Intervalos de confianza

Kevin Piterman

2 de Julio, 2020

Resumen.

Intervalos de confianza: Sean $X_1, \ldots, X_n$ iid con distribución $\mathcal{D}_\theta$. 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
$$P(a(X_1, \ldots, X_n) < \theta < b(X_1, \ldots, X_n)) = 1 - \alpha.$$

¿Qué pasa si no encontramos una función pivote o bien su distribución no es conocida?

Respuesta: Usamos TCL y aproximamos por una normal o alguna distribución que aparezca a partir de la estandarización. Así, tratamos de despejar el intervalo a partir de una aproximación por estandarización. El nivel de confianza del intervalo es asintótico, porque vale aproximadamente, en el límite. En general se comporta bastante bien.

¿En qué cambia el cálculo del intervalo asintótico?

Respuesta: La idea en general es que no voy a conocer la distribución del pivote T, pero asintóticamente sí. Luego me baso en esto para construir el intervalo de la misma manera que hacíamos con las distribuciones conocidas. Al final, no vamos a tener un intervalo de nivel exacto $1 - \alpha$, pero va a ser de nivel aproximado $1 - \alpha$ (y lo decimos de esa manera "intervalo de confianza asintótico de nivel aproximado $1 - \alpha$").

¿Qué intervalo elegimos si hay más de uno?

Respuesta: Elegimos siempre el que tenga menor longitud esperada y el más exacto. Es decir, si $I = (a, b)$ es el intervalo (con a y b aleatorios), entonces su longitud es $L(I) = b - a$ y la longitud esperada es $\mathbb{E}(L(I)) = \mathbb{E}(b - a)$.

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 $\sqrt{Z} \sim \chi^2_1$.

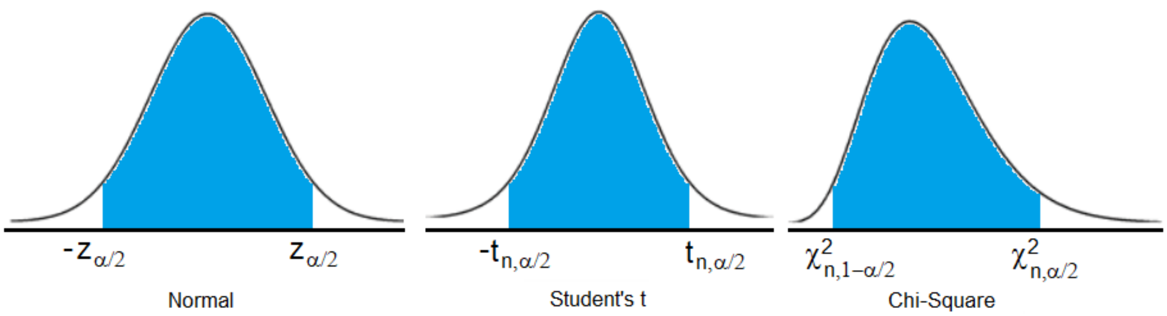

Figure 1: Cuantiles para las distribuciones conocidas. Las zonas en azul concentran una probabilidad de $1 - \alpha$.

Ejemplo 1. Se ha obtenido una muestra de 30 vendedores de una editorial para estimar el valor medio de las ventas por trabajador en la empresa. La media y varianza de la muestra (en cientos de miles de pesos) son 5 y 2, respectivamente. Hallar

1. Intervalo de confianza asintótico para la venta media por trabajador en la editorial de nivel 90%.
2. Asumiendo que la venta de cada vendedor tiene distribución Normal, hallar un intervalo de confianza para la varianza de las ventas por trabajador en la editorial de nivel 90%.

Solución. Podemos suponer que los trabajadores realizan ventas con la misma distribución y que además las ventas son independientes entre trabajadores. Lo que nos están pidiendo es el valor esperado de ventas por trabajador, o sea, la media de esta distribución (que no conocemos!).

