Tópico 5: Estimação pontual de parâmetros

Ben Dêivide

6 de outubro de 2021

Descrição sobre o ponto - Estimação de parâmetros: Método dos momentos, Método da máxima verossimilhança, método do mínimos quadrados; Teorema de Rao-Blackwell; Estatísticas suficientes e completas; Teorema de Lehman-Scheffé; Informação de Fisher: Desigualdade de Cramer-Rao; Propriedades assintóticas: eficiência, consistência e normalidade assintótica.

1 Conceitos iniciais

Assumimos $X_1, X_2, \ldots, X_n$ uma amostra aleatória com função densidade de probabilidade (fdp) ou função de probabilidade (fp) $f_X(x;\theta)$, em que a forma de $f_X(x;\theta)$ é conhecida, mas o parâmetro θ é desconhecido. Considere o termo amostra aleatória como um conjunto de variáveis aleatórias independentes e identicamente distribuídas (iid). Ainda podemos assumir que θ pode ser um vetor de parâmetros $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]'$. Considere Θ o espaço paramétrico, denotando o conjunto dos possíveis valores que o parâmetro θ pode assumir.

O objetivo é encontrar funções da amostra $X_1, X_2, \ldots, X_n$ para serem usadas como estimadores de θ_j, $j = 1, 2, \ldots, k$. Ou mais geral, nosso objetivo será tentar encontrar estimadores de certas funções, ditas $\tau_1(\boldsymbol{\theta})$, $\tau_2(\boldsymbol{\theta})$, ..., $\tau_r(\boldsymbol{\theta})$ de $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]'$. O ramo da estatística do qual buscamos essas funções da amostra é a inferência estatística.

Definição 1 (Inferência estatística)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com função densidade de probabilidade (fdp) ou função de probabilidade (fp) $f_X(x;\theta)$, em que o parâmetro $\theta \in \Theta$ é desconhecido. Chamamos de inferência estatística o problema que consiste em especificar um ou mais valores para θ, baseado em um conjunto de valores observados $x_1, x_2, \ldots, x_n$ de $X_1, X_2, \ldots, X_n$.*

Definição 2 (Estatística)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x;\theta)$, com $\theta \in \Theta$ em que Θ é o espaço paramétrico. Então qualquer função do tipo $T = t(X_1, X_2, \ldots, X_n)$ que não depende de θ é chamado de estatística.* □

Algumas estatísticas conhecidas são: $\bar{X}$, S^2, $X_{(1)}$, $X_{(n)}$, $X_{(n)} - X_{(1)}$, $\bar{X}/S^2$. Então a afirmação "...não depender de θ...", significa que $\theta\bar{X}$ não é uma estatística. Entretanto, as estatísticas têm distribuições que podem depender do parâmetro θ desconhecido.

Definição 3 (Estimador)**.** *Qualquer estatística cujos valores são usados para estimar $\tau(\theta)$, em que $\tau(.)$ é alguma função do parâmetro θ, é definida ser um estimador de $\tau(\theta)$.* □

Um estimador é sempre uma estatística que é uma função de uma amostra aleatória e que portanto, também é uma variável aleatória. Usaremos como notação $\hat{\theta}$ para representar um estimador de θ, ou mais geral, $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ é um vetor de estimadores de $(\theta_1, \theta_2, \ldots, \theta_k)$.

Nesse material dissertaremos sobre diversos métodos para encontrar estimadores que têm sido propostos. Apresentaremos três métodos: método dos momentos, método da máxima verossimilhança e método dos mínimos quadrados. Posteriormente, mostraremos algumas propriedade desses estimadores e a forma de avaliá-los.

2 Métodos para encontrar estimadores

2.1 Método dos momentos

Definição 4 (Momentos amostrais). *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x; \boldsymbol{\theta})$, com $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]' \in \Theta$ em que Θ é o espaço paramétrico. O k-ésimo momento amostral, denotado por M_k, é definido por*

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \tag{1}$$

e o k-ésimo momento amostral em torno da média amostral $\bar{X}$, denotado por M'_k, é definido por

$$M'_k = \frac{1}{n}\sum_{i=1}^{n}(X - \bar{X})^k. \tag{2}$$

□

Definição 5 (Momentos populacionais). *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x; \boldsymbol{\theta})$, com $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]' \in \Theta$ em que Θ é o espaço paramétrico. O k-ésimo momento populacional, denotado por μ_k, é definido por*

$$\mu_k = E[X^k], \tag{3}$$

e o k-ésimo momento populacional em torno da média populacional $\mu = E[X]$, denotado por μ'_k, é definido por

$$\mu'_k = E[(X - \mu)^k]. \tag{4}$$

□

Geralmente μ_k ou μ'_k é uma função conhecida e função dos k parâmetros $\theta_1, \theta_2, \ldots, \theta_k$. Portanto, poderíamos reescrever $\mu_k = \mu_k(\theta_1, \theta_2, \ldots, \theta_k)$ e $\mu'_k = \mu_k(\theta_1, \theta_2, \ldots, \theta_k)$.

Definição 6 (Método dos momentos). *O método dos momentos consiste em igualar* (1) *e* (3) *ou* (2) *e* (4), *isto é,*

$$M_j = \mu_j(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \text{ para } j = 1, 2, \ldots, k, \tag{5}$$

ou

$$M'_j = \mu'_j(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k), \text{ para } j = 1, 2, \ldots, k, \tag{6}$$

e obter a solução para as k equações. Diremos que $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ são os estimadores de $\theta_1, \theta_2, \ldots, \theta_k$ pelo método dos momentos. □

Exemplo 1. *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma normal com média μ e variância σ^2. Denote $(\theta_1, \theta_2) = \mu, \sigma^2$. Estimando os parâmetros μ e σ^2 pelo método dos momentos, igualamos*

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} = \mu_1 = E[X] = \mu$$
$$\bar{X} = \mu.$$

Igualando o segundo momento usando (2) *e* (4)*, temos*

$$M_2' = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \mu_2' = E[(X-\mu)^2] = \hat{\sigma}^2.$$

Assim, os estimadores de μ e σ^2 pelo método dos momentos são $\hat{\mu} = \bar{X}$ e $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. □

Exemplo 2. *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma distribuição Poisson com parâmetro λ. Há somente um parâmetro, então há somente uma equação, que é*

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} = \mu_1 = E[X] = \lambda$$
$$\bar{X} = \lambda.$$

Então o estimador de λ pelo método dos momentos é $\hat{\lambda} = \bar{X}$. □

Exemplo 3. *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma distribuição exponencial com densidade $f_X(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$. O estimador pelo método dos momentos é*

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} = \mu_1 = E[X] = \frac{1}{\theta}$$
$$\bar{X} = \frac{1}{\theta}.$$

Assim, $\hat{\theta} = \frac{1}{\bar{X}}$. □

2.2 Método da máxima verossimilhança

Definição 7 (Função de verossimilhança)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp conjunta $f_{\mathbf{X}}(\boldsymbol{x};\theta)$, com $\theta \in \Theta$ em que Θ é o espaço paramétrico. Considere ainda $x_1, x_2, \ldots, x_n$ a realização da amostra aleatória $X_1, X_2, \ldots, X_n$, então a função de verossimilhança é definida por*

$$L(\theta; x_1, x_2, \ldots, x_n) = L(\theta; \boldsymbol{x}) = f_{\mathbf{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta). \tag{7}$$

□

Definição 8 (Método da máxima verossimilhança)**.** *Seja uma função de verossimilhança $L(\theta; x_1, x_2, \ldots, x_n)$ para uma amostra aleatória $X_1, X_2, \ldots, X_n$. Então o método da máxima verossimilhança é a forma de encontrar um $\hat{\theta} = \vartheta(x_1, x_2, \ldots, x_n)$, uma função das observações $x_1, x_2, \ldots, x_n$, que é o valor estimado de $\theta \in \Theta$ que maximiza $L(\theta; x_1, x_2, \ldots, x_n)$. Dizemos que $\hat{\theta} = \vartheta(X_1, X_2, \ldots, X_n)$ é o estimador de máxima verossimilhança de θ.* □

Para maximizar $L(\theta; x_1, x_2, \ldots, x_n)$, tomamos a sua derivada em relação a θ, igualamos a zero e resolvemos o sistema para obtenção de $\hat{\theta} = \vartheta(X_1, X_2, \ldots, X_n)$. Posteriormente, devemos identificar se a segunda derivada de $L(\theta; x_1, x_2, \ldots, x_n)$ é negativa para saber se $\hat{\theta}$ é um ponto de máximo. Muitas vezes esse processo torna-se complicado. Uma solução é usar a função de Log-verossimilhança.

Definição 9 (Função de Log-verossimilhança)**.** *Se $L(\theta; x_1, x_2, \ldots, x_n)$, expressão* (7)*, é a função de verossimilhança, então*

$$l(\theta; \boldsymbol{x}) = l(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x}), \tag{8}$$

é a função de log-verossimilhança, para $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]'$. □

Como a função logaritmo é monótona crescente, então $l(\theta; \boldsymbol{x})$ e $L(\theta; \boldsymbol{x})$ levam ao mesmo máximo de θ. Se denotarmos $h(\theta) = L(\theta; \boldsymbol{x})$ e $g(y) = \log(y)$, temos

$$0 = \frac{d}{d\theta}(g \circ h(\theta)) = g'(h(\theta))h'(\theta) = \frac{h'(\theta)}{h(\theta)},$$

logo as raízes dessa expressão são as mesmas que $h'(\theta) = 0$. Assim, como é mais fácil fazer manipulações algébricas com a função logaritmo, o problema antes intratável, agora pode ser resolvido.

Considerando um caso geral, $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_k]'$, para determinar o estimador de máxima verossimilhança usando a função de log-verossimilhança, temos que usar a função escore dada por $U(\boldsymbol{\theta})$ com componentes dados por

$$U_j(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_j} = 0, \; j = 1, 2, \ldots, k, \tag{9}$$

Neste caso as condições de segunda ordem para garantir que a solução da função escore seja um ponto de máximo referem-se à matriz hessiana H da função de log-verossimilhança, isto é, a condição é de que a matriz

$$H = \left. \frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}}, \tag{10}$$

seja negativa definida, $\boldsymbol{z}'H\boldsymbol{z} < 0, \forall \boldsymbol{z} \neq \boldsymbol{0}$, sendo cada elemento de H dado por

$$h_{ij} = \frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_i \partial \theta_j}. \tag{11}$$

Exemplo 4 (Verossimilhança perfilhada, Bolfarine p. 47)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma normal com média μ e variância σ^2. Vamos determinar os estimadores desses parâmetros pelo método de máxima verossimilhança, denotando $\boldsymbol{\theta} = (\mu, \sigma^2)$. A função de log-verossimilhança pode ser dada por:*

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{x}) &= \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right) \\ &= \log \left(\left[\frac{1}{2\pi\sigma^2} \right]^{n/2} e^{-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}} \right) \\ &= \frac{n}{2} \log \left(\frac{1}{2\pi\sigma^2} \right) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}. \end{aligned}$$

Tomando-se as primeiras derivadas, temos

$$0 = \frac{\partial}{\partial \mu} l(\mu; \boldsymbol{x}, \sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})}{\sigma^2}.$$

Isolando $\hat{\mu}$ segue que

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.$$

Derivando $l(\sigma^2; \boldsymbol{x}, \mu)$ com relação a σ^2, temos

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

Isolando $\hat{\sigma}^2$ avaliado em $\hat{\mu}$ segue que

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}.$$

Para verificar se $\hat{\mu}$ e $\hat{\sigma}^2$ são os estimadores de máxima verossimilhança, calculemos a matriz H,

$$H = \left.\frac{\partial^2 l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = \begin{bmatrix} \frac{\partial^2}{\partial \mu^2} l(\boldsymbol{\theta}; \boldsymbol{x}) & \frac{\partial^2}{\partial \mu \partial \sigma^2} l(\boldsymbol{\theta}; \boldsymbol{x}) \\ \frac{\partial^2}{\partial \sigma^2 \partial \hat{\mu}} l(\boldsymbol{\theta}; \boldsymbol{x}) & \frac{\partial^2}{\partial (\sigma^2)^2} l(\boldsymbol{\theta}; \boldsymbol{x}) \end{bmatrix}.$$

Calculando as segundas derivadas

$$\frac{\partial^2}{\partial \mu^2} l(\boldsymbol{\theta}; \boldsymbol{x}) = \frac{-n}{\hat{\sigma}^2} < 0;$$

$$\frac{\partial^2}{\partial \mu \partial \sigma^2} l(\boldsymbol{\theta}; \boldsymbol{x}) = -\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})}{(\hat{\sigma}^2)^2} = 0.$$

$$\begin{aligned} \frac{\partial}{\partial (\sigma^2)^2} l(\boldsymbol{\theta}; \boldsymbol{x}) &= \frac{n}{2(\hat{\sigma}^2)^2} - \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{(\hat{\sigma}^2)^3} \\ &= \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\hat{\sigma}^2}{(\hat{\sigma}^2)^3} \\ &= \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n}{(\hat{\sigma}^2)^2} \\ &= -\frac{n}{2(\hat{\sigma}^2)^2} < 0. \end{aligned}$$

Logo, $\hat{\mu}$ e $\hat{\sigma}^2$ são os estimadores de máxima verossimilhança.

