ESTIMAÇÃO PARAMÉTRICA PONTUAL

March 13, 2003

1 Introdução

Assuma que alguma característica dos elementos de uma população amostrada possa ser representada por uma v.a. X cuja densidade (ou função de probabilidade) é $f(\cdot,\theta)$, onde a forma da densidade é assumida ser conhecida, exceto por um parâmetro desconhecido θ. Nesta situação decidimos tomar uma amostra de tamanho n de X ($X_1,\ldots,X_n$ i.i.d. com densidade $f(\cdot,\theta)$) e, com base nos valores observados $x_1,\ldots,x_n$, deseja-se um bom "chute" do valor θ ou de uma função $\tau(\theta)$.

Exemplo 1.1: É razoável supor que o número de clientes que vão ao Banespa no horário das 12 às 14h é uma v.a. Poisson com média (desconhecida) λ. A fim de dimensionar o número de pessoas (caixas), que devem trabalhar nesse horário, observamos o movimento do banco durante 10 dias e com base nessas observações desejamos estimar λ.

Exemplo 1.2: Na produção de esponjas Scotch-Brite, a fim de fazer o controle de qualidade, a cada 3 horas, 100 esponjas são selecionadas e o número de defeituosas são verificadas para se controlar o valor do parâmetro $p =$ proporção de defeituosas. Sabe-se que

$$X_i \sim b(100,p).$$

A estimação do parâmetro θ pode ser feita de dois modos:

(i) **Estimação Pontual:** Tomamos o valor de alguma estatística $t(X_1,\ldots,X_n)$ para representar, ou estimar, $\tau(\theta)$. Tal estimativa é chamada estimador pontual;

(ii) **Estimação por intervalo:** Definimos duas estatísticas $t_1(X_1,\ldots,_n)$ e $t_2(X_1,\ldots,X_n)$ onde

$$t_1(X_1,\ldots,_n) < t_2(X_1,\ldots,X_n)$$

de modo que $[t_1(X_1,\ldots,_n), t_2(X_1,\ldots,X_n)]$ constitui um intervalo aleatório para o qual é possível se calcular a probabilidade que este intervalo contenha $\tau(\theta)$. Este intervalo é chamado de intervalo de confiança.

Exemplo 1.3: Se queremos pesar um objeto qualquer e ter uma idéia da "confiança"nas medidas fazemos n medições com este objeto; chame os resultados de $X_1,\ldots,X_n$. É razoável supor que $X_i \sim N(\mu,\sigma^2)$ e $\theta = (\mu,\sigma^2)$ é o parâmetro desconhecido, $\tau(\theta) = \mu$ e

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

é um estimador pontual para μ e

$$\left[\bar{X} - 2\sqrt{\frac{S^2}{n}}; \bar{X} + 2\sqrt{\frac{S^2}{n}}\right]$$

é um intervalo de confiança para μ.

Problemas: (i) Como encontrar um "bom" estimador?
(ii) Como selecionar o "melhor" estimador?

1.1 Métodos para se encontrar estimadores

Assuma que $X_1, \ldots, X_n$ é uma amostra aleatória de uma distribuição $f(\cdot, \theta)$ (densidade ou função de probabilidade) e $\theta = (\theta_1, \ldots, \theta_k)$ é um vetor de números reais (podemos ter $k = 1$).

Definição 1.1: Espaço Paramétrico : O conjunto de valores possíveis que θ pode assumir é chamado de espaço paramétrico, e geralmente denotado por Θ.

Objetivo: Queremos, a partir das informações da amostra estimar alguma função de θ, isto é, queremos estimar $\tau(\theta)$. Isto é feito através de funções da amostra, que não dependem de valores desconhecidos, e estas funções foram definidas anteriormente como Estatística. Logo,

Definição 1.2: Estimador: Qualquer estatística cujos valores são usados para estimar $\tau(\theta)$ é dita ser um **estimador** de $\tau(\theta)$.

Exemplo 1.4: Suponha que os pesos dos frangos de um certo galpão possam ser considerados como tendo distribuição aproximadamente normal. Podemos estar interessados em estimar o peso médio dos frangos, a variância ou a proporção de frangos acima de um peso P_0. Caso n frangos sejam escolhidos aleatoriamente, chamando de X_i o peso do i-ésimo frango, podemos considerar $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$. Temos como parâmetro $\theta = (\mu, \sigma^2)$ e o espaço paramétrico $\Theta = \{(\mu, \sigma^2); \mu > 0, \sigma^2 > 0\}$. Neste caso as funções $\tau(\theta)$ seriam iguais a μ, σ^2 e $P[N(\mu, \sigma^2) > P_0]$. Como estimadores podemos utilizar $\bar{X}$ para estimar μ, S^2 para estimar σ^2 e $P[N(\bar{X}, S^2) > P_0]$ para estimar a proporção.

1.2 Método dos momentos

Este método é o mais antigo, proposto por Karl Pearson em 1894. Este é um método simples que produz resultados "razoáveis" na maioria dos casos.

Seja X uma v.a. com distribuição $f(\cdot, \theta_1, \ldots, \theta_k)$. Definimos

$$\mu_r = \mathbf{E}[X^r]$$

o r-ésimo momento de X. Em geral, μ_r é função de $\theta_1, \ldots, \theta_k$. Seja $X_1, \ldots, X_n$ uma amostra aleatória de $f(\cdot, \theta)$ e denote

$$M_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r$$

o r-ésimo momento amostral. Sabemos que

$$\mathbf{E}[M_r] = \mu_r,$$

daí é intuitivo se utilizar de valores de

$$M_r = \mu_r(\hat{\theta_1}, \ldots, \hat{\theta_k})$$

para estimar os parâmetros.

Existem vrias formas de definir estes estimadores; comearemos pelos mais simples, que so utilizados nos textos mais introdutrios.

Definição 1.3.a: O estimador pelo método dos momentos (RMM) de θ é dada pela solução do sistema de equações:

$$m_i = \mu_i(\theta_1, \cdots, \theta_k), \quad i = 1, \cdots, k,$$

isto é, iguala-se os k primeiros momentos amostrais aos k primeiros momentos populacionais.

Esta definição tem a vantagem de levar, quase sempre, a uma solução única. No entanto, nem sempre leva a uma solução. Por exemplo, tome uma amostra aleatória de uma população normal com média conhecida, ou então a distribuição exponencial dupla. Para evitar o problema de não levar a nenhum estimador pode-se generalizar a definição escolhendo-se de forma adequada k momentos amostrais e populacionais para serem igualados. A definição fica

Definição 1.3.b: Um estimador pelo método dos momentos é qualquer solução de um sistema de equações dado por

$$m_i = \mu_i(\theta_1, \cdots, \theta_k), \quad i \in I, \; I = \{i_{i1}, \cdots, i_{ik}\}$$

para uma escolha adequada de forma que se tenha uma solução única.

Em geral procura-se utilizar os momentos de mais baixa ordem. Isto se deve ao fato dos momentos amostrais com menor ordem terem menor variabilidade e serem menos afetados por valores aberrantes. Dada a liberdade de escolha dos momentos utilizados esta definição não produz um estimador único.

Exemplo 1.5: Seja $X_1, \ldots, X_n$ uma a.a. de uma distribuição $N(\mu, \sigma^2)$. Neste caso,

$$\mu_1 = \mu, \quad \sigma^2 = \mu_2 - \mu_1^2.$$

Daí,

$$M_1 = \hat{\mu}, \quad M_2 = \hat{\sigma}^2 + \hat{\mu}^2$$

e

$$\hat{\mu} = M_1 = \bar{X}$$

e

$$\hat{\sigma} = \sqrt{M_2 - \bar{X}^2} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}.$$

Exemplo 1.6: Seja $X_1, \ldots, X_n$ uma a.a. de uma Poisson(λ). Queremos estimar λ pelo método de momentos. Como temos somente um parâmetro, temos somente uma equação

$$M_1 = \bar{X} = \hat{\lambda}.$$

Exemplo 1.7: Seja $X_1, \ldots, X_n$ uma a.a. de uma $\exp(\theta)$. Lembre-se que $\mu_1 = 1/\theta$. Queremos estimar θ pelo método de momentos. Como temos somente um parâmetro, uma nica equação pode ser suficiente. Tomando-se o primeiro momento temos:

$$M_1 = \bar{X} = 1/\hat{\theta} \longrightarrow \hat{\theta} = \frac{1}{\bar{X}}.$$

Verifique a soluo caso fosse escolhido o segundo momento.

Exemplo 1.8: Sejam $X_1, \dots, X_n$ i.i.d. $U[a,b]$, o parâmetro de interesse é $(\theta_1, \theta_2) = (a, b)$. Neste caso,

$$\mu_1 = \frac{a+b}{2}, \quad \mu_2 = \frac{a^2+ab+b^2}{3}.$$

Daí,

$$M_1 = \hat{\mu}_1 = \frac{\hat{a}+\hat{b}}{2}, \quad M_2 = \frac{\hat{a}^2+\hat{a}\hat{b}+\hat{b}^2}{3}.$$

Exemplo 1.9: Sejam $X_1, \dots, X_n$ i.i.d. $U[0,\theta]$, o parâmetro de interesse é θ. Neste caso,

$$\mu_1 = \frac{\theta}{2} \quad \Rightarrow \quad \hat{\theta} = 2\bar{X}.$$

Suponha que $x_1 = 4, x_2 = 6, x_3 = 50$ e assim $\bar{x} = 20$. Assim,

$$\hat{\theta} = 40.$$

Este é um resultado absurdo pois sabemos que $\theta \geq 50 = x_{(n)}$; ou seja, o método dos momentos pode produzir péssimas estimativas. Um estimador melhor seria, por exemplo, $T'(\underline{X}) = \max(X_{(n)}, 2\overline{X})$.

Definição 1.3.c: Suponha que queremos estimar $\gamma = \gamma(\theta)$ e que ela possa ser expressa como uma função contínua dos r primeiros momentos populacionais, isto é, que

$$\gamma(\theta) = g(\mu_1(\theta), \cdots, \mu_r(\theta)).$$

Neste caso dizemos que um estimador de γ pelo método dos momentos é dado por

$$T(\underline{X}) = g(m_1(\underline{X}), \cdots, m_r(\underline{X}))$$

Exemplo 1.10: Considere uma amostra aleatória de tamanho n de uma Poisson com média λ. Dê alguns estimadores pelo método dos momentos.

Sabemos que $m_1 = \lambda$, $m_2 = \lambda + \lambda^2$. Logo, alguns dos estimadores pelo método dos momento são dados por

$$\begin{aligned} \hat{\lambda} &= m_1 \\ \hat{\lambda} &= m_2 - m_1^2 \\ \hat{\lambda} &= m_2/m_1 - 1 \end{aligned}$$

Note que não é especificado que r deve ser o mínimo valor para a qyual existe uma função g. Caso isto fosse especificado no exemplo anterior teríamos um único estimador pelo método dos momentos. No entanto, a unicidade nem sempre ocorreria mesmo tendo esta restrição.

Exemplo 1.11: Considere uma amostra aleatória de tamanho n de uma uma distribuição logística, isto é, da densidade:

$$f(x;\theta) = \frac{e^{-(y-\theta)}}{\{1+e^{-(y-\theta)}\}^2}$$

Como a esperança existe ele é igual ao ponto de simetria, isto é, θ. Logo um estimador pelo método dos momentos é dado pela média amostral. Este é um caso típico onde a estimativa pelo método dos momentos é facilmente encontrado, mas não o de máxima verossimilhança, que não tem solução analítica. Neste caso, a estimativa pelo método dos momentos pode ser utilizada como ponto inicial em uma rotina de maximização para encontrar a estimativa de máxima verossimilhança.

