Análisis estadístico de datos simulados
Estimadores

Patricia Kisbye

FaMAF

11 de mayo, 2010

Análisis estadístico

Inferencia estadística:

- Elegir una distribución en base a los datos observados.
- Estimar los parámetros de la distribución (EMV).
- Pruebas de bondad de ajuste.

Estimación de parámetros

- Varianza del estimador. Var($\hat{\theta}$).
- Error cuadrático medio del estimador. $E[(\hat{\theta} - \theta)^2]$.
- Estimadores por intervalo e intervalos de confianza.
- Pruebas de hipótesis. Nivel de significación α. Valor p.

Media muestral

Dadas n observaciones: $X_1, X_2, \ldots, X_n$, con una misma distribucón, la media muestral se define por

$$\overline{X}(n) = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

La media muestral se utiliza como un estimador de la media θ, es decir, de $\theta = E[X_i]$.

Estimador insesgado.

$$E[\overline{X}(n)] = E\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] = \sum_{i=1}^{n} \frac{E[X_i]}{n} = \frac{n\theta}{n} = \theta.$$

Error cuadrático medio

- $\hat{\theta}$: estimador del parámetro θ de una distribución F
- Se define el error cuadrático medio (ECM) de $\hat{\theta}$ con respecto al parámtero θ como

$$ECM(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2].$$

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2] + (E[\hat{\theta}] - \theta)^2 \\ &= \text{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2 \end{aligned}$$

- El error cuadrático medio de un estimador es igual a su varianza más el sesgo al cuadrado.
- Si el estimador es insesgado, su ECM es igual a la varianza.

ECM de la media muestral respecto de la media

Muestra de X: $X_1, X_2, \ldots, X_n, \qquad E[X_i] = \theta$

$$\begin{aligned} ECM(\overline{X}(n), \theta) &= E[(\overline{X}(n) - \theta)^2] \\ &= \text{Var}(\overline{X}(n)) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

La media muestral es un buen estimador de $E[X]$ si $\sigma/\sqrt{n}$ es pequeño.

- El ECM depende de la distribución de X_i y del tamaño de la muestra.
- Teorema central del límite. Si $Z \sim N(0,1)$ y n es grande:

$$P\left(\frac{|\overline{X}(n) - \theta|}{\sigma/\sqrt{n}} > c\right) \approx P\{|Z| > c\}.$$

Varianza muestral

El indicador $\frac{\sigma^2}{n}$ como estimación del error en la media muestral, tiene el inconveniente que σ es en general desconocida.

Para estimar la varianza se utiliza el estimador

$$S^2(n) = \frac{\sum_{i=1}^{n}(X_i - \overline{X}(n))^2}{n-1}.$$

- Estimador insesgado de la varianza
- Fórmula a utilizar:

$$E\left[S^2(n)\right] = \mathrm{Var}(X)$$

$$\sum_{i=1}^{n}(X_i - \overline{X}(n))^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2(n)$$

Varianza muestral

$$
\begin{aligned}
E[X_i^2] &= \text{Var}(X_i) + (E[X_i])^2 = \sigma^2 + \theta^2. \\
E[\overline{X}^2(n)] &= \frac{\sigma^2}{n} + \theta^2. \\
(n-1)E[S^2(n)] &= nE[X_1^2] - nE[\overline{X}^2(n)] = n(\sigma^2 + \theta^2) - n(\frac{\sigma^2}{n} + \theta^2) \\
E[S^2(n)] &= \sigma^2
\end{aligned}
$$

Utilizaremos $S(n) = \sqrt{S^2(n)}$ como estimador de la desviación estándar.

- Error del estimador $\overline{X}(n)$: σ^2/n.
- Simulación de datos: Si el objetivo es estimar la media, para disminuir el error deben generarse muestras de tamaño n, n grande.

Media muestral

- Elegir un valor aceptable d para la desviación estándar del estimador.
- Generar (n) datos hasta que $\sigma/\sqrt{n} < d$. ($S/\sqrt{n} < d$)
- Conviene generar al menos 100 datos para:
  - asegurar normalidad de la distribución de $\overline{X}(n)$.
  - para disminuir la varianza de S.
- La estimación de θ estará dada por el último valor de $\overline{X}(n)$.
- El algoritmo implica calcular en cada paso $\overline{X}(n)$ y $S(n)$.
- Es posible calcularlo recursivamente.