Exemplo 5. *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma normal com média μ e variância 1. Observe que $\hat{\mu} = \bar{X}$ é o estimador de máxima verossimilhança de μ. Ver exemplo anterior.*

Podemos também obter o estimador de máxima verossimilhança para uma de fdp ou fp que pertencem a família exponencial.

Definição 10 (Família Exponencial)**.** *Uma família de fdp ou fp $f_X(x;\theta)$ pertence a família exponencial se:*

I) Caso uniparamétrico (θ):

$$f_X(x;\theta) = a(\theta)b(x)\exp\{c(\theta)d(x)\} \tag{12}$$

II) Caso multiparamétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, $p \leq k$:

$$f_X(x;\boldsymbol{\theta}) = a(\boldsymbol{\theta})b(x)\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})d_j(x)\right\}, \tag{13}$$

em que a e d são funções de θ, c e b função de x que não dependem de θ. □

Exemplo 6. *Se $f_X(x;\theta) = \theta e^{-\theta x} I_{(0,\theta)}(x)$, então $f_X(x;\theta)$ pertence a família exponencial, pois $a(\theta) = \theta$, $b(x) = I_{(0,\infty)}(x)$, $c(\theta) = -\theta$ e $d(x) = x$. Observe que poderíamos reparametrizar a expressão* (12) *e dizer que $f_X(x;\theta)$ poderia ser membro da família exponencial se*

$$f_X(x;\theta) = \frac{b(x)}{a(\theta)}\exp\{c(\theta)d(x)\}, \tag{14}$$

de modo que $a(\theta) = 1/\theta$ pela nova reparametrização. Assim, queremos mostrar que diversas formas de reparametrização podem ser realizadas para membros da família exponencial.

2.2.1 Método da máxima verossimilhança para a família exponencial

Reparametrizando (13) para o caso multiparamétrico, temos

$$f_X(x;\boldsymbol{\theta}) = \frac{b(x)}{a(\boldsymbol{\theta})}\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})d_j(x)\right\}. \tag{15}$$

Sabemos que (15) é uma fdp ou fp. Considerando para o caso contínuo, temos que

$$\int f_X(x;\boldsymbol{\theta})dx = 1$$

$$\int \frac{b(x)}{a(\boldsymbol{\theta})}\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})d_j(x)\right\}dx = 1$$

$$a(\boldsymbol{\theta}) = \int b(x)\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})d_j(x)\right\}dx. \tag{16}$$

Assim $a(\boldsymbol{\theta})$ funciona como uma constante de normalização. A função de verossimilhança (15) é dada por:

$$L(\boldsymbol{\theta};\boldsymbol{x}) = \prod_{i=1}^{n} f_X(x_i;\boldsymbol{\theta}) = \frac{\prod_{i=1}^{n} b(x)}{a^n(\boldsymbol{\theta})}\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})\sum_{i=1}^{n} d_j(x)\right\}. \tag{17}$$

Aplicando o logaritmo em (17), temos

$$l(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{i=1}^{n} \log[b(x_i)] - n\log[a(\boldsymbol{\theta})] + \sum_{j=1}^{k} c_j(\boldsymbol{\theta}) \sum_{i=1}^{n} d_j(x). \tag{18}$$

Para determinar o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ devemos maximizar (18). Para isso, usaremos a função escore $U(\boldsymbol{\theta}) = \mathbf{0}$, Expressão (9), em que suas componentes são dadas por

$$0 = U_j(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta};\boldsymbol{x})}{\partial \theta_j} = -\frac{n}{a(\boldsymbol{\theta})}\left(\frac{\partial \boldsymbol{\theta}}{\partial \theta_j} a(\boldsymbol{\theta})\right) + \sum_{j=1}^{k}\left(\frac{\partial \boldsymbol{\theta}}{\partial \theta_j} c_j(\boldsymbol{\theta})\right) S_j(\boldsymbol{x}),\ j = 1, \ldots, k, \tag{19}$$

sendo $S_j(\boldsymbol{x}) = \sum_{i=1}^{n} d_j(x)$. Entretanto a expressão (19) geralmente são não lineares e têm que ser resolvidas numericamente por processos iterativos do tipo Newton-Raphson.

O método de Newton, também chamado de método Newton-Raphson, devido a Isaac Newton e Joseph Raphson, tem por objetivo encontrar aproximações para as raízes de uma função real, ou seja,

$$x : f(x) = 0.$$

Pode-se deduzir o algoritmo do método Newton-Raphson, baseado na Figura 1.

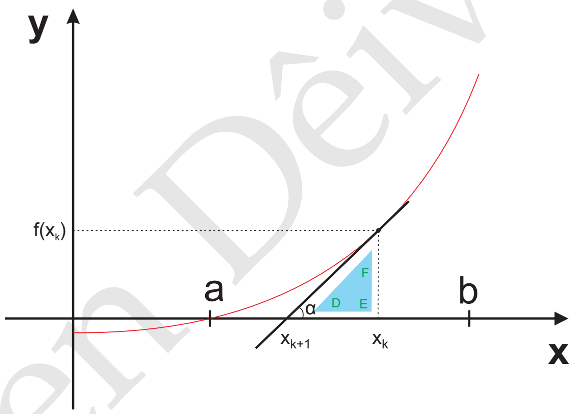

Figura 1: Forma geométrica do método Newton-Raphson

Suponha $f : [a, b] \to \mathbb{R}$ é uma função diferenciável definida no intervalo $[a, b]$ com valores nos reais $\mathbb{R}$. A fórmula para a convergência pode ser facilmente encontrada, pois a derivada da função f no ponto x_k é igual a tangente do ângulo α entre a reta tangente e a curva no ponto x_k. Usando a relação sobre o triângulo retângulo, tem-se:

$$\begin{aligned} f'(x) &= \tan(\alpha) = \frac{\Delta y}{\Delta x}, \\ &= \frac{f(x_k) - 0}{x_k - x_{k+1}}, \end{aligned}$$

em que f' é a derivada da função f. Assim, com uma simples álgebra, pode-se derivar

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ \ k = 0, 1, 2, \ldots. \tag{20}$$

Começando o processo com um valor arbitrário inicial x_0, em que quanto mais perto esse ponto for da raiz da função, mais rápido será a convergência da iteração, considerando $f'(x_0) \neq 0$. O valor da estimativa inicial (x_0) deve ser um ponto no qual a função tenha o mesmo sinal de sua derivada segunda.

Assim, para determinarmos a solução do sistema de equações em (19), usaremos a versão multivariada do método em (20), isto é, $x_{k+1} = \hat{\boldsymbol{\theta}}^{(m+1)}$, $x_k = \hat{\boldsymbol{\theta}}^{(m)}$, $f(x_k) = U(\hat{\boldsymbol{\theta}})^{(m)}$ e $f'(x_k) = U'^{(m)}(\hat{\boldsymbol{\theta}}) = \frac{\partial^2 l(\hat{\boldsymbol{\theta}};\boldsymbol{x})}{\partial\theta\partial\theta'} = \boldsymbol{H}^{(m)}$. Assim,

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} - [\boldsymbol{H}^{(m)}]^{-1}U(\hat{\boldsymbol{\theta}})^{(m)}, \tag{21}$$

sendo $\hat{\boldsymbol{\theta}}^{(m+1)}$ e $\hat{\boldsymbol{\theta}}^{(m)}$ os vetores de parâmetros estimados nos passos m e $m+1$.

Se considerarmos a matriz de informação observada de Fisher dada por $\boldsymbol{I}(\boldsymbol{\theta})$[1] $= -\frac{\partial^2 l(\hat{\boldsymbol{\theta}};\boldsymbol{x})}{\partial\theta\partial\theta'}$ com elementos $-\frac{\partial^2 l(\boldsymbol{\theta};\boldsymbol{x})}{\partial\theta_i\partial\theta_j}$, então (21) pode ser reescrito como

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} + [\boldsymbol{I}^{(m)}(\boldsymbol{\theta})]^{-1}U(\hat{\boldsymbol{\theta}})^{(m)}, \tag{22}$$

Quando as derivadas parciais de segunda ordem são avaliadas facilmente, o método Newton-Raphson é bastante útil. Quando há problemas na inversa da matriz Hessiana, pode-se utilizar a matriz de informação esperada de Fisher dada por $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 l(\hat{\boldsymbol{\theta}};\boldsymbol{x})}{\partial\theta\partial\theta'}\right]$. Assim, ao invés de utilizar a matriz hessiana ou a matriz de informação observada de Fisher, segue

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \hat{\boldsymbol{\theta}}^{(m)} + [\boldsymbol{\mathcal{I}}^{(m)}(\boldsymbol{\theta})]^{-1}U(\hat{\boldsymbol{\theta}})^{(m)}. \tag{23}$$

Os processos (21), (22) e (23) se encerram quando $|\hat{\boldsymbol{\theta}}^{(m+1)} - \hat{\boldsymbol{\theta}}^{(m)}| < \epsilon$, em que ϵ é especificado arbitrariamente.

2.2.2 Método da máxima verossimilhança para a família exponencial

Se considerarmos em (13) que $c(\boldsymbol{\theta}) = \boldsymbol{\theta}$, isto é, é um parâmetro natural, então poderemos encontrar o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ em função das estatísticas suficientes. Segue que

$$f_X(x;\boldsymbol{\theta}) = \frac{b(x)}{a(\boldsymbol{\theta})}\exp\left\{\sum_{j=1}^{k}\theta_j d_j(x)\right\}. \tag{24}$$

Sabemos que (24) é uma fdp ou fp. Considerando para o caso contínuo, temos que

$$\begin{aligned} \int f_X(x;\boldsymbol{\theta})dx &= 1 \\ \int \frac{b(x)}{a(\boldsymbol{\theta})}\exp\left\{\sum_{j=1}^{k}\theta_j d_j(x)\right\}dx &= 1 \\ a(\boldsymbol{\theta}) &= \int b(x)\exp\left\{\sum_{j=1}^{k}\theta_j d_j(x)\right\}dx. \end{aligned} \tag{25}$$

[1]As condições de regularidade referem-se à verossimilhanca ser derivável em todo o espaço paramétrico e à troca dos sinais de derivação e integração.