Definição 1.3.d - Método dos Momentos Generalizados: Vamos considerar agora que temos uma amostra $\underline{X}$, com f.d.p. $f^*(\underline{x}|\theta)$. A amostra pode ser aleatória ou não. Considere adicionalmente que temos funções $g_i(\underline{X},\theta)$ com média igual a zero. Desta forma, dada uma amostra gostaríamos que $g_i(\underline{x},\theta)$ fosse o mais próximo possível de zero. Se considerarmos mais funções g do que a dimesnão de θ não conseguiremos fazer esta distância igual a zero. O método dos momentos generalizados diz que qualquer valor de θ que minimiza a distância:

$$g(x,\theta)' S(x,y)^{-1} g(x,\theta)$$

onde $g(x,\theta) = (g_1(x,\theta), \cdots, g_k(x,\theta))$, é um estimador pelo método dos momentos. $S(x,y)$ de ser tal que convirja em probabilidade para um valor não aleatório S_0 quando o tamanho da amostra vai para infinito. Em geral escolhe-se $S(x,y)$ uma matriz simétrica positiva definida.

Observe ques se tomarmos $g_i(X,\theta) = m_i - \mu_i(\theta)$, $i = 1, \cdots, k$ e a matriz S igual a identidade temos a Definição 1.3.a de estimador do método dos momentos. Em nenhuma das definições o método dos momentos é invariante em relação a transformações não lineares. Por exemplo considere uma amostra aleatória de uma $U(0,\theta)$. Neste caso vimos que um estimador pelo método dos momentos é $2\overline{X}$. Procure agora o estimador pelo método dos momentos se você tivesse observado o quadrado das observações e trabalhasse com a distribuição do quadrado da $U(0,\theta)$.

Exercício 1.1: Nos exemplos anteriores procure outros estimadores pelo método dos momentos.

1.3 Método de máxima verossimilhança

O método de máxima verossimilhança para gerar estimadores de um parâmetro desconhecido foi introduzido por Sir R.A. Fisher.

Este método geralmente produz muito "bons" estimadores. Veremos mais tarde as boas propriedades dos estimadores de máxima verossimilhança e alguns exemplos onde o método produz péssimos estimadores.

Considere o seguinte problema: temos duas moedas, uma é honesta e a outra é viciada (tem probabilidade de cara igual a 0.70). O problema é que misturamos as duas moedas e não sabemos diferenciá-las. Para decidir isto, tomamos uma das moedas e jogamos n vezes. Seja:

$$X = \text{número de caras nas } n \text{ repetições;}$$

Daí, $X \sim b(n,p)$, isto é:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = f(k,n)$$

Aqui, $p = 0.5$ ou $p = 0.7$, isto é, $\Theta = \{.5; .7\}$. Se $n = 3$, temos

Valores Possíveis k	0	1	2	3
f(k;0.5)	0.125	0.375	0.375	0.125
f(k;0.7)	0.027	0.189	0.441	0.343

Note que se tiramos 3 caras em 3 lançamentos da moeda não acreditamos muito que $p = 0.5$, é mais "acreditável" (verossímil) que $p = 0.7$. Por outro lado, se tirássemos 0 caras em 3 lançamentos

$p = 0.5$ seria mais verossímil, embora a probabilidade de sair este resultado ainda seja baixa. O que importa, portanto, so os valores relativos.

Neste caso,
Se tiramos 0 ou 1 cara dizemos que $\hat{p} = 0.5$;
Se tiramos 2 ou 3 caras dizemos que $\hat{p} = 0.7$
Isto é, escolhemos $\hat{p}$ que faz com que $f(k, \hat{p})$ seja máximo:

$$\hat{p} = \arg\max_{p \in \Theta} f(k, p)$$

Da forma geral para um tamanho de amostra n e $\Theta = [0, 1]$ temos

$$f(k; p) = P(X = k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}$$

Queremos $\hat{p} = \arg\max_{p \in \Theta} f(k; p)$, para tanto derivamos $f(k; p)$, igualamos a derivada a zero e procuramos os pontos críticos no intervalo paramtrico.

$$\begin{aligned} \frac{d}{dp} f(k; p) &= \begin{pmatrix} n \\ k \end{pmatrix} k p^{k-1} (1-p)^{n-k} - \begin{pmatrix} n \\ k \end{pmatrix} p^k (n-k)(1-p)^{n-k-1} \\ &= \begin{pmatrix} n \\ k \end{pmatrix} p^{k-1} (1-p)^{n-k-1} [k(1-p) - (n-k)p] \\ &= \begin{pmatrix} n \\ k \end{pmatrix} p^{k-1} (1-p)^{n-k-1} [k - np] \end{aligned}$$

Igualando a zero e resolvendo a equação temos como raízes os pontos 0, 1 e k/n. Se $0 < k < n$, analisando a segunda derivada nestes pontos temos que 0 e 1 são pontos de mínimo. Analise a funo de verossimilhana quando k igual a zero ou n. Voc verificar que em todos os casos a soluo de ponto de mximo pode ser escrita como $\hat{p} = k/n$. Portanto, o estimador de máxima verossimilhança é:

$$\hat{p} = \frac{K}{n}$$

Definição 1.4: Função de Verossimilhança:
(1) Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma variável aleatória discreta X com função de probabilidade $f(\cdot, \theta)$ dependendo de um parâmetro desconhecido θ. Se $x_1, \ldots, x_n$ são os valores observados, a **função de verossimilhança** da amostra é

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \theta) \ldots f(x_n, \theta), \qquad \theta \in \Theta$$

(2) Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma variável aleatória contínua X com densidade $f(\cdot, \theta)$ dependendo de um parâmetro desconhecido θ. Se $x_1, \ldots, x_n$ são os valores observados, a **função de verossimilhança** da amostra é funo de θ e numericamente dada por:

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \theta) \ldots f(x_n, \theta), \qquad \theta \in \Theta.$$

Definição 1.5: Estimador de Máxima Verossimilhança: Seja $L(\theta) = L(\theta; x_1, \ldots, x_n)$ a função de verossimilhança para as v.a.'s $X_1, \ldots, X_n$. Se $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ é uma função das observações e é o valor de θ no espaço paramétrico Θ que maximiza $L(\theta)$, então $\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$ é a **estimativa de máxima verossimilhança** de θ e $\hat{\Theta} = \hat{\theta}(X_1, \ldots, X_n)$ é o **estimador de máxima verossimilhança** de θ.
Antes de olhar alguns exemplos, vamos relembrar um teorema de cálculo que é muito útil para

encontrar máximos de funções. Geralmente, como $L(\theta)$ é um produto de funções de probabilidade ou densidades, é sempre positiva. Assim, $l(\theta) = \log(L(\theta))$ sempre pode ser definida e o valor de θ que maximiza $L(\theta)$ também maximiza $l(\theta)$.

Exemplo 1.12: Suponha que retiramos uma amostra aleatória de tamanho n de uma distribuição de Bernoulli

$$f(x,p) = p^x(1-p)^{1-x} I_{\{0,1\}}(x), \quad 0 \leq p \leq 1$$

Os valores amostrais $x_1, \ldots, x_n$ serão uma sequência de 0's e 1's e a função de verossimilhança é:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Podemos definir,

$$l(p) = \sum x_i \log(p) + (n - \sum x_i)\log(1-p)$$

Como l é uma função contínua de p, se existir um valor $(\hat{p})$ tal que

$$\frac{d}{dp} l(\hat{p}) = 0, \; \frac{d^2}{dp^2} l(\hat{p}) < 0$$

então este valor maximiza a função l:

$$\frac{d}{dp} l(p) = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}$$

Assim,

$$\frac{\sum x_i}{\hat{p}} - \frac{n - \sum x_i}{1-\hat{p}} = 0$$

Para $\sum x_i$ diferentes de zero e n temos,

$$\hat{p} = \frac{\sum x_i}{n}$$

Como,

$$\frac{d^2}{dp^2} = -\frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} < 0$$

para todos os valores de p temos que $\hat{p}$ corresponde a um ponto de máximo. Portanto, o estimador de máxima verossimilhança de θ é:

$$\hat{P} = \frac{\sum X_i}{n}$$

Discuta, de forma anloga a realizada no Exemplo da binomial quando temos $\sum x_i$ igual a zero ou n. Note que, s eo espao paramtrico fosse $(0,1)$ o estimador de mxima verossimilhana no existiria para $\sum x_i$ igual a zero ou n.

Exemplo 1.13: Suponha que retiramos uma amostra aleatória de tamanho n de uma distribuição normal com média μ e variância 1. Se $X_1, \ldots, X_n$ é a amostra aleatória, a função de verossimilhança da amostra é:

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} f(x_i, \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} \\ &= (2\pi)^{-n/2} \exp[-\sum (x_i - \mu)^2/2] \end{aligned}$$

cujo logaritmo é:

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \sum \frac{(x_i-\mu)^2}{2}$$

e

$$\frac{d}{d\mu}l(\mu) = \sum (x_i - \mu) = \sum x_i - n\mu$$

$$\frac{d^2}{d\mu^2}l(\mu) = -n < 0$$

Assim,

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}$$

é a estimativa de máxima verossimilhança de θ e o estimador de máxima verossimilhança é:

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X}$$

Se a função de verossimilhança contém k parâmetros, isto é, se:

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

então os estimadores de máxima verossimilhança são as estatísticas $\hat{\theta}_1(X_1,\ldots,X_n),\ldots,\hat{\theta}_k(X_1,\ldots,X_n)$ onde $\hat{\theta}_1,\ldots,\hat{\theta}_k$ são os valores em Θ que maximizam $L(\theta_1,\ldots,\theta_k)$.

