INFERÊNCIA ESTATÍSTICA I

Ralph dos Santos Silva

http://www.ralphsilva.org/inferenciaestatistica1.html

Instituto de Matemática
Centro de Ciências Matemáticas e da Natureza
Universidade Federal do Rio de Janeiro

Tópicos

- **Função de verossimilhança.**
- **Distribuição a priori e distribuição a posteriori.**
- **Funções perda e estimadores bayesianos.**
- **Estimador de máxima verossimilhança.**
- **Propriedades de estimadores de máxima verossimilhança.**
- **Estatísticas suficientes e teorema de fatoração.**
- **Estatísticas suficientes minimais.**
- **Propriedades frequentistas de estimadores.**
- **Consistência.**
- **Erro quadrático médio.**
- **Estimadores não tendenciosos (não viciados).**
- **Distribuição amostral de estatísticas.**
- **Distribuições derivadas da distribuição normal.**
- **Distribuições qui-quadrado, t-Student e F-Snedecor (Fisher-Snedecor).**
- **Distribuição conjunta de média e variância amostrais.**
- **Intervalos de confiança e de credibilidade.**
- **Intervalos de predição.**

Referências

- *Probability and Statistics*, 4^{th} Edition - DeGroot, M. H. and Schervish, M. J., 2011, Addison-Wesley.
- *Introduction to the Theory of Statistics*, 3^{rd} Edition - Mood, A. M., Graybill, F. A. and Boes, D. C., 1974, McGraw Hill.
- *Statistical Inference: An Integrated Approach*, 2^{nd} Edition - Migon, H. S., Gamerman, D. and Louzada, F., 2014, Chapman & Hall/CRC.
- *Statistical Inference*, 2^{nd} Edition - Casella, G. and Berger, R. L., 2002, Thomson Learning.
- *Introduction to Probability Theory and Statistical Inference*, 3^{rd} Edition, Larson, H. J., 1982, John Wiley.

Inferência Estatística

É um procedimento que produz afirmações probabilísticas sobre um modelo estatístico.

- O termo **inferência estatística** pode ser entendido como o processo de fazer determinadas afirmações sobre quantidades desconhecidas de um modelo estatístico após observar dados que supomos ter informações relevantes.

Exemplo

Uma máquina produz itens defeituosos com uma probabilidade desconhecida de uma distribuição conhecida. Depois de inspecionar n itens, gostaríamos de fazer afirmações sobre esta probabilidade desconhecida.

Exemplo

Partículas radioativas atingem um alvo de acordo com um processo de Poisson com parâmetro λ. Depois de observar n choques em um determinado intervalo de tempo, gostaríamos de fazer afirmações sobre este parâmetro λ.

Definição

População é uma coleção de objetos que possuem uma ou mais características de interesse e consiste na totalidade das observações possíveis de um fenômeno em estudo.

Definição

Amostra é um subconjunto de observações selecionadas de uma população.

Definição

O **espaço paramétrico (parâmetro)** é uma combinação de características que determinam a distribuição conjunta da variável aleatória de interesse.

Notação: Ω

Definição

Uma **Estatística** é qualquer função da amostra. Os elementos de uma amostra são tratados como uma coleção de variáveis aleatórias (antes da amostra ser observada de fato).

Exemplo

Suponha que $(X_1, X_2, \ldots, X_n)$ sejam variáveis aleatórias observáveis (uma amostra aleatória). Seja r uma função real arbitrária das n variáveis aleatórias. Então, a variável aleatória

$$T = r(X_1, X_2, \ldots, X_n),$$

é uma **Estatística**.

Outros exemplos de estatísticas:

- $T = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, a média amostral; e
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$, a variância amostral.

Definição (distribuição a priori)

Suponha um modelo estatístico indexado por um parâmetro θ. Se θ for tratado como uma variável aleatória, então a distribuição de probabilidade de θ antes de observar as realizações das outras variáveis aleatórias de interesse é chamada de **distribuição a priori**. Notação: $f(\theta)$.

Exemplo

Considere o lançamento de uma moeda que é honesta ou tem duas caras. Além disso, considere a probabilidade de se obter cara igual a θ. Agora, suponha que a probabilidade a priori da moeda ser honesta é 0,8. Qual é a função de probabilidade a priori de θ?

$$\Omega = \{1/2,\, 1\}, \quad f(\theta) = \begin{cases} 0,8; & \text{se } \theta = 1/2; \\ 0,2; & \text{se } \theta = 1. \end{cases} \quad \text{e} \quad (X|\theta) \sim \mathcal{BR}(\theta).$$

Exemplo

Uma máquina produz itens defeituosos com uma probabilidade desconhecida θ de uma distribuição conhecida. Podemos, ter a distribuição a priori dada por

$$f(\theta) = \begin{cases} 1; & \text{se } 0 < \theta < 1; \\ 0; & \text{caso contrário.} \end{cases} \quad \text{ou} \quad \theta \sim \mathcal{U}(0;\, 1).$$

Exemplo[1]

Suponha que o tempo de vida, em horas, de um determinado componente eletrônico possa ser modelado através da distribuição exponencial com parâmetro λ. Agora, considere para λ uma distribuição a priori gama com parâmetros a e b tal que $\mathrm{E}(\lambda) = 0,0002$ e $\mathrm{DP}(\lambda) = 0,0001$.

Temos que $\frac{a}{b} = 0,0002$ e $\frac{a}{b^2} = 0,0001^2 \Rightarrow a = 4$ e $b = 20.000$. Então,

$$f(\lambda) = \frac{(20000)^4}{\Gamma(4)} \lambda^3 \exp\{-20000\lambda\} \boldsymbol{I}_{[0,\infty)}(\lambda).$$

[1] Consulte apendice.pdf

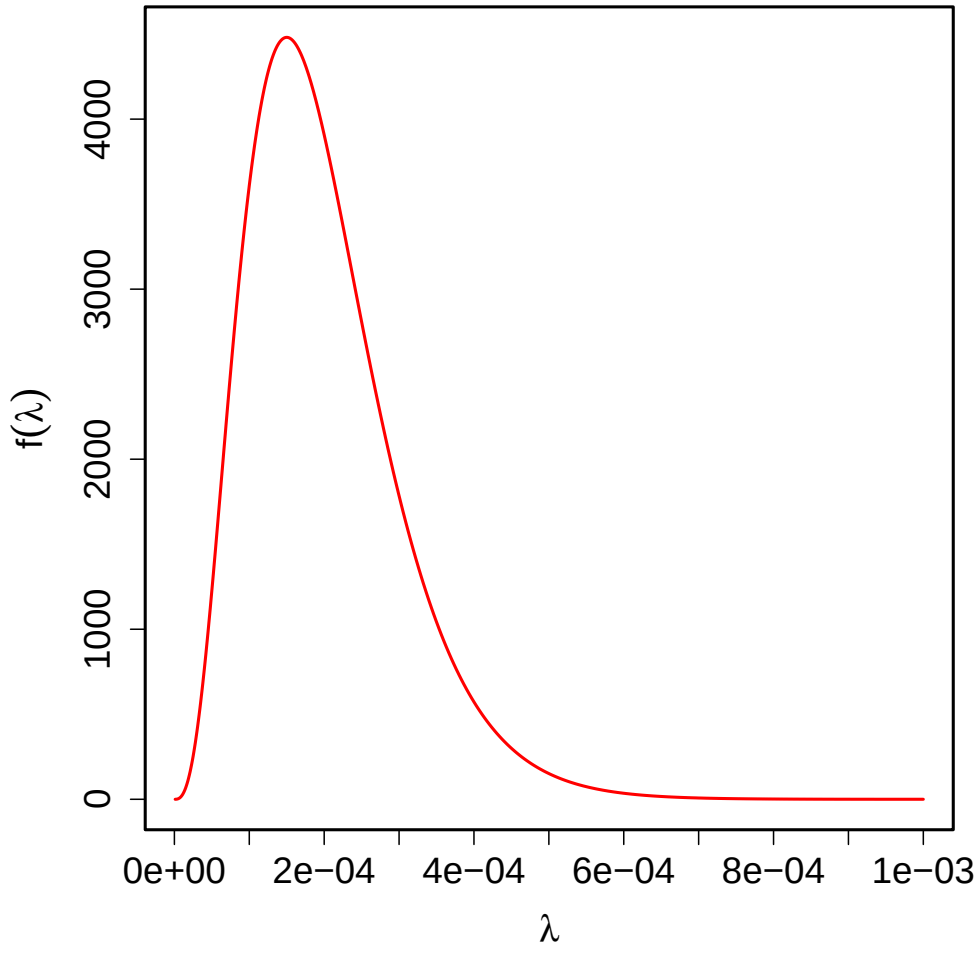


Figura: Distribuição a priori gama com parâmetros $a = 4$ e $b = 20000$.

Ver arquivo exemplo_01.r

Exemplo[2]

Suponha que $(X_1, X_2, \ldots, X_n)$ formam uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$, isto é, as variáveis aleatórias são independentes e identicamente distribuídas tal que $(X_i|\lambda) \sim \mathcal{E}(\lambda)$ para $i = 1, 2, \ldots, n$. Então, a distribuição conjunta é dada por

$$\begin{aligned} f(\boldsymbol{x}|\lambda) &= f(x_1, x_2, \ldots, x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda x_i\} \boldsymbol{I}_{[0,\infty)}(x_i) \\ &= \lambda^n \exp\left\{-\lambda \sum\nolimits_{i=1}^{n} x_i\right\} \left[\prod_{i=1}^{n} \boldsymbol{I}_{[0,\infty)}(x_i)\right] \\ &= \lambda^n \exp\{-\lambda n \overline{x}\} \left[\prod_{i=1}^{n} \boldsymbol{I}_{[0,\infty)}(x_i)\right], \end{aligned}$$

em que $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

[2] Consulte apendice.pdf

Definição (distribuição a posteriori)

Considere um processo estatístico com parâmetro θ. Condicionalmente a θ, sejam $(X_1, X_2, \ldots, X_n)$ variáveis aleatórias a serem observadas. A distribuição condicional de θ dado $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ é chamada de **distribuição a posteriori de** θ. Notação: $f(\theta|\boldsymbol{x})$.

Teorema de Bayes

Suponha que n variáveis aleatórias $(X_1, X_2, \ldots, X_n)$ formam uma amostra aleatória da distribuição $(X|\theta)$ com função (de densidade) de probabilidade $f(x|\theta)$. Suponha também que θ seja desconhecido e que a distribuição a priori de θ seja dada por $f(\theta)$. Então, a função (de densidade) de probabilidade a posteriori de θ é dada por

$$f(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)f(\theta)}{f(\boldsymbol{x})} = \frac{f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)f(\theta)}{f(\boldsymbol{x})} \propto f(\boldsymbol{x}|\theta)f(\theta),$$

em que $\theta \in \Omega$ e $f(\boldsymbol{x})$ é a distribuição marginal de $\boldsymbol{x}$. Além disso, $1/f(\boldsymbol{x})$ é a constante de normalização da posteriori.

Continuando, de forma geral temos que

$$f(\boldsymbol{x}) = \int_{\Omega} f(\boldsymbol{x}|\theta) dF(\theta).$$

Se θ for unidimensional e contínua, podemos escrever

$$f(\boldsymbol{x}) = \int_{\Omega} f(\boldsymbol{x}|\theta) f(\theta) d\theta.$$

Definição (função de verossimilhança)

Quando a função (de densidade) de probabilidade conjunta das observações, $f(\boldsymbol{x}|\theta)$, é considerada como função de θ para particulares valores $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, essa função é chamada de **função de verossimilhança**. Notação: $f(\boldsymbol{x}|\theta)$.

Em muitos livros se utiliza a notação $L(\theta) = f(\boldsymbol{x}|\theta)$. Assim, o teorema de Bayes pode ser reescrito como

$$f(\theta|\boldsymbol{x}) = \frac{L(\theta) f(\theta)}{f(\boldsymbol{x})} \propto L(\theta) f(\theta).$$

Exemplo (tempos de vida; continuação)

Suponha que $(X|\lambda) \sim \mathcal{E}(\lambda)$ e $\lambda \sim \mathcal{G}(a;\ b)$. Qual é a distribuição a posteriori de λ? É possível encontrar a distribuição marginal de $\boldsymbol{x}$? Em caso afirmativo, qual seria?

Vimos que

$$f(\boldsymbol{x}|\lambda) = \lambda^n \exp\{-\lambda n\overline{x}\} \quad \text{e} \quad f(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp\{-b\lambda\}\boldsymbol{I}_{[0,\infty)}(\lambda).$$

Note que a distribuição a posteriori é uma função de λ. Assim, temos que

$$\begin{aligned} f(\lambda|\boldsymbol{x}) &\propto \lambda^n \exp\{-\lambda n\overline{x}\}\,\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp\{-b\lambda\}\boldsymbol{I}_{[0,\infty)}(\lambda) \\ &\propto \lambda^{(n+a)-1}\exp\{-(b+n\overline{x})\lambda\}\boldsymbol{I}_{[0,\infty)}(\lambda) \end{aligned}$$

Então, reconhecendo a forma deste núcleo, temos que

$$f(\lambda|\boldsymbol{x}) = \frac{(b+n\overline{x})^{(n+a)}}{\Gamma(n+a)}\lambda^{(n+a)-1}\exp\{-(b+n\overline{x})\lambda\}\boldsymbol{I}_{[0,\infty)}(\lambda),$$

ou seja, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n+a;\ b+n\overline{x})$.

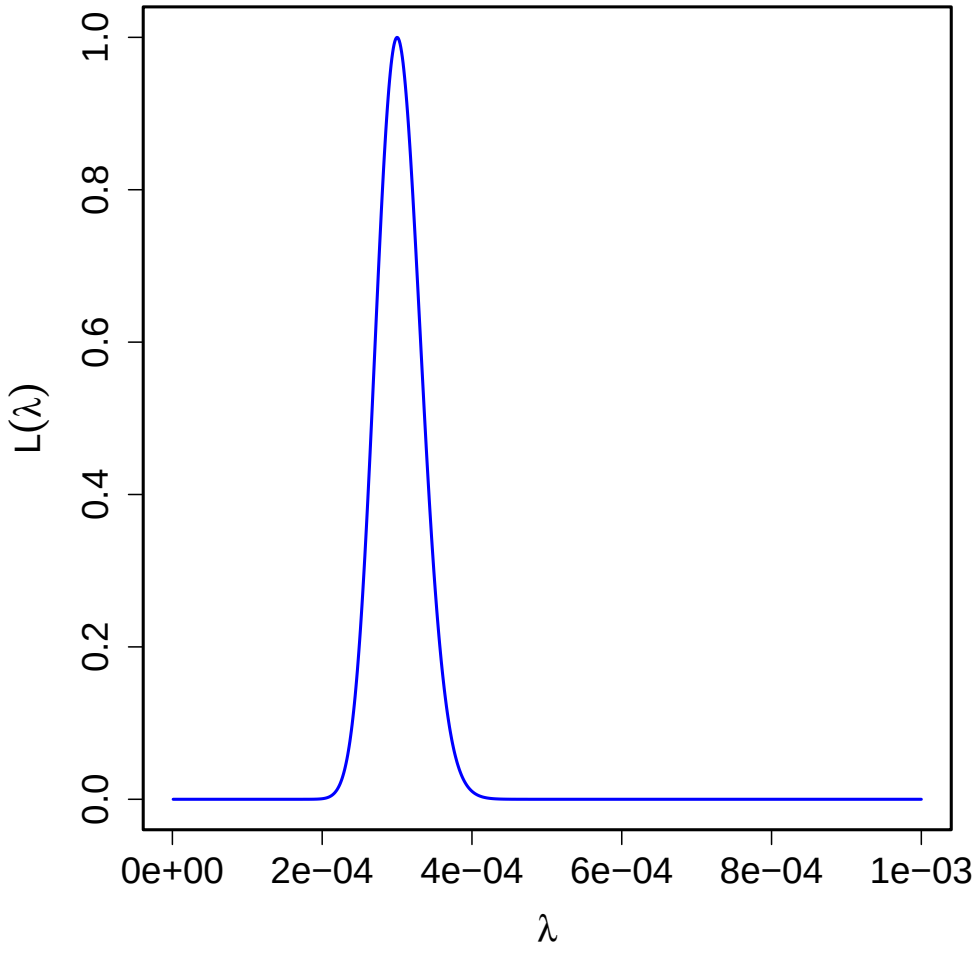


Figura: Exemplo de função de verossimilhança do modelo exponencial.

Ver arquivo exemplo_02.r

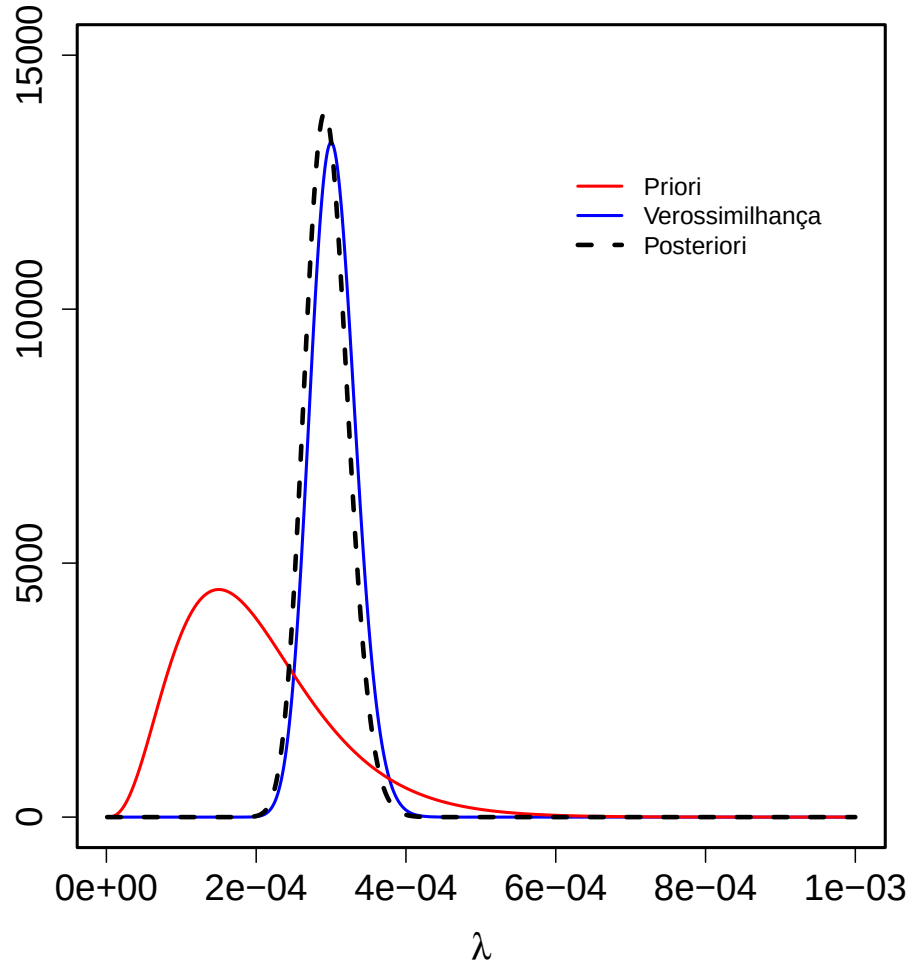


Figura: Exemplo de distribuição a priori, de função de verossimilhança e de distribuição a posteriori do modelo exponencial.

A distribuição marginal de $\boldsymbol{x}$ é dada por

$$
\begin{aligned}
f(\boldsymbol{x}) &= \int_0^\infty f(\boldsymbol{x}|\lambda) f(\lambda) d\lambda \\
&= \int_0^\infty \lambda^n \exp\{-\lambda n\overline{x}\} \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp\{-b\lambda\} d\lambda \\
&= \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{(n+a)-1} \exp\{-(b+n\overline{x})\lambda\} d\lambda \\
&= \frac{b^a}{\Gamma(a)} \frac{\Gamma(n+a)}{(b+n\overline{x})^{(n+a)}} \int_0^\infty \frac{(b+n\overline{x})^{(n+a)}}{\Gamma(n+a)} \lambda^{(n+a)-1} \exp\{-(b+n\overline{x})\lambda\} d\lambda \\
&= \frac{b^a}{\Gamma(a)} \frac{\Gamma(n+a)}{(b+n\overline{x})^{(n+a)}} \\
&= \left(\frac{b}{b+n\overline{x}}\right)^a \left(\frac{1}{b+n\overline{x}}\right)^n \frac{\Gamma(n+a)}{\Gamma(a)}.
\end{aligned}
$$

Note que poderíamos encontrar $f(\boldsymbol{x})$ e, depois, fazer os cálculos para determinar a posteriori de λ.

Observações sequenciais e previsão

Em muitos experimentos, as observações $(X_1, X_2, \ldots, X_n)$ que formam a amostra aleatória são obtidas de forma sequencial.

Suponha que se observou $X_1 = x_1$. Então, pelo teorema de Bayes, temos que

$$f(\theta|x_1) \propto f(x_1|\theta)f(\theta).$$

Agora, observou-se $X_2 = x_2$. Podemos utilizar a posteriori anterior, $f(\theta|x_1)$, como a priori de θ tal que

$$\begin{aligned} f(\theta|x_1, x_2) &\propto f(x_2|\theta)f(\theta|x_1) \\ &\propto f(x_2|\theta)f(x_1|\theta)f(\theta) \end{aligned}$$

Assim, temos que

$$\begin{aligned} f(\theta|x_1, x_2, \ldots, x_n) &\propto f(x_n|\theta)f(\theta|x_1, x_2, \ldots, x_{n-1}) \\ &\propto f(x_n|\theta)f(x_{n-1}|\theta) \cdots f(x_2|\theta)f(x_1|\theta)f(\theta) \\ f(\theta|\boldsymbol{x}) &\propto \left[\prod_{i=1}^{n} f(x_i|\theta)\right] f(\theta). \end{aligned}$$

Segue-se que

$$f(\theta|\boldsymbol{x}) = f(\theta|x_1, x_2, \ldots, x_n) = c \times f(x_n|\theta) f(\theta|x_1, x_2, \ldots, x_{n-1}),$$

com

$$c^{-1} = \int_{\Omega} f(x_n|\theta) f(\theta|x_1, x_2, \ldots, x_{n-1}) d\theta = f(x_n|x_1, x_2, \ldots, x_{n-1}).$$

A distribuição $f(x_n|x_1, x_2, \ldots, x_{n-1})$ pode ser utilizada para fazer previsão.

Após observar as $n-1$ primeiras observações, gostaríamos de prever a n-ésima observação.

Exemplo (tempos de vida; continuação)

Deseja-se fazer a previsão de X_6 após ter observado $\boldsymbol{x}_5 = (x_1, x_2, \ldots, x_5) = (2911;\ 3403;\ 3237;\ 3509;\ 3118)$ em horas.
Sabemos que

$$f(\lambda|\boldsymbol{x}_{n-1}) = \frac{(b+(n-1)\overline{x}_{n-1})^{(n-1+a)}}{\Gamma(n-1+a)}\lambda^{(n-1+a)-1}\exp\{-(b+(n-1)\overline{x}_{n-1})\lambda\}.$$

Daí, temos que

$$\begin{aligned}
f(x_n|\boldsymbol{x}_{n-1}) &= \int_0^\infty \lambda\exp\{-\lambda x_n\}\frac{(b+(n-1)\overline{x}_{n-1})^{(n-1+a)}}{\Gamma(n-1+a)} \\
&\times \lambda^{(n-1+a)-1}\exp\{-(b+(n-1)\overline{x}_{n-1})\lambda\}d\lambda \\
&= \frac{(b+(n-1)\overline{x}_{n-1})^{(n-1+a)}}{\Gamma(n-1+a)} \times \frac{\Gamma(n+a)}{(b+n\overline{x}_n)^{(n+a)}} \\
&\times \int_0^\infty \frac{(b+n\overline{x}_n)^{(n+a)}}{\Gamma(n+a)}\lambda^{(n+a)-1}\exp\{-(b+n\overline{x}_n)\lambda\}d\lambda \\
&= \frac{(b+(n-1)\overline{x}_{n-1})^{(n-1+a)}}{\Gamma(n-1+a)} \times \frac{\Gamma(n+a)}{(b+n\overline{x}_n)^{(n+a)}}.
\end{aligned}$$

Lembre-se que $a = 4$ e $b = 20.000$.

Temos que $\boldsymbol{x}_5 = (x_1, x_2, \ldots, x_5) = (2911;\ 3403;\ 3237;\ 3509;\ 3118)$. Logo, $\sum_{i=1}^{5} x_i = 16.178$.

Assim,

$$\begin{aligned} f(x_6|\boldsymbol{x}_5) &= \frac{(20.000 + 16.178)^{(5+4)}}{\Gamma(5+4)} \times \frac{\Gamma(6+4)}{(20.000 + 16.178 + x_6)^{(6+4)}} \\ &= \frac{\Gamma(10)}{\Gamma(9)} \times \frac{36.178^9}{(36.178 + x_6)^{10}} \\ &= \frac{9 \times 36.178^9}{(36.178 + x_6)^{10}}, \quad \text{para} \quad x_6 > 0. \end{aligned}$$

Isto permite calcular probabilidades como

$$\Pr(X_6 > 3000|\boldsymbol{x}_5) = \int_{3.000}^{\infty} \frac{9 \times 36.178^9}{(36.178 + x_6)^{10}}\, dx_6 = \frac{9 \times 36.178^9}{9 \times 39.178^9} = 0,4882.$$

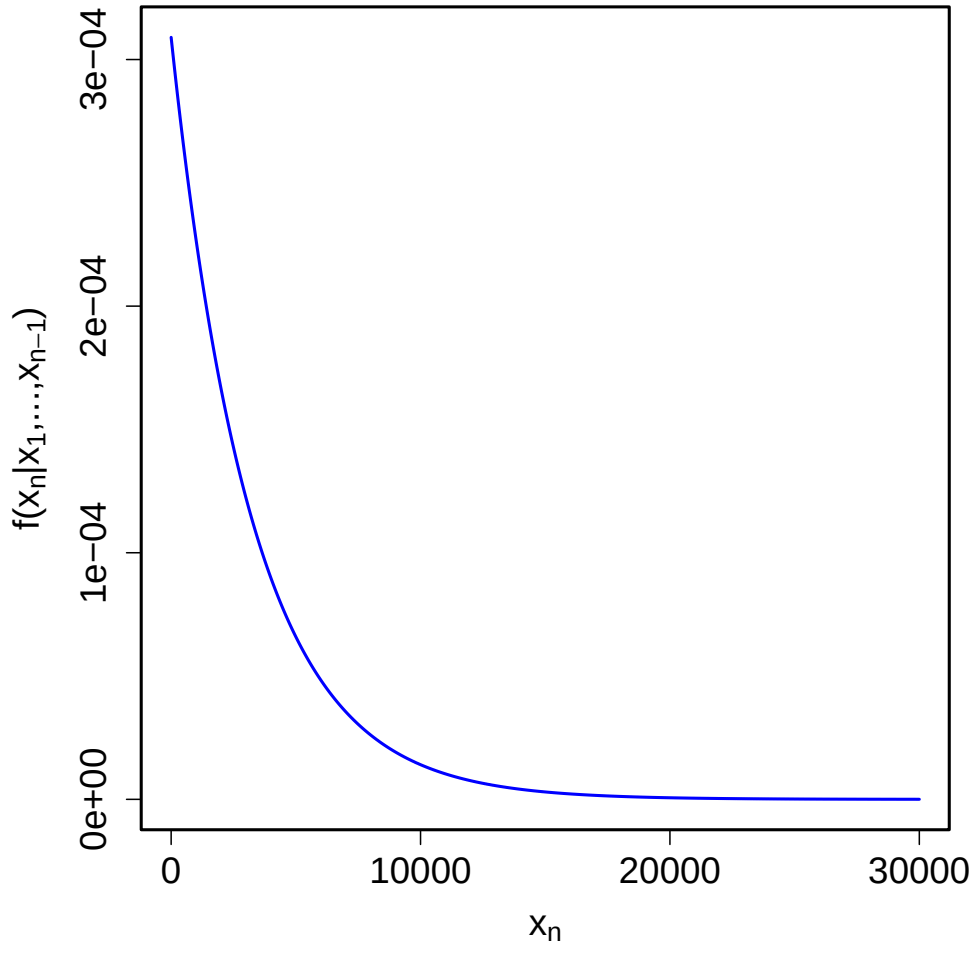


Figura: Exemplo de distribuição preditiva resultante do modelo exponencial.

Ver arquivo exemplo_03.r

Distribuição a priori conjugada

Definição

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória da distribuição de $(X|\theta)$ com função (de densidade) de probabilidade conjunta $f(\boldsymbol{x}|\theta)$. Seja Ψ uma família de possíveis distribuições sob o espaço paramétrico Ω. Suponha que, independente da escolha da distribuição a priori, escolhemos $f(\theta)$ da família Ψ. Se independente do tamanho amostral n e dos valores observados $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ a distribuição a posteriori $f(\theta|\boldsymbol{x})$ pertencer a família Ψ, então Ψ é chamada de família de distribuição a **priori conjugada** para amostras de $f(\boldsymbol{x}|\theta)$. Diz-se também que Ψ é fechada sob amostragem das distribuições $f(\boldsymbol{x}|\theta)$.

Exemplo

- Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$.
- Assim, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\pi) = \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})}$$

- A distribuição a posteriori é dada por

$$f(\pi|\boldsymbol{x}) \propto \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})} f(\pi).$$

- Podemos considerar a distribuição a priori $\mathcal{BT}(a; b)$, isto é,

$$f(\pi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1} \boldsymbol{I}_{[0,1]}(\pi).$$

- Portanto, a distribuição a posteriori é dada por

$$(\pi|\boldsymbol{x}) \sim \mathcal{BT}(a + n\overline{x}; b + n(1-\overline{x})).$$

- Assim, $\mathcal{BT}(a; b)$ é conjugada para amostras de $f(\boldsymbol{x}|\pi)$ (Bernoulli).

Exemplo

- Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$.
- Assim, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\lambda) = \prod_{i=1}^{n} \left[\frac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!} \right] = \frac{\exp\{-n\lambda\}\lambda^{n\overline{x}}}{\prod_{i=1}^{n} \Gamma(x_i + 1)}$$

- A distribuição a posteriori é dada por

$$f(\lambda|\boldsymbol{x}) \propto \lambda^{n\overline{x}} \exp\{-n\lambda\} f(\lambda).$$

- Podemos considerar a distribuição a priori $\mathcal{G}(a; b)$, isto é,

$$f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp\{-b\lambda\} \boldsymbol{I}_{[0,\infty)}(\lambda).$$

- Portanto, a distribuição a posteriori é dada por

$$(\lambda|\boldsymbol{x}) \sim \mathcal{G}(a + n\overline{x}; b + n).$$

- Assim, $\mathcal{G}(a; b)$ é conjugada para amostras de $f(\boldsymbol{x}|\lambda)$ (Poisson).

Exemplo

- Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$.
- Assim, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\lambda) = \lambda^n \exp\{-\lambda n\overline{x}\}.$$

- A distribuição a posteriori é dada por

$$f(\lambda|\boldsymbol{x}) \propto \lambda^n \exp\{-\lambda n\overline{x}\} f(\lambda).$$

- Podemos considerar a distribuição a priori $\mathcal{G}(a; b)$, isto é,

$$f(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp\{-b\lambda\} \boldsymbol{I}_{[0,\infty)}(\lambda).$$

- Portanto, a distribuição a posteriori é dada por

$$(\lambda|\boldsymbol{x}) \sim \mathcal{G}(a + n; b + n\overline{x}).$$

- Assim, $\mathcal{G}(a; b)$ é conjugada para amostras de $f(\boldsymbol{x}|\lambda)$ (Exponencial).

Exemplo

- Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu; \sigma_0^2)$ com σ_0^2 conhecido.
- Assim, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\mu) = (2\pi)^{-n/2}(\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

Temos que

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 - 2(\mu - \overline{x})\sum_{i=1}^{n}(x_i - \overline{x}) + n(\mu - \overline{x})^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\mu - \overline{x})^2,
\end{aligned}$$

pois

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

- Agora, podemos escrever a função de verossimilhança como

$$f(\boldsymbol{x}|\mu) \propto \exp\left\{-\frac{n}{2\sigma_0^2}(\mu - \overline{x})^2\right\}.$$

- A distribuição a posteriori é dada por

$$f(\mu|\boldsymbol{x}) \propto \exp\left\{-\frac{n}{2\sigma_0^2}(\mu - \overline{x})^2\right\} f(\mu).$$

- Mostraremos que a distribuição a priori conjugada é $\mathcal{N}(a; b^2)$.
- Então, considere

$$f(\mu) \propto \exp\left\{-\frac{1}{2b^2}(\mu - a)^2\right\}.$$

- A distribuição a posteriori é dada por

$$\begin{aligned} f(\mu|\boldsymbol{x}) &\propto \exp\left\{-\frac{n}{2\sigma_0^2}(\mu-\overline{x})^2\right\} f(\mu) \\ &\propto \exp\left\{-\frac{n}{2\sigma_0^2}(\mu-\overline{x})^2 - \frac{1}{2b^2}(\mu-a)^2\right\}. \end{aligned}$$

Agora, temos que

$$\begin{aligned} \frac{1}{(\sigma_0^2/n)}(\mu-\overline{x})^2 &= \frac{\mu^2}{\sigma_0^2/n} - 2\frac{\mu\overline{x}}{\sigma_0^2/n} + \frac{\overline{x}^2}{\sigma_0^2/n} \\ \frac{1}{b^2}(\mu-a)^2 &= \frac{\mu^2}{b^2} - 2\frac{\mu a}{b^2} + \frac{a^2}{b^2}. \end{aligned}$$

Somando-se ambas as equações e dispensando as partes que não dependem de μ, temos

$$\mu^2\left[\frac{n}{\sigma_0^2} + \frac{1}{b^2}\right] - 2\mu\left[\frac{n\overline{x}}{\sigma_0^2} + \frac{a}{b^2}\right].$$

Agora, definimos

$$\tau^2 = [n\sigma_0^{-2} + b^{-2}]^{-1} \quad \text{e} \quad \xi = \tau^2 \left[\frac{n\overline{x}}{\sigma_0^2} + \frac{a}{b^2} \right].$$

Assim, temos que

$$\mu^2 \left[\frac{n}{\sigma_0^2} + \frac{1}{b^2} \right] - 2\mu \left[\frac{n\overline{x}}{\sigma_0^2} + \frac{a}{b^2} \right] = \frac{\mu^2}{\tau^2} - \frac{2\mu\xi}{\tau^2}.$$

Estamos buscando o núcleo de uma normal. Portanto, completamos quadrados para obter

$$\frac{\mu^2}{\tau^2} - \frac{2\mu\xi}{\tau^2} + \frac{\xi^2}{\tau^2} = \frac{1}{\tau^2}(\mu - \xi)^2.$$

Note que o termo introduzido não depende de μ.

- A distribuição a posteriori é dada por

$$f(\mu|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2\tau^2}(\mu-\xi)^2\right\}.$$

- Resumindo, temos a distribuição a posteriori dada por

$$(\mu|\boldsymbol{x}) \sim \mathcal{N}(\xi;\tau^2),$$

com

$$\tau^2 = [n\sigma_0^{-2} + b^{-2}]^{-1} \quad \text{e} \quad \xi = \tau^2\left[\frac{n\bar{x}}{\sigma_0^2} + \frac{a}{b^2}\right].$$

- Para um modelo normal com média μ e variância conhecida, a distribuição a priori normal para μ é uma priori conjugada, isto é, $\mu \sim \mathcal{N}(a; b^2)$ é uma priori conjugada.

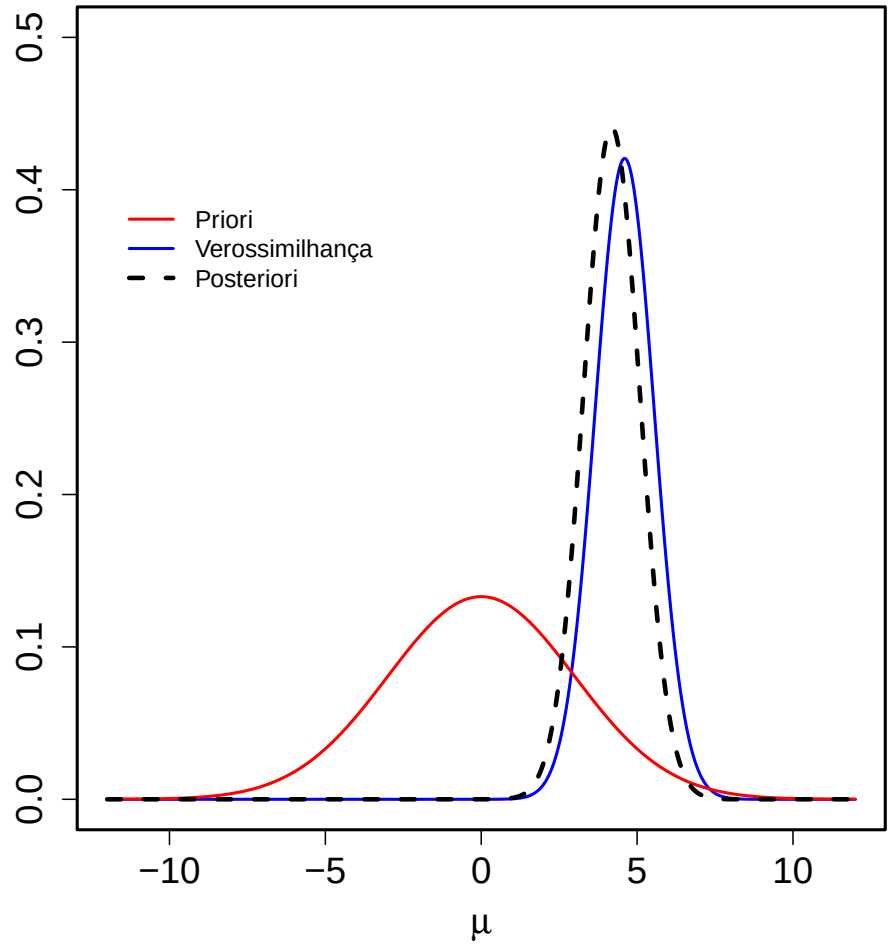


Figura: Exemplo de distribuição a priori, de função de verossimilhança e de distribuição a posteriori do modelo normal.

Ver arquivo exemplo_04.r

Distribuição a priori imprópria

Definição

Seja f uma função não negativa cujo domínio inclui o espaço paramétrico de um modelo estatístico. Suponha que

$$\int_{\Omega} f(\theta) d\theta = \infty.$$

Se utilizarmos $f(\theta)$ como distribuição a priori para θ, então $f(\theta)$ **é uma distribuição a priori imprópria para** θ.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$. Então, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\lambda) = \lambda^n \exp\{-\lambda n\overline{x}\}.$$

Agora, considere

$$f(\lambda) \propto \frac{1}{\lambda} \boldsymbol{I}_{[0,\infty)}(\lambda), \quad \text{em que} \quad \int_0^\infty f(\lambda) d\theta = \int_0^\infty \frac{\mathbf{c}}{\lambda} d\lambda = \Big[\mathbf{c} \ln(\lambda)\Big]_0^\infty = \infty.$$

Claramente, $f(\lambda)$ **é uma distribuição a priori imprópria para** λ.

A função de densidade de probabilidade a posteriori é dada por

$$f(\lambda|\boldsymbol{x}) \propto \lambda^{n-1} \exp\{-\lambda n\overline{x}\},$$

ou seja, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n;\, n\overline{x})$ que é uma **distribuição própria**. Note que

$$f(\boldsymbol{x}) = \int_0^\infty \mathbf{c}\lambda^{n-1} \exp\{-\lambda n\overline{x}\}\, d\lambda = \mathbf{c} \times \frac{\Gamma(n)}{[n\overline{x}]^n}, \quad \text{com} \quad \mathbf{c} > 0.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$. Então, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\pi) = \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})}.$$

Agora, considere a seguinte **distribuição a priori imprópria para** π:

$$f(\pi) \propto \frac{1}{\pi(1-\pi)} \boldsymbol{I}_{[0,1]}(\pi), \quad \text{em que}$$

$$\int_0^1 \mathbf{c} \frac{1}{\pi(1-\pi)} d\pi = \int_0^1 \mathbf{c} \left[\frac{1}{\pi} + \frac{1}{1-\pi}\right] d\pi = \Big[\mathbf{c}\{\ln(\pi) - \ln(1-\pi)\}\Big]_0^1 = \infty.$$

A função de densidade de probabilidade a posteriori é dada por

$$f(\pi|\boldsymbol{x}) \propto \pi^{n\overline{x}-1}(1-\pi)^{n(1-\overline{x})-1}, \quad \text{i.e.,} \quad (\pi|\boldsymbol{x}) \sim \mathcal{BT}(n\overline{x}; n(1-\overline{x})) \text{ (é própria)}.$$

Note que

$$f(\boldsymbol{x}) = \int_0^1 \mathbf{c}\pi^{n\overline{x}-1}(1-\pi)^{n(1-\overline{x})-1} d\pi = \mathbf{c} \times \frac{\Gamma(n\overline{x})\Gamma(n(1-\overline{x}))}{\Gamma(n)}, \quad \text{com} \quad \mathbf{c} > 0.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu; \sigma_0^2)$ com σ_0^2 conhecido. Então, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\mu) \propto \exp\left\{-\frac{n}{2\sigma_0^2}(\mu - \overline{x})^2\right\}.$$

Agora, considere a seguinte **distribuição a priori imprópria para** μ:

$$f(\mu) \propto 1.$$

A função de densidade de probabilidade a posteriori é dada por

$$f(\mu|\boldsymbol{x}) \propto \exp\left\{-\frac{n}{2\sigma_0^2}(\mu - \overline{x})^2\right\}, \text{ ou seja, } (\mu|\boldsymbol{x}) \sim \mathcal{N}\left(\overline{x}; \frac{\sigma_0^2}{n}\right) \text{ (é própria).}$$

Estimadores de Bayes

Definição

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra da função (de densidade) de probabilidade $f(x|\theta)$ com o parâmetro θ desconhecido tal que $\theta \in \Omega \subseteq \mathbb{R}$. **Um estimador do parâmetro θ é uma função real $\delta(X_1, X_2, \ldots, X_n)$.** Além disso, **$\delta(x_1, x_2, \ldots, x_n)$ é chamada de estimativa de θ.**

- Todo estimador é uma Estatística.
- Como $\theta \in \Omega$, então $\delta(X_1, X_2, \ldots, X_n) \in \Omega$.
- Lembre-se que um estimador é uma variável aleatória.
- Lembre-se que uma estimativa é uma particular realização da variável aleatória.

Definição

Uma **função perda** é uma função real de duas variáveis $L(\theta, a)$ em que $\theta \in \Omega$ e a é um número real. Seja $f(\theta)$ uma distribuição a priori para $\theta \in \Omega$. Então, dado um determinado valor a, temos que a perda esperada (a priori) é dada por

$$\mathrm{E}(L(\theta, a)) = \int_{\Omega} L(\theta, a) f(\theta) d\theta.$$

Após observar os dados $(x_1, x_2, \ldots, x_n)$, temos que a perda esperada (a posteriori) é dada por

$$\mathrm{E}(L(\theta, a)|\boldsymbol{x}) = \int_{\Omega} L(\theta, a) f(\theta|\boldsymbol{x}) d\theta,$$

para cada escolha de a possível.

Seja $a = \delta^\star(x_1, x_2, \ldots, x_n)$ uma estimativa cuja perda esperada a posteriori é mínima. Então, a função $\delta(X_1, X_2, \ldots, X_n)$ cujos valores são especificados dessa forma é um estimador de θ.

Definição

Seja $L(\theta, a)$ uma função perda para cada possível valor $\boldsymbol{x}$ de $\boldsymbol{X}$. Seja $\delta^\star(\boldsymbol{x})$ um valor de a tal que $\mathrm{E}(L(\theta, a)|\boldsymbol{x})$ é minimizado. Então, **$\delta^\star(\boldsymbol{X})$ é chamado de estimador de Bayes para θ** e **$\delta^\star(\boldsymbol{x})$ é chamado de estimativa de Bayes para θ**.

Definição

A função perda $L(\theta, a) = (\theta - a)^2$ é chamada de função **perda quadrática**.

Corolário

Seja θ um parâmetro com valores nos reais. Suponha que a função perda quadrática seja utilizada e que a média a posteriori de θ, $\mathrm{E}(\theta|\boldsymbol{x})$, seja finita. Então, o estimador de Bayes de θ é $\delta^\star(\boldsymbol{X}) = \mathrm{E}(\theta|\boldsymbol{X})$.

Prova

Suponha que $\mathrm{E}(L(\theta, a)|\boldsymbol{x}) = \mathrm{E}([\theta - a]^2|\boldsymbol{x})$ exista e seja finito. Queremos $\min_a \mathrm{E}(L(\theta, a)|\boldsymbol{x})$. Então, supondo θ contínuo, temos que

$$\begin{aligned}
\mathrm{E}(L(\theta, a)|\boldsymbol{x}) &= \int_\Omega (\theta - a)^2 f(\theta|\boldsymbol{x}) d\theta \\
\frac{\partial \mathrm{E}(L(\theta, a)|\boldsymbol{x})}{\partial a} &= \frac{\partial}{\partial a} \int_\Omega (\theta - a)^2 f(\theta|\boldsymbol{x}) d\theta = \int_\Omega \frac{\partial}{\partial a} (\theta - a)^2 f(\theta|\boldsymbol{x}) d\theta \\
&= -2 \int_\Omega (\theta - a) f(\theta|\boldsymbol{x}) d\theta = -2[\mathrm{E}(\theta|\boldsymbol{x}) - a] = 0
\end{aligned}$$

Logo, $a = \mathrm{E}(\theta|\boldsymbol{x})$ é ponto crítico de $\mathrm{E}(L(\theta, a)|\boldsymbol{x})$ como função de a.

Note que

$$\frac{\partial^2 \mathrm{E}(L(\theta, a)|\boldsymbol{x})}{\partial a^2} = 2.$$

Portanto, $a = \mathrm{E}(\theta|\boldsymbol{x})$ é ponto de mínimo.

Existe também uma prova para o caso θ discreto.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$ e $\lambda \sim \mathcal{G}(a; b)$. Então,

$$(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n + a;\ b + n\overline{x}).$$

Assim, o estimador de Bayes utilizando a perda quadrática é dado por

$$\mathrm{E}(\lambda|\boldsymbol{X}) = \frac{n + a}{b + n\overline{X}}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$ e $\pi \sim \mathcal{BT}(a; b)$. Então,

$$(\pi|\boldsymbol{x}) \sim \mathcal{BT}(a + n\overline{x}; b + n(1 - \overline{x})).$$

Assim, o estimador de Bayes utilizando a perda quadrática é dado por

$$\mathrm{E}(\pi|\boldsymbol{X}) = \frac{a + n\overline{X}}{a + b + n}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$ e $f(\lambda) \propto 1/\lambda$. Então,

$$(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n;\ n\overline{x}).$$

Assim, o estimador de Bayes utilizando a perda quadrática é dado por

$$\mathrm{E}(\lambda|\boldsymbol{X}) = \frac{1}{\overline{X}}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$ e $f(\pi) \propto \pi^{-1}(1-\pi)^{-1}$. Então,

$$(\pi|\boldsymbol{x}) \sim \mathcal{BT}(n\overline{x}; n(1-\overline{x})).$$

Assim, o estimador de Bayes utilizando a perda quadrática é dado por

$$\mathrm{E}(\pi|\boldsymbol{X}) = \overline{X}.$$

Definição

A função perda $L(\theta, a) = |\theta - a|$ é chamada de função **perda absoluta**.