Assim $a(\boldsymbol{\theta})$ funciona como uma constante de normalização. A função de verossimilhança (15) é dada por:

$$L(\boldsymbol{\theta};\boldsymbol{x}) = \prod_{i=1}^{n} f_X(x_i;\boldsymbol{\theta}) = \frac{\prod_{i=1}^{n} b(x)}{a^n(\boldsymbol{\theta})} \exp\left\{\sum_{j=1}^{k} \theta_j \sum_{i=1}^{n} d_j(x_i)\right\}. \tag{26}$$

Aplicando o logaritmo em (77), temos

$$l(\boldsymbol{\theta};\boldsymbol{x}) = \sum_{i=1}^{n} \log[b(x_i)] - n\log[a(\boldsymbol{\theta})] + \sum_{j=1}^{k} \theta_j \sum_{i=1}^{n} d_j(x_i). \tag{27}$$

Para obtermos o estimador de máxima verossimilhança usamos a função escore. Assim,

$$0 = U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta};\boldsymbol{x})}{\partial \theta_j} = -\frac{n}{a(\boldsymbol{\theta})}\left(\frac{\partial}{\partial \theta_j} a(\boldsymbol{\theta})\right) + \sum_{i=1}^{n} d_j(x_i),\ j = 1, 2, \ldots, k. \tag{28}$$

Observe que, sob condições de regularidade[2], temos

$$\begin{aligned}
-\frac{n}{a(\boldsymbol{\theta})}\left(\frac{\partial}{\partial \theta_j} a(\boldsymbol{\theta})\right) &= -\frac{n}{a(\boldsymbol{\theta})}\left(\frac{\partial}{\partial \theta_j}\int b(x)\exp\left\{\sum_{j=1}^{k}\theta_j d_j(x)\right\} dx\right)\\
&= -\frac{n}{a(\boldsymbol{\theta})}\left(\int b(x)\frac{\partial}{\partial \theta_j}\exp\left\{\sum_{j=1}^{k}\theta_j d_j(x)\right\} dx\right)\\
&= -n\left(\int d_j(x)\,\underbrace{\frac{b(x)}{a(\boldsymbol{\theta})}\exp\left\{\sum_{j=1}^{k}\theta_j d_j(x)\right\}dx}_{f_X(x;\boldsymbol{\theta})}\right)\\
&= -nE[d_j(x)].
\end{aligned} \tag{29}$$

Aplicando (29) em (28), logo

$$E[d_j(x)] = \frac{\sum_{i=1}^{n} d_j(x_i)}{n} = \frac{S_j(\boldsymbol{x})}{n}, \tag{30}$$

em que $S_j(\boldsymbol{x}) = \sum_{i=1}^{n} d_j(x_i)$ é a j-ésima estatística suficiente. Não deve ser surpresa que o resultado envolve as observações somente via estatística suficiente. Isso dar um significado operacional para a suficiência: para a proposta de estimar os parâmetros usamos somente a estatística suficiente. Para distribuições em que $d_j(x) = x$, que inclui a distribuição de Bernoulli, distribuição de Poisson e a distribuição multinomial, o resultado (29) mostra que a média amostral é o estimador para a média.

Como a segunda derivada de (27) é negativa, então o estimador $\frac{\sum_{i=1}^{n} d_j(x_i)}{n}$ de θ_j tem ponto de máximo.

[2] As condições de regularidade referem-se à verossimilhanca ser derivável em todo o espaço paramétrico e à troca dos sinais de derivação e integração.

2.3 Método dos mínimos quadrados

O método dos mínimos quadrados consiste em estimar parâmetros de um modelo de regressão, expresso por

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}, \tag{31}$$

em que $\underset{\sim}{Y}$ é um vetor de dimensões $n \times 1$ da variável aleatória Y; $\underset{\sim}{X}$ é a matriz de dimensões $n \times p'$ conhecida do delineamento, assumindo que $n > p'$ e que $\underset{\sim}{X}$ é de posto completo p', sendo $p' = p + 1$; $\underset{\sim}{\theta}$ é o vetor de parâmetros de dimensão $p' \times 1$; $\underset{\sim}{\varepsilon}$ é o vetor de dimensões $n \times 1$ dos erros aleatórios. Assim, estes podem ser expressos da seguinte forma:

$$\underset{\sim}{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \underset{\sim}{X}_{n\times p'} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \underset{\sim}{\theta}_{p'\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ e}$$

$$\underset{\sim}{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

As pressuposições para esse modelo são:

1. $E[\underset{\sim}{\varepsilon}] = \underset{\sim}{0}$, sendo $\underset{\sim}{0}$ um vetor de zeros de dimensão $n \times 1$, ou equivalentemente $E[\underset{\sim}{Y}] = \underset{\sim}{X}\underset{\sim}{\theta}$;

2. $cov[\underset{\sim}{\varepsilon}] = \underset{\sim}{I}\sigma^2$, sendo $\underset{\sim}{I}$ uma matriz identidade de dimensão $n \times n$, ou equivalentemente $cov[\underset{\sim}{Y}] = \underset{\sim}{I}\sigma^2$;

3. $cov[\varepsilon_i, \varepsilon_j] = 0$ para todo $i \neq j$, ou equivalentemente, $cov[Y_i, Y_j] = 0$.

Teorema 1 (Método de mínimos quadrados de $\underset{\sim}{\theta}$). *Se $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\theta} + \underset{\sim}{\varepsilon}$, em que $\underset{\sim}{X}$ é $n \times p'$ de posto $p' < n$, então o valor de $\hat{\underset{\sim}{\theta}} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p]'$ que minimiza $\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$ é igual a*

$$\hat{\underset{\sim}{\theta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}. \tag{32}$$

Assim, $\hat{\underset{\sim}{\theta}}$ é conhecido como estimador de mínimos quadrados de $\underset{\sim}{\theta}$. □

Demonstração. Podemos escrever $\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$ como

$$\begin{aligned} \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} &= (\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\theta})'(\underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\theta}) \\ &= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{Y}'\underset{\sim}{X}\underset{\sim}{\theta} + \underset{\sim}{\theta}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\theta}. \end{aligned}$$

Para encontrarmos $\hat{\underset{\sim}{\theta}}$ que minimiza $\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$, calculamos a diferencial $\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$ em relação a $\underset{\sim}{\theta}$:

$$\frac{\partial \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}}{\partial \underset{\sim}{\theta}} = 0 - 2\underset{\sim}{X}'\underset{\sim}{Y} + 2\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\theta}.$$

Igualando a zero, obtemos o sistema de equações normais:

$$\underset{\sim}{X}'\underset{\sim}{X}\hat{\underset{\sim}{\theta}} = \underset{\sim}{X}'\underset{\sim}{Y}. \tag{33}$$

Como $\underset{\sim}{X}$ tem posto completo, $\underset{\sim}{X}'\underset{\sim}{X}$ é não singular e portanto invertível. Assim, a solução (33) é (32). □

Obviamente, que $\underset{\sim}{\hat{\theta}}$ é um ponto de mínimo, pois

$$\frac{\partial \underset{\sim}{\varepsilon}' \underset{\sim}{\varepsilon}}{\partial \underset{\sim}{\theta} \partial \underset{\sim}{\theta}'} = 2\underset{\sim}{X}' \underset{\sim}{X} > 0.$$

Exemplo 7. *Seja um modelo de regressão linear simples do tipo* $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ *para* $i = 1, 2, \ldots, n$*. De modo matricial, temos*

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Os termos matriciais podem ser expressos:

$$\underset{\sim}{X}' \underset{\sim}{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}, \; (\underset{\sim}{X}' \underset{\sim}{X})^{-1} = \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{bmatrix} \Rightarrow$$

$$\underset{\sim}{\hat{\theta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{bmatrix} \sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i \\ -\sum_{i=1}^n X_i \sum_{i=1}^n Y_i + n\sum_{i=1}^n X_i Y_i \end{bmatrix}$$

Assim, os estimadores de mínimos quadrados podem ser dados por

$$\begin{aligned} \hat{\beta}_1 &= \frac{-\sum_{i=1}^n X_i \sum_{i=1}^n Y_i + n\sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{n}{n} \\ &= \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} = \frac{SPXY}{SQX}. \end{aligned}$$

$$\begin{aligned} \hat{\beta}_0 &= \frac{\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{n}{n} \\ &= \frac{n\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2 \sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{n\sum_{i=1}^n X_i^2 \sum_{i=1}^n Y_i - \sum_{i=1}^n Y_i (\sum_{i=1}^n X_i)^2 + \sum_{i=1}^n Y_i (\sum_{i=1}^n X_i)^2 - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2 \sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{\sum_{i=1}^n Y_i [n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2] + \sum_{i=1}^n Y_i (\sum_{i=1}^n X_i)^2 - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2 \sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{\sum_{i=1}^n Y_i [n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]}{n[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]} + \frac{\sum_{i=1}^n Y_i (\sum_{i=1}^n X_i)^2 - n\sum_{i=1}^n X_i \sum_{i=1}^n X_i Y_i}{n^2 \sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i)^2} \\ &= \frac{\sum_{i=1}^n Y_i [n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]}{n[n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2]} + \frac{\sum_{i=1}^n Y_i (\sum_{i=1}^n X_i) - n\sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{\sum_{i=1}^n X_i}{n} \\ &= \frac{\sum_{i=1}^n Y_i}{n} - \frac{-\sum_{i=1}^n Y_i (\sum_{i=1}^n X_i) + n\sum_{i=1}^n X_i Y_i}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \times \frac{\sum_{i=1}^n X_i}{n} \\ &= \bar{Y} - \hat{\beta}_1 \bar{X}. \end{aligned}$$

3 Propriedade dos estimadores

Após apresentar alguns métodos de estimação pontual de parâmetros, nos perguntamos, qual dos estimadores de θ é o melhor? Sabemos que muitos desses métodos, apresentam os mesmos estimadores para um determinado parâmetro, ver Exemplo 1 e Exemplo 4. Outros apresentam resultados completamente diferentes.

Assim, precisamos de algum critério que possa nos informar qual dos estimadores é o que melhor estima θ.

Para uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x;\theta)$, $f_X(x;\theta)$, com $\theta \in \Theta$ em que Θ é o espaço paramétrico, poderíamos obter uma escolha inicial do melhor estimador $T = t(X_1, X_2, \ldots, X_n)$ de $\tau(\theta)$ com base na seguinte Definição,

Definição 11 (Estimador não-viesado)**.** *Um estimador $T = t(X_1, X_2, \ldots, X_n)$ é definido um estimador não-viesado de $\tau(\theta)$ se e somente se*

$$E_\theta[T] = E[t(X_1, X_2, \ldots, X_n)] = \tau(\theta), \forall \theta \in \Theta. \tag{34}$$

□

Entretanto, existe uma classe muito grande de estimadores não viesados. Uma outra Definição é baseado no princípio da suficência.

Definição 12 (Estatística suficiente)**.** *Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória com fdp ou fp $f_X(x;\theta)$, com $\theta \in \Theta$, sendo Θ o espaço paramétrico. Uma estatística $S = s(X_1, X_2, \ldots, X_n)$ é suficiente se e somente se, a distribuição condicional de $X_1, X_2, \ldots, X_n$ dado $S = s(x_1, x_2, \ldots, x_n)$ não depende de θ.* □

O princípio da suficiência diz que se S é uma estatística suficiente para θ, então qualquer inferência sobre θ deverá depender da amostra $X_1, X_2, \ldots, X_n$ somente pelo valor de $S = s(X_1, X_2, \ldots, X_n)$.

Exemplo 8. *Material escrito... (Exemplo da Binomial)*

Determinar uma estatística suficiente pela Definição 12 não é nada fácil. Mas, observe a seguinte afirmação da Definição 12 em níveis probabilísticos, sendo uma amostra $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ com fdp ou fp $f_X(x;\theta)$ sendo $S = s(X_1, X_2, \ldots, X_n) = s(\boldsymbol{x})$ uma estatística suficiente

$$f_{\boldsymbol{X}|S(\boldsymbol{X})}(x|s(x_1, \ldots, x_n);\theta) = \frac{f_{\boldsymbol{X},S(\boldsymbol{X})}(\boldsymbol{x}, s(\boldsymbol{x});\theta)}{f_{S(\boldsymbol{X})}(\boldsymbol{x};\theta)} = h(\boldsymbol{x}),$$

isto é, $h(\boldsymbol{x})$ é o resultado da distribuição de $\boldsymbol{X}$ dado $S(\boldsymbol{x})$ que não depende de θ, mas da amostra. Percebendo $f_{\boldsymbol{X},S(\boldsymbol{X})}(\boldsymbol{x}, s(\boldsymbol{x});\theta) = f_{\boldsymbol{X}}(\boldsymbol{x};\theta)$, poderíamos então expressar a distribuição conjunta de $X_1, X_2, \ldots, X_n$ da seguinte forma:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = f_{S(\boldsymbol{X})}(\boldsymbol{x};\theta)h(\boldsymbol{x}).$$

Daí, poderemos obter uma estatística suficiente facilmente.