Se certas condições de regularidade são satisfeitas, o ponto onde a função de verossimilhança é máxima é a solução das k equações:

$$\frac{\partial}{\partial\theta_1}L(\theta_1,\ldots,\theta_k) = 0,\ldots,\frac{\partial}{\partial\theta_k}L(\theta_1,\ldots,\theta_k) = 0$$

ou equivalentemente,

$$\frac{\partial}{\partial\theta_1}l(\theta_1,\ldots,\theta_k) = 0,\ldots,\frac{\partial}{\partial\theta_k}l(\theta_1,\ldots,\theta_k) = 0$$

Exemplo 1.14: Uma amostra aleatória de tamanho n da distribuição normal de média μ e desvio padrão σ tem densidade:

$$f(x_1,\ldots,x_n,\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(x_i-\mu)^2}$$

e

$$L(\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2\}$$

seu logaritmo sendo:

$$l(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum (x_i-\mu)^2$$

onde $\Theta = \{(\mu,\sigma^2); -\infty < \mu < \infty, \sigma^2 > 0\}$. Portanto,

$$\frac{\partial}{\partial\mu}l(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum (x_i-\mu)$$

$$\frac{\partial}{\partial\sigma^2}l(\mu,\sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\sum (x_i-\mu)^2$$

Daí,

$$\frac{1}{\hat{\sigma}^2}\sum(x_i-\hat{\mu})=0 \Rightarrow \sum(x_i-\hat{\mu})=0 \Rightarrow \hat{\mu}=\frac{\sum x_i}{n}$$
$$-\frac{n}{2}\frac{1}{\hat{\sigma}^2}+\frac{1}{\hat{\sigma}^4}\sum(x_i-\hat{\mu})^2=0 \Rightarrow \hat{\sigma}^2=\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n}$$

e os estimadores de máxima verossimilhança são:

$$\hat{\mu}=\frac{\sum X_i}{n} \qquad \text{e} \qquad \hat{\sigma}^2=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n}$$

Exemplo 1.15: Seja uma variável aleatória tendo densidade uniforme dada por:

$$f(x,\theta)=I_{[\theta-0.5;\theta+0.5]}(x)$$

onde $\Theta=(-\infty,\infty)$. A função de verossimilhança para uma amostra aleatória de tamanho n é dada por:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} I_{[\theta-0.5;\theta+0.5]}(x_i) \\ &= I_{[x_{(n)}-0.5;x_{(1)}+0.5]}(\theta) \end{aligned}$$

onde $x_{(1)}=\min\{x_1,\ldots,x_n\}$ e $x_{(n)}=\max\{x_1,\ldots,x_n\}$ e temos a última igualdade pois

$$\begin{aligned} \prod_{i=1}^{n} I_{[\theta-0.5;\theta+0.5]}(x_i)=1 \quad &\Leftrightarrow \quad x_i\in[\theta-0.5;\theta+0.5], \text{ para todo } i=1,\ldots,n \\ &\Leftrightarrow \quad \theta-0.5\le x_{(1)} \text{ e } \theta+0.5\ge x_{(n)} \\ &\Leftrightarrow \quad \theta\le x_{(1)}+0.5 \text{ e } \theta\ge x_{(n)}-0.5 \end{aligned}$$

Daí,

$$L(\theta)=\begin{cases} 1, & \text{se } x_{(n)}-0.5\le\theta\le x_{(1)}+0.5 \\ 0, & \text{caso contrário} \end{cases}$$

Assim, qualquer estatística com valor $\hat{\theta}$ satisfazendo $X_{(n)}-0.5\le\hat{\theta}\le X_{(1)}+0.5$ é estimador de máxima verossimilhança de θ. Por exemplo, $X_{(n)}-0.5$, $X_{(1)}+0.5$ ou $(X_{(1)}+X_{(n)})/2$, etc...; ou seja, o estimador de máxima verossimilhança neste caso não é único.

Exemplo 1.16: Seja X uma variável aleatória com densidade uniforme no intervalo $[0,\theta]$. Encontre o EMV de θ.

$$f(x,\theta)=\frac{1}{\theta}I_{[0;\theta]}(x)$$

onde $\Theta=(0,\infty)$. A função de verossimilhança para uma amostra aleatória de tamanho n é dada por:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i,\theta) \\ &= \prod_{i=1}^{n}\frac{1}{\theta}I_{[0;\theta]}(x_i) \\ &= \theta^{-n}I_{[0;\theta]}(x_{(n)}) \\ &= \theta^{-n}I_{[x_{(n)};\infty]}(\theta) \end{aligned}$$

onde $x_{(n)} = \max\{x_1, \ldots, x_n\}$. Daí,

$$L(\theta) = \begin{cases} \theta^{-n}, & \text{se } x_{(n)} \leq \theta \\ 0, & \text{caso contrário} \end{cases};$$

ou seja, $L(\theta) = 0$ para $\theta < x_{(n)}$; igual a $x_{(n)}^{-n}$ (valor positivo) no ponto $x_{(n)}$ e depois decresce a partir deste ponto. Assim, o valor de θ que maximiza $L(\theta)$ é $\hat{\theta} = x_{(n)}$ e portanto o EMV de θ é $X_{(n)}$.

Teorema 1.1: Propriedade de Invariância dos Estimadores de Máxima Verossimilhança: Seja $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ o estimador de máxima verossimilhança de θ. Se $\tau(\theta) = (\tau_1(\theta), \ldots, \tau_r(\theta))$, $1 \leq r \leq k$, é uma transformação no espaço paramétrico Θ, então o estimador de máxima verossimilhança de $\tau(\theta)$ é: $\tau(\hat{\theta}) = (\tau_1(\hat{\theta}), \ldots, \tau_r(\hat{\theta}))$.

Exemplo 17: Na densidade normal, seja $\theta = (\mu, \sigma^2)$. Suponha $\tau(\theta) = \mu + z_q\sigma$ onde z_q é tal que $\Phi(z_q) = q$, portanto $\tau(\theta)$ é o q-ésimo quartil. Portanto, o estimador de máxima verossimilhança de $\tau(\theta)$ é:

$$\bar{X} + z_q\sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}$$

Exemplo 18: Considere uma amostra de tamanho 1 de uma população que tem densidade massa discreta igual a $e^{-\lambda/2}$ no ponto zero e uma parte contínua em $(0, \infty)$ com densidade dada por:

$$f_\lambda(t) = \frac{1}{t}e^{-(\lambda+t)/2}\sum_{k=1}^{\infty}\frac{(\lambda t/4)^k}{k!(k-1)!}$$

A verossimilhança é dada por

$$L(t; \lambda) = \begin{cases} f_\lambda(t), & \text{se } t > 0 \\ e^{-\lambda/2}, & \text{se } t = 0 \end{cases};$$

Se o valor da observação for igual a zero então a verossimilhança é igual a $e^{-\lambda/2}$ que é maximizado quando $\lambda = 0$. Caso a observação seja um valor $t > 0$ então o EMV é dado pela solução única da equação:

$$\frac{df_\lambda(t)}{d\lambda} = \sum_{k=1}^{\infty}\frac{(\lambda/2)^{k-1}(t/2)^k}{(k-1)!(k-1)!}\left[1 - \frac{\lambda}{2k}\right] = 0$$

vemos que para para qualquer valor de $t > 0$ a derivada será negativa para $\lambda < 2$ pois todos os termos serão negativos. Logo a solução de máxima verossimilhança é maior do que 2. Este estimador tem umas características estranhas. Embora λ possa adotar qualquer valor positivo o estimador nunca assume valores no intervalo $(0, 2)$. Além disso

$$\lim_{t \to 0^+}\lambda(t) \geq 2, \quad \text{mas} \quad \lambda(0) = 0,$$

ou seja, existe um ponto de descontinuidade no ponto zero.

Exercício 1.2: Suponha que X seja uma variável normal com média 10 e variância σ^2 desconhecida. Qual o EMV do primeiro quartil baseado em uma amostra aleatória de n observações de X?

Exercício 1.3: Suponha $X \sim P(\lambda)$. Dada uma amostra aleatória de tamanho n de X, qual o EMV de $P[X > 0]$?

Exercício 1.4: Se $X \sim \text{Geom}(p)$ qual o EMV de $Var(X)$ baseado em uma amostra de tamanho n?

Exercício 1.5: Se $X \sim \text{Exp}(\lambda)$ qual o EMV de $P[X > t_0]$ baseado em uma amostra aleatória de n observações?

Exercício 1.6: Considere uma população com três tipos de elementos denominados 1, 2 e 3 que ocorrem com a proporção de Hardy-Weinberg; i.e

$$p(1,\theta) = \theta^2, \quad p(2,\theta) = 2\theta(1-\theta), \quad p(3,\theta) = (1-\theta)^2$$

onde $0 < \theta < 1$ e $p(1,\theta)$ é a probabilidade de um elemento ser do tipo i. Dada uma amostra aleatória de tamanho n onde se verifica qual o tipo do elemento selecionado encontre o EMV de θ.

1.4 Outros métodos

Existem outros métodos para se encontrar estimadores. Entre eles podemos citar o Método Bayesiano, o Método dos Mínimos Quadrados, o Método de Mínimo Qui-Quadrado, e o Método da Distância Mínima. Nesta subseção discutiremos rapidamente os dois últimos métodos. Existe também uma classe de estimadores, os Estimadores Não Viciados de Mínima Variância, para o qual existe toda uma metodologia para encontrá-los e que será objeto de estudo da Seção 3.

Definição 1.6: Estimador pelo Método do Mínimo Qui-Quadrado): Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma densidade dada por $f_X(x;\theta)$, e seja $P_1, \ldots, X_k$ uma partição do conjunto de variabilidade de X. A probabilidade de que uma observação caia na cela P_i, $\quad i = 1, \ldots, k$ denotada por $p_i(\theta)$ pode ser encontrada. Por exemplo, se $f_X(x;\theta)$ é uma função densidade de uma variável aleatória contínua, então $p_i(\theta) = P[X \text{ caia na cela } P_i] = \int_{P_i} f_x(x;\theta)dx$. Seja N_j o número de X 's que caem na cela P_j, $j = 1, \ldots, k$; então $n = \sum_{j=1}^{n} N_j$ é o tamanho amostral. A estimativa de mínimo qui-quadrado de θ é o valor $\hat{\theta}$ que minimiza a seguinte soma:

$$\chi^2 = \sum_{j=1}^{k} \frac{[n_j - np_j(\theta)]^2}{np_j(\theta)}$$

O estimador de mínimo qui-quadrado é o valor de θ que faz com que o valor esperado de observações na cela P_j seja o "mais próximo" possível do valor observado. A medida de proximidade é dada pela fórmula acima. Mais tarde veremos o motivo do nome qui-quadrado. O estimador depende da partição utilizada. Muitas vezes fica difícil encontrar o estimador e a causa desta dificuldade está no denominador, daí ter se proposto na literatura uma modificação onde no denominador aparece o valor n_j , isto é, o valor esperado é substituido pelo valor observado.

Definição 1.7: Estimador Pelo Método da Distância Mínima : Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição dada pela função distribuição acumulada $F_X(x;\theta) = F(x;\theta)$, e seja $d(F,G)$ uma medida da distância entre duas funções distribuições acumuladas F e G. Um exemplo de medida de distância é $d(F,G) = sup|F(x) - G(x)|$, que é a maior distância vertical entre F e G. A estimativa de mínima distância de θ é o valor de θ, pertencente ao espaço paramétrico para o qual $d(F(x;\theta), F_n(x))$ é minimizada, onde $F_n(x)$ é a função distribuição acumulada empírica. Embora este estimador tenha um apelo intuitivo bastante forte já que $F_n(x)$ converge para $F_X(x)$, ele não é prático porque é muito difícil de ser encontrado.

2 Propriedades de estimadores pontuais

Já vimos vários métodos de construção de estimadores pontuais para parâmetros desconhecidos. Em muitos casos os dois métodos obtêm o mesmo estimador, mas em muitos outros casos importantes não. Também há outros métodos ainda não estudados para a obtenção de estimadores. As questões que nos vêm a mente agora são: "Qual estimador devo utilizar?", "Como selecionar o melhor estimador?", Quais as propriedades que um bom estimador deve ter?". Se pudéssemos encontrar uma escala de "bondade" de estimadores, sempre poderíamos escolher o melhor estimador para cada caso. Entretanto, não há uma escala universal de "bondade".