Lo que sí sabemos por el teorema central del límite y álgebra de límites de convergencia en distribución y probabilidad, es que, asintóticamente,

$$\frac{\bar{X}-\mu}{S}\sqrt{n} \approx t_{n-1}$$

Más aún, esto se parece a la $N(0,1)$ por el siguiente motivo. Como $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \xrightarrow{P} N(0,1)$ y $\frac{\sigma}{S} \xrightarrow{P} 1$, por el álgebra de límites de convergencia en probabilidad,

$$\frac{\bar{X}-\mu}{S}\sqrt{n} = \left(\frac{\bar{X}-\mu}{\sigma}\sqrt{n}\right)\left(\frac{\sigma}{S}\right) \xrightarrow{P} N(0,1).1 = N(0,1).$$

La aproximación por t de Student es correcta pues $t_{n-1} \approx N(0,1)$ si n es grande. Notemos que usamos S en lugar del desvío estándar de las X_i porque queremos estimar la media μ y no conocemos su varianza (o desvío).

Sabemos que $n = 30$ y $\alpha = 0.10$, por lo que $\alpha/2 = 0.05$. Luego $t_{n-1,\alpha/2} = 1.699127$. El código de R para obtener este valor es:

```
qt(0.05, df = 29, lower.tail = FALSE)
```

Ahora construimos el intervalo de confianza de nivel asintótico a partir de los datos obtenidos:

$$\begin{aligned} 1-\alpha \sim P&\left(-t_{n-1,\alpha/2} \leq \frac{\bar{X}-\mu}{S}\sqrt{n} \leq t_{n-1,\alpha/2}\right) \\ &= P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \bar{X}-\mu \leq t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\ &= P\left(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right). \end{aligned}$$

Luego el intervalo de confianza asintótico de nivel 90% para la media (con varianza desconocida) es:

$$I(X_1,\ldots,X_n) = \left(\bar{X}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Reemplazamos los datos que tenemos y nos queda el siguiente intervalo:

$$I = \left(5-1.699127\frac{\sqrt{2}}{\sqrt{30}}, 5+1.699127\frac{\sqrt{2}}{\sqrt{30}}\right) = (4.5612873, 5.4387127).$$

Luego el promedio de ventas pertenece a este intervalo con una confianza de 90%.

Si hubiéramos usado que $\frac{\bar{X}-\mu}{S}\sqrt{n} \approx N(0,1)$, entonces el intervalo sería el mismo reemplazando donde dice $t_{n-1,\alpha/2}$ por $z_{\alpha/2}$. En este caso, $z_{\alpha/2} = 1.644854$ (que es muy similar al valor obtenido en la t de Student). El intervalo final queda:

$$I = \left(5 - 1.644854\frac{\sqrt{2}}{\sqrt{30}}, 5 + 1.644854\frac{\sqrt{2}}{\sqrt{30}}\right) = (4.575301, 5.424699).$$

Notar que este intervalo tiene menor longitud que el anterior, por lo que la aproximación por $N(0,1)$ es que mejor que la aproximación por t de Student. Este nos dice que es preferible elegir la aproximación por normal que por t de Student para este intervalo.

Para el segundo ítem, notamos que ahora sí conocemos la distribución de las ventas por trabajador: es una normal con varianza y media desconocida. Para ello utilizamos la siguiente función pivote:

$$T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Notemos que en esta expresión de la función pivote no aparece el parámetro desconocido correspondiente a la media. Ahora construimos el intervalo a partir de esta función pivote. En este caso $\alpha = 0.10$ y $n = 30$. Luego, $\chi^2_{n-1,1-\alpha/2} = 17.70837 =: a$ y $\chi^2_{n-1,\alpha/2} = 42.55697 =: b$.

$$\begin{aligned}
1-\alpha &= P(a \leq T \leq b) \\
&= P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) \\
&= P\left(\frac{1}{b} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{a}\right) \\
&= P\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) \\
&= P\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right).
\end{aligned}$$