Teorema 2 (Critério de fatoração de Neyman-Fisher - estatística suficiente simples)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x;\theta)$, com $\theta \in \Theta$ em que Θ é o espaço* *paramétrico e θ pode ser um vetor. Então a estatística $S = s(X_1, X_2, \ldots, X_n)$ é suficiente se e somente se a fdp ou fp conjunta de $X_1, X_2, \ldots, X_n$ fatorar como*

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta) = h(x_1, x_2, \ldots, x_n) g_\theta(s(x_1, x_2, \ldots, x_n); \theta), \tag{35}$$

em que $h(.)$ é uma função não negativa que não depende de θ e $g_\theta(.)$ uma função não negativa que depende de $X_1, X_2, \ldots, X_n$ através de $S = s(X_1, X_2, \ldots, X_n)$. □

Demonstração. Vamos provar para o caso discreto. Suponha que $S = s(X_1, X_2, \ldots, X_n)$ é uma estatística suficiente. A escolha para $g_\theta(s(x_1, x_2, \ldots, x_n) = P_\theta(S = s(x_1, x_2, \ldots, x_n))$ e $h(x_1, x_2, \ldots, x_n) = P((X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)|S = s(x_1, x_2, \ldots, x_n))$ que não depende de θ. Assim, denotando $\boldsymbol{X} = X_1, X_2, \ldots, X_n$ e $\boldsymbol{x} = x_1, x_2, \ldots, x_n$, temos

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) &= P_\theta(\boldsymbol{X} = \boldsymbol{x}) \\ &= P_\theta(\boldsymbol{X} = \boldsymbol{x} \text{ e } S(\boldsymbol{X}) = s(\boldsymbol{x})). \end{aligned}$$

Isso sempre ocorre quando usamos uma estatística é suficiente. Pense num experimento de Bernoulli em que temos uma amostra de três elementos e o resultado é:

$$\boldsymbol{X} = (X_1 = 0, X_2 = 1, X_3 = 1).$$

Considere que $S(\boldsymbol{X}) = \sum_{i=1}^{3} X_i = 2$ é uma estatística suficiente. Então diversos arranjos X_1, X_2, X_3 são possíveis para que $S = 2$,

$$\begin{cases} X_1 = 0, X_2 = 1, X_3 = 1 \Rightarrow S = 2 \\ X_1 = 1, X_2 = 0, X_3 = 1 \Rightarrow S = 2 \\ X_1 = 1, X_2 = 1, X_3 = 0 \Rightarrow S = 2 \end{cases}$$

Com esse exemplo, percebemos então que $\{\boldsymbol{X} = \boldsymbol{x}\} = \{\boldsymbol{X} = \boldsymbol{x}\} \cap \{S(\boldsymbol{X}) = s(\boldsymbol{x})\}$

Dessa forma,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) &= P_\theta(\boldsymbol{X} = \boldsymbol{x} \text{ e } S(\boldsymbol{X}) = S(\boldsymbol{s})) \\ &= P((\boldsymbol{X}) = (\boldsymbol{x})|S(\boldsymbol{X}) = s(\boldsymbol{x}))P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x})) \qquad \text{(Suficência)} \\ &= h(\boldsymbol{x}) g_\theta(s(\boldsymbol{x})). \end{aligned}$$

Agora, assumimos que a fatoração (35) existe. Vamos provar agora que $P((\boldsymbol{X}) = (\boldsymbol{x})|S(\boldsymbol{X}) = s(\boldsymbol{x}))$ não depende de θ. Temos,

$$\begin{aligned} P((\boldsymbol{X}) = (\boldsymbol{x})|S(\boldsymbol{X}) = s(\boldsymbol{x})) &= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x} \text{ e } S(\boldsymbol{X}) = s(\boldsymbol{s}))}{P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x}))} \\ &= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x})}{P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x}))} \\ &= \frac{h(\boldsymbol{x}) g_\theta(s(\boldsymbol{x}))}{P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x}))} \qquad \text{(uma vez que (35) é safisfeito)} \end{aligned}$$

Considerando que (35) é satisfeito, então podemos expressar $P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x}))$ da seguinte forma:

$$P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x})) = \sum_{\{\boldsymbol{x}:S(\boldsymbol{X})=s(\boldsymbol{x})\}} h(\boldsymbol{x})g_\theta(s(\boldsymbol{x})).$$

Pense num experimento de Bernoulli em que temos uma amostra de três elementos e o resultado é:

$$\boldsymbol{X} = (X_1 = 0, X_2 = 1, X_3 = 1).$$

Considere que $S(\boldsymbol{X}) = \sum_{i=1}^{3} X_i = 2$ é uma estatística suficiente. Então diversos arranjos X_1, X_2, X_3 para S,

$$\begin{cases} A_1 = \{X_1 = 0, X_2 = 1, X_3 = 1 \Rightarrow S = 2\} \\ A_2 = \{X_1 = 1, X_2 = 0, X_3 = 1 \Rightarrow S = 2\} \\ A_3 = \{X_1 = 1, X_2 = 1, X_3 = 0 \Rightarrow S = 2\} \\ A_4 = \{X_1 = 0, X_2 = 0, X_3 = 0 \Rightarrow S = 0\} \\ A_5 = \{X_1 = 1, X_2 = 0, X_3 = 0 \Rightarrow S = 1\} \\ A_6 = \{X_1 = 0, X_2 = 1, X_3 = 0 \Rightarrow S = 1\} \\ A_7 = \{X_1 = 0, X_2 = 0, X_3 = 1 \Rightarrow S = 1\} \\ A_8 = \{X_1 = 1, X_2 = 1, X_3 = 1 \Rightarrow S = 3\} \end{cases}$$

Entretanto, queremos calcular apenas os eventos em que $S(\boldsymbol{X}) = \sum_{i=1}^{3} X_i = 2$, assim em termos de probabilidade, temos

$$\begin{aligned} P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x})) &= P(A_1) + P(A_2) + P(A_3) \\ &= \sum_{\{\boldsymbol{x}:S(\boldsymbol{X})=2\}} P(\boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{\{\boldsymbol{x}:S(\boldsymbol{X})=s(\boldsymbol{x})\}} P(\boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{\{\boldsymbol{x}:S(\boldsymbol{X})=s(\boldsymbol{x})\}} h(\boldsymbol{x})g_\theta(s(\boldsymbol{x})). \end{aligned}$$

Portanto,

$$\begin{aligned} P((\boldsymbol{X}) = (\boldsymbol{x}) | S(\boldsymbol{X}) = s(\boldsymbol{x})) &= \frac{h(\boldsymbol{x})g_\theta(s(\boldsymbol{x}))}{P_\theta(S(\boldsymbol{X}) = s(\boldsymbol{x}))} \\ &= \frac{h(\boldsymbol{x})g_\theta(s(\boldsymbol{x}))}{\sum_{\{\boldsymbol{x}:S(\boldsymbol{X})=s(\boldsymbol{x})\}} h(\boldsymbol{x})g_\theta(s(\boldsymbol{x}))} \\ &= \frac{h(\boldsymbol{x})}{\sum_{\{\boldsymbol{x}:S(\boldsymbol{X})=s(\boldsymbol{x})\}} h(\boldsymbol{x})}, \end{aligned}$$

como a proporção não depende de θ, $S(\boldsymbol{X})$ é uma estatística suficiente para θ. Prova concluída. $\square$

Teorema 3 (Critério de fatoração de Neyman-Fisher - estatísticas suficientes conjuntas)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x; \boldsymbol{\theta})$, com $\boldsymbol{\theta} \in \Theta$ em que Θ é o espaço paramétrico e $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_d]'$. Então um conjunto de estatísticas $S_j = s_j(X_1, X_2, \ldots, X_n)$, $j = 1, 2, \ldots, k$, é conjuntamente suficientes se e somente se a fdp ou fp conjunta de $X_1, X_2, \ldots, X_n$ fatorar como*

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i; \boldsymbol{\theta}) = h(\boldsymbol{x}) g_{\boldsymbol{\theta}}(s_1(\boldsymbol{x}), \ldots, s_k(\boldsymbol{x}); \boldsymbol{\theta}), \tag{36}$$

em que $h(.)$ é uma função não negativa que não depende de $\boldsymbol{\theta}$ e $g_{\boldsymbol{\theta}}(.)$ uma função não negativa que depende de $X_1, X_2, \ldots, X_n$ através de $S_j = s_j(X_1, X_2, \ldots, X_n)$ e $d \leq k$. □

Algo interessante é que o número de estatísticas suficientes não corresponde ao número de parâmetros necessariamente. Assim, se considerarmos $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, com $d \leq k$, sendo $S_1(\boldsymbol{X}), S_1(\boldsymbol{X}), \ldots, S_k(\boldsymbol{X})$, então o número de estatísticas conjuntamente suficientes é no mínimo igual ao número de parâmetros. (Exemplo 5.2.15, Casella, port. p.251).

O critério de fatoração pode ser estendido para uma classe de distribuições da família exponencial.

Teorema 4 (Estatística suficiente para a família exponencial)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ com fdp ou fp $f_X(x; \theta)$ pertencente a família exponencial, com $\theta \in \Theta$ em que Θ é o espaço paramétrico e θ pode ser um vetor. Então a estatística $S = s(X_1, X_2, \ldots, X_n)$ é suficiente se e somente se a fdp ou fp conjunta de $X_1, X_2, \ldots, X_n$ fatorar como:*

I) Caso uniparamétrico:

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f_X(x_i; \theta) &= a^n(\theta) \left[\prod_{i=1}^{n} b(x_i)\right] \exp\left\{c(\theta) \sum_{i=1}^{n} d(x_i)\right\} \\ &= h(\boldsymbol{x}) g_{\theta}(s(\boldsymbol{x}); \theta), \end{aligned} \tag{37}$$

em que a e d são funções de θ, c e b função de X que não dependem de θ, sendo $S = \sum_{i=1}^{n} d(X_i)$ é um estatística suficiente;

II) Caso múltiparamétrico:

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i; \boldsymbol{\theta}) = a^n(\boldsymbol{\theta}) \left[\prod_{i=1}^{n} b(x_i)\right] \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) \sum_{i=1}^{n} d_j(x_i)\right\}; \tag{38}$$

em que a e d são funções de θ, c e b função de X que não dependem de θ, sendo $S_1 = \sum_{i=1}^{n} d_1(x_i)$, $S_2 = \sum_{i=1}^{n} d_2(x_i)$, ..., $S_k = \sum_{i=1}^{n} d_k(x_i)$ um conjunto de estatísticas suficientes. □