O estimador ($\hat{\Gamma}$) de um parâmetro desconhecido (γ) é uma estatística e como tal uma v.a. que tem uma lei de probabilidade; portanto, é sujeita a variabilidade e não é razoável de se esperar que a estimativa $\hat{\gamma}$ seja igual ao valor verdadeiro do parâmetro γ para todas as amostras retiradas. Se consideramos dois estimadores $\hat{\Gamma}$ e $\tilde{\Gamma}$ para o mesmo parâmetro γ, podemos derivar as leis de probabilidade dos estimadores e compará-las de algum modo. Por exemplo, se $\hat{\Gamma} \sim U(\gamma - 0.5; \gamma + 0.5)$ e $\tilde{\Gamma} \sim U(\gamma - 0.01; \gamma + 0.01)$, certamente preferiríamos $\tilde{\Gamma}$ como estimador de γ. Infelizmente as comparações não são tão diretas e fáceis como nesse caso.

Intuitivamente, queremos um estimador que seja "próximo" do verdadeiro valor do parâmetro. Há várias maneiras de se definir "próximo". Seja $\hat{\Gamma} = \hat{\Gamma}(X_1, \ldots, X_n)$ uma v.a. e portanto com uma distribuição de probabilidade. A distribuição de $\hat{\Gamma}$ nos diz como os valores observados (estimativas) $\hat{\gamma}$ estão distribuídos, e gostaríamos de ter valores de $\hat{\Gamma}$ distribuídos próximos de γ. Sabemos que a média e a variância de uma distribuição são medidas de locação e dispersão, daí o sentido de $\hat{\Gamma}$ ser "próximo" de γ poderia ser:

- $\mathbf{E}(\hat{\Gamma})$ "próxima" de γ;

- $\mathrm{Var}(\hat{\Gamma})$ "próxima" de 0.

Uma propriedade desejável para um estimador é que sua média seja o valor verdadeiro do parâmetro.

Definição 2.1: Estimador Não Viciado: Um estimador $\hat{\Gamma}$ de um parâmetro γ é **não viciado** se $\mathbf{E}(\hat{\Gamma}) = \gamma, \quad$ para todo $\gamma \in \Gamma$. Alguns autores utilizam os nomes **Estimador Não Tendencioso** e **Não Viesado**.

Exemplo 2.1: Se $X_1, \ldots, X_n$ forma uma amostra aleatória de uma distribuição tal que $\mathbf{E}(X_i) = \mu$ então sabemos que $\mathbf{E}(\bar{X}) = \mu$. Portanto, $\bar{X}$ é um estimador não viciado de μ se $X_i \sim N(\mu, \sigma^2)$, de p se $X_i \sim b(1,p)$, de λ se $X_i \sim \text{Poisson}(\lambda)$.

A propriedade de ser não viciado, embora desejável para um estimador, não deve ser o único critério utilizado para se comparar estimadores; também devemos ter estimadores mais "concentrados" em torno do verdadeiro valor do parâmetro. Para isto eles dever ter variância pequena.

Definição 2.2: Estimador Mais Eficiente: Se $\hat{\Gamma}$ e $\tilde{\Gamma}$ são dois estimadores não viciados de γ, dizemos que $\hat{\Gamma}$ **é mais eficiente** que $\tilde{\Gamma}$ se

$$\mathrm{Var}(\hat{\Gamma}) < \mathrm{Var}(\tilde{\Gamma}).$$

Exemplo 2.2: Suponha que $X_1, \ldots, X_n$ é uma amostra aleatória de uma distribuição Poisson(λ).

Portanto, $\hat{\Lambda} = \bar{X}$ e $\tilde{\Lambda} = (X_1 + X_2)/2$ são ambos estimadores não viciados de λ, entretanto,

$$\mathrm{Var}(\hat{\Lambda}) = \frac{\lambda}{n}, \quad \mathrm{Var}(\tilde{\Lambda}) = \frac{\lambda}{2}$$

Assim, se $n > 2$, $\hat{\Lambda}$ é mais eficiente que $\tilde{\Lambda}$.

Exemplo 2.3: Considere X uma amostra de tamanho 1 de uma Poisson com média $\lambda > 0$. Mostre que a estatística $T(X) = (-2)^X$ é um estimador não viciado de $\tau(\lambda) = e^{-3\lambda}$.

$$\begin{aligned} E(T) &= \sum_{x=0}^{\infty} (-2)^x \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-2\lambda)^x}{x!} \\ &= e^{-\lambda} e^{-2\lambda} = e^{-3\lambda} \end{aligned}$$

Este estimador é ridículo porque assume valores negativos quando o valor observado é ímpar. Por exemplo, se o valor observado for 10 temos uma estimativa igual a 1024 enquanto se o valor observado for 11 a estimativa é igual a -2048, o que é ridículo. O pior é que, como veremos mais tarde no *Exemplo 3.19*, ele é o único estimador não viciado de $e^{-3\lambda}$.

Exemplo 2.4: Considere $X_1, \ldots, X_n$ funções indicadoras de n ensaios de Bernoulli independentes com probabilidade de sucesso θ. Já vimos que este modelo pode ser utilizado em várias situações. A função $\tau(\theta) = \theta/(1-\theta)$ é chamada de risco relativo e bastante utilizado em bioestatística e epidemiologia. Mostre que não existe um estimador não viciado para o risco relativo.
Os possíveis resultados do experimento são as 2^n distintas combinações de zeros e uns. Qualquer estatística T define um valor real t_j para cada um dos pontos do espaço amostral, onde $j = 1, \ldots, 2^n$ enumera os resultados possíveis. Para esta estatística geral a esperança é dada por:

$$E(T) = \sum_{j=1}^{2^n} t_j p^{n_j} (1-p)^{n-n_j},$$

onde n_j é o número de sucessos obtidos no j-ésimo ponto do espaço amostral. Para que T seja não viciado precisamos ter a seguinte condição:

$$\sum_{j=1}^{2^n} t_j p^{n_j} (1-p) n - n_j = \frac{\theta}{1-\theta} \quad \text{para todo } \theta \in (0,1)$$

Esta igualdade nos diz que um polinômio em θ de ordem 2^n deve ser igual a $\theta/(1-\theta)$ para todo θ em um intervalo. Claramente isto não pode ocorrer e, portanto, não podemos ter um estimador não viciado para o risco relativo.

Os dois exemplos anteriores mostram que nem sempre podemos, ou é desejável, nos restringirmos aos estimadores não viciados.

2.1 Erro Quadrático Médio

Nem sempre um estimador viciado é ruim Às vezes, o que perdemos por ter um vício pequeno pode ser compensado pela concentração em torno do valor verdadeiro. De alguma forma temos que combinar os dois fatores mencionados anteriormente: $\mathbf{E}(\hat{\Gamma})$ "próxima" de γ e $\mathrm{Var}(\hat{\Gamma})$ "próxima" de 0. Isto pode ser obtido através de uma medida muito útil de proximidade chamada *erro quadrático* *médio* (EQM).

Definição 2.3: Erro Quadrático Médio: Seja $\hat{\Gamma} = \hat{\Gamma}(X_1, \ldots, X_n)$ um estimador de γ baseado em uma amostra aleatória $X_1, \ldots, X_n$. O **erro quadrático médio (EQM)** de $\hat{\Gamma}$ é:

$$\mathrm{EQM}(\hat{\Gamma}, \gamma) = \mathbf{E}_\gamma[(\hat{\Gamma} - \gamma)^2].$$

Obs.: Para v.a.'s contínuas com densidade $f(\cdot, \gamma)$,

$$\mathbf{E}_\gamma[(\hat{\Gamma} - \gamma)^2] = \int \ldots \int [(\hat{\gamma}(x_1, \ldots, x_n) - \gamma)^2] f(x_1, \gamma) \ldots f(x_n, \gamma) dx_1 \ldots dx_n.$$

Como $\mathrm{EQM}(\hat{\Gamma}, \gamma) = \mathrm{Var}_\gamma \hat{\Gamma} + [\gamma - \mathrm{E}_\gamma \hat{\Gamma}]^2$ então se $\hat{\Gamma}$ é não viciado, $\mathrm{EQM}(\hat{\Gamma}, \gamma) = \mathrm{Var}_\gamma(\hat{\Gamma})$ e $\mathrm{EQM}(\hat{\Gamma}, \gamma)$ pode ser pensado como uma medida de espalhamento de $\hat{\Gamma}$ em torno de γ.

Se formos comparar estimadores baseados em seus EQM, naturalmente iremos preferir aquele com menor EQM. Geralmente, EQM depende de γ (parâmetro desconhecido) e não temos um estimador com EQM uniformemente menor. A situação a seguir em geral é a mais comum:

Se $\gamma \in [a, b]$ dizemos que Γ_1 é melhor que Γ_2;
Se $\gamma \notin [a, b]$ dizemos que Γ_2 é melhor que Γ_1.

Não temos base para escolher um estimador em detrimento do outro.

Exemplo 2.5: Sejam $X_1, X_2, \ldots, X_n$ i.i.d. $\exp(\beta)$ e tome

$$T_1 = (\sum_{i=1}^n X_i)/n$$

e

$$T_2 = \sum_{i=1}^n a_i X_i$$

onde $\sum_{i=1}^n a_i = 1$. Portanto, T_1 e T_2 são estimadores de $\tau(\beta) = 1/\beta$.
Calcule $\mathrm{EQM}(T_1, \beta)$ e $\mathrm{EQM}(T_2, \beta)$ e verifique se preferimos T_1 ou T_2 com base neste critério.

Como T_1 e T_2 são não viciados temos que

$$\begin{aligned}
\text{EQM}(T_1,\beta) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n}\text{Var}(X_1) = \frac{1}{n\beta^2}
\end{aligned}$$

e

$$\begin{aligned}
\text{EQM}(T_2,\beta) &= \text{Var}(\sum_{i=1}^{n} a_i X_i) \\
&= \sum_{i=1}^{n} a_i^2 \text{Var}(X_i) \\
&= \frac{1}{\beta^2}\sum_{i=1}^{n} a_i^2
\end{aligned}$$

Daí, $\text{EQM}(T_1,\beta) \leq \text{EQM}(T_2,\beta)$ se $(1/n) \leq \sum_{i=1}^{n} a_i^2$. Mas $\min \sum_{i=1}^{n} a_i^2$ sujeito a $\sum_{i=1}^{n} a_i = 1$ ocorre quando $a_i = 1/n$ para todo $i = 1, \ldots, n$. Portanto, T_1 é sempre melhor que T_2.

Exemplo 2.6: Sejam $X_1, X_2, \ldots, X_n$ i.i.d. Poisson(λ). Sejam $T_1 = 1$ e $T_2 = \bar{X}$ dois estimadores de λ. Daí,

$$\text{EQM}(T_1,\lambda) = \mathbf{E}_\lambda(1-\lambda)^2 = (1-\lambda)^2$$

$$\text{EQM}(T_2,\lambda) = \mathbf{E}_\lambda(\bar{X}-\lambda)^2 = \text{Var}(\bar{X}) = \lambda/n$$

Vamos supor que $n = 2$, assim,

e temos que se:

- Se $\lambda \in [1/2; 2]$ temos T_1 preferível a T_2;
- Se $\lambda \notin [1/2; 2]$ temos T_2 preferível a T_1.

- Mas, em $\lambda = 1$, $\text{EQM}(T_1, 1) = 0 < \text{EQM}(T, 1)$ para qualquer estimador $T \neq 1$ de λ. Assim, quando $\lambda = 1$ o estimador $T_1 = 1$ será preferível a qualquer estimador. Assim vemos que não existe um estimador $\hat{\Gamma}$ de γ que possa ser o melhor de todos considerando-se o critério de EQM.

