Inferencia Estadística

Marisol García Peña

Departamento de Matemáticas
Pontificia Universidad Javeriana

Bogotá, 2022

Ejemplo

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una $N(\mu, \sigma^2)$. Construya un intervalo del $\gamma = (1 - \alpha)\,\%$ para σ^2 con μ conocido.

Estimación por intervalos (cont.)

Ejemplo

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una $N(\mu, \sigma^2)$. Construya un intervalo del $(1-\alpha)\,\%$ para μ con σ^2 desconocido.

Estimación por intervalos (cont.)

Ejemplo

Suponga que se obtiene una única observación de Y de una distribución exponencial con media θ. Use Y para formar un intervalo de confianza para θ con $(1-\alpha) = 0.90$

Ejemplo

Suponga que se toma una muestra aleatoria de tamaño $n = 1$ de una distribución uniforme definida en $[0, \theta]$, donde θ es desconocido. Encuentre un limite inferior de confianza del 95 % para θ.

Ejemplo

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $U(0, \theta)$. Construya un intervalo del 90 % para θ e interprete. Inicie con el estimador de máxima verosimilitud de θ.

Estimación por intervalos (cont.)

Método estadístico

- $T = t(X_1, \ldots, X_n)$ una estadística.
- Selección de la estadística de distintas maneras.
- Estadística suficiente, sino existe $\Longrightarrow$ estimador puntual (EMV).
- La selección de T depende de la facilidad de las operaciones que se necesiten para encontrar el intervalo de confianza.
- Una de las opercaciones $\Longrightarrow$ f.d.p. de T.

Estimación por intervalos (cont.)

- $f_T(t, \theta)$ f.d.p. de T.
- Caso continuo, se definen dos funciones $h_1(\theta)$ y $h_2(\theta)$.
- $\displaystyle\int_{-\infty}^{h_1(\theta)} f_T(t, \theta)dt = p_1$ y $\displaystyle\int_{h_2(\theta)}^{\infty} f_T(t, \theta)dt = p_2$.
- p_1 y p_2 son números fijos tales que $0 < p_1$, $0 < p_2$ y $p_1 + p_2 < 1$.
- h_1 y h_2 son estrictamente monótonas, se asume que $h_1(\theta) < h_2(\theta)$.

Estimación por intervalos (cont.)

- $P[h_1(\theta) < t(X_1, \ldots, X_n) < h_2(\theta)] = 1 - p_1 - p_2$.
- $P[v_1(X_1, \ldots, X_n) < \theta < v_2(X_1, \ldots, X_n)] = 1 - p_1 - p_2$.
- $P[V_1 < \theta < V_2]$ es el IC para θ.

Intervalos de confianza asintóticos/muestras grandes

- En la práctica es díficil encontrar las cantidades pivotales.
- Una solución es utilizar la distribución asintótica de los EMV.
- $T_n = t_n(X_1, \ldots, X_n)$ una secuencia de estimadores de θ cuya densidad es asintóticamente distribuida normal con media θ y varianza $\sigma_n^2(\theta)$, una función de θ.
- Para muestras grandes $\Longrightarrow$ EMV de θ se distribuye aproximadamente normal alrededor de θ.
- Varianza del EMV para muestras grandes

$$\sigma_n^2(\theta) = \frac{1}{nE\left[\left\{\frac{\partial}{\partial\theta}\log f(X,\theta)\right\}^2\right]} = -\frac{1}{nE\left[\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)\right]}$$

Estimación por intervalos (cont.)

- $\dfrac{T_n - \theta}{\sigma_n(\theta)} \Longrightarrow$ Cantidad pivotal.

- $-z < \dfrac{T_n - \theta}{\sigma_n(\theta)} < z.$

- $z = z_{(1+\gamma)/2} = z_{(\alpha/2)}$ definido como $\Phi(z_{(\alpha/2)}) = \alpha/2$ o $\Phi(z) - \Phi(-z) = 1 - \alpha = \gamma$.

- Otra forma de verlo $\sqrt{n}(T - \theta) \xrightarrow[n\to\infty]{D} N(0, I_F^{-1}(\theta))$.

- $$z = \frac{\sqrt{n}(T - \theta)}{\sqrt{I_F^{-1}(\theta)}} \xrightarrow[n\to\infty]{D} N(0, 1).$$

- Si se está interesado en encontrar IC para una función de θ, $\tau(\theta) \Longrightarrow$ Usar método Delta.