Para todo valor observado $\boldsymbol{x}$, a estimativa de Bayes $\delta^\star(\boldsymbol{x})$ será o valor de a que minimiza $\mathrm{E}(|\theta - a||\boldsymbol{x})$.

Corolário

Seja θ um parâmetro com valores nos reais. Suponha que a função perda absoluta seja utilizada. Então, o estimador de Bayes de θ é $\delta^\star(\boldsymbol{X}) = \mathrm{m}(\theta|\boldsymbol{X})$ (mediana da distribuição a posteriori de θ).

Definição

A função perda $L(\theta, a) = 1 - \boldsymbol{I}_{[0,\epsilon)}(|\theta - a|)$ é chamada de função **perda 0-1**.

Para todo valor observado $\boldsymbol{x}$, a estimativa de Bayes $\delta^{\star}(\boldsymbol{x})$ será o valor de $a = \max_{\theta} f(\theta|\boldsymbol{x})$, tomando-se $\epsilon \to 0$.

Corolário

Seja θ um parâmetro com valores nos reais. Suponha que a função perda 0-1 seja utilizada.Então, o estimador de Bayes de θ é $\delta^{\star}(\boldsymbol{X}) = \max_{\theta}(f(\theta|\boldsymbol{X}))$ (moda da distribuição a posteriori de θ).

Outras funções perda

- Perda quadrática: $L(\theta, a) = (\theta - a)^2$ **(média a posteriori)**.
- Perda absoluta: $L(\theta, a) = |\theta - a|$ **(mediana a posteriori)**.
- Perda 0-1: $L(\theta, a) = 1 - \boldsymbol{I}_{[0,\epsilon)}(|\theta - a|)$ **(moda a posteriori)**.
- Perda de mistura:

$$L(\theta, a) = \begin{cases} (a-\theta)^2, & \text{se } |a-\theta| < k; \\ 2k|a-\theta| - k^2, & \text{caso contrário.} \end{cases}$$

- Perda multilinear:

$$L_{k_1,k_2}(\theta, a) = \begin{cases} k_2(\theta - a), & \text{se } \theta > a; \\ k_1(a-\theta), & \text{caso contrário.} \end{cases}$$

Corolário

Seja θ um parâmetro com valores nos reais. Suponha que a função **perda multilinear** seja utilizada. Então, o estimador de Bayes de θ é $\delta^\star(\boldsymbol{X}) = \mathrm{Q}_\kappa(\theta|\boldsymbol{X})$, com $\kappa = k_2/(k_1 + k_2)$ e Q_κ representando **κ-ésimo percentil da distribuição a posteriori de θ**.

Prova

Temos que

$$\begin{aligned} \mathrm{E}(L_{k_1,k_2}(\theta,a)|\boldsymbol{x}) &= k_1\int_{-\infty}^{a}(a-\theta)f(\theta|\boldsymbol{x})d\theta + k_2\int_{a}^{+\infty}(\theta-a)f(\theta|\boldsymbol{x})d\theta \\ &= k_1\int_{-\infty}^{a}\Pr(\theta<s|\boldsymbol{x})ds - k_2\int_{a}^{+\infty}\Pr(\theta>s|\boldsymbol{x})ds, \end{aligned}$$

com a última igualdade obtida por integração por partes. Derivando-se em relação a a, obtemos

$$k_1\Pr(\theta<a|\boldsymbol{x}) - k_2\Pr(\theta>a|\boldsymbol{x}) = 0, \quad \text{isto é,} \quad \Pr(\theta<a|\boldsymbol{x}) = \frac{k_2}{k_1+k_2}.$$

Lembrando a fórmula de integração por partes:

$$\int u\ dv = u\times v - \int v\ du$$

Estimativas de Bayes para grandes amostras

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$ e $\lambda \sim \mathcal{G}(a; b)$. Então, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n + a;\ b + n\overline{x})$ e o estimador de Bayes utilizando a perda quadrática é dado por

$$\mathrm{E}(\lambda|\boldsymbol{X}) = \frac{n + a}{b + n\overline{X}}.$$

Agora, suponha que $n = 100$ e $\overline{x} = 0,48$. Então, para

- $a = b = 1$, temos que $\mathrm{E}(\lambda|\boldsymbol{x}) = \dfrac{100 + 1}{48 + 1} = 2,061$;
- $a = 2$ e $b = 1$, temos que $\mathrm{E}(\lambda|\boldsymbol{x}) = \dfrac{100 + 1}{48 + 2} = 2,02$; e
- $a = b = 0,01$, temos que $\mathrm{E}(\lambda|\boldsymbol{x}) = \dfrac{100 + 0,01}{48 + 0,01} = 2,083$.

Além disso, temos que

$$\lim_{n\to\infty} \mathrm{E}(\lambda|\boldsymbol{x}) = \lim_{n\to\infty} \frac{n + a}{b + n\overline{x}} = \frac{1}{\overline{x}}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$ e $f(\lambda) \propto 1/\lambda$. Então, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n;\, n\overline{x})$ e o estimador de Bayes utilizando a perda quadrática é dado por

$$\mathrm{E}(\lambda|\boldsymbol{X}) = \frac{1}{\overline{X}}.$$

Agora, suponha que $n = 100$ e $\overline{x} = 0,48$. Então,

$$\mathrm{E}(\lambda|\boldsymbol{x}) = \frac{100}{48} = 2,08333.$$

Consistência do estimador de Bayes

Definição

Uma sequência de estimadores que converge em probabilidade para o desconhecido valor do parâmetro a ser estimado, quando $n \to \infty$, é chamado de uma **sequência consistente de estimadores**.

Exemplo

Suponha $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n+a;\ b+n\overline{x})$ tal que

$$\delta(\boldsymbol{X}_n) = \mathrm{E}(\lambda|\boldsymbol{X}) = \frac{n+a}{b+n\overline{X}} \quad \text{e} \quad \lim_{n\to\infty} \mathrm{E}(\lambda|\boldsymbol{X}) = \lim_{n\to\infty} \frac{n+a}{b+n\overline{X}} = \frac{1}{\overline{X}},$$

com $\boldsymbol{X}_n = \boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. Contudo, pela lei dos grandes números, temos que

$$\lim_{n\to\infty} \Pr\left(\left|\overline{X} - \frac{1}{\lambda}\right| < \epsilon\right) = 1,$$

e, pelo teorema de Slutsky,

$$\lim_{n\to\infty} \Pr(|\delta(\boldsymbol{X}_n) - \lambda| < \epsilon) = 1, \quad \text{isto é,} \quad \delta(\boldsymbol{X}_n) \xrightarrow[n\to\infty]{} \lambda, \quad \text{em probabilidade.}$$

Definição

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição indexada por um vetor paramétrico $\boldsymbol{\theta}$, tal que $\boldsymbol{\theta} \subseteq \Omega \subseteq \mathbb{R}^d$. Seja h uma função de Ω no espaço d dimensional. Defina $\Psi = h(\boldsymbol{\theta})$. Um **estimador** de Ψ é uma função $\delta(X_1, X_2, \ldots, X_n)$ que assume valores no espaço d dimensional. Chamamos $\delta(x_1, x_2, \ldots, x_n)$ de **estimativa** de Ψ em que $(x_1, x_2, \ldots, x_n)$ são os valores observados de $(X_1, X_2, \ldots, X_n)$.

Exemplo

Suponha $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n+a;\ b+n\overline{x})$ com $n=3$, $a=1$, $b=2$ e $\sum_{i=1}^{3} x_i = 6,6$. Então,

$$\delta_\lambda = \mathrm{E}(\lambda|\boldsymbol{x}) = \frac{3+1}{2+6,6} = \frac{4}{8,6} = 0,4651.$$

Agora, considere o interesse em estimar $\psi = 1/\lambda$ sob a perda quadrática. Assim,

$$\begin{aligned}
\delta_\psi(\boldsymbol{x}) &= \mathrm{E}(\psi|\boldsymbol{x}) = \mathrm{E}(\lambda^{-1}|\boldsymbol{x}) = \int_0^\infty \lambda^{-1} f(\lambda|\boldsymbol{x}) d\lambda \\
&= \frac{[b+n\overline{x}]^{n+a}}{\Gamma(n+a)} \times \frac{\Gamma(n+a-1)}{[b+n\overline{x}]^{n+a-1}} \\
&\times \int_0^\infty \frac{[b+n\overline{x}]^{n+a-1}}{\Gamma(n+a-1)} \lambda^{(n+a-1)-1} \exp\{-(b+n\overline{x})\lambda\} d\lambda \\
&= \frac{b+n\overline{x}}{n+a-1}.
\end{aligned}$$

Finalizando, temos que $\delta_\psi(\boldsymbol{x}) = \dfrac{8,6}{4-1} = 2,867$ (é diferente de $1/\mathrm{E}(\lambda|\boldsymbol{x})$).

Estimação por máxima verossimilhança

Definição (relembrando)

Se a função (de densidade) de probabilidade $f(\boldsymbol{x}|\theta)$ das observações de uma amostra aleatória é vista como uma função de θ para dado valor de $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, então esta é chamada de **função de verossimilhança**.

Notação: $f(\boldsymbol{x}|\theta)$ ou $L(\theta)$. (Não confundir com a função perda $L(\theta, a)$!)

Definição

Para cada possível valor de $\boldsymbol{x}$, seja $\delta(\boldsymbol{x}) \in \Omega$ o valor de $\theta \in \Omega$ que maximiza a função de verossimilhança $f(\boldsymbol{x}|\theta)$. Seja $\hat{\theta} = \delta(\boldsymbol{X})$ o estimador definido desta forma. O estimador $\hat{\theta}$ é chamado de **estimador de máxima verossimilhança** de θ. Após observar $\boldsymbol{X} = \boldsymbol{x}$, o valor $\delta(\boldsymbol{x})$ é chamado de **estimativa de máxima verossimilhança**.

Observação: às vezes $\hat{\theta}$ é utilizado como notação para estimador e para estimativa.

Note que $\max_z h(z) = \max_z \ln(h(z))$.

Exemplo

Suponha que $(X_1, X_2, \ldots, X_n)$ formam uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$. Então, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} [\lambda \exp\{-\lambda x_i\}] = \lambda^n \exp\{-\lambda n\overline{x}\},$$

em que $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Então,

$$\begin{aligned}
\ell(\lambda) &= \ln f(\boldsymbol{x}|\lambda) = n \ln(\lambda) - \lambda n\overline{x} \\
\frac{d}{d\lambda}\ell(\lambda) &= \frac{n}{\lambda} - n\overline{x} = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{\overline{x}} \quad \text{(estimativa)} \\
\frac{d^2}{d\lambda^2}\ell(\lambda) &= -\frac{1}{\lambda^2} < 0.
\end{aligned}$$

Então,

- o estimador de máxima verossimilhança de λ é dado por $\hat{\lambda} = \frac{1}{\overline{X}}$; e
- a estimativa de máxima verossimilhança de λ é dada por $\hat{\lambda} = \frac{1}{\overline{x}}$.

  Por exemplo, suponha $n = 3$ e $\boldsymbol{x} = (3;\ 1,5;\ 2,1)$. Então, a estimativa de máxima verossimilhança de λ é dada por

$$\hat{\lambda} = \frac{3}{3 + 1,5 + 2,1} = 0,4545.$$

Note que o ponto crítico é realmente o ponto de máximo da função de verossimilhança.

Exemplo

Suponha que um laboratório esteja oferecendo um teste médico para COVID-19. O teste é 90% confiante, isto é, se uma pessoa tem COVID-19, há uma probabilidade de 0,9 do teste fornecer um resultado positivo; por outro lado, se a pessoa não tem COVID-19, há uma probabilidade de apenas 0,1 do teste fornecer resultado positivo.

Seja X a variável aleatória representando o resultado do teste tal que

$$X = \begin{cases} 1, & \text{se positivo;} \\ 0, & \text{se negativo.} \end{cases}$$

Note que $\Omega = \{0,1;\ 0,9\}$. Temos também que $(X|\pi) \sim \mathcal{BR}(\pi)$, tal que

$$f(x|\pi) = \pi^x(1-\pi)^{1-x}, \quad \text{para} \quad x = \{0;\ 1\}.$$

Suponha que observamos $X = 0$ tal que

$$f(0|\pi) = \begin{cases} \pi^0(1-\pi)^{1-0} = 0,9, & \text{se } \pi = 0,1; \\ \pi^0(1-\pi)^{1-0} = 0,1, & \text{se } \pi = 0,9. \end{cases}$$

Assim, a estimativa de máxima verossimilhança é dado por $\pi = 0,1$ que maximiza $f(0|\pi)$.

Agora, suponha que observamos $X = 1$ tal que

$$f(1|\pi) = \begin{cases} \pi^1(1-\pi)^{1-1} = 0,1, & \text{se } \pi = 0,1; \\ \pi^1(1-\pi)^{1-1} = 0,9, & \text{se } \pi = 0,9. \end{cases}$$

Logo, a estimativa de máxima verossimilhança é dado por $\pi = 0,9$ que maximiza $f(1|\pi)$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$. Assim, temos que

$$
\begin{aligned}
f(\boldsymbol{x}|\pi) &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})} \\
\ell(\pi) &= n\overline{x}\ln(\pi) + n(1-\overline{x})\ln(1-\pi) \\
\frac{d}{d\pi}\ell(\pi) &= \frac{n\overline{x}}{\pi} - \frac{n(1-\overline{x})}{1-\pi} = 0 \quad \Rightarrow \quad \hat{\pi} = \overline{x} \quad \text{para} \quad \overline{x} \notin \{0;\ 1\} \\
\frac{d^2}{d\pi^2}\ell(\pi) &= -\frac{n\overline{x}}{\pi^2} - \frac{n(1-\overline{x})}{(1-\pi)^2} < 0.
\end{aligned}
$$

Então,

- o estimador de máxima verossimilhança de π é dado por $\hat{\pi} = \overline{X}$;
- a estimativa de máxima verossimilhança de π é dada por $\hat{\pi} = \overline{x}$;
- se $\overline{x} = 0$, então $\ell(\pi)$ é uma função decrescente de π com máximo em $\pi = 0$; e
- se $\overline{x} = 1$, então $\ell(\pi)$ é uma função crescente de π com máximo em $\pi = 1$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu, \sigma_0^2)$ com σ_0^2 conhecido. Assim, temos que

$$
\begin{aligned}
f(\boldsymbol{x}|\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(x_i-\mu)^2\right\} \\
&= (2\pi)^{-n/2}(\sigma_0^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
\ell(\mu) &= -\frac{n}{2}\left[\ln(2\pi)+\ln(\sigma_0^2)\right] - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{d}{d\mu}\ell(\mu) &= \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{n\overline{x}}{\sigma_0^2} - \frac{n\mu}{\sigma_0^2} = 0 \quad \Rightarrow \quad \hat{\mu} = \overline{x} \\
\frac{d^2}{d\mu^2}\ell(\pi) &= -\frac{n}{\sigma_0^2} < 0.
\end{aligned}
$$

- o estimador de máxima verossimilhança de μ é dado por $\hat{\mu} = \overline{X}$; e
- a estimativa de máxima verossimilhança de μ é dada por $\hat{\mu} = \overline{x}$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$. Assim, para (μ, σ^2), temos que

$$\begin{aligned}
\ell(\mu, \sigma^2) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{n\overline{x}}{\sigma^2} - \frac{n\mu}{\sigma^2} = 0 \\
\frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.
\end{aligned}$$

Resolvendo o sistema de equações acima, obtemos que

- o estimador de MV de (μ, σ^2) é $(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{X};\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$; e
- a estimativa de MV de (μ, σ^2) é $(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{x};\ \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)$.

Continuando, temos que

$$\begin{aligned} h_{11}(\mu,\sigma^2) &= \frac{\partial^2}{\partial\mu^2}\ell(\mu,\sigma^2) = -\frac{n}{\sigma^2} \\ h_{22}(\mu,\sigma^2) &= \frac{\partial^2}{\partial(\sigma^2)^2}\ell(\mu,\sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2 \\ h_{12}(\mu,\sigma^2) &= \frac{\partial^2}{\partial\mu\partial\sigma^2}\ell(\mu,\sigma^2) = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu). \end{aligned}$$

Logo, o determinante da matriz hessiana no ponto crítico $(\hat{\mu},\hat{\sigma}^2)$ é

$$h_{11}(\hat{\mu},\hat{\sigma}^2)h_{22}(\hat{\mu},\hat{\sigma}^2) - [h_{12}(\hat{\mu},\hat{\sigma}^2)]^2 = \left[-\frac{n}{\hat{\sigma}^2}\right]\left[-\frac{n}{2\hat{\sigma}^4}\right] - 0^2 = \frac{n^2}{2\hat{\sigma}^6} > 0.$$

Além disso, note que

- $h_{11}(\hat{\mu},\hat{\sigma}^2) < 0$; e
- $h_{22}(\hat{\mu},\hat{\sigma}^2) < 0$.

Assim, $(\hat{\mu},\hat{\sigma}^2)$ é ponto de máximo da função de verossimilhança.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(0, \theta)$, isto é,

$$f(x_i|\theta) = \frac{1}{\theta} \boldsymbol{I}_{[0,\theta]}(x_i) \quad \textbf{(devemos utilizar a função indicadora)}.$$

Assim, temos que

$$L(\theta) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \left[\frac{1}{\theta} \boldsymbol{I}_{[0,\theta]}(x_i) \right] = \frac{1}{\theta^n} \boldsymbol{I}_{[0,\theta]}(x_{(n)}) = \frac{1}{\theta^n} \boldsymbol{I}_{[x_{(n)},\infty)}(\theta).$$

- $L(\theta) = 1/\theta^n$ é uma função decrescente em θ.
- Procuramos o menor valor de θ tal que $\theta \geqslant x_i$ para todo $i = 1, 2, \ldots, n$.
- Então, a estimativa é dada por $\hat{\theta} = x_{(n)} = \max\limits_{1 \leqslant i \leqslant n} x_i$.

Portanto,

- o estimador de MV de θ é dado por $\hat{\theta} = X_{(n)} = \max\limits_{1 \leqslant i \leqslant n} X_i$;
- a estimativa de MV de θ é dada por $\hat{\theta} = x_{(n)} = \max\limits_{1 \leqslant i \leqslant n} x_i$.

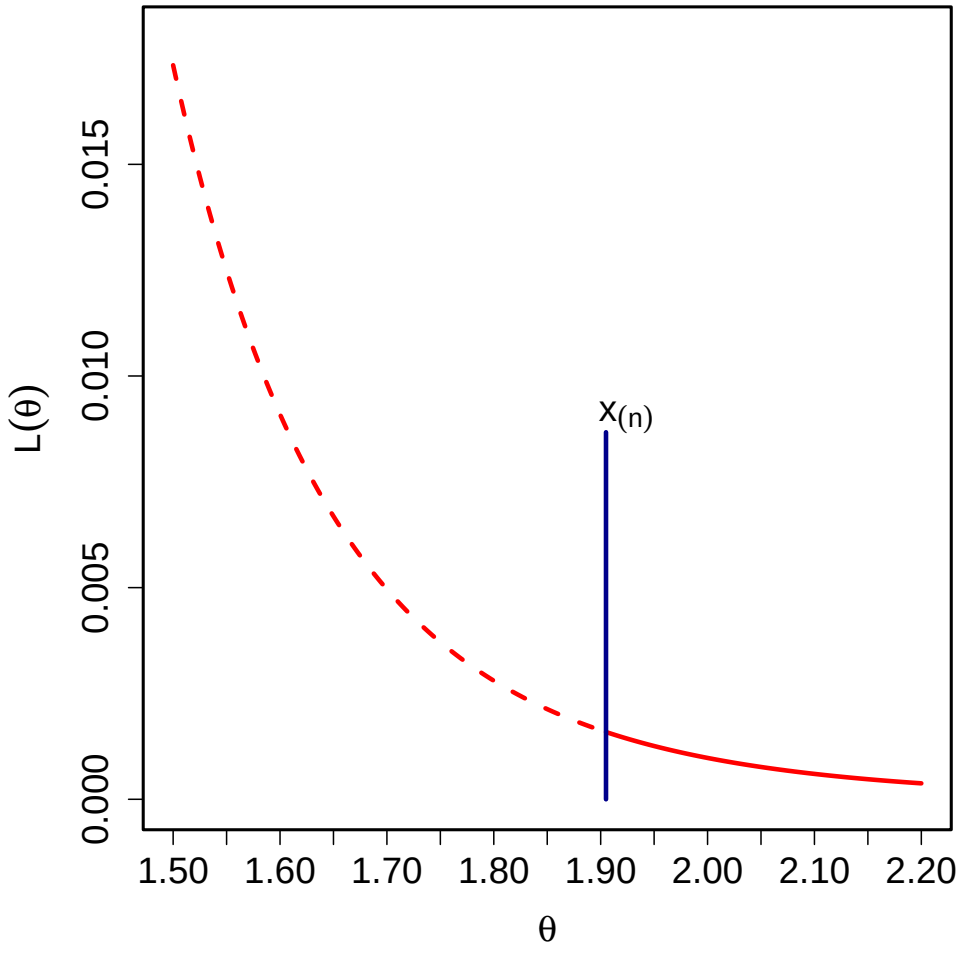


Figura: Função de verossimilhança de θ no modelo $X \sim \mathcal{U}(0, \theta)$ para $n = 10$.

Ver exemplo_05.r

Estimação por máxima verossimilhança

Para um determinado problema de inferência estatística, o estimador de máxima verossimilhança pode não existir ou não ser único.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(\theta;\ \theta + 1)$, isto é,

$$f(x_i|\theta) = \boldsymbol{I}_{[\theta,\theta+1]}(x_i) \quad \text{e} \quad L(\theta) = \prod_{i=1}^{n} \boldsymbol{I}_{[\theta,\theta+1]}(x_i) = \boldsymbol{I}_{[x_{(n)}-1,x_{(1)}]}(\theta),$$

pois $\theta \leqslant x_i$ e $\theta + 1 \geqslant x_i$ para todo $i = 1, 2, \ldots, n$.

Note que $L(\theta)$ é constante no intervalo $[x_{(n)} - 1, x_{(1)}]$. Então, o estimador de máxima verossimilhança é qualquer valor neste intervalo.

Portanto, o **estimador de máxima verossimilhança** de θ **não é único!**

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória da distribuição dada por

$$f(x_i|\mu, \sigma^2) = \frac{1}{2\sqrt{2\pi}} \left[\exp\left\{ -\frac{x_i^2}{2} \right\} + \frac{1}{\sqrt{\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\} \right],$$

(mistura discreta de "$\mathcal{N}(0;\ 1)$" e "$\mathcal{N}(\mu;\ \sigma^2)$" com pesos iguais).

Temos que

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2).$$

Se tomarmos $\mu = x_k$, para $x_k \in \{x_1, x_2, \ldots, x_n\}$, e $\sigma^2 \to 0$, então $L(\mu, \sigma^2) \to \infty$ pois

$f(x_k|\mu, \sigma^2) \to \infty$ e $f(x_i|\mu, \sigma^2) \to \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{x_i^2}{2} \right\}$ para todo $i \neq k$.

Contudo, sabemos que $\sigma^2 > 0$.

Portanto, o **estimador de máxima verossimilhança** de (μ, σ^2) **não existe!**

Propriedade de invariância do estimador de máxima verossimilhança

Teorema

Se $\hat{\theta}$ é o estimador de máxima verossimilhança de θ e se g é uma função biunívoca, então $g(\hat{\theta})$ é o estimador de máxima verossimilhança de $g(\theta)$.

Prova

Considere $\psi = g(\theta)$ como novo parâmetro e seja Ψ o novo espaço paramétrico referente a imagem de Ω sob a função g. Seja também $\theta = h(\psi)$ a função inversa de g. Então, $f(x|h(\psi))$ é a função (de densidade) de probabilidade expressa em termos de ψ e a função de verossimilhança é dada por $f(\boldsymbol{x}|h(\psi))$.

O estimador de máxima verossimilhança $\hat{\psi}$ de ψ será igual ao valor de ψ que maximiza $f(\boldsymbol{x}|h(\psi))$. Como $f(\boldsymbol{x}|\theta)$ é maximizada quando $\theta = \hat{\theta}$, então segue que $f(\boldsymbol{x}|h(\psi))$ é maximizada quando $h(\psi) = \hat{\theta}$. Portanto, o estimador de máxima verossimilhança $\hat{\psi}$ deve satisfazer a relação
$\hat{\theta} = h(\hat{\psi}) \quad \Leftrightarrow \quad \hat{\psi} = g(\hat{\theta})$. □

Definição

Seja $g(\theta)$ uma função arbitrária do parâmetro θ, e seja G a imagem de Ω sob a função g. Para cada $t \in G$, defina $G_t = \{\theta : g(\theta) = t\}$ e

$$L^\star(t) = \max_{\theta \in G_t} \ln f(\boldsymbol{x}|\theta).$$

Finalmente, defina o estimador de máxima verossimilhança de $g(\theta)$ como $\hat{t}$, em que

$$L^\star(\hat{t}) = \max_{t \in G_t} L^\star(t).$$

Teorema

Seja $\hat{\theta}$ o estimador de máxima verossimilhança de θ e seja $g(\theta)$ uma função de θ. Então, o estimador de máxima verossimilhança de $g(\theta)$ é $g(\hat{\theta})$.

Prova

Considere as quantidades definidas na página anterior: θ, $g(\theta)$, G, t, G_t, $L^\star(t)$, $\hat{t}$ e $L^\star(\hat{t})$. Como $L^\star(t)$ é o máximo de $\ln f(\boldsymbol{x}|\theta)$ sobre θ em um subconjunto de Ω e, como $\ln f(\boldsymbol{x}|\hat{\theta})$ é o máximo sob todo o vetor θ, então sabemos que $L^\star(t) \leqslant \ln f(\boldsymbol{x}|\hat{\theta})$ para todo $t \in G$. Seja $\hat{t} = g(\hat{\theta})$. Note que $\hat{\theta} \in G_{\hat{t}}$ já que $\hat{\theta}$ maximiza $f(\boldsymbol{x}|\theta)$ para todo o vetor θ e, também, maximiza $f(\boldsymbol{x}|\theta)$ para $\theta \in G_{\hat{t}}$. Portanto, $L^\star(\hat{t}) = \ln f(\boldsymbol{x}|\hat{\theta})$ e $\hat{t} = g(\hat{\theta})$ é o estimador de máxima verossimilhança de $g(\theta)$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$.
Desejamos encontrar os estimadores de máxima verossimilhança de σ e de $\mathrm{E}(X^2|\mu, \sigma^2) = \mu^2 + \sigma^2$.

Sabemos que $(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$.

Assim, temos que

- $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ é o estimador de MV de σ; e
- $\widehat{\mathrm{E}(X^2|\mu, \sigma^2)} = \widehat{\mu^2 + \sigma^2} = \hat{\mu}^2 + \hat{\sigma}^2$ é o estimador de MV de $\mathrm{E}(X^2|\mu, \sigma^2)$.

Note que $\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Consistência

Definição

Suponha que para todo n suficientemente grande existe um único estimador de máxima verossimilhança de θ.

Então, sob certas condições, a sequência de estimadores de máxima verossimilhança é uma sequência consistente de estimadores de θ.

Em outras palavras, a sequência de estimadores de máxima verossimilhança converge em probabilidade para o valor desconhecido do parâmetro θ quando $n \to \infty$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu, \sigma_0^2)$ com σ_0^2 conhecido. Sabemos que $\hat{\mu} = \overline{X}$ é o estimador de máxima verossimilhança de μ. Além disso, temos que

$$0 \leqslant \lim_{n\to\infty} \Pr(|\overline{X} - \mu| > \varepsilon) \leqslant \lim_{n\to\infty} \frac{\sigma_0^2}{n\varepsilon^2} = 0.$$

Portanto, $\overline{X}$ é um estimador consistente de μ.

Computação numérica

Existem casos em que o estimador (estimativa) de máxima verossimilhança é única, mas não existe forma fechada conhecida como função dos dados.

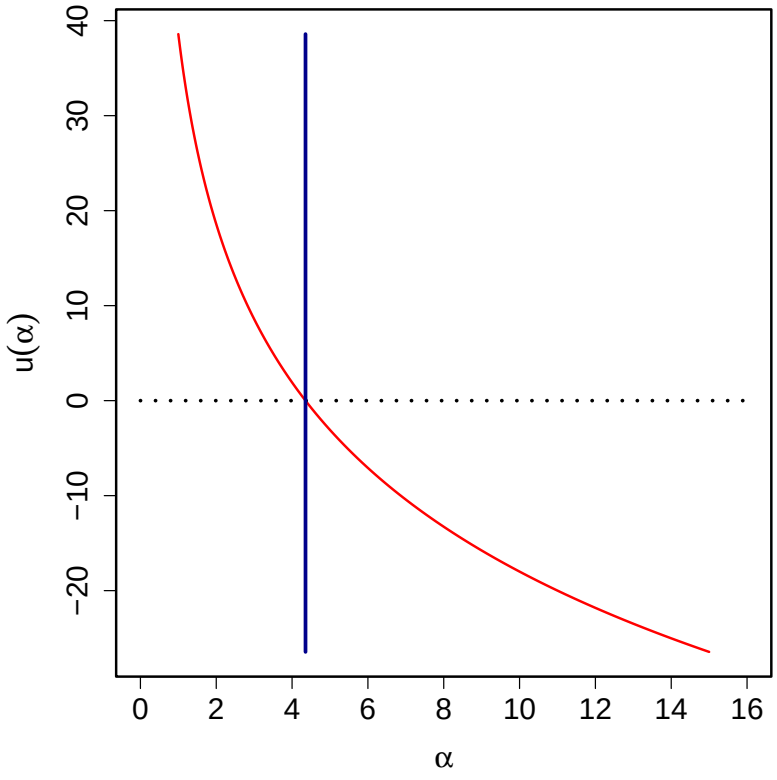
Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha) \sim \mathcal{G}(\alpha;\ 1)$. Temos que

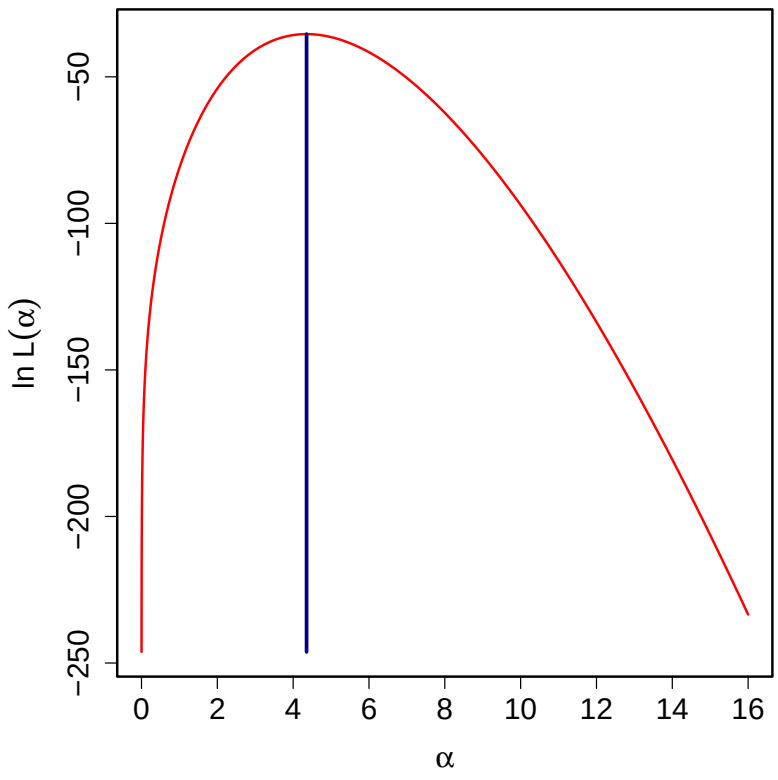
$$
\begin{aligned}
f(\boldsymbol{x}|\alpha) &= \prod_{i=1}^{n}\left[\frac{1}{\Gamma(\alpha)}x_i^{\alpha-1}\exp\{-x_i\}\right] = \Gamma(\alpha)^{-n}\left[\prod_{i=1}^{n}x_i^{\alpha-1}\right]\exp\{-n\overline{x}\} \\
\ell(\alpha) &= -n\ln\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\ln(x_i) - n\overline{x} \\
\frac{d\ell(\alpha)}{d\alpha} &= -n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\ln(x_i) = 0.
\end{aligned}
$$

cuja solução explícita não existe. Contudo, podemos recorrer a métodos numéricos.

Lembre-que $\Gamma(\alpha) = \int_0^\infty s^{\alpha-1}\exp\{-s\}ds$ e $\varphi(\alpha) = \dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ é conhecida como a função digama.



(a) $u(\alpha) = \dfrac{d\ell(\alpha)}{d\alpha}$



(b) $\ell(\alpha) = \ln L(\alpha)$

Figura: Método da bisseção para encontrar a estimativa de máxima verossimilhança.

Ver exemplo_06.r

Estimação pelo Método dos Momentos

Seja $(X_1, X_2, \ldots, X_n)$ uma amosta aleatória de uma distribuição que é indexada por um vetor paramétrico $\boldsymbol{\theta}$ de dimensão k, e esta distribuição tem pelo menos k momentos finitos.

Para $j = 1, 2, \ldots, k$, seja $\mu_j = \mathrm{E}(X^j|\boldsymbol{\theta})$.

Suponha que $\mu(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \mu_2(\boldsymbol{\theta}), \ldots, \mu_k(\boldsymbol{\theta}))$ seja uma função bijetiva de $\boldsymbol{\theta}$.

Seja $H(\mu_1, \mu_2, \ldots, \mu_k)$ a função inversa, isto é,

$$\boldsymbol{\theta} = H(\mu_1(\boldsymbol{\theta}), \mu_2(\boldsymbol{\theta}), \ldots, \mu_k(\boldsymbol{\theta})).$$

Defina os momentos amostrais (ordinários) por

$$M_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad \text{para} \quad j = 1, 2, \ldots, k.$$

O estimador de θ pelo método dos momentos é $H(M_1, M_2, \ldots, M_k)$.

Observação: quando um momento de ordem inferior não depender do(s) parâmetro(s), devemos utilizar o momento de ordem imediatamente superior (deve existir e ser finito). Por exemplo, veja exercício sobre $\mathcal{U}(-\theta, \theta)$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha) \sim \mathcal{G}(\alpha;\ 1)$ tal que $\mathrm{E}(X|\alpha) = \alpha$. Portanto, o estimador de α pelo método dos momentos é dado por $\hat{\alpha} = \overline{X}$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha, \beta) \sim \mathcal{G}(\alpha;\ \beta)$ tal que

$$\mathrm{E}(X|\alpha, \beta) = \frac{\alpha}{\beta} \quad \text{e} \quad \mathrm{E}(X^2|\alpha, \beta) = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}.$$

Igualando-se a M_1 e a M_2, respectivamente, obtemos

$$M_1 = \frac{\alpha}{\beta} \quad \text{e} \quad M_2 = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}.$$

Resolvendo o sistema em termos de α e β, o estimador de (α, β) pelo método dos momentos é dado por

$$\hat{\alpha} = \frac{M_1^2}{M_2 - M_1^2} \quad \text{e} \quad \hat{\beta} = \frac{M_1}{M_2 - M_1^2}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(\theta;\ \theta + 1)$ tal que $\mathrm{E}(X|\theta) = \theta + \dfrac{1}{2}$. Portanto, o estimador de θ pelo método dos momentos é dado por $\hat{\theta} = \overline{X} - \dfrac{1}{2}$.

Como visto anteriormente, temos que

$$x_{(n)} - 1 \leqslant \theta \leqslant x_{(1)}.$$

Agora, suponha $\boldsymbol{x} = (0,2;\ 0,99;\ 0,01)$ uma amostra aleatória de tamanho $n = 3$. É fácil verificar que $\overline{x} = 0,4$, $x_{(1)} = 0,01$ e $x_{(3)} = 0,99$ tal que

$$-0,01 \leqslant \theta \leqslant 0,01.$$

Então, a estimativa de θ pelo método dos momentos é dada por

$$\hat{\theta} = \overline{x} - \frac{1}{2} = 0,4 - 0,5 = -0,1,$$

e, consequentemente, **está fora do espaço paramétrico**.

Teorema

Suponha que $(X_1, X_2, \ldots, X_n)$ seja uma amosta aleatória de uma distribuição que é indexada por um vetor paramétrico θ de dimensão k, que esta distribuição tem pelo menos k momentos finitos e que a função inversa H definida anteriormente seja contínua. Então, a sequência de estimadores pelo método dos momentos baseados em $(X_1, X_2, \ldots, X_n)$ é uma sequência consistente de estimadores de θ.

Prova

Aplicações seguidas da lei dos grandes números e do teorema de Slutsky.

Estatística Suficiente

Definição

Se T é uma estatística e t é qualquer particular valor de T, então a distribuição condicional de $(X_1, X_2, \ldots, X_n)$ dado $T = t$ pode ser calculada a partir da distribuição conjunta $f(\boldsymbol{x}|\theta)$. Em geral, essa distribuição dependerá de θ. Assim, para cada valor de t, haverá uma família de possíveis distribuições condicionais correspondendo aos diferentes possíveis valores de $\theta \in \Omega$.

Contudo, pode acontecer que, para cada possível valor de t, a distribuição condicional conjunta de $(X_1, X_2, \ldots, X_n)$ dado $T = t$, seja a mesma para todos os valores de $\theta \in \Omega$ e, portanto, ela não depende do valor de θ. Neste caso, diz-se que **T é uma estatística suficiente para o parâmetro θ**.

T é uma estatística suficiente se $f(\boldsymbol{x}|T = t, \theta) = f(\boldsymbol{x}|T = t)$.

Critério da fatoração

Teorema

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$ com θ desconhecido e $\theta \in \Omega$. Uma estatística $T = r(X_1, X_2, \ldots, X_n)$ é uma **estatística suficiente** se, e somente se, a função (de densidade) de probabilidade $f(\boldsymbol{x}|\theta)$ de $(X_1, X_2, \ldots, X_n)$ puder ser fatorada da seguinte forma

$$f(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v(r(\boldsymbol{x}), \theta),$$

para todos os valores $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ e $\theta \in \Omega$.

- As funções u e v são não negativas.
- A função u pode depender de $\boldsymbol{x}$, mas não de θ.
- A função v dependerá de θ e dos valores observados $\boldsymbol{x}$ somente através do valor da estatística $t = r(\boldsymbol{x})$.

Prova

Consideraremos apenas o caso em que as variáveis aleatórias são discretas tal que a distribuição de $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ é dada por

$$f(\boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta).$$

($\Rightarrow$) Suponha que $f(\boldsymbol{x}|\theta)$ possa ser fatorada como $f(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v(r(\boldsymbol{x}), \theta)$ para todos os valores $\boldsymbol{x} \in \mathbb{R}^n$ e $\theta \in \Omega$.Para cada possível valor de t de T, seja $A(t)$ o conjunto de todos os pontos $\boldsymbol{x} \in \mathbb{R}^n$ tal que $r(\boldsymbol{x}) = t$. Para qualquer valor dado de $\theta \in \Omega$, devemos determinar a distribuição condicional de $\boldsymbol{X}$ dado $T = t$. Para qualquer ponto $\boldsymbol{x} \in A(t)$,

$$\begin{aligned}
\Pr(\boldsymbol{X} = \boldsymbol{x}|T = t, \theta) &= \frac{\Pr(\boldsymbol{X} = \boldsymbol{x}|\theta)}{\Pr(T = t|\theta)} = \frac{f(\boldsymbol{x}|\theta)}{\sum_{\boldsymbol{y} \in A(t)} f(\boldsymbol{y}|\theta)} \\
&= \frac{u(\boldsymbol{x})v(r(\boldsymbol{x}), \theta)}{\sum_{\boldsymbol{y} \in A(t)} u(\boldsymbol{y})v(r(\boldsymbol{y}), \theta)} = \frac{u(\boldsymbol{x})v(t, \theta)}{\sum_{\boldsymbol{y} \in A(t)} u(\boldsymbol{y})v(t, \theta)} \\
&= \frac{u(\boldsymbol{x})}{\sum_{\boldsymbol{y} \in A(t)} u(\boldsymbol{y})}
\end{aligned}$$

Finalmente, para qualquer ponto $\boldsymbol{x} \notin A(t)$,

$$\Pr(\boldsymbol{X} = \boldsymbol{x} | T = t, \theta) = 0.$$

Logo, a distribuição de $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ condicional a $T = t$ e θ não depende de θ e, portanto, T é uma estatística suficiente para θ.

($\Leftarrow$) Por outro lado, suponha que T seja uma estatística suficiente. Então, para qualquer t de T, $\boldsymbol{x} \in A(t)$ e qualquer $\theta \in \Omega$, $\Pr(\boldsymbol{X} = \boldsymbol{x} | T = t, \theta)$ não dependerá de θ. Portanto, temos que

$$\Pr(\boldsymbol{X} = \boldsymbol{x} | T = t, \theta) = \Pr(\boldsymbol{X} = \boldsymbol{x} | T = t) = u(\boldsymbol{x}).$$

Se fizermos $v(t, \theta) = \Pr(T = t | \theta)$, então

$$f(\boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x}|\theta) = \Pr(\boldsymbol{X} = \boldsymbol{x} | T = t, \theta) \Pr(T = t|\theta) = u(\boldsymbol{x}) v(t, \theta).$$

Para qualquer valor de $\boldsymbol{x}$ tal que $f(\boldsymbol{x}|\theta) = 0$ para todos os valores de $\theta \in \Omega$, o valor da função $u(\boldsymbol{x})$ pode ser escolhido igual a zero. $\square$

Portanto, quando este critério for utilizado, basta verificar que a fatorização da definição é satisfeita para todo $\boldsymbol{x}$ tal que $f(\boldsymbol{x}|\theta) > 0$, para pelo menos um valor de $\theta \in \Omega$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$ tal que $\mathrm{E}(X|\lambda) = \lambda$. Seja $t = r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$. Mostraremos que $T = r(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ é uma estatística suficiente para λ. Temos que

$$f(\boldsymbol{x}|\lambda) = \prod_{i=1}^{n} \left[\frac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!} \right] = \left[\prod_{i=1}^{n} \frac{1}{x_i!} \right] \exp\{-n\lambda\}\lambda^{r(\boldsymbol{x})}.$$

Sejam

$$u(\boldsymbol{x}) = \left[\prod_{i=1}^{n} \frac{1}{x_i!} \right] \quad \text{e} \quad v(t, \lambda) = \exp\{-n\lambda\}\lambda^{r(\boldsymbol{x})} = \exp\{-n\lambda\}\lambda^{t}.$$

Então, pelo teorema da fatoração, temos que

$$T = \sum_{i=1}^{n} X_i \quad \textbf{é uma estatística suficiente para } \lambda\textbf{.}$$

Corolário

Uma estatística $T = r(\boldsymbol{X})$ é suficiente se, e somente se, qualquer que seja a distribuição a priori, a distribuição a posteriori de θ depende dos dados apenas através do valor de T.

(Veja, por exemplo, pag. 13, aula_01.pdf; pag. 8 a 15, e aula_02.pdf.)

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória da distribuição com função de densidade de probabilidade dada por

$$f(x|\theta) = \theta x^{\theta-1} \boldsymbol{I}_{[0,1]}(x), \quad \text{para} \quad \theta > 0,$$

com θ desconhecido.

Seja $t = r(\boldsymbol{x}) = \prod_{i=1}^{n} x_i$. Mostraremos que $T = r(\boldsymbol{X}) = \prod_{i=1}^{n} X_i$ é uma estatística suficiente para θ. Temos que

$$f(\boldsymbol{x}|\theta) = \theta^n \left[\prod_{i=1}^{n} x_i\right]^{\theta-1} \left[\prod_{i=1}^{n} \boldsymbol{I}_{[0,1]}(x_i)\right] = \theta^n [r(\boldsymbol{x})]^{\theta-1} \left[\prod_{i=1}^{n} \boldsymbol{I}_{[0,1]}(x_i)\right].$$

Sejam

$$u(\boldsymbol{x}) = \prod_{i=1}^{n} \boldsymbol{I}_{[0,1]}(x_i) \quad \text{e} \quad v(t,\theta) = \theta^n [r(\boldsymbol{x})]^{\theta-1} = \theta^n t^{\theta-1}.$$

Então, pelo teorema da fatoração, temos que

$$T = \prod_{i=1}^{n} X_i \quad \textbf{é uma estatística suficiente para } \theta.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(0, \theta)$ com $\theta > 0$ e desconhecido tal que $f(x|\theta) = \dfrac{1}{\theta} \boldsymbol{I}_{[0,\theta]}(x)$.

Seja $t = r(\boldsymbol{x}) = x_{(n)}$. Mostraremos que $T = r(\boldsymbol{X}) = X_{(n)}$ é uma estatística suficiente para θ. Temos que

$$f(\boldsymbol{x}|\theta) = \theta^{-n} \left[\prod_{i=1}^{n} \boldsymbol{I}_{[0,\theta]}(x_i) \right] = \theta^{-n} \boldsymbol{I}_{[0,\theta]}(x_{(n)}) = \theta^{-n} \boldsymbol{I}_{[x_{(n)},\infty)}(\theta).$$

Sejam

$$u(\boldsymbol{x}) = 1 \quad \text{e} \quad v(t, \theta) = \theta^{-n} \boldsymbol{I}_{[x_{(n)},\infty)}(\theta) = \theta^{-n} \boldsymbol{I}_{[t,\infty)}(\theta).$$

Então, pelo teorema da fatoração, temos que

$$T = \max_{1 \leqslant i \leqslant n} X_i = X_{(n)} \quad \textbf{é uma estatística suficiente para } \theta.$$

Estatísticas Conjuntamente Suficientes

Definição

Suponha que para cada θ e cada possível valor $(t_1, t_2, \ldots, t_k)$ de $(T_1, T_2, \ldots, T_k)$, a distribuição conjunta condicional de $(X_1, X_2, \ldots, X_n)$ dado $(T_1, T_2, \ldots, T_k) = (t_1, t_2, \ldots, t_k)$ não depende de θ. Então, $(T_1, T_2, \ldots, T_k)$ são estatísticas conjuntamente suficientes para θ.