Multiplicadores de Lagrange: Pode-se mostrar que o mínimo da função $g(\mathbf{x})$ sujeito a $h(\mathbf{x}) = K$ é encontrado achando-se o mínimo da função $g(\mathbf{x}) - \lambda h(\mathbf{x})$. Não vamos provar isto aqui, mas é possível se provar que se $\mathbf{y}$ satisfaz $h\mathbf{y} = K$ e minimiza $g(\mathbf{x}) - \lambda h(\mathbf{x})$ para algum λ, então para qualquer outro $\mathbf{x}$ tal que $h(\mathbf{x}) = K$,

$$g(\mathbf{x}) - \lambda h(\mathbf{x}) \geq g(\mathbf{y}) - \lambda h(\mathbf{y}),$$

ou, como $h(\mathbf{x}) = h(\mathbf{y})$,

$$g(\mathbf{x}) \geq g(\mathbf{y}).$$

Assim, $\mathbf{y}$ é o ponto de mínimo. No caso que queremos minimizar $\sum a_i^2$ sujeito à $\sum a_i$, minimizamos a função $f(\mathbf{a}) = \sum a_i^2 - \lambda \sum a_i = 1$. A derivada desta função com respeito a cada a_i deve ser zero:

$$2a_i - \lambda = 0, \quad j = 1, \ldots, n$$

Portanto, os valores de a_i que minimizam $f(\mathbf{a})$ são todos iguais e como eles devem somar 1 devem ser todos iguais a $1/n$. Portanto, a média amostral é o estimador linear não viciado mais eficiente (de mínima variância).

O problema de se encontrar um estimador que tenha uniformemente o menor EQM não tem solução (uniformemente significa para qualquer valor do parâmetro pertencente ao espaço paramétrico). Já vimos que o "pior" estimador possível, tem um EQM de zero para um valor particular do parâmetro. Isto ocorre porque estamos procurando estimadores numa classe muito ampla. Algumas vezes pode-se encontrar estimadores com mínima variância na classe dos estimadores não viciados (veja ENVUMV); mas exceto pelo fato de que nesta classe o problema de minimalidade de EQM tem solução, a restrição a estimadores não viciados algumas vezes excluem estimadores que são bons.

Exemplo 2.7: Já vimos que se temos uma amostra aleatória de uma distribuição $N(\mu, \sigma^2)$, o estimador de máxima verossimilhança de σ^2 é $\hat{\sigma}^2 = (1/n) \sum (X_i - \bar{X})^2$ e

$$\mathbf{E}[\hat{\sigma}^2] = \sigma^2 (1 - \frac{1}{n})$$

portanto, $\hat{\sigma}^2$ tem um pequeno vício. Seu erro quadrático médio é:

$$\begin{aligned} EQM(\hat{\sigma}^2; \mu, \sigma^2) &= \text{Var}(\hat{\sigma}^2) + (\mathbf{E}(\hat{\sigma}^2) - \sigma^2)^2 \\ &= \frac{2\sigma^4 (n-1)}{n^2} + (-\frac{\sigma^2}{n})^2 \\ &= \frac{2n-1}{n^2} \sigma^4 \end{aligned}$$

Um estimador não viciado de σ^2 é $S^2 = (n-1)^{-1} \sum (X_i - \bar{X})^2$ (a variância amostral) e seu EQM é:

$$\begin{aligned} EQM(S^2; \mu, \sigma^2) &= \text{Var}(S^2) \\ &= \frac{1}{(n-1)^2} 2(n-1)\sigma^4 \\ &= \frac{2\sigma^4}{n-1} > \frac{2n-1}{n^2} \sigma^4 \end{aligned}$$

Portanto, neste caso, o estimador não viciado tem um EQM maior que um estimador "um pouco" viciado.

Apesar de sua dependência nos parâmetros desconhecidos o EQM é útil quando estamos estudando a "performance" dos estimadores para grandes amostras. Neste caso, estamos procurando estimadores cujos EQM's sejam próximos a zero quando o tamanho cresce.

2.2 Consistência

Um estimador, em geral, depende do tamanho da amostra. Por exemplo, os momentos amostrais dependem de n e são definidos para todos os tamanhos amostrais, e.g., $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Assim temos uma sequência de estimadores $\hat{\Gamma}_n$ que dependem do tamanho da amostra. É intuitivo desejar que quanto maior a amostra melhor seja o nosso estimador; assim um bom estimador $\hat{\Gamma}_n$ tem EQM que decresce a 0 quanto mais elementos contiver a amostra, daí a definição de consistência em média quadrática.

Definição 2.4: Estimador Consistente em Média Quadrática: Uma sequência de estimadores $\{\hat{\Gamma}_n\}$ é dita ser **consistente em média quadrática** se a seguinte condição ocorre:

$$\lim_{n\to\infty} EQM(\hat{\Gamma}_n, \gamma) = \lim_{n\to\infty} \mathbf{E}(\hat{\Gamma}_n - \gamma)^2 = 0.$$

Note que a condição é verdadeira se, e somente se, o vício do estimador **e** a variância do estimador tende a 0 quando $n \to \infty$. Uma outra condição um pouco mais fraca é dada pela definição seguinte.

Teorema 2.1: Se uma sequência de estimadores é Consistente em média quadrática então ela é consistente (convergência em probabilidade), mas o inverso não é necessariamente verdadeiro.

Prova: Seja $\{\hat{\Gamma}_n\}$ uma sequência de estimadores de γ.

$$\begin{aligned} \mathbf{P}[|\hat{\Gamma}_n - \gamma| < \epsilon] &= \mathbf{P}[|\hat{\Gamma}_n - \gamma|^2 < \epsilon^2] \\ \text{pela desigualdade de Chebyshev} \qquad &\geq 1 - \frac{\mathrm{E}_\gamma[(\hat{\Gamma}_n - \gamma)^2]}{\epsilon^2} \end{aligned}$$

Como $\{\hat{\Gamma}_n\}$ é consistente em média quadrática temos que $\mathrm{E}_\gamma[(\hat{\Gamma}_n - \gamma)^2]$ vai para zero quando n tende ao infinito. Logo

$$\lim_{n\to\infty} \mathbf{P}(|\hat{\Gamma}_n - \gamma| \geq \epsilon) = 0, \quad \text{para todo } \epsilon > 0$$

Para mostrar que o contrário não é necessariamente verdadeiro basta dar um contra-exemplo. Veja o exemplo 5.15 do livro de Romano e Siegel (Counterexamples in Probability and Statistics).

Exemplo 2.8: Os momentos amostrais $M_{n,k} = (1/n)\sum_{i=1}^{n} X_i^k$ são consistentes em média quadrática dos correspondentes momentos populacionais μ_k pois são não viciados e a variância tende a zero.

$$\begin{aligned} \mathbf{E}(M_{n,k}) &= \mathbf{E}(\frac{1}{n}\sum_{i=1}^{n} X_i^k) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}(X_i^k) = \mu_k \end{aligned}$$

portanto, o vício é zero. Mais ainda,

$$\begin{aligned} \mathrm{Var}(M_{n,k}) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i^k) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i^k) \\ &= \frac{1}{n}\mathrm{Var}(X_1^k) \to 0 \end{aligned}$$

quando $n \to 0$. Em particular, $\bar{X}$ é um estimador consistente de μ e $\hat{\sigma}^2$ é estimador consistente de σ^2. A variância amostral também é um estimador consistente de σ^2 (por quê?).

2.3 Normalidade Assintótica

Novamente vamos considerar uma sequência de estimadores $\hat{\Gamma}_n$ do parâmetro desconhecido γ.

Definição 2.5: Melhor Sequência Assintoticamente Normal: Uma sequência de estimadores $\hat{\Gamma}_n$ de γ é definida como sendo a **melhor sequência assintóticamente normal** (best asymptotically normal, BAN) se, e somente se, as 3 condições abaixo são satisfeitas:

- (i) $\sqrt{n}(\hat{\Gamma}_n - \gamma) \approx N(0, \sigma^2(\gamma))$, quando $n \to \infty$;
- (ii) Para todo $\epsilon > 0$
$$\lim_{n\to\infty} \mathbf{P}_\gamma[|\hat{\Gamma}_n - \gamma| > \epsilon] = 0$$
para todo γ. ($\hat{\Gamma}_n$ é fracamente consistente).
- (iii) Seja S_n uma outra sequência de estimadores fracamente consistentes de γ tal que
$$\sqrt{n}(S_n - \gamma) \approx N(0, \tilde{\sigma}^2(\gamma))$$
quando $n \to \infty$, Então $\sigma^2(\gamma) < \tilde{\sigma}^2(\gamma)$, para todo γ.

A utilidade desta definição se deriva parcialmente dos teoremas que garantem a existência de estimadores BAN e do fato que estimadores razoáveis e comuns são assintoticamente normalmente distribuídos.

Exemplo: 2.9 $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ é BAN para μ. De fato,

$$\mathbf{P}[|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0, \text{quando } n \to \infty$$

e

$$\sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2), \text{quando } n \to \infty$$

e nenhum outro estimador com essas propriedades possui variância assintótica menor que σ^2. Mas há muitos outros estimadores S_n que também são BAN, e.g.

$$S_n = \frac{1}{n+1}\sum_{i=1}^{n} X_i$$

também é BAN para μ.

3 ENVUMV - Estimadores Não Viciados Uniformemente de Mínima Variância

Se nos restringimos a estudar os estimadores não viciados, podemos colocar como critério de optimalidade aminimalização do EQM, que neste caso, é igual a variância.

Portanto, gostaríamos de poder ter um método de construir estimadores não viciados uniformemente de mínima variância (ENVUMV). Para isto precisamos introduzir os conceitos de suficiência, minimalidade, completude e famílias exponenciais. Embora estes conceitos sejam utilizados aqui basicamente para encontrar ENVUMV sua importância é mais abrangente em estatística e por isto em muitos livros textos eles merecem capítulos próprios.

3.1 Suficiência

Quando realizamos um experimento e nos deparamos com uma amostra, em geral, temos um conjunto de dados os quais não estão suficientemente organizados para que possamos tirar qualquer informação. Neste ponto, fazemos uma análise exploratória de dados e reduzimos os dados de tal forma que possamos entender o fenômeno de interesse. Por exemplo, calculamos as medidas de tendência central e de dispersão, fazemos ramo e folhas, histogramas, etc. Sabemos há uma perda de informação, mas a pergunta a ser feita é se podemos reduzir os dados em uma coleção de estatísticas sem perda de informação sobre o parâmetro. Isto é, antes de continuarmos nossa busca do melhor estimador, temos que introduzir o conceito de estatísticas **suficientes**. Nos problemas de estimação temos um conjunto de dados observados $x_1, \ldots, x_n$ e queremos condensar esta informação em alguns números sem perder informação sobre o parâmetro de interesse. Isto é, queremos ser capazes de encontrar uma função da amostra que nos diga tudo sobre o parâmetro θ como se olhássemos a amostra como um todo. Tal função seria suficiente para propósitos de inferência e é chamada **estatística suficiente**.

Seja $X_1, \ldots, X_n$ uma amostra aleatória de alguma distribuição, com densidade ou função de probabilidade $f(\cdot, \theta)$.

Uma estatística é uma função da amostra: $T = t(X_1, \ldots, X_n)$ e é também uma v.a. unidimensional, ela condensa as n v.a.'s $X_1, \ldots, X_n$ em uma única v.a. T. Tal condensamento é desejável pois é muito mais fácil trabalhar com números do que com vetores.

