Estimador de Máxima Verossimilhança

5.2

Introdução

- Assuma que deseja-se conhecer um parâmetro de interesse θ de certa característica dos elementos de uma população que possa ser representada por uma variável aleatória X com função densidade $f(x;\theta)$, em que θ é desconhecido.

- Assuma também que os valores $x_1,x_2,...,x_n$ da amostra aleatória $X_1,X_2,...,X_n$ de $f(x;\theta)$ foram observados.

- Baseado nos valores observados da amostra aleatória, queremos estimar o valor do parâmetro desconhecido θ.

- Na **estimação pontual**, o valor de alguma estatística $t(X_1,X_2,...,X_n)$ representa, ou **estima**, o parâmetro desconhecido θ.

Exemplo

- Suponha que uma urna contém bolas pretas e brancas e que a razão entre elas é de 3/1, mas não sabemos se há mais bolas pretas ou brancas. Assim, a probabilidade de retirar uma bola preta é 1/4 ou 3/4 .

- Se n bolas são retiradas da urna com reposição, a distribuição de X = número de bolas pretas é dada pela distribuição binomial

$$P(x; p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

em que p = 1/4 ou p = 3/4.

- Iremos retirar uma amostra de três bolas (n = 3) com reposição e tentar estimar o parâmetro desconhecido p da distribuição.

Resultados de x	0	1	2	3
$P(X = x; \frac{3}{4})$	$\frac{1}{64}$	$\frac{9}{64}$	$\frac{27}{64}$	$\frac{27}{64}$
$P(X = x; \frac{1}{4})$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$

- Se encontrarmos x = 0 em uma amostra de 3, a estimativa 0,25 para p deve ser preferida sobre 0,75 porque a probabilidade 27/64 é maior que 1/64, isto é, porque uma amostra com x = 0 e mais verossímil (no sentido de ter maior probabilidade) ter surgido de uma população com p = 1/4 do que de uma com p = 3/4 .

- O estimador pode ser definido como

$$\hat{p} = \hat{p}(x) = \begin{cases} 0{,}25, & \text{para } x = 0, 1 \\ 0{,}75, & \text{para } x = 2, 3. \end{cases}$$

- O estimador então seleciona para cada possível x o valor de p, dito $\hat{p}$, de tal forma que

$$P(x; \hat{p}) > P(x; p'),$$

em que p' e o valor alternativa de p.

Função de Verossimilhança

Definição: A **função de verossimilhança** de n variáveis aleatórias $X_1,X_2,...,X_n$ é definida como a densidade conjunta das n variáveis aleatórias, isto e, $L(\theta; x_1,x_2,...,x_n) = f(x_1,x_2,...,x_n ;\theta)$, que é considerada ser uma função de θ, com $\theta \in \Theta$, em que Θ é o espaço paramétrico.

Em particular, se $X_1,X_2,...,X_n$ é uma amostra aleatória da densidade $f(x;\theta)$, então a função de verossimilhança é $L(\theta; x_1,x_2,...,x_n) = \prod_{i=1}^{n} f(x_i;\theta)$

Interpretação: A função de verossimilhança $L(\theta; x_1,x_2,...,x_n)$ nos dá a verossimilhança relativa à um particular valor $x_1,x_2,...,x_n$ que a variável aleatória assume. Supondo θ conhecido, então um particular valor da variável aleatória é mais "provável que ocorra", ou mais "verossímil", quando o valor da função for o máximo.

O **Princípio da Verossimilhança** postula que para fazer inferência sobre uma quantidade de interesse θ só importa aquilo que foi realmente observado e não aquilo que "poderia" ter ocorrido mas efetivamente não ocorreu.

Estimador de Máxima Verossimilhança

Definição: Seja $L(\theta) = L(\theta; x_1,x_2,...,x_n)$ a função de verossimilhança das variáveis aleatórias $X_1,X_2,...,X_n$. Se $\hat{\theta}$ (em que $\hat{\theta}$ é função das observações $x_1,x_2,...,x_n$) é o valor de $\theta \in \Theta$ que maximiza $L(\theta)$, então $\hat{\theta}$ é o **estimador de máxima verossimilhança (EMV)** de θ, ou seja,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; x_1,x_2,...,x_n).$$

Se a função de verossimilhança conter k parâmetros, ou seja,

$$L(\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_k)$$

então os estimadores de máxima verossimilhança de $\theta_1,\theta_2,...,\theta_k$ são variáveis aleatórias (que dependem da amostra), em que $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ são valores em Θ que maximizam $L(\theta_1,\theta_2,...,\theta_k)$.