Teorema

Sejam $(r_1, r_2, \ldots, r_k)$ funções de n variáveis reais. As estatísticas $T_j = r_j(\boldsymbol{X})$, para $j = 1, 2, \ldots, k$, são estatísticas conjuntamente suficientes para θ se, e somente se, a função (de densidade) de probabilidade $f(\boldsymbol{x}|\theta)$ puder ser fatorada, para todos os valores de $\boldsymbol{x} \in \mathbb{R}^n$ e $\theta \in \Omega$, como

$$f(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v(r_1(\boldsymbol{x}), r_2(\boldsymbol{x}), \ldots, r_k(\boldsymbol{x}), \theta).$$

- As funções u e v são não negativas.
- A função u pode depender de $\boldsymbol{x}$, mas não de θ.
- A função v dependerá de θ e dos valores observados $\boldsymbol{x}$ somente através dos valores das k estatísticas $t_j = r_j(\boldsymbol{x})$, para $j = 1, 2, \ldots, k$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$. Temos que

$$\begin{aligned} f(\boldsymbol{x}|\mu, \sigma^2) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2}\right\}. \end{aligned}$$

Sejam

$$u(\boldsymbol{x}) = (2\pi)^{-n/2} \quad \text{e} \quad v(t_1, t_2, \mu, \sigma^2) = (\sigma^2)^{-n/2} \exp\left\{-\frac{t_2}{2\sigma^2} + \frac{t_1\mu}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right\},$$

com $t_1 = \sum_{i=1}^{n} x_i$ e $t_2 = \sum_{i=1}^{n} x_i^2$. Seja $(T_1, T_2) = \left(\sum_{i=1}^{n} X_i;\ \sum_{i=1}^{n} X_i^2\right)$.

Então, pelo teorema da fatoração, temos que

(T_1, T_2) são estatísticas conjuntamente suficientes para (μ, σ^2).

Corolário

Suponha que as estatísticas $(T_1, T_2, \ldots, T_k)$ são conjuntamente suficientes para o vetor paramétrico $\boldsymbol{\theta}$. Se outras k estatísticas $(T_1', T_2', \ldots, T_k')$ são obtidas como transformação biunívoca de $(T_1, T_2, \ldots, T_k)$, então $(T_1', T_2', \ldots, T_k')$ também são conjuntamente suficientes para $\boldsymbol{\theta}$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$. Temos que $(T_1, T_2) = \left(\sum_{i=1}^n X_i;\ \sum_{i=1}^n X_i^2\right)$ são estatísticas conjuntamente suficientes para (μ, σ^2).

Considere $(T_1', T_2') = \left(\overline{X};\ \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right)$. Claramente, $T_1 = nT_1'$ e

$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 \quad \Rightarrow \quad T_2 = nT_2' + n[T_1']^2.$$

Como (T_1', T_2') tem relação biunívoca com (T_1, T_2), então

(T_1', T_2') são estatísticas conjuntamente suficientes para (μ, σ^2).

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha, \beta) \sim \mathcal{U}(\alpha, \beta)$ com $-\infty < \alpha < \beta < \infty$ desconhecidos. Temos que

$$\begin{aligned} f(\boldsymbol{x}|\alpha,\beta) &= \prod_{i=1}^{n}\left[\frac{1}{(\beta-\alpha)}\boldsymbol{I}_{[\alpha,\beta]}(x_i)\right] = \frac{1}{(\beta-\alpha)^n}\boldsymbol{I}_{[\alpha,\beta]}(x_{(1)})\boldsymbol{I}_{[\alpha,\beta]}(x_{(n)}) \\ &= \frac{1}{(\beta-\alpha)^n}\boldsymbol{I}_{(-\infty,x_{(1)}]}(\alpha)\boldsymbol{I}_{[x_{(n)},\infty)}(\beta) \quad \text{com} \quad x_{(1)} < x_{(n)}, \quad n \geqslant 2. \end{aligned}$$

Sejam

$$u(\boldsymbol{x}) = 1 \quad \text{e} \quad v(t_1, t_2, \alpha, \beta) = \frac{1}{(\beta-\alpha)^n}\boldsymbol{I}_{(-\infty,t_1]}(\alpha)\boldsymbol{I}_{[t_2,\infty)}(\beta),$$

com $t_1 = x_{(1)} = \min_{1\leqslant i\leqslant n} x_i$ e $t_2 = x_{(n)} = \max_{1\leqslant i\leqslant n} x_i$. Seja

$$(T_1, T_2) = \left(\min_{1\leqslant i\leqslant n} X_i;\ \max_{1\leqslant i\leqslant n} X_i\right) = \left(X_{(1)}, X_{(n)}\right).$$

Então, pelo teorema da fatoração, temos que

(T_1, T_2) são estatísticas conjuntamente suficientes para (α, β).

Definição

Suponha que $(X_1, X_2, \ldots, X_n)$ formam uma amostra aleatória de alguma distribuição. Sejam $X_{(1)}$ a menor observação, $X_{(2)}$ a segunda menor, até $X_{(n)}$ a maior observação da amostra. **As variáveis $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ são chamadas de estatísticas de ordem da amostra.**

Teorema

Suponha que $(X_1, X_2, \ldots, X_n)$ formam uma amostra aleatória de uma distribuição para a qual a função (de densidade) de probabilidade é $f(\boldsymbol{x}|\theta)$. Então, as estatísticas de ordem são conjuntamente suficientes para θ.

Prova

Sejam $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ os valores das estatísticas de ordem. A função (de densidade) de probabilidade conjunta é dada por

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Como a ordem dos fatores no produto é irrelevante, podemos escrever

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_{(i)}|\theta).$$

Como $f(\boldsymbol{x}|\theta)$ depende de $\boldsymbol{x}$ somente através dos valores $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, então $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ são conjuntamente suficiente para θ. □

Para algumas distribuições as estatísticas de ordem formam o conjunto mais simples de estatísticas conjuntamente suficientes.

Estatística Suficiente Minimal

Definição

Uma estatística T é uma estatística suficiente minimal se:

- T é estatística suficiente; e
- T é função de qualquer outra estatística suficiente T'.

Dado uma amostra aleatória $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, definimos $T = r(\boldsymbol{X})$ e $T' = r'(\boldsymbol{X})$.

Dizer que T é uma função de T' significa que se $r'(\boldsymbol{x}) = r'(\boldsymbol{y})$, então $r(\boldsymbol{x}) = r(\boldsymbol{y})$.

Definição

Um vetor de estatísticas $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$ são estatísticas conjuntamente suficientes minimais se:

- as coordenadas de $\boldsymbol{T}$ são estatística suficientes; e
- $\boldsymbol{T}$ é função de qualquer outra estatística conjuntamente suficiente.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$ com λ desconhecido.

Temos que

$$T_1 = \sum_{i=1}^{n} X_i \quad \text{e} \quad T_1' = \overline{X}$$

são estatísticas suficientes com T_1' como função de T_1. Logo, T_1 é uma estatística suficiente minimal. Por outro lado, T_1' também é uma estatística suficiente minimal.

Note que $T_1 = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_{(i)}$, isto é, T_1 é uma função das estatísticas de ordem $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ que são conjuntamente suficientes. Portanto, T_1 é minimal.

Contudo, as estatísticas de ordem $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ não são estatísticas conjuntamente suficientes minimais porque não é possível escrevê-las como função de T_1.

Teorema

Seja $f(\boldsymbol{x}|\theta)$ a função (de densidade) de probabilidade conjunta de uma amostra aleatória $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. Suponha que existe uma função $T(\boldsymbol{x})$ tal que, para quaisquer dois pontos amostrais $\boldsymbol{x}$ e $\boldsymbol{y}$, a razão $f(\boldsymbol{x}|\theta)/f(\boldsymbol{y}|\theta)$ é uma constante como função de θ se, e somente se, $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Então, $T(\boldsymbol{X})$ é uma estatística suficiente minimal para θ.

Prova

Veja *Theorem 6.2.13* e sua prova na Seção 6.2 da página 281 de *Statistical Inference*, 2^{nd} Edition, Casella, G. and Berger, R. L., 2002, Thomson Learning.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$ e considere dois pontos amostrais $\boldsymbol{x}$ e $\boldsymbol{y}$. Então,

$$\begin{aligned}
\frac{f(\boldsymbol{x}|\lambda)}{f(\boldsymbol{y}|\lambda)} &= \frac{\prod_{i=1}^{n}\left[\dfrac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!}\right]}{\prod_{i=1}^{n}\left[\dfrac{\exp\{-\lambda\}\lambda^{y_i}}{y_i!}\right]} = \frac{\left[\prod_{i=1}^{n}\dfrac{1}{x_i!}\right]\exp\{-n\lambda\}\lambda^{r(\boldsymbol{x})}}{\left[\prod_{i=1}^{n}\dfrac{1}{y_i!}\right]\exp\{-n\lambda\}\lambda^{r(\boldsymbol{y})}} \\
&= \left[\prod_{i=1}^{n}\frac{y_i!}{x_i!}\right]\lambda^{r(\boldsymbol{x})-r(\boldsymbol{y})},
\end{aligned}$$

com $r(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$ e $r(\boldsymbol{y}) = \sum_{i=1}^{n} y_i$.

A razão $f(\boldsymbol{x}|\lambda)/f(\boldsymbol{y}|\lambda)$ será uma função constante de λ se, e somente se, $r(\boldsymbol{x}) = r(\boldsymbol{y})$. Portanto, $T = r(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ é uma estatística suficiente minimal para λ.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\ \sigma^2)$ e considere dois pontos amostrais $\boldsymbol{x}$ e $\boldsymbol{y}$ tal que $(\overline{x}, s^2_{\boldsymbol{x}})$ e $(\overline{y}, s^2_{\boldsymbol{y}})$ são as respectivas médias e variâncias amostrais. Então,

$$\begin{aligned}
\frac{f(\boldsymbol{x}|\mu,\sigma^2)}{f(\boldsymbol{y}|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{x}-\mu)^2+(n-1)s^2_{\boldsymbol{x}}\right]\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{y}-\mu)^2+(n-1)s^2_{\boldsymbol{y}}\right]\right\}} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[n(\overline{x}^2-\overline{y}^2)-2\mu(\overline{x}-\overline{y})+(n-1)(s^2_{\boldsymbol{x}}-s^2_{\boldsymbol{y}})\right]\right\}
\end{aligned}$$

Esta razão será uma função constante de (μ, σ^2) se, e somente se, $(\overline{x}, s^2_{\boldsymbol{x}}) = (\overline{y}, s^2_{\boldsymbol{y}})$. Portanto, $\boldsymbol{T} = (\overline{X}, S^2_{\boldsymbol{X}})$ é uma estatística suficiente minimal para (μ, σ^2).

Lembre-se que $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $S^2_{\boldsymbol{X}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Estimadores de Bayes e de máxima verossimilhança como estatísticas suficientes

Nos teoremas a seguir, sejam $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$ em que o valor de θ é desconhecido e $\theta \in \Omega$.

Os teoremas a seguir podem ser estendidos para o caso do vetor paramétrico $\boldsymbol{\theta}$ e das estatísticas conjuntamente suficientes $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$.

Teorema

Seja $T = r(X_1, X_2, \ldots, X_n)$ uma estatística suficiente para θ. Então, o estimador de máxima verossimilhança $\hat{\theta}$ de θ depende das observações $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ somente através de T. Além disso, se $\hat{\theta}$ é uma estatística suficiente, então $\hat{\theta}$ é uma estatística suficiente minimal.

Prova

Pelo critério da fatoração $f(\boldsymbol{x}|\theta) = u(\boldsymbol{x})v(r(\boldsymbol{x}), \theta)$. O estimador de máxima verossimilhança é o valor de θ que maximiza $f(\boldsymbol{x}|\theta)$ e neste caso o valor que maximiza $v(r(\boldsymbol{x}), \theta)$. Como $v(r(\boldsymbol{x}), \theta)$ depende da amostra somente através de $r(\boldsymbol{x})$, então $\hat{\theta}$ dependerá da amostra somente através de $r(\boldsymbol{x})$. Assim, o estimador $\hat{\theta}$ é função de $T = r(\boldsymbol{X})$. Se $\hat{\theta}$ é estatística suficiente, então $\hat{\theta}$ será estatística suficiente minimal.

Exemplo

Suponha que $(X_1, X_2, \ldots, X_n)$ seja uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$. O estimador de máxima verossimilhança de (μ, σ^2) é dado por

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

que são estatísticas conjuntamente suficientes.

Note que $(\hat{\mu}, \hat{\sigma}^2)$ são funções da estatística conjuntamente suficiente

$$(T_1, T_2) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right).$$

Portanto, $(\hat{\mu}, \hat{\sigma}^2)$ são estatísticas conjuntamente suficientes minimais.

Teorema

Seja $T = r(X_1, X_2, \ldots, X_n)$ uma estatística suficiente para θ. Então, todo estimador de Bayes $\hat{\theta}$ de θ depende das observações $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ somente através de T. Além disso, se $\hat{\theta}$ é uma estatística suficiente, então $\hat{\theta}$ é uma estatística suficiente minimal.

Prova

Seja $f(\theta)$ a distribuição a priori de θ. A distribuição a posteriori de θ pode ser escrita como

$$f(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)f(\theta) = u(\boldsymbol{x})v(r(\boldsymbol{x}), \theta)f(\theta) \propto v(r(\boldsymbol{x}), \theta)f(\theta).$$

Portanto, $f(\theta|\boldsymbol{x})$ depende de $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ somente através de $r(\boldsymbol{x})$. Como o estimador de Bayes é calculado a partir da posteriori, o estimador de Bayes também dependerá da amostra através de $r(\boldsymbol{X})$. Ou seja, o estimador de Bayes $\hat{\theta}$ é uma função de $T = r(\boldsymbol{X})$ e, se $\hat{\theta}$ é uma estatística suficiente, então será estatística suficiente minimal.

Exemplo (tempos de vida)

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$ e $\lambda \sim \mathcal{G}(a, b)$. Temos que

$$f(\boldsymbol{x}|\lambda) = \lambda^n \exp\{-\lambda n\overline{x}\} = u(\boldsymbol{x})v(r(\boldsymbol{x}), \lambda),$$

com $u(\boldsymbol{x}) = 1$ e $v(r(\boldsymbol{x}), \lambda) = \lambda^n \exp\{-\lambda n\overline{x}\}$. Logo, $r(\boldsymbol{X}) = \overline{X}$ é uma estatística suficiente minimal.

Note que a distribuição a posteriori é uma função de λ. Assim, temos que

$$f(\lambda|\boldsymbol{x}) \propto \lambda^n \exp\{-\lambda n\overline{x}\}\, \lambda^{a-1} \exp\{-b\lambda\} \propto \lambda^{(n+a)-1} \exp\{-(b + n\overline{x})\lambda\} \boldsymbol{I}_{[0,\infty)}(\lambda).$$

Então, reconhecendo a forma deste núcleo, temos que

$$f(\lambda|\boldsymbol{x}) = \frac{(b + n\overline{x})^{(n+a)}}{\Gamma(n + a)} \lambda^{(n+a)-1} \exp\{-(b + n\overline{x})\lambda\} \boldsymbol{I}_{[0,\infty)}(\lambda),$$

ou seja, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n + a;\ b + n\overline{x})$. Assim, média, mediana, moda, etc. a posteriori dependerá somente de $r(\boldsymbol{X}) = \overline{X}$. Por exemplo, $\mathrm{E}(\lambda|\boldsymbol{X}) = \dfrac{n + a}{b + n\overline{X}}$ é uma estatística suficiente minimal.

Erro quadrático médio de um estimador

Suponha que $(X_1, X_2, \ldots, X_n)$ seja uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$, com θ desconhecido e $\theta \in \Omega$.

Para cada variável $Z = g(\boldsymbol{X})$, seja $\mathrm{E}_\theta(Z) = \mathrm{E}(Z|\theta)$ o valor esperado de Z calculado com respeito à função (de densidade) de probabilidade $f(\boldsymbol{x}|\theta)$. Por exemplo, se $f(\boldsymbol{x}|\theta)$ é uma função de densidade de probabilidade, então

$$\mathrm{E}_\theta(g(\boldsymbol{X})) = \mathrm{E}(g(\boldsymbol{X})|\theta) = \int\int\cdots\int g(\boldsymbol{x}) f(\boldsymbol{x}|\theta) dx_1 dx_2 \ldots dx_n.$$

Suponha que desejamos estimar uma função $h(\theta)$. Para cada estimador $\delta = \delta(\boldsymbol{X})$ e todo o valor de $\theta \in \Omega$, denote por $R(\theta, \delta)$ **o erro quadrático médio de** δ calculado com respeito ao valor dado de θ. Assim,

$$R(\theta, \delta) = \mathrm{E}_\theta\left([\delta(\boldsymbol{X}) - h(\theta)]^2\right) = \mathrm{E}\left([\delta(\boldsymbol{X}) - h(\theta)]^2 \middle| \theta\right).$$

Desejamos encontrar um estimador $\delta = \delta(\boldsymbol{X})$ para o qual o erro quadrático médio é pequeno para tal valor de $\theta \in \Omega$ ou, pelo menos, para um grande número de valores de θ.

Esperança condicional quando uma estatística suficiente é conhecida

Seja $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$, com $T_j = r_j(\boldsymbol{X})$ para $j = 1, 2, \ldots, k$, uma estatística conjuntamente suficiente para θ. Defina o estimador $\delta_0(\boldsymbol{T})$ através da seguinte esperança condicional:

$$\delta_0(\boldsymbol{T}) = \mathrm{E}_\theta(\delta(\boldsymbol{X})|\boldsymbol{T}) = \mathrm{E}(\delta(\boldsymbol{X})|\boldsymbol{T}, \theta).$$

Como $\boldsymbol{T}$ é uma estatística suficiente, então $f(\boldsymbol{x}|\boldsymbol{T} = \boldsymbol{t}, \theta) = f(\boldsymbol{x}|\boldsymbol{T} = \boldsymbol{t})$, para todo $\theta \in \Omega$. Portanto, $\mathrm{E}_\theta(\delta(\boldsymbol{X})|\boldsymbol{T})$ será a mesma para todo θ, ou seja, $\delta_0(\boldsymbol{T})$ depende de $\boldsymbol{T}$, mas não de θ. Assim, concluímos que

$$\delta_0(\boldsymbol{T}) = \mathrm{E}(\delta(\boldsymbol{X})|\boldsymbol{T}),$$

é um estimador de θ.

Teorema

Seja $\delta = \delta(\boldsymbol{X})$ um estimador de θ e $\boldsymbol{T} = (T_1, T_2, \ldots, T_k)$, com $T_j = r_j(\boldsymbol{X})$ para $j = 1, 2, \ldots, k$, uma estatística conjuntamente suficiente para θ. Defina $\delta_0(\boldsymbol{T}) = \mathrm{E}(\delta(\boldsymbol{X})|\boldsymbol{T})$. Então, para todo $\theta \in \Omega$, temos que

$$R(\theta, \delta_0) \leqslant R(\theta, \delta).$$

Prova

Se $R(\theta, \delta)$ é infinito para um dado valor de θ, então a desigualdade é automaticamente satisfeita. Suponha que $R(\theta, \delta) < \infty$. Sabemos que

$$\mathrm{E}_\theta\left([\delta(\boldsymbol{X}) - \theta]^2\right) \geqslant [\mathrm{E}_\theta(\delta(\boldsymbol{X})) - \theta]^2,$$

pela desigualdade de Jensen para funções convexas ($g(x) = (x - \theta)^2$). Pelo mesmo argumento,

$$\mathrm{E}_\theta\left([\delta(\boldsymbol{X}) - \theta]^2 \middle| \boldsymbol{T}\right) \geqslant [\mathrm{E}_\theta(\delta(\boldsymbol{X})|\boldsymbol{T}) - \theta]^2 = (\delta_0(\boldsymbol{T}) - \theta)^2.$$

Tomando-se o valor esperado E_θ de cada lado, temos que

$$\begin{aligned} R(\theta, \delta_0) &= \mathsf{E}_\theta\left((\delta_0(\boldsymbol{T}) - \theta)^2\right) \leqslant \mathsf{E}_\theta\left(\mathsf{E}_\theta\left([\delta(\boldsymbol{X}) - \theta]^2 \middle| \boldsymbol{T}\right)\right) \\ &= \mathsf{E}_\theta\left([\delta(\boldsymbol{X}) - \theta]^2\right) = R(\theta, \delta), \end{aligned}$$

para todo valor de $\theta \in \Omega$. □

O teorema acima pode ser estendido ao caso de vetor paramétrico $\boldsymbol{\theta}$.

Um resultado similar vale se $R(\theta, \delta)$ é definido como erro absoluto médio, ou seja, $\mathsf{E}_\theta\left(\left|\delta(\boldsymbol{X}) - \theta\right|\right)$.

Estimador inadmissível

Definição

Assuma que $R(\theta, \delta) = \mathrm{E}_\theta\left([\delta(\boldsymbol{X}) - h(\theta)]^2\right)$ ou $R(\theta, \delta) = \mathrm{E}_\theta\left(|\delta(\boldsymbol{X}) - h(\theta)|\right)$. Dizemos que **um estimador $\delta(\boldsymbol{X})$ é inadmissível se existe outro estimador $\delta_0(\boldsymbol{X})$ tal que** $R(\theta, \delta_0) \leqslant R(\theta, \delta)$, para todo $\theta \in \Omega$ e na desigualdade estrita para pelo menos um vetor de θ.

Sob essas condições, dizemos que o estimador $\delta_0(\boldsymbol{X})$ domina $\delta(\boldsymbol{X})$. **Um estimador $\delta_0(\boldsymbol{x})$ é admissível se não é dominado por nenhum outro estimador.**

Então, o teorema anterior pode ser lido da seguinte forma:

- Um estimador $\delta(\boldsymbol{X})$ que não é função apenas da estatística suficiente $\boldsymbol{T}$ é inadmissível se o estimador $\delta_0(\boldsymbol{T}) = \mathrm{E}\left(\delta(\boldsymbol{X})|\boldsymbol{T}\right)$ domina $\delta(\boldsymbol{X})$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu, \sigma_0^2)$ com μ desconhecido e σ_0^2 conhecido. Mostre que a mediana amostral, $m(\boldsymbol{X})$, é um estimador inadmissível para μ.

Temos que

- $T_1 = \sum_{i=1}^{n} X_i$ é uma estatística suficiente para μ; e
- $T_2 = \overline{X}$ é uma estatística suficiente para μ.

Como T_2 é função de T_1, então $\overline{X}$ é uma estatística suficiente minimal para μ. Como $m(\boldsymbol{X})$ (mediana amostral) não é uma estatística suficiente, ela será dominada por $\overline{X}$.

Assim, de acordo com o teorema anterior, qualquer função de $\delta(\boldsymbol{X}) = m(\boldsymbol{X})$ (mediana amostral) será dominada por alguma função de $\overline{X}$. Portanto, $\delta(\boldsymbol{X}) = m(\boldsymbol{X})$ é um estimador inadmissível (para μ).

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$ com μ e σ^2 desconhecidos. Mostraremos que a amplitude amostral, $A(\boldsymbol{X}) = X_{(n)} - X_{(1)}$, é um estimador inadmissível para σ.

Temos que

- $(T_1, T_2) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$ são estatísticas conjuntamente suficientes para (μ, σ^2); e
- $(T_1', T_2') = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2\right)$ são estatísticas conjuntamente suficientes para (μ, σ^2).

Além disso,

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

é uma estatística suficiente para σ.

Assim, de acordo com o teorema anterior, $A(\boldsymbol{X}) = X_{(n)} - X_{(1)}$ é um estimador inadmissível (para σ).

Distribuição amostral de uma estatística

Definição

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com parâmetro desconhecido θ. Como a estatística $T = r(X_1, X_2, \ldots, X_n)$ é uma função de variáveis aleatórias, então T é uma variável aleatória e sua distribuição pode ser obtida a partir da distribuição conjunta de $(X_1, X_2, \ldots, X_n)$. Essa distribuição é conhecida como **distribuição amostral da estatística T.**

Como $\hat{\theta}$ é uma estatística e um estimador de θ, então é possível, em princípio, encontrar a distribuição amostral de $\hat{\theta}$.

Exemplo[3]

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu;\, \sigma_0^2)$ com μ desconhecido e σ_0^2 conhecido. Sabemos que $\hat{\mu} = \overline{X}$ é o estimador de μ pelo método dos momentos e pelo método da máxima verossimilhança. Como $\overline{X}$ é uma combinação linear de normais, temos que

$$\overline{X} \sim \mathcal{N}\left(\mu;\, \frac{\sigma_0^2}{n}\right).$$

Exemplo[1]

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(0;\, \theta)$ com θ desconhecido. Sabemos que $\hat{\theta} = X_{(n)}$ é o estimador de máxima verossimilhança de θ. Defina $W = X_{(n)}$. Então,

$$f_W(w) = n f_X(w)[F_X(w)]^{n-1} = n\frac{1}{\theta}\left[\frac{w}{\theta}\right]^{n-1} = \frac{n w^{n-1}}{\theta^n}, \quad \text{para} \quad 0 \leqslant w \leqslant \theta.$$

[3] Consulte apendice.pdf

Para que serve a distribuição amostral?

Suponha que o interesse em determinado problema de inferência esteja na distribuição a posteriori de θ. Assim, poderíamos calcular a seguinte probabilidade:

$$\Pr(|\hat{\theta} - \theta| < \varepsilon).$$

Contudo, antes de se observar a amostra, ambos $\hat{\theta}$ e θ são aleatórios e a probabilidade acima envolve a distribuição conjunta de $\hat{\theta}$ e θ.

A distribuição amostral nada mais é do que a distribuição de $\hat{\theta}$ dado θ. Pela lei da probabilidade total, temos que

$$\Pr(|\hat{\theta} - \theta| < \varepsilon) = \mathrm{E}\left(\Pr(|\hat{\theta} - \theta| < \varepsilon|\theta)\right).$$

Para que serve a distribuição amostral?

A distribuição amostral pode ser útil quando desejamos selecionar um tamanho de amostra a ser utilizado.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu;\, 4)$ com μ desconhecido e $\sigma_0^2 = 4$ (conhecido). Qual deve ser o tamanho da amostra para que $\mathrm{E}(|\overline{X} - \mu|^2) \leqslant 0,1$ para todo μ?

Sabemos que $\overline{X} \sim \mathcal{N}\left(\mu;\, \frac{4}{n}\right)$.

Note que $\mathrm{E}(|\overline{X} - \mu|^2) = \mathrm{Var}(\overline{X}) = \frac{4}{n} \leqslant 0,1$.

Portanto, a condição será satisfeita se $n \geqslant 40$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu; 4)$ com μ desconhecido e $\sigma_0^2 = 4$ (conhecido). Qual deve ser o tamanho da amostra para que $\mathrm{E}(|\overline{X} - \mu|) \leqslant 0,1$ para todo μ?

Defina a variável aleatória padronizada $Z = \sqrt{n}(\overline{X} - \mu)/2$ e note que

$$\mathrm{E}(|\overline{X} - \mu|) = \frac{2}{\sqrt{n}}\mathrm{E}(|Z|) \leqslant 0,1.$$

Assim,

$$\begin{aligned}
\mathrm{E}(|Z|) &= \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} z \exp\left(-\frac{z^2}{2}\right) dz \\
&= \sqrt{\frac{2}{\pi}} \left[-\exp\left(-\frac{z^2}{2}\right) \Big|_0^{\infty} \right] = \sqrt{\frac{2}{\pi}}, \quad \text{e} \\
\mathrm{E}(|\overline{X} - \mu|) &= \frac{2}{\sqrt{n}} \times \sqrt{\frac{2}{\pi}} = 2\sqrt{\frac{2}{n\pi}} \leqslant 0,1.
\end{aligned}$$

Portanto, a condição será satisfeita se $n \geqslant 255$.

Distribuição qui-quadrado

Seja X uma variável aleatória com distribuição qui-quadrada com (parâmetro) n graus de liberdade. Então, a função de densidade de probabilidade é dada por

$$f_X(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\{-x/2\} I_{[0,\infty]}(x).$$

$$\begin{aligned} \mathrm{E}(X) &= n \\ \mathrm{Var}(X) &= 2n \\ \psi(s) &= (1-2s)^{-n/2} \\ \varphi(t) &= (1-2it)^{-n/2}. \end{aligned}$$

Denotaremos a distribuição qui-quadrada com n graus de liberdade por χ^2_n.

Note que $\chi^2_n \equiv \mathcal{G}(n/2;\ 1/2)$ para n inteiro positivo.

Teorema

Se $X_1, X_2, \ldots, X_K$ são variáveis aleatórias independentes tal que $X_j \sim \chi^2_{n_j}$, para $j = 1, 2, \ldots, K$, e $W = X_1 + X_2 + \cdots + X_K$, então $W \sim \chi^2_m$ em que $m = n_1 + n_2 + \cdots + n_K$.

Prova

Poderíamos provar por indução utilizando transformadas de variáveis aleatórias, mas utilizaremos função característica. Segue-se que

$$\begin{aligned}
\varphi_W(t) &= \mathrm{E}_W(\exp\{itW\}) = \mathrm{E}_{X_1,X_2,\ldots,X_K}(\exp\{it(X_1 + X_2 + \cdots + X_K)\}) \\
&= \mathrm{E}_{X_1}(\exp\{itX_1\})\mathrm{E}_{X_2}(\exp\{itX_2\}) \ldots \mathrm{E}_{X_K}(\exp\{itX_K\}) \\
&= \varphi_{X_1}(t)\varphi_{X_2}(t) \ldots \varphi_{X_K}(t) \\
&= (1-2it)^{-n_1/2}(1-2it)^{-n_2/2} \cdots (1-2it)^{-n_K/2} \\
&= (1-2it)^{-m/2} = \left(\frac{1/2}{1/2 - it}\right)^{m/2},
\end{aligned}$$

em que $m = n_1 + n_2 + \cdots + n_K$.

Portanto, $W \sim \mathcal{G}(m/2;\ 1/2) \equiv \chi^2_m$.

Teorema

Se $Z \sim \mathcal{N}(0;\, 1)$ e $Y = Z^2$, então $Y \sim \chi^2_1$.

Prova

$$\begin{aligned} F_Y(y) &= \Pr(Y \leqslant y) = \Pr(Z^2 \leqslant y) = \Pr(-\sqrt{y} \leqslant Z \leqslant \sqrt{y}) \\ &= \Pr(Z \leqslant \sqrt{y}) - \Pr(Z \leqslant -\sqrt{y}) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y}). \end{aligned}$$

Logo,

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} = \frac{1}{2\sqrt{y}} f_Z(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_Z(-\sqrt{y}) \\ &= \frac{y^{-1/2}}{2}(2\pi)^{-1/2} \exp\left\{-\frac{(\sqrt{y})^2}{2}\right\} + \frac{y^{-1/2}}{2}(2\pi)^{-1/2} \exp\left\{-\frac{(-\sqrt{y})^2}{2}\right\} \\ &= \frac{(1/2)^{1/2}}{\Gamma(1/2)} y^{1/2-1} \exp\left\{-\frac{y}{2}\right\} \boldsymbol{I}_{[0,\infty)}(y). \end{aligned}$$

Portanto, $Y \sim \mathcal{G}(1/2;\ 1/2) \equiv \chi^2_1$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\sigma^2) \sim \mathcal{N}(\mu_0;\ \sigma^2)$ com μ_0 conhecido e σ^2 desconhecido. Sabemos que $M^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2$ é o estimador de máxima verossimilhança de σ^2.

Defina $Z_i = (X_i - \mu_0)/\sigma$, para $i = 1, 2, \ldots, n$. Então, $(Z_1, Z_2, \ldots, Z_n)$ é uma amostra aleatória de $Z_i \sim \mathcal{N}(0;\ 1)$ e, consequentemente,

$$W^2 = \frac{nM^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n \equiv \mathcal{G}(n/2;\ 1/2).$$

É fácil mostrar que $M^2 = \dfrac{\sigma^2 W^2}{n} \sim \mathcal{G}\left(\dfrac{n}{2};\ \dfrac{n}{2\sigma^2}\right)$ (veremos a seguir) .

Evitamos o uso da notação $\hat{\sigma}^2$ como estimador de máxima verossimilhança de σ^2 para facilitar o desenvolvimento dos resultados.

Distribuição de M^2

Sejam $Y = \dfrac{nM^2}{\sigma^2} \sim \mathcal{G}\left(\dfrac{n}{2};\ \dfrac{1}{2}\right)$ e $V = \dfrac{\sigma^2 Y}{n} = M^2$.

Então,

$$
\begin{aligned}
F_V(v) &= \Pr(V \leqslant v) = \Pr\left(\frac{\sigma^2 Y}{n} \leqslant v\right) = \Pr\left(Y \leqslant \frac{nv}{\sigma^2}\right) = F_Y\left(\frac{nv}{\sigma^2}\right). \\
f_V(v) &= \frac{dF_V(v)}{dv} = f_Y\left(\frac{nv}{\sigma^2}\right)\frac{n}{\sigma^2} \\
&= \frac{n}{\sigma^2} \times \frac{(1/2)^{n/2}}{\Gamma[n/2]}\left[\frac{nv}{\sigma^2}\right]^{n/2-1} \exp\left\{-\frac{nv}{2\sigma^2}\right\} \\
&= \frac{[n/(2\sigma^2)]^{n/2}}{\Gamma[n/2]} v^{n/2-1} \exp\left\{-\frac{nv}{2\sigma^2}\right\}.
\end{aligned}
$$

Portanto, $V \sim \mathcal{G}\left(\dfrac{n}{2};\ \dfrac{n}{2\sigma^2}\right)$, isto é, $M^2 \sim \mathcal{G}\left(\dfrac{n}{2};\ \dfrac{n}{2\sigma^2}\right)$.

Distribuição conjunta da média amostral e da variância amostral

Teorema

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\ \sigma^2)$. Então,

$$\overline{X} \sim \mathcal{N}\left(\mu;\ \frac{\sigma^2}{n}\right) \quad \text{e} \quad \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1},$$

e são variáveis aleatórias independentes.

Note que as distribuições são condicionais ao conhecimento dos parâmetros, (μ, σ^2), embora os mesmos sejam desconhecidos.

A seguir, alguns conceitos para conseguirmos provar este teorema.

Matrizes ortogonais

Definição

Uma matriz $\boldsymbol{A}$, de dimensão $n \times n$, é ortogonal se $\boldsymbol{A}^{-1} = \boldsymbol{A}^{\top}$, em que $\boldsymbol{A}^{\top}$ é a matriz transposta de $\boldsymbol{A}$.

Em outras palavras, temos que

- uma matriz $\boldsymbol{A}$ é ortogonal se, e somente se, $\boldsymbol{A}\boldsymbol{A}^{\top} = \boldsymbol{A}^{\top}\boldsymbol{A} = \boldsymbol{I}$, em que $\boldsymbol{I}$ é a matriz identidade de dimensão $n \times n$; e
- uma matriz $\boldsymbol{A}$ é ortogonal se, e somente se, a soma dos quadrados dos elementos de cada coluna é igual a 1, e a soma dos produtos dos elementos correspondentes em quaisquer colunas diferentes é igual a 0. O mesmo vale para as linhas de $\boldsymbol{A}$.

Propriedades de matrizes ortogonais

Teorema

Se $\boldsymbol{A}$ é uma matriz ortogonal, então $|\det \boldsymbol{A}| = 1$.

Prova

Para qualquer matriz quadrada $\boldsymbol{A}$, temos que $\det \boldsymbol{A} = \det \boldsymbol{A}^\top$. Além disso, se $\boldsymbol{B}$ é uma outra matriz quadrada, então $\det \boldsymbol{AB} = (\det \boldsymbol{A})(\det \boldsymbol{B})$. Assim, $\det \boldsymbol{AA}^\top = (\det \boldsymbol{A})(\det \boldsymbol{A}^\top) = (\det \boldsymbol{A})^2$. Por outro lado, se $\boldsymbol{A}$ é uma matriz ortogonal, então $\boldsymbol{AA}^\top = \boldsymbol{I}$. Segue-se que $\det \boldsymbol{AA}^\top = \det \boldsymbol{I} = 1$. Portanto, $(\det \boldsymbol{A})^2 = 1$, ou equivalentemente, $|\det \boldsymbol{A}| = 1$.

□

Propriedades de matrizes ortogonais

Teorema

Considere dois vetores aleatórios de tamanho n,

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \text{e} \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix},$$

e suponha que $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, em que $\boldsymbol{A}$ é uma matriz ortogonal. Então,

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i^2.$$

Prova

Como $\boldsymbol{A}$ é uma matriz ortogonal, temos que

$$\sum_{i=1}^{n} Y_i^2 = \boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{X}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^\top \boldsymbol{X} = \sum_{i=1}^{n} X_i^2.$$

□

Essas duas propriedades implicam que se um vetor aleatório $\boldsymbol{Y}$ é obtido a partir de uma transformação ortogonal de um outro vetor aleatório $\boldsymbol{X}$, $\boldsymbol{Y} = \boldsymbol{AX}$, então o valor absoluto do jacobiano da transformação é 1 e $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$.

Teorema

Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes tal que $X_i \sim \mathcal{N}(0; 1)$, para $i = 1, 2, \ldots, n$. Sejam $\boldsymbol{A}$ uma matriz ortogonal de dimensão $n \times n$ e $\boldsymbol{Y} = \boldsymbol{AX}$. Então, as variáveis $Y_1, Y_2, \ldots, Y_n$ também são independentes tal que $Y_i \sim \mathcal{N}(0; 1)$, para $i = 1, 2, \ldots, n$, e $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$.

Prova

Temos que

$$f(\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum\nolimits_{i=1}^{n} x_i^2\right\},$$

para $x_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$. Ao aplicar a transformada linear, obtemos que

$$g(\boldsymbol{y}) = \frac{1}{|\det \boldsymbol{A}|} f(\boldsymbol{A}^{-1}\boldsymbol{y}).$$

Seja $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$. Como $\boldsymbol{A}$ é ortogonal, então $|\det \boldsymbol{A}| = 1$ (jacobiano) e $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$. Portanto,

$$g(\boldsymbol{y}) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum\nolimits_{i=1}^{n} y_i^2\right\}.$$

□

Prova (do Teorema da pag. 116)

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $X \sim \mathcal{N}(0;\, 1)$ e defina o vetor linha de tamanho n por

$$\boldsymbol{u} = \left[\frac{1}{\sqrt{n}} \;\; \frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}}\right].$$

Como $\sum_{i=1}^{n} u_i^2 = 1$, então é possível construir uma matriz ortogonal $\boldsymbol{A}$ de modo que $\boldsymbol{u}$ seja a primeira linha de $\boldsymbol{A}$ (método de Gram-Schmidt). Assumindo que $\boldsymbol{A}$ tenha sido construída, definimos $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ tal que $\boldsymbol{Y} = \boldsymbol{AX}$. Assim,

$$Y_1 = \boldsymbol{uX} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} X_i = \sqrt{n}\,\overline{X}. \tag{1}$$

Pelo teorema anterior, sabemos que $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} X_i^2$. Portanto,

$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

ou seja,

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

Como $(Y_1, Y_2, \ldots, Y_n)$ são independentes, então Y_1 e $\sum_{i=2}^{n} Y_i^2$ também são independentes. Segue-se das Equações (1) e (2) que $\overline{X}$ e $\sum_{i=1}^{n}(X_i - \overline{X})^2$ são independentes.

Pelo teorema anterior, $(Y_1, Y_2, \ldots, Y_n)$ são independentes e $Y_i \sim \mathcal{N}(0; 1)$, para $i = 1, 2, \ldots, n$. Logo, $\sum_{i=2}^{n} Y_i^2 \sim \chi^2_{n-1}$ e isto implica que $\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2_{n-1}$.

Agora, seja $(X_i|\mu, \sigma^2) \sim \mathcal{N}(\mu; \sigma^2)$ e defina $Z_i = \dfrac{X_i - \mu}{\sigma}$, para $i = 1, 2, \ldots, n$. Sabemos que $\overline{Z} = (\overline{X} - \mu)/\sigma$ e $\sum_{i=1}^{n}(Z_i - \overline{Z})^2 = \left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]/\sigma^2$ são independentes com $\overline{Z} \sim \mathcal{N}(0; 1/n)$ e $\sum_{i=1}^{n}(Z_i - \overline{Z})^2 \sim \chi^2_{n-1}$.

Logo, $\overline{X}$ e $\dfrac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$ são independentes e

$$\overline{X} \sim \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right) \quad \textbf{e} \quad \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$.
Sabemos que a função característica de X é dada por

$$\varphi_X(t) = \exp\left\{it\mu - \frac{\sigma^2 t^2}{2}\right\}.$$

Agora, considere $\overline{X}$ como estimador de μ. Então,

$$\begin{aligned}
\varphi_{\overline{X}}(t) &= \mathrm{E}_{\overline{X}}(\exp\{it\overline{X}\}) = \mathrm{E}_{X_1,X_2,\ldots,X_n}(\exp\{it(X_1 + X_2 + \cdots + X_n)/n\}) \\
&= \mathrm{E}_{X_1}(\exp\{i(t/n)X_1\})\mathrm{E}_{X_2}(\exp\{i(t/n)X_2\}) \cdots \mathrm{E}_{X_n}(\exp\{i(t/n)X_n\}) \\
&= \varphi_{X_1}(t/n)\varphi_{X_2}(t/n) \cdots \varphi_{X_n}(t/n) \\
&= \prod_{i=1}^{n} \exp\left\{i(t/n)\mu - \frac{\sigma^2(t/n)^2}{2}\right\} = \exp\left\{it\mu - \frac{(\sigma^2/n)t^2}{2}\right\}.
\end{aligned}$$

Portanto, $(\overline{X}|\mu, \sigma^2) \sim \mathcal{N}\left(\mu;\, \frac{\sigma^2}{n}\right)$.

Note que $\mathrm{E}(\overline{X}|\mu, \sigma^2) = \mu$ e $\mathrm{Var}(\overline{X}|\mu, \sigma^2) = \frac{\sigma^2}{n}$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Agora, considere o estimador

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

para o parâmetro σ^2.

Vimos que $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$. Portanto, temos que

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

Segue-se que

$$\mathrm{E}\left(\frac{(n-1)S^2}{\sigma^2}\right) = n-1 \quad \Rightarrow \quad \mathrm{E}(S^2|\sigma^2) = \sigma^2.$$

$$\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1) \quad \Rightarrow \quad \mathrm{Var}(S^2|\sigma^2) = \frac{2\sigma^4}{n-1}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Agora, considere o estimador de máxima verossimilhança dado por

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

para o parâmetro σ^2 (com μ desconhecido).

Vimos que $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$. Portanto, temos que

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Segue-se que

$$\mathrm{E}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = n-1 \quad \Rightarrow \quad \mathrm{E}(\hat{\sigma}^2|\sigma^2) = \frac{(n-1)\sigma^2}{n}.$$

$$\mathrm{Var}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = 2(n-1) \quad \Rightarrow \quad \mathrm{Var}(\hat{\sigma}^2|\sigma^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

Distribuição de S^2

Agora, sejam $Y = \dfrac{(n-1)S^2}{\sigma^2} \sim \mathcal{G}\left(\dfrac{n-1}{2};\ \dfrac{1}{2}\right)$ e $V = \dfrac{\sigma^2 Y}{n-1} = S^2$.

Então,

$$
\begin{aligned}
F_V(v) &= \Pr(V \leqslant v) = \Pr\left(\frac{\sigma^2 Y}{n-1} \leqslant v\right) = \Pr\left(Y \leqslant \frac{(n-1)v}{\sigma^2}\right) \\
&= F_Y\left(\frac{(n-1)v}{\sigma^2}\right). \\
f_V(v) &= \frac{dF_V(v)}{dv} = f_Y\left(\frac{(n-1)v}{\sigma^2}\right)\frac{(n-1)}{\sigma^2} \\
&= \frac{(n-1)}{\sigma^2}\frac{(1/2)^{(n-1)/2}}{\Gamma([n-1]/2)}\left[\frac{(n-1)v}{\sigma^2}\right]^{(n-1)/2-1}\exp\left\{-\frac{(n-1)v}{2\sigma^2}\right\} \\
&= \frac{[(n-1)/(2\sigma^2)]^{(n-1)/2}}{\Gamma([n-1]/2)}v^{(n-1)/2-1}\exp\left\{-\frac{(n-1)v}{2\sigma^2}\right\}.
\end{aligned}
$$

Portanto, $V \sim \mathcal{G}\left(\dfrac{n-1}{2};\ \dfrac{n-1}{2\sigma^2}\right)$.

Estimação da média e do desvio padrão

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\ \sigma^2)$ com

$$(\hat{\mu}, \hat{\sigma}) = \left(\overline{X}, \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \right),$$

o estimador de máxima verossimilhança de (μ, σ).

Qual é o menor tamanho de amostra n para que

$$\Pr\left(|\hat{\mu} - \mu| < \frac{\sigma}{5};\ |\hat{\sigma} - \sigma| < \frac{\sigma}{5} \right) \geqslant \frac{1}{2} \ ?$$

Temos que $\hat{\mu}$ e $\hat{\sigma}$ são independentes. Logo,

$$\Pr\left(|\hat{\mu}-\mu|<\frac{\sigma}{5};\ |\hat{\sigma}-\sigma|<\frac{\sigma}{5}\right)=\Pr\left(|\hat{\mu}-\mu|<\frac{\sigma}{5}\right)\Pr\left(|\hat{\sigma}-\sigma|<\frac{\sigma}{5}\right)\geqslant\frac{1}{2}.$$

Segue-se que

$$\begin{aligned}\Pr\left(|\hat{\mu}-\mu|<\frac{\sigma}{5}\right) &= \Pr\left(\frac{|\hat{\mu}-\mu|}{\sigma}<\frac{1}{5}\right)=\Pr\left(\frac{\sqrt{n}|\hat{\mu}-\mu|}{\sigma}<\frac{\sqrt{n}}{5}\right)\\ &= \Pr\left(|Z|<\frac{\sqrt{n}}{5}\right)=2\Phi\left(\frac{\sqrt{n}}{5}\right)-1=p_1,\end{aligned}$$

em que Φ é a f.d.a. da normal padrão, e

$$\begin{aligned}\Pr\left(|\hat{\sigma}-\sigma|<\frac{\sigma}{5}\right) &= \Pr\left(-\frac{\sigma}{5}<\hat{\sigma}-\sigma<\frac{\sigma}{5}\right)=\Pr\left(\frac{4\sigma}{5}<\hat{\sigma}<\frac{6\sigma}{5}\right)\\ &= \Pr\left(\frac{16n}{25}\leqslant\frac{n\hat{\sigma}^2}{\sigma^2}\leqslant\frac{36n}{25}\right)\\ &= \Pr\left(0,64n\leqslant\mathcal{X}\leqslant 1,44n\right)=p_2,\end{aligned}$$

em que $\mathcal{X}$ é uma variável aleatória tal que $\mathcal{X}\sim\chi^2_{n-1}$. Então, queremos encontrar o menor valor de n tal que $p_1\times p_2\geqslant 0,5$.

