Análisis estadístico de datos simulados
Estimadores puntuales

Georgina Flesia

FaMAF

3 de mayo, 2016

Análisis estadístico

Modelización 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

- Estimador puntual.
- Varianza del estimador. $\text{Var}(\hat{\theta})$.
- Error cuadrático medio del estimador. $E[(\hat{\theta} - \theta)^2]$.
- Estimadores por intervalo e intervalos de confianza.

Estimación de parámetros

Dada una muestra de n datos observados, se llama estimador $\hat{\theta}$ del parámetro θ a cualquier función de los datos observados.

Propiedades de un buen estimador puntual

- Insesgabilidad: se dice que el estimador es insesgado si $E[\hat{\theta}] = \theta$.
- Consistencia: si al aumentar la muestra, el estimador se aproxima al parámetro.
- Eficiencia: se calcula comparando su varianza con la de otro estimador.
- Suficiencia: utiliza toda la información obtenida de la muestra.

Media muestral

Dadas n observaciones: $X_1, X_2, \ldots, X_n$, con una misma distribución, 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 $\theta = E[X_i]$, si la media es finita.

Estimador insesgado.

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

Media muestral

Dadas n observaciones: $X_1, X_2, \ldots, X_n$, con una misma distribución con media finita θ,

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

Estimador consistente.

$$\lim_{n \to \infty} [\overline{X}(n)] \quad = \quad \theta.$$

por la ley de los grandes números.

Media muestral

Dadas n observaciones: $X_1, X_2, \ldots, X_n$, con una misma distribución con media finita y varianza finita

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

$$Z = \sqrt{n}(\overline{X}(n) - \theta)/\sigma.$$

Estimador asintóticamente normal.

$$\lim_{n \to \infty} [F_Z(n)(x)] \quad = \quad \Phi(x).$$

por el teorema central del límite.

Métodos de estimación mas comunes

- Estimador de máxima verosimilitud
- Estimador de momentos

Estimador de máxima verosimilitud

Si la distribución supuesta es discreta para los datos observados, y se desconoce un parámetro θ.
Sea $p_\theta(x)$ la probabilidad de masa para dicha distribución.

Dado que se han observado datos $X_1, X_2, \ldots, X_n$, se define la función de máxima verosimilitud $L(\theta)$ como sigue:

$$L(\theta) = p_\theta(X_1) \cdot p_\theta(X_2) \cdots p_\theta(X_n).$$

El estimador de máxima verosimilitud es el valor $\hat{\theta}$ que maximiza $L(\theta)$:

$$L(\hat{\theta}) \geq L(\theta), \qquad \theta \text{ valor posible.}$$

Estimador de máxima verosimilitud

Si la distribución supuesta es continua, y $f_\theta(x)$ es la densidad para dicha distribución.

Dado que se han observado datos $X_1, X_2, \ldots, X_n$, se define la función de máxima verosimilitud $L(\theta)$ como sigue:

$$L(\theta) = f_\theta(X_1) \cdot f_\theta(X_2) \cdots f_\theta(X_n).$$

El estimador de máxima verosimilitud es el valor $\hat{\theta}$ que maximiza $L(\theta)$:

$$L(\hat{\theta}) \geq L(\theta), \qquad \theta \text{ valor posible.}$$

Estimador de máxima verosimilitud

El estimador de máxima verosimilitud tiene, en general, las siguientes propiedades:

1. Es único: $L(\hat{\theta}) > L(\theta)$ para cualquier otro valor de θ.
2. La distribución asintótica de $\hat{\theta}$ tiene media θ.
3. Es invariante: $\phi = h(\theta)$, entonces $\hat{\phi} = h(\hat{\theta})$.
4. La distribución asintótica es la normal.
5. Es fuertemente consistente: $\lim_{n\to\infty} \hat{\theta} = \theta$.