Procedimento usual

- O logaritmo natural da função de verossimilhança de θ é denotado por

$$\ell(\theta) = log[L(\theta)].$$

- Como o logaritmo é uma função crescente e contínua, o valor de θ que maximiza L(θ) também maximiza $\ell(\theta)$. Se $\ell(\theta)$ é variável, o estimador de máxima verossimilhança pode ser encontrado como a raiz da equação de verossimilhança

$$\ell'(\theta) = \frac{d\ell(\theta)}{d\theta} = 0.$$

- Para se concluir que e um ponto de máximo, é necessário verificar se

$$\ell''(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0.$$

Outras formas de encontrar o estimador

- Nos casos em que o suporte da distribuição de X depende de θ ou o máximo ocorre na fronteira de Θ, o estimador de máxima verossimilhança é em geral obtido inspecionando o gráfico da função de verossimilhança.

- Quando não é possível encontrar analiticamente o ponto de máximo da função de verossimilhança, podemos utilizar métodos numéricos, como o método de Newton-Raphson. Computacionalmente, no R temos a rotina "*mle*" do pacote "*stats4*" ou o "*optim*" do pacote "*stats*".

Exemplo

Exponencial

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória da variável aleatória X ~Exp(θ) com densidade

$$f(x;\theta) = \theta e^{-\theta x},$$

$\theta > 0$ e $x \geq 0$. Encontre o estimador de máxima verossimilhança para θ.

$$f(x;\theta) = \theta e^{-\theta x} \qquad \boxed{\text{III}}$$

$$L(\theta) = L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

$$\ell(\theta) = \log(L(\theta)) = n \log(\theta) - \theta \sum_{i=1}^{n} x_i$$

$$\frac{d\ell(\theta)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \rightarrow \frac{n}{\theta} = \sum_{i=1}^{n} x_i \qquad \hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}}$$

$$\frac{d^2\ell(\theta)}{d\theta^2} = \qquad < 0 \qquad \hat{\theta} = \frac{1}{\bar{X}} \text{ EMV p/ } \theta$$

Exemplo

Normal

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória da variável aleatória $X \sim N(\mu, \sigma^2)$, onde μ e σ^2 são desconhecidos. Temos então que $\boldsymbol{\theta} = (\mu, \sigma^2)$ e

$$f(x;\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$-\infty < x < \infty$, $-\infty < \mu < \infty$ e $\sigma^2 > 0$. Encontre o estimador de maxima verossimilhanca para $\boldsymbol{\theta}$.

$$L(\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \mathrm{e}^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \ell(\theta)}{\partial \mu} = 2\sum_{i=1}^{n} \frac{(x_i-\mu)}{2\sigma^2} = 0 \qquad \rightarrow \qquad \hat{\mu} = \bar{X}.$$

$$\ell(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{2\sigma^2}$$

$$\frac{\partial \ell(\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i-\bar{x})^2}{2\sigma^4} = 0 \qquad \rightarrow \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (X_i-\bar{X})^2.$$

Exemplo

Weibull

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória da variável aleatória X ~ Weibull(β,λ), onde β e λ são desconhecidos. Temos então que $\boldsymbol{\theta} = (\beta, \lambda)$

$$f(x;\theta) = \beta\lambda(\lambda x)^{\beta-1} e^{-(\lambda x)^{\beta}}$$

$x \geq 0$, $\beta > 0$ e $\lambda > 0$. Encontre o estimador de maxima verossimilhanca de $\boldsymbol{\theta}$.