Demonstração. Para o caso uniparamétrico, temos

$$
\begin{aligned}
\prod_{i=1}^{n} f_X(x_i;\theta) &= a^n(\theta)\left[\prod_{i=1}^{n} b(x_i)\right]\exp\left\{c(\theta)\sum_{i=1}^{n} d(x_i)\right\}\\
&= \left[\prod_{i=1}^{n} b(x_i)\right]\exp\{n\log[a(\theta)]\}\exp\left\{c(\theta)\sum_{i=1}^{n} d(x_i)\right\}\\
&= \left[\prod_{i=1}^{n} b(x_i)\right]\exp\left\{c(\theta)\sum_{i=1}^{n} d(x_i)+n\log[a(\theta)]\right\}\\
&= \left[\prod_{i=1}^{n} b(x_i)\right]\exp\left\{c(\theta)S(\boldsymbol{x})+n\log[a(\theta)]\right\}\\
&= h(\boldsymbol{x})g_\theta(s(\boldsymbol{x})),
\end{aligned}
$$

sendo $h(\boldsymbol{x}) = [\prod_{i=1}^{n} b(x_i)]$, $g_\theta(s(\boldsymbol{x})) = \exp\{c(\theta)S(\boldsymbol{x}) + n\log[a(\theta)]\}$ e $S(\boldsymbol{x}) = \sum_{i=1}^{n} d(x_i)$. Pelo critério de fatoração, Teorema 2, $S(\boldsymbol{x}) = \sum_{i=1}^{n} d(x_i)$ é uma estatística suficiente. Para o caso multiparamétrico, temos

$$
\begin{aligned}
\prod_{i=1}^{n} f_X(x_i;\boldsymbol{\theta}) &= a^n(\boldsymbol{\theta})\left[\prod_{i=1}^{n} b(x_i)\right]\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})\sum_{i=1}^{n} d_j(x_i)\right\}\\
&= \left[\prod_{i=1}^{n} b(x_i)\right]\exp\{n\log[a(\boldsymbol{\theta})]\}\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})\sum_{i=1}^{n} d_j(x_i)\right\}\\
&= \left[\prod_{i=1}^{n} b(x_i)\right]\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})\sum_{i=1}^{n} d_j(x_i)+n\log[a(\boldsymbol{\theta})]\right\}\\
&= \left[\prod_{i=1}^{n} b(x_i)\right]\exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})S_j(\boldsymbol{x})+n\log[a(\boldsymbol{\theta})]\right\}\\
&= h(\boldsymbol{x})g_{\boldsymbol{\theta}}(s_1(\boldsymbol{x}),\ldots,s_k(\boldsymbol{x})),
\end{aligned}
$$

sendo $h(\boldsymbol{x}) = [\prod_{i=1}^{n} b(x_i)]$, $g_{\boldsymbol{\theta}}(s_1(\boldsymbol{x}),\ldots,s_k(\boldsymbol{x}))$ e $S(\boldsymbol{x}) = \sum_{i=1}^{n} d(x_i) = \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})S_j(\boldsymbol{x}) + n\log[a(\boldsymbol{\theta})]\right\}$, sendo $S_1 = \sum_{i=1}^{n} d_1(x_i)$, $S_2 = \sum_{i=1}^{n} d_2(x_i)$, ..., $S_k = \sum_{i=1}^{n} d_k(x_i)$ um conjunto de estatísticas suficientes, pelo critério de fatoração, Teorema 3. □

Se $S = s(X_1, X_2, \ldots, X_n)$ é uma estatística suficiente, existe uma função $h(.)$ e uma estatística T tal que $S = h(T)$, em que T não pode conter menos informação de θ que S, sendo que T também é uma estatística suficiente. Além disso S fornece um maior grau de compreensão dos dados do que T, a menos que h seja uma função 1-a-1, nesse caso S e T são equivalentes.

Exemplo 9. *Seja $X_1, X_2, \ldots, X_n$ iid $N(\mu, \sigma^2)$. A densidade conjunta é*

$$
f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\mu^2\right\}. \tag{39}
$$

Pelo Teorema da fatoração $S = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$ é equivalente a $S' = (\bar{X}, S^2)$, em que $\bar{X} = \sum_{i=1}^{n} X_i/n$ e $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. □

Exemplo 10. *Seja* $X_1, X_2, \ldots, X_n$ *iid* $N(0, \sigma^2)$. *Então as estatísticas*

$$\begin{aligned}
S_1(\boldsymbol{X}) &= (X_1, X_2, \ldots, X_n) \\
S_2(\boldsymbol{X}) &= (X_1^2, X_2^2, \ldots, X_n^2) \\
S_3(\boldsymbol{X}) &= (X_1^2 + X_2^2 + \ldots + X_m^2,\ X_{m+1}^2 + \ldots + X_n^2) \\
S_4(\boldsymbol{X}) &= X_1^2 + X_2^2 + \ldots + X_n^2,
\end{aligned}$$

são todas suficientes para σ^2. S_i *fornece um grau de compreensão dos dados à medida que i cresce.* □

É natural se perguntar, dado que S é uma estatística suficiente que condensa os dados sem perder a informação do parâmetro, existe algum S que condense os dados mais do que qualquer outra estatística suficiente?

Definição 13 (Estatística suficiente mínima)**.** *Uma estatística suficiente S é mínima se para qualquer estatística suficiente* S' *existe uma função h tal que* $S = h(S')$. □

Essas informações sobre a estatística suficiente serão extremamente importante na sequência, pois a partir dela, iremos obter melhores estimadores.

Definição 14 (Estimador Não Viesado de Variância Mínima Uniformemente (UNVVMU))**.** *Seja uma amostra aleatória* $X_1, X_2, \ldots, X_n$ *com fdp ou fp* $f_X(x;\theta)$. *Um estimador* $W = w(X_1, X_2, \ldots)$ *de* $\tau(\theta)$ *é definido como um estimador não viesado de variância mínima uniformemente se*

i) $E_\theta[W] = \tau(\theta)$;

ii) $Var_\theta[W] \leq Var_\theta[W^*]$, *para qualquer outro estimador não viesado* W^*.

O grande problema na classe dos estimadores não viesados de $\tau(\theta)$, é saber o que tem menor variância. Entretanto, o Teorema a seguir mostra que existe um limite inferior para a variância dos estimadores. Assim, se um estimador atinge esse limite, ele é o melhor estimador de variância mínima uniformente de $\tau(\theta)$. Assumiremos a prova para o próximo Teorema para o caso contínuo. A desigualdade de Cramer-Rao também se aplica para o caso de variáveis aleatórias discretas. Neste caso, consideraremos $f(\mathbf{x}|\theta)$ a função de probabilidade ao invés da função densidade e, observamos que basta substituir a integral pelo somatório. Apesar da função de probabilidade não ser diferenciável em $\boldsymbol{x}$, ela o é em θ.

Teorema 5. ***(Desigualdade de Cramér-Rao)*** *Seja* $X_1, ..., X_n$ *uma amostra aleatória com fdp* $f(\mathbf{x}|\theta)$, *e seja* $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ *qualquer estimador que satisfaça*

$$\frac{d}{d\theta} E_\theta [W(\mathbf{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial \theta} [W(\mathbf{x}) f(\mathbf{x}|\theta)]\, d\mathbf{x}, \text{ sendo } \mathcal{X} \text{ o suporte de } \boldsymbol{X}, \tag{40}$$

e

$$Var_\theta [W(\mathbf{X})] < \infty.$$

Então

$$Var_\theta [W(\mathbf{X})] \geq \frac{\left(\frac{d}{d\theta} E_\theta [W(\mathbf{X})]\right)^2}{E_\theta \left[\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}|\theta)\right)^2\right]}. \tag{41}$$

Demonstração. A prova desse Teorema é elegantemente simples e utiliza aplicação da desigualdade de Cauchy-Schwarz. Considere duas variáveis aleatórias X e Y contínuas,

$$[COV\,(X, Y)]^2 \leq Var\,[X]\,Var\,[Y] \tag{42}$$

rearranjando a expressão (42), podemos obter um limite inferior para a variância de X dado por

$$Var\,[X] \geq \frac{[COV\,(X, Y)]^2}{Var\,[Y]}. \tag{43}$$

A chave desse Teorema segue da escolha de X como sendo o estimador $W\,(\mathbf{X})$ e Y como sendo a quantidade $\frac{\partial}{\partial\theta}\log f\,(\mathbf{X}\,|\theta)$ e aplicando na desigualdade (43).

Primeiro, note que

$$\begin{aligned}
\frac{d}{d\theta}E_\theta\,[W\,(\mathbf{X})] &= \frac{d}{d\theta}\left(\int \dots \int W\,(x_1, \dots, x_n)\, f_{X_1,\dots,X_n}\,(x_1, \dots, x_n\,|\theta) dx_1, \dots, dx_n\right)\\
&= \int \dots \int W\,(x_1, \dots, x_n)\,\frac{d}{d\theta} f_{X_1,\dots,X_n}\,(x_1, \dots, x_n\,|\theta) dx_1, \dots, dx_n\\
&= \int \dots \int W\,(\mathbf{x}) \left[\frac{d}{d\theta} f_{\mathbf{x}}\,(\mathbf{x}\,|\theta)\right] \frac{f_{\mathbf{X}}\,(\mathbf{x}\,|\theta)}{f_{\mathbf{X}}\,(\mathbf{x}\,|\theta)} dx_1, \dots, dx_n\\
&= E_\theta\left[W\,(\mathbf{X})\,\frac{\frac{d}{d\theta} f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)}{f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)}\right]\\
&= E_\theta\left[W\,(\mathbf{X})\,\frac{d}{d\theta}\log\left(f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)\right)\right]
\end{aligned}$$

o qual sugere uma covariância entre $W\,(\mathbf{X})$ e $\frac{d}{d\theta}\log\left(f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)\right)$. Para isso ser uma covariância, é necessário subtrair o produto dos valores esperados, isto é,

$$\begin{aligned}
Cov_\theta\left[W\,(\mathbf{X})\,\frac{d}{d\theta}\log\left(f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)\right)\right] = E_\theta&\left[W\,(\mathbf{X})\,\frac{d}{d\theta}\log\left(f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)\right)\right]\\
&- E_\theta\,[W\,(\mathbf{X})]\,E_\theta\left[\frac{d}{d\theta}\log\left(f_{\mathbf{X}}\,(\mathbf{X}\,|\theta)\right)\right].
\end{aligned}$$

Entretanto, observe que

$$\begin{aligned}
\int_{-\infty}^{\infty} f'\,(x)\,dx = \int_{-\infty}^{\infty}\frac{f'\,(x)}{f\,(x)} f\,(x)\,dx &= \int_{-\infty}^{\infty}\frac{d}{dx}\log\left(f\,(x)\right) f\,(x)\,dx\\
&= E\left[\frac{d}{dx}\log\left(f\,(x)\right)\right]
\end{aligned}$$

como $f\,(x;\,\theta)$ é fdp,

$$\begin{aligned}
\int_{-\infty}^{\infty}\frac{d}{d\theta} f\,(x;\,\theta)\,dx = \int_{-\infty}^{\infty}\frac{\frac{d}{d\theta} f\,(x;\,\theta)}{f\,(x;\,\theta)} f\,(x;\,\theta)\,dx &= \int_{-\infty}^{\infty}\left[\frac{d}{d\theta}\log\left(f\,(x;\,\theta)\right)\right] f\,(x;\,\theta)\,dx\\
&= E_\theta\left[\frac{d}{d\theta}\log\left(f\,(X;\,\theta)\right)\right]
\end{aligned}$$

ainda, note que

$$\int_{-\infty}^{\infty} \frac{d}{d\theta} f\left(x;\,\theta\right) dx = \frac{d}{d\theta} \int_{-\infty}^{\infty} f\left(x;\,\theta\right) dx = \frac{d}{d\theta}\left(1\right) = 0$$

logo, a v.a. $\frac{d}{d\theta} \log\left(f\left(X;\,\theta\right)\right)$ tem média zero, para qualquer que seja o parâmetro θ.

Portanto $COV\left[W\left(\mathbf{X}\right), \frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right]$ é igual a esperança do produto, logo

$$COV\left[W\left(\mathbf{X}\right), \frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right] = E_{\theta}\left[W\left(\mathbf{X}\right) \frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right] = \frac{d}{d\theta} E_{\theta}\left[W\left(\mathbf{X}\right)\right]. \tag{44}$$

Também, uma vez que $E_{\theta}\left[Y = \frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right] = 0$. Sabendo que $Var\left[Y\right] = E\left[Y^2\right] - \left(E\left[Y\right]\right)^2$ temos

$$Var_{\theta}\left[\frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right] = E_{\theta}\left[\left(\frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right)^2\right] \tag{45}$$

Usando a desigualdade de Cauchy-Schwarz juntamente com 44 e 45, obtemos

$$Var_{\theta}\left[W\left(\mathbf{X}\right)\right] \geq \frac{\left(\frac{d}{d\theta} E_{\theta}\left[W\left(\mathbf{X}\right)\right]\right)^2}{E_{\theta}\left[\left(\frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right)^2\right]}$$

provando o teorema. □

A prova do Teorema 5 foi demonstrado para o caso contínuo, sendo que para o caso discreto a prova é análoga. Se adicionarmos a suposição de amostras independentes, então o cálculo do limite inferior é simplificado. A esperança no denominador da expressão 41 recai a cálculos univariados, conforme será mostrado no corolário a seguir.