Ejemplo

Sea $X_1, \ldots, X_n$ uma muestra aleatoria de una población con $f(x,\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$. Encuentre el IC para θ.

Intervalo de confianza para la media con σ conocida

Intervalo de confianza para μ con σ conocida

$$\overline{X} \pm Z_{(1-\frac{\alpha}{2})}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- Cantidad pivotal $\Longrightarrow \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$.
- σ conocido.
- n "grande", $n \geq 30$.
- Distribución simétrica sin valores atípicos.
- Muestra de una distribución normal.
- $Z = 1.645, \alpha = 0.1$; $Z = 1.960, \alpha = 0.05$; $Z = 2.576, \alpha = 0.01$.

Intervalo de confianza para la media con σ conocida (cont.)

Ejemplo

El volumen envasado en botellas de $\frac{1}{2}$ litro de cierta gaseosa es normalmente distribuido. De experimentos anteriores se sabe que $\sigma = 1.20ml$. Una muestra de 10 botellas proporciona una media de $\overline{X} = 503.4ml$.

Intervalo de confianza para la media con σ conocida (cont.)

- Población normal $\Longrightarrow$ Media muestral normalmente distribuida para cualquier tamaño de muestra.

- IC $(90\,\%)$ $503.4 \pm 1.645\left(\frac{1.20}{\sqrt{10}}\right) \Longrightarrow 502.78 \leq \mu \leq 504.02$

- IC $(95\,\%)$ $503.4 \pm 1.960\left(\frac{1.20}{\sqrt{10}}\right) \Longrightarrow 502.66 \leq \mu \leq 504.14$

- IC $(99\,\%)$ $503.4 \pm 2.576\left(\frac{1.20}{\sqrt{10}}\right) \Longrightarrow 502.42 \leq \mu \leq 504.38$

- En la realidad, es poco común que σ sea conocida.

- Procesos de control de calidad puede ser razonable asumir que σ se mantenga constante a lo largo del tiempo.

Intervalo de confianza para la media con σ desconocida

- Población normal, pero σ desconocida $\Longrightarrow$ distribución t.
- Tamaño de muestra pequeño.
- Distribución t es simétrica y tiene forma similar a la distribución normal.
- La distribución t depende del tamaño de la muestra.
- Muestra grande $\Longrightarrow$ valores de la distribución t, se aproximan a los de la normal.
- Para un nivel/coeficiente de confianza dado $\Longrightarrow$ t siempre es mayor que z $\Longrightarrow$ IC más amplio.
- Cantidad pivotal $\Longrightarrow \dfrac{\overline{X}-\mu}{S/\sqrt{n}}$.

Intervalo de confianza para la media con σ desconocida (cont.)

Condiciones para uso de la distribución t

- Distribución de la población es normal.
- Distribución de la muestra es simétrica, unimodal, sin valores atípicos (outliers) y tamaño de muestra ≤ 15.
- Distribución de la muestra es moderadamente asimétrica, unimodal, sin valores atípicos y el tamaño de muestra $16 \leq n \leq 30$.
- Tamaño de muestra es mayor de 30, sin valores atípicos (en este caso, puede usarse la distribución normal).

Intervalo de confianza para la media con σ desconocida (cont.)

Intervalo de confianza para μ con σ desconocida

$$\overline{X} \pm t_{(1-\frac{\alpha}{2};n-1)} \left(\frac{S}{\sqrt{n}} \right)$$

Ejemplo

Muestra aleatoria de puntaciones para un MBA, obtenidas por 20 candidatos a este programa.

530	450	600	570	360
550	640	490	460	550
480	440	530	470	560
500	430	640	420	530

Intervalo de confianza del 90 % para la puntuación media de todos los candidatos al MBA.

Intervalo de confianza para la media con σ desconocida (cont.)

- $\overline{X} = 510$, $S = 73.77$.
- $n = 20 \Longrightarrow v = 19$
- σ desconocida $\Longrightarrow t_{0.90} = 1.729$.
- `=inv.t.2c(prob=`α`;g.l.)`.
- `=inv.t(prob=`$1-\frac{\alpha}{2}$`;g.l.)`.
- IC $(90\,\%) \Longrightarrow 510 \pm 1.729\left(\dfrac{73.77}{\sqrt{20}}\right)$
- $481.48 \leq \mu \leq 538.52$

Intervalo de confianza para la media con σ desconocida (cont.)