$$L(\theta;X) = \prod_{i=1}^{n} \beta\lambda(\lambda x_i)^{\beta-1} e^{-(\lambda x_i)^{\beta}}$$

$$\ell(\theta;X) = n\log(\beta) + n\beta\log(\lambda) + (\beta-1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}(\lambda x_i)^{\beta}$$

$$\frac{\partial \ell(\theta;X)}{\partial \lambda} = \frac{n\beta}{\lambda} - \beta\lambda^{\beta-1}\sum_{i=1}^{n}(x_i)^{\beta}$$

$$\frac{\partial \ell(\theta;X)}{\partial \beta} = \frac{n}{\beta} + n\log(\lambda) + \sum_{i=1}^{n}\log(x_i) - \sum_{i=1}^{n}(\lambda x_i)^{\beta}\log(\lambda x_i).$$

Utilizando o R

```
> library(stats)
> x=c(3, 5, 6, 7, 8, 9, 10, 10, 12, 15, 15,18, 19, 20, 22, 25, 28, 30, 40, 45)
> n=length(x)
> log.vero.W=function(l){
+ lambda=l[1]
+ beta=l[2]
+ -( log(beta)*n + (beta-1)*sum(log(x)) + beta*log(lambda)*n - sum( (lambda*x)^beta )
+ )
+ }
> optim(c(1,1),log.vero.W)
$par
[1] 0.05132946 1.62517976

$value
[1] 73.91645

$counts
function gradient
     119       NA

$convergence
[1] 0
```

Propriedades dos Estimadores de Máxima Verossimilhança

O método da máxima verossimilhança é um método estabelecido de estimação de parâmetros de modelos estatístico, sendo utilizado por muitos estatísticos teóricos e práticos. O uso generalizado do método se deve às propriedades probabilísticas das estimativas produzidas por ele.

Assumindo-se que a função de verossimilhança satisfaz algumas propriedades matemáticas básicas, que são frequentemente alcançadas pelos modelos mais utilizados, os EMV tem as seguintes propriedades:

Consistência: os EMV são consistentes, ou seja, elas convergem em probabilidade para o valor do parâmetro. Ou seja, os EMV **são não-viciados para grandes amostras** ($n \rightarrow \infty$).

Eficiência Assintótica: O Teorema do Limite Inferior de Cramer-Rao afirma que, para um dado parâmetro qualquer, existe um limite inferior para a variância das estimativas não-viciadas. **Para grandes amostras**, os EMV atingem esse limite e, portanto, **têm a menor variância possível dentre as estimativas não-viciadas**.

Invariância: os EMV são **invariantes sob transformações monotônicas**. Por exemplo, seja $\widehat{\mu}$ um EMV que pode ser transformada para:

$$\widehat{\theta}_1 = \log(\widehat{\mu}),$$
$$\widehat{\theta}_2 = \sqrt{\widehat{\mu}} \text{ e}$$
$$\widehat{\theta}_3 = \mathrm{e}^{\widehat{\mu}},$$

então as estimativas $\widehat{\theta}_1$, $\widehat{\theta}_2$ e $\widehat{\theta}_3$ também são EMV.

Normalidade Assintótica: os EMV convergem em distribuição para distribuição normal. Ou seja, para **grandes amostras** os EMV **tem distribuição aproximadamente normal**.

Teorema (O princípio da invariância):

Sejam $X_1, X_2, \ldots, X_n$ uma amostra aleatória da variável aleatória X com função de densidade (ou de probabilidade) $f(x;\theta)$. Se $\hat{\theta}$ é um estimador de máxima verossimilhança de θ, então $g(\hat{\theta})$ é um estimador de máxima verossimilhança de $g(\theta)$.

Prova:

Provamos o resultado para o caso em que g e 1:1. Sendo g(.) uma função 1:1, temos que g(.) e inversível, de modo que $\theta = g^{-1}(g(\theta))$. Assim,

$$L(\theta;X) = L(\, g^{-1}(g(\theta))\, ;X),$$

de modo que $\hat{\theta}$ maximiza os dois lados da equação acima. Logo

$$\hat{\theta} = g^{-1}\left(\widehat{g(\theta)}\right),$$

portanto,

$$\widehat{g(\theta)} = g(\hat{\theta}),$$

ou seja, o estimador de maxima verossimilhanca de $g(\theta)$ e $g(\hat{\theta})$.