Corolário 1. ***(Caso iid para a Desigualdade de Cramér-Rao)*** *Se as suposiçõs do Teorema 5 são satisfeitas e considerando agora que $X_1, \ldots, X_n$ são iid (independentes e identicamente distribuídas) com pdf $f\left(x\left|\theta\right.\right)$, então*

$$Var_{\theta}\left[W\left(\mathbf{X}\right)\right] \geq \frac{\left(\frac{d}{d\theta} E_{\theta}\left[W\left(\mathbf{X}\right)\right]\right)^2}{nE_{\theta}\left[\left(\frac{d}{d\theta} \log\left(f_{X}\left(X\left|\theta\right.\right)\right)\right)^2\right]}. \tag{46}$$

Demonstração. Precisamos mostrar apenas que

$$E_{\theta}\left[\left(\frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(\mathbf{X}\left|\theta\right.\right)\right)\right)^2\right] = nE_{\theta}\left[\left(\frac{d}{d\theta} \log\left(f_{\mathbf{X}}\left(X\left|\theta\right.\right)\right)\right)^2\right]$$

Usando o fato de a amostra $X_1, \ldots, X_n$ ser independente, temos que

$$\begin{aligned}\frac{d}{d\theta} \log\left(f_{X_1,\ldots,X_n}\left(X_1,\,\ldots,\,X_n;\,\theta\right)\right) &= \frac{d}{d\theta} \log\left(\prod_{i=1}^{n} f\left(X_i;\,\theta\right)\right)\\ &= \frac{d}{d\theta} \sum_{i=1}^{n} \log\left(f\left(X_i;\,\theta\right)\right)\end{aligned}$$

elevando ambos os membros ao quadrado

$$
\begin{aligned}
\left[\frac{d}{d\theta}\log\left(f_{X_1,\ldots,X_n}\left(X_1,\,\ldots,\,X_n;\,\theta\right)\right)\right]^2 &= \left[\frac{d}{d\theta}\sum_{i=1}^{n}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\\
&= \sum_{i=1}^{n}\left[\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\\
&+ 2\sum_{i<j}\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\frac{d}{d\theta}\log\left(f\left(X_j;\,\theta\right)\right)
\end{aligned}
$$

note que a passagem da primeira expressão para a segunda se deve ao fato de aparecer somas de termos quadráticos e somas de termos com produtos cruzados. Aplicando a esperança em ambos os lados temos que

$$
\begin{aligned}
&E_\theta\left[\left[\frac{d}{d\theta}\log\left(f_{X_1,\ldots,X_n}\left(X_1,\,\ldots,\,X_n;\,\theta\right)\right)\right]^2\right] = E_\theta\left[\left[\frac{d}{d\theta}\sum_{i=1}^{n}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\right]\\
&= E_\theta\left[\sum_{i=1}^{n}\left[\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\right] + 2E_\theta\left[\sum_{i<j}\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\frac{d}{d\theta}\log\left(f\left(X_j;\,\theta\right)\right)\right]\\
&= \sum_{i=1}^{n}E_\theta\left[\left[\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\right] + 2\sum_{i<j}E_\theta\left[\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\right]E_\theta\left[\frac{d}{d\theta}\log\left(f\left(X_j;\,\theta\right)\right)\right]
\end{aligned}
$$

Note que da expressão acima o produto das esperanças é zero, uma vez que as variáveis aleatórias *i* e *j* são independentes. Note ainda que o termo $\sum_{i=1}^{n}E_\theta\left[\left[\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\right]$ nada mais é do que a soma da esperança de uma mesma variável aleatória. Portanto,

$$
\sum_{i=1}^{n}E_\theta\left[\left[\frac{d}{d\theta}\log\left(f\left(X_i;\,\theta\right)\right)\right]^2\right] = nE_\theta\left[\left[\frac{d}{d\theta}\log\left(f\left(X;\,\theta\right)\right)\right]^2\right]
$$

finalizando a prova do corolário. □

A cota de Cramér-Rao foi apresentada para variáveis contínuas, mas também é aplicada à variáveis aleatórias discretas.

A quantidade $\mathcal{I}\left(\theta\right) = E_\theta\left[\left[\frac{d}{d\theta}\log\left(f\left(X;\,\theta\right)\right)\right]^2\right]$ é chamada de *matriz de informação de Fisher* ou *número de informação de Fisher* da amostra. Essa terminologia reflete o fato de que o número de informação fornece um limite para a variância do melhor estimador não viesado de θ. Conforme o número de informação se torna maior e temos mais informação sobre θ, temos um menor limite para a variância do melhor estimador não viesado.

Teorema 6. *Seja $X_1, \ldots, X_n$ uma amostra com fdp $f\left(\mathbf{x}\,|\theta\right)$, e seja $W\left(\mathbf{X}\right) = W\left(X_1,\,\ldots,\,X_n\right)$ qualquer estimador que satisfaça*

$$
\frac{d}{d\theta}E_\theta\left[W\left(\mathbf{X}\right)\right] = \int_{\mathcal{X}}\frac{\partial}{\partial\theta}\left[W\left(\mathbf{x}\right)f\left(\mathbf{x}\,|\theta\right)\right]d\mathbf{x},\ \textit{sendo } \mathcal{X} \textit{ o suporte de } \mathbf{X}, \tag{47}
$$

e

$$Var_\theta\left[W\left(\mathbf{X}\right)\right] < \infty.$$

Então

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_X(X|\theta)\right)^2\right] = -E_\theta\left[\left(\frac{\partial^2}{\partial\theta^2}\log f_X(X|\theta)\right)\right] \tag{48}$$

Demonstração. Considere a função $l'(x;\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{f'(X;\theta)}{f(x;\theta)}$. Então

- $\int f'(x;\theta)dx = \int \frac{\partial}{\partial\theta}f(x;\theta)dx = \frac{\partial}{\partial\theta}\underbrace{\int f(x;\theta)dx}_{=1} = 0.$

- $\int f''(x;\theta)dx = \int \frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\underbrace{\int f(x;\theta)dx}_{=1} = 0.$

- $\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{\partial}{\partial\theta}\left[\frac{f'(x;\theta)}{f(x;\theta)}\right] = \frac{f''(x;\theta)f(x;\theta)-[f'(x;\theta)]^2}{[f(x;\theta)]^2} = \frac{f''(x;\theta)}{f(x;\theta)} - [l'(x;\theta)]^2.$

Assim,

$$\begin{aligned}
-E_\theta\left[\left(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right)\right] &= -\int\left(\frac{f''(x;\theta)}{f(x;\theta)} - [l'(x;\theta)]^2\right)f(x;\theta)dx,\\
&= -\int\left(\frac{f''(x;\theta)}{f(x;\theta)} - \left[\frac{\partial}{\partial\theta}\log f(x;\theta)\right]^2\right)f(x;\theta)dx,\\
&= -\underbrace{\int f''(x;\theta)dx}_{=0} + \int\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)dx,\\
&= E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right].
\end{aligned}$$

□

A desigualdade de Cramér-Rao é muito útil na comparação do desempenho de estimadores. Para uma função diferenciável $\tau(\theta)$, temos agora um limite inferior da variânia de qualquer estimador W, tal que $E_\theta[W] = \tau(\theta)$. A cota depende apenas de $\tau(\theta)$ e $f(x|\theta)$ e é uma cota inferior uniforme sobre a variância. Qualquer estimador candidato satisfazendo $E_\theta[W] = \tau(\theta)$ e alcançando esse limite inferior é o melhor estimador não viesado de $\tau(\theta)$.

Uma forma mais simples da desigualdade de Cramér-Rao é se o estimador W for uma identidade, ou seja, se $E_\theta[W] = \tau(\theta) = \theta$. Nesse caso a expressão do Corolário 1 se reduz a

$$
\begin{aligned}
Var_\theta\left[W\left(\mathbf{X}\right)\right] &\geq \frac{\left(\tau'\left(\theta\right)\right)^2}{nE_\theta\left[\left(\frac{d}{d\theta}\log\left(f_X\left(x\,|\theta\right)\right)\right)^2\right]} \\
&\geq \frac{1}{nE_\theta\left[\left(\frac{d}{d\theta}\log\left(f_X\left(x\,|\theta\right)\right)\right)^2\right]}
\end{aligned}
$$

que fica apenas em termos da fdp de X.

Para o entendimento do Teorema Rao-Blackwell, vamos relembrar sobre a esperança condicional de Y dado $X = x$.

Teorema 7. *Sejam (X, Y) duas variáveis aleatórias bidimensionais em $(\Omega, \mathcal{F}, P)$, então*

$$E[g(Y)] = E[E[g(Y)|X]] \tag{49}$$

e em particular

$$E[Y] = E[E[Y|X]]. \tag{50}$$

$\square$

Demonstração.

$$
\begin{aligned}
E[E[g(Y)|X]] = E[h(X)] &= \int_{-\infty}^{\infty} h(x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} E[g(Y)|x] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) dy\right] f_X(x) dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) f_X(x) dy dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y) f_{X,Y}(x, y) dy dx \\
&= E[g(Y)].
\end{aligned}
$$

Para o outro resultado, basta substituir $g(Y)$ por Y. $\square$

Definição 15. *A variância de Y dado $X = x$ é definida por*

$$Var[Y|X = x] = E[Y^2|X = x] - (E[Y|X = x])^2. \tag{51}$$

$\square$

Teorema 8. $Var[Y] = E[Var[Y|X]] + Var[E[Y|X]]$. $\square$

Demonstração.

$$
\begin{aligned}
E[Var[Y|X]] &= E[E[Y^2|X]] - E[(E[Y|X])^2] \\
&= E[Y^2] - E[(E[Y|X])^2] \\
&= Var[Y] + (E[Y])^2 - E[(E[Y|X])^2] \\
&= Var[Y] + (E[E[Y|X]])^2 - E[(E[Y|X])^2] \\
&= Var[Y] - Var[E[Y|X]].
\end{aligned}
$$

$\square$

Com base em uma estatística suficiente podemos encontrar um UNVVMU pelo seguinte Teorema,

Teorema 9 (Rao-Blackwell). *Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória com função densidade de probabilidade (fdp) ou função de probabilidade (fp) $f_X(x;\theta)$. Considere $T = t(X_1, X_2, \ldots, X_n)$ um estimador não viesado de $\tau(\theta)$ e $S = s(X_1, X_2, \ldots, X_n)$ uma estatística suficiente. Se definirmos*

$$\phi(s) = E_\theta[T|S], \tag{52}$$

então

a) $E_\theta[\phi(s)] = \tau(\theta)$;

b) $Var_\theta[\phi(s)] \leq Var_\theta[T]$.

Isto é, $\phi(s)$ é um ENVVMU. □

Demonstração. a) Por (49), temos

$$\tau(\theta) = E[T] = E[E[T|S]] = E[\phi(t)].$$

b) Pelo Teorema 8, temos

$$Var[T] = E[Var[T|S]] + Var[E[T|S]] = E[Var[T|S]] + Var[\phi(t)] > Var[\phi(t)].$$

□

A questão surge agora é se temos $E_\theta[\phi] = \tau(\theta)$ e ϕ é baseado em uma estatística suficiente, como saber se ϕ é o melhor estimador não viesado de $\tau(\theta)$? Naturalmente, se ϕ atinge o limite inferior de Cramer-Rao, então é o melhor estimador, mas se não atinge, o que podemos concluir? Por exemplo, se $\phi^* = E[T^*|S]$ é um outro estimador não viesado de $\tau(\theta)$, como ϕ^* se compara a ϕ? O Teorema a seguir mostra que um melhor estimador não viesado é único.