Estamos interessados em que não haja perda de informação sobre o parâmetro de interesse em tal condensamento. A única informação sobre o parâmetro θ na densidade (função de probabilidade) $f(\cdot, \theta)$ da qual amostramos está contida na amostra $X_1, \ldots, X_n$, assim quando dizemos que a estatística suficiente não perde informação, queremos dizer que ela contém toda informação sobre θ que está contida na amostra. Enfatizamos que o tipo de informação que estamos falando é o tipo de informação sobre θ **dado que** conhecemos a forma da distribuição (e.g., normal, exponencial, Poisson, etc).

Definição 3.1 Estatística Suficiente: Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $f(\cdot, \theta)$, onde θ pode ser um vetor. Uma estatística $S = s(X_1, \ldots, X_n)$ é dita ser uma *estatística suficiente* para θ se, e somente se, a distribuição condicional de $X_1, \ldots, X_n$ dado que $S = s$ não depende de θ qualquer que seja o valor de s.

Note que a idéia é que se soubermos o valor da estatística suficiente, então os valores amostrais por si mesmos não são mais necessários e não nos dã nenhuma informação adicional a respeito de

θ. Não podemos esperar aprender nada sobre θ amostrando uma distribuição que não depende de θ.

Exemplo 3.1: Seja X_1, X_2, X_3 uma amostra de tamanho 3 de uma distribuição de Bernoulli, isto é:

$$f_{X_1,X_2,X_3}(t_1,t_2,t_3) = p^{t_1+t_2+t_3}(1-p)^{3-t_1-t_2-t_3}$$

para $t_i = 0$ ou 1, $i = 1, 2, 3$.

Considere as estatísticas:

$$S = X_1 + X_2 + X_3$$
$$T = X_1X_2 + X_3$$

Vamos mostrar que S é suficiente mas T não é.

$X_1, X_2, X_3)$	S	T	$f_{X_1,X_2,X_3\|S}$	$f_{X_1,X_2,X_3\|T}$
(0,0,0)	0	0	1	$\frac{1-p}{1+p}$
(0,0,1)	1	1	1/3	$\frac{1-p}{1+2p}$
(0,1,0)	1	0	1/3	$\frac{p}{1+p}$
(1,0,0)	1	0	1/3	$\frac{p}{1+p}$
(0,1,1)	2	1	1/3	$\frac{p}{1+2p}$
(1,0,1)	2	1	1/3	$\frac{p}{1+2p}$
(1,1,0)	2	1	1/3	$\frac{p}{1+2p}$
(1,1,1)	3	2	1	1

A densidade condicional dada por:

$$\begin{aligned}
f_{X_1,X_2,X_3|S=1}(0,1,0) &= \mathbf{P}(X_1=0, X_2=1, X_3=0|S=1) \\
&= \frac{\mathbf{P}(X_1=0,_2=1, X_3=0, S=1)}{\mathbf{P}(S=1)} \\
&= \frac{\mathbf{P}(X_1=0, X_2=1, X_3=0)}{\mathbf{P}(S=1)} \\
&= \frac{(1-p)p(1-p)}{3p(1-p)^2} = \frac{1}{3}
\end{aligned}$$

e

$$\begin{aligned}
f_{X_1,X_2,X_3|T=1}(1,0,1) &= \mathbf{P}(X_1=1, X_2=0, X_3=1|T=1) \\
&= \frac{\mathbf{P}(X_1=1, X_2=0, X_3=1, T=1)}{\mathbf{P}(T=1)} \\
&= \frac{\mathbf{P}(X_1=1, X_2=0, X_3=1)}{\mathbf{P}(T=1)} \\
&= \frac{p^2(1-p)}{p(1-p)^2+3p^2(1-p)} \\
&= \frac{p}{1-p+3p} = \frac{p}{1+2p}
\end{aligned}$$

Pois,

$$\begin{aligned}
\mathbf{P}(T=1) &= \mathbf{P}(X_1=0, X_2=0, X_3=1) + \mathbf{P}(X_1=0, X_2=1, X_3=1) \\
&\quad + \mathbf{P}(X_1=1, X_2=1, X_3=0) + \mathbf{P}(X_1=1, X_2=0, X_3=1) \\
&= p(1-p)^2 + p^2(1-p) + p^2(1-p) + p^2(1-p)
\end{aligned}$$

A distribuição condicional de (X_1, X_2, X_3) dado os valores de S é independente de p, assim S é *estatística suficiente.*

A distribuição condicional de (X_1, X_2, X_3) dado os valores de T não independe de p, portanto, T não é estatística suficiente.

A distribuição condicional só é trabalhável em poucos casos. Primeiro, nós temos que "chutar" uma estatística a ser tratada e depois calcular a distribuição condicional que não é muito fácil, principalmente no caso contínuo.

Temos assim que achar alguns critérios que nos ajudem a encontrar estatísticas suficientes. Antes, note que se tivermos mais de um parâmetro, é improvável que uma única estatística possa ser suficiente para $(\theta_1, \ldots, \theta_k)$. entretanto, sempre existirá um conjunto de estatísticas que serão conjuntamente suficientes.

Definição 3.2 Estatísticas Conjuntamente Suficientes: Seja $X_1, \ldots, X_n)$ uma amostra aleatória da distribuição $f(\cdot, \theta)$. As estatísticas $S_1, \ldots, S_r$ são ditas serem *conjuntamente suficientes* se, e somente se, a distribuição condicional de $X_1, \ldots, X_n$ dado $S_1 = s_1, \ldots, S_r = s_r$ não depende de θ.

Generalizando o resultado do exercício anterior seja $X_1, \ldots, X_n$ uma amostra aleatória de uma Bernouille; por exemplo, para testar a proporção de peças defeituosas. Seja θ a proporção real que queremos descobrir. Neste caso o número de peças defeituosas concentra toda a informação? Isto, no entanto, somente é correto se partirmos do princípio que o modelo é adequado. Se quisermos verificar se o modelo está correto, isto é, independência e probabilidade constante precisaríamos ter cada uma das respostas. No entanto, se o modelo for considerado correto não perderemos nenhuma informação.
Observe que o conjunto de variação de $\mathbf{X} = (X_1, \ldots, X_n$ é a coleção de todos os vetores n-dimensionais com os componentes iguais a 0 ou 1. A Estatística define uma partição de $\mathbf{X}$. (Uma partição de $\mathbf{X}$ é um conjunto de subconjuntos disjuntos cuja união é $\mathbf{X}$). Desta forma, para n = 3, o subconjunto relativo ao valor da estatística igual a 1 é dado por $\{(0,0,1), (0,1,0), (1,0,0)\}$; ou seja, se a Estatística T é suficiente e $t = 1$, a informação de qual dos três possíveis pontos ocorreu não traz nenhuma informação adicional sobre o parâmetro. Como cada estatística induz uma partição podemos também falar em partição suficiente. Observe que uma estatística induz apenas uma partição, mas que uma mesma partição pode ser induzida por diferentes estatísticas. Por exemplo qualquer estatística $T' = g(T)$ tal que $g(\cdot)$ seja uma função $1-1$ no conjunto de variação das estatísticas induz a mesma partição. Um resultado equivalente é dado pelo teorema seguinte.

Teorema 3.3: Se $S_1, \ldots, S_r$ são conjuntamente suficientes para θ e $h : \mathbf{R}^d \to \mathbf{R}^d$ é uma função bijetora então $(T_1, \ldots, T_r) = h(S_1, \ldots, S_r)$ também são conjuntamente suficientes.

Por exemplo, se $S_1 = \sum_{i=1}^n X_i$ e $S_2 = \sum_{i=1}^n X_i^2$ são conjuntamente suficientes para (θ_1, θ_2) então $\bar{X}$ e S^2 também são conjuntamente suficientes para (θ_1, θ_2). Mesmo que μ possa ser estimado somente pela estatística S_1 (mais tarde veremos que S_1 é o ENVUMV de μ) náo podemos dizer que S_1 seja suficiente para μ. O conceito de suficiência está relacionado com todos os parâmetros do modelo. seja

3.2 Critério da Fatorização

A definição de estatística suficiente e conjuntamente suficientes são dadas por definições não muito fáceis de serem verificadas, portanto precisamos de um critério mais fácil.

Teorema 3.2: Seja $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n de uma distribuição $f(\cdot, \theta)$, onde θ pode ser um vetor. Um conjunto de estatísticas $S_1 = s_1(X_1, \ldots, X_n), \ldots, S_r = s_r(X_1, \ldots, X_n)$ são conjuntamente suficientes para θ se, e somente se, a distribuição conjunta de $X_1, \ldots, X_n$ pode ser fatorada como:

$$f(x_1, \ldots, x_n, \theta) = g(s_1(x_1, \ldots, x_n), \ldots, s_r(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)$$

onde $h(x_1, \ldots, x_n)$ é uma função não negativa e não envolve o parâmetro θ e a função $g(t_1, \ldots, t_r, \theta)$ é não negativa e depende de $x_1, \ldots, x_n$ somente através das funcões $s_1, \ldots, s_r$.

Obs.: Se $r = 1$ temos uma estatística suficiente unidimensional.

Note que há muitos conjuntos possíveis de estatísticas suficientes. O teorema acima nos dá um método relativamente fácil de verificar se uma certa estatística é suficiente ou se um conjunto de estatísticas é conjuntamente suficiente. Entretanto, o método não nos diz quando uma estatística **não** é suficiente pois isso pode se dever ao fato de que não termos conseguido fatorar a distribuição conjunta e não pelo fato de não existir tal fatoração.

O teorema acima é útil para se descobrir estatísticas suficientes.

Exercício 3.1: Qual a estatística suficiente para o parâmetro θ nos casos a seguir?
(1) $X_1, \ldots, X_n$ i.i.d. $b(1, \theta)$;
(2) $X_1, \ldots, X_n$ i.i.d. $N(\mu, 1)$, $\theta = \mu$;
(3) $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$;
(4) $X_1, \ldots, X_n$ i.i.d. $U(\theta_1, \theta_2)$, $\theta = (\theta_1, \theta_2$.

Teorema 3.3: Se o estimador de máxima verossimilhança é único ele só depende da amostra através das estatísticas suficientes.
Prova: Se $S_1 = s_1(X_1, \ldots, X_n), \ldots, S_r = s_r(X_1, \ldots, X_n)$ são conjuntamente suficientes para θ então a função de verossimilhança pode ser escrita como

$$L(\theta; x_1, \ldots, x_n) = g(s_1(x_1, \ldots, x_n), \ldots, s_r(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)$$

e $\max L(\theta)$ será atingido no mesmo ponto que $\max g(s_1(x_1, \ldots, x_n), \ldots, s_r(x_1, \ldots, x_n); \theta)$. Caso o estimador seja único ele depende de $x_1, \ldots, x_n$ somente através das funções $s_1, \ldots, s_r$.

O último teorema está anunciado incorretamente no livro; está faltando a condição de unicidade. Como contra-exemplo considere uma uniforme $U(\theta - 1/2, \theta + 1/2)$. É fácil verificar que (X_1, X_n) é suficiente para θ, e que qualquer valor no intervalo $(X_n - 0,5, X_1 + 0,5)$ é um estimador de máxima verossimilhança. Em particular, a estatística $\{(X_n - 0.5) + [cos(X_2)]^2 (X_1 - X_n + 1)\}$ é um estimador de máxima verossimilhança porque seus valores estão no intervalo $(X_n - 0,5, X_1 + 0,5)$, mas ele depende de X_2.
A necessidade da unicidade vem do fato de que no caso de termos mais de um ponto de máximo a escolha do ponto de máximo pode não ser uma função das estatísticas suficientes. Este é o caso no exemplo.