Distribuição em grandes amostras:

No caso em que a amostra é grande, o suporte A(x) = {x, $f(x;\theta) > 0$} seja independente de θ e que seja possível a troca de ordens das operações de derivação e de integração sob a distribuição da variável aleatória X, temos que

$$\sqrt{n}(\hat{\theta} - \theta) \overset{a}{\sim} N\left(0, \frac{1}{I_F(\theta)}\right),$$

e

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \overset{a}{\sim} N\left(0, \frac{(g'(\theta))^2}{I_F(\theta)}\right).$$

Assim, para grandes amostras, os estimadores de máxima verossimilhança de θ e $g(\theta)$ são aproximadamente não viciados, cujas variâncias coincidem com os correspondentes limites inferiores das variâncias dos estimadores não viciados de θ e $g(\theta)$. Neste caso, em grandes amostras o estimador de máxima verossimilhança é **eficiente**.

O estimador de máxima verossimilhança **nem sempre é não viciado**, mas muitas vezes ele pode ser modificado para que se torne não viciado.

Erro Quadrático Médio

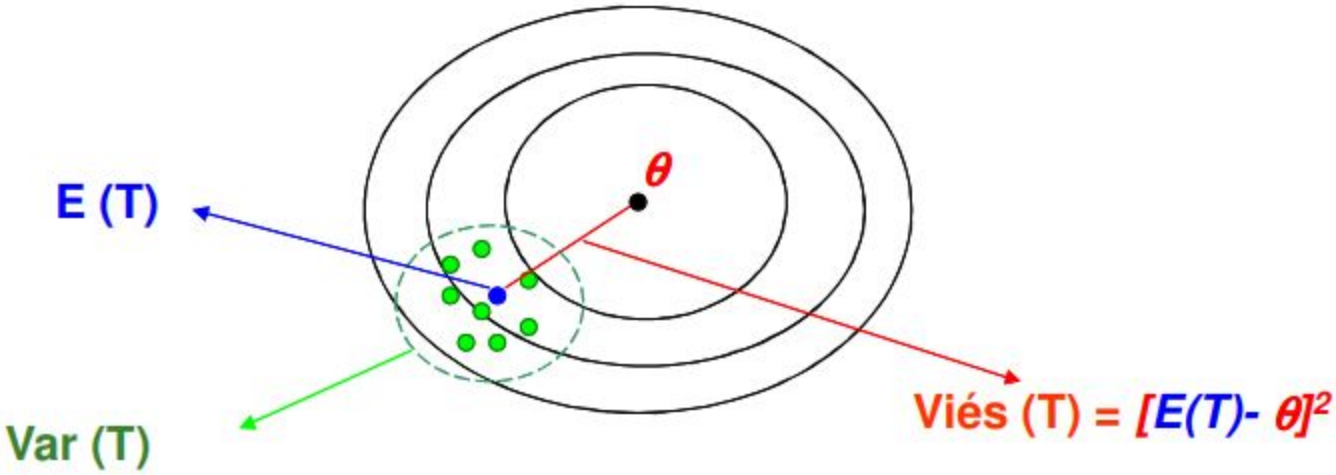
Nem sempre um estimador viciado não é bom, às vezes o que perdemos por ter um viés pequeno pode ser compensado pela concentração em torno do valor verdadeiro. De alguma forma temos que combinar os dois fatores: E($\hat{\theta}$) "próxima" de θ e Var($\hat{\theta}$) "próxima" de 0. Isto pode ser obtido através de uma medida muito útil de proximidade chamada erro quadrático médio (EQM).

Definição: Seja $\hat{\theta}$ um estimador de θ baseado em uma amostra aleatória $X_1, X_2, \ldots, X_n$. O **erro quadrático médio** (EQM) de $\hat{\theta}$ é:

$$EQM(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2] = Vício(\hat{\theta}, \theta)^2 + Var(\hat{\theta}),$$

em que Vício($\hat{\theta}$,θ) é o vício de $\hat{\theta}$ e Var($\hat{\theta}$) é a variância de $\hat{\theta}$.