Ejemplo

En un determinado período, un hospital tuvo 8261 casos en maternidad. El diagnóstico más común es parto normal sin complicaciones, con código 373. Durante el período de estudio se encontraton 4409 de estos casos.

Es necesario conocer la media del tiempo de hospitalización para que pueda planearse la capacidad de camas en la unidad de maternidad y los turnos para las enfermeras.

Son seleccionados aleatoriamente $n = 25$ registros hospitalarios de nacimientos, que proporciona una estimación de la media de hospitalización de $\overline{X} = 39.144$ horas con una desviación estándar de $S = 16.204$ horas.

¿Cuál es el intervalo de confianza del 95 % para la media poblacional?

Intervalo de confianza para la media con σ desconocida (cont.)

- Se asume población normal.
- $\alpha = 0.05,\ t_{24} = 2.064.$
- IC $(95\,\%) \Longrightarrow 32.455 \leq \mu \leq 45.833$

Longitud del intervalo

Depende del tamaño de la muestra, nivel de confianza y la desviación estándar.
Para disminuir la longitud del intervalo $\Longrightarrow$ aumentar tamaño de la muestra o disminuir el nivel de confianza.

Distribución t asume normalidad $\left\{\begin{array}{l}\text{Siempre y cuando la distribución no} \\ \text{sea excesivamente asimétrica.} \\ \text{Si es cuestionable, puede aumentarse} \\ \text{el tamaño de la muestra.}\end{array}\right.$

Intervalo de confianza para la proporción π

$$p \pm Z_{(1-\frac{\alpha}{2})}\sqrt{\frac{p(1-p)}{n}}$$

Ejemplo

Una muestra de 75 compras al por menor mostró que 24 fueron pagadas en efectivo. Construir un intervalo de confianza del 95 % de todas las compras al por menor pagadas en efectivo.

- Proporción muestral de compras pagadas en efectivo.
- $p = \dfrac{x}{n} = \dfrac{24}{75} = 0.32.$
- p se distribuye normalmente $\Longrightarrow np = 75(0.32) = 24$ y $n(1-p) = 75(0.68) = 51$, son mayores de 10.

- $\sigma_p = \sqrt{\frac{0.32(1-0.32)}{75}}$.
- IC $(95\,\%) \Longrightarrow 0.214 \leq \pi \leq 0.426$

Disminuyendo el intervalo

- Longitud del intervalo $\Longrightarrow$ tamaño de muestra, coeficiente de confianza y proporción muestral.
- Aumentar n, disminuir el nivel de confianza.

Ejemplo

Una muestra aleatoria de 200 páginas del directorio telefónico reveló que 30 de las páginas seleccionadas contenían por lo menos un aviso con más de un color. ¿Cuál es el IC del 90 % para la proporción de todas las páginas con por lo menos uno de estos avisos?

- $p = \frac{30}{200} = 0.15$, proporción muestral de páginas con por lo menos un aviso con más de un color.
- $np = 200(0.15) = 30$ y $n(1-p) = 200(0.85) = 170 \Longrightarrow$ Asume normalidad.
- $Z = 1.645 \Longrightarrow$ IC $(90\,\%) \Longrightarrow 0.108 \leq \pi \leq 0.192$.

Intervalo de confianza de score para proporción

Considere una muestra aleatoria de tamaño n de una población con parámetro π que indica la verdadera proporción de una característica binaria, $0 < \pi < 1$. X denota el número en la muestra con esa característica y sea $p = \frac{X}{n}$ la proporción muestral. El intervalo de confianza $(1-\alpha)100\,\%$ para π es

$$Lim.Inf. = \frac{p + q^2/(2n) - q\sqrt{p(1-p)/n + q^2/(4n^2)}}{1 + q^2/n}$$

$$Lim.Sup. = \frac{p + q^2/(2n) + q\sqrt{p(1-p)/n + q^2/(4n^2)}}{1 + q^2/n}$$

donde q es el percentil $(1 - \frac{\alpha}{2})$ de la distribución $N(0,1)$.
También se le conoce como intervalo de score de Wilson.

Intervalo de confianza para la varianza σ^2

Si se toma una muestra de tamaño n de una población normal.

$$\frac{(n-1)S^2}{\chi^2_U} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_L}$$

- $\chi^2_L = \chi^2_{\frac{\alpha}{2};n-1}$ y $\chi^2_U = \chi^2_{1-\frac{\alpha}{2};n-1}$, percentiles de la cola inferior y superior de la distribución Chi-cuadrado.
- Se asume población normal.
- Si se quiere encontrar un IC para $\sigma \Longrightarrow$ tomar la raíz de los límites del IC para σ^2.
- `inv.chicuad.cd(prob=`$\frac{\alpha}{2}$ `o` $1-\frac{\alpha}{2}$`;g.l.)`

Intervalo de confianza para la varianza σ^2 (cont.)