Por lo tanto, el intervalo de confianza de nivel exacto 0.90 para la varianza de una normal con media desconocida es

$$I(X_1,\ldots,X_n) = \left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right).$$

Reemplazando los valores que tenemos, el intervalo final para la varianza es

$$I = \left(\frac{29(2)}{42.55697}, \frac{29(2)}{17.70837}\right) = (1.36288, 3.27529).$$

Nota: R por defecto calcula probabilidades, cuantiles, etc., de las distribuciones utilizando la cola inferior. Es decir, calcula a partir de la función de distribución acumulada $P(X \leq t) = p$. Si queremos que calcule mirando la cola superior, es decir $P(X \geq t) = p$ tenemos que especificar el parámetro `lower.tail = FALSE`, como hicimos en este ejemplo.

Ejemplo 2. Sean $X_1, \dots, X_n$ variables aleatorias iid $\mathcal{E}(\lambda)$.

1. Hallar dos intervalos de confianza de nivel asintótico $1-\alpha$ para λ.
2. Ídem para $\frac{1}{\lambda}$. ¿Cuál de los dos intervalos propuestos conviene si se usa como criterio elegir al de menor longitud esperada?

Solución. Notemos que $\mathbb{E}(\mathcal{E}(\lambda)) = \frac{1}{\lambda}$. Recordemos que $\sum_i X_i \sim \Gamma(n, \lambda)$, y que $2\lambda \sum_i X_i \sim \Gamma(n, \frac{1}{2}) = \Gamma(\frac{2n}{2}, \frac{1}{2}) = \chi^2_{2n}$. Luego el intervalo de confianza de nivel (exacto) $1-\alpha$ utilizando como función pivote $T = 2\lambda \sum_i X_i$ es:

$$I_1(X_1, \dots, X_n) = \left(\frac{\chi^2_{2n,1-\alpha/2}}{2\sum_i X_i}, \frac{\chi^2_{2n,\alpha/2}}{2\sum_i X_i} \right).$$

Ahora busquemos un intervalo asintótico teniendo en mente el teorema central del límite. Proponemos

$$T = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}.$$

Notemos que en este caso, como las variables son exponenciales, $\mu = \frac{1}{\lambda}$ y $\sigma = \frac{1}{\lambda}$, por lo que

$$T = \frac{\bar{X} - \frac{1}{\lambda}}{\frac{1}{\lambda}}\sqrt{n} = (\lambda \bar{X} - 1)\sqrt{n} \approx N(0,1) \quad \text{(TCL)}.$$

Ahora calculemos el intervalo asintótico:

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} \le N(0,1) \le z_{\alpha/2}) \\
&\approx P(-z_{\alpha/2} \le T \le z_{\alpha/2}) \\
&= P(-z_{\alpha/2} \le (\lambda\bar{X} - 1)\sqrt{n} \le z_{\alpha/2}) \\
&= P\left(\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right)\frac{1}{\bar{X}} \le \lambda \le \left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\frac{1}{\bar{X}} \right).
\end{aligned}$$

Por lo tanto, un intervalo de confianza de nivel asintótico $1-\alpha$ para λ es

$$I_2(X_1, \dots, X_n) = \left(\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right)\frac{1}{\bar{X}}, \left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)\frac{1}{\bar{X}} \right).$$

Para $\frac{1}{\lambda}$ podemos dar vuelta los extremos de los intervalos e invertirlos:

$$J_1(X_1, \dots, X_n) = \left(\frac{2\sum_i X_i}{\chi^2_{2n,\alpha/2}}, \frac{2\sum_i X_i}{\chi^2_{2n,1-\alpha/2}} \right).$$

$$J_2(X_1, \dots, X_n) = \left(\frac{\bar{X}}{\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)}, \frac{\bar{X}}{\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right)} \right).$$

Para poder decidir entre los dos intervalos, tengo que calcular la esperanza de la longitud de cada intervalo y quedarme con el que tenga la menor.