Distribución exponencial

Ejemplo

Para la distribución exponencial, $\theta = 1/\lambda$ ($\lambda > 0$) y $f_\lambda(x) = \lambda e^{-x\lambda}$ para $x \geq 0$.

$$\begin{aligned} L(\lambda) &= \left(\lambda e^{-X_1\lambda}\right)\left(\lambda e^{-X_2\lambda}\right)\cdots\left(\lambda e^{-X_n\lambda}\right) \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} X_i\right) \end{aligned}$$

Distribución exponencial

$$\begin{aligned}
\ln(L(\lambda)) &= \ln(\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} X_i\right)) \\
&= n\ln(\lambda) - \lambda \sum_{i=1}^{n} X_i \\
\frac{d}{d\lambda}\ln(L(\lambda)) &= \frac{n}{\lambda} - \sum_{i=1}^{n} X_i \\
&= 0
\end{aligned}$$

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}(n)} = \frac{1}{\text{Media muestral.}}$$

$$\hat{\theta} = \frac{1}{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}(n) = \text{ Media muestral.}$$

Distribución geométrica

Ejemplo

Para la distribución geométrica, $\theta = p$ y $p_p(x) = p(1-p)^{x-1}$ para $x = 1, 2, \ldots$.

$$\begin{aligned} L(p) &= p(1-p)^{x_1-1} \ldots p(1-p)^{x_n-1} \\ &= p^n (1-p)^{\sum_{i=1}^n (x_i-1)} \\ &= p^n (1-p)^{\sum_{i=1}^n x_i} (1-p)^{-n} \\ &= \left(\frac{p}{1-p}\right)^n (1-p)^{\sum_{i=1}^n x_i} \end{aligned}$$

$$\begin{aligned} \ln(L(p)) &= n \ln\left(\frac{p}{1-p}\right) + \ln(1-p) \sum_{i=1}^n X_i \\ &= n \ln(p) - n \ln(1-p) + \ln(1-p) \sum_{i=1}^n X_i \end{aligned}$$

Distribución geométrica

$$\begin{aligned}\frac{d}{dp}\ln(L(p)) &= \frac{d}{dp}[n\ln(p)+\ln(1-p)[\sum_{i=1}^{n}X_i - n]] \\ &= \frac{n}{p} - \frac{1}{1-p}(\sum_{i=1}^{n}X_i - n) = 0\end{aligned}$$

$$\begin{aligned}\frac{n}{p} &= \frac{1}{1-p}(\sum_{i=1}^{n}X_i) - n \\ 1-p &= p(\frac{1}{n}\sum X_i - 1) \\ 1 &= p + p(\frac{1}{n}\sum X_i) - p\end{aligned}$$

$$\hat{p} = \left(\frac{1}{n}\sum X_i\right)^{-1}$$

Estimadores de máxima verosimilitud:

Distribuciones continuas:

- Uniforme: $\hat{a} = \min\{X_i\}$, $\hat{b} = \max\{X_i\}$.
- Exponencial: $\hat{\theta} = \overline{X}(n)$.
- Gamma, Weibull: $\hat{\alpha}$ y $\hat{\beta}$ se resuelven numéricamente.
- Normal:

$$\hat{\mu} = \overline{X}(n), \quad \hat{\sigma} = \left[\frac{n-1}{n} S^2(n)\right]^{1/2} = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^{1/2}.$$

- Lognormal:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \log(X_i)}{n}, \qquad \hat{\sigma} = \left[\frac{\sum_{i=1}^{n}(\log(X_i) - \hat{\mu})^2}{n}\right]^{1/2}.$$

Estimadores de máxima verosimilitud

Distribuciones discretas:

- Binomial (t, p): si t es conocido, $\hat{p} = \overline{X}(n)/t$.
- Bernoulli: Caso binomial con $t = 1$ e igual p.
- Geométrica: $\hat{p} = \frac{1}{\overline{X}(n)}$.
- Binomial negativa (s, p): número de ensayos hasta el s-ésimo éxito. Si s es conocido: $\hat{p} = \frac{s}{\overline{X}(n)}$.
- Poisson: $\hat{\lambda} = \overline{X}(n)$.

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 parámetro 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

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

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 $E[X]$

Algorithm 1: Estimación de la media M de X con error d

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

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

for $2 \leq j \leq J_0$ **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

$j \leftarrow J_0$;

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

Algorithm 2: Estimación de la probabilidad p de X con error d

Generar X 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