Tabela: Valores de p_1, p_2 e $p_1 \times p_2$ para diferentes valores de n.

n	p_1	p_2	$p_1 \times p_2$
2	0,2227026	0,1682130	0,03746147
3	0,2709655	0,2675678	0,07250164
4	0,3108435	0,3406564	0,10589083
5	0,3452792	0,3992418	0,13784988
6	0,3757939	0,4483605	0,16849112
7	0,4032988	0,4906854	0,19789284
8	0,4283924	0,5278313	0,22611891
9	0,4514938	0,5608617	0,25322554
10	0,4729107	0,5905216	0,27926402
11	0,4928775	0,6173583	0,30428204
12	0,5115777	0,6417880	0,32832441
13	0,5291583	0,6641368	0,35143349
14	0,5457398	0,6846659	0,37364942
15	0,5614220	0,7035891	0,39501036
16	0,5762892	0,7210835	0,41555264
17	0,5904133	0,7372985	0,43531085
18	0,6038561	0,7523613	0,45431798
19	0,6166715	0,7663813	0,47260550
20	0,6289066	0,7794535	0,49020345
21	**0,6406032**	**0,7916608**	**0,50714049**
22	0,6517983	0,8030766	0,52344400
23	0,6625250	0,8137657	0,53914014
24	0,6728131	0,8237858	0,55425387
25	0,6826895	0,8331886	0,56880908
26	0,6921785	0,8420206	0,58282858
27	0,7013024	0,8503238	0,59633418
28	0,7100815	0,8581362	0,60934671
29	0,7185345	0,8654924	0,62188612
30	0,7266783	0,8724238	0,63397146

Ver exemplo_07.r

Distribuição *t*-Student

Seja X uma variável aleatória com distribuição *t*-Student com parâmetros localização (posição) μ, de escala σ^2 e ν graus de liberdade. Então, a função de densidade de probabilidade é dada por

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{1}{2}\right)\nu^{1/2}\sigma}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R},$$

em que $\mu \in \mathbb{R}$, $\nu > 0$ e $\sigma^2 > 0$.

$$\begin{aligned}
\text{Mediana}(X) &= \mu \\
\text{Moda}(X) &= \mu \\
\text{E}(X) &= \mu, \qquad \text{se} \quad \nu > 1 \\
\text{Var}(X) &= \frac{\nu}{\nu-2}\sigma^2, \qquad \text{se} \quad \nu > 2.
\end{aligned}$$

Denotaremos a distribuição *t*-Student por $t_\nu(\mu, \sigma^2)$.

Teorema

Se Z e W são variáveis aleatórias independentes tal que $Z \sim \mathcal{N}(0;\, 1)$, $W \sim \chi^2_n$ e

$$T = \frac{Z}{\sqrt{W/n}},$$

então $T \sim t_n(0;\, 1)$.

Prova

$$\begin{aligned} F_T(t) &= \Pr(T \leqslant t) = \Pr(Z \leqslant t\sqrt{W/n}) \\ &= \int_0^\infty \Pr(Z \leqslant t\sqrt{w/n} \,\Big|\, W = w) f_W(w)\, dw \\ &= \int_0^\infty \Phi(t\sqrt{w/n}) f_W(w)\, dw. \end{aligned}$$

$$
\begin{aligned}
f_T(t) &= \frac{\partial F_T(t)}{\partial t} = \frac{\partial}{\partial t}\int_0^\infty \Phi(t\sqrt{w/n}) f_W(w) dw \\
&= \int_0^\infty \frac{\partial}{\partial t}\Phi(t\sqrt{w/n}) f_W(w) dw = \int_0^\infty \phi(t\sqrt{w/n})\sqrt{\frac{w}{n}} f_W(w) dw \\
&= \frac{1}{\Gamma(n/2)\Gamma(1/2)2^{(n+1)/2} n^{1/2}} \int_0^\infty w^{(n+1)/2-1} \exp\left\{-w\frac{1}{2}\left(1+\frac{t^2}{n}\right)\right\} dw \\
&= \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2) n^{1/2}} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad t \in \mathbb{R}.
\end{aligned}
$$

Portanto, $T \sim t_n(0; 1)$.

Teorema

Suponha que $T_n \sim t_n(0; 1)$ seja uma sequência de variáveis aleatórias.

Então, $T_n \xrightarrow{D} Z$ quando $n \to \infty$ em que $Z \sim \mathcal{N}(0; 1)$.

Prova

$$\begin{aligned}
\lim_{n\to\infty} f_T(t) &= \lim_{n\to\infty} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2)n^{1/2}} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \\
&= \frac{1}{2^{1/2}\Gamma(1/2)} \lim_{m\to\infty} \frac{\Gamma(m+1/2)}{\Gamma(m)m^{1/2}} \left(1+\frac{t^2}{2m}\right)^{-m} \left(1+\frac{t^2}{2m}\right)^{-1/2} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\}, \qquad \text{pois}
\end{aligned}$$

$$\lim_{m\to\infty} \frac{\Gamma(m+1/2)}{\Gamma(m)m^{1/2}} = 1 \quad \text{(Fórmula de Stirling)}$$

$$\lim_{m\to\infty} \left(1+\frac{t^2}{2m}\right)^{-m} = \exp\left\{-\frac{t^2}{2}\right\} \quad \left(\lim_{m\to\infty}(1+f(x)/m)^{-m} = \exp\{-f(x)\}\right)$$

$$\lim_{m\to\infty} \left(1+\frac{t^2}{2m}\right)^{-1/2} = 1.$$

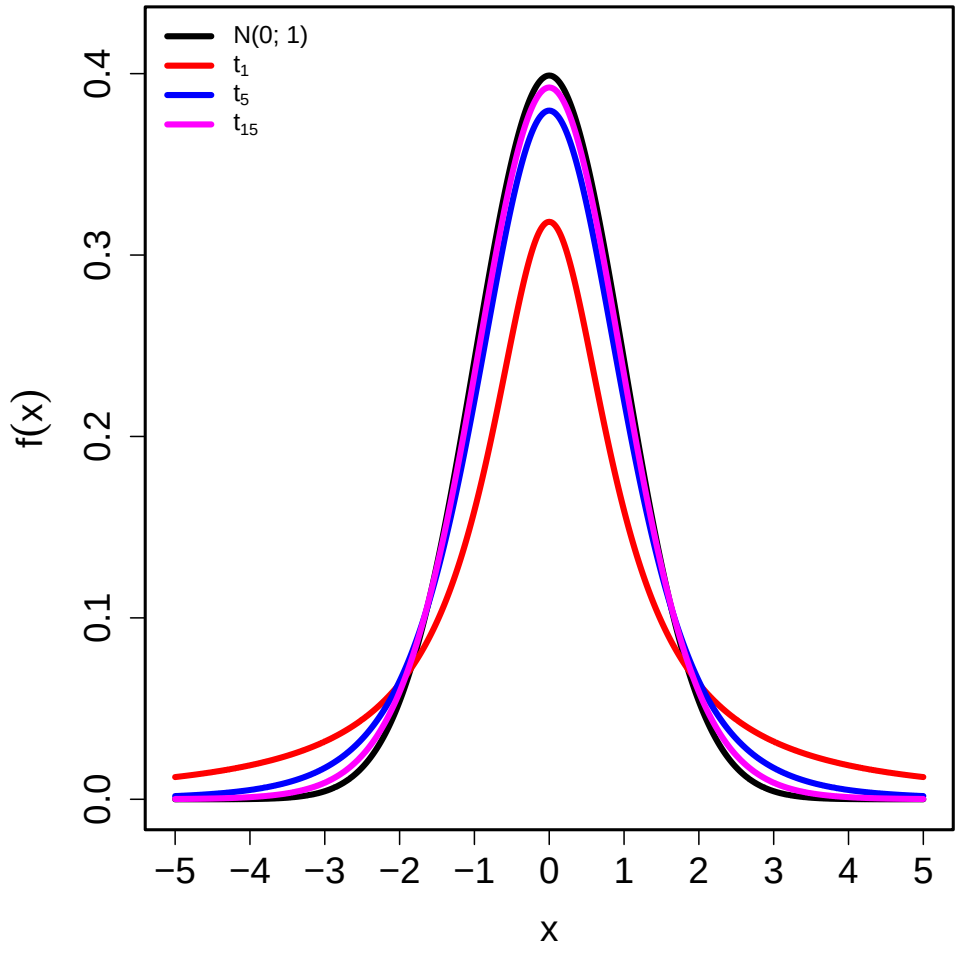


Figura: Distribuição *t*-Student com diferentes graus de liberdade e a normal padrão.

Ver exemplo_08.r

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Agora, sabemos que

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim \mathcal{N}(0;\, 1) \quad \text{e} \quad Y = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1},$$

com Z e Y independentes.

Assim,

$$T = \frac{Z}{\sqrt{\dfrac{Y}{n-1}}} \sim t_{n-1}(0;\, 1).$$

Outras formas:

$$T = \frac{\dfrac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\sqrt{\dfrac{n\hat{\sigma}^2}{(n-1)\sigma^2}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{\dfrac{n\hat{\sigma}^2}{n-1}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}} \sim t_{n-1}(0;\, 1).$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\ \sigma^2)$. Considere $(\overline{X};\ S^2)$ como estimador de (μ, σ^2). Sabemos que

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim \mathcal{N}(0;\ 1) \quad \text{e} \quad Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

com Z e Y independentes. Novamente, temos que

$$T = \frac{Z}{\sqrt{\dfrac{Y}{n-1}}} \sim t_{n-1}(0;\ 1).$$

Por outro lado, segue-se que

$$T = \frac{\dfrac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\sqrt{\dfrac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}(0;\ 1).$$

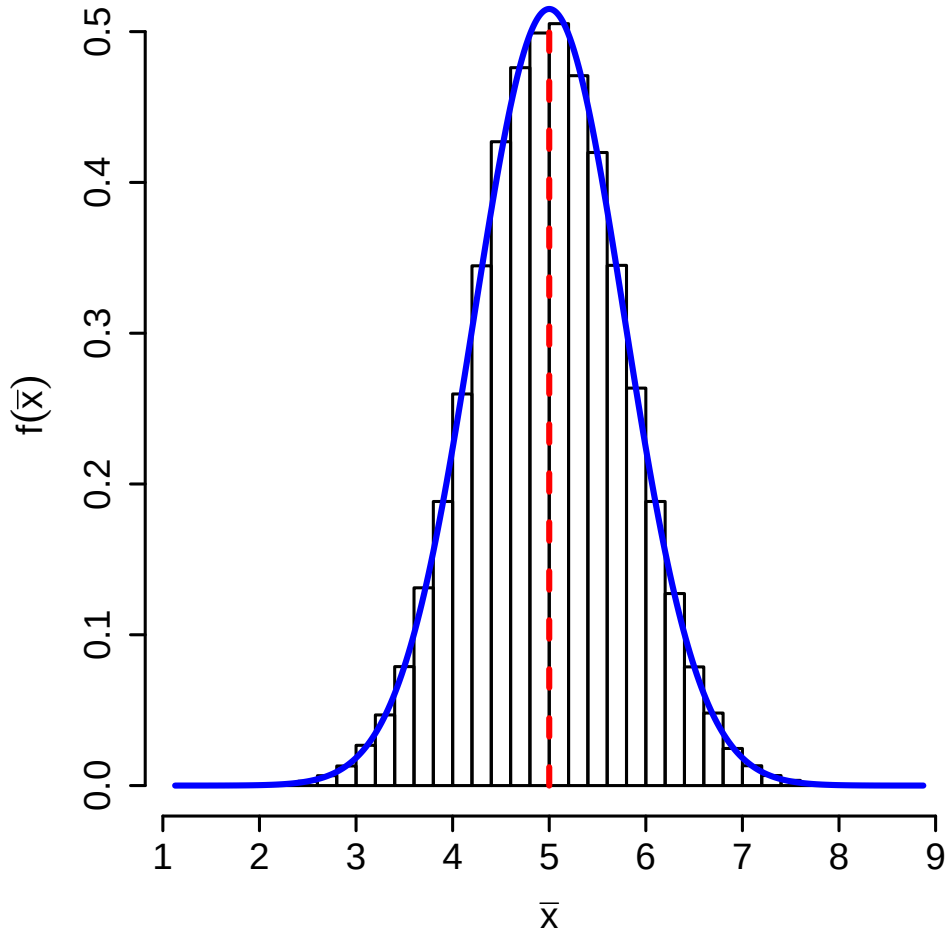


Figura: Distribuição amostral de $\overline{X}$ dado (μ, σ^2): $\overline{X} \sim \mathcal{N}\left(\mu;\ \frac{\sigma^2}{n}\right)$ para $n = 15$.

Ver exemplo_09.r

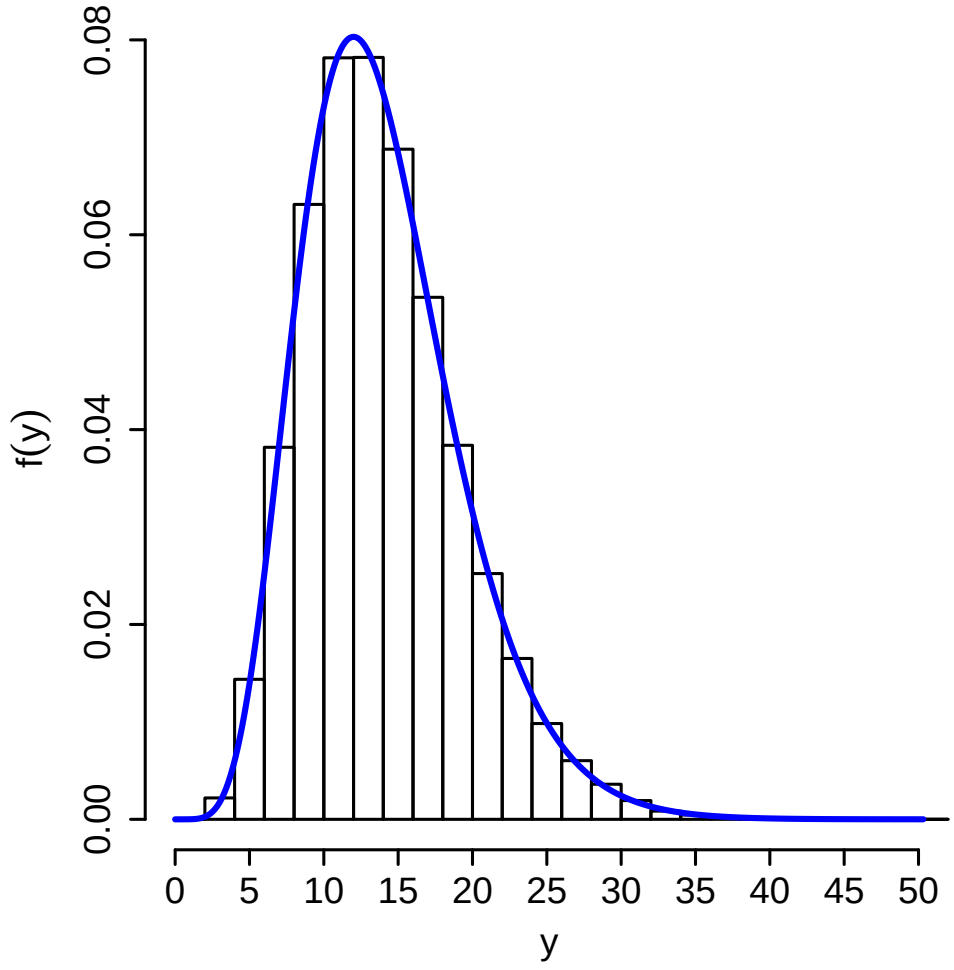


Figura: Distribuição amostral de $Y = \dfrac{(n-1)S^2}{\sigma^2}$ dado σ^2: $Y = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ para $n = 15$.

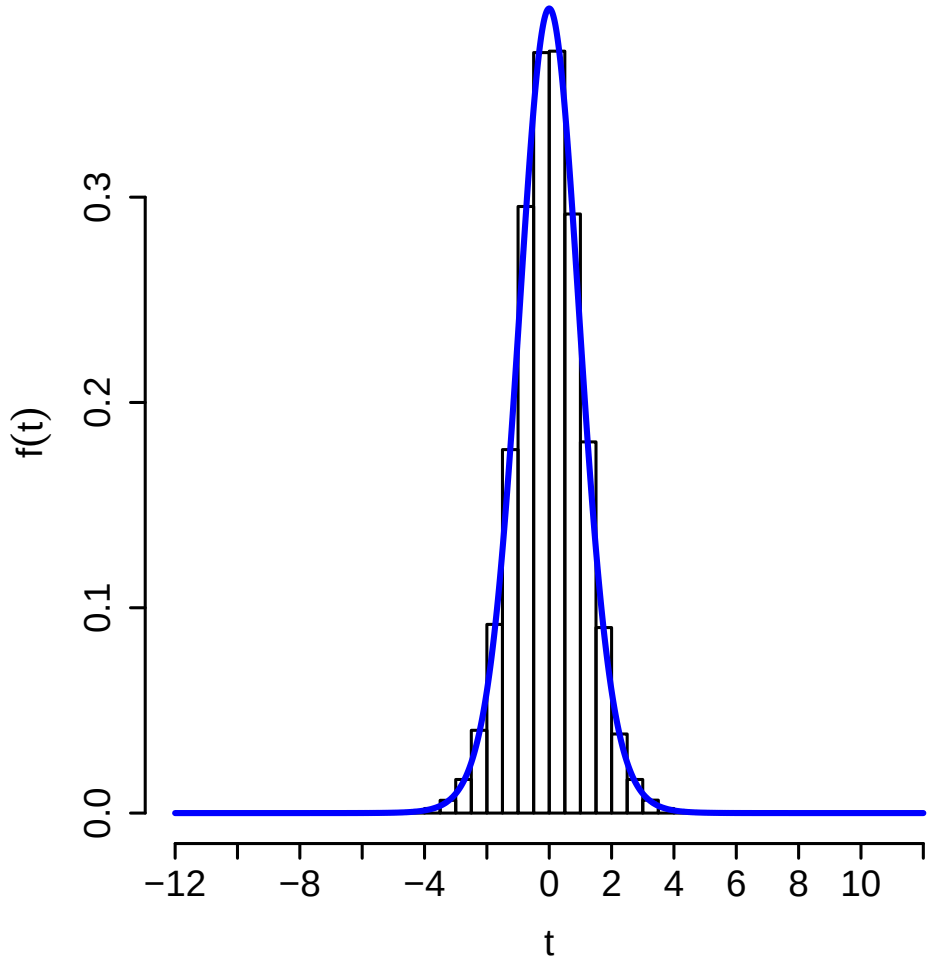


Figura: Distribuição amostral de $\overline{X}$ com σ^2 desconhecido: $\dfrac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}(0;\, 1)$ para $n = 15$.

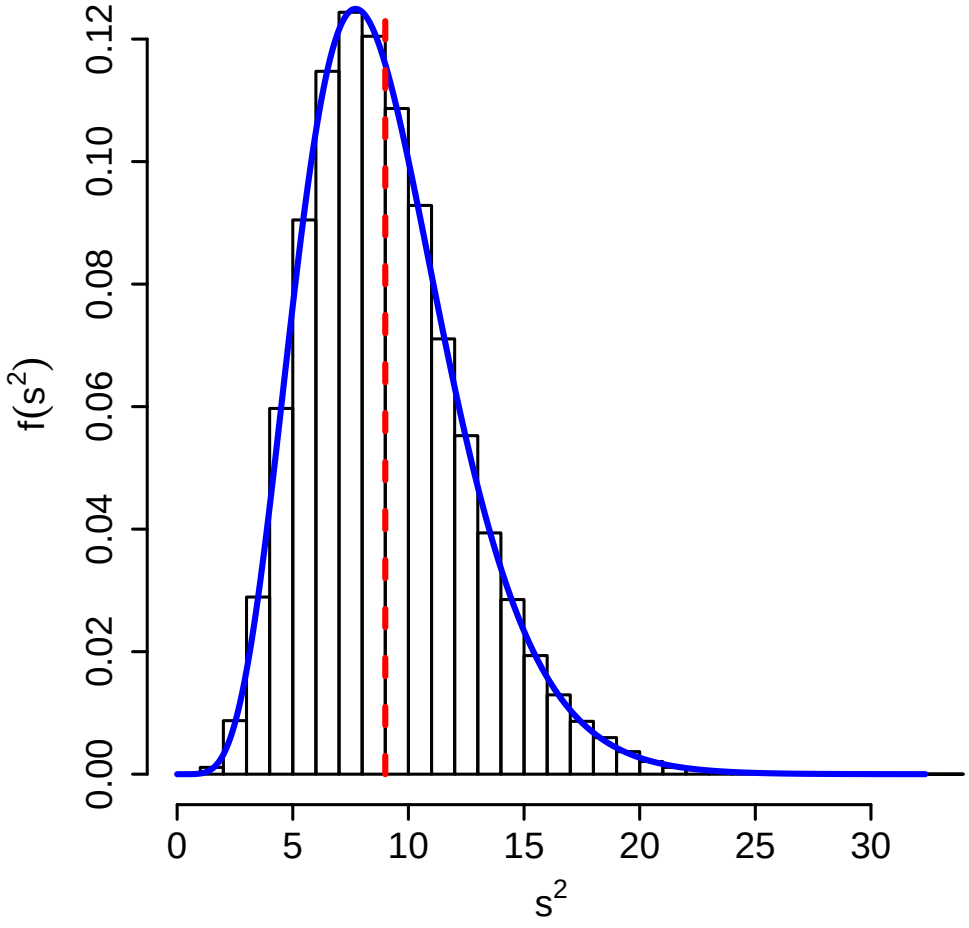


Figura: Distribuição amostral de S^2 dado σ^2: $S^2 \sim \mathcal{G}\left(\frac{n-1}{2};\ \frac{n-1}{2\sigma^2}\right)$ para $n = 15$.

Intervalos de Confiança

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Sabemos que

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t_{n-1}(0;\, 1).$$

Assim, podemos determinar um valor c tal que para um dado γ, com $0 < \gamma < 1$, que

$$\Pr(-c < T < c) = \gamma.$$

Note que

$$\begin{aligned}
\gamma &= \Pr(-c < T < c) = \Pr\left(-c < \frac{\sqrt{n}(\overline{X} - \mu)}{S} < c\right) \\
&= \Pr\left(-\frac{cS}{\sqrt{n}} < \overline{X} - \mu < \frac{cS}{\sqrt{n}}\right) = \Pr\left(-\overline{X} - \frac{cS}{\sqrt{n}} < -\mu < -\overline{X} + \frac{cS}{\sqrt{n}}\right) \\
&= \Pr\left(\overline{X} - \frac{cS}{\sqrt{n}} < \mu < \overline{X} + \frac{cS}{\sqrt{n}}\right).
\end{aligned}$$

Note que $\overline{X}$ e S são variáveis aleatórias enquanto μ é fixo.

Definição

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição que depende de um parâmetro θ (ou vetor paramétrico $\boldsymbol{\theta}$). Seja $g(\theta)$ uma função real de θ. Sejam $A \leqslant B$ duas estatísticas que possuem a propriedade de que para todo θ,

$$\Pr(A < g(\theta) < B) \geqslant \gamma.$$

Então, o intervalo aleatório (A, B) é chamado de **intervalo de confiança de nível γ para $g(\theta)$** ou **intervalo de confiança de $100\gamma\%$ para $g(\theta)$**.

Após observar $\boldsymbol{X} = \boldsymbol{x}$, os valores $A = a(\boldsymbol{x})$ e $B = b(\boldsymbol{x})$ são calculados e o intervalo (a, b) é chamado de **valor observado do intervalo de confiança.**

Teorema

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Para cada $0 < \gamma < 1$, o intervalo (A, B) para a média μ, com variância σ^2 desconhecida, é dado por

$$A = \overline{X} - T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{S}{\sqrt{n}} \qquad \text{e} \qquad B = \overline{X} + T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)\frac{S}{\sqrt{n}}.$$

Note que

$$\begin{aligned} \gamma &= \Pr(-c < T < c) = T_{n-1}(c) - T_{n-1}(-c) = 2T_{n-1}(c) - 1 \\ &\Rightarrow T_{n-1}(c) = \frac{1+\gamma}{2} \quad \Leftrightarrow \quad c = T_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right), \end{aligned}$$

em que T_{n-1} representa a função de distribuição acumulada de uma variável aleatória com distribuição t-Student com $n-1$ graus de liberdade e T_{n-1}^{-1} sua inversa (o percentil da distribuição).

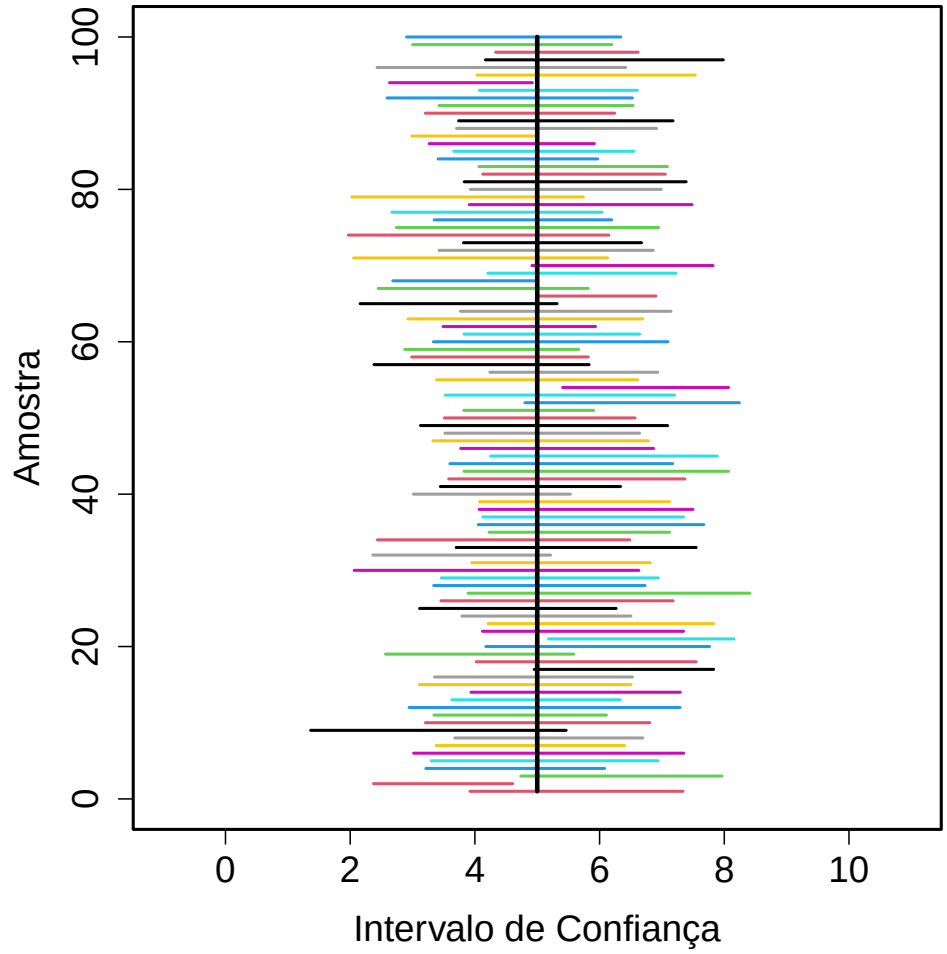


Figura: Interpretação de intervalos de confiança com $n = 15$ (valores observados).

Ver exemplo_10.r

Para um dado nível de confiança γ, é possível construir vários intervalos de confiança para μ.

Por exemplo, sejam $0 < \gamma_1 < \gamma_2 < 1$ tal que $\gamma_2 - \gamma_1 = \gamma$. Então,

$$\gamma = \Pr\left(T_{n-1}^{-1}(\gamma_1) < T < T_{n-1}^{-1}(\gamma_2)\right),$$

e

$$A = \overline{X} - T_{n-1}^{-1}(\gamma_1)\frac{S}{\sqrt{n}} \qquad \text{e} \qquad B = \overline{X} + T_{n-1}^{-1}(\gamma_2)\frac{S}{\sqrt{n}}.$$

Entre todos os intervalos de confiança de $100\gamma\%$, o intervalo simétrico com $\gamma_1 = 1 - \gamma_2$ é o de menor comprimento.

Intervalos de Confiança Unilaterais

Definição

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição que depende de um parâmetro θ (ou vetor paramétrico $\boldsymbol{\theta}$). Seja $g(\theta)$ uma função real de θ. Seja A uma estatística tal que para todo θ,

$$\Pr(A < g(\theta)) \geqslant \gamma.$$

Então, o intervalo aleatório (A, ∞) é chamado de **intervalo de confiança unilateral (inferior) de nível γ para $g(\theta)$**. Similarmente, se B é uma estatística tal que

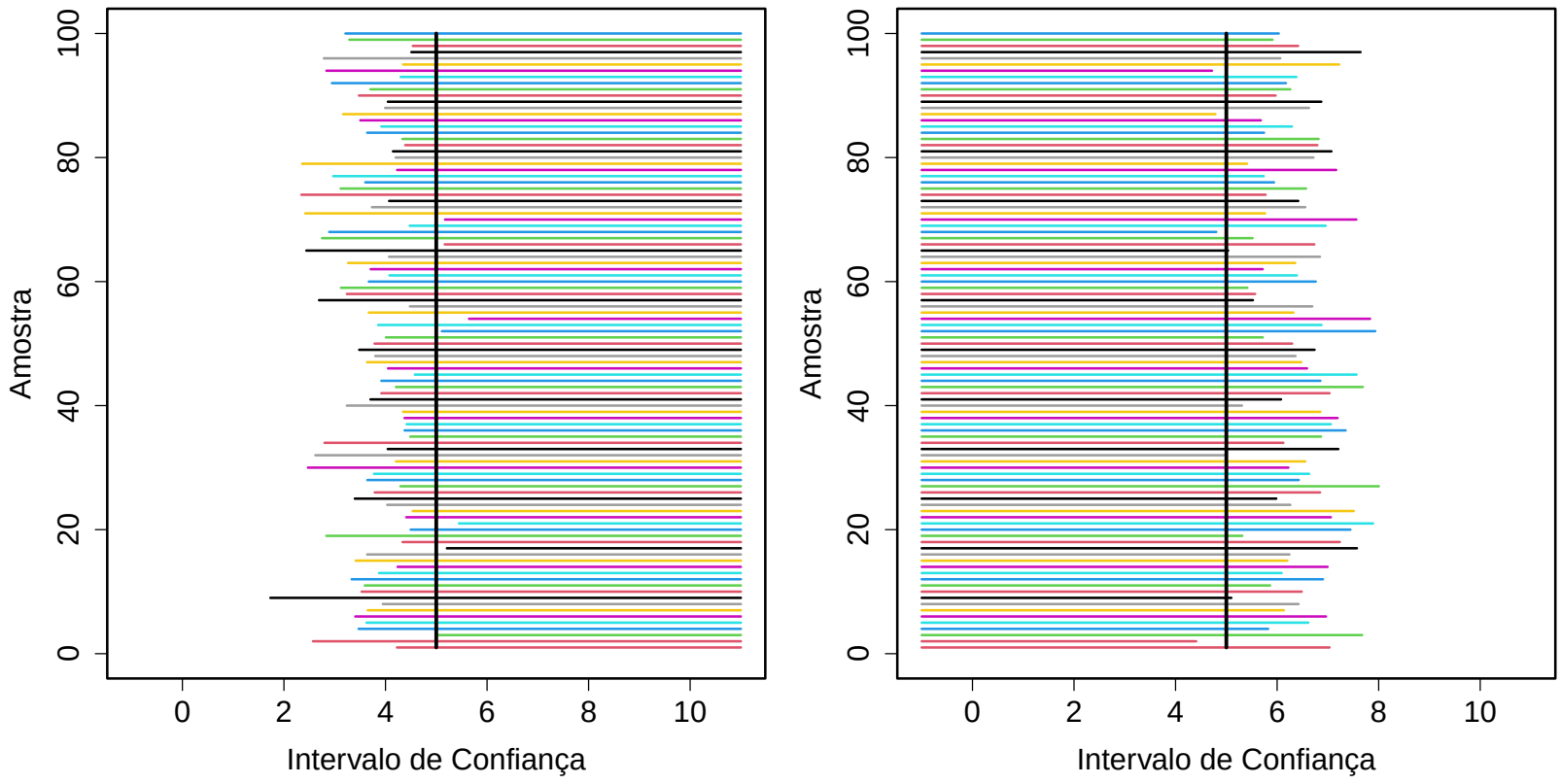
$$\Pr(g(\theta) < B) \geqslant \gamma,$$

então o intervalo aleatório $(-\infty, B)$ é chamado de **intervalo de confiança unilateral (superior) de nível γ para $g(\theta)$**.

Teorema

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Para todo $0 < \gamma < 1$, as seguintes estatísticas são, respectivamente, os limites inferior e superior dos intervalos de confiança unilaterais de nível γ para μ:

$$\begin{aligned} A &= \overline{X} - T_{n-1}^{-1}(\gamma)\,\frac{S}{\sqrt{n}} \\ B &= \overline{X} + T_{n-1}^{-1}(\gamma)\,\frac{S}{\sqrt{n}}. \end{aligned}$$



(a) IC inferior

(b) IC superior

Figura: Interpretação de intervalos de confiança inferior e superior com $n = 15$ (valores observados).

Ver exemplo_11.r

Definição

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição que depende de um parâmetro θ (ou vetor paramétrico $\boldsymbol{\theta}$). Seja $V(\boldsymbol{X}, \theta)$ uma variável aleatória cuja distribuição é a mesma para todo θ. Então, V é chamada de **quantidade pivotal**.

Para utilizar uma quantidade pivotal para construir um intervalo de confiança para $g(\theta)$, é necessário inverter esta quantidade pivotal. Isto é, necessita-se de uma função $r(v, \boldsymbol{x})$ tal que

$$r(V(\boldsymbol{X}, \theta), \boldsymbol{X}) = g(\theta).$$

Se uma função assim existe, então ela pode ser usada para construir um intervalo de confiança.

Teorema

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição que depende de um parâmetro θ (ou vetor paramétrico $\boldsymbol{\theta}$). Suponha que exista uma quantidade pivotal $V(\boldsymbol{X}, \theta)$. Seja G a função de distribuição acumulada de $V(\boldsymbol{X}, \theta)$ e assuma que G é contínua. Assuma também que existe uma função $r(v, \boldsymbol{x})$ tal que $r(V(\boldsymbol{X}, \theta), \boldsymbol{X}) = g(\theta)$ e que $r(v, \boldsymbol{x})$ é estritamente crescente em v para cada $\boldsymbol{x}$. Sejam $0 < \gamma < 1$ e $0 < \gamma_1 < \gamma_2 < 1$ tal que $\gamma = \gamma_2 - \gamma_1$. Então, as estatísticas

$$A = r(G^{-1}(\gamma_1), \boldsymbol{X}) \qquad \text{e} \qquad B = r(G^{-1}(\gamma_2), \boldsymbol{X})$$

são os limites do intervalo de confiança de nível γ para $g(\theta)$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$.
Sabemos que

$$V(\boldsymbol{X}, \sigma^2) = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Então,

$$\begin{aligned}
\gamma &= \Pr\left(c_1 < V(\boldsymbol{X}, \sigma^2) < c_2\right) = \Pr\left(c_1 < \frac{(n-1)S^2}{\sigma^2} < c_2\right) \\
&= \Pr\left(\frac{c_1}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{c_2}{(n-1)S^2}\right) \\
&= \Pr\left(\frac{(n-1)S^2}{c_2} < \sigma^2 < \frac{(n-1)S^2}{c_1}\right),
\end{aligned}$$

com c_1 e c_2 obtidos através de

$$\Pr(\chi^2_{n-1} < c_1) = \gamma_1 \qquad \text{e} \qquad \Pr(\chi^2_{n-1} < c_2) = \gamma_2.$$

Usualmente, escolhe-se $\gamma_1 = 1 - \gamma_2$ embora talvez não ressulte no intervalo de menor comprimento.

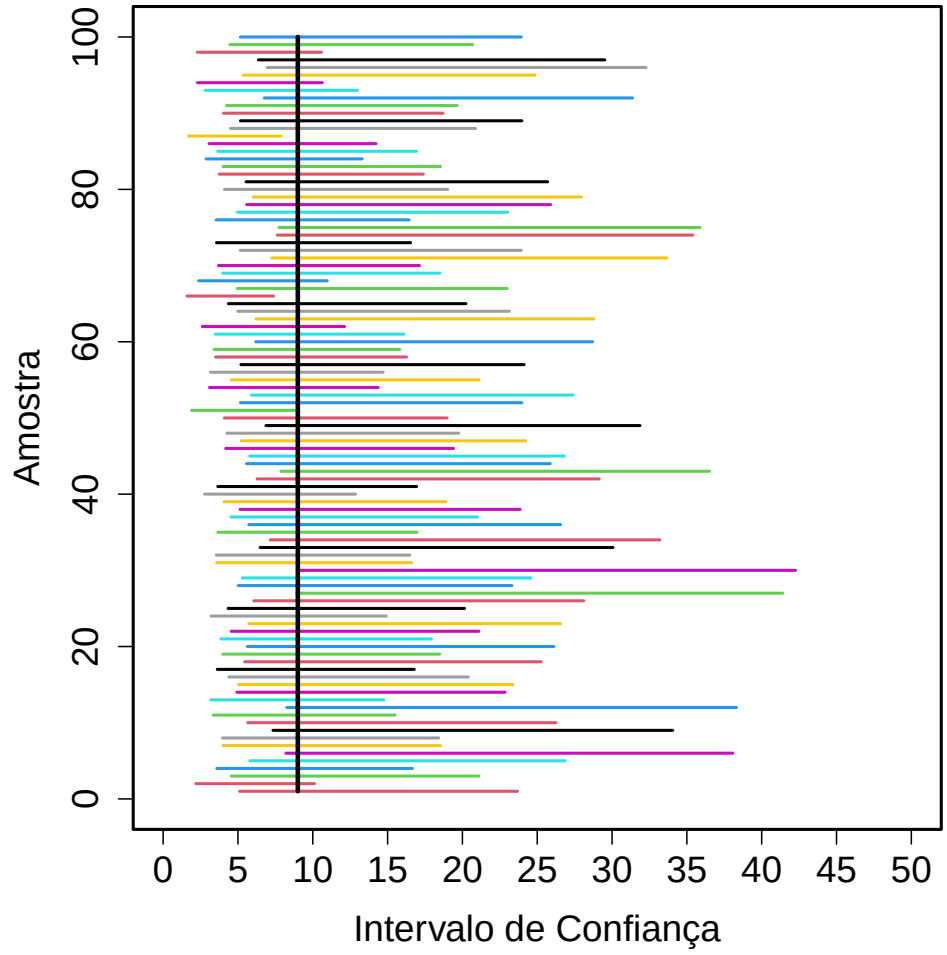


Figura: Interpretação de intervalos de confiança com $n = 15$ (valores observados).

Ver exemplo_12.r

Exemplo

Suponha que duas observações X_1 e X_2 são obtidas de $\mathcal{U}(\theta - 1/2,\ \theta + 1/2)$ com $\theta \in \mathbb{R}$ (desconhecido). Sejam $X_{(1)} = \min(X_1, X_2)$ e $X_{(2)} = \max(X_1, X_2)$.

$$\begin{aligned} \Pr(X_{(1)} < \theta < X_{(2)}) &= \Pr(X_1 < \theta < X_2) + \Pr(X_2 < \theta < X_1) \\ &= \Pr(X_1 < \theta)\Pr(X_2 > \theta) + \Pr(X_2 < \theta)\Pr(X_1 > \theta) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

Portanto, se observarmos $X_{(1)} = x_{(1)}$ e $X_{(2)} = x_{(2)}$, então o intervalo $(x_{(1)}, x_{(2)})$ será um intervalo de confiança de nível 0,5 para θ.

Note que as duas observações X_1 e X_2 devem ser maiores que $\theta - 1/2$ e menores que $\theta + 1/2$. Sabemos com certeza que $X_{(1)} > \theta - 1/2$ e $X_{(2)} < \theta + 1/2$. Em outras palavras, sabemos com certeza que

$$X_{(2)} - 1/2 < \theta < X_{(1)} + 1/2. \tag{3}$$

Agora, suponha que se observou $(X_{(2)} - X_{(1)}) \geqslant 1/2$.

Então, $X_{(1)} \leqslant X_{(2)} - 1/2$ e da Equação (3), temos que $X_{(1)} < \theta$.

Além disso, como $X_{(1)} + 1/2 \leqslant X_{(2)}$, e da Equação (3), temos que $\theta < X_{(2)}$.

Assim, se $X_{(2)} - X_{(1)} \geqslant 1/2$, então $X_{(1)} < \theta < X_{(2)}$.

Mesmo quando $(X_{(2)} - X_{(1)}) \leqslant 1/2$, quanto mais próximo o valor de $(X_{(2)} - X_{(1)})$ estiver de 1/2, mais certo nós estaremos que o intervalo $(X_{(1)}, X_{(2)})$ inclui θ.

Estimadores não tendenciosos

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com parâmetro θ. Se desejamos estimar $g(\theta)$, esperamos que, com alta probabilidade, o estimador $\delta(\boldsymbol{X})$ esteja próximo de $g(\theta)$.

Definição

Um estimador $\delta(\boldsymbol{X})$ é não tendencioso para $g(\theta)$ se

$$\mathbf{E}_{\boldsymbol{X}}(\delta(\boldsymbol{X})|\theta) = g(\theta).$$

Se $\mathbf{E}_{\boldsymbol{X}}(\delta(\boldsymbol{X})|\theta) \neq g(\theta)$, então dizemos que $\delta(\boldsymbol{X})$ é um estimador tendencioso de $g(\theta)$.

Define-se a tendenciosidade (viés) de um estimador $\delta(\boldsymbol{X})$ como

$$B(\delta(\boldsymbol{X})) = \mathbf{E}_{\boldsymbol{X}}(\delta(\boldsymbol{X})|\theta) - g(\theta).$$

Observação: um estimador é não tendencioso se, e somente se, seu viés é igual a zero.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\ \sigma^2)$.
Sabemos que a função característica de X_i é dada por

$$\varphi_{X_i}(t) = \exp\left\{it\mu - \frac{\sigma^2 t^2}{2}\right\}.$$

Agora, considere $\overline{X}$ como estimador de μ. Então,

$$\begin{aligned}
\varphi_{\overline{X}}(t) &= \mathrm{E}_{\overline{X}}(\exp\{it\overline{X}\}) = \mathrm{E}_{X_1,x_2,\ldots,x_n}(\exp\{it(X_1 + X_2 + \cdots + X_n)/n\}) \\
&= \mathrm{E}_{X_1}(\exp\{i(t/n)X_1\})\mathrm{E}_{X_2}(\exp\{i(t/n)X_2\}) \cdots \mathrm{E}_{X_n}(\exp\{i(t/n)X_n\}) \\
&= \varphi_{X_1}(t/n)\varphi_{X_2}(t/n) \cdots \varphi_{X_n}(t/n) \\
&= \prod_{i=1}^{n} \exp\left\{i(t/n)\mu - \frac{\sigma^2(t/n)^2}{2}\right\} = \exp\left\{it\mu - \frac{(\sigma^2/n)t^2}{2}\right\}.
\end{aligned}$$

Portanto, $(\overline{X}|\mu, \sigma^2) \sim \mathcal{N}\left(\mu;\ \frac{\sigma^2}{n}\right)$. Note que $\mathrm{E}(\overline{X}|\mu, \sigma^2) = \mu$ e

$\mathrm{Var}(\overline{X}|\mu, \sigma^2) = \frac{\sigma^2}{n}$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Agora, considere o estimador

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

para o parâmetro σ^2.

Vimos que $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$. Portanto, temos que

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

Segue-se que

$$\mathrm{E}\left(\frac{(n-1)S^2}{\sigma^2}\right) = n-1 \quad \Rightarrow \quad \mathrm{E}(S^2|\sigma^2) = \sigma^2.$$

$$\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1) \quad \Rightarrow \quad \mathrm{Var}(S^2|\sigma^2) = \frac{2\sigma^4}{n-1}.$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Agora, considere o estimador de máxima verossimilhança dado por

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

para o parâmetro σ^2. Vimos que $\dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$. Portanto, temos que $\dfrac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$. Segue-se que

$$\mathrm{E}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = n-1 \quad \Rightarrow \quad \mathrm{E}(\hat{\sigma}^2|\sigma^2) = \frac{(n-1)\sigma^2}{n}.$$

$$\mathrm{Var}\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = 2(n-1) \quad \Rightarrow \quad \mathrm{Var}(\hat{\sigma}^2|\sigma^2) = \frac{2(n-1)\sigma^4}{n^2}.$$

Note que

$$B(\hat{\sigma}^2) = \mathrm{E}(\hat{\sigma}^2|\sigma^2) - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n}.$$

Erro quadrático médio

Uma maneira utilizada para medir a variabilidade de um estimador é o erro quadrático médio:

$$\text{EQM}(\delta(\boldsymbol{X})|\theta) = \text{E}([\delta(\boldsymbol{X}) - g(\theta)]^2|\theta).$$

Se $\delta(\boldsymbol{X})$ é um estimador não tendencioso de $g(\theta)$, então

$$\text{EQM}(\delta(\boldsymbol{X})|\theta) = \text{Var}(\delta(\boldsymbol{X})|\theta).$$

Corolário

$$\text{EQM}(\delta(\boldsymbol{X})|\theta) = \text{Var}(\delta(\boldsymbol{X})|\theta) + [\text{B}(\delta(\boldsymbol{X}))]^2 .$$

Prova

$$\begin{aligned}
\text{EQM}(\delta(\boldsymbol{X})|\theta) &= \text{E}([\delta(\boldsymbol{X}) - g(\theta)]^2|\theta) \\
&= \text{E}([\delta(\boldsymbol{X}) - \text{E}(\delta(\boldsymbol{X})|\theta) + \text{E}(\delta(\boldsymbol{X})|\theta) - g(\theta)]^2|\theta) \\
&= \text{E}([\delta(\boldsymbol{X}) - \text{E}(\delta(\boldsymbol{X})|\theta)]^2|\theta) + [\text{E}(\delta(\boldsymbol{X})|\theta) - g(\theta)]^2 \\
&= \text{Var}(\delta(\boldsymbol{X})|\theta) + [\text{B}(\delta(\boldsymbol{X}))]^2 .
\end{aligned}$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Vimos que

$$\mathrm{E}(S^2|\sigma^2) = \sigma^2 \quad \text{e} \quad \mathrm{Var}(S^2|\sigma^2) = \frac{2\sigma^4}{n-1}$$

$$\mathrm{E}(\hat{\sigma}^2|\sigma^2) = \frac{(n-1)\sigma^2}{n}, \quad \mathrm{Var}(\hat{\sigma}^2|\sigma^2) = \frac{2(n-1)\sigma^4}{n^2} \quad \text{e} \quad B(\hat{\sigma}^2) = -\frac{\sigma^2}{n}.$$

Portanto, o erro quadrático médio dos estimadores é dada por

$$\begin{aligned}
\mathrm{EQM}(S^2|\sigma^2) &= \mathrm{Var}(S^2|\sigma^2) = \frac{2\sigma^4}{n-1} \\
\mathrm{EQM}(\hat{\sigma}^2|\sigma^2) &= \mathrm{Var}(\hat{\sigma}^2|\sigma^2) + \left[B(\hat{\sigma}^2)\right]^2 = \frac{2(n-1)\sigma^4}{n^2} + \left[-\frac{\sigma^2}{n}\right]^2 \\
&= \frac{(2n-1)\sigma^4}{n^2}.
\end{aligned}$$

Podemos mostrar que EQM$(S^2|\sigma^2) >$ EQM$(\hat{\sigma}^2|\sigma^2)$ para todo $n \geqslant 2$.

Exemplo (Estimadores não tendenciosos)

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$.
Sabemos que

$$\begin{aligned} \widehat{\pi}_{MM} &= \overline{X}, \quad \text{(estimador pelo método dos momentos)} \\ \widehat{\pi}_{MV} &= \overline{X}, \quad \text{(estimador pelo método da máxima verossimilhança).} \end{aligned}$$

Assim,

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

Portanto, $\overline{X}$ é um estimador não tendencioso de π.