Ejemplo

Se desea estimar la varianza del nivel de nistamina en un ungüento. Se conoce por larga experiencia que sigue una distribución normal. Se toma una muestra de 9 ungüentos, dando el nivel siguiente (en millones de unidades/gr): $1; 0.9; 1.5; 2.8; 3.1; 3.2; 2.5; 1.9; 2$. Estimar la varianza mediante dos intervalos de confianza, al 99 % y 95 %.

- $S^2 = 0.74$
- Caso 1. $1-\alpha = 0.99$, $\frac{\alpha}{2} = 0.005$, $\chi^2_U = \chi^2_{1-\frac{\alpha}{2};n-1} = 21.9549$, $\chi^2_L = \chi^2_{\frac{\alpha}{2};n-1} = 1.3444$.
- IC $(99\,\%)$ $0.2696 \leq \sigma^2 \leq 4.4034$.
- Caso 2. $1-\alpha = 0.95$, $\frac{\alpha}{2} = 0.025$, $\chi^2_U = \chi^2_{1-\frac{\alpha}{2};n-1} = 17.5345$, $\chi^2_L = \chi^2_{\frac{\alpha}{2};n-1} = 2.1797$.
- IC $(95\,\%)$ $0.3376 \leq \sigma^2 \leq 2.7159$.

Tamaño de muestra para la media

Se quiere estimar la media poblacional con error admisible $\pm E$ (precisión).

$$\overline{X} \pm Z_{(1-\frac{\alpha}{2})}\left(\frac{\sigma}{\sqrt{n}}\right) \Longrightarrow \overline{X} \pm E$$

- Se usa Z en lugar de t, porque la distribución t necesita conocer n.
- Igualando el error admisible (E) a la semiamplitud del IC.
- n se aproxima al mayor entero más próximo.

Tamaño de muestra para la media

$$E = Z\frac{\sigma}{\sqrt{n}} \Longrightarrow E^2 = \frac{Z^2\sigma^2}{n} \Longrightarrow n = \frac{Z^2\sigma^2}{E^2}$$

$$n = \left(\frac{z\sigma}{E}\right)^2.$$

¿Cómo estimar σ?

Método 1 Muestra piloto. Seleccionar una muestra, calcular S y reemplazar por σ.

Método 2 Asumir una población uniforme. $\sigma = \left[\frac{(b-a)^2}{12}\right]^{1/2}$.

El peso de un camión liviano varía entre 1500 y 3500 libras, entonces, $\sigma = 577$ libras.

La distribución uniforme no tiene tendencia central.

El verdadero σ puede ser más pequeño que el supuesto inicial.

Se obtiene un valor para n más grande del necesario.

¿Cómo estimar σ?

Método 3 Asumir una población normal. $\sigma = \frac{b-a}{4}$.

Con la mayor parte de los datos dentro de $\mu + 2\sigma$ y $\mu - 2\sigma$, de tal forma que la amplitud es de 4σ.

Tomando el ejemplo anterior, $a = 1500$ libras y $b = 3500$ libras, $\sigma = 500$ libras.

Método 4 Llegadas Poisson. $\sigma = \sqrt{\mu}$.

Si se cree que la tasa de llegadas es de 20 clientes por hora, entonces, $\sigma = \sqrt{20} = 4.47$.

Ejemplo

Un gerente de producción quiere estimar el peso medio de cebollas españolas que serán entregadas por un proveedor, con 95 % de confianza y un error de ± 1 onza. Muestra piloto de 12 cebollas presenta una desviación estándar de 3.60 onzas. ¿Cuál es el tamaño de muestra adecuado?

- $Z = 1.96$, $S = 3.60$, $E = 1$
- $n = \left[\frac{1.96(3.60)}{1}\right]^2 = 49.79 \approx 50$ cebollas.
- Una pequeña modificación en $E \Longrightarrow$ efecto en n.
- Suponga $E = 0.5$ onzas $\Longrightarrow$ Estimación más precisa.
- $n = \left[\frac{1.96(3.60)}{0.5}\right]^2 = 199.1 \approx 200$ cebollas.