$$\mathbb{E}(L(J_1)) = \mathbb{E}\left(2\sum_i X_i \left(\frac{1}{\chi^2_{2n,1-\alpha/2}} - \frac{1}{\chi^2_{2n,\alpha/2}} \right) \right) = \frac{2n}{\lambda}\left(\frac{1}{\chi^2_{2n,1-\alpha/2}} - \frac{1}{\chi^2_{2n,\alpha/2}} \right).$$

Sea $z = z_{\alpha/2}$.

$$\mathbb{E}(L(J_2)) = \mathbb{E}\left(\bar{X}\left(\frac{1}{1-\frac{z}{\sqrt{n}}} - \frac{1}{1+\frac{z}{\sqrt{n}}}\right)\right) = \frac{1}{\lambda}\left(\frac{1}{1-\frac{z}{\sqrt{n}}} - \frac{1}{1+\frac{z}{\sqrt{n}}}\right).$$

En este punto para comparar las longitudes, podemos tomar el cociente de las dos esperanzas y notar que no depende de λ, solo de α y de los cuantiles de $\alpha/2$ para la χ^2_{2n} y para la $N(0,1)$. Notemos que

$$\frac{1}{1-\frac{z}{\sqrt{n}}} - \frac{1}{1+\frac{z}{\sqrt{n}}} = \frac{2nz}{n-z^2}.$$

El cociente de $\mathbb{E}(L(J_1))$ por $\mathbb{E}(L(J_2))$ nos queda

$$\frac{\mathbb{E}(L(J_1))}{\mathbb{E}(L(J_2))} = \frac{\frac{2n}{\lambda}\left(\frac{1}{\chi^2_{2n,1-\alpha/2}} - \frac{1}{\chi^2_{2n,\alpha/2}}\right)}{\frac{1}{\lambda}\left(\frac{2nz}{n-z^2}\right)} = \left(\frac{n-z^2}{z}\right)\left(\frac{1}{\chi^2_{2n,1-\alpha/2}} - \frac{1}{\chi^2_{2n,\alpha/2}}\right).$$

Notar que ambas longitudes tienden a 0 cuando n es grande. Si analizamos los diferentes valores de este cociente:

- Si es $= 1$: ambos tienen la misma longitud.
- Si es > 1: $\mathbb{E}(L(J_2))$ tiende a 0 más rápido que $\mathbb{E}(L(J_1))$ y por lo tanto se espera que J_2 sea el de menor longitud.
- Si es < 1: $\mathbb{E}(L(J_1))$ tiende a 0 más rápido que $\mathbb{E}(L(J_2))$ y por lo tanto se espera que J_1 sea el de menor longitud.

Podemos hacer una función en R que nos devuelva el valor de este cociente para diferentes valores de n para un α fijo. Por ejemplo, podemos emplear el siguiente código:

```
cociente <- function(n,alpha){
z <- qnorm(alpha/2, lower.tail = FALSE);
chi1 <- qchisq(alpha/2, df = 2*n, lower.tail = FALSE);
chi2 <- qchisq(1-alpha/2, df = 2*n, lower.tail = FALSE);
return((n-z*z)/z * (1/chi2  - 1/chi1));
}
```

Obtenemos los siguientes valores al corregir el código anterior:

```
> cociente(20,0.10)
[1] 0.2080578
> cociente(20,0.05)
[1] 0.1984951
> cociente(30,0.05)
[1] 0.1694645
> cociente(50,0.05)
[1] 0.1355284
```

Por lo tanto podemos decir que J_1 tendrá menor longitud esperada que J_2.

Un análisis similar puede hacer con I_1 e I_2. Para poder decir con total certeza cuál es el de menor longitud esperada, deberíamos ser capaces de encontrar una función inversa a la acumulada de manera de obtener los valores de $z_{\alpha/2}$ y $\chi^2_{2n,\alpha/2}, \chi^2_{2n,1-\alpha/2}$ explícitamente en función de n y α. ¿Se les ocurre cómo se podría hacer?