Note que os estimadores pelo método dos momentos podem não ser função somente de estatísticas suficientes.

3.3 Estatísticas Suficientes Minimais

Quando introduzimos o conceito de suficiência, nosso objetivo era condensar a informação contida na amostra sem perder informação sobre o parâmetro. Já vimos que há mais de um conjunto de estatísticas suficientes. Por exemplo, no caso da normal, temos que as estatísticas de ordem $(X_{(1)}, \ldots, X_{(n)})$ são conjuntamente suficientes para μ e σ^2, mas também $\bar{X}$ e S^2 também o são. Mas estas últimas condensam mais a informação. Pergunta: Será que podemos condensar os dados mais ainda? Resposta: Não. Estas são *estatísticas suficientes minimais.*

Definição 3.3: Estatística suficiente minimal: Um conjunto de estatísticas conjuntamente suficientes é dito ser *minimal* se, e somente se, é uma função de todo outro conjunto de estatísticas suficientes.

Note que a definição acima é inútil para se realmente encontrar as estatísticas suficientes minimais. Uma definição equivalente pode ser conseguida através de partição mais "grossa". Este conceito está por trás de um dos teoremas que auxiliam a procurar estatísticas suficientes minimais. Mais tarde estudaremos uma classe de distribuições, a família exponencial, onde o fato da densidade ser propriamente fatorada, nos dá um conjunto de estatísticas suficientes minimais. Inicialmente veremos o teorema

Teorema 3.4: Teorema: Seja $(X_1, \ldots, X_n)$ uma amostra aleatória de tamanho n de uma densidade $f(\cdot;\theta)$. Suponha que exista uma função $T(X_1, \ldots, X_n)$ tal que para dois pontos $\mathbf{x} = \{x_1, \ldots, x_n\}$ e $\mathbf{y} = \{y_1, \ldots, y_n\}$ a razão $f_c(x;\theta) = f_c(y;\theta)$ é constante como função de θ se e somente se $T(\mathbf{x}) = T(\mathbf{y})$. Então $T(\mathbf{X})$ é uma estatística suficiente minimal. $f_c(\cdot;\theta)$ é a função densidade conjunta.

Exemplo 3.2: Seja $(X_1, \ldots, X_n)$ uma amostra aleatória de uma $N(\mu, \sigma)$, onde os dois parâmetros são desconhecidos. Sejam $\mathbf{x}$ e $\mathbf{y}$ dois pontos amostrais e $(\bar{\mathbf{x}}, s^2_{\mathbf{x}})$ e $(\bar{\mathbf{y}}, s^2_{\mathbf{y}})$ as médias e variâncias amostrais dos pontos $\mathbf{x}$ e $\mathbf{y}$, respectivamente. Então a razão das duas densidades conjuntas é dada por

$$\frac{f_c(\mathbf{x};\mu,\sigma^2)}{f_c(\mathbf{y};\mu,\sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\{-[(n\bar{x}-\mu)^2+(n-1)s_x^2]/(2\sigma^2)\}}{(2\pi\sigma^2)^{-n/2}\exp\{-[(n\bar{y}-\mu)^2+(n-1)s_y^2]/(2\sigma^2)\}}$$

A razão será constante como uma função de μ e σ se e somente se $\bar{x} = \bar{y}$ e $s_x^2 = s_y^2$. Então pelo teorema anterior $(\bar{X}, S^2)$ é um conjunto de estatísticas suficiente minimal para (μ, σ).

3.4 Famílias Exponenciais

Muitas das distribuições que estamos interessados em estudar têm características e propriedades comuns e são agrupadas em uma classe de distribuições chamada *família exponencial.* (não confundir com a distribuição exponencial que será um caso particular desta classe). Para uma distribuição na classe da família exponencial será muito fácil encontrar a estatística suficiente minimal completa (que veremos a seguir) e pa partir delas encontrar ENVUMV's. Estes modelos também são importantes pois eles têm muita coisa em comum quando queremos fazer inferências a respeito deles. Reconhecê-los como casos especiais de modelos mais gerais torna possível derivar os resultados em comum que teriam que ser obtidos caso a caso.

Definição 3.4: Família Exponencial de Distribuições Uniparamétrica: Uma família de distribuições uniparamétrica que pode ser escrita (através de uma escolha adequada de funções) como:

$$f(x;\theta) = B(\theta)h(x)\exp[Q(\theta)T(x)]$$

é dita pertencer à *família exponencial de distribuições uniparamétrica.*

A maioria das distribuições que encontramos até o presente momento pertencem a esta família.

Exemplo 3.3:

(1) **Bernoulli:** $f(x;p) = p^x(1-p)^{1-x}$.

$$B(p) = 1-p; \quad Q(p) = \log \frac{p}{1-p}; \quad T(x) = x; \quad h(x) = 1.$$

(2) **Binomial:** $f(x;p) = \binom{n}{x} p^x(1-p)^{n-x}$.

$$B(p) = (1-p)^n; \quad Q(p) = \log \frac{p}{1-p}; \quad T(x) = x; \quad h(x) = \binom{n}{x}.$$

(3) **Geométrica:** $f(x;p) = p(1-p)^x$.

$$B(p) = p; \quad Q(p) = \log(1-p); \quad T(x) = x; \quad h(x) = 1.$$

(4) **Binomial Negativa:** $f(x;p) = \binom{r+x-1}{x} p^r(1-p)^{x-1}$.

$$B(p) = p^r(1-p)^{-1}; \quad Q(p) = \log(1-p); \quad T(x) = x; \quad h(x) = \binom{r+x-1}{x}.$$

(5) **Poisson:** $f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$.

$$B(\lambda) = e^{-\lambda}; \quad Q(\lambda) = \log\lambda; \quad T(x) = x; \quad h(x) = \frac{1}{x!}.$$

(6) **Exponencial:** $f(x;\lambda) = \lambda e^{-\lambda x}$.

$$B(\lambda) = \lambda; \quad Q(\lambda) = -\lambda; \quad T(x) = x; \quad h(x) = 1.$$

(7) **Normal ($N(0,\theta)$):** $f(x;\theta) = (2\pi\theta)^{-1/2}\exp[-x^2/2\theta]$.

$$B(\theta) = (2\pi\theta)^{-1/2}; \quad Q(\theta) = -(2\theta)^{-1}; \quad T(x) = x^2; \quad h(x) = 1.$$

(8) **Normal ($N(\theta,1)$) :** $f(x;\theta) = (2\pi)^{-1/2}\exp[-\frac{1}{2}(x-\theta)^2]$.

$$B(\theta) = (2\pi)^{-1/2}\exp[-\frac{1}{2}\theta^2]; \quad Q(\theta) = \theta; \quad T(x) = x; \quad h(x) = \exp[-\frac{1}{2}x^2].$$

(9) **Gama:** $f(x;\lambda) = \lambda^n x^{n-1}\frac{e^{-\lambda x}}{(n-1)!}$.

$$B(\lambda) = \frac{\lambda^n}{(n-1)!}; \quad Q(\lambda) = -\lambda; \quad T(x) = x; \quad h(x) = x^{n-1}.$$

(10) **Raleigh:** $f(x;\theta) = \frac{x}{\theta^2}\exp[-x^2/2\theta^2]$.

$$B(\theta) = \frac{1}{\theta^2}; \quad Q(\theta) = -(2\theta^2)^{-1}; \quad T(x) = x^2; \quad h(x) = x.$$

Definição 3.5: Família Exponencial de Distribuições: Uma família de distribuições indexada por um parâmetro $\theta = (\theta_1, \ldots, \theta_k)$ que pode ser escrita (através de uma escolha adequada de funções) como:

$$f(x;\theta) = B(\theta)h(x)\exp[(Q_1(\theta)T_1(x) + \cdots + Q_k(\theta)T_k(x))]$$

é dita pertencer à *família exponencial* de distribuições.

Na definição X pode ser uma variável aleatória multivariada.

Exemplo 3.4: Considere a distribuição normal $N(\mu, \sigma^2)$ indexada pelo parâmetro $\theta = (\mu, \sigma^2)$:

$$\begin{aligned} f(x; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\mu^2}{2\sigma^2}) \exp[-\frac{x^2}{2\sigma^2} + x\frac{\mu}{\sigma^2}]. \end{aligned}$$

A qual pertence à família exponencial com a seguinte identificação:

$$B(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\mu^2}{2\sigma^2}); \quad h(x) = 1;$$

$$Q_1(\mu, \sigma^2) = -\frac{1}{2\sigma^2}; \quad R_1(x) = x^2;$$

$$Q_2(\mu, \sigma^2) = \frac{\mu}{\sigma^2}; \quad T_2(x) = x.$$

Exercício 3.2: Verifique que a distribuição *Beta* com densidade:

$$f(x; r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} x^{r-1}(1-x)^{s-1}, \ 0 < x < 1$$

pertence à família exponencial.

Note que nem todas as distribuições pertencem a família exponencial. Alguns exemplos são a distribuição de Cauchy e a distribuição uniforme. Na verdade, qualquer famía de densidades na qual o conjunto de valores, para os quais a densidade não é negativa depende de θ, não pertence à classe exponencial.

Teorema 3.5: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma de uma densidade exponencial dada por

$$f(\mathbf{x}; \theta) = B(\theta)h(\mathbf{x}) \exp[(Q_1(\theta)T_1(\mathbf{x}) + \cdots + Q_k(\theta)T_k(\mathbf{x}))].$$

então

$$(S_1, \ldots, S_k) = (\sum_{i=1}^n T_1(x_i), \ldots, \sum_{i=1}^n T_k(x_i))$$

é conjuntamente suficiente para θ.

Prova: A densidade conjunta de n observações independentes a densidade conjunta é dada por:

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= \prod_{i=1}^n B(\theta)h(x_i) \exp[(Q_1(\theta)T_1(x_i) + \cdots + Q_k(\theta)T_k(x_i))] \\ &= B(\theta)^n \prod_{i=1}^n h(x_i) \exp[(Q_1(\theta) \sum_{i=1}^n T_1(x_i) + \cdots + Q_k(\theta) \sum_{i=1}^n T_k(x_i))] \end{aligned}$$

e também pertence à família exponencial e mais ainda, se

$$h(x_1, \ldots, x_n) = \prod_{i=1}^n h(x_i)$$

e

$$g(s_1, \dots, s_k; \theta) = B(\theta)^n \exp[(Q_1(\theta)s_1 + \cdots + Q_k(\theta)s_k)]$$

temos que

$$f(x_1, \dots, x_n; \theta) = g(s_1, \dots, s_k; \theta)h(x_1, \dots, x_n)$$

e portanto

$$(S_1, \dots, S_k) = (\sum_{i=1}^{n} T_1(x_i), \dots, \sum_{i=1}^{n} T_k(x_i))$$

é conjuntamente suficiente para θ.
Pode-se mostrar também que é conjuntamente suficiente completa (definição dada a seguir) e minimal caso a região de variação de $(Q_1(\theta, \dots, Q_k(\theta) contenhauminteriornãovazio. Pode-setambémmostrarqueseuma$

Exemplo 3.5: No caso da normal vimos que o vetor (x, x^2) é conjuntamente suficiente para (μ, σ^2). Dada uma amostra aleatória, pela propriedade anterior, $(\sum x_i, \sum x_i^2)$ é um vetor conjuntamente suficiente minimal para (μ, σ^2). Temos que este vetor é uma transformação 1-1 do vetor $(\bar{X}, S^2)$ e portanto não existe contradição com o Exemplo onde deduzimos que este vetor era suficiente minimal. Observe que embora $\bar{X}$ seja utilizado para estimar μ e S^2 para estimar σ^2 não podemos dizer que $\bar{X}$ seja suficeinte para μ e S^2 seja suficente para σ^2. Verifique se poderíamos fazer estas afirmações se um dos estimadores é conhecido.