Note que $\mathrm{Var}(X_i) = \pi(1-\pi)$ e $\mathrm{Var}(\overline{X}) = \dfrac{\pi(1-\pi)}{n}$. Logo,

$$\mathrm{EQM}(\overline{X}) = \mathrm{Var}(\overline{X}) = \frac{\pi(1-\pi)}{n}.$$

Finalmente, concluímos que $\overline{X}$ é um estimador consistente de π.

Exemplo (Estimadores não tendenciosos)

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|m, \pi) \sim \mathcal{BN}(m;\, \pi)$ com m conhecido. Sabemos que

$$\widehat{\pi}_{MM} = \frac{\overline{X}}{m} \quad \text{e} \quad \widehat{\pi}_{MV} = \frac{\overline{X}}{m}.$$

Assim, $\mathrm{E}\left(\dfrac{\overline{X}}{m}\right) = \dfrac{1}{mn}\sum_{i=1}^{n} \mathrm{E}(X_i) = \dfrac{1}{mn}\sum_{i=1}^{n} m\pi = \pi.$

Portanto, $\dfrac{\overline{X}}{m}$ é um estimador não tendencioso de π.

Note que $\mathrm{Var}(X_i) = m\pi(1-\pi)$ e $\mathrm{Var}\left(\dfrac{\overline{X}}{m}\right) = \dfrac{\pi(1-\pi)}{mn}$. Logo,

$$\mathrm{EQM}\left(\frac{\overline{X}}{m}\right) = \mathrm{Var}\left(\frac{\overline{X}}{m}\right) = \frac{\pi(1-\pi)}{mn}.$$

Finalmente, concluímos que $\dfrac{\overline{X}}{m}$ é um estimador consistente de π.

Exemplo (Estimadores não tendenciosos)

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{GM}(\pi)$. Sabemos que

$$\widehat{\pi}_{MM} = \frac{1}{\overline{X}} \quad \text{e} \quad \widehat{\pi}_{MV} = \frac{1}{\overline{X}}.$$

Contudo,

$$\mathrm{E}(\overline{X}) = \frac{1}{\pi} \quad \Rightarrow \quad \mathrm{E}\left(\frac{1}{\overline{X}}\right) \neq \pi,$$

pois $\mathrm{E}(g(Y)) \neq g(\mathrm{E}(Y))$ para $g(y) = 1/y$. Portanto, $\dfrac{1}{\overline{X}}$ é um estimador **tendencioso** de π.

Pela lei forte dos grandes números $\overline{X} \to \dfrac{1}{\pi}$ (em probabilidade quando $n \to \infty$) e pelo teorema de Slutsky $\dfrac{1}{\overline{X}} \to \pi$ (em probabilidade quando $n \to \infty$). Concluímos que $\dfrac{1}{\overline{X}}$ é um estimador consistente de π.

Exemplo (Estimadores não tendenciosos)

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(0;\, \theta)$ com $\theta > 0$. Sabemos que

$$\widehat{\theta}_{MM} = 2\overline{X} \quad \text{e} \quad \widehat{\theta}_{MV} = X_{(n)}.$$

Agora, $\mathrm{E}(2\overline{X}) = 2\dfrac{\theta}{2} = \theta$. Portanto, $2\overline{X}$ é um estimador **não tendencioso** de θ e também **consistente** para θ, mas é um estimador **inadmissível**.

Por outro lado, para $W = X_{(n)}$, temos que

$$f(w) = \frac{n w^{n-1}}{\theta^n} \boldsymbol{I}_{[0,\theta]}(w) \quad \Rightarrow \quad \mathrm{E}(W) = \mathrm{E}(X_{(n)}) = \frac{n\theta}{n+1}.$$

Então, $X_{(n)}$ é um estimador **tendencioso** de θ com viés dado por $B(X_{(n)}) = -\dfrac{\theta}{n+1}$ (subestima o valor de θ).

Contudo, $X_{(n)}$ é um estimador **consistente** de θ (vide lista de exercícios).

Estimador não tendencioso da variância

Teorema

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com parâmetro θ tal que $\text{Var}(X_i) = g(\theta)$. Então,

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

é um estimador **não tendencioso** de $g(\theta)$.

Prova

Sejam $\text{E}(X_i) = \mu$ e $\text{Var}(X_i) = \sigma^2 = g(\theta)$. Então,

$$\begin{aligned}
\text{E}(S^2) &= \frac{1}{n-1}\text{E}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n-1}\text{E}\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\text{E}(X_i^2) - n\text{E}(\overline{X}^2)\right] \\
&= \frac{1}{n-1}\left[n(\mu^2+\sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\
&= \frac{n-1}{n-1}\sigma^2 = \sigma^2 = g(\theta).
\end{aligned}$$

Se $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ é uma amostra aleatória de uma distribuição específica, como por exemplo a Poisson, então é possível e desejável considerar outro estimador não tendencioso de $g(\theta)$.

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$.

Sabemos que $\mathrm{E}(X_i) = \mathrm{Var}(X_i) = \lambda$.

Então, tanto $\overline{X}$ quanto S^2 são estimadores **não tendenciosos** de λ.

Assim, qualquer combinação

$$\delta(\boldsymbol{X}) = \alpha\overline{X} + (1-\alpha)S^2, \quad \text{com} \quad 0 \leqslant \alpha \leqslant 1,$$

é **não tendencioso** para λ, pois

$$\mathrm{E}(\delta(\boldsymbol{X})) = \alpha\mathrm{E}(\overline{X}) + (1-\alpha)\mathrm{E}(S^2) = \alpha\lambda + (1-\alpha)\lambda = \lambda.$$

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\beta) \sim \mathcal{E}(1/\beta)$.

Sabemos que $\mathrm{E}(X_i) = \beta$ e $\mathrm{Var}(X_i) = \beta^2$.

Então, $\overline{X}$ é um estimador **não tendenciosos** de β.

Temos que S^2 é um estimador **não tendenciosos** de β^2, mas S é um estimador **tendencioso** de β.

Contudo, S é um estimador consistente de β.

Na escolha de um estimador é importante levar em consideração se o estimador é não tendencioso, seu erro quadrático médio e se é um estimador consistente.

Informação de Fisher

- Mede a quantidade de informação contida em uma amostra a respeito de um parâmetro desconhecido.
- Esta medida tem a propriedade intuitiva que:
  - mais dados fornecem mais informação; e
  - dados mais precisos fornecem mais informação.
- A medida de informação pode ser utilizada para encontrar limites da variância do estimador.

Informação de Fisher de uma variável aleatória

Definição

Seja X uma variável aleatória com função (de densidade) de probabilidade $f(x|\theta)$, $\theta \in \Omega$, um intervalo aberto da reta. Assuma que **os valores de x que satisfazem $f(x|\theta) > 0$ são os mesmos para todo θ** e que $\ln f(x|\theta)$ é duas vezes diferenciável como função de θ. A **Informação de Fisher** é definida como

$$I(\theta) = \mathrm{E}([\ell'(\theta|X)]^2|\theta) = \int [\ell'(\theta|x)]^2 f(x|\theta) dx,$$

em que $\ell(\theta|x) = \ln f(x|\theta)$.

Se X é variável aleatória discreta, a integral é substituída por um somatório.

Note que $\ell'(\theta|x) = \dfrac{d}{d\theta} \ln f(x|\theta)$.

Teorema

Assuma as condições da definição anterior e que as duas derivadas de $\int f(x|\theta)dx$ com respeito a θ podem ser calculadas ao reverter a ordem de integração e diferenciação. Então, a **informação de Fisher** pode ser reescrita como

$$I(\theta) = \mathrm{E}(-\ell''(\theta|X)|\theta),$$

com $\ell''(\theta|x) = \dfrac{d^2}{d\theta^2} \ln f(x|\theta)$.

Uma forma alternativa da informação de Fisher é $I(\theta) = \mathbf{Var}(\ell'(\theta|X)|\theta)$.

Prova

Temos que

$$\int f(x|\theta)dx = 1 \;\Rightarrow\; \frac{d}{d\theta}\int f(x|\theta)dx = \int \frac{d}{d\theta} f(x|\theta)dx = \int f'(x|\theta)dx = 0,$$

e, consequentemente,

$$\frac{d^2}{d\theta^2}\int f(x|\theta)dx = 0 \;\Rightarrow\; \int f''(x|\theta)dx = 0.$$

Por outro lado,

$$\ell'(\theta|x) = \frac{d}{d\theta}\ln f(x|\theta) = \frac{f'(x|\theta)}{f(x|\theta)}.$$

Logo,

$$\mathrm{E}(\ell'(\theta|X)|\theta) = \int \ell'(\theta|x) f(x|\theta)dx = \int f'(x|\theta)dx = 0.$$

Dado que $\mathrm{E}(\ell'(\theta|X)|\theta) = 0$, temos que

$$I(\theta) = \mathrm{E}([\ell'(\theta|X)]^2|\theta) = \mathrm{Var}(\ell'(\theta|X)|\theta).$$

Segue-se que

$$\begin{aligned}
\ell''(\theta|x) &= \frac{d^2}{d\theta^2} \ln f(x|\theta) = \frac{d}{d\theta}\left[\frac{f'(x|\theta)}{f(x|\theta)}\right] \\
&= \frac{f''(x|\theta)f(x|\theta) - [f'(x|\theta)]^2}{[f(x|\theta)]^2} \\
&= \frac{f''(x|\theta)}{f(x|\theta)} - [\ell'(\theta|x)]^2.
\end{aligned}$$

Portanto,

$$\mathrm{E}(\ell''(\theta|X)|\theta) = \int f''(x|\theta)dx - I(\theta) \quad \Rightarrow \quad I(\theta) = \mathrm{E}(-\ell''(\theta|X)|\theta),$$

pois $\int f''(x|\theta)dx = 0$. □

Exemplo

Suponha que $(X|\pi) \sim \mathcal{BR}(\pi)$. Temos

$$
\begin{aligned}
f(x|\pi) &= \pi^x(1-\pi)^{1-x} \\
\ell(\pi|x) &= x\ln\pi + (1-x)\ln(1-\pi) \\
\frac{d\ell(\pi|x)}{d\pi} &= \frac{x}{\pi} - \frac{(1-x)}{1-\pi} \\
\mathrm{E}\left(\frac{d\ell(\pi|X)}{d\pi}\right) &= \frac{\mathrm{E}(X)}{\pi} - \frac{(1-\mathrm{E}(X))}{(1-\pi)} = 0 \\
\frac{d^2\ell(\pi|x)}{d\pi^2} &= -\frac{x}{\pi^2} - \frac{(1-x)}{(1-\pi)^2} \\
\mathrm{E}\left(-\frac{d^2\ell(\pi|X)}{d\pi^2}\right) &= \frac{\mathrm{E}(X)}{\pi^2} + \frac{(1-\mathrm{E}(X))}{(1-\pi)^2} = \frac{1}{\pi(1-\pi)} \\
I(\pi) &= \frac{1}{\pi(1-\pi)}.
\end{aligned}
$$

Note que $\mathrm{Var}(X|\pi) = \pi(1-\pi)$, quanto maior a variância menor a informação e quanto menor a variância maior a informação.

Exemplo

Suponha que $(X|\mu) \sim \mathcal{N}(\mu;\, \sigma_0^2)$ com σ_0^2 conhecido. Temos

$$
\begin{aligned}
f(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma_0^2}\right\} \\
\ell(\mu|x) &= -\frac{1}{2}\left[\ln(2\pi) + \ln(\sigma_0^2)\right] - \frac{(x-\mu)^2}{2\sigma_0^2} \\
\frac{d\ell(\mu|x)}{d\mu} &= \frac{(x-\mu)}{\sigma_0^2} \quad \Rightarrow \quad \mathrm{E}\left(\frac{d\ell(\mu|X)}{d\mu}\right) = \frac{1}{\sigma_0^2}\mathrm{E}(X-\mu) = 0 \\
\frac{d^2\ell(\mu|x)}{d\mu^2} &= -\frac{1}{\sigma_0^2} \quad \Rightarrow \quad \mathrm{E}\left(-\frac{d^2\ell(\mu|X)}{d\mu^2}\right) = \frac{1}{\sigma_0^2} \\
I(\mu) &= \frac{1}{\sigma_0^2}.
\end{aligned}
$$

Note que $\mathrm{Var}(X|\mu) = \sigma_0^2$, quanto menor a variância maior a informação.

Informação de Fisher de uma amosta aleatória

Definição

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$, $\theta \in \Omega$, um intervalo aberto da reta. Seja $f(\boldsymbol{x}|\theta)$ a função (de densidade) de probabilidade conjunta de $\boldsymbol{X}$. Defina

$$\ell(\theta|\boldsymbol{x}) = \ln f(\boldsymbol{x}|\theta).$$

Assuma que os valores de x que satisfazem $f(\boldsymbol{x}|\theta) > 0$ são os mesmos para todo θ e que $\ell(\theta|\boldsymbol{x})$ é duas vezes diferenciável com respeito a θ. A **informação de Fisher** de uma amostra aleatória $\boldsymbol{X}$ é definida como

$$I_n(\theta) = \mathrm{E}([\ell'(\theta|\boldsymbol{X})]^2|\theta) = \int\int\cdots\int [\ell'(\theta|\boldsymbol{x})]^2 f(\boldsymbol{x}|\theta)\, dx_1\, dx_2 \ldots dx_n.$$

Além disso, sob certas condições, temos que

$$I_n(\theta) = \mathrm{E}(-\ell''(\theta|\boldsymbol{X})|\theta) = \mathrm{Var}(\ell'(\theta|\boldsymbol{X})|\theta).$$

Teorema

Sob as condições das duas definições anteriores, temos que

$$I_n(\theta) = nI(\theta).$$

Interpretação: quanto maior o tamanho da amostra, maior a informação.

Prova

Temos que $\ell(\theta|\boldsymbol{x}) = \ln f(\boldsymbol{x}|\theta)$ e

$$\begin{aligned}
f(\boldsymbol{x}|\theta) &= f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) \quad \Rightarrow \\
\ell(\theta|\boldsymbol{x}) &= \ell(\theta|x_1) + \ell(\theta|x_2) + \cdots + \ell(\theta|x_n) = \sum_{i=1}^{n} \ell(\theta|x_i) \quad \Rightarrow \\
\ell'(\theta|\boldsymbol{x}) &= \ell'(\theta|x_1) + \ell'(\theta|x_2) + \cdots + \ell'(\theta|x_n) = \sum_{i=1}^{n} \ell'(\theta|x_i) \quad \Rightarrow \\
\ell''(\theta|\boldsymbol{x}) &= \ell''(\theta|x_1) + \ell''(\theta|x_2) + \cdots + \ell''(\theta|x_n) = \sum_{i=1}^{n} \ell''(\theta|x_i) \quad \Rightarrow \\
\mathrm{E}(-\ell''(\theta|\boldsymbol{X})) &= \sum_{i=1}^{n} \mathrm{E}(-\ell''(\theta|X_i)) = \sum_{i=1}^{n} I(\theta) = nI(\theta).
\end{aligned}$$

□

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$. Temos

$$
\begin{aligned}
L(\pi) &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})} \\
\ell(\pi) &= n\overline{x}\ln\pi + n(1-\overline{x})\ln(1-\pi) \\
\frac{d\ell(\pi)}{d\pi} &= \frac{n\overline{x}}{\pi} - \frac{n(1-\overline{x})}{1-\pi}, \qquad \frac{d\ell(\pi)}{d\pi} = 0 \Rightarrow \widehat{\pi} = \overline{x} \\
\mathrm{E}\left(\frac{d\ell(\pi)}{d\pi}\right) &= \frac{n\mathrm{E}(\overline{X})}{\pi} - \frac{n(1-\mathrm{E}(\overline{X}))}{(1-\pi)} = \frac{n\pi}{\pi} - \frac{n(1-\pi)}{1-\pi} = 0 \\
\frac{d^2\ell(\pi)}{d\pi^2} &= -\frac{n\overline{x}}{\pi^2} - \frac{n(1-\overline{x})}{(1-\pi)^2} < 0, \quad \widehat{\pi} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\pi)}{d\pi^2}\right) &= \frac{n\mathrm{E}(\overline{X})}{\pi^2} + \frac{n(1-\mathrm{E}(\overline{X}))}{(1-\pi)^2} = \frac{n}{\pi(1-\pi)} \\
I_n(\pi) &= \frac{n}{\pi(1-\pi)} \quad (\textbf{A informação cresce com } n.).
\end{aligned}
$$

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$. Temos

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left[\frac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!} \right] = \frac{\exp\{-n\lambda\}\lambda^{n\overline{x}}}{\prod_{i=1}^{n} \Gamma(x_i+1)} \\
\ell(\lambda) &= -n\lambda + n\overline{x} \ln \lambda - \sum_{i=1}^{n} \ln \Gamma(x_i+1) \\
\frac{d\ell(\lambda)}{d\lambda} &= -n + \frac{n\overline{x}}{\lambda}, \qquad \frac{d\ell(\lambda)}{d\lambda} = 0 \Rightarrow \widehat{\lambda} = \overline{x} \\
\mathrm{E}\left(\frac{d\ell(\lambda)}{d\lambda} \right) &= -n + \frac{n\mathrm{E}(\overline{X})}{\lambda} = -n + \frac{n\lambda}{\lambda} = 0 \\
\frac{d^2\ell(\lambda)}{d\lambda^2} &= -\frac{n\overline{x}}{\lambda^2} < 0, \quad \widehat{\lambda} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\lambda)}{d\lambda^2} \right) &= \frac{n\mathrm{E}(\overline{X})}{\lambda^2} = \frac{n}{\lambda} \\
I_n(\lambda) &= \frac{n}{\lambda} \qquad \text{(Note que } \mathrm{Var}(X|\lambda) = \lambda\text{).}
\end{aligned}
$$

Desigualdade (da Informação) de Cramér-Rao

A informação de Fisher pode ser utilizada para calcular um limite inferior para a variância de um estimador de um determinado parâmetro θ.

Teorema

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$. Suponha que todas as condições anteriores são satisfeitas. Seja $T = r(\boldsymbol{X})$ uma estatística com variância finita e seja $m(\theta) = \mathrm{E}(T|\theta)$. Assuma que $m(\theta)$ é diferenciável. Então,

$$\mathrm{Var}(T|\theta) \geqslant \frac{[m'(\theta)]^2}{nI(\theta)}.$$

Além disso,

$$\mathrm{Var}(T|\theta) = \frac{[m'(\theta)]^2}{nI(\theta)}$$

se, e somente se, existem funções $u(\theta)$ e $v(\theta)$ - que não dependem de $\boldsymbol{X}$ - tal que

$$T = u(\theta)\ell'(\theta|\boldsymbol{X}) + v(\theta).$$

Prova

A desigualdade de Schwarz é dada por

$$[\text{Cov}(X, Y)]^2 \leqslant \text{Var}(X)\text{Var}(Y).$$

Ao aplicar em T e $\ell'(\theta|\boldsymbol{X})$, temos que

$$[\text{Cov}(T, \ell'(\theta|\boldsymbol{X})|\theta)]^2 \leqslant \text{Var}(T|\theta)\text{Var}(\ell'(\theta|\boldsymbol{X})|\theta) \Rightarrow$$

$$\frac{[\text{Cov}(T, \ell'(\theta|\boldsymbol{X})|\theta)]^2}{\text{Var}(\ell'(\theta|\boldsymbol{X})|\theta)} \leqslant \text{Var}(T|\theta) \Rightarrow$$

$$\frac{[\text{Cov}(T, \ell'(\theta|\boldsymbol{X})|\theta)]^2}{nI(\theta)} \leqslant \text{Var}(T|\theta).$$

Além disso, sabemos que $\text{E}(\ell'(\theta|\boldsymbol{X})|\theta) = 0$. Assim,

$$\text{Cov}(T, \ell'(\theta|\boldsymbol{X})|\theta) = \text{E}(T\ell'(\theta|\boldsymbol{X})|\theta).$$

Segue-se que

$$\begin{aligned}
\mathrm{E}(T\ell'(\theta|\boldsymbol{X})|\theta) &= \int\int\cdots\int r(\boldsymbol{x})\ell'(\theta|\boldsymbol{x})f(\boldsymbol{x}|\theta)dx_1\,dx_2\ldots dx_n \\
&= \int\int\cdots\int r(\boldsymbol{x})f'(\boldsymbol{x}|\theta)dx_1\,dx_2\ldots dx_n.
\end{aligned}$$

Agora, defina

$$m(\theta) = \int\int\cdots\int r(\boldsymbol{x})f(\boldsymbol{x}|\theta)dx_1\,dx_2\ldots dx_n$$

e suponha que a ordem da derivada e da integral podem ser trocadas tal que

$$\begin{aligned}
m'(\theta) &= \frac{d}{d\theta}m(\theta) = \frac{d}{d\theta}\int\int\cdots\int r(\boldsymbol{x})f(\boldsymbol{x}|\theta)dx_1\,dx_2\ldots dx_n \\
&= \int\int\cdots\int r(\boldsymbol{x})\frac{d}{d\theta}f(\boldsymbol{x}|\theta)dx_1\,dx_2\ldots dx_n \\
&= \int\int\cdots\int r(\boldsymbol{x})f'(\boldsymbol{x}|\theta)dx_1\,dx_2\ldots dx_n.
\end{aligned}$$

Então, $\mathrm{Cov}(T,\ell'(\theta|\boldsymbol{X})|\theta) = m'(\theta) \quad\Rightarrow\quad \mathrm{Var}(T|\theta) \geqslant \dfrac{[m'(\theta)]^2}{nI(\theta)}.$

Agora, $[\text{Cov}(X, Y)]^2 = \text{Var}(X)\text{Var}(Y)$ se, e somente se, $|\text{Cor}(X, Y)| = 1$. Em outras palavras se, e somente se, X é uma combinação linear de Y, ou seja, $X = aY + b$ para algum $a \neq 0$.

Assim, $T = a\ell'(\theta|\boldsymbol{X}) + b$ se, e somente se,

$$\text{Var}(T|\theta) = \frac{[m'(\theta)]^2}{nI(\theta)}.$$

Mas a análise é feita condicional a θ. Portanto, podemos tomar $a = u(\theta)$ e $b = v(\theta)$ tal que $T = u(\theta)\ell'(\theta|\boldsymbol{X}) + v(\theta)$. □

Corolário

Considere as suposições do teorema anterior. Se $T = r(\boldsymbol{X})$ é um estimador não tendencioso para θ, então

$$\text{Var}(T|\theta) \geqslant \frac{1}{nI(\theta)}.$$

Prova

Se $m(\theta) = \text{E}(T|\theta) = \theta$ (não tendencioso), então $m'(\theta) = 1$. □

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$ para $n > 2$. Temos

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} [\lambda \exp\{-\lambda x\}] = \lambda^n \exp\{-\lambda n \overline{x}\} \\
\ell(\lambda) &= n \ln \lambda - \lambda n \overline{x} \\
\frac{d\ell(\lambda)}{d\lambda} &= \frac{n}{\lambda} - n\overline{x}, \qquad \frac{d\ell(\lambda)}{d\lambda} = 0 \Rightarrow \widehat{\lambda} = \frac{1}{\overline{x}} \\
\mathrm{E}\left(\frac{d\ell(\lambda)}{d\lambda}\right) &= \frac{n}{\lambda} - n\mathrm{E}(\overline{X}) = \frac{n}{\lambda} - \frac{n}{\lambda} = 0 \\
\frac{d^2\ell(\lambda)}{d\lambda^2} &= -\frac{n}{\lambda^2} < 0, \quad \widehat{\lambda} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\lambda)}{d\lambda^2}\right) &= \frac{n}{\lambda^2} \\
I_n(\lambda) &= nI(\lambda) = \frac{n}{\lambda^2}.
\end{aligned}
$$

Agora, note que $(X|\lambda) \sim \mathcal{E}(\lambda) \equiv \mathcal{G}(1,\lambda)$. Segue-se que

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{G}(n, n\lambda) \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{\overline{X}} \sim \mathcal{IG}(n, n\lambda).$$

Consequentemente, temos que

$$m(\lambda) = \mathrm{E}(\hat{\lambda}|\lambda) = \frac{n\lambda}{n-1}, \quad m'(\lambda) = \frac{n}{n-1} \quad \text{e} \quad \mathrm{Var}(\hat{\lambda}|\lambda) = \frac{n^2\lambda^2}{(n-1)^2(n-2)}.$$

Pela desigualdade de Cramér-Rao, temos que

$$\mathrm{Var}(\hat{\lambda}|\lambda) = \frac{n^2\lambda^2}{(n-1)^2(n-2)} \geqslant \frac{[m'(\lambda)]^2}{nI(\lambda)} = \frac{n^2\lambda^2}{(n-1)^2 n} \quad \text{para} \quad n > 2.$$

A igualdade não vale pois $\hat{\lambda}$ é uma função não linear de $\ell'(\lambda) = n \ln \lambda - \lambda n\overline{X}$.

Estimador eficiente

Um estimador cuja variância é igual no limite inferior de Cramér-Rao, isto é, $\frac{[m'(\theta)]^2}{nI(\theta)}$ faz o uso mais eficiente da informação que vem dos dados de alguma forma.

Definição

Um estimador $T = r(\boldsymbol{X})$ é eficiente para sua média $m(\theta)$ se, e somente se,

$$\text{Var}(T|\theta) = \frac{[m'(\theta)]^2}{nI(\theta)}.$$

Observações:

- Pode não haver um estimador eficiente para uma particular função $m(\theta)$;
- Frequentemente existem estimadores eficientes para muitas funções interessantes de θ; e
- É possível que o único estimador eficiente de θ seja uma constante.

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$. Temos

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left[\frac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!} \right] = \frac{\exp\{-n\lambda\}\lambda^{n\overline{x}}}{\prod_{i=1}^{n} \Gamma(x_i+1)} \\
\ell(\lambda) &= -n\lambda + n\overline{x} \ln \lambda - \sum_{i=1}^{n} \ln \Gamma(x_i+1) \\
\frac{d\ell(\lambda)}{d\lambda} &= -n + \frac{n\overline{x}}{\lambda}, \qquad \frac{d\ell(\lambda)}{d\lambda} = 0 \Rightarrow \widehat{\lambda} = \overline{x} \\
\mathrm{E}\left(\frac{d\ell(\lambda)}{d\lambda} \right) &= -n + \frac{n\mathrm{E}(\overline{X})}{\lambda} = -n + \frac{n\lambda}{\lambda} = 0 \\
\frac{d^2\ell(\lambda)}{d\lambda^2} &= -\frac{n\overline{x}}{\lambda^2} < 0, \quad \widehat{\lambda} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\lambda)}{d\lambda^2} \right) &= \frac{n\mathrm{E}(\overline{X})}{\lambda^2} = \frac{n}{\lambda} \\
I_n(\lambda) &= nI(\lambda) = \frac{n}{\lambda}.
\end{aligned}
$$

Além disso, sabemos que $\widehat{\lambda} = \overline{X}$ e S^2 são estimadores não tendenciosos para λ.

Defina $u(\lambda) = \dfrac{\lambda}{n}$ e $v(\lambda) = \lambda$. Logo,

$$\widehat{\lambda} = u(\lambda)\ell'(\lambda) + v(\lambda) = \frac{\lambda}{n}\left[-n + \frac{n\overline{X}}{\lambda}\right] + \lambda = \overline{X}.$$

Em outras palavras, $\widehat{\lambda}$ é uma combinação linear de $\ell'(\lambda)$. Portanto, $\widehat{\lambda}$ tem a variância igual ao limite inferior de Cramér-Rao, sendo assim um **estimador eficiente** para λ.

Por outro lado, não é possível escrever S^2 como combinação linear de $\ell'(\lambda)$. Portanto, a variância de S^2 será sempre maior que a variância de $\widehat{\lambda}$. Assim, S^2 **não** é um **estimador eficiente** para λ.

Estimador não tendencioso com variância mínima

Sejam $T_1 = r_1(\boldsymbol{X})$ e $T_2 = r_2(\boldsymbol{X})$ estimadores não tendenciosos de $m(\theta)$. Suponha que para qualquer valor de θ,

$$\text{Var}(T_1|\theta) = \frac{[m'(\theta)]^2}{nI(\theta)} \quad \text{e} \quad \text{Var}(T_2|\theta) \geqslant \frac{[m'(\theta)]^2}{nI(\theta)}.$$

Isto implica que

$$\text{Var}(T_1|\theta) \leqslant \text{Var}(T_2|\theta).$$

Assim, um estimador não tendencioso de variância mínima é um estimador eficiente.

Distribuição assintótica de um estimador eficiente

Teorema

Assuma as mesmas condições do último teorema (desigualdade de Cramér-Rao). Seja $T = r(\boldsymbol{X})$ um estimador eficiente de sua média $m(\theta)$ e que $m'(\theta) \neq 0$ para todo θ. Então, a distribuição assintótica de

$$\frac{[nI(\theta)]^{1/2}}{m'(\theta)}(T - m(\theta))$$

é uma distribuição normal padrão.

Prova

Considere

$$\ell(\theta|\boldsymbol{X}) = \sum_{i=1}^{n} \ell(\theta|X_i) \quad \Rightarrow \quad \ell'(\theta|\boldsymbol{X}) = \sum_{i=1}^{n} \ell'(\theta|X_i),$$

em que $\ell'(\theta|X_i)$, $i = 1, 2, \ldots, n$, são variáveis aleatórias independentes e identicamente distribuídas. Sabemos que

$$\mathrm{E}(\ell'(\theta|X_i)|\theta) = 0 \quad \text{e} \quad \mathrm{Var}(\ell'(\theta|X_i)|\theta) = I(\theta),$$

são finitas e iguais. Portanto, pelo teorema central do limite de Lindeberg e Lévy,

$$\frac{\ell'(\theta|\boldsymbol{X})}{[nI(\theta)]^{1/2}} \xrightarrow{D} \mathcal{N}(0,1), \quad \text{quando} \quad n \longrightarrow \infty.$$

Como T é um estimador eficiente de $m(\theta)$, temos que

$$\mathrm{E}(T|\theta) = m(\theta) \quad \text{e} \quad \mathrm{Var}(T|\theta) = \frac{[m'(\theta)]^2}{nI(\theta)}.$$

Por outro lado, temos que

$$T = u(\theta)\ell'(\theta|\boldsymbol{X}) + v(\theta),$$

tal que

$$\mathrm{E}(T|\theta) = v(\theta) \quad \text{e} \quad \mathrm{Var}(T|\theta) = [u(\theta)]^2 nI(\theta),$$

e, consequentemente,

$$v(\theta) = m(\theta) \quad \text{e} \quad |u(\theta)| = \frac{|m'(\theta)|}{nI(\theta)}.$$

Agora, sem perda de generalidade, tomamos $u(\theta) = \dfrac{m'(\theta)}{nI(\theta)}$ tal que

$$T = \frac{m'(\theta)}{nI(\theta)}\ell'(\theta|\boldsymbol{X}) + m(\theta),$$

e, consequentemente,

$$\frac{[nI(\theta)]^{1/2}}{m'(\theta)}(T - m(\theta)) = \frac{\ell'(\theta|\boldsymbol{X})}{[nI(\theta)]^{1/2}} \xrightarrow{D} \mathcal{N}(0,1), \quad \text{quando} \quad n \longrightarrow \infty.$$

□

Distribuição assintótica do estimador de máxima verossimilhança

Teorema

Suponha que um determinado estimador de máxima verossimilhança $\hat{\theta}$ é encontrado ao resolver a equação $\ell'(\theta|\boldsymbol{x}) = 0$ e que $\ell''(\theta|\boldsymbol{x})$ e $\ell'''(\theta|\boldsymbol{x})$ existam e satisfaçam a certas condições. Então, a distribuição assintótica de

$$[nI(\theta)]^{1/2}(\hat{\theta} - \theta)$$

é um distribuição normal padrão.

Equivalentemente, temos que $\hat{\theta} \overset{a}{\sim} \mathcal{N}(\theta, [nI(\theta)]^{-1})$.

Dizemos que $\hat{\theta}$ é um estimador assintoticamente eficiente.

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\sigma^2) \sim \mathcal{N}(0;\, \sigma^2)$. O estimador de máxima verossimilhança de σ (desvio padrão) é dado por

$$\hat{\sigma} = \left[\frac{1}{n}\sum\nolimits_{i=1}^{n} X_i^2\right]^{1/2}$$

Pode ser mostrado que $I(\sigma) = \dfrac{2}{\sigma^2}$. Portanto, se o tamanho da amostra n é grande,

$$\hat{\sigma} \overset{a}{\sim} \mathcal{N}\left(\sigma;\, \frac{\sigma^2}{2n}\right).$$

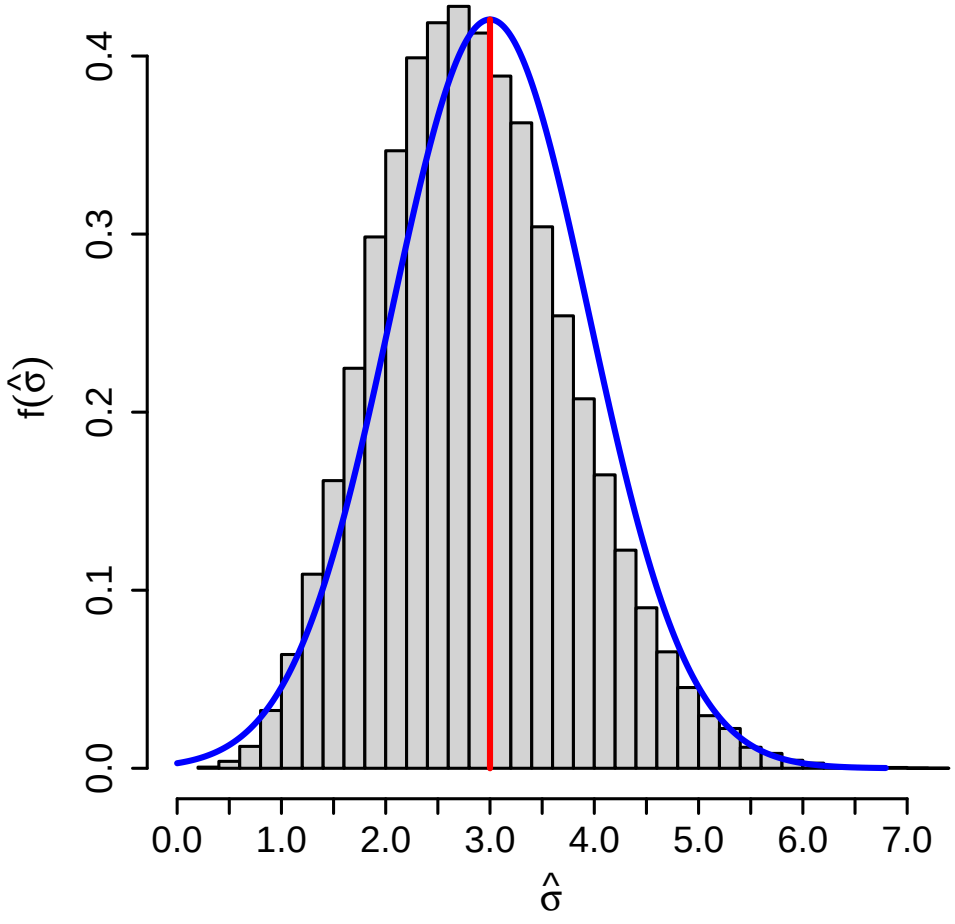


Figura: Distribuição amostral assintótica de $\hat{\sigma}$ para $n = 5$.

Ver exemplo_13.r

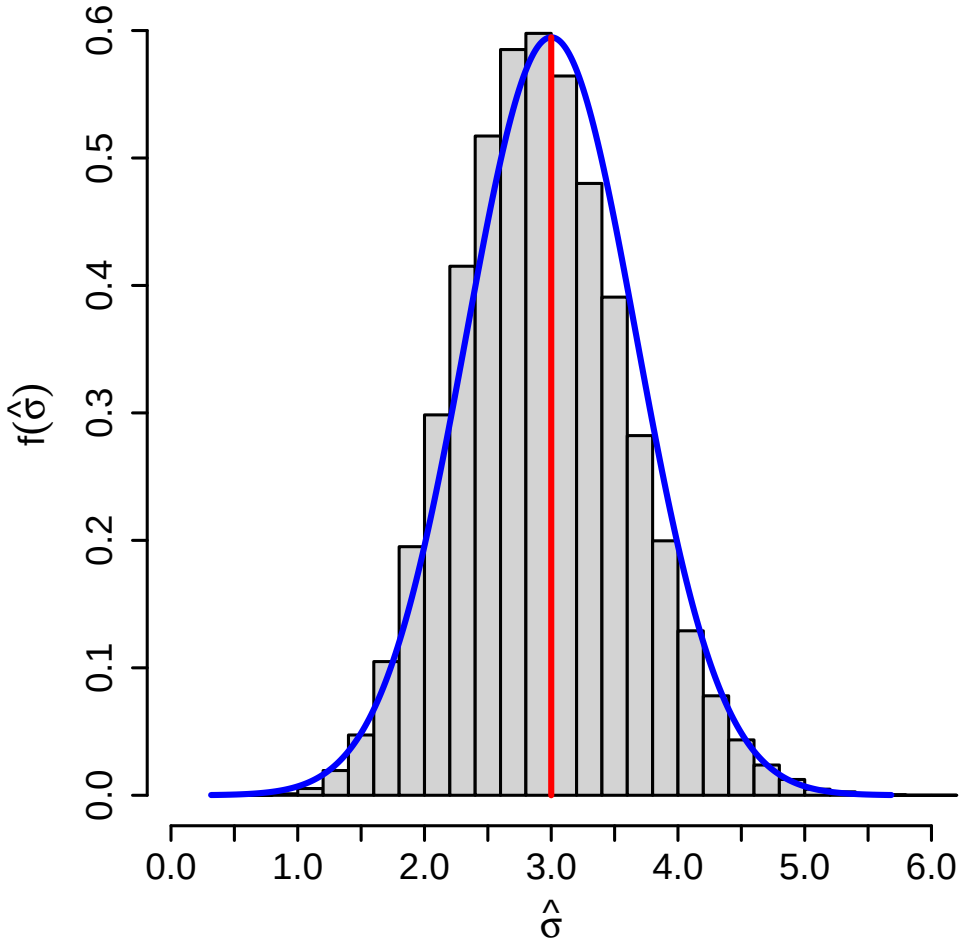


Figura: Distribuição amostral assintótica de $\hat{\sigma}$ para $n = 10$.

Ver exemplo_13.r

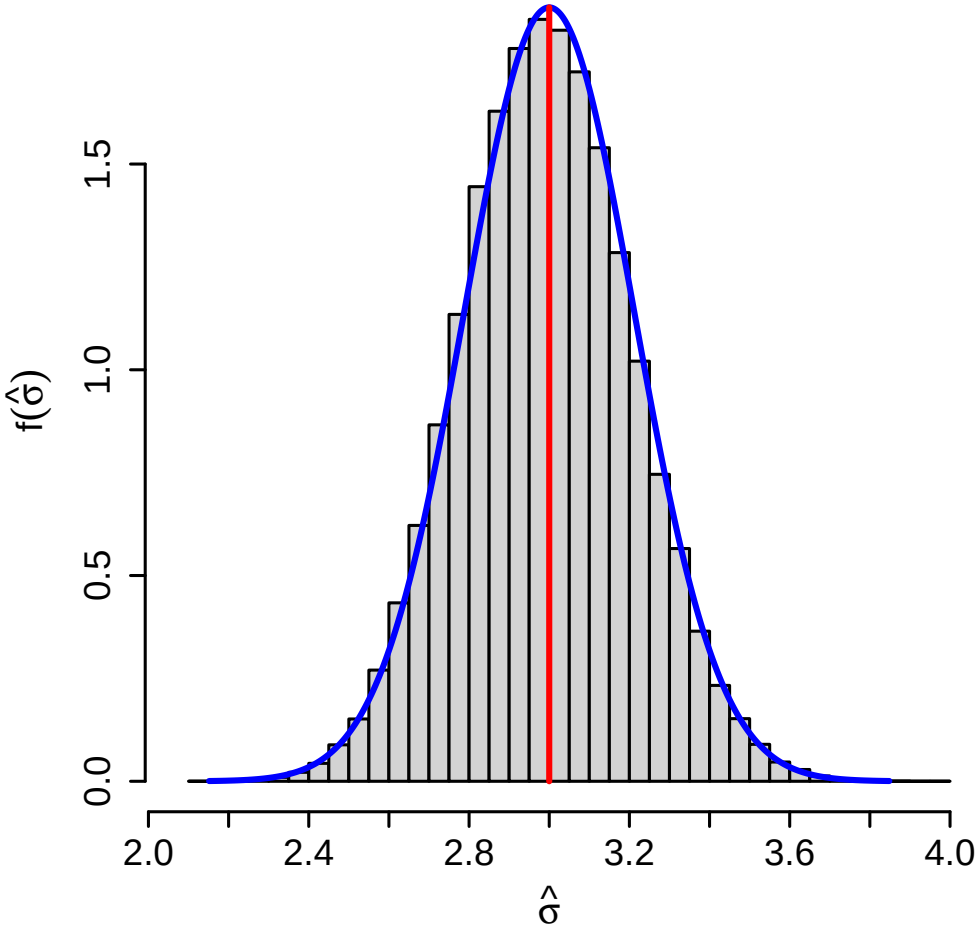


Figura: Distribuição amostral assintótica de $\hat{\sigma}$ para $n = 100$.

Ver exemplo_13.r

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$. Temos

$$
\begin{aligned}
L(\pi) &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})} \\
\ell(\pi) &= n\overline{x}\ln\pi + n(1-\overline{x})\ln(1-\pi) \\
\frac{d\ell(\pi)}{d\pi} &= \frac{n\overline{x}}{\pi} - \frac{n(1-\overline{x})}{1-\pi}, \qquad \frac{d\ell(\pi)}{d\pi} = 0 \Rightarrow \widehat{\pi} = \overline{x} \\
\frac{d^2\ell(\pi)}{d\pi^2} &= -\frac{n\overline{x}}{\pi^2} - \frac{n(1-\overline{x})}{(1-\pi)^2} < 0, \quad \widehat{\pi} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\pi)}{d\pi^2}\right) &= \frac{n\mathrm{E}(\overline{X})}{\pi^2} + \frac{n(1-\mathrm{E}(\overline{X}))}{(1-\pi)^2} = \frac{n}{\pi(1-\pi)}.
\end{aligned}
$$

Pelo teorema anterior, temos que

$$
\frac{\sqrt{n}(\overline{X}-\pi)}{\sqrt{\pi(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0;\, 1).
$$

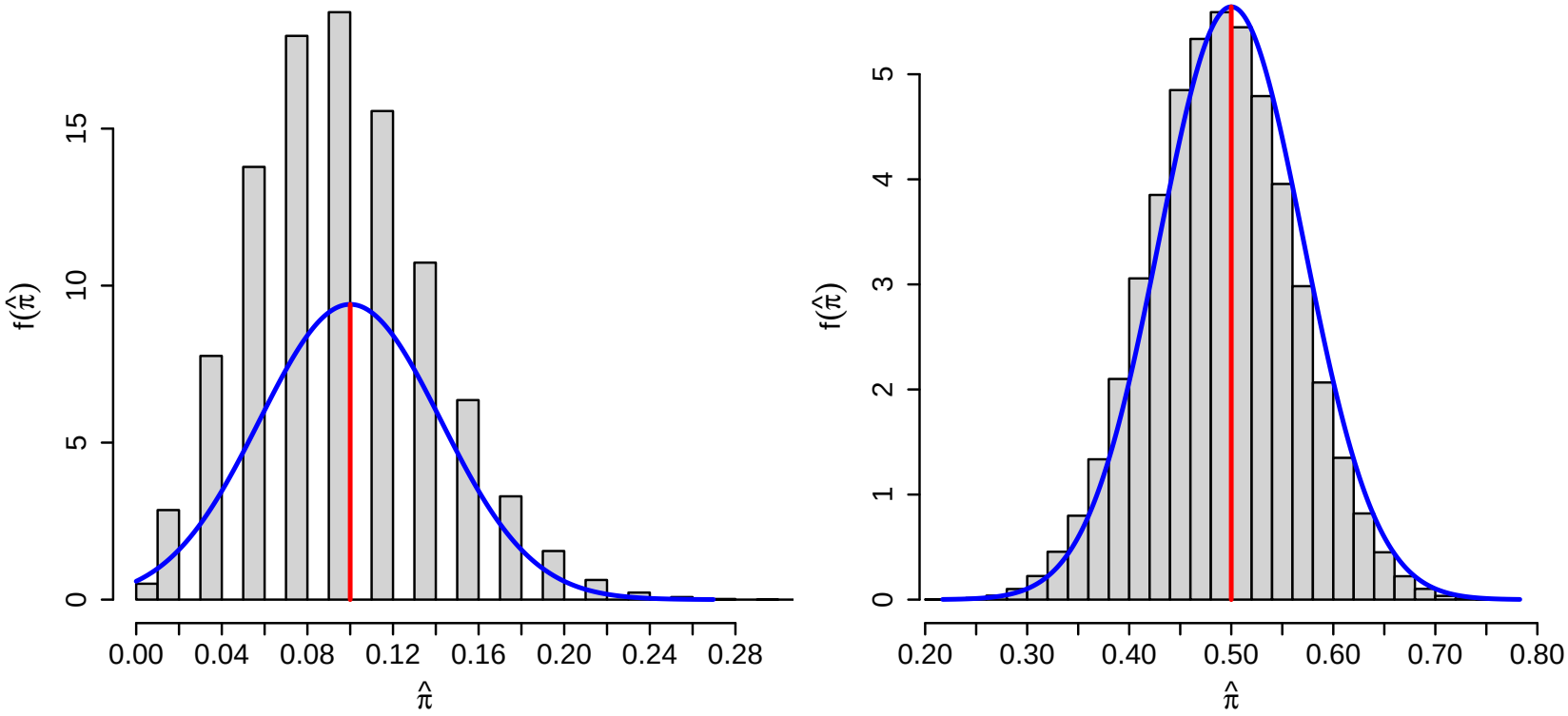


Figura: Distribuição amostral assintótica de $\hat{\pi}$ para $n = 50$ com $\pi = 0,1$ na esquerda e $\pi = 0,5$ na direita.