Media muestral

Cálculo recursivo de $\overline{X}(n)$ y $S^2(n)$

- $\overline{X}(1) = X_1$,
- $S^2(1) = 0$.

$$\begin{aligned} \overline{X}(j+1) &= \overline{X}(j) + \frac{X_{j+1} - \overline{X}(j)}{j+1} \\ S^2(j+1) &= \left(1 - \frac{1}{j}\right) S^2(j) + (j+1)(\overline{X}(j+1) - \overline{X}(j))^2 \end{aligned}$$

Estimación de una proporción

El estimador $\overline{X}(n)$ puede utilizarse también para estimar la proporción de casos en una población.

$$X_i = \begin{cases} 1 & \text{probabilidad } p \\ 0 & \text{probabilidad } 1-p. \end{cases}$$

- $\overline{X}(n)$ es un estimador insesgado de p.
- $E[(\overline{X}(n) - p)^2] = \text{Var}(\overline{X}(n)) = \dfrac{p(1-p)}{n}$
- En este caso, se estima la varianza del estimador $\overline{X}(n)$ por:

$$\frac{\overline{X}(n)(1-\overline{X}(n))}{n}.$$

Algoritmo: Cálculo de $E[X]$

Estimación de la media M de X con error d

Generar X, $M \leftarrow X$ $\qquad$ $M = \overline{X}(1) = X_1$;

$S^2 \leftarrow 0$ $\qquad$ $S^2 = S^2(1) = 0$;

for $1 < j \leq 100$ **do**

  Generar X; $A \leftarrow M$;

  $M \leftarrow M + (X - M)/j$;

  $S^2 \leftarrow (1 - 1/(j-1))S^2 + j(M - A)^2$

end

$j \leftarrow 100$;

while $\sqrt{S^2/j} > d$ **do**

  $j \leftarrow j + 1$;

  Generar X;

  $A \leftarrow M$;

  $M \leftarrow M + (X - M)/j$;

  $S^2 \leftarrow (1 - 1/(j-1))S^2 + j(M - A)^2$

end

return M

Algoritmo: Cálculo de una probabilidad

Estimación de la probabilidad p de X con error d

Generar X $\quad$ X es 0 o 1;
$p \leftarrow X$;
for $1 < j \leq 100$ **do**
    Generar X;
    $p \leftarrow p + (X - p)/j$
end
$j \leftarrow 100$;
while $\sqrt{p(1-p)/j} > d$ **do**
    $j \leftarrow j + 1$;
    Generar X;
    $p \leftarrow p + (X - p)/j$;
end
return p

Estimador por intervalos

Un estimador por intervalo de un parámetro es un intervalo para el que se predice que el parámetro está contenido en él.
La confianza que se da al intervalo es la probabilidad de que el intervalo contenga al parámetro.

Estimador por intervalo de la media poblacional

- $\overline{X}(n)$ es un estimador puntual de la media.
- Si la población es normal con media θ y d.s. σ,

$$\frac{\overline{X}(n) - \theta}{\sigma/\sqrt{n}} \sim Z = N(0,1)$$

- $P(Z > z_\alpha) = \alpha$, para $0 < \alpha < 1$.
- Si el nivel de confianza deseado es $1 - \alpha$, utilizamos $\pm z_{\alpha/2}$.
- Ejemplo: nivel de confianza del 95%: $\alpha = 0.025$, y $z_\alpha = 1.96$.

Estimador por intervalos

$$P\left(\frac{|\overline{X}(n)-\theta|}{\sigma\sqrt{n}} \leq 1.96\right) = 0.95.$$

$$P\left(\overline{X}(n) - 1.96\frac{\sigma}{\sqrt{n}} \leq \theta \leq \overline{X}(n) + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

- El intervalo con extremos

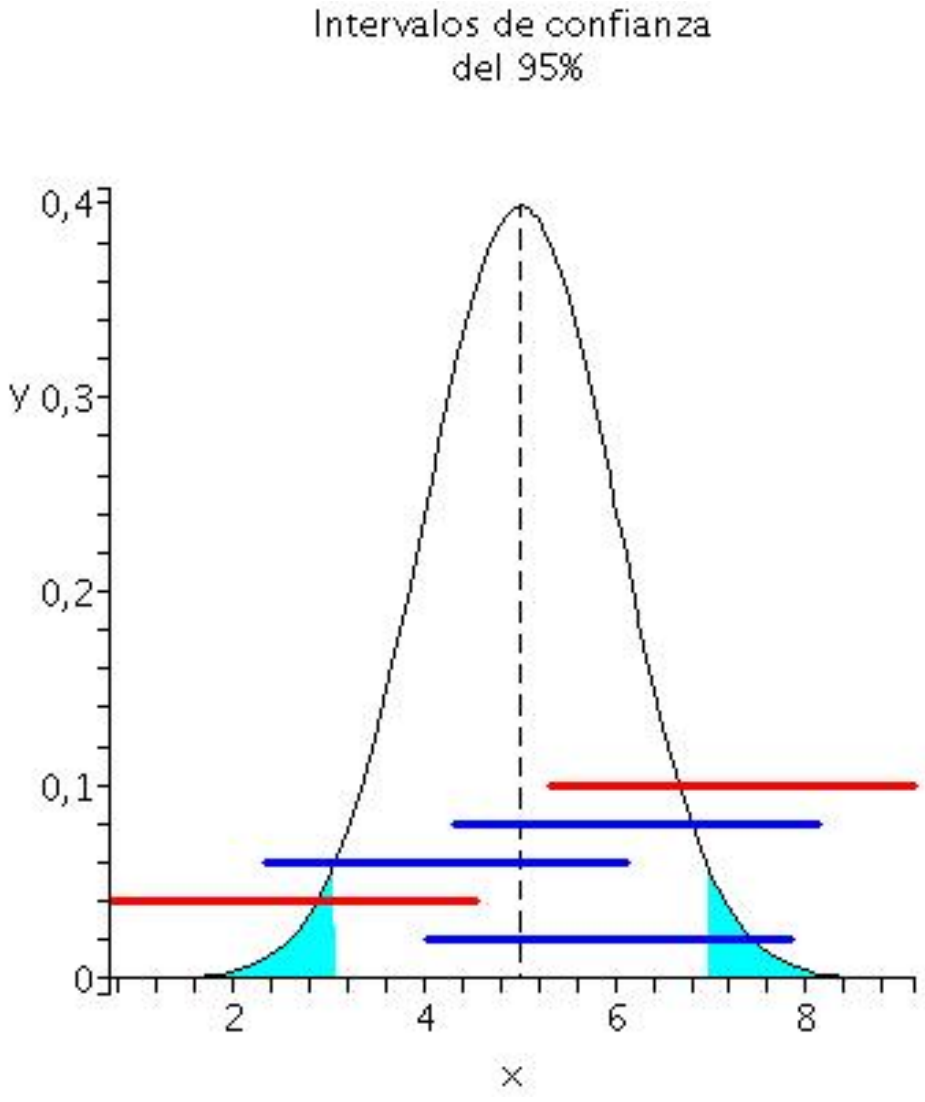
$$\overline{X}(n) - 1.96\,\sigma/\sqrt{n} \qquad \text{y} \qquad \overline{X}(n) + 1.96\,\sigma/\sqrt{n}$$