Teorema 10. *Se W é o melhor estimador não viesado de $\tau(\theta)$, então W é único.* □

Demonstração. Feito em Casella, port. p. 306 e complemento na p. 156; no material de Devanil - Inf I 2015-2016. □

Entretanto, como saber quando um estimador não viesado é o melhor dentre os estimadores não viesados de $\tau(\theta)$? Suponha que W satisfaça $E_\theta[W] = \tau(\theta)$ e temos um outro estimador U tal que $E_\theta[U] = 0$ para todo θ, isto é, U é um estimador não viesado de 0. Então,

$$\phi_a = W + aU, \tag{53}$$

em que a é uma constante, satisfaz $E_\theta[\phi_a] = \tau(\theta)$. É possível que ϕ_a seja melhor que W? A variância ϕ_a é

$$Var_\theta[\phi_a] = Var_\theta[W + aU] = Var_\theta[W] + 2aCov_\theta[W, U] + a^2 Var_\theta[U]. \tag{54}$$

Agora, para algum $\theta = \theta_0$ assuma que $Cov_{\theta_0}[W, U] < 0$, então podemos tornar $2aCov_\theta[W, U] + a^2 Var_\theta[U] < 0$ escolhendo $a \in (0, -2aCov_\theta[W, U]/Var_{\theta_0}[U])$. Deste modo, ϕ_a será melhor que W em $\theta = \theta_0$, e W não poderá ser o melhor estimador não viesado. Da mesma forma acontecerá para $Cov_{\theta_0}[W, U] > 0$ (Ver Casella, p. 307, 156; Magalhaes p. 257). A uníca situação em que W é o melhor estimador é a condição $Cov_{\theta_0}[W, U] = 0$. Assim, a relação de W com estimadores não viesados de 0 é crucial para determinar se W será o melhor estimador de $\tau(\theta)$.

Teorema 11. *Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória com função densidade de probabilidade (fdp) ou função de probabilidade (fp) $f_X(x;\theta)$ e uma estatística $W = w(X_1, X_2, \ldots, X_n)$. W é o melhor estimador não viesado de $\tau(\theta)$ se e somente se W não estiver correlacionado com todos os estimadores de 0 não viesados.* □

Demonstração. Feita em Casella, port. p. 307. □

Para contornar esse problema, vamos definir uma estatística completa.

Definição 16 (Estatística completa)**.** *Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória com função densidade de probabilidade (fdp) ou função de probabilidade (fp) $f_X(x;\theta)$ e uma estatística $T = t(X_1, X_2, \ldots, X_n)$. Uma família de $f_T(t;\theta)$ de T é completa se e somente $E_\theta[g(T)] = 0$ o que implica $P_\theta(g(T) = 0) = 1$ para todo θ, em que $g(T)$ é uma estatística. Assim, a estatística T é completa se e somente se a sua família de densidades for completa.* □

A estatística completa elimina o problema em (53), pois não poderá haver nenhum estimador U de $\tau(\theta)$ relacionado com W tal que $E_\theta[U] = 0$, mas apenas as situações tal que $U^* = 0$ em que $E_\theta[U^*] = 0$ com $P_\theta[U^* = 0] = 1$. Isto é, só poderá haver estimadores não viesados de 0 se esse estimador for 0. Dessa forma se tivermos uma estatística completa W que é um estimador não viesado de $\tau(\theta)$, W nunca estará relacionado com estimadores iguais a U, e daí não teremos problema para afirmar se W é um melhor estimador não viesado de $\tau(\theta)$.

Com a introdução de uma estatística completa, poderemos apresentar o Teorema de Lehmann-Scheffé.

Teorema 12 (Lehmann-Scheffé)**.** *Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória com função densidade de probabilidade (fdp) ou função de probabilidade (fp) $f_X(x;\theta)$. Se $S = s(X_1, X_2, \ldots, X_n)$ é uma estatística suficiente e completa e $T = t(X_1, X_2, \ldots, X_n)$ um estimador não viesado de $\tau(\theta)$, então*

$$\varphi(s) = E[T|S], \tag{55}$$

é o único ENVVMU de $\tau(\theta)$.

Demonstração. (Prova 1). Assuma que S seja uma estatística suficiente e completa, e $\varphi(s)$ um estimador não viesado de $\tau(\theta)$, então pelo Teorema 9 sabemos que $\varphi(s)$ é o melhor estimador não viesado de $\tau(\theta)$, e a partir do Teorema 10 sabemos que $\varphi(s)$ é único.

(Prova 2) Poderemos provar esse Teorema sem mencionar o Teorema 10, mas simplesmente com a Definição 16. Seja $\varphi^*(s) = g(S)$ tal que $E_\theta[\varphi^*(s)] = \tau(\theta)$. Então $E_\theta[\varphi(s) - \varphi^*(s)] = 0$ para todo θ. Pela completicidade temos que $P_\theta(\varphi(s) - \varphi^*(s) = 0) = 1 \Rightarrow P_\theta(\varphi(s) = \varphi^*(s)) = 1$ para todo θ, haverá somente um único estimador de θ que é função de S. Pelo Teorema 9, $Var_\theta[\varphi(s)] \leq Var_\theta[\varphi^*(s)]$, $\varphi(s)$ é ENVVMU. □

A afirmação que $\varphi(s)$ é o único ENVVMU de $\tau(\theta)$ pode ser redundante, pois o Teorema 10 mostra que se $\varphi(s)$ é ENVVMU, é único. Entretanto, tentamos enfatizar o fato de que o estimador $\varphi(s)$ não terá problemas do tipo encontrado em (53).

Em muitas situações não haverá candidato óbvio para um estimador não viesado de $\tau(\theta)$, muito menos um candidato para melhor estimador não viesado. Entretanto, com a presença da completude, o Teorema 12 nos diz que pudermos encontrar um estimador não viesado de θ, poderemos encontrar o melhor estimador não viesado.

4 Propriedade dos estimadores de máxima verossimilhança

Poderemos observar a seguir algumas propriedades interessantes dos estimadores de máxima verossimilhança.

Teorema 13 (Princípio da invariância). *Material Antigo...*

4.1 Propriedades assintóticas

Veremos na sequência que as vezes é possível encontrar uma sequência de estimadores $W_n(X_1, X_2, \ldots, X_n)$ que assintoticamente tem distribuição normal com média θ e variância $\sigma_n^2(\theta)$, em que $\sigma_n^2(\theta)$ indica que a variância é uma função de θ e do tamanho da amostra n. Em particular, temos os estimadores de máxima verossimilhança (EMV), denotado por $\hat{\theta}_n(X_1, X_2, \ldots, X_n)$, que apresentam essa propriedade.

4.1.1 Revisão de alguns Teoremas úteis

Teorema 14 (Lei Fraca dos Grandes Números). *Seja $(X_n)_{n\in\mathbb{N}}$ uma sequência de variáveis aleatórias, independente e identicamente distribuídas (iid), tal que $E[X] = \mu$ e $Var[X] = \sigma^2 < \infty$, definidas no espaço de probabilidade $(\Omega, \mathcal{F}, P)$. Então, para cada $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \leq \epsilon) = 1, \tag{56}$$

isto é, $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ converge em probabilidade para μ, denotada por $\bar{X}_n \xrightarrow{p} \mu$. □

Teorema 15 (Teorema do Limite Central (TLC)). *Seja $(X_n)_{n\in\mathbb{N}}$ uma sequência de variáveis aleatórias, independente e identicamente distribuídas (iid), definidas no espaço de probabilidade $(\Omega, \mathcal{F}, P)$, tal que $E[X] = \mu$ e $0 < Var[X] = \sigma^2 < \infty$. Então*

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\text{d}} Z \sim N(0,1), \tag{57}$$

sendo $\bar{X}_n$ a média amostral. Assim, dizemos que Z_n converge em distribuição para Z. □

Teorema 16 (Teorema de Slutsky). *Se $X_n \xrightarrow{\text{d}} X$ em distribuição e $Y_n \xrightarrow{p} a$, uma constante, em probabilidade, então*

a) $Y_n X_n \xrightarrow{\text{d}} aX$ em distribuição;

b) $X_n + Y_n \xrightarrow{\text{d}} X + a$ em distribuição.

4.1.2 Eficiência, consistência e normalidade assintótica

Teorema 17 (Eficiência e consistência assintótica dos EMV). *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ iid com fp ou fdp $f_X(x;\theta)$. Supondo que $\hat{\theta}$ denote o EMV de θ e que $\tau(\theta)$ seja uma função contínua de θ, sob condições de regularidade de $f_X(x;\theta)$, então*

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \xrightarrow{\text{d}} N(0, \sigma_n^2(\theta)), \tag{58}$$

em que $\sigma_n^2(\theta)$ é o limite inferior da cota de Cramér-Rao, dado por:

$$\sigma_n^2(\theta) = \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\right)^2\right]}.$$

Dizemos que $\tau(\hat{\theta})$ é um estimador consistente e assintoticamente eficiente de $\tau(\theta)$. □

Demonstração. Vamos fazer a prova considerando o EMV $\hat{\theta}$ e X uma v.a. contínua. Considerando que $\ell(\theta; X_1, X_2, \ldots, X_n) = \sum_{i=1}^n \log f_X(X_i;\theta)$ é a função log de verossimilhança, denote $\ell'(\theta, \boldsymbol{X})$ a primeira derivada da função log verossimilhança com relação a θ. Expanda essa derivada em torno do verdadeiro valor do parâmetro θ, denotado por θ_0, isto é,

$$\ell'(\theta, \boldsymbol{X}) = \ell'(\theta_0, \boldsymbol{X}) + (\theta - \theta_0)\ell''(\theta_0, \boldsymbol{X}). \tag{59}$$

Agora, substitua o EMV $\hat{\theta}$ para θ. Como $\ell'(\hat{\theta}, \boldsymbol{X}) = 0$, então

$$(\hat{\theta} - \theta_0) = \frac{-\ell'(\theta_0, \boldsymbol{X})}{\ell''(\theta_0, \boldsymbol{X})}. \tag{60}$$

Pré-multiplicando $\sqrt{n}$ em (60), em ambos os lados, temos

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_0) &= \sqrt{n}\frac{-\ell'(\theta_0, \boldsymbol{X})}{\ell''(\theta_0, \boldsymbol{X})} \\
&= \frac{\sqrt{n}\sqrt{n}}{\sqrt{n}}\frac{-\ell'(\theta_0, \boldsymbol{X})}{\ell''(\theta_0, \boldsymbol{X})} \\
&= \frac{(\sqrt{n})^2}{\sqrt{n}}\frac{-\ell'(\theta_0, \boldsymbol{X})}{\ell''(\theta_0, \boldsymbol{X})} \\
&= \frac{n}{\sqrt{n}}\frac{-\ell'(\theta_0, \boldsymbol{X})}{\ell''(\theta_0, \boldsymbol{X})} \\
&= \frac{-\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})}{\frac{1}{n}\ell''(\theta_0, \boldsymbol{X})} \qquad (61) \\
&= \frac{\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})}{-\frac{1}{n}\ell''(\theta_0, \boldsymbol{X})} \qquad (62)
\end{aligned}$$

Usando primeiro a expressão do numerador de (61), temos que

$$
\begin{aligned}
E\left[\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})\right] &= \frac{1}{\sqrt{n}}E\left[\ell'(\theta_0, \boldsymbol{X})\right] \\
&= \frac{1}{\sqrt{n}}E\left[\frac{\partial}{\partial\theta_0}\sum_{i=1}^{n}\log f_X(X_i;\theta_0)\right] \\
&= \frac{1}{\sqrt{n}}E\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta_0}\log f_X(X_i;\theta_0)\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[\frac{\partial}{\partial\theta_0}\log f_X(X_i;\theta_0)\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[\frac{\partial}{\partial\theta_0}\log f_X(X_i;\theta_0)\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\int_{\mathcal{X}}\frac{\partial}{\partial\theta_0}\log(f_X(t_i;\theta_0))f_X(t_i;\theta_0)dx_i\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\int_{\mathcal{X}}\frac{\frac{\partial}{\partial\theta_0}f_X(t_i;\theta_0)}{f_X(t_i;\theta_0)}f_X(t_i;\theta_0)dx_i\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\int_{\mathcal{X}}\frac{\partial}{\partial\theta_0}f_X(t_i;\theta_0)dx_i\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta_0}\underbrace{\int_{\mathcal{X}}f_X(t_i;\theta_0)dx_i}_{=1}\right) = 0.
\end{aligned} \tag{63}
$$

A variância pode ser expressa da seguinte forma:

$$
\begin{aligned}
Var\left[\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})\right] &= E\left[\left(\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})\right)^2\right] - \left(E\left[\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})\right]\right)^2 \\
&= E\left[\left(\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})\right)^2\right] \\
&= \frac{1}{n}E\left[\left(\ell'(\theta_0, \boldsymbol{X})\right)^2\right] \\
&= \frac{1}{n}E\left[\left(\frac{\partial}{\partial\theta_0}\log L(\theta_0; \boldsymbol{X})\right)^2\right].
\end{aligned} \tag{64}
$$

Existe um resultado para amostras *iid* que $E\left[\left(\frac{\partial}{\partial\theta_0}\log L(\theta_0;\boldsymbol{X})\right)^2\right] = nE\left[\left(\frac{\partial}{\partial\theta_0}\log f_X(X;\theta_0)\right)^2\right]$.