Se $\hat{\theta}$ é não viciado, então EQM($\hat{\theta}$, θ) = Var($\hat{\theta}$). Se $\hat{\theta}$ é viciado, então EQM($\hat{\theta}$, θ) pode ser pensado como uma medida de espalhamento de $\hat{\theta}$ em torno de θ.

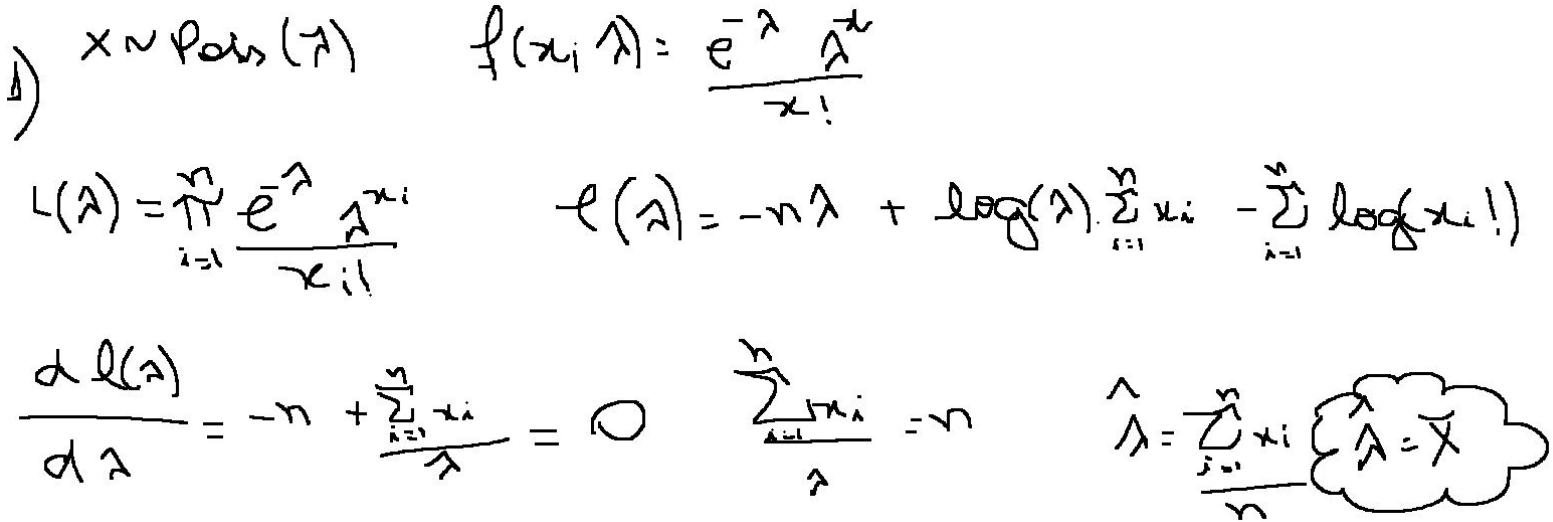


Fonte: https://edisciplinas.usp.br/pluginfile.php/2250126/mod_resource/content/0/Aula20-2016-SLIDES.pdf

Exemplo

Seja $X_1, X_2, \ldots, X_n$ a.a. com n observações de uma população $X \sim Poisson(\lambda)$.

1. Encontre o EMV de λ.
2. Compare o estimador encontrado anteriormente com o estimador T = 1.



2) $EQM(\hat{\lambda};\lambda) = E\left[(\bar{x}-\lambda)^2\right] = Var(\bar{X}) + Vício(\bar{x};\lambda)^2 = \frac{\lambda}{n} - 0 = \frac{\lambda}{n}$

$Vício(\bar{x};\lambda) = E(\bar{X}) - \lambda = \frac{\sum_{i=1}^{n} E(X_i)}{n} - \lambda = \frac{n\lambda}{n} - \lambda = 0$

$Var(\bar{x}) = \frac{n\,Var(X_i)}{n^2} = \frac{\lambda}{n}$

$EQM(1;\lambda) = E\left[(1-\lambda)^2\right] = (1-\lambda)^2$