Probabilidad y estadística

Clase 6

Intervalos de confianza

Motivación

Hasta ahora habíamos visto estimadores puntuales, que, dada un muestra, nos devuelven un único valor $\hat{\theta}$ que se aproxima al valor verdadero del parámetro deseado θ.

Una forma de obtener información sobre la precisión de la estimación, en el caso de que θ sea unidimensional, es proporcionar un intervalo $[a(X), b(X)]$ de manera que la probabilidad de que dicho intervalo contenga el verdadero valor θ sea alta, por ejemplo, 0.95.

Región de confianza

Def: Dada una m.a. $\underset{\sim}{X}$ con distribución perteneciente a una familia $F_\theta(x)$, con $\theta \in \Theta$, una región de confianza $S(\underset{\sim}{X})$ para θ con nivel de confianza $1-\alpha$ será un conjunto tal que

$$\mathbb{P}(\theta \in S(\underset{\sim}{X})) = 1-\alpha. \text{ (*)}$$

Obs: θ **no** es aleatorio, lo aleatorio es (*) es $S(\underset{\sim}{X})$.

Obs: Si $S(\underset{\sim}{X}) = (a(\underset{\sim}{X}), b(\underset{\sim}{X}))$ diremos que es un intervalo de confianza.

Si $S(\underset{\sim}{X}) = (\min(\Theta), b(\underset{\sim}{X}))$ diremos que es una cota superior.

Si $S(\underset{\sim}{X}) = (a(\underset{\sim}{X}), \max(\Theta))$ diremos que es una cota inferior.

Juguemos un poquito

Usemos la siguiente api para entender mejor qué es un IC:

http://rossmanchance.com/applets/2021/confsim/ConfSim.html

Método del pivote

Teorema: Sea $\underline{X}$ una muestra aleatoria con distribución perteneciente a una familia $F_\theta(x)$, con $\theta \in \Theta$, y sea $U = g(\underline{X}, \theta)$ una variable cuya distribución **no** depende de θ. Sean a y b tales que $\mathbb{P}(a \leq U \leq b) = 1 - \alpha$. Luego,

$$S(\underline{X}) = \{\theta : a < g(\underline{X}, \theta) \leq b\}$$

es una región de confianza para θ. A U se lo llama pivote.

Ejercicio 1

Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de tamaño n de una población con distribución normal de media μ y varianza 4. Hallar una cota inferior del 95% para μ .

Suponer n=20 y μ=3, simular la muestra y obtener el valor de la cota.

Algunos resultados importantes

Teorema: Sea $\underline{X} = X_1, \ldots, X_n$ una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$

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

$$W = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

V y W son independientes

Si $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $U = \sqrt{n}\frac{(\bar{X}-\mu)}{S} \sim t_{n-1}$

Obs: en general vale que si $X \sim \mathcal{N}(0,1)$ y $Y \sim \chi^2_n$, con X e Y independientes vale que $\frac{X}{\sqrt{Y/n}} \sim t_n$

Algunos pivotes para variables normales

Dada $\underline{X}_n$ una m.a. de una distribución $\mathcal{N}(\mu, \sigma^2)$ definimos algunos pivotes:

- Para la media con varianza conocida: $U(\underline{X}, \mu) = \frac{(\overline{X} - \mu)}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$
- Para la media con varianza desconocida: $U(\underline{X}, \mu) = \frac{(\overline{X} - \mu)}{S}\sqrt{n} \sim t_{n-1}$
- Para el desvío con media conocida: $U(\underline{X}, \sigma) = \frac{\sum_{i-1}^{n}(X_i - \mu)^2}{\sigma} \sim \chi^2_n$
- Para el desvío con media desconocida: $U(\underline{X}, \sigma) = \frac{\sum_{i-1}^{n}(X_i - \overline{X})^2}{\sigma} \sim \chi^2_{n-1}$

Dada también $\underline{Y}_m$ una m.a. de una distribución $\mathcal{N}(\lambda, \sigma^2)$ y sea :

- Comparación de medias con varianzas conocidas: $U(\underline{X}, \Delta) = \frac{\overline{X} - \overline{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$
- Comparación de medias con varianzas desconocidas e iguales:

$U(\underline{X}, \Delta) = \frac{\overline{X} - \overline{Y} - \Delta}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$, con $S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{n+m-2}$

Ejercicio 2

Dada una muestra aleatoria $\underline{X} = (X_1, \ldots, X_n)$ de una población con distribución normal con media y varianza desconocidas, hallar el intervalo de confianza de nivel 0.99 para la media de la población.