Ver Casella (2001, port. p.300-301) e no material escrito de inf II. Assim,

$$
\begin{aligned}
Var\left[\frac{1}{\sqrt{n}}\ell'(\theta_0,\boldsymbol{X})\right] &= \frac{n}{n}E\left[\left(\frac{\partial}{\partial\theta_0}\log f_X(X;\theta_0)\right)^2\right] \\
&= E\left[\left(\frac{\partial}{\partial\theta_0}\log f_X(X;\theta_0)\right)^2\right] \qquad (65)\\
&= \frac{1}{\sigma_n^2(\theta_0)}. \qquad (66)
\end{aligned}
$$

Pelo Teorema Central do limite, temos que

$$
\frac{\frac{1}{\sqrt{n}}\ell'(\theta_0,\boldsymbol{X})-0}{\sqrt{1/\sigma_n^2(\theta_0)}} \xrightarrow{\text{d}} N(0,1), \qquad (67)
$$

ou

$$
\frac{1}{\sqrt{n}}\ell'(\theta_0,\boldsymbol{X}) \xrightarrow{\text{d}} N(0,1/\sigma_n^2(\theta_0)). \qquad (68)
$$

Se considerarmos o denominador de (61), temos

$$
\begin{aligned}
-\frac{1}{n}\ell''(\theta_0,\boldsymbol{X}) &= -\frac{1}{n}\left(\frac{\partial^2}{\partial\theta_0^2}\sum_{i=1}^{n}\log f_X(X_i;\theta_0)\right)\\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2}{\partial\theta_0^2}\log f_X(X_i;\theta_0)\right)
\end{aligned}
\qquad (69)
$$

Observe que $\frac{\partial^2}{\partial\theta_0^2}\log f_X(X_i;\theta_0)$ pode ser encarada como uma variável aleatória. Se denotarmos $\frac{\partial^2}{\partial\theta_0^2}\log f_X(X_i;\theta_0)=Y_i$, então

$$
-\frac{1}{n}\ell''(\theta_0,\boldsymbol{X}) = -\frac{1}{n}\sum_{i=1}^{n}Y_i = -\bar{Y}. \qquad (70)
$$

Pela Lei Fraca dos Grandes números,

$$
-\bar{Y} \xrightarrow{p} -E\left[\frac{\partial^2}{\partial\theta_0^2}\log f_X(X_i;\theta_0)\right] = E\left[\left(\frac{\partial}{\partial\theta_0}\log L(\theta_0;X)\right)^2\right] = \frac{1}{\sigma_n^2(\theta_0)}. \qquad (71)
$$

Portanto, pelo Teorema de Slutsky, item (a), como

$$
-\frac{1}{n}\ell''(\theta_0,\boldsymbol{X}) \xrightarrow{p} \frac{1}{\sigma_n^2(\theta_0)}
$$

e

$$
\frac{1}{\sqrt{n}}\ell'(\theta_0,\boldsymbol{X}) \xrightarrow{\text{d}} N(0,1/\sigma_n^2(\theta_0)),
$$

então considerando que $W \sim N(0, 1/\sigma_n^2(\theta_0))$, logo

$$\frac{\frac{1}{\sqrt{n}}\ell'(\theta_0, \boldsymbol{X})}{-\frac{1}{n}\ell''(\theta_0, \boldsymbol{X})} \xrightarrow{\mathrm{d}} \sigma_n^2(\theta_0)W. \tag{72}$$

Dessa forma, $\sigma_n^2(\theta_0)W$ também tem distribuição normal com parâmetros

$$E[\sigma_n^2(\theta_0)W] = \sigma_n^2(\theta_0)E[W] = 0,$$

e

$$Var[\sigma_n^2(\theta_0)W] = \sigma_n^4(\theta_0)Var[W] = \frac{\sigma_n^4(\theta_0)}{\sigma_n^2(\theta_0)} = \sigma_n^2(\theta_0),$$

isto é, $\sigma_n^2(\theta_0)W \sim N(0, \sigma_n^2(\theta_0))$. Logo,

$$\sqrt{n}(\theta - \theta_0) \xrightarrow{\mathrm{d}} N(0, \sigma_n^2(\theta_0)),$$

provando o Teorema. □

Exemplo 11 (Normalidade e consistência assintótica). *O Teorema 17 mostra que estimadores EMV $\tau(\hat{\theta})$ de $\tau(\theta)$ são assintoticamente normal, e por consequência eficientes. Ainda mais, a normalidade assintótica implica em consistência. Suponha que*

$$\sqrt{n}\frac{W_n - \mu}{\sigma} \xrightarrow{\mathrm{d}} Z \textit{ em distribuição,}$$

em que $Z \sim N(0, 1)$. Aplicando o Teorema de Slutsky, temos

$$W_n - \mu = \underbrace{\left(\frac{\sigma}{\sqrt{n}}\right)}_{\xrightarrow{p}\left(\frac{\sigma}{\sqrt{n}}\right)} \underbrace{\left(\sqrt{n}\frac{W_n - \mu}{\sigma}\right)}_{\xrightarrow{\mathrm{d}} Z} \xrightarrow{\mathrm{d}} \lim_{n\to\infty}\left(\frac{\sigma}{\sqrt{n}}\right) Z = 0,$$

deste modo, $W_n - \mu \to 0$ converge em distribuição. e o Teorema (Casella, port. pag. 211) mostra que a convergência em distribuição para uma constante1 implica em convergência em probabilidade. Logo, $W_n \xrightarrow{p} \mu$, isto é, W_n é um estimador consistente. □

Como $\sigma_n^2(\theta)$ depende de θ, uma aproximação (Método Delta) para a variância pode ser expresso por

$$\sigma_n^2(\hat{\theta}|\theta) = \sigma_n^2(\hat{\theta}) \approx \frac{\left(\frac{d}{d\theta}\tau(\theta)\right)^2 |_{\theta=\hat{\theta}}}{E_\theta\left[\left(\frac{\partial}{\partial\theta}\log L(\theta; \boldsymbol{X})\right)^2\right] |_{\theta=\hat{\theta}}}, \tag{73}$$

em que $L(\theta; \boldsymbol{X}) = L(\theta; X_1, X_2, \ldots, X_n) = \prod_{i=1}^n f_X(X_i; \theta)$ é a função de verossimilhança.

A quantidade,

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log L(\theta; \boldsymbol{X})\right)^2\right] \tag{74}$$

é conhecida como número de informação ou informação de Fisher. Uma outra forma de apresentar (74) é

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log L(\theta;\boldsymbol{X})\right)^2\right]=-E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L(\theta;\boldsymbol{X})\right]. \tag{75}$$

Considerando uma amostra iid, a expressão (74) pode ser dada por

$$E_\theta\left[\left(\frac{\partial}{\partial\theta}\log L(\theta;\boldsymbol{X})\right)^2\right]=nE_\theta\left[\left(\frac{\partial}{\partial\theta}\log f_X(X;\theta)\right)^2\right]. \tag{76}$$

A prova desses resultados está no material escrito de Inf II. Na prática,

$$\begin{aligned}
\sqrt{n}[\tau(\hat\theta)-\tau(\theta)] &\to N(0,\sigma_n^2(\theta))\\
\frac{\sqrt{n}[\tau(\hat\theta)-\tau(\theta)]}{\sqrt{\sigma_n^2(\theta)}} &\to N(0,1)\\
\tau(\hat\theta)-\tau(\theta) &\to N(0,\sigma_n^2(\theta)/n)\\
\tau(\hat\theta) &\to N(\tau(\theta),\sigma_n^2(\theta)/n)
\end{aligned}$$

4.2 Aplicações

Com essas informações, poderemos agora construir intervalos de confiança para grandes amostras. Usando a aproximação em (73), temos

$$\frac{\sqrt{n}[\tau(\hat\theta)-\tau(\theta)]}{\sqrt{\sigma_n^2(\hat\theta)}}\xrightarrow{\text{d}}N(0,1), \tag{77}$$

pois, pelo Teorema 17 sabemos que $\sqrt{n}[\tau(\hat\theta)-\tau(\theta)]\xrightarrow{\text{d}}N(0,\sigma_n^2(\theta))$. Pelo mesmo Teorema, sabemos que os estimadores de EMV são consistentes assintoticamente, e ainda sabendo pelo princípio da invariância (Mood, 1974, p. 284; Casela, 2001, port p. 285) que se $\hat\theta$ é um EMV de θ, então $\sigma_n^2(\hat\theta)$ também é um EMV de $\sigma_n^2(\theta)$. Logo, $\sigma_n^2(\hat\theta)\xrightarrow{p}\sigma_n^2(\theta)$. Assim, pelo Teorema de Slutsky fica provado a convergência em distribuição de (77).

Assim, um intervalo de confiança aproximado é

$$\tau(\hat\theta)-z_{\frac{\alpha}{2}}\sqrt{\sigma_n^2(\hat\theta)}\le\tau(\theta)\le\tau(\hat\theta)+z_{\frac{\alpha}{2}}\sqrt{\sigma_n^2(\hat\theta)}, \tag{78}$$

sendo $z_{\frac{\alpha}{2}}$ o quantil superior $100(\alpha/2)\%$ com distribuição normal padrão.

Exemplo 12 (Intervalos de confiança para grandes amostras)**.** *Seja uma amostra aleatória $X_1,X_2,\ldots,X_n$ de uma população com distribuição de Bernoulli(p). Construa um intervalo de confiança aproximado para p. Sabemos que o estimador EMV de p é $\hat p_n=\bar X$ (Casella, port. p.*

283). Para calcular $\sigma_n^2(p)$, usaremos a aproximação de (73), isto é,

$$
\begin{aligned}
\sigma_n^2(\hat{p}_n) &\approx \frac{\left(\frac{d}{dp}\tau(p)\right)^2|_{p=\hat{p}_n}}{E_p\left[\left(\frac{\partial}{\partial p}\log L(p;\boldsymbol{X})\right)^2\right]|_{p=\hat{p}_n}} \\
&\approx \frac{1}{E_p\left[\left(\frac{\partial}{\partial p}\log L(p;\boldsymbol{X})\right)^2\right]|_{p=\hat{p}_n}} \\
&\approx \frac{1}{\frac{\partial^2}{\partial p^2}\log L(p;\boldsymbol{X})|_{p=\hat{p}_n}} \\
&\approx \frac{\hat{p}_n(1-\hat{p}_n)}{n}, \text{ para detalhes ver Inf II (Lucas, p. 46)}
\end{aligned}
$$

Assim, um intervalo de confiança com base em (77) é

$$
\hat{p}_n - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \le p \le \hat{p}_n + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}. \tag{79}
$$

sendo $z_{\frac{\alpha}{2}}$ o quantil superior $100(\alpha/2)\%$ com distribuição normal padrão.

Exemplo 13 (Testes binomiais para grandes amostras)**.** *Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$ de uma população com distribuição de Bernoulli(p). Obtenha um teste de hipótese para $\mathcal{H}_0 : p \le p_0$ versus $\mathcal{H}_1 : p > p_0$, para $0 < p_0 < 1$. Se tivermos quaisquer estatísticas W e V e um parâmetros θ de modo que à medida que $n \to \infty$,*

$$
\frac{W-\theta}{V} \xrightarrow{\text{d}} N(0,1), \text{ Ver detalhes, Casella, port. p. 440} \tag{80}
$$

conhecido como teste de Wald. Assim, considerando $W = \hat{p}_n$ e $V = \sigma^2(\hat{p}_n)$ e $\theta = p_0$ sob $\mathcal{H}_0$, o teste de Wald para grandes amostras rejeita $\mathcal{H}_0$ se $Z_n > z_\alpha$, sendo $Z_n = \frac{\hat{p}_n - p_0}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$, e z_α é o quantil superior $100\alpha\%$ de uma distribuição normal.