Intervalos de confianza

Los intervalos de confianza son una forma alternativa de utilizar un estimador $\hat{\theta}$ cuando deseamos estimar un parámetro desconocido θ. Podemos encontrar un intervalo (A, B) que pensamos tiene una alta probabilidad de contener al parámetro.

Supongamos que $X_1, \ldots, X_n$ constituye una muestra aleatoria de una distribución con parámetro θ. Supongamos también que es posible encontrar estadísticos $A(X_1, \ldots, X_n)$ y $B(X_1, \ldots, X_n)$ tales que:

$$P(A(X_1, \ldots, X_n) < \theta < B(X_1, \ldots, X_n)) = \gamma$$

donde γ es una probabilidad fija $(0 < \gamma < 1)$. Si los valores observados de A y B son a y b, diremos que el intervalo (a, b) es un *intervalo de confianza para* θ con coeficiente de confianza γ.

No es correcto afirmar que θ está en el intervalo (a, b) con probabilidad γ (las variables aleatorias son los extremos del intervalo; una vez observada la muestra, desaparece la aleatoriedad).

Interpretación del Intervalo de Confianza:

Antes de tomar una muestra, hay una probabilidad γ de que el intervalo que se va a construir a partir de la muestra incluya el valor desconocido de θ. Una vez que se observa la muestra, no es posible asignar una probabilidad al suceso $\theta \in (a, b)$ sin considerar a θ como variable aleatoria, lo cual entonces que θ tendrá una distribución de probabilidades.

Limitaciones del Intervalo de Confianza

Aunque los valores de la muestra particular que se observan den mayor información al experimentador sobre si el intervalo realmente incluye a θ, no existe un método estándar para ajustar el coeficiente de confianza γ partiendo de esta información.

Intervalo de Confianza para la media de una distribución normal

Sean $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución normal con parámetros μ y σ desconocidos.

Es conocido que:

$$\frac{\bar{X}_n - \mu}{\sqrt{s^2/n}} \sim t_{n-1}$$

Sea $g_{n-1}(x)$ la f.d.p de una t_{n-1}.

Sea c una constante tal que :

$$\int_{-c}^{c} g_{n-1}(x)dx = \gamma$$

$$\Rightarrow G_{n-1}(c) - G_{n-1}(-c) = G_{n-1}(c) - (1 - G_{n-1}(c)) = 2G_{n-1}(c) - 1$$

$$\Rightarrow G_{n-1}(c) = \frac{1-\gamma}{2}$$

Llamando $1 - \gamma = \alpha$; $c = t_{n-1}^{\alpha/2}$

$$\Rightarrow P(-t_{n-1}^{\alpha/2} < \frac{\bar{X}_n - \mu}{\sqrt{s^2/n}} < t_{n-1}^{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P(\bar{X}_n - t_{n-1}^{\alpha/2}\frac{s}{n^{1/2}} < \mu < \bar{X}_n + t_{n-1}^{\alpha/2}\frac{s}{n^{1/2}})$$

$$\Rightarrow (\bar{X}_n - t_{n-1}^{\alpha/2}\frac{s}{n^{1/2}}, \bar{X}_n + t_{n-1}^{\alpha/2}\frac{s}{n^{1/2}})$$

es intervalo de confianza de coeficiente de confianza γ para μ.

Intervalos de confianza (Continuación)

Nótese que si $P(c_1 < T < c_2) = \gamma$, ($T \sim t_{n-1}$), el intervalo:

$$(\bar{X}_n - \frac{c_2 s}{n^{1/2}}, \bar{X}_n + \frac{c_1 s}{n^{1/2}})$$

es también un intervalo de confianza con coeficiente de confianza γ para μ; sin embargo puede demostrarse que el intervalo simétrico con respecto a $\bar{X}_n$ es el de menor longitud.

Ejemplo: Variables aleatorias uniformes en un intervalo de longitud uno

Sean X_1 y X_2 son variables aleatorias tomadas de una distribución uniforme en el intervalo $[\theta - 1/2, \theta + 1/2]$, donde el valor de θ es desconocido ($-\infty < \theta < \infty$).

Sean $Y_1 = min(X_1, X_2)$ y $Y_2 = max(X_1, X_2)$.

$$\begin{aligned}
P(Y_1 < \theta < Y_2) &= P(X_1 < \theta < X_2) + P(X_2 < \theta < X_1) \\
&= P(X_1 < \theta)P(X_2 > \theta) + P(X_2 < \theta)P(X_1 > \theta) \\
&= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}
\end{aligned}$$

Para valores observados de y_1 y y_2 el intervalo (y_1, y_2) será un intervalo de confianza para θ con coeficiente de confianza $1/2$.

Continuación del ejemplo anterior

Las observaciones X_1 y X_2 deben ser al menos $\theta - (1/2)$ y ambas deben ser máximo $\theta + (1/2)$. Sabemos con certeza que $y_1 \geq \theta - (1/2)$ y $y_2 \leq \theta + (1/2)$. En otras palabras:

$$y_2 - (1/2) \leq \theta \leq y_1 + (1/2) \quad (*)$$

Supongamos que $(y_2 - y_1) > 1/2$; entonces $y_1 < y_2 - (1/2)$ y de (*) se obtiene que $y_1 < \theta$. Además como $y_1 + (1/2) < y_2$, entonces de (*) se obtiene que $\theta < y_2$. Esto nos dice con certeza que $y_1 < \theta < y_2$; es decir, el intervalo (y_1, y_2) incluye al valor desconocido de θ cuando $(y_2 - y_1) > 1/2$, aunque el coeficiente de confianza sea solamente $1/2$.

Notar que el coficiente de confianza no depende de los valores observados y_1 y y_2.

Intervalos de Probabilidad

Sean c_1 y c_2 cantidades tales que:

$$P(c_1 < \theta < c_2|\mathbf{x}) = \gamma$$

En este caso diremos que (c_1, c_2) es un intervalo de probabilidad γ para θ (**Nótese la diferencia de interpretación**!).

Claramente existen muchos valores posibles de c_1 y c_2, y por lo tanto muchos posibles intervalos de probabilidad. Se suele usar el que tiene mínima longitud, el cuál corresponde al **intervalo de probabilidad posterior máxima** (HPD).

Idealmente a uno le gustaría reportar una región de valores de θ que sea tan pequeña como sea posible, pero que tenga tanta probabilidad como sea posible.