Ver exemplo_14.r

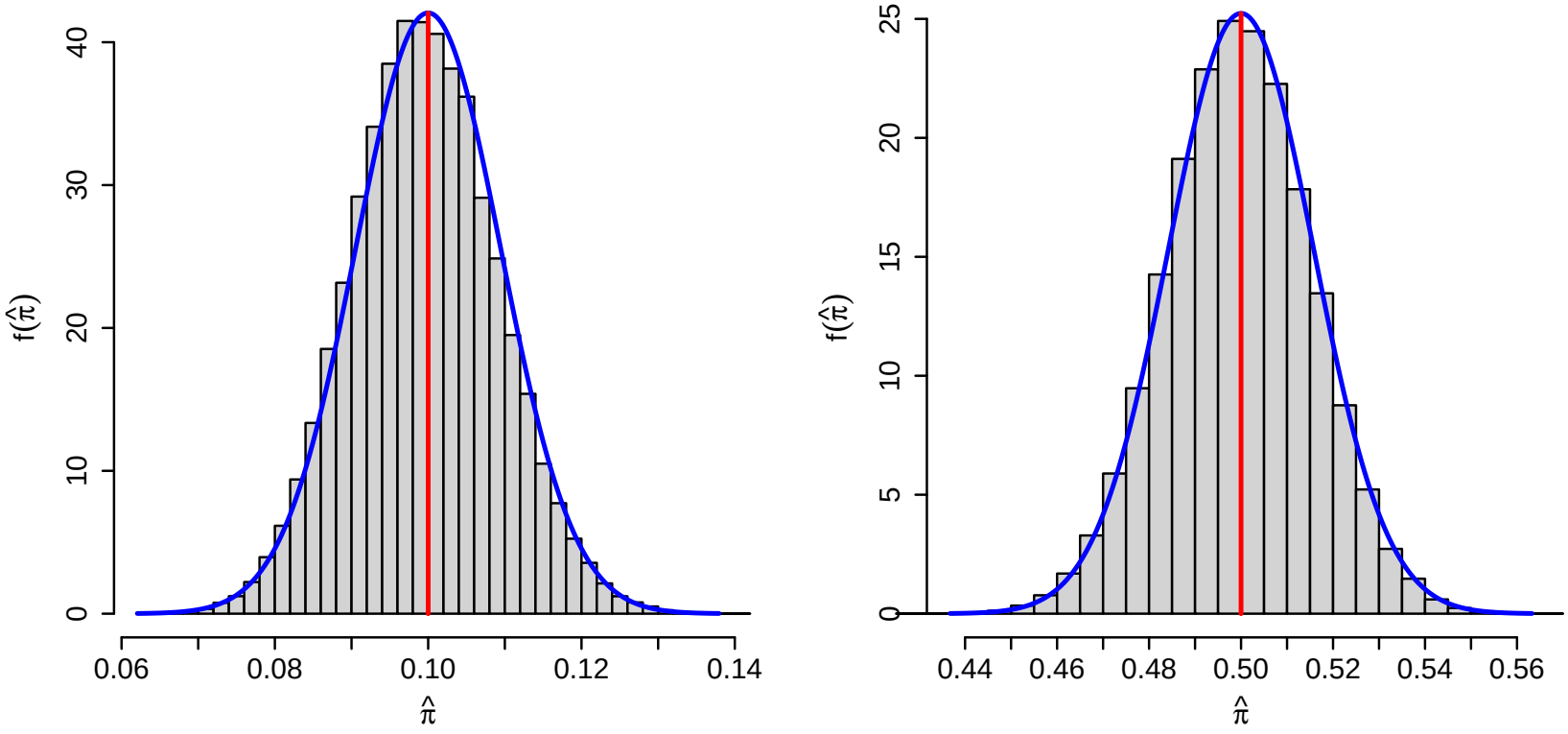


Figura: Distribuição amostral assintótica de $\hat{\pi}$ para $n = 1.000$ com $\pi = 0,1$ na esquerda e $\pi = 0,5$ na direita.

Ver exemplo_14.r

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{GM}(\pi)$. Temos

$$
\begin{aligned}
L(\pi) &= \prod_{i=1}^{n} \pi(1-\pi)^{x_i - 1} = \pi^n (1-\pi)^{n(\overline{x}-1)} \\
\ell(\pi) &= n \ln \pi + n(\overline{x} - 1) \ln(1-\pi) \\
\frac{d\ell(\pi)}{d\pi} &= \frac{n}{\pi} - \frac{n(\overline{x}-1)}{1-\pi}, \qquad \frac{d\ell(\pi)}{d\pi} = 0 \Rightarrow \widehat{\pi} = \frac{1}{\overline{x}} \\
\frac{d^2\ell(\pi)}{d\pi^2} &= -\frac{n}{\pi^2} - \frac{n(\overline{x}-1)}{(1-\pi)^2} < 0, \quad \widehat{\pi} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\pi)}{d\pi^2}\right) &= \frac{n}{\pi^2} + \frac{n(\mathrm{E}(\overline{X})-1)}{(1-\pi)^2} = \frac{n}{\pi^2(1-\pi)}.
\end{aligned}
$$

A informação é menor para $\pi = 2/3$. Pelo teorema anterior, temos que

$$
\frac{\sqrt{n}\left(\dfrac{1}{\overline{X}} - \pi\right)}{\sqrt{\pi^2(1-\pi)}} \overset{a}{\sim} \mathcal{N}(0;\, 1).
$$

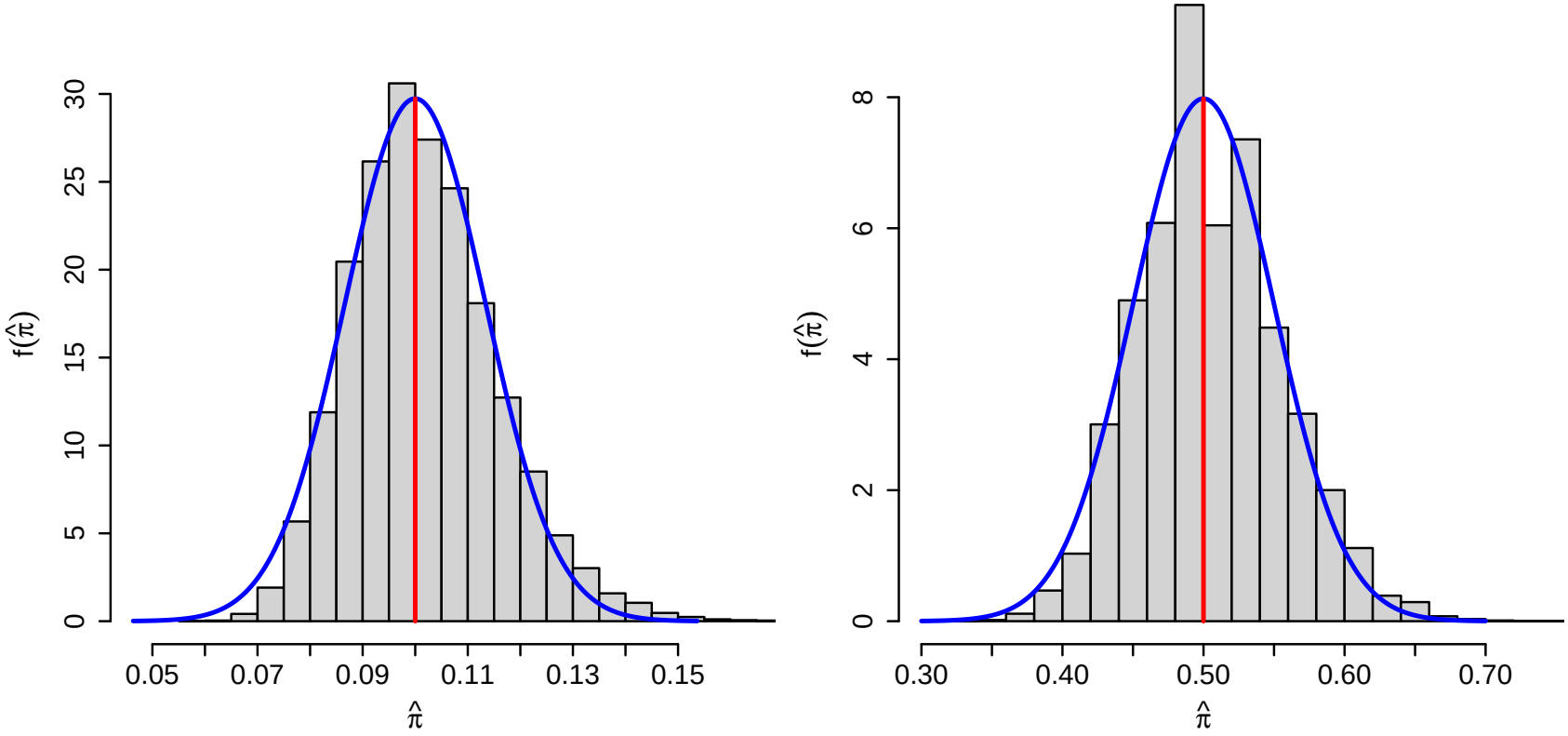


Figura: Distribuição amostral assintótica de $\hat{\pi}$ para $n = 50$ com $\pi = 0,1$ na esquerda e $\pi = 0,5$ na direita.

Ver exemplo_15.r

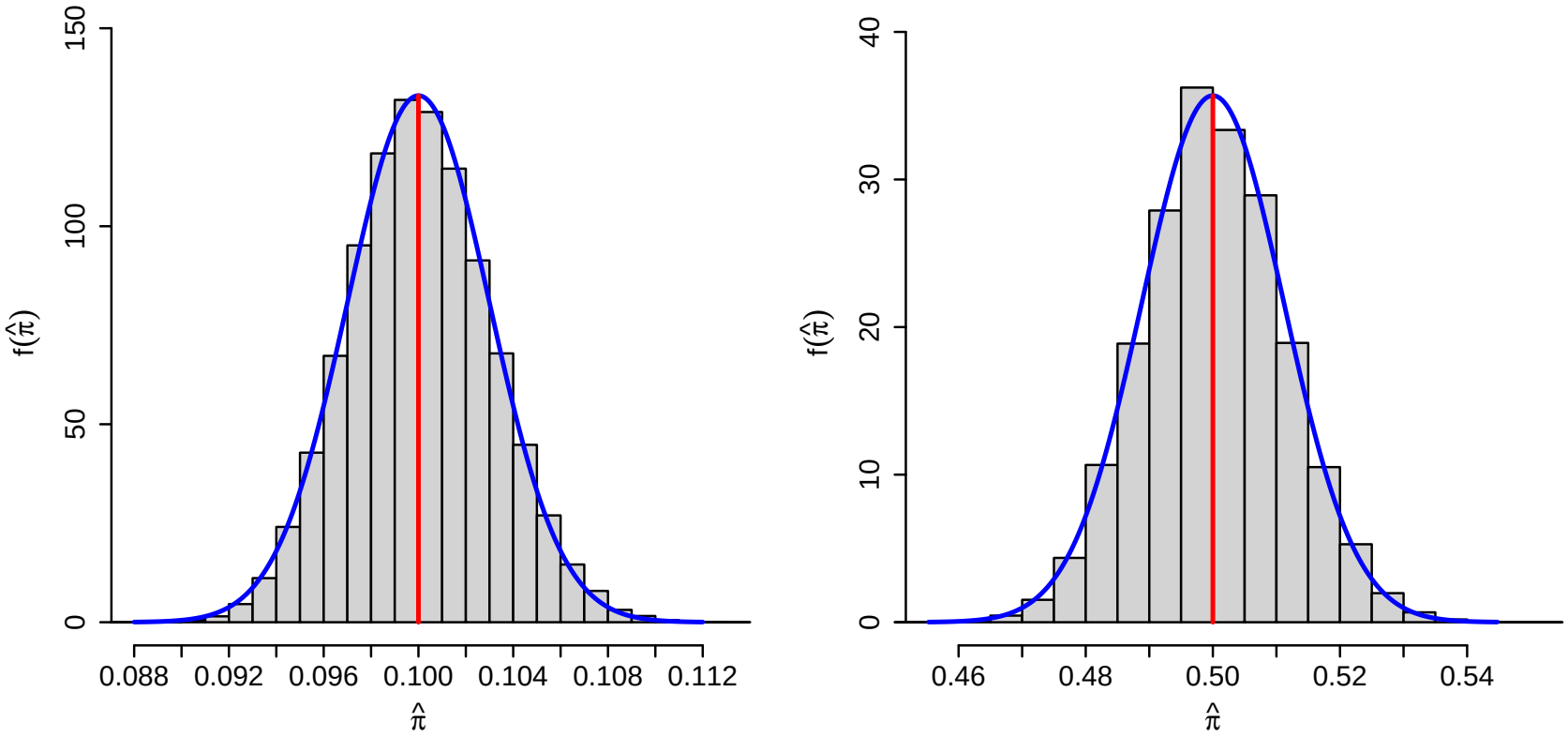


Figura: Distribuição amostral assintótica de $\hat{\pi}$ para $n = 1.000$ com $\pi = 0,1$ na esquerda e $\pi = 0,5$ na direita.

Ver exemplo_15.r

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória simples de $(X|\lambda) \sim \mathcal{P}(\lambda)$.
Temos

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left[\frac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!} \right] = \frac{\exp\{-n\lambda\}\lambda^{n\overline{x}}}{\prod_{i=1}^{n} \Gamma(x_i+1)} \\
\ell(\lambda) &= -n\lambda + n\overline{x} \ln \lambda - \sum_{i=1}^{n} \ln \Gamma(x_i+1) \\
\frac{d\ell(\lambda)}{d\lambda} &= -n + \frac{n\overline{x}}{\lambda}, \qquad \frac{d\ell(\lambda)}{d\lambda} = 0 \Rightarrow \widehat{\lambda} = \overline{x} \\
\frac{d^2\ell(\lambda)}{d\lambda^2} &= -\frac{n\overline{x}}{\lambda^2} < 0, \quad \widehat{\lambda} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{d^2\ell(\lambda)}{d\lambda^2} \right) &= \frac{n\mathrm{E}(\overline{X})}{\lambda^2} = \frac{n}{\lambda}.
\end{aligned}
$$

Pelo teorema anterior, temos que

$$
\frac{\sqrt{n}(\overline{X} - \lambda)}{\sqrt{\lambda}} \overset{a}{\sim} \mathcal{N}(0;\, 1).
$$

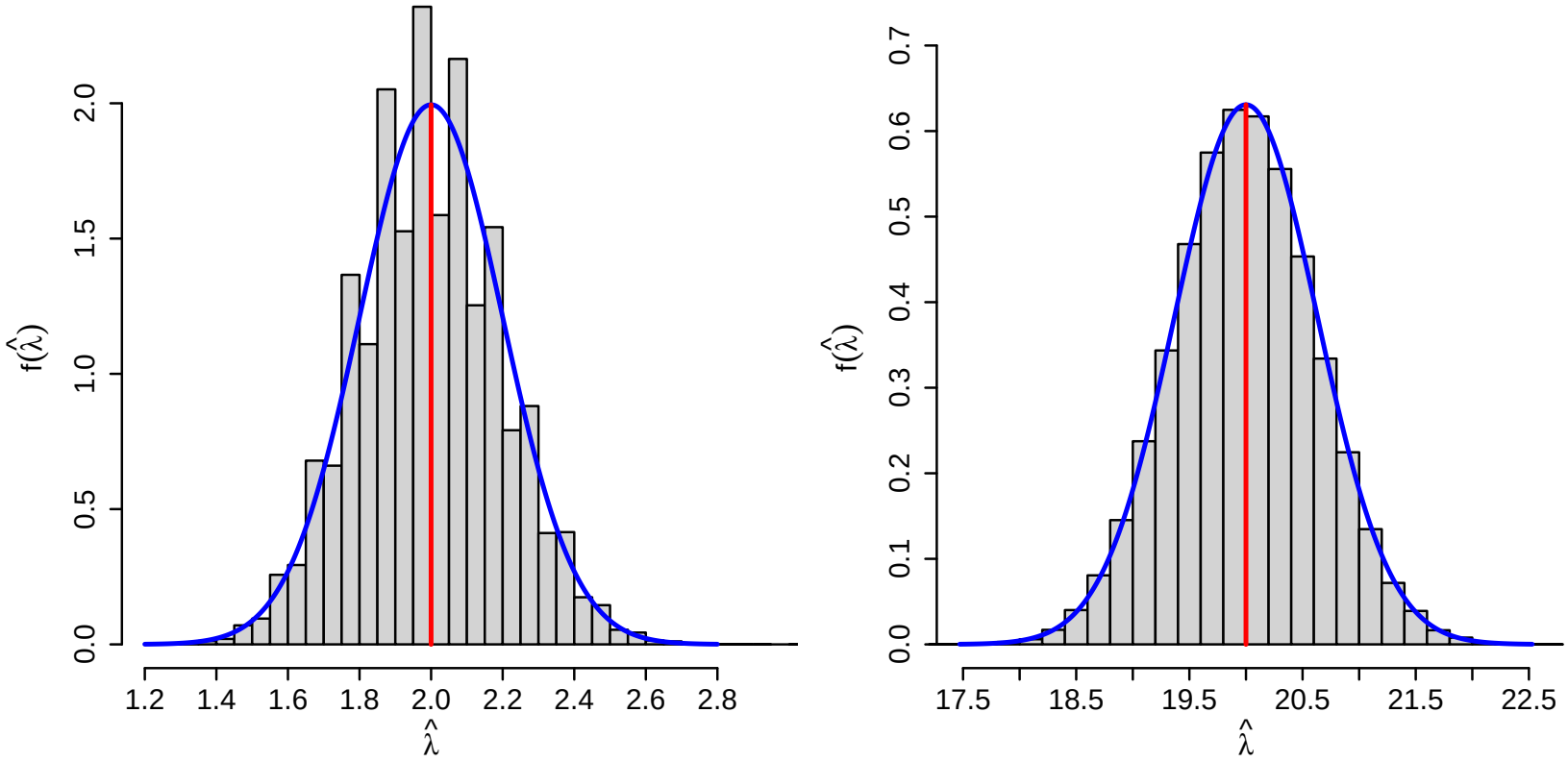


Figura: Distribuição amostral assintótica de $\hat{\lambda}$ para $n = 50$ com $\lambda = 2$ na esquerda e $\lambda = 20$ na direita.

Ver exemplo_13.r

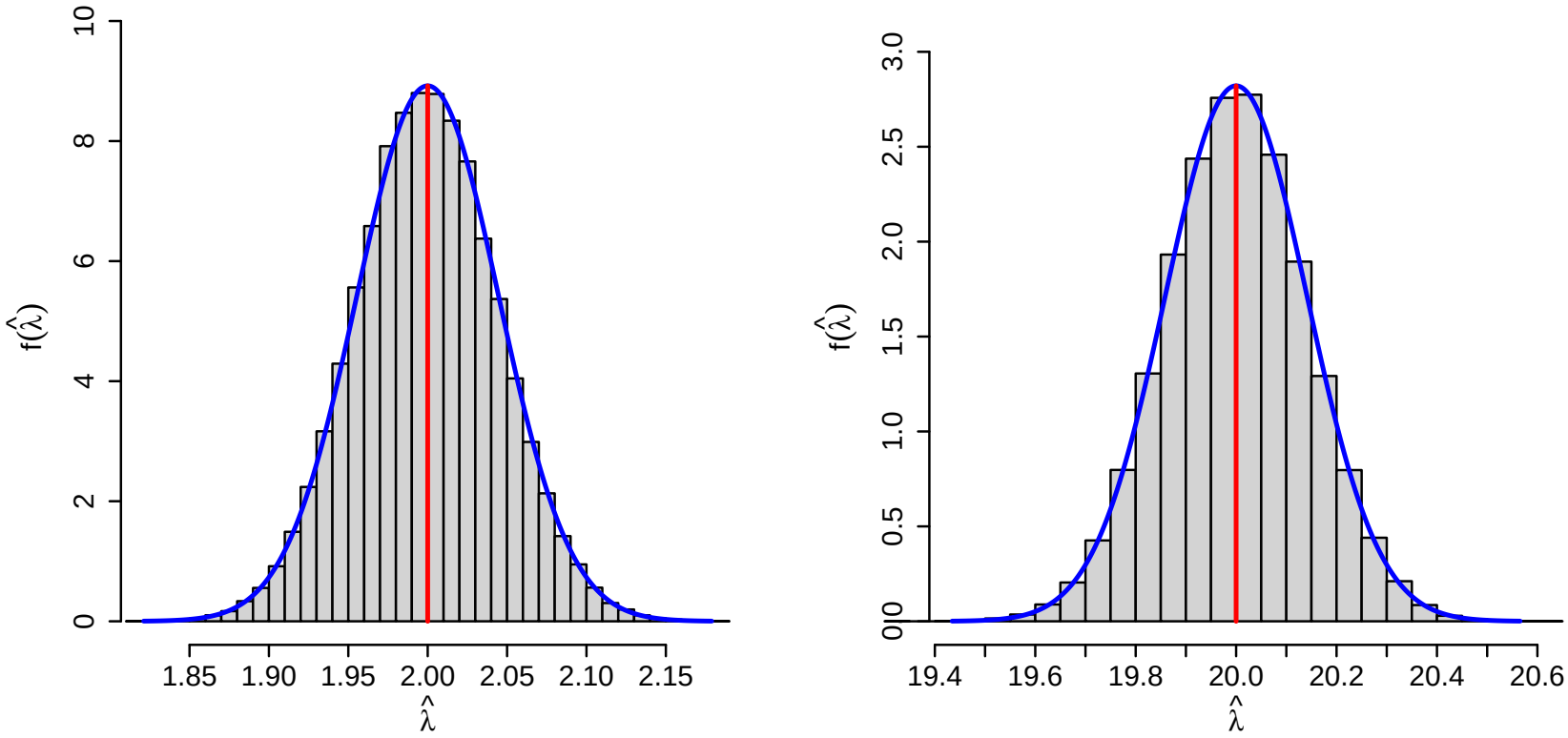


Figura: Distribuição amostral assintótica de $\hat{\pi}$ para $n = 1.000$ com $\lambda = 2$ na esquerda e $\lambda = 20$ na direita.

Ver exemplo_13.r

A estimação bayesiana (assintótica)

Suponha que a distribuição a priori de θ pode ser representada por uma função de densidade de probabilidade positiva e diferenciável sob o intervalo Ω e que o tamanho da amostra n é grande. Então, sob certas condições,

$$(\theta|\boldsymbol{x}) \overset{a}{\sim} \mathcal{N}\left(\hat{\theta};\ [nI(\hat{\theta})]^{-1}\right).$$

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\sigma^2) \sim \mathcal{N}(0;\ \sigma^2)$. Suponha que $f(\sigma) > 0$ e diferenciável para $\sigma > 0$. Então,

$$(\sigma|\boldsymbol{x}) \overset{a}{\sim} \mathcal{N}\left(\hat{\sigma};\ \frac{\hat{\sigma}^2}{2n}\right).$$

Informação de Fisher para múltiplos parâmetros

Definição

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\boldsymbol{\theta})$ com $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ pertencente a um subconjunto aberto Ω do espaço real de dimensão k. Seja $f(\boldsymbol{x}|\boldsymbol{\theta})$ a função (de densidade) de probabilidade conjunta de $\boldsymbol{X}$. Defina

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \ln f(\boldsymbol{x}|\boldsymbol{\theta}).$$

Assuma que o conjunto $\boldsymbol{x}$ para o qual $f(\boldsymbol{x}|\boldsymbol{\theta}) > 0$ é o mesmo para todo $\boldsymbol{\theta}$ e que $\ln f(\boldsymbol{x}|\boldsymbol{\theta})$ é duas vezes diferenciável com respeito a $\boldsymbol{\theta}$.

A **matriz de informação de Fisher $\boldsymbol{I}(\theta)$** em uma amostra aleatória $\boldsymbol{X}$ é definida como uma matriz $k \times k$ com elementos (r, s) dados por

$$I_{rs}(\boldsymbol{\theta}) = \text{Cov}\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_r}, \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_s} \middle| \boldsymbol{\theta} \right).$$

Observação: ainda vale o teorema da aproximação assintótica normal.

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$ tal que $\boldsymbol{\theta} = (\mu, \sigma^2)$. Temos

$$
\begin{aligned}
f(\boldsymbol{x}|\mu, \sigma^2) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\
\ell(\mu, \sigma^2|\boldsymbol{x}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{x})}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \\
\mathrm{E}\left(\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{X})}{\partial \mu}\right) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{E}(X_i - \mu) = 0 \\
\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{x})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\mathrm{E}\left(\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{X})}{\partial \sigma^2}\right) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}\mathrm{E}\left(\sum_{i=1}^{n}\left[\frac{X_i - \mu}{\sigma}\right]^2\right) = 0,
\end{aligned}
$$

pois $Z_i = \dfrac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$, $Z_i^2 \sim \chi_1^2$ e $\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$.

Portanto, temos que

$$I_{rs}(\boldsymbol{\theta}) = \text{Cov}\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_r}, \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_s}\middle|\boldsymbol{\theta}\right) = \text{E}\left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_r} \times \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_s}\middle|\boldsymbol{\theta}\right).$$

Agora,

$$\begin{aligned}
I_{11}(\mu, \sigma^2) &= \text{E}\left(\left[\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{X})}{\partial \mu}\right]^2\right) = \text{Var}\left(\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{X})}{\partial \mu}\right) \\
&= \frac{1}{\sigma^2}\text{Var}\left(\sum\nolimits_{i=1}^{n}\left[\frac{X_i - \mu}{\sigma}\right]\right) = \frac{n}{\sigma^2} \\
I_{22}(\mu, \sigma^2) &= \text{E}\left(\left[\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{X})}{\partial \sigma^2}\right]^2\right) = \text{Var}\left(\frac{\partial \ell(\mu, \sigma^2|\boldsymbol{X})}{\partial \sigma^2}\right) \\
&= \frac{1}{4\sigma^4}\text{Var}\left(\sum\nolimits_{i=1}^{n}\left[\frac{X_i - \mu}{\sigma}\right]^2\right) = \frac{2n}{4\sigma^4} = \frac{n}{2\sigma^4}.
\end{aligned}$$

$$
\begin{aligned}
I_{12}(\mu, \sigma^2) &= \mathrm{E}\left(\frac{\partial \ell(\mu, \sigma^2 | \boldsymbol{X})}{\partial \mu} \times \frac{\partial \ell(\mu, \sigma^2 | \boldsymbol{X})}{\partial \sigma^2} \right) \\
&= \mathrm{E}\left(\left\{ \frac{1}{\sigma^2} \sum_{i=1}^{n} [X_i - \mu] \right\} \left\{ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} [X_i - \mu]^2 \right\} \right) \\
&= -\frac{n}{2\sigma^4} \mathrm{E}\left(\sum_{i=1}^{n} [X_i - \mu] \right) + \frac{1}{2\sigma^6} \mathrm{E}\left(\sum_{i=1}^{n} [X_i - \mu]^3 \right) = 0.
\end{aligned}
$$

Portanto,

$$
\boldsymbol{I}_n(\mu, \sigma^2) = \begin{bmatrix} \dfrac{n}{\sigma^2} & 0 \\ 0 & \dfrac{n}{2\sigma^4} \end{bmatrix}.
$$

Assim, a distribuição assintótica do estimador de máxima verossimilhança de (μ, σ^2) é dado por

$$
\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}; \begin{bmatrix} \dfrac{\sigma^2}{n} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{bmatrix} \right).
$$

A matriz de informação de Fisher será utilizada em outras disciplinas.

Informação de Fisher

- Todos os resultados para a versão unidimensional têm versões para o caso de dimensão k.
- $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$.
- Defina a função escore ("vetor gradiente aleatório")

$$\boldsymbol{u} = \boldsymbol{u}(\theta, \boldsymbol{X}) = \left(\frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_1}, \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_2}, \ldots, \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_k} \right)^\top.$$

  Então, $\boldsymbol{I}_n(\boldsymbol{\theta}) = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top)$.

- Defina

$$I^\star_{rs}(\boldsymbol{\theta}) = \mathrm{E}\left(-\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{X})}{\partial \theta_r \partial \theta_s} \middle| \boldsymbol{\theta} \right),$$

  o negativo do valor esperado das derivadas segundas parciais. Então,

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = \mathrm{E}(\boldsymbol{u}\boldsymbol{u}^\top) = \begin{bmatrix} I^\star_{11}(\boldsymbol{\theta}) & I^\star_{12}(\boldsymbol{\theta}) & \cdots & I^\star_{1k}(\boldsymbol{\theta}) \\ I^\star_{21}(\boldsymbol{\theta}) & I^\star_{22}(\boldsymbol{\theta}) & \cdots & I^\star_{2k}(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots \\ I^\star_{k1}(\boldsymbol{\theta}) & I^\star_{k2}(\boldsymbol{\theta}) & \cdots & I^\star_{kk}(\boldsymbol{\theta}) \end{bmatrix}$$

- Seja $T = r(\boldsymbol{X})$ uma estatística com variância finita e seja $m(\boldsymbol{\theta}) = \mathrm{E}(T|\boldsymbol{\theta})$. Defina

$$\boldsymbol{m}(\boldsymbol{\theta}) = \left(\frac{\partial m(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial m(\boldsymbol{\theta})}{\partial \theta_2}, \ldots, \frac{\partial m(\boldsymbol{\theta})}{\partial \theta_k} \right)^\top.$$

  Suponha que $\boldsymbol{I}_n(\boldsymbol{\theta})$ seja uma matriz inversível. Então, a desigualdade de Cramér-Rao é dada por

$$\mathrm{Var}(T|\boldsymbol{\theta}) \geqslant \boldsymbol{m}(\boldsymbol{\theta})^\top \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) \boldsymbol{m}(\boldsymbol{\theta}).$$

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha, \beta) \sim \mathcal{G}(\alpha;\ \beta)$. Defina $\psi(\alpha) = \ln \Gamma(\alpha)$. Temos que

$$
\begin{aligned}
f(x|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} \\
\ell(\alpha,\beta) &= \ln f(x|\alpha,\beta) = \alpha \ln \beta - \psi(\alpha) + (\alpha - 1) \ln x - \beta x \\
\frac{\partial \ell(\alpha,\beta)}{\partial \alpha} &= \ln \beta - \psi'(\alpha) + \ln x \\
\frac{\partial \ell(\alpha,\beta)}{\partial \beta} &= \frac{\alpha}{\beta} - x \\
\frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha^2} &= -\psi''(\alpha) = -\frac{(\Gamma''(\alpha)\Gamma(\alpha) - [\Gamma'(\alpha)]^2)}{[\Gamma(\alpha)]^2} \\
\frac{\partial^2 \ell(\alpha,\beta)}{\partial \beta^2} &= -\frac{\alpha}{\beta^2} \\
\frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha \partial \beta} &= \frac{1}{\beta}.
\end{aligned}
$$

Assim, a matriz de informação de Fisher é dada por

$$\boldsymbol{I}_n(\alpha,\beta) = n \begin{bmatrix} \psi'(\alpha) & -\dfrac{1}{\beta} \\ -\dfrac{1}{\beta} & \dfrac{\alpha}{\beta^2} \end{bmatrix}.$$

Segue-se que a distribuição assintótica do estimador de máxima verossimilhança de (α,β) é dado por

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}; \frac{1}{n(\alpha\psi'(\alpha)-1)} \begin{bmatrix} \alpha & \beta \\ \beta & \beta^2\psi'(\alpha) \end{bmatrix} \right).$$

Podemos utilizar os conhecimentos adquiridos para calcular $\mathrm{E}(\ln X|\alpha,\beta)$.

Note que $\mathrm{E}\left(\dfrac{\partial \ell(\alpha,\beta)}{\partial \alpha}\right) = 0$.

Logo, $\ln\beta - \psi'(\alpha) + \mathrm{E}(\ln X|\alpha,\beta) = 0$.

Portanto, $\mathrm{E}(\ln X|\alpha,\beta) = \psi'(\alpha) - \ln\beta$.

Inferência bayesiana: distribuição a priori

Definição (distribuição a priori)

Suponha um modelo estatístico indexado por um parâmetro θ. Se θ for tratado como uma variável aleatória, então a distribuição de probabilidade de θ antes de observar as realizações das outras variáveis aleatórias de interesse é chamada de **distribuição a priori**. Notação: $f(\theta)$.

Ver página 7

Definição (extensão para vetor de parâmetros)

Suponha um modelo estatístico indexado por um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Se $\boldsymbol{\theta}$ for tratado como um vetor aleatório, então a distribuição de probabilidade de $\boldsymbol{\theta}$ antes de observar as realizações das outras variáveis aleatórias de interesse é chamada de **distribuição a priori**. Notação: $f(\boldsymbol{\theta})$.

Distribuição a priori conjugada

Definição

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória da distribuição de $(X|\theta)$ com função (de densidade) de probabilidade conjunta $f(\boldsymbol{x}|\theta)$. Seja Ψ uma família de possíveis distribuições sob o espaço paramétrico Ω. Suponha que, independente da escolha da distribuição a priori, escolhemos $f(\theta)$ da família Ψ. Se independente do tamanho amostral n e dos valores observados $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ a distribuição a posteriori $f(\theta|\boldsymbol{x})$ pertencer a família Ψ, então Ψ é chamada de família de distribuição a **priori conjugada** para amostras de $f(\boldsymbol{x}|\theta)$. Diz-se também que Ψ é fechada sob amostragem das distribuições $f(\boldsymbol{x}|\theta)$.

Ver página 22

Distribuição a priori conjugada

Definição (extensão para vetor de parâmetros)

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória da distribuição de $(X|\boldsymbol{\theta})$ com função (de densidade) de probabilidade conjunta $f(\boldsymbol{x}|\boldsymbol{\theta})$ tal que $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ e $\boldsymbol{\theta} \in \Omega$. Seja Ψ uma família de possíveis distribuições sob o espaço paramétrico Ω. Suponha que, independente da escolha da distribuição a priori, escolhemos $f(\boldsymbol{\theta})$ da família Ψ. Se independente do tamanho amostral n e dos valores observados $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ a distribuição a posteriori $f(\boldsymbol{\theta}|\boldsymbol{x})$ pertencer a família Ψ, então Ψ é chamada de família de distribuição a **priori conjugada** para amostras de $f(\boldsymbol{x}|\boldsymbol{\theta})$. Diz-se também que Ψ é fechada sob amostragem das distribuições $f(\boldsymbol{x}|\boldsymbol{\theta})$.

Distribuição a priori imprópria

Definição

Seja f uma função não negativa cujo domínio inclui o espaço paramétrico de um modelo estatístico. Suponha que

$$\int_{\Omega} f(\theta) d\theta = \infty.$$

Se utilizarmos $f(\theta)$ como distribuição a priori para θ, então $f(\theta)$ **é uma distribuição a priori imprópria para** θ.

Ver página 32

Distribuição a priori imprópria

Definição (extensão para vetor de parâmetros)

Seja f uma função não negativa cujo domínio inclui o espaço paramétrico de um modelo estatístico indexado por um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Suponha que

$$\int\int \cdots \int_{\Omega} f(\boldsymbol{\theta}) d\theta_1 d\theta_2 \ldots d\theta_k = \infty.$$

Se utilizarmos $f(\boldsymbol{\theta})$ como distribuição a priori para $\boldsymbol{\theta}$, então $f(\theta)$ **é uma distribuição a priori imprópria para** θ.

Distribuição a priori objetiva

- Informação a priori subjetiva pode não ser aceitável em um contexto científico.
- Inferência completamente objetiva ao invés da utilização de distribuição a priori subjetiva.
- Abordaremos o conceito de distribuição a priori objetiva.
- Uma primeira ideia foi utilizar distribuição uniforme.

Distribuição a priori (objetiva) de Jeffreys

Definição

Seja X uma variável aleatória com função (de densidade) de probabilidade $f(x|\boldsymbol{\theta})$ tal que $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ e $\boldsymbol{\theta} \in \Omega$. A distribuição a priori (objetiva) de Jeffreys tem função (de densidade) de probabilidade dada por

$$f(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}, \quad \text{para} \quad \boldsymbol{\theta} \in \Omega,$$

em que $I(\boldsymbol{\theta})$ é a matriz de informação de Fisher.

Muitas vezes nos referimos somente a **distribuição a priori de Jeffreys**.

Uma representação alternativa é dada por

$$f(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}, \quad \text{para} \quad \boldsymbol{\theta} \in \Omega,$$

em que $|\boldsymbol{A}| = \det \boldsymbol{A}$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$. Temos

$$
\begin{aligned}
L(\pi) &= \prod_{i=1}^{n} \pi^{x_i}(1-\pi)^{1-x_i} = \pi^{n\overline{x}}(1-\pi)^{n(1-\overline{x})} \\
\ell(\pi) &= n\overline{x}\ln\pi + n(1-\overline{x})\ln(1-\pi) \\
\frac{\partial \ell(\pi)}{\partial \pi} &= \frac{n\overline{x}}{\pi} - \frac{n(1-\overline{x})}{1-\pi}, \qquad \frac{\partial \ell(\pi)}{\partial \pi} = 0 \Rightarrow \widehat{\pi} = \overline{x} \\
\frac{\partial^2 \ell(\pi)}{\partial \pi^2} &= -\frac{n\overline{x}}{\pi^2} - \frac{n(1-\overline{x})}{(1-\pi)^2} < 0, \quad \widehat{\pi} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{\partial^2 \ell(\pi)}{\partial \pi^2}\right) &= \frac{n\mathrm{E}(\overline{X})}{\pi^2} + \frac{n(1-\mathrm{E}(\overline{X}))}{(1-\pi)^2} = \frac{n}{\pi(1-\pi)}.
\end{aligned}
$$

A distribuição a priori de Jeffreys e a respectiva posteriori são dadas por

$$
\begin{aligned}
f(\pi) &\propto \pi^{-1/2}(1-\pi)^{-1/2}, \quad \text{ou} \quad \pi \sim \mathcal{BT}(1/2;\ 1/2), \quad \text{e} \\
(\pi|\boldsymbol{x}) &\sim \mathcal{BT}\left(n\overline{x} + \frac{1}{2};\ n(1-\overline{x}) + \frac{1}{2}\right).
\end{aligned}
$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BN}(m, \pi)$. Temos

$$
\begin{aligned}
L(\pi) &= \pi^{n\overline{x}}(1-\pi)^{n(m-\overline{x})} \prod_{i=1}^{n} \frac{m!}{x_i!(m-x_i)!} \\
\ell(\pi) &= n\overline{x} \ln \pi + n(m-\overline{x}) \ln(1-\pi) + \sum\nolimits_{i=1}^{n} \ln \left[\frac{m!}{x_i!(m-x_i)!} \right] \\
\frac{\partial \ell(\pi)}{\partial \pi} &= \frac{n\overline{x}}{\pi} - \frac{n(m-\overline{x})}{1-\pi}, \qquad \frac{\partial \ell(\pi)}{\partial \pi} = 0 \Rightarrow \widehat{\pi} = \frac{\overline{x}}{m} \\
\frac{\partial^2 \ell(\pi)}{\partial \pi^2} &= -\frac{n\overline{x}}{\pi^2} - \frac{n(m-\overline{x})}{(1-\pi)^2} < 0, \quad \widehat{\pi} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{\partial^2 \ell(\pi)}{\partial \pi^2} \right) &= \frac{n\mathrm{E}(\overline{X})}{\pi^2} + \frac{n(m-\mathrm{E}(\overline{X}))}{(1-\pi)^2} = \frac{nm}{\pi(1-\pi)}.
\end{aligned}
$$

A distribuição a priori de Jeffreys e a respectiva posteriori são dadas por

$$
\begin{aligned}
f(\pi) &\propto \pi^{-1/2}(1-\pi)^{-1/2}, \quad \text{ou} \quad \pi \sim \mathcal{BT}(1/2;\ 1/2), \quad \text{e} \\
(\pi|\boldsymbol{x}) &\sim \mathcal{BT}\left(n\overline{x} + \frac{1}{2};\ n(m-\overline{x}) + \frac{1}{2} \right).
\end{aligned}
$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{GM}(\pi)$. Temos

$$
\begin{aligned}
L(\pi) &= \prod_{i=1}^{n} \pi(1-\pi)^{x_i - 1} = \pi^n (1-\pi)^{n(\overline{x}-1)} \\
\ell(\pi) &= n \ln \pi + n(\overline{x}-1) \ln(1-\pi) \\
\frac{\partial \ell(\pi)}{\partial \pi} &= \frac{n}{\pi} - \frac{n(\overline{x}-1)}{1-\pi}, \qquad \frac{\partial \ell(\pi)}{\partial \pi} = 0 \Rightarrow \widehat{\pi} = \frac{1}{\overline{x}} \\
\frac{\partial^2 \ell(\pi)}{\partial \pi^2} &= -\frac{n}{\pi^2} - \frac{n(\overline{x}-1)}{(1-\pi)^2} < 0, \quad \widehat{\pi} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{\partial^2 \ell(\pi)}{\partial \pi^2}\right) &= \frac{n}{\pi^2} + \frac{n(\mathrm{E}(\overline{X})-1)}{(1-\pi)^2} = \frac{n}{\pi^2(1-\pi)}.
\end{aligned}
$$

A distribuição a priori de Jeffreys e a respectiva posteriori são dadas por

$$
\begin{aligned}
f(\pi) &\propto \pi^{-1}(1-\pi)^{-1/2}, \quad \text{(distribuição a priori imprópria)}, \quad \text{e} \\
(\pi|\boldsymbol{x}) &\sim \mathcal{BT}\left(n;\, n(\overline{x}-1) + \frac{1}{2}\right).
\end{aligned}
$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{P}(\lambda)$. Temos

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{n} \left[\frac{\exp\{-\lambda\}\lambda^{x_i}}{x_i!} \right] = \frac{\exp\{-n\lambda\}\lambda^{n\overline{x}}}{\prod_{i=1}^{n} \Gamma(x_i+1)} \\
\ell(\lambda) &= -n\lambda + n\overline{x} \ln \lambda - \sum_{i=1}^{n} \ln \Gamma(x_i+1) \\
\frac{\partial \ell(\lambda)}{\partial \lambda} &= -n + \frac{n\overline{x}}{\lambda}, \qquad \frac{\partial \ell(\lambda)}{\partial \lambda} = 0 \Rightarrow \widehat{\lambda} = \overline{x} \\
\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} &= -\frac{n\overline{x}}{\lambda^2} < 0, \quad \widehat{\lambda} \text{ é ponto de máximo} \\
\mathrm{E}\left(-\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} \right) &= \frac{n\mathrm{E}(\overline{X})}{\lambda^2} = \frac{n}{\lambda}.
\end{aligned}
$$

A distribuição a priori de Jeffreys e a respectiva posteriori são dadas por

$$
\begin{aligned}
f(\lambda) &\propto \lambda^{-1/2}, \quad \text{(distribuição a priori imprópria)}, \quad \text{e} \\
(\lambda|\boldsymbol{x}) &\sim \mathcal{G}\left(n\overline{x} + \frac{1}{2};\, n \right).
\end{aligned}
$$

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha, \beta) \sim \mathcal{G}(\alpha, \beta)$. Defina $\psi(\alpha) = \ln \Gamma(\alpha)$. Temos que

$$
\begin{aligned}
f(x|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} \\
\ell(\alpha,\beta) &= \alpha \ln \beta - \psi(\alpha) + (\alpha - 1) \ln x - \beta x \\
\frac{\partial \ell(\alpha,\beta)}{\partial \alpha} &= \ln \beta - \psi'(\alpha) + \ln x \\
\frac{\partial \ell(\alpha,\beta)}{\partial \beta} &= \frac{\alpha}{\beta} - x \\
\frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha^2} &= -\psi''(\alpha) = -\frac{(\Gamma''(\alpha)\Gamma(\alpha) - [\Gamma'(\alpha)]^2)}{[\Gamma(\alpha)]^2} \\
\frac{\partial^2 \ell(\alpha,\beta)}{\partial \beta^2} &= -\frac{\alpha}{\beta^2} \\
\frac{\partial^2 \ell(\alpha,\beta)}{\partial \alpha \partial \beta} &= \frac{1}{\beta}.
\end{aligned}
$$

Assim, a matriz de informação de Fisher é dada por

$$I(\alpha,\beta) = \begin{bmatrix} \psi''(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix}.$$

A distribuição a priori de Jeffreys é dada por

$$f(\alpha,\beta) \propto \frac{1}{\beta}\sqrt{\alpha\psi''(\alpha) - 1}, \quad \text{para} \quad \alpha > 0 \quad \text{e} \quad \beta > 0,$$

sendo uma distribuição a priori imprópria.

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$ tal que $\boldsymbol{\theta} = (\mu, \sigma^2)$. Então, a matriz de informação de Fisher é dada por

$$I_n(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.$$

Ver páginas 209, 210 e 211

Assim, a distribuição a priori de Jeffreys é dada por

$$f(\mu, \sigma^2) \propto \frac{1}{\sigma^3} \quad \text{ou} \quad f(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{3/2}},$$

sendo uma distribuição a priori imprópria.

Reparametrização: caso unidimensional

Definição

Se θ e ϕ são duas possíveis reparametrizações de um modelo estatístico, e θ é uma função contínua e diferenciável de ϕ, dizemos que a distribuição a priori $f(\theta)$ é **invariante** sob reparametrização se

$$f_\phi(\phi) = f_\theta(\theta) \left| \frac{d\theta}{d\phi} \right| .$$

Lema

A priori de Jeffreys é invariante sob reparametrização.

Prova

Note que

$$
\begin{aligned}
I_\theta(\theta) &= \mathrm{E}([\ell'(\theta|X)]^2) = \mathrm{E}\left(\left[\frac{d\ell(\theta|X)}{d\theta}\right]^2\right) \\
I_\phi(\phi) &= \mathrm{E}\left(\left[\frac{d\ell(\phi|X)}{d\phi}\right]^2\right) = \mathrm{E}\left(\left[\frac{d\ell(\theta(\phi)|X)}{d\theta} \times \frac{d\theta}{d\phi}\right]^2\right) \\
&= \mathrm{E}\left(\left[\frac{d\ell(\theta(\phi)|X)}{d\theta}\right]^2\right)\left(\frac{d\theta}{d\phi}\right)^2 = I_\theta(\theta)\left(\frac{d\theta}{d\phi}\right)^2.
\end{aligned}
$$

Como $f_\phi(\phi) \propto \sqrt{I_\phi(\phi)}$ e $f_\theta(\theta) \propto \sqrt{I_\theta(\theta)}$, segue-se que

$$
f_\phi(\phi) = f_\theta(\theta)\left|\frac{d\theta}{d\phi}\right|.
$$

□

Reparametrização: caso multidimensional

Definição

Se $\boldsymbol{\theta}$ e ϕ são duas possíveis reparametrizações de um modelo estatístico, e $\boldsymbol{\theta}$ é uma função contínua e diferenciável de ϕ, dizemos que a distribuição a priori $f(\boldsymbol{\theta})$ é **invariante** sob reparametrização se

$$f_{\phi}(\phi) = f_{\boldsymbol{\theta}}(\boldsymbol{\theta})|\det \boldsymbol{J}|,$$

em que $\boldsymbol{J}$ é a matriz jacobiana da transformação com entradas

$$J_{ik} = \frac{\partial \theta_i}{\partial \phi_k}.$$

Lema

A priori de Jeffreys é invariante sob reparametrização.

Prova

Para o caso multidimensional (vetor de parâmetros), temos que

$$I_{\phi}(\phi) = \boldsymbol{J}^{\top} I_{\theta}(\boldsymbol{\theta}) \boldsymbol{J},$$

tal que

$$\det I_{\phi}(\phi) = [\det I_{\theta}(\boldsymbol{\theta})] \times [\det \boldsymbol{J}]^2.$$

Como $f_{\phi}(\phi) \propto \sqrt{\det I_{\phi}(\phi)}$ e $f_{\theta}(\boldsymbol{\theta}) \propto \sqrt{\det I_{\theta}(\boldsymbol{\theta})}$, segue-se que

$$f_{\phi}(\phi) = f_{\theta}(\boldsymbol{\theta}) |\det \boldsymbol{J}|.$$

□

Análise bayesiana de uma amostra da distribuição normal

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$.