  se dice que es un estimador por intervalo, con un 95% de confianza para la media θ.
- Si $\overline{x}$ es un valor observado de $\overline{X}(n)$, el intervalo con extremos

$$\overline{x} - 1.96\,\sigma/\sqrt{n} \qquad \text{y} \qquad \overline{x} + 1.96\,\sigma/\sqrt{n}$$

  es el valor estimado del estimador por intervalo de θ, con un 95% de confianza.

Estimador por intervalos



- $(\overline{X} - \frac{1.96\sigma}{\sqrt{n}}, \overline{X} + \frac{1.96\sigma}{\sqrt{n}})$.
- $z_{0.025} = 1.96$.
- El 95% de los intervalos cubren la media.

Estimador por intervalos

- Si la varianza σ^2 es desconocida, utilizamos el estimador $S^2(n)$.
- Para determinar un intervalo de confianza, es necesario conocer la distribución del estadístico:

$$\sqrt{n}\,\frac{\overline{X}(n)-\theta}{S(n)}$$

Distribuciones derivadas de la normal

- χ^2 de Pearson con k grados de libertad: si $Z_1, Z_2, \ldots, Z_k$ son v.a. N(0,1), independientes:

$$\chi^2_k = Z_1^2 + \cdots + Z_k^2$$

- T_k de Student, con k grados de libertad: (W. S. Gosset)

$$T_k = \frac{Z}{\sqrt{\frac{\chi^2_k}{k}}}$$

Intervalos de confianza

- El estadístico tiene una distribución T_{n-1}:

$$\sqrt{n}\frac{\overline{X}(n)-\theta}{S(n)} \sim T_{n-1}$$

- Sea t_α tal que $\quad P(|T_{n-1}| > t_\alpha) = 1-\alpha.$

$$P\left(\overline{X}(n) - t_{\alpha/2}\frac{S(n)}{\sqrt{n}} \leq \theta \leq \overline{X}(n) + t_{\alpha/2}\frac{S(n)}{\sqrt{n}}\right) = 1-\alpha.$$

- Para $n > 120$, puede usarse la distribución normal, es decir, $t_\alpha \approx z_\alpha$.

Intervalos de confianza para proporciones

- $X_1, X_2, \ldots, X_n$: Bernoulli, independientes, con probabilidad p de éxito.
- Para n suficientemente grande tal que np y $n(1-p)$ es mayor que 5,

$$X_1 + \cdots + X_n = Bi(n,p) \sim N(np, np(1-p)).$$

- Si p es desconocido, podemos estimar p con la media muestral:

$$\hat{p} = \overline{X}(n) \quad \text{y} \quad \text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}.$$

- Intervalos de confianza del $100(1-\alpha)\%$:

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Longitud del intervalo de confianza

- Estimación de la media: $s(n)$: valor observado de la varianza muestral.

$$2\,\frac{z_{\alpha/2}\sigma}{\sqrt{n}} \quad \text{o} \quad 2\,\frac{z_{\alpha/2}s(n)}{\sqrt{n}}.$$

- Estimación de la proporción:

$$2z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- La longitud del intervalo de confianza al $100(1-\alpha)\%$ depende del tamaño de la muestra.