Suponer n=50, $\mu = 2, \sigma = 3$, simular la muestra y calcular el IC resultante de la misma.

Regiones de confianza asintóticas

Def: Sea $\underline{X}_n = X_1, \ldots, X_n$ una m.a de una población con distribución perteneciente a la flía. $F_\theta(x)$, con $\theta \in \Theta$. Se dice que $S_n(\underline{X}_n)$ es una sucesión de regiones de confianza de nivel asintótico $1-\alpha$ si:

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in S_n(\underline{X}_n)) = 1-\alpha$$

Teorema: Sea $\underline{X}_n$ una m.a. de una población con distribución $F_\theta(x)$, con $\theta \in \Theta$. Supongamos que para cada n se tiene $U_n = g(\underline{X}_n, \theta)$ que converge en distribución a U, donde U es una v.a. cuya distribución no depende de θ. Entonces si a y b son tales que $\mathbb{P}(a < U < b) = 1-\alpha$ se tiene que $S_n(\underline{X}_n) = \{\theta : a < U_n < b\}$ es una región de confianza de nivel asintótico $1-\alpha$ para θ.

Ejercicio 3

Se arroja 50 veces una moneda con probabilidad p de salir cara. Hallar un intervalo de confianza asintótico de nivel 0.95 para p basado en la observación $x=50$.

IC para la media de una población desconocida

En general, dada una m.a $\underline{X}_n$ de una población desconocida, una buena forma de aproximarse a la media de dicha población es considerar el promedio de las muestras ($\bar{X}_n$).

Por TCL, sabemos que $\bar{X}_n$ tiende en distribución a una v.a. normal. En particular,

$$\frac{\bar{X}_n - \mathbb{E}[X]}{\sqrt{var(X)/n}} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

Se puede probar que si se desconoce también la varianza de la población (que es lo más común) vale que

$$\frac{\bar{X}_n - \mathbb{E}[X]}{S/\sqrt{n}} \overset{(a)}{\sim} \mathcal{N}(0,1)$$

Ejercicio 4

De un experimento en los efectos de un medicamento para la ansiedad se midió el puntaje en un test de memoria antes y después de tomar el medicamento. A partir de los datos que se encuentran en el archivo Islander_data.csv hallar un IC para la media del tiempo de respuesta después de consumir el medicamento.

Intervalos de confianza aproximados (plug-in)

Intervalos basados en la Normal: $T(\widehat{F}_n) \approx N(T(F), \widehat{\mathsf{se}}^2)$

Entonces, un intervalo aproximado de nivel $1-\alpha$ está dado por $T(\widehat{F}_n) \pm z_{\alpha/2}\,\widehat{\mathsf{se}}$.

Ejemplo: un intervalo para la media $\mu = T(F) = \int x\,dF(x)$ está dado por $\overline{X}_n \pm z_{\alpha/2}\,\widehat{\mathsf{se}}$. donde $\mathsf{se} = \sqrt{\mathbb{V}(\overline{X}_n)} = \sigma/\sqrt{n}$

Bootstrap: estimar la varianza

La idea es estimar $\mathbb{V}_{\widehat{F}_n}(T_n)$ mediante una simulación:

$$
\begin{array}{lcccccc}
\text{Real world} & F & \Longrightarrow & X_1, \ldots, X_n & \Longrightarrow & T_n = g(X_1, \ldots, X_n) \\
\text{Bootstrap world} & \widehat{F}_n & \Longrightarrow & X_1^*, \ldots, X_n^* & \Longrightarrow & T_n^* = g(X_1^*, \ldots, X_n^*)
\end{array}
$$

Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$.
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left(T_{n,b}^* - \frac{1}{B} \sum_{r=1}^{B} T_{n,r}^* \right)^2 .$$

Bibliografía

- "Notas de Estadística", Graciela Boente y Víctor Yohai, FCEyN, UBA.
- “All of Statistic: A concise Course in Statistical Inference”, Larry Wasserman