A função de verossimilhança é dada por

$$f(\boldsymbol{x}|\mu, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

Vimos que

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \mu)^2 &= \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - \mu)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 - 2(\mu - \overline{x})\sum_{i=1}^{n}(x_i - \overline{x}) + n(\mu - \overline{x})^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\mu - \overline{x})^2,
\end{aligned}$$

pois

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Inferência bayesiana: μ desconhecido e σ^2 conhecido

Neste caso, podemos escrever a função de verossimilhança como

$$f(\boldsymbol{x}|\mu) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu - \overline{x})^2\right\}.$$

A distribuição a priori conjugada é $\mathcal{N}(a;\, b^2)$ tal que

$$f(\mu) \propto \exp\left\{-\frac{1}{2b^2}(\mu - a)^2\right\}.$$

Para encontrar a distribuição a posteriori, temos que

$$\begin{aligned}
\frac{1}{(\sigma^2/n)}(\mu - \overline{x})^2 &= \frac{\mu^2}{\sigma^2/n} - 2\frac{\mu\overline{x}}{\sigma^2/n} + \frac{\overline{x}^2}{\sigma^2/n} \\
\frac{1}{b^2}(\mu - a)^2 &= \frac{\mu^2}{b^2} - 2\frac{\mu a}{b^2} + \frac{a^2}{b^2}.
\end{aligned}$$

Somando-se ambas as equações e dispensando as partes que não dependem de μ, temos

$$\mu^2\left[\frac{n}{\sigma^2} + \frac{1}{b^2}\right] - 2\mu\left[\frac{n\overline{x}}{\sigma^2} + \frac{a}{b^2}\right].$$

Agora, definimos

$$\tau^2 = [n\sigma^{-2} + b^{-2}]^{-1} \quad \text{e} \quad \xi = \tau^2 \left[\frac{n\overline{x}}{\sigma^2} + \frac{a}{b^2} \right].$$

Assim, temos que

$$\mu^2 \left[\frac{n}{\sigma^2} + \frac{1}{b^2} \right] - 2\mu \left[\frac{n\overline{x}}{\sigma^2} + \frac{a}{b^2} \right] = \frac{\mu^2}{\tau^2} - \frac{2\mu\xi}{\tau^2}.$$

Finalmente, completamos quadrados para obter

$$\frac{\mu^2}{\tau^2} - \frac{2\mu\xi}{\tau^2} + \frac{\xi^2}{\tau^2} = \frac{1}{\tau^2}(\mu - \xi)^2.$$

Note que o termo introduzido não depende de μ.

Assim, a distribuição a posteriori é dada por

$$f(\mu|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2\tau^2}(\mu-\xi)^2\right\}.$$

Em outras palavras, **a distribuição a posteriori** é dada por

$$(\mu|\boldsymbol{x}) \sim \mathcal{N}(\xi;\, \tau^2),$$

com

$$\tau^2 = [n\sigma^{-2} + b^{-2}]^{-1} \quad \text{e} \quad \xi = \tau^2\left[\frac{n\bar{x}}{\sigma^2} + \frac{a}{b^2}\right].$$

Se μ **desconhecido** e σ^2 **conhecido**, temos que

$$\begin{aligned}
f(\boldsymbol{x}|\mu) &= \text{cst} \times \exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\mu)^2\right\} \\
\ell(\mu|\boldsymbol{x}) &= \ln \text{cst} - \frac{n}{2\sigma^2}(\overline{x}-\mu)^2 \\
\frac{d\ell(\mu|\boldsymbol{x})}{d\mu} &= \frac{n}{\sigma^2}(\overline{x}-\mu) \\
I(\mu) &= \text{E}\left(\left[\frac{d\ell(\mu|\boldsymbol{X})}{d\mu}\right]^2\right) = \text{E}\left(\left[\frac{n}{\sigma^2}(\overline{X}-\mu)\right]^2\right) \\
&= \frac{n^2}{\sigma^4}\text{E}([\overline{X}-\mu]^2) = \frac{n^2\sigma^2}{n\sigma^4} = \frac{n}{\sigma^2}.
\end{aligned}$$

Assim, **a distribuição a priori de Jeffreys** é dada por

$$f(\mu) \propto 1 \quad \text{(uma distribuição a priori imprópria)},$$

e **a distribuição a posteriori** é dada por

$$(\mu|\boldsymbol{x}) \sim \mathcal{N}\left(\overline{x};\ \frac{\sigma^2}{n}\right).$$

Inferência bayesiana: μ **conhecido** e σ^2 **desconhecido**

Neste caso, podemos escrever a função de verossimilhança como

$$f(\boldsymbol{x}|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right]\right\}$$

A distribuição a priori conjugada é $\mathcal{IG}(a;\, b)$ tal que

$$f(\sigma^2) \propto (\sigma^2)^{-(a+1)} \exp\left\{-\frac{b}{\sigma^2}\right\}.$$

Assim, **a distribuição a posteriori** dada por

$$(\sigma^2|\boldsymbol{x}) \sim \mathcal{IG}\left(a+\frac{n}{2};\, b+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Se μ **conhecido** e σ^2 **desconhecido**, temos que

$$
\begin{aligned}
f(\boldsymbol{x}|\sigma^2) &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
\ell(\sigma^2|\boldsymbol{x}) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{d\ell(\sigma^2|\boldsymbol{x})}{d\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \\
I(\sigma^2) &= \text{Var}\left(\frac{d\ell(\sigma^2|\boldsymbol{X})}{d\sigma^2}\right) = \frac{1}{4\sigma^4}\text{Var}\left(\sum_{i=1}^{n}\left[\frac{X_i-\mu}{\sigma}\right]^2\right) = \frac{n}{2\sigma^4}.
\end{aligned}
$$

Assim, **a distribuição a priori de Jeffreys** é dada por

$$f(\sigma^2) \propto \frac{1}{\sigma^2} \quad \text{ou} \quad f(\sigma^2) \propto (\sigma^2)^{-1} \quad \text{(uma distribuição a priori imprópria)},$$

e **a distribuição a posteriori** é dada por

$$(\sigma^2|\boldsymbol{x}) \sim \mathcal{IG}\left(\frac{n}{2};\ \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Definição

O parâmetro de precisão ϕ de uma distribuição normal é definido como o recíproco da variância, isto é, $\phi = \frac{1}{\sigma^2}$.

Agora, podemos escrever a função de verossimilhança como

$$f(\boldsymbol{x}|\phi) \propto \phi^{n/2} \exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

A distribuição a priori conjugada é $\mathcal{G}(a, b)$ tal que

$$f(\phi) \propto \phi^{a-1} \exp\{-b\phi\}.$$

Assim, **a distribuição a posteriori** é dada por

$$(\phi|\boldsymbol{x}) \sim \mathcal{G}\left(a + \frac{n}{2};\ b + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

Se μ **conhecido** e ϕ **desconhecido**, temos que

$$\begin{aligned}
f(\boldsymbol{x}|\phi) &= (2\pi)^{-n/2}\phi^{n/2}\exp\left\{-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\
\ell(\phi|\boldsymbol{x}) &= -\frac{n}{2}\ln(2\pi)+\frac{n}{2}\ln\phi-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{d\ell(\phi|\boldsymbol{x})}{d\phi} &= \frac{n}{2\phi}-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{d^2\ell(\phi|\boldsymbol{x})}{d\phi^2} &= -\frac{n}{2\phi^2} \\
I(\phi) &= \mathrm{E}\left(-\frac{d^2\ell(\phi|\boldsymbol{X})}{d\phi^2}\right)=\frac{n}{2\phi^2}.
\end{aligned}$$

Assim, **a distribuição a priori de Jeffreys** é dada por

$$f(\phi)\propto\frac{1}{\phi} \quad \text{ou} \quad f(\phi)\propto\phi^{-1} \quad \text{(uma distribuição a priori imprópria)},$$

e **a distribuição a posteriori** é dada por

$$(\phi|\boldsymbol{x})\sim\mathcal{G}\left(\frac{n}{2};\ \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Note que

$$I_\phi(\phi) = I_{\sigma^2}(\sigma^2(\phi)) \left(\frac{d\sigma^2}{d\phi}\right)^2 \quad \text{e} \quad f_\phi(\phi) = f_{\sigma^2}(\sigma^2(\phi)) \left|\frac{d\sigma^2}{d\phi}\right|.$$

Em outras palavras, vimos que **a priori de Jeffreys é invariante sob reparametrização.**

Considere $\phi = \dfrac{1}{\sigma^2} \quad \Leftrightarrow \quad \sigma^2 = \dfrac{1}{\phi}$ tal que $\dfrac{d\sigma^2}{d\phi} = -\dfrac{1}{\phi^2}$.

Assim,

$$f_\phi(\phi) = f_{\sigma^2}\left(\frac{1}{\phi}\right)\left|\frac{1}{\phi^2}\right| = \phi \times \frac{1}{\phi^2} = \frac{1}{\phi},$$

isto é,

$$f(\phi) \propto \frac{1}{\phi} \quad \text{ou} \quad f(\phi) \propto \phi^{-1} \quad \text{(uma distribuição a priori imprópria)}.$$

Distribuição normal-gama e normal-inversa gama

Definição

Sejam μ e ϕ duas variáveis aleatórias. Suponha que

$$(\mu|\phi) \sim \mathcal{N}(a;\, [b\phi]^{-1}) \quad \text{e} \quad \phi \sim \mathcal{G}(c;\, d),$$

tal que $b\phi$ seja a precisão da normal. Então, dizemos que a distribuição conjunta de (μ, ϕ) é uma **distribuição normal-gama com parâmetros** a, b, c **e** d**.**

Definição

Sejam μ e σ^2 duas variáveis aleatórias. Suponha que

$$(\mu|\sigma^2) \sim \mathcal{N}(a;\, b^{-1}\sigma^2) \quad \text{e} \quad \sigma^2 \sim \mathcal{IG}(c;\, d).$$

Então, dizemos que a distribuição conjunta de (μ, σ^2) é uma **distribuição normal-inversa gama com parâmetros** a, b, c **e** d**.**

Inferência bayesiana: μ e σ^2 **desconhecidos**

Neste caso, podemos escrever a função de verossimilhança como

$$f(\boldsymbol{x}|\mu,\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\mu-\overline{x})^2\right]\right\}.$$

A distribuição a priori conjugada é $(\mu|\sigma^2) \sim \mathcal{N}(a;\, b^{-1}\sigma^2)$ **e** $\sigma^2 \sim \mathcal{IG}(c/2;\, d/2)$, isto é,

$$\begin{aligned} f(\mu,\sigma^2) &\propto (\sigma^2)^{-1/2} \exp\left\{-\frac{b}{2\sigma^2}(\mu-a)^2\right\} \times (\sigma^2)^{-(c/2+1)} \exp\left\{-\frac{d}{2\sigma^2}\right\} \\ &= (\sigma^2)^{-(c/2+1/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[d + b(\mu-a)^2\right]\right\}. \end{aligned}$$

Assim, **a distribuição a posteriori** é dada por

$$\begin{aligned} f(\mu,\sigma^2|\boldsymbol{x}) &\propto (\sigma^2)^{-(c/2+n/2+1/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[d + (n-1)s^2\right]\right\} \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left[b(\mu-a)^2 + n(\mu-\overline{x})^2\right]\right\}. \end{aligned}$$

Agora, temos que

$$\begin{aligned} f(\mu|\sigma^2, \boldsymbol{x}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[b(\mu-a)^2 + n(\mu-\overline{x})^2\right]\right\} \\ &\propto \exp\left\{-\frac{b^\star}{2\sigma^2}(\mu-a^\star)^2\right\}, \end{aligned}$$

isto é, $(\mu|\sigma^2, \boldsymbol{x}) \sim \mathcal{N}(a^\star, [b^\star]^{-1}\sigma^2)$ com

$$b^\star = b + n \quad \text{e} \quad a^\star = \frac{ba + n\overline{x}}{b+n}.$$

Além disso,

$$(\sigma^2|\mu, \boldsymbol{x}) \sim \mathcal{IG}\left(\frac{n+c+1}{2};\ \frac{1}{2}\left[d + (n-1)s^2 + b(\mu-a)^2 + n(\mu-\overline{x})^2\right]\right).$$

Para os cálculos a seguir, considere

$$g(\sigma^2) = (\sigma^2)^{-(c/2+n/2+1/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[d + (n-1)s^2\right]\right\}.$$

Agora, calcularemos a distribuição marginal de σ^2 a posteriori. Então,

$$
\begin{aligned}
f(\sigma^2|\boldsymbol{x}) &= \int_{-\infty}^{\infty} f(\mu, \sigma^2|\boldsymbol{x}) d\mu \\
&\propto g(\sigma^2) \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2\sigma^2} \left[b(\mu - a)^2 + n(\mu - \overline{x})^2 \right] \right\} d\mu \\
&= g(\sigma^2) \int_{-\infty}^{\infty} \exp\left\{ -\frac{b^\star}{2\sigma^2} (\mu - a^\star)^2 \right\} d\mu \\
&\times \exp\left\{ -\frac{1}{2\sigma^2} \left[ba^2 + n\overline{x}^2 - (a^\star)^2 \right] \right\} \\
&\propto g(\sigma^2) \times (\sigma^2)^{1/2} \times \exp\left\{ -\frac{1}{2\sigma^2} \left[ba^2 + n\overline{x}^2 - (a^\star)^2 \right] \right\} \\
&= (\sigma^2)^{-(c/2+n/2+1)} \\
&\times \exp\left\{ -\frac{1}{2\sigma^2} \left[d + (n-1)s^2 + ba^2 + n\overline{x}^2 - (a^\star)^2 \right] \right\},
\end{aligned}
$$

tal que

$$(\sigma^2|\boldsymbol{x}) \sim \mathcal{IG}(c^\star/2;\, d^\star/2),$$

com $c^\star = c + n$ e $d^\star = d + (n-1)s^2 + ba^2 + n\overline{x}^2 - (a^\star)^2$.

Assim, temos que

$$(\sigma^2|\boldsymbol{x}) \sim \mathcal{IG}(c^\star/2;\, d^\star/2) \quad \text{e} \quad (\mu|\sigma^2, \boldsymbol{x}) \sim \mathcal{N}(a^\star;\, [b^\star]^{-1}\sigma^2),$$

com

$$\begin{aligned} a^\star &= \frac{ba + n\bar{x}}{b+n} \\ b^\star &= b + n \\ c^\star &= c + n \\ d^\star &= d + (n-1)s^2 + ba^2 + n\bar{x}^2 - (a^\star)^2. \end{aligned}$$

Concluímos que $(\mu, \sigma^2|\boldsymbol{x})$ tem distribuição normal-inversa gama com parâmetros $a^\star$, $b^\star$, $c^\star/2$ e $d^\star/2$.

Agora, calcularemos a distribuição marginal de μ a posteriori. Então,

$$
\begin{aligned}
f(\mu|\boldsymbol{x}) &= \int_0^{\infty} f(\mu, \sigma^2|\boldsymbol{x}) d\sigma^2 \\
&\propto \int_0^{\infty} (\sigma^2)^{-(c/2+n/2+1/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[d + (n-1)s^2\right]\right\} \\
&\times \exp\left\{-\frac{b^\star}{2\sigma^2}(\mu - a^\star)^2\right\} \\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left[ba^2 + n\overline{x}^2 - (a^\star)^2\right]\right\} d\sigma^2 \\
&\propto \left[d^\star + b^\star(\mu - a^\star)^2\right]^{-\frac{c^\star+1}{2}} \propto \left[1 + \frac{(\mu - a^\star)^2}{c^\star r^2}\right]^{-\frac{c^\star+1}{2}},
\end{aligned}
$$

tal que

$$(\mu|\boldsymbol{x}) \sim t_{c^\star}(a^\star;\, r^2),$$

com $r^2 = \dfrac{d^\star}{b^\star c^\star}$.

Inferência bayesiana: priori de Jeffreys para (μ, σ^2)

Já vimos que

$$I_n(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}, \quad f(\mu, \sigma^2) \propto \frac{1}{\sigma^3} \quad \text{ou} \quad f(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{3/2}}.$$

Como μ e σ^2 são **desconhecidos**, a função de verossimilhança é dada por

$$f(\boldsymbol{x}|\mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\mu - \overline{x})^2\right]\right\}.$$

Consequentemente, **a distribuição a posteriori** é dada por

$$f(\mu, \sigma^2|\boldsymbol{x}) \propto (\sigma^2)^{-(n/2+1/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\mu - \overline{x})^2\right]\right\}.$$

Daí, segue-se que

$$(\mu|\sigma^2, \boldsymbol{x}) \sim \mathcal{N}\left(\overline{x};\ \frac{\sigma^2}{n}\right) \text{ e } (\sigma^2|\mu, \boldsymbol{x}) \sim \mathcal{IG}\left(\frac{n+1}{2};\ \frac{(n-1)s^2 + n(\mu - \overline{x})^2}{2}\right).$$

Agora, calcularemos a distribuição marginal de μ a posteriori. Então,

$$
\begin{aligned}
f(\mu|\boldsymbol{x}) &= \int_0^{\infty} f(\mu, \sigma^2|\boldsymbol{x}) d\sigma^2 \\
&\propto \int_0^{\infty} (\sigma^2)^{-(n/2+1/2+1)} \exp\left\{ -\frac{1}{2\sigma^2} \left[(n-1)s^2 + n(\mu - \overline{x})^2 \right] \right\} d\sigma^2 \\
&\propto \left[(n-1)s^2 + n(\mu - \overline{x})^2 \right]^{-\frac{n+1}{2}} \propto \left[1 + \frac{(\mu - \overline{x})^2}{n(\hat{\sigma}^2/n)} \right]^{-\frac{n+1}{2}},
\end{aligned}
$$

tal que

$$
(\mu|\boldsymbol{x}) \sim t_n\left(\overline{x};\ \frac{\hat{\sigma}^2}{n}\right),
$$

com $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Agora, calcularemos a distribuição marginal de σ^2 a posteriori. Então,

$$
\begin{aligned}
f(\sigma^2|\boldsymbol{x}) &= \int_{-\infty}^{\infty} f(\mu, \sigma^2|\boldsymbol{x}) d\mu \\
&\propto \int_{-\infty}^{\infty} (\sigma^2)^{-(n/2+1/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\mu-\overline{x})^2\right]\right\} d\mu \\
&= (\sigma^2)^{-(n/2+1)} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \\
&\times \int_{-\infty}^{\infty} (\sigma^2)^{-1/2} \exp\left\{-\frac{n}{2\sigma^2}(\mu-\overline{x})^2\right\} d\mu \\
&= (\sigma^2)^{-(n/2+1)} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\},
\end{aligned}
$$

tal que

$$
(\sigma^2|\boldsymbol{x}) \sim \mathcal{IG}\left(\frac{n}{2}; \frac{(n-1)s^2}{2}\right) \quad \text{ou} \quad (\sigma^2|\boldsymbol{x}) \sim \mathcal{IG}\left(\frac{n}{2}; \frac{n\hat{\sigma}^2}{2}\right).
$$

Estimação por Mínimos Quadrados

Sejam $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ observações de uma variável aleatória X com média finita.

Suponha que o interesse seja estimar $\mathrm{E}(X) = \mu$.

Podemos utilizar o método dos momentos: $\hat{\mu} = \overline{x}$.

Por outro lado, podemos utilizar a **estimação por mínimos quadrados**. Defina

$$Q(\mu) = \sum_{i=1}^{n} (x_i - \mu)^2.$$

O objetivo é minimizar a soma dos quadrados das distâncias entre as observações e a média.

Então,

$$\frac{dQ(\mu)}{d\mu} = -2\sum_{i=1}^{n} (x_i - \mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \overline{x}.$$

Uma outra possibilidade é utilizar a **estimação por mínimos absolutos**, isto é,

$$A(\mu) = \sum_{i=1}^{n} |x_i - \mu| \quad \Rightarrow \quad \hat{\mu} = \mathrm{med}(x_1, x_2, \ldots, x_n).$$

Regressão linear

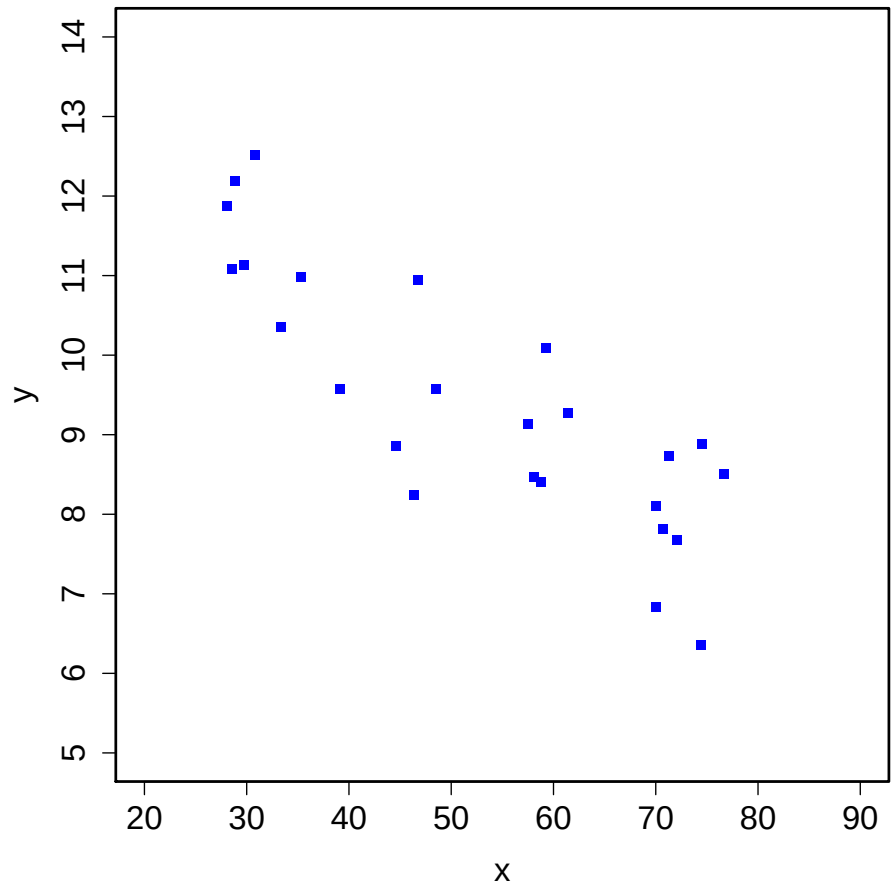


Figura: Relação entre x e y.

Regressão linear

O termo linear se refere aos parâmetros do modelo (os $\beta's$). Por exemplo,

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \varepsilon; \\ y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon; \text{ e} \\ y &= \beta_0 + \beta_1 \exp(x_1) + \beta_2 \log(x_2^2) + \varepsilon, \end{aligned}$$

são exemplos de modelos de regressão linear.

Em geral, tratamos ε e y como variáveis aleatórias enquanto $\boldsymbol{x}$ é dado, isto é, a análise é condicional ao conhecimento de $\boldsymbol{x}$ (regressoras).

Por hipótese, temos que a média de ε é zero.

Então, começaremos com o modelo de **regressão linear simples** dado por

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

O objetivo é obter a "melhor" equação da reta que descreve a relação entre x e y.

A ideia de mínimos quadrados

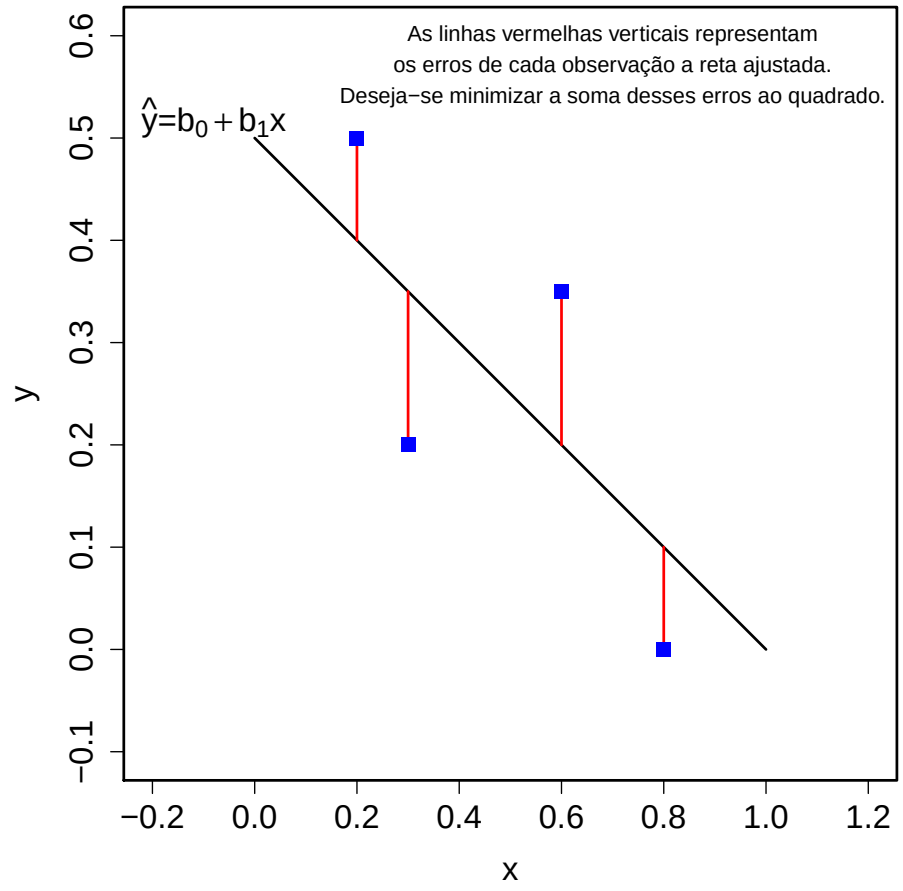


Figura: Ilustração da ideia de mínimos quadrados.

A equação estimada é denotada por $\widehat{y} = b_0 + b_1 x = \widehat{\beta}_0 + \widehat{\beta}_1 x$, em que $b_0 = \widehat{\beta}_0$ e $b_1 = \widehat{\beta}_1$ são **estimativas pontuais** de β_0 e β_1, respectivamente.

Mínimos quadrados

Suponha que uma amostra aleatória $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ de tamanho n seja obtida tal que

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

Definimos a função **soma de quadrados** como

$$Q = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Queremos minimizar esta soma de quadrados. Logo,

$$\begin{aligned} \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) , \\ \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) . \end{aligned}$$

Agora, precisamos resolver o sistema de equações dado por

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) &= 0, \\ \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) &= 0. \end{aligned}$$

Assim, temos que

$$\begin{aligned} \sum_{i=1}^{n} y_i - nb_0 - b_1 \sum_{i=1}^{n} x_i &= 0, \\ \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 &= 0, \end{aligned}$$

ou

$$\begin{aligned} nb_0 + b_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned}$$

Ambas as representações são chamadas de **equações normais** (perpendicular ou ortogonal).

A solução do sistema acima é dado por

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}, \quad \text{e} \\
b_0 &= \overline{y} - b_1\overline{x},
\end{aligned}
$$

em que

$$
\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \text{e}
$$

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) &= \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_i + n\overline{x}\,\overline{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \\
&= \sum_{i=1}^{n} x_i y_i - \frac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}.
\end{aligned}
$$

Utilizaremos as seguintes definições:

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) = \sum_{i=1}^{n}(x_i-\overline{x})y_i = \sum_{i=1}^{n}x_i(y_i-\overline{y}) \\
&= \sum_{i=1}^{n}x_iy_i - \frac{\left[\sum_{i=1}^{n}x_i\right]\left[\sum_{i=1}^{n}y_i\right]}{n} = \sum_{i=1}^{n}x_iy_i - n\overline{x}\,\overline{y}, \\
S_{xx} &= \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{i=1}^{n}(x_i-\overline{x})x_i \\
&= \sum_{i=1}^{n}x_i^2 - \frac{\left[\sum_{i=1}^{n}x_i\right]^2}{n} = \sum_{i=1}^{n}x_i^2 - n\overline{x}^2, \quad \text{e} \\
S_{yy} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i-\overline{y})y_i \\
&= \sum_{i=1}^{n}y_i^2 - \frac{\left[\sum_{i=1}^{n}y_i\right]^2}{n} = \sum_{i=1}^{n}y_i^2 - n\overline{y}^2.
\end{aligned}
$$

Assim, temos que $b_1 = \dfrac{S_{xy}}{S_{xx}}$ e $b_0 = \overline{y} - b_1\overline{x}$.

Então, agora é possível escrever a **equação ajustada** ou **predita** como

$$\widehat{y} = b_0 + b_1 x.$$

Agora, substituindo $b_0 = \overline{y} - b_1 \overline{x}$ na equação acima temos que

$$\widehat{y} = \overline{y} + b_1 (x - \overline{x}).$$

Observe que $x = \overline{x} \Rightarrow \widehat{y} = \overline{y}$.

Exemplo

i	y_i	x_i
1	10,98	35,3
2	11,13	29,7
3	12,51	30,8
4	8,40	58,8
5	9,27	61,4
6	8,73	71,3
7	6,36	74,4
8	8,50	76,7
9	7,82	70,7
10	9,14	57,5
11	8,24	46,4
12	12,19	28,9
13	11,88	28,1
14	9,57	39,1
15	10,94	46,8
16	9,58	48,5
17	10,09	59,3
18	8,11	70,0
19	6,83	70,0
20	8,88	74,5
21	7,68	72,1
22	8,47	58,1
23	8,86	44,6
24	10,36	33,4
25	11,08	28,6

$$
\begin{aligned}
n &= 25 \\
\sum_{i=1}^{25} y_i &= 10,98 + 11,13 + \cdots + 11,08 = 235,60 \\
\overline{y} &= \frac{235,60}{25} = 9,424 \\
\sum_{i=1}^{25} x_i &= 35,3 + 29,7 + \cdots + 28,6 = 1.315 \\
\overline{x} &= \frac{1.315}{25} = 52,60 \\
\sum_{i=1}^{25} x_i y_i &= (10,98)(35,3) + (11,13)(29,7) + \cdots + (11,08)(28,6) \\
&= 11.821,432 \\
\sum_{i=1}^{25} x_i^2 &= (35,3)^2 + (29,7)^2 + \cdots + (28,6)^2 = 76.323,42 \\
b_1 &= \frac{11.821,432 - \dfrac{(1315)(235,6)}{25}}{76.323,42 - \dfrac{(1315)^2}{25}} = \frac{-571,128}{7.154,42} = -0,079829
\end{aligned}
$$

A equação ajustada é dada por

$$\begin{aligned} \widehat{y} &= \overline{y} + b_1(x - \overline{x}) = 9,4240 - 0,079829(x - 52,60) \\ &= 13,623 - 0,079829x. \end{aligned}$$

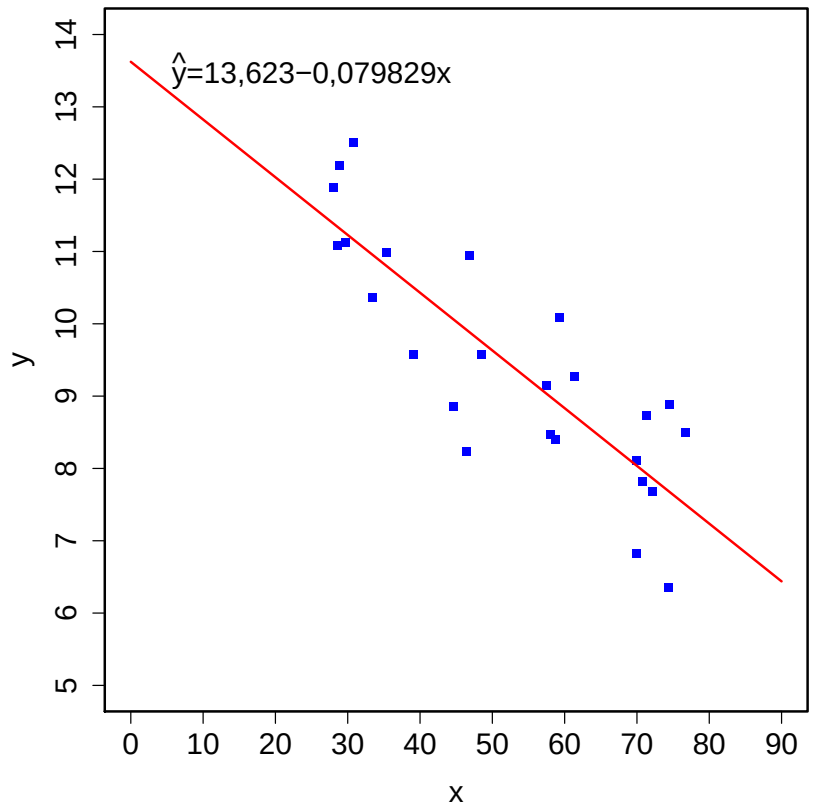


Figura: Equação da reta ajustada.

Mínimos quadrados em **redes neurais**

Considere um conjunto de valores $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$ e observações y_i, $y_i \in \{0; 1\}$, de uma variável aleatória Y (binária), para $i = 1, 2, \ldots, n$.

O objetivo é utilizar $(x_{1i}, x_{2i}, \ldots, x_{ki})$ para explicar a ocorrência de $Y_i = y_i$ tal que $y_i \in \{0; 1\}$ ou, em outros termos, explicar a $\Pr(Y_i = 1|\boldsymbol{x}_i)$.

Então, considere $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$.

Claramente, $\eta_i \in \mathbb{R}$ e $y_i \in \{0, 1\}$. Assim, considere, por exemplo, a **função de ativação**

$$\sigma(z) = \frac{\exp\{z\}}{1 + \exp\{z\}} \quad (\textbf{sigmoide}).$$

Utilizando a **função custo perda quadrática**, queremos minimizar

$$Q = Q(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \sigma(\eta_i))^2 .$$

Em redes neurais, existe uma sequência encaixada de estruturas como esta e a solução é feita de forma numérica (via gradiente descendente e derivada via retropropagação).

Distribuições da Família Exponencial

Definição

Uma família de distribuições de probabilidade que depende de um único parâmetro θ é chamada de **família exponencial** Υ se, e somente se, a função (de densidade) de probabilidade é dada por

$$f(x|\theta) = \exp\{a(\theta)s(x) + b(\theta) + c(x)\}$$

ou

$$f(x|\theta) = h(x)\exp\{a(\theta)s(x) + b(\theta)\}$$

em que

- $a(\theta)$ e $b(\theta)$ são funções de θ (somente) e conhecidas; e
- $s(x)$, $c(x)$ e $h(x)$ são funções de x (somente) e conhecidas.

Note que $h(x) \geqslant 0$ para todo $x \in \mathbb{R}$. No suporte de X, temos que $h(x) > 0$.

Exemplo

A família de distribuições exponencial com parâmetro λ é uma componente da família exponencial Υ. É fácil verificar que

$$f(x|\lambda) = \lambda \exp\{-\lambda x\} \boldsymbol{I}_{[0,\infty)}(x) = \exp\{-\lambda x + \ln(\lambda) + \ln(\boldsymbol{I}_{[0,\infty)}(x))\},$$

com

- $s(x) = x$;
- $a(\lambda) = -\lambda$;
- $b(\lambda) = \ln(\lambda)$; e
- $c(x) = \ln(\boldsymbol{I}_{[0,\infty)}(x))$ (poderia ser omitido e considerar $c(x) = 0$)
- Na notação alternativa, $h(x) = \boldsymbol{I}_{[0,\infty)}(x)$.

Exemplo

A família de distribuições Bernoulli com parâmetro π é uma componente da família exponencial Υ. Note que

$$f(x|\pi) = \pi^x(1-\pi)^{1-x} = \exp\left\{\ln\left(\frac{\pi}{1-\pi}\right)x + \ln(1-\pi)\right\},$$

com

- $s(x) = x$;
- $a(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$;
- $b(\pi) = \ln(1-\pi)$; e
- $c(x) = 0$ (ou $h(x) = 1$ na notação alternativa).

Exemplo

A família de distribuições Poisson com parâmetro λ é uma componente da família exponencial Υ. Podemos verificar que

$$f(x|\lambda) = \frac{\exp\{-\lambda\}\lambda^x}{x!} = \exp\{x \ln(\lambda) - \lambda - \ln \Gamma(x+1)\},$$

com

- $s(x) = x$;
- $a(\lambda) = \ln(\lambda)$;
- $b(\lambda) = -\lambda$; e
- $c(x) = -\ln \Gamma(x+1)$.
- Na notação alternativa, $h(x) = \frac{1}{x!} = \frac{1}{\Gamma(x+1)}$.

Teorema

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma variável aleatória X com função (de densidade) de probabilidade $f(x|\theta)$. Se X tem distribuição pertencente à família exponencial Υ, então $T = r(\boldsymbol{X}) = \sum_{i=1}^{n} s(X_i)$ é uma estatística suficiente para θ.

Prova

Se $f(x|\theta)$ é uma função (de densidade) de probabilidade da família exponencial Υ, então

$$\begin{aligned} f(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \exp\{a(\theta)s(x_i) + b(\theta) + c(x_i)\} \\ &= \exp\left\{\sum_{i=1}^{n} c(x_i)\right\} \exp\left\{a(\theta)\sum_{i=1}^{n} s(x_i) + nb(\theta)\right\}, \end{aligned}$$

tal que, pelo critério da fatoração[4], $u(\boldsymbol{x}) = \exp\left\{\sum_{i=1}^{n} c(x_i)\right\}$ e $v(r(\boldsymbol{x}), \theta) = \exp\left\{a(\theta)\sum_{i=1}^{n} s(x_i) + nb(\theta)\right\}$.

Assim, $T = r(\boldsymbol{X}) = \sum_{i=1}^{n} s(X_i)$ é uma estatística suficiente.

[4] Ver página 3 da aula_06.pdf

Exemplo (continuação)

A família de distribuições exponencial com parâmetro λ é uma componente da família exponencial Υ. É fácil verificar que

$$f(x|\lambda) = \lambda \exp\{-\lambda x\} \boldsymbol{I}_{[0,\infty)}(x) = \exp\{-\lambda x + \ln(\lambda) + \ln(\boldsymbol{I}_{[0,\infty)}(x))\},$$

com

- $s(x) = x$;
- $a(\lambda) = -\lambda$;
- $b(\lambda) = \ln(\lambda)$; e
- $c(x) = \ln(\boldsymbol{I}_{[0,\infty)}(x))$ (poderia ser omitido e considerar $c(x) = 0$)
- Na notação alternativa, $h(x) = \boldsymbol{I}_{[0,\infty)}(x)$.

Logo, de acordo com o critério da fatoração para a família exponencial, temos que

$$T = r(\boldsymbol{X}) = \sum_{i=1}^{n} s(X_i) = \sum_{i=1}^{n} X_i$$

é uma estatística suficiente para λ.

Exemplo (continuação)

A família de distribuições Bernoulli com parâmetro π é uma componente da família exponencial Υ. Note que

$$f(x|\pi) = \pi^x(1-\pi)^{1-x} = \exp\left\{\ln\left(\frac{\pi}{1-\pi}\right)x + \ln(1-\pi)\right\},$$

com

- $s(x) = x$;
- $a(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$;
- $b(\pi) = \ln(1-\pi)$; e
- $c(x) = 0$ (ou $h(x) = 1$ na notação alternativa).

Portanto, de acordo com o critério da fatoração para a família exponencial, temos que

$$T = r(\boldsymbol{X}) = \sum_{i=1}^{n} s(X_i) = \sum_{i=1}^{n} X_i$$

é uma estatística suficiente para π.

Exemplo (continuação)

A família de distribuições Poisson com parâmetro λ é uma componente da família exponencial Υ. Podemos verificar que

$$f(x|\lambda) = \frac{\exp\{-\lambda\}\lambda^x}{x!} = \exp\{x\ln(\lambda) - \lambda - \ln\Gamma(x+1)\},$$

com

- $s(x) = x$;
- $a(\lambda) = \ln(\lambda)$;
- $b(\lambda) = -\lambda$; e
- $c(x) = -\ln\Gamma(x+1)$.
- Na notação alternativa, $h(x) = \dfrac{1}{x!} = \dfrac{1}{\Gamma(x+1)}$.

Novamente, de acordo com o critério da fatoração para a família exponencial, temos que

$$T = r(\boldsymbol{X}) = \sum_{i=1}^{n} s(X_i) = \sum_{i=1}^{n} X_i$$

é uma estatística suficiente para λ.

Definição (família exponencial multiparamétrica)

Uma família de distribuições de probabilidade que depende de um vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ é chamada de **família exponencial** Υ **(multiparamétrica)** se, e somente se, a função (de densidade) de probabilidade é dada por

$$\begin{aligned} f(x|\boldsymbol{\theta}) &= \exp\left\{\sum_{j=1}^{K} a_j(\theta_1, \theta_2, \ldots, \theta_K) s_j(x) + b(\theta_1, \theta_2, \ldots, \theta_K) + c(x)\right\} \\ &= \exp\left\{\sum_{j=1}^{K} a_j(\boldsymbol{\theta}) s_j(x) + b(\boldsymbol{\theta}) + c(x)\right\} \end{aligned}$$

ou

$$f(x|\theta) = h(x) \exp\left\{\sum_{j=1}^{K} a_j(\boldsymbol{\theta}) s_j(x) + b(\boldsymbol{\theta})\right\}$$

em que

- $a_j(\boldsymbol{\theta})$, $j = 1, 2, \ldots, K$, e $b(\boldsymbol{\theta})$ são funções de $\boldsymbol{\theta}$ e conhecidas; e
- $s_j(x)$, $j = 1, 2, \ldots, K$, $c(x)$ e $h(x)$ são funções de x e conhecidas.

Exemplo

Seja $(X|\mu,\sigma^2) \sim \mathcal{N}(\mu;\ \sigma^2)$. Temos que

$$\begin{aligned} f(x|\mu,\sigma^2) &= (2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= (2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[x^2-2x\mu+\mu^2\right]\right\} \\ &= \exp\left\{-\frac{x^2}{2\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\ln(\sigma^2)-\frac{1}{2}\ln(2\pi)\right\}. \end{aligned}$$

Logo,

- $a_1(\mu,\sigma^2) = -\dfrac{1}{2\sigma^2}$; $\quad a_2(\mu,\sigma^2) = \dfrac{\mu}{\sigma^2}$; $\quad s_1(x) = x^2$; $\quad s_2(x) = x$;
- $b(\mu,\sigma^2) = -\dfrac{\mu^2}{2\sigma^2} - \dfrac{1}{2}\ln(\sigma^2)$; $\quad$ e $\quad c(x) = -\dfrac{1}{2}\ln(2\pi)$.

Portanto, **a distribuição normal pertence a família exponencial** Υ.

Exemplo

Seja $(X|\alpha,\beta) \sim \mathcal{G}(\alpha;\ \beta)$. Temos que

$$\begin{aligned} f(x|\alpha,\beta) &= \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} I_{[0,\infty)}(x) \\ &= \exp\left\{-\beta x + \alpha \ln(x) + \alpha \ln(\beta) - \ln\Gamma(\alpha) - \ln(x)\right\}. \end{aligned}$$

Logo,

- $a_1(\alpha,\beta) = -\beta$; $a_2(\alpha,\beta) = \alpha$; $s_1(x) = x$; $s_2(x) = \ln(x)$;
- $b(\alpha,\beta) = \alpha \ln(\beta) - \ln\Gamma(\alpha)$; e $c(x) = -\ln(x)$.

Portanto, **a distribuição gama pertence a família exponencial** Υ.

Teorema

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma variável aleatória X com função (de densidade) de probabilidade $f(x|\boldsymbol{\theta})$ com $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$. Se X tem distribuição pertencente à família exponencial Υ, então

$$\begin{aligned} T_1 &= r_1(\boldsymbol{X}) = \sum_{i=1}^{n} s_1(X_i) \\ T_2 &= r_2(\boldsymbol{X}) = \sum_{i=1}^{n} s_2(X_i) \\ &\vdots \quad \vdots \quad \vdots \\ T_K &= r_K(\boldsymbol{X}) = \sum_{i=1}^{n} s_K(X_i) \end{aligned}$$

são estatísticas conjuntamente suficiente para $\boldsymbol{\theta}$.

Prova

Fazer como exercício.[5]

[5] Ver página 10 da aula_06.pdf

Exemplo (continuação)

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu, \sigma^2) \sim \mathcal{N}(\mu;\, \sigma^2)$. Temos que

$$\begin{aligned} f(x|\mu,\sigma^2) &= (2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= (2\pi)^{-1/2}(\sigma^2)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\left[x^2-2x\mu+\mu^2\right]\right\} \\ &= \exp\left\{-\frac{x^2}{2\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\ln(\sigma^2)-\frac{1}{2}\ln(2\pi)\right\}. \end{aligned}$$

Logo,

- $a_1(\mu,\sigma^2) = -\dfrac{1}{2\sigma^2}$; $\quad a_2(\mu,\sigma^2) = \dfrac{\mu}{\sigma^2}$; $\quad s_1(x) = x^2$; $\quad s_2(x) = x$;
- $b(\mu,\sigma^2) = -\dfrac{\mu^2}{2\sigma^2} - \dfrac{1}{2}\ln(\sigma^2)$; $\quad$ e $\quad c(x) = -\dfrac{1}{2}\ln(2\pi)$.

Portanto, **a distribuição normal pertence a família exponencial** Υ.

Segue-se que $T_1 = \sum_{i=1}^{n} X_i^2$ e $T_2 = \sum_{i=1}^{n} X_i$ são estatísticas conjuntamente suficientes para (μ, σ^2).

Exemplo (continuação)

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\alpha, \beta) \sim \mathcal{G}(\alpha;\ \beta)$. Temos que

$$\begin{aligned} f(x|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} I_{[0,\infty)}(x) \\ &= \exp\left\{-\beta x + \alpha \ln(x) + \alpha \ln(\beta) - \ln\Gamma(\alpha) - \ln(x)\right\}. \end{aligned}$$

Logo,

- $a_1(\alpha,\beta) = -\beta$; $a_2(\alpha,\beta) = \alpha$; $s_1(x) = x$; $s_2(x) = \ln(x)$;
- $b(\alpha,\beta) = \alpha \ln(\beta) - \ln\Gamma(\alpha)$; e $c(x) = -\ln(x)$.

Portanto, **a distribuição gama pertence a família exponencial** Υ.

Segue-se que $T_1 = \sum_{i=1}^n X_i$ e $T_2 = \sum_{i=1}^n \ln(X_i)$ são estatísticas conjuntamente suficientes para (α, β).

Teste de Hipóteses

Considere um problema estatístico envolvendo o parâmetro θ cujo valor é desconhecido e $\theta \in \Omega$.

Suponha que Ω pode ser particionado em dois conjuntos disjuntos Ω_0 e Ω_1, isto é, $\Omega = \Omega_0 \cup \Omega_1$ e $\varnothing = \Omega_0 \cap \Omega_1$.

Denotaremos por H_0 a hipótese que $\theta \in \Omega_0$ e H_1 a hipótese que $\theta \in \Omega_1$.

Como Ω_0 e Ω_1 são disjuntos, então apenas uma hipótese pode ser verdadeira.

O estatístico deve decidir qual hipótese é a verdadeira. A esse procedimento chamamos de **teste de hipóteses**.

Temos que

		Verdade	
		Ω_0	Ω_1
Decisão	Ω_0	0	Erro Tipo II
	Ω_1	Erro Tipo I	0

Definição

A hipótese H_0 é chamada de **hipótese nula** e a hipótese H_1 é chamada de **hipótese alternativa**. Quando realizamos um teste, se decidimos que θ está em Ω_1, dizemos que **rejeitamos** H_0. Se decidimos que θ está em Ω_0, dizemos que **não rejeitamos** H_0.

