entonces:

$$P\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\bar{X}-p}{\sqrt{\bar{X}\left(1-\bar{X}\right)}} \leq z_{\alpha/2}\right) \longrightarrow 1-\alpha$$

El IC asintótico de nivel $1-\alpha$ para p queda:

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

Con los datos:

$$\left[0{,}6-1{,}6449\sqrt{\frac{0{,}6(0{,}4)}{1000}};0{,}6+1{,}6449\sqrt{\frac{0{,}6(0{,}4)}{1000}}\right]=[0{,}5745;0{,}6255]$$

La longitud del intervalo es $2z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$.

Si quiero n tal que esa longitud sea menor que un valor dado (0.02 en este caso), tengo un problema pues $\bar{X}$ depende del tamaño de muestra.

Necesito acotar $\sqrt{\bar{X}\left(1-\bar{X}\right)}$ de alguna forma. Pero esto ya lo hicimos antes: $\bar{X}\left(1-\bar{X}\right) \leq 1/4$. ¿Por qué?

Entonces:

$$2z_{\alpha/2}\sqrt{\frac{\bar{X}\left(1-\bar{X}\right)}{n}} \leq 2z_{\alpha/2}\sqrt{\frac{1}{4n}} = \frac{z_{\alpha/2}}{\sqrt{n}}.$$

Basta entonces hallar n tal que

$$\frac{z_{\alpha/2}}{\sqrt{n}} \leq 0{,}02.$$

Con los datos:

$$n \geq (1{,}6449/0{,}02)^2 = 6763{,}859.$$

▸ Simulación