Exemplo

Um estudo reporta as medidas de crânios humanos antigos, no Egito, de várias dimensões em diversos períodos no tempo. Em certo período, pode-se considerar as medidas - em milímetros (mm) - das larguras do crânios como uma variável aleatória normal com média μ e variância 26. O interesse pode estar em comparar a média das larguras dos crânios antigos com as dos crânios humanos atuais que é cerca de 140mm. O espaço paramétrico são os números reais positivos. Então, podemos testar algo como

$$\begin{cases} H_0: \ \mu \geqslant 140 \\ H_1: \ \mu < 140. \end{cases}$$

Assim, $\Omega_0 = [140, \infty)$ e $\Omega_1 = (0, 140)$.

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$ com $\theta \in \Omega$ tal que $\Omega = \Omega_0 \cup \Omega_1$ e $\varnothing = \Omega_0 \cap \Omega_1$. Deseja-se testar

$$\begin{cases} H_0 : \ \theta \in \Omega_0 \\ H_1 : \ \theta \in \Omega_1. \end{cases}$$

Para $i = 0$ ou $i = 1$, o conjunto Ω_i pode conter somente um valor de θ ou uma coleção maior.

Definição

Se Ω_i contém apenas um valor de θ, então H_i é uma **hipótese simples**. Se Ω_i contém mais de um valor de θ, então H_i é uma **hipótese composta**.

Definição

Seja θ um parâmetro unidimensional. **Hipóteses unilaterais** são da forma $H_0 : \ \theta \leqslant \theta_0$ ou $H_0 : \ \theta \geqslant \theta_0$ com as hipóteses alternativas unilateriais $H_1 : \ \theta > \theta_0$ ou $H_1 : \ \theta < \theta_0$. Quando H_0 é simples, usualmente H_1 é uma **hipótese bilateral**, $H_1 : \ \theta \neq \theta_0$.

Teste para a média de uma normal com variância conhecida

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu;\, \sigma_0^2)$ com σ_0^2 conhecido. Queremos testar as hipóteses

$$\begin{cases} H_0 : \ \mu = \mu_0 \\ H_1 : \ \mu \neq \mu_0. \end{cases}$$

Note que deve ser razoável rejeitar H_0 se $\overline{x}$ estiver a uma distância maior que c de μ_0.

Podemos dividir o conjunto de todos os dados $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ possíveis por

$$S_0 = \{\boldsymbol{x} : \ -c \leqslant \overline{x} - \mu_0 \leqslant c\} \quad \text{e} \quad S_1 = S_0^c$$

Então, rejeitamos H_0 se $\boldsymbol{x} \in S_1$ e não rejeitamos H_0 se $\boldsymbol{x} \in S_0$.

Podemos definir a estatística $T = |\overline{X} - \mu_0|$ e rejeitar H_0 se $T > c$.

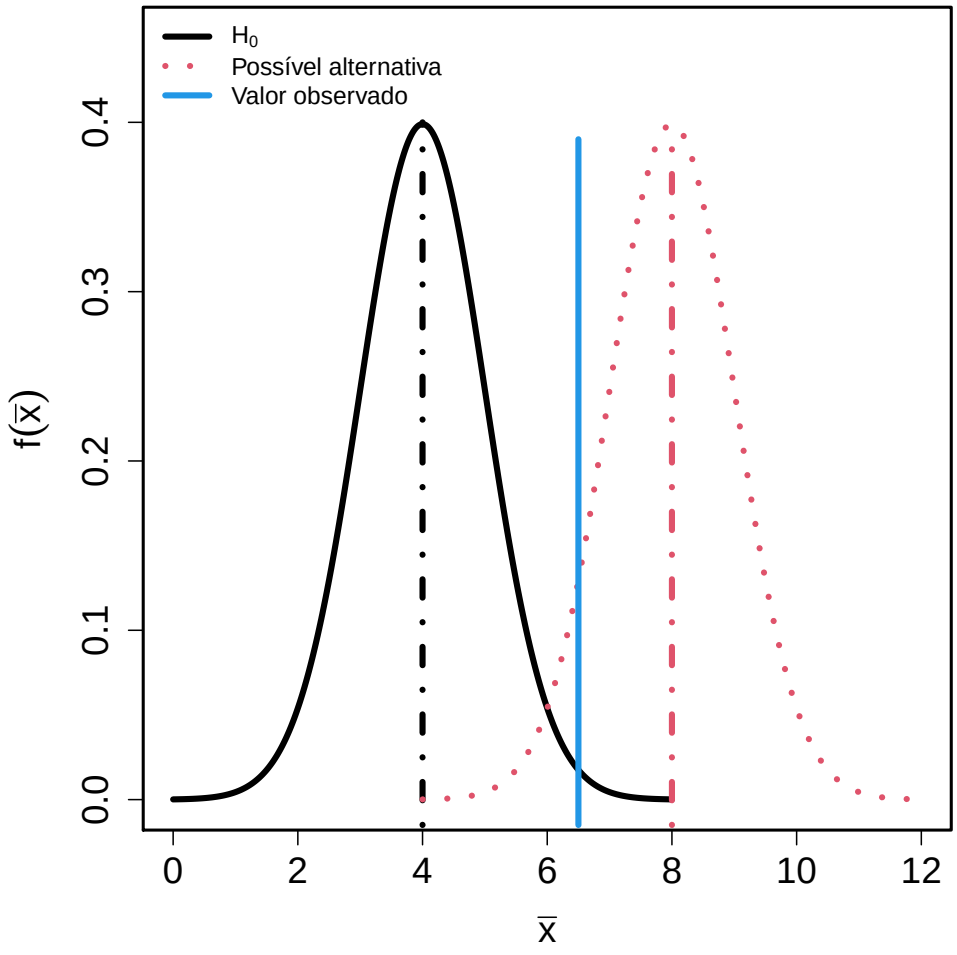


Figura: Teste de hipóteses para média de uma normal com variância conhecida.

Região crítica

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma distribuição com função (de densidade) de probabilidade $f(x|\theta)$ com $\theta \in \Omega$ tal que $\Omega = \Omega_0 \cup \Omega_1$ e $\varnothing = \Omega_0 \cap \Omega_1$. Deseja-se testar

$$\begin{cases} H_0 : \ \theta \in \Omega_0 \\ H_1 : \ \theta \in \Omega_1. \end{cases}$$

Seja S o conjunto de todos os resultados possíveis da amostra $\boldsymbol{X}$.

Defina S_0 e S_1 os subconjuntos de S para o qual H_0 não será rejeitada e H_0 será rejeitada, respectivamente.

Definição

O conjunto S_1 é chamado de **região crítica** do teste.

Um procedimento de teste é determinado ao especificar a região crítica do teste.

O complemento da região crítica deve conter todos os resultados possíveis para o qual H_0 não será rejeitada.

Em geral, a região crítica é definida em termos de uma estatística $T = r(\boldsymbol{X})$.

Definição

Seja $\boldsymbol{X}$ uma amostra aleatória de uma distribuição que depende do parâmetro θ. Seja $T = r(\boldsymbol{X})$ uma estatística e seja R um subconjunto dos reais, $\mathbb{R}$. Suponha que se deseja testar

$$\begin{cases} H_0 : \ \theta \in \Omega_0 \\ H_1 : \ \theta \in \Omega_1, \end{cases}$$

tal que o procedimento seja da forma "rejeita-se H_0 se $T \in R$". Então, chamamos T **de estatística de teste** e R **de região de rejeição ou região crítica** do teste.

Por exemplo, podemos considerar $S_1 = \{\boldsymbol{x} : \ r(\boldsymbol{x}) \in R\}$.

Função poder

Definição

Seja δ um procedimento de teste. Defina $\pi(\theta|\delta)$ como a probabilidade de rejeitar H_0 para $\theta \in \Omega$. A função $\pi(\theta|\delta)$ é chamada de **função poder** do teste δ. Se S_1 denota a região crítica de δ, então a função poder $\pi(\theta|\delta)$ é determinada pela relação

$$\pi(\theta|\delta) = \Pr(\boldsymbol{X} \in S_1|\theta), \quad \theta \in \Omega.$$

Se δ é descrito em termos de uma estatística de teste T e uma região de rejeição R, então a função poder é dada por

$$\pi(\theta|\delta) = \Pr(T \in R|\theta), \quad \theta \in \Omega.$$

A função poder $\pi(\theta|\delta)$ especifica a probabilidade que δ rejeitará H_0, para cada possível valor do parâmetro θ.

Assim, gostaríamos que

- $\pi(\theta|\delta) = 0$ para todo $\theta \in \Omega_0$, e
- $\pi(\theta|\delta) = 1$ para todo $\theta \in \Omega_1$.

Teste para a média de uma normal com variância conhecida

Exemplo

No exemplo anterior, o teste δ é baseado na estatística $T = |\overline{X} - \mu_0|$ com região de rejeição $R = [c, \infty)$. Sabemos que $\overline{X} \sim \mathcal{N}(\mu;\, \sigma_0^2/n)$. A função poder pode ser calculada a partir desta distribuição como

$$
\begin{aligned}
\Pr(T \in R|\mu) &= \Pr(|\overline{X} - \mu_0| \geqslant c|\mu) = 1 - \Pr(|\overline{X} - \mu_0| \leqslant c|\mu) \\
&= 1 - \Pr(-c \leqslant \overline{X} - \mu_0 \leqslant c|\mu) \\
&= 1 - \Pr(\mu_0 - c \leqslant \overline{X} \leqslant \mu_0 + c|\mu) \\
&= 1 - \left[\Pr(\overline{X} \leqslant \mu_0 + c|\mu) - \Pr(\overline{X} \leqslant \mu_0 - c|\mu)\right] \\
&= 1 - \Pr(\overline{X} \leqslant \mu_0 + c|\mu) + \Pr(\overline{X} \leqslant \mu_0 - c|\mu) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(\mu_0 + c - \mu)}{\sigma_0}\right) + \Phi\left(\frac{\sqrt{n}(\mu_0 - c - \mu)}{\sigma_0}\right).
\end{aligned}
$$

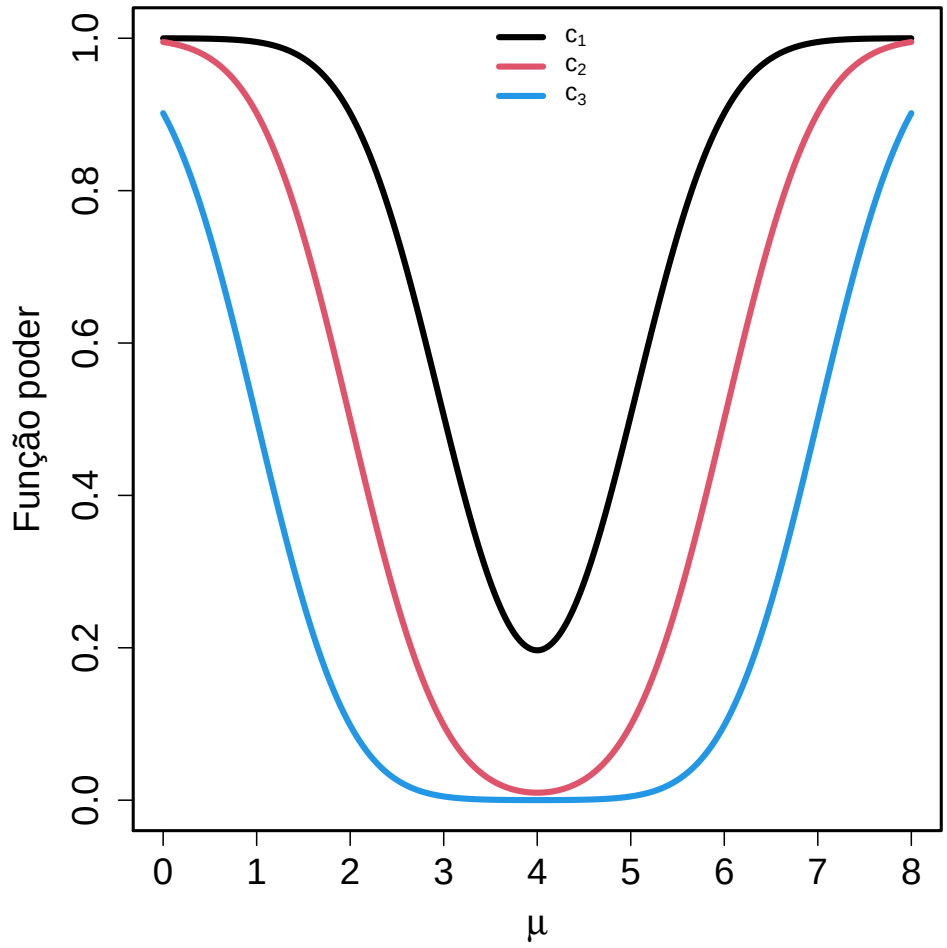


Figura: Função poder para $\mu_0 = 4$, $n = 15$, $\sigma_0^2 = 9$ e variando $c_1 = 1$, $c_2 = 2$ e $c_3 = 3$.

Conforme visto anteriormente, temos que

		Verdade	
		Ω_0	Ω_1
Decisão	Ω_0	0	Erro Tipo II
	Ω_1	Erro Tipo I	0

ou

		Verdade	
		$\theta \in \Omega_0$	$\theta \in \Omega_1$
Decisão	$\theta \in \Omega_0$	0	Erro Tipo II
	$\theta \in \Omega_1$	Erro Tipo I	0

Definição

Uma decisão errada de rejeitar a hipótese nula - quando ela é verdadeira - é um **Erro tipo I** ou um erro do primeiro tipo. Uma decisão errada de não rejeitar a hipótese nula - quando ela é falsa - é um **Erro tipo II** ou um erro do segundo tipo.

Temos que

$$\begin{aligned}
\alpha &= \text{Pr(Erro tipo I)} \\
&= \text{Pr(Rejeitar } H_0 | H_0 \text{ é verdadeira)} \\
&= \text{Pr}(T \in R | \theta \in \Omega_0) \\
&= \pi(\theta|\delta), \quad \text{para} \quad \theta \in \Omega_0.
\end{aligned}$$

$$\begin{aligned}
\beta &= \text{Pr(Erro tipo II)} \\
&= \text{Pr(Não rejeitar } H_0 | H_0 \text{ é falsa)} \\
&= \text{Pr}(T \notin R | \theta \in \Omega_1) \\
&= 1 - \text{Pr}(T \in R | \theta \in \Omega_1) \\
&= 1 - \pi(\theta|\delta), \quad \text{para} \quad \theta \in \Omega_1.
\end{aligned}$$

O método mais popular que procura balancear entre dois objetivos (minimizar $\pi(\theta|\delta)$ para todo $\theta \in \Omega_0$ - baixo poder em Ω_0, e maximizar $\pi(\theta|\delta)$ para todo $\theta \in \Omega_1$ - alto poder em Ω_1) é escolher um número $\alpha_0 \in (0, 1)$ tal que

$$\pi(\theta|\delta) \leqslant \alpha_0, \quad \theta \in \Omega_0.$$

Definição

Um teste que satisfaz

$$\pi(\theta|\delta) \leqslant \alpha_0, \quad \theta \in \Omega_0,$$

é chamado de **teste de nível de significância α_0.** Além disso, o tamanho $\alpha(\delta)$ de um teste δ é definido como

$$\alpha(\delta) = \sup_{\theta \in \Omega_0} \pi(\theta|\delta).$$

Corolário

Um teste δ é um teste de nível α_0 se, e somente se, seu tamanho é no máximo α_0 (isto é, $\alpha(\delta) \leqslant \alpha_0$). Se a hipótese nula é simples, isto é, $H_0 : \ \theta = \theta_0$, então o tamanho de δ será $\alpha(\delta) = \pi(\theta_0|\delta)$.

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\theta) \sim \mathcal{U}(0;\, \theta)$ com $\theta > 0$. Suponha que desejamos testar as seguintes hipóteses:

$$\begin{cases} H_0 : \ 3 \leqslant \theta \leqslant 4 \\ H_1 : \ \theta < 3 \text{ ou } \theta > 4. \end{cases}$$

- Sabemos que $X_{(n)}$ é o estimador de máxima verossimilhança para θ.
- Sabemos que $X_{(n)} \leqslant \theta$, mas $X_{(n)}$ tem alta probabilidade de estar perto de θ se o tamanho da amostra n for grande.
- Suponha que o teste δ não rejeita H_0 se $2,9 < X_{(n)} < 4$ e o teste δ rejeita H_0 se $X_{(n)} \leqslant 2,9$ ou $X_{(n)} \geqslant 4$.
- A região crítica do teste δ deve conter todos os valores $\boldsymbol{X}$ tal que $X_{(n)} \leqslant 2,9$ ou $X_{(n)} \geqslant 4$.
- Em termos da estatística de teste $X_{(n)}$, a região de rejeição é $R = (-\infty; 2,9] \cup [4; \infty)$.

- A função poder de δ é dada por

$$\pi(\theta|\delta) = \Pr(X_{(n)} \leqslant 2,9|\theta) + \Pr(X_{(n)} \geqslant 4|\theta).$$

- Se $\theta \leqslant 2,9$, então $\Pr(X_{(n)} \leqslant 2,9|\theta) = 1$ e $\Pr(X_{(n)} \geqslant 4|\theta) = 0$. Isto implica que $\pi(\theta|\delta) = 1$ para $\theta \leqslant 2,9$.
- Se $2,9 < \theta \leqslant 4$, então $\Pr(X_{(n)} \leqslant 2,9|\theta) = [2,9/\theta]^n$ e $\Pr(X_{(n)} \geqslant 4|\theta) = 0$. Isto implica que $\pi(\theta|\delta) = [2,9/\theta]^n$ para $2,9 < \theta \leqslant 4$.
- Se $\theta > 4$, então $\Pr(X_{(n)} \leqslant 2,9|\theta) = [2,9/\theta]^n$ e $\Pr(X_{(n)} \geqslant 4|\theta) = 1 - [4/\theta]^n$. Isto implica que $\pi(\theta|\delta) = [2,9/\theta]^n + 1 - [4/\theta]^n$ para $\theta > 4$.

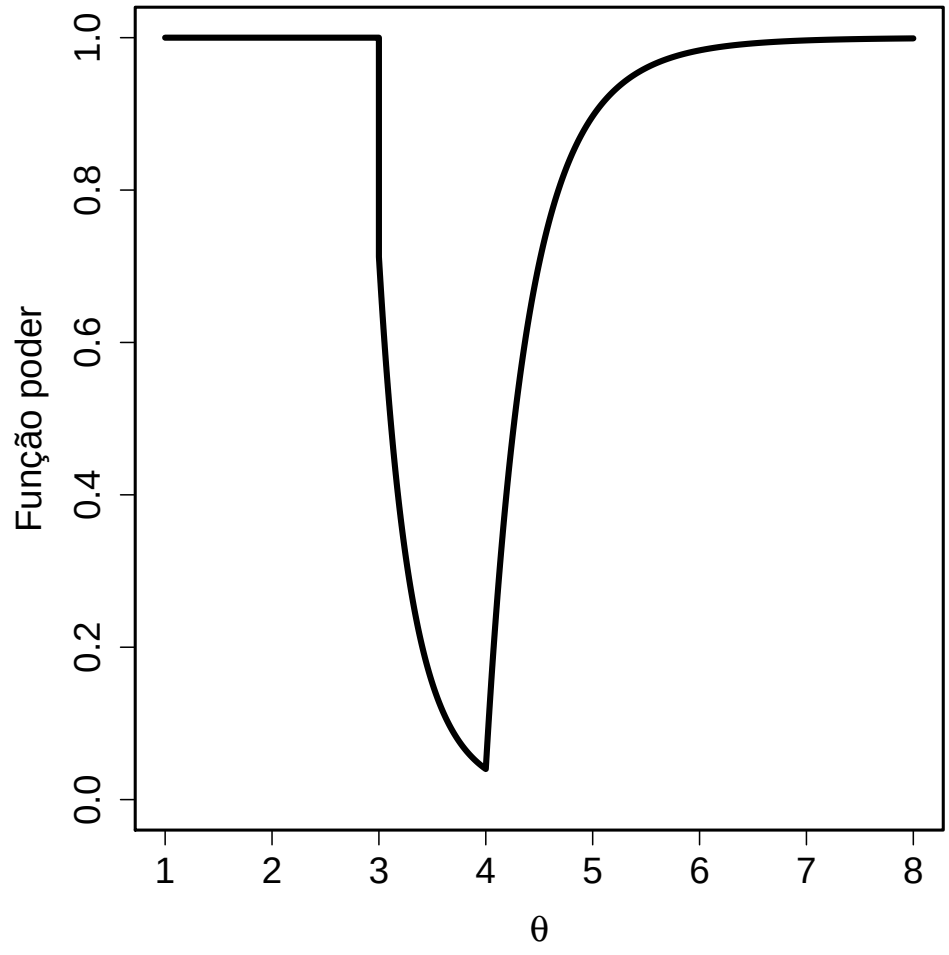


Figura: Função poder para θ ($n = 10$).

Nível de significância específico para um teste

Suponha que desejamos testar as hipóteses:

$$\begin{cases} H_0 : \theta \in \Omega_0 \\ H_1 : \theta \in \Omega_1. \end{cases}$$

Seja T a estatística de teste e suponha que o teste rejeita H_0 se $T \geqslant c$, para alguma constante c.

Suponha também que desejamos ter o teste com nível de significância α_0.

A função poder do teste é $\pi(\theta|\delta) = \Pr(T \geqslant c|\theta)$ e queremos que

$$\sup_{\theta \in \Omega_0} \Pr(T \geqslant c|\theta) \leqslant \alpha_0.$$

A função poder é não crescente em c. A desigualdade acima será satisfeita para valores grandes de c, mas não para valores pequenos.

Se desejamos que a função poder seja tão grande quanto possível para $\theta \in \Omega_1$, então devemos ter c tão pequeno quanto possível enquanto satisfaz a desigualdade anterior.

Se T tem distribuição contínua, então é geralmente simples de encontrar um c apropriado.

Teste para a média de uma normal com variância conhecida

Exemplo (continuação)

O teste δ é baseado na estatística $T = |\overline{X} - \mu|$ com região de rejeição $R = [c, \infty)$. Logo,

$$
\begin{aligned}
\alpha_0 &= \pi(\mu_0|\delta) = \Pr(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = \Pr(T \geqslant c | \mu = \mu_0) \\
&= \Pr(|\overline{X} - \mu| \geqslant c | \mu = \mu_0) = 1 - \Pr(|\overline{X} - \mu_0| \leqslant c) \\
&= 1 - \Pr\left(\frac{\sqrt{n}|\overline{X} - \mu_0|}{\sigma} \leqslant \frac{c\sqrt{n}}{\sigma}\right) = 1 - \Pr\left(|Z| \leqslant \frac{c\sqrt{n}}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right) + \Phi\left(-\frac{c\sqrt{n}}{\sigma}\right) = 2\left[1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right] \quad \Rightarrow \\
\frac{\alpha_0}{2} &= 1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right) \quad \Rightarrow \\
c &= \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right).
\end{aligned}
$$

Então, rejeitamos H_0 se $T \geqslant c$, isto é, se $|z| \geqslant \Phi^{-1}\left(1 - \frac{\alpha_0}{2}\right)$.

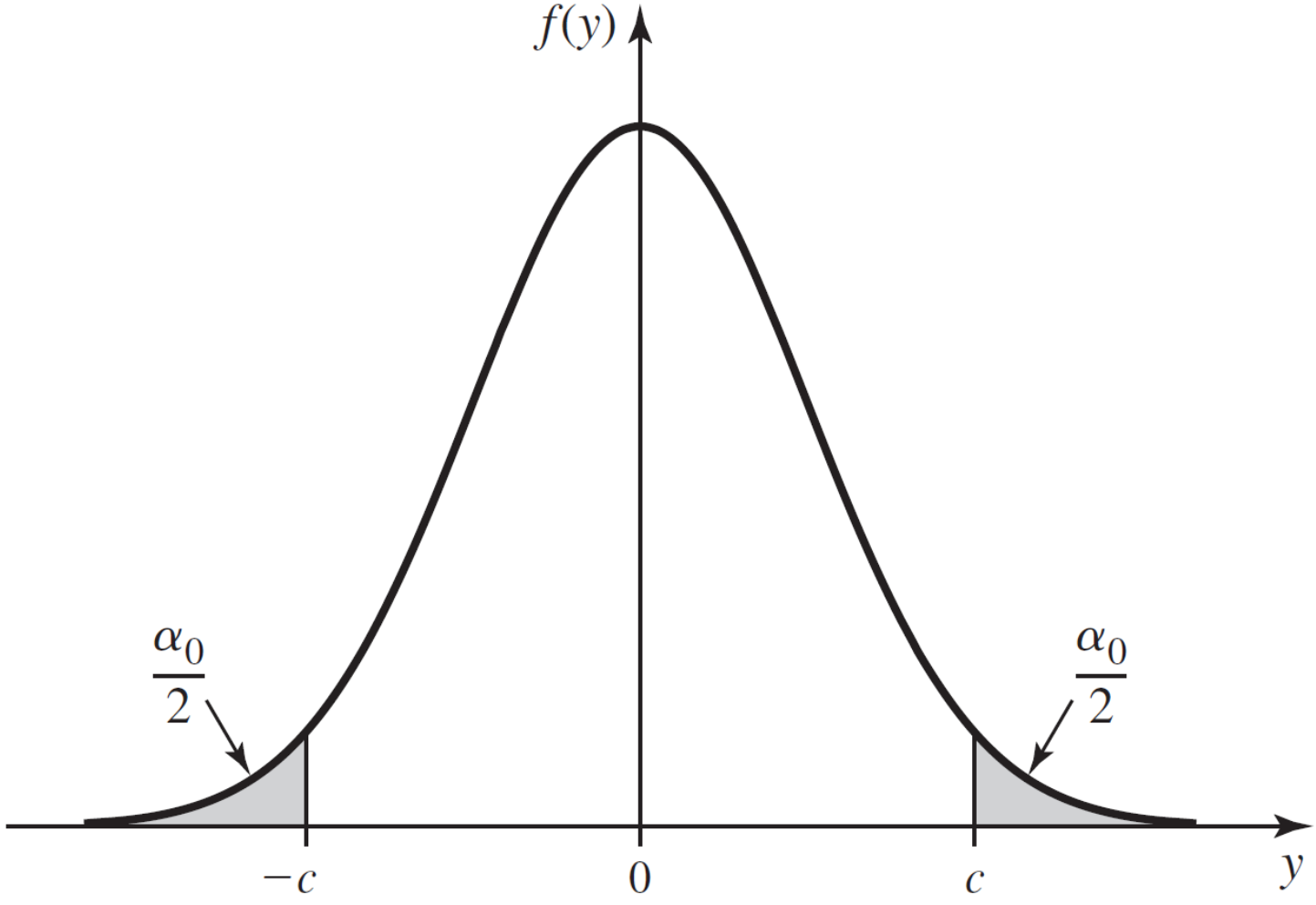


Figura: Função de densidade de probabilidade para $Y = \overline{X} - \mu_0$. A área sombreada é a probabilidade de $|Y| \geqslant c$.

Teste para o parâmetro da Bernoulli

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$ e suponha que desejamos testar

$$\begin{cases} H_0 : \pi \leqslant \pi_0 \\ H_1 : \pi > \pi_0. \end{cases}$$

- Sabemos que $Y = \sum_{i=1}^{n} X_i \sim \mathcal{BN}(n; \pi)$.
- Valores grandes de π induzem valores altos de Y.
- Rejeitaremos H_0 se $Y \geqslant c$ para alguma constante c.
- Queremos que o tamanho do teste esteja tão perto de α_0 quando possível (sem ultrapassar).
- É fácil verificar que $\Pr(Y \geqslant c|\pi)$ é uma função crescente de π. Assim, o tamanho do teste será $\Pr(Y \geqslant c|\pi_0)$.
- Portanto, c deve ser o menor número tal que $\Pr(Y \geqslant c|\pi_0) \leqslant \alpha_0$.
- Por exemplo, para $n = 10$, $\pi_0 = 0,3$ e $\alpha_0 = 0,1$, temos que

$$\sum_{y=6}^{10} \Pr(Y = y|\pi = 0,3) = 0,047 \text{ e } \sum_{y=5}^{10} \Pr(Y = y|\pi = 0,3) = 0,15.$$

- Concluímos que para o teste ter um tamanho de no máximo 0,1, devemos escolher $c > 5$. Qualquer valor em $(5, 6]$ produz o mesmo resultado.

Valor *p* de um teste

Tipicamente, um analista não especifica α_0 e, então, simplesmente reporta se H_0 foi ou não rejeitada ao nível de significância α_0.

É comum reportar, em adição ao valor da estatística de teste, todos os valores de α_0' para os quais o teste de nível α_0 levariam a rejeição de H_0.

Exemplo

Considere o exemplo anterior do teste para média da normal com variância conhecida. Suponha que $z = 2,78$ (valor observado).

Então, H_0 será rejeitada para cada nível de significância α_0 tal que $2,78 \geqslant \Phi^{-1}(1 - \alpha_0/2)$. Isto implica que $\alpha_0 \geqslant 0,0054$.

O valor $0,0054$ é chamado de valor *p* para os dados observados e as hipóteses testadas.

Se $\alpha_0 = 0,01$, então como $0,01 > 0,0054$, segue-se que H_0 é rejeitada.

Valor *p*

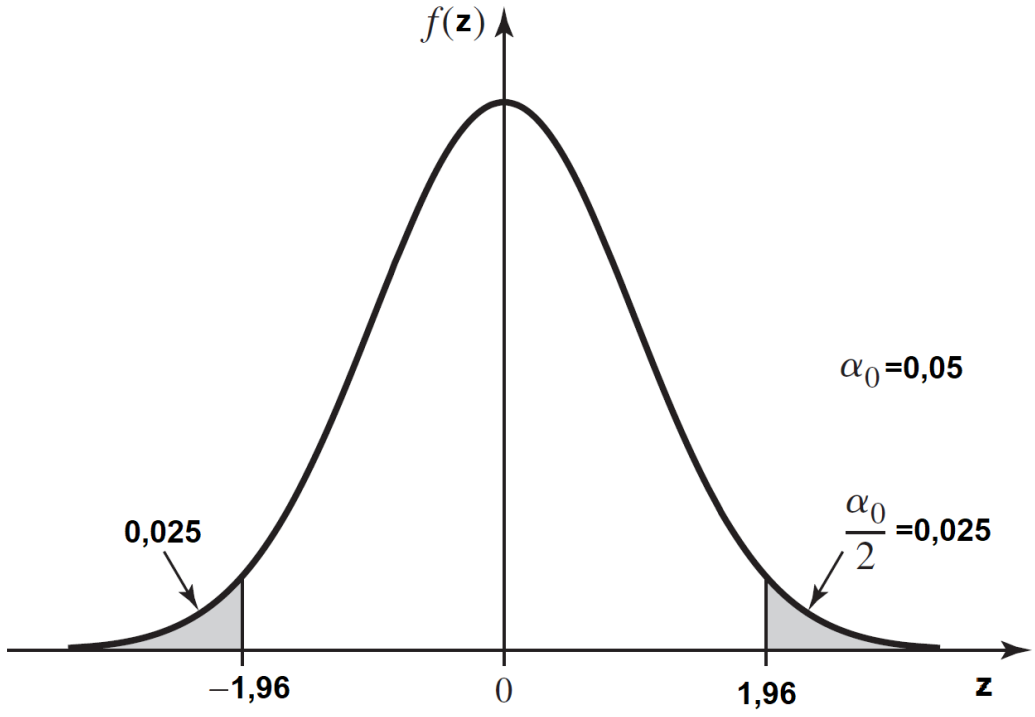


Figura: Comparando com a $\mathcal{N}(0,1)$.

Definição

O **valor** *p* é o menor nível de significância α_0 tal que rejeitamos H_0 ao nível α_0 com os dados observados.

Calculando o valor p

Mais geralmente, para cada t, seja δ_t o teste que rejeita H_0 se $T \geqslant t$. Então, o valor p, quando $T = t$ é observado, é o tamanho do teste δ_t.

Assim, o valor p é obtido através de

$$\sup_{\theta \in \Omega_0} \pi(\theta|\delta_t) = \sup_{\theta \in \Omega_0} \Pr(T \geqslant t|\theta).$$

Tipicamente, $\pi(\theta|\delta_t)$ é maximizado em algum valor θ_0 na fronteira entre Ω_0 e Ω_1.

Exemplo (parâmetro da Bernoulli)

- No exemplo, rejeitamos H_0 se $Y \geqslant c$.
- O valor p, quando $Y = y$ é observado, é $\sup_{\pi \leqslant \pi_0} \Pr(Y \geqslant y|\pi)$.
- É fácil verificar que $\Pr(Y \geqslant y|\pi)$ é uma função crescente em π.
- Assim, o valor p é $\Pr(Y \geqslant y|\pi_0)$.
- Por exemplo, para $\pi_0 = 0,3$ e $n = 10$, se $Y = 6$ é observado, então $\Pr(Y \geqslant 6|\pi = 0,3) = 0,0473$ (o valor p).

Exemplo

Seja $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\mu) \sim \mathcal{N}(\mu, 1)$. Suponha que para $n = 10$, obteve-se $\overline{x} = 2,5$.

Desejamos testar as hipóteses

$$\begin{cases} H_0 : \ \mu \leqslant 2 \\ H_1 : \ \mu > 2. \end{cases}$$

Então, o valor p é dado por

$$\begin{aligned} \pi(2|\delta) &= \Pr(\overline{X} \geqslant 2,5|\mu = 2) = \Pr\left(\frac{\overline{X} - 2}{\sqrt{1/10}} \geqslant \frac{2,5 - 2}{\sqrt{1/10}}\right) \\ &= \Pr(Z \geqslant 1,5811) = 0,0569. \end{aligned}$$

Suponha que o nível de significância tenha sido fixado por três pesquisadores diferentes em $\alpha_0^A = 0,1$, $\alpha_0^B = 0,025$ e $\alpha_0^C = 0,01$, respectivamente.

- Se $\alpha_0^A = 0,1$, $\Pr(\overline{X} \geqslant c | \mu = 2) = \Pr\left(Z \geqslant \dfrac{c-2}{\sqrt{1/10}}\right)$ e $c = 2,405$.

  Como $\overline{x} = 2,5 > c = 2,405$, então **rejeitamos** H_0.

- Se $\alpha_0^B = 0,025$, $\Pr(\overline{X} \geqslant c | \mu = 2) = \Pr\left(Z \geqslant \dfrac{c-2}{\sqrt{1/10}}\right)$ e $c = 2,619$.

  Como $\overline{x} = 2,5 < c = 2,619$, então **não rejeitamos** H_0.

- Se $\alpha_0^C = 0,01$, $\Pr(\overline{X} \geqslant c | \mu = 2) = \Pr\left(Z \geqslant \dfrac{c-2}{\sqrt{1/10}}\right)$ e $c = 2,736$.

  Como $\overline{x} = 2,5 < c = 2,736$, então **não rejeitamos** H_0.

Distribuição F-Snedecor

Seja X uma variável aleatória contínua com distribuição F-Snedecor com (parâmetros) graus de liberdades n e m. Então, a função de densidade de probabilidade é dada por

$$f_X(x) = \left(\frac{n}{m}\right)^{n/2} \frac{\Gamma((m+n)/2)}{\Gamma(n/2)\Gamma(m/2)} x^{n/2-1} \left(1 + \frac{n}{m}x\right)^{-(n+m)/2} \boldsymbol{I}_{[0,\infty)}(x),$$

em que $n > 0$ e $m > 0$.

Denotaremos a distribuição F-Snedecor com n e m graus de liberdade por $\mathcal{F}_{n,m}$.

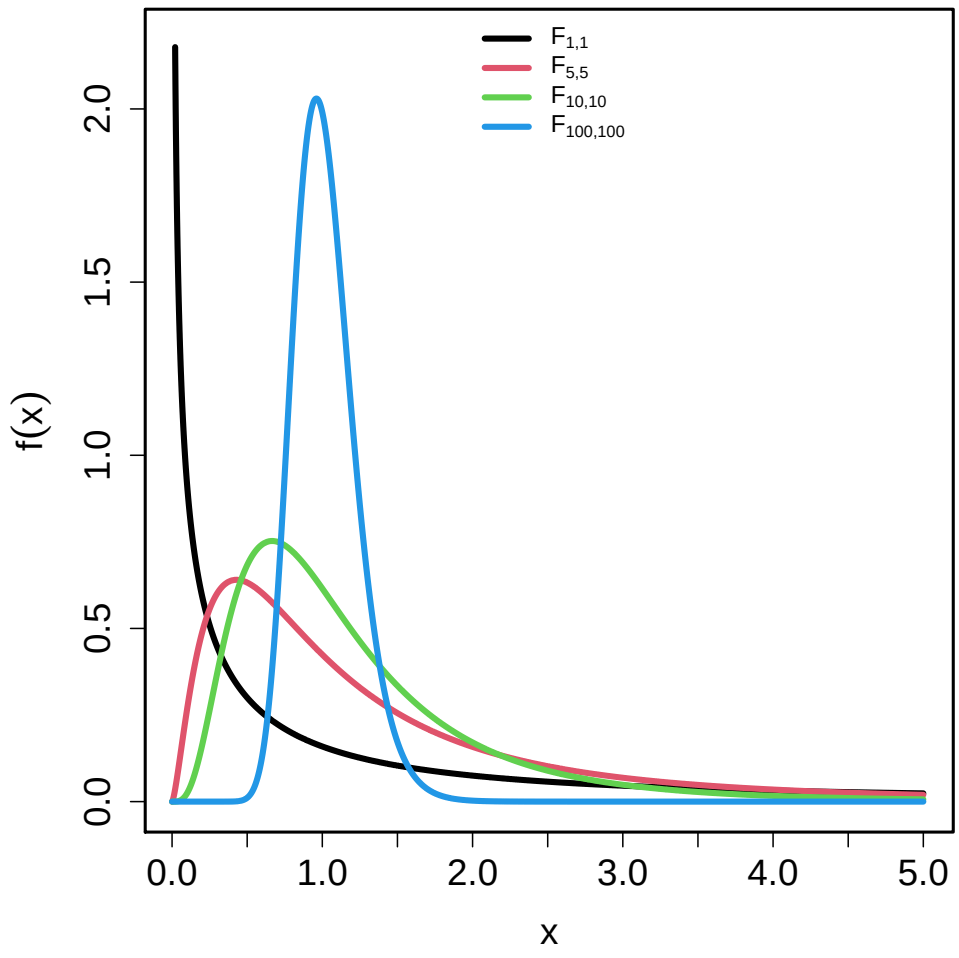


Figura: Exemplos da função de densidade de probabilidade da F-Snedecor.

Teorema

Sejam X e Y variáveis aleatórias independentes com $X \sim \chi^2_n$ e $Y \sim \chi^2_m$, então

$$W = \frac{X/n}{Y/m} \sim \mathcal{F}_{n,m}.$$

Prova

$$\begin{aligned}
F_W(w) &= \Pr(W \leqslant w) = \Pr\left(\frac{X/n}{Y/m} \leqslant w\right) = \Pr\left(X \leqslant w\frac{nY}{m}\right) \\
&= \int_0^\infty \Pr\left(X \leqslant w\frac{ny}{m}\middle| Y = y\right) f_Y(y)dy \\
f_W(w) &= \frac{\partial F_W(w)}{\partial w} = \frac{\partial}{\partial w}\int_0^\infty \Pr\left(X \leqslant w\frac{ny}{m}\middle| Y = y\right) f_Y(y)dy \\
&= \int_0^\infty \frac{\partial}{\partial w} F_X\left(w\frac{ny}{m}\right) f_Y(y)dy = \int_0^\infty f_X\left(w\frac{ny}{m}\right)\frac{ny}{m} f_Y(y)dy \\
&= \int_0^\infty \frac{(n/m)^{n/2} w^{n/2-1}}{\Gamma(n/2)\Gamma(m/2)2^{(n+m)/2}} y^{(n+m)/2-1} \exp\left\{-y\frac{1}{2}\left(1 + \frac{nw}{m}\right)\right\} dy \\
&= \left(\frac{n}{m}\right)^{n/2} \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} w^{n/2-1}\left(1 + \frac{nw}{m}\right)^{-(n+m)/2}, \quad w \geqslant 0.
\end{aligned}$$

Portanto, $W \sim \mathcal{F}_{n,m}$.

Note que

$$\frac{1}{W} = \frac{Y/m}{X/n} \sim \mathcal{F}_{m,n}.$$

Teorema

Se $T \sim t_n(0;\, 1)$ e $Y = T^2$, então $Y \sim \mathcal{F}_{1,n}$.

Prova

$$\begin{aligned}
F_Y(y) &= \Pr(Y \leqslant y) = \Pr(T^2 \leqslant y) = \Pr(-\sqrt{y} \leqslant T \leqslant \sqrt{y}) \\
&= \Pr(T \leqslant \sqrt{y}) - \Pr(T \leqslant -\sqrt{y}) = F_T(\sqrt{y}) - F_T(-\sqrt{y}) \\
f_Y(y) &= \frac{\partial F_Y(y)}{\partial y} = f_T(\sqrt{y})\frac{1}{2\sqrt{y}} + f_T(-\sqrt{y})\frac{1}{2\sqrt{y}} \\
&= \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2)n^{1/2}} y^{1/2-1}\left(1 + \frac{y}{n}\right)^{-(n+1)/2}, \quad y \geqslant 0.
\end{aligned}$$

Portanto, $Y \sim \mathcal{F}_{1,m}$.

Aplicação da distribuição F-Snedecor

Sejam $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ e $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$ amostras aleatórias (independentes) de $(X|\mu_X, \sigma_Y^2) \sim \mathcal{N}(\mu_X;\, \sigma_X^2)$ e $(Y|\mu_Y, \sigma_Y^2) \sim \mathcal{N}(\mu_Y;\, \sigma_Y^2)$, respectivamente.

Suponha que nosso interesse esteja em testar as hipóteses

$$\begin{cases} H_0 : \ \sigma_X^2 = \sigma_Y^2 \\ H_1 : \ \sigma_X^2 \neq \sigma_Y^2 \end{cases} \iff \begin{cases} H_0 : \ \dfrac{\sigma_X^2}{\sigma_Y^2} = 1 \\ \\ H_1 : \ \dfrac{\sigma_X^2}{\sigma_Y^2} \neq 1. \end{cases}$$

Considere os estimadores de σ_X^2 e σ_Y^2 como

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{e} \quad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_i - \overline{Y})^2,$$

respectivamente, tal que

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_n^2 \quad \text{e} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_m^2.$$

Note que as distribuições das estatísticas S_X^2 e S_Y^2 não dependem de μ_X e μ_Y. Assim, não importa, em princípio, se as médias são iguais ou diferentes.

Agora, se H_0 é verdadeira, isto é, se $\sigma_X^2/\sigma_Y^2 = 1$ for verdadeira, esperamos que S_X^2 e S_Y^2 produzam valores observados muito próximos tal que S_X^2/S_Y^2 esteja próximo de 1. Assim, podemos utilizar esta estatística de teste.

A distribuição da estatística de teste, sob H_0 em que $\sigma_X^2 = \sigma_Y^2$, é dada por

$$\frac{\dfrac{(n-1)S_X^2}{\sigma_X^2} \Big/ (n-1)}{\dfrac{(m-1)S_Y^2}{\sigma_Y^2} \Big/ (m-1)} = \frac{\sigma_Y^2}{\sigma_X^2} \times \frac{S_X^2}{S_Y^2} = \frac{S_X^2}{S_Y^2} \sim \mathcal{F}_{n-1,m-1}.$$

Note que sob H_0

$$\frac{S_Y^2}{S_X^2} \sim \mathcal{F}_{m-1,n-1}.$$

Intervalos de credibilidade

Definição

Considere a distribuição a posteriori $f(\boldsymbol{\theta}|\boldsymbol{x})$ e que

$$\Pr(\boldsymbol{\theta} \in \boldsymbol{A}|\boldsymbol{x}) \geqslant \gamma \quad \text{para} \quad 0 < \gamma < 1.$$

Dizemos que *A* é a região de credibilidade de $100\gamma\%$ para θ.

Quando θ é unidimensional e *A* é um intervalo da reta, dizemos que *A* é um intervalo de credibilidade.

Existem os intervalos de credibilidade que consideram a "maior densidade a posteriori" possível, cujo conceito está em encontrar o menor intervalo tal que a probabilidade a posteriori corresponda a γ.

Se a posteriori for uma distribuição simétrica, então este tipo de intervalo é fácil de ser encontrado.

Contudo, em geral, escolheremos intervalos de credibilidade que colocam $\alpha/2$ de probabilidade nas caudas inferior e superior com $\alpha = 1 - \gamma$.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\lambda) \sim \mathcal{E}(\lambda)$. Suponha que $\lambda \sim \mathcal{G}(a, b)$. Então,

$$f(\boldsymbol{x}|\lambda) = \lambda^n \exp\{-\lambda n\overline{x}\}, \quad f(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp\{-b\lambda\}\boldsymbol{I}_{[0,\infty)}(\lambda) \quad \text{e}$$

$$f(\lambda|\boldsymbol{x}) = \frac{(b+n\overline{x})^{(n+a)}}{\Gamma(n+a)}\lambda^{(n+a)-1}\exp\{-(b+n\overline{x})\lambda\}\boldsymbol{I}_{[0,\infty)}(\lambda),$$

ou seja, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(n+a,\ b+n\overline{x})$.

Suponha que $a = b = 0,1$, $n = 10$ e $\overline{x} = 7,5$. Então, $(\lambda|\boldsymbol{x}) \sim \mathcal{G}(10,1;\ 75,1)$. Agora, podemos calcular um intervalo de credibilidade de 0,95 para λ. Precisamos encontrar

$$\lambda_1^* = F^{-1}(0,025|10,1;\ 75,1) \quad \text{e} \quad \lambda_2^* = F^{-1}(0,975|10,1;\ 75,1),$$

em que F^{-1} é a função de distribuição acumulada inversa de $\mathcal{G}(10,1;\ 75,1)$.

Utilizando a função **`qgamma`** do **`R`**, obtemos que

$$(\lambda_1^\star;\ \lambda_2^\star) = (0,06477;\ 0,22924).$$

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de $(X|\pi) \sim \mathcal{BR}(\pi)$. Suponha que $f(\pi) \propto \pi^{-1}(1-\pi)^{-1}$. Então,

$$(\pi|\boldsymbol{x}) \sim \mathcal{BT}(n\overline{x};\ n(1-\overline{x})).$$

Suponha que $n = 50$ e $\overline{x} = 0,32$. Então, $(\pi|\boldsymbol{x}) \sim \mathcal{BT}(16;\ 34)$. Agora, podemos calcular um intervalo de credibilidade de 0,95 para π. Precisamos encontrar

$$\pi_1^* = F^{-1}(0,025|16;\ 34) \quad \text{e} \quad \pi_2^* = F^{-1}(0,975|16;\ 34),$$

em que F^{-1} é a função de distribuição acumulada inversa de $\mathcal{BT}(16;\ 34)$.

Utilizando a função **qbeta** do **R**, obtemos que

$$(\pi_1^\star;\ \pi_2^\star) = (0,1995;\ 0,4542).$$