Existem várias formas de se escrever uma densidade pertencente à família exponencial. Por exemplo, fazendo-se uma reparametrização, $\eta = \eta(\theta$, a densidade de uma distribuição da família exponencial pode ser escrita da forma:

$$f(x; \eta) = A(\eta)b(x) \exp[\sum_{i=1}^{k} \eta_i d_i(x)],$$

esta reparametrização é chamada de **parametrização natural da família exponencial**.

Outra forma de representar uma densidade de uma **família exponencial uniparamétrica na forma natural** é dada por

$$f(x; \eta) = \{\exp[\eta T(x) + d(\eta) + S(x)\} I_A(x)$$

Neste caso temos que, se η é um ponto interior, a função geratriz de momentos de $T(X)$ existe e é dada por

$$\psi(s) = exp[d(\eta) - d(s + \eta)]$$

para s em alguma vizinhança de 0. Uma aplicação imediata é que

$$E[T(X)] = -d'(\eta) \quad \text{e} \quad V[T(X)] = -d''(h).$$

Exemplo 3.6: Seja $X_1, \dots, X_n$ uma amostra aleatória de uma distribuição de Rayleigh (utilizada para modelar tempo de falha de certos equipamentos)

$$\begin{aligned} f(x; \theta) &= (x/\theta^2) \exp(-x/2\theta^2) \quad , x > 0 \ , \quad \theta > 0 \\ &= \exp\{-(2\theta^2\}^{-1} x^2 - \log(\theta) + \log x\} I_{(0,\infty)}(x) \end{aligned}$$

na forma anterior temos $\eta = -1/(2\theta^2$, ou seja $\theta^2 = -1/2\eta$, $d(\eta) = n \log(-2\eta)$. Portanto, $E(X_i^2) = -1/\eta = 2\theta^2$ e $V(X) = 1/\eta^2 = 4\theta^4$.

3.5 Estatísticas Ancilares

Ao introduzir o conceito de estatística suficiente comentamos que certas estatísticas dão informação sobre o parâmetro porque sua distribuição depende do parâmetro θ. Devemos tomar cuidado, no entanto, para olhar para toda a informação. Considere os seguintes exemplos:

Exemplo 3.7: Considere a distribuição uniforme $U(\theta, \theta+1)$. Neste caso temos que $(X_n - X_1, X_n + X_1)$ é uma estatística suficiente minimal. No entanto é fácil verificar que a distribuição da diferença entre as estatísticas de ordem extremais independe de θ, mas que condicionado em toda a informação na amostra ela depende de θ. Estas estatísticas cuja distribuição independe do parâmetro são chamadas de estatísticas ancilares, e desempenham um papel importante em certas áreas de inferência estatística.

Exemplo 3.8 : Suponha que uma variável aleatória X tem a mesma probabilidade de vir de uma normal $N(\mu, \sigma_1^2)$ ou da $N(\mu, \sigma_2^2\ )$. Considere a variável aleatória C que irá decidir de onde virá a variável aleatória. Se ela for igual a 1 a amostra virá da primeira normal e se C for igual a 2 virá da segunda. Temos $P[C = 1] = P[C = 2] = 1/2$. A verossimilhança no caso é dada por

$$f_{C,X}(c, x) = \frac{\exp\left[-\frac{(x-\mu)^2}{2\sigma_c^2}\right]}{2(2\pi)^{0.5}\sigma_c^2}$$

Pelo teorema da fatoração temos que (C, X) é suficiente para μ quando as variâncias são conhecidas. Embora a distribuição de C seja fixa e conhecida temos que X não é suficiente. Observe que a razão $f_{(C,X)}(1, x)/f_{(C,X)}(2, x)$ deveria ser idependente de μ para qualquer x se X fosse suficiente. Calcule a razão para X igual a zero e vefifique se esta condição é satisfeita. A razão deveria ser constante porque deveríamos estar no mesmo subconjunto da partição gerada pela estatística suficiente X.

3.6 Cota Inferior Para Variância

Como já vimos estimadores que tenham uniformemente mínimo EQM não existem e se queremos usar este critério devemos restringir a classe de estimadores sob estudo. Vamos nos restringir a classe de *estimadores não viciados.*

Definição 3.6: Estimador Não Viciado Uniformemente de Mínima Variância (ENVUMV): Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $f(x; \theta)$. Um estimador $\Gamma^* = \gamma^*(X_1, \ldots, X_n)$ de $\tau(\theta)$ é definido como sendo um *estimador não viciado uniformemente de mínima variância (ENVUMV)* se, e somente se:
(i) $\mathbf{E}_\theta(\Gamma^*) = \tau(\theta)$, isto é, Γ^* é não viciado;
(ii) $\text{Var}_\theta[\Gamma^*] \leq \text{Var}_\theta[S]$ para qualquer outro estimador S não viciado de $\tau(\theta)$.

O problema agora é como encontrar um ENVUMV. Um bom início seria termos uma idéia da mínima variância que poderia ser atingido pelos estimadores não viciados. Se tivermos este limite e encontrarmos um estimador não viciado que atinge este limite teremos encontrado um ENVUMV. Este limite existe, mas como veremos mais tarde ele é mais importante em outras aplicações e não para encontrar o ENVUMV.

Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $f(x; \theta)$, onde $\theta \in \Theta$. Assuma que θ é univariado. Seja $T = t(X_1, \ldots, X_n)$, um estimador não viciado de $\tau(\theta)$. Suponha primeiramente que $f(x; \theta)$ é uma densidade de probabilidade (você deveria tentar fazer o desenvolvimento análogo se $f(x; \theta)$ é função de probabilidade). Sempre assumiremos que as seguintes *condições de*

regularidade são satisfeitas:
(i) $\frac{\partial}{\partial\theta}f(x;\theta)$ existe para todo x e θ;
(ii)

$$\frac{\partial}{\partial\theta}\int\cdots\int\prod_{i=1}^{n}f(x_i;\theta)dx_1\ldots dx_n$$
$$=\int\cdots\int\frac{\partial}{\partial\theta}\prod_{i=1}^{n}f(x_i;\theta)dx_1\ldots dx_n$$

(iii)

$$\frac{\partial}{\partial\theta}\int\cdots\int t(x_1,\ldots,x_n)\prod_{i=1}^{n}f(x_i;\theta)dx_1\ldots dx_n$$
$$=\int\cdots\int t(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}\prod_{i=1}^{n}f(x_i;\theta)dx_1\ldots dx_n$$

(iv) $0<\mathbf{E}[[\frac{\partial}{\partial\theta}\log f(X;\theta)]^2]<\infty$, para todo $\theta\in\Theta$.

Teorema 3.6: Desigualdade de Cramér- Rao. sob as condições de regularidade acima temos que:

$$\text{Var}_\theta[T]\geq\frac{[\tau'(\theta)]^2}{n\mathbf{E}[[\frac{\partial}{\partial\theta}\log f(X;\theta)]^2]}$$

onde $t=t(X_1,\ldots,X_n)$ é um estimador não viciado de $\tau(\theta)$.

A equação acima é chamada de **desigualdade de Cramér-Rao** e a expressão à direita é chamada **cota inferior de Cramér-Rao** para a variância de estimadores não viciados de $\tau(\theta)$.

O teorema pode ser utilizado de duas formas:
(1) O teorema nos dá uma cota inferior para a variância de estimadores não viciados e portanto se temos um estimador não viciado cuja variância atinge a cota inferior de Cramér-Rao, sabemos que temos o ENVUMV. Infelizmente este nem sempre é o caso.
(2) Por outro lado, se não conseguimos achar o ENVUMV, mas temos um estimador não viciado cuja variância esteja perto da cota inferior de Cramér-Rao sabemos que temos um "bom" estimador. (3) O fato de um estimador ter uma variância muito longe da cota inferior de Cramér-Rao não significa que ele seja ruim porque a variância do ENVUMV também pode estar longe da quota inferior. Este comentário vale para pequenas amostras porque como veremos mais tarde, dentro de certas condições de regularidade, para grandes amostras os estimadores de máxima verossimilhança tem vício pequeno e sua variância é próxima à quota inferior.

Exemplo 3.9: Seja $X_1,\ldots,X_n$ uma amostra aleatória de uma distribuição exponencial com parâmetro θ, suponha que desejamos estimar θ. Pode-se mostrar facilmente que as condições de regularidade são satisfeitas (tente!). Neste caso, $\tau'(\theta)=1$ e

$$\text{Var}_\theta[T]\geq\frac{1}{n\mathbf{E}[[\frac{\partial}{\partial\theta}\log f(X;\theta)]^2]}.$$

Note que, $\frac{\partial}{\partial\theta}\log f(x;\theta)=\frac{\partial}{\partial\theta}(\log\theta-\theta x)=1/\theta-x$, e portanto,

$$\mathbf{E}[[\frac{\partial}{\partial\theta}\log f(X;\theta)]^2]=\mathbf{E}[(\frac{1}{\theta}-X)^2]=\text{Var}(X)=\frac{1}{\theta^2}$$

Assim, a cota inferior de Cramér-Rao para a variância de estimadores não viciados de θ é dada por:

$$\mathrm{Var}_\theta[T] \geq \frac{1}{n(1\theta^2)} = \frac{\theta^2}{n}.$$

Teorema 3.7: Se a estimativa de máxima verossimilhança de θ, digamos, $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ é dada pela solução da equação

$$\frac{\partial}{\partial\theta} \log L(\theta; x_1, \ldots, x_n) = \frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(x_i; \theta) = 0$$

e se Γ^* é um estimador não viciado de $\tau(\theta)$ que atinge a cota inferior de Cramér-Rao, então $\Gamma^* = \tau(\hat{\theta}(X_1, \ldots, X_n))$.

Exercício 3.3: Ache a cota inferior de Cramér-Rao para a variância de estimadores não viciados de $\frac{1}{\theta}$ e verifique que $\bar{X}$ é ENVUMV para $\frac{1}{\theta}$.

Nem sempre a cota inferior de Cramér-Rao pode ser atingida, portanto precisamos de outros métodos para encontrar ENVUMV's. Para isto precisamos do conceito de estatística suficiente e completa.

3.7 Suficiência e Completitude

Uma ilustração para o tipo de problemas que podemos resolver com os resultados desta seção. Suponha que $X_1, \ldots, X_n$ seja uma amostra aleatória de uma distribuição $N(\mu, \sigma^2)$ onde $\theta = (\mu, \sigma)$. Em 1920, uma aposta surgiu entre o físico A. Eddington e um dos fundadores da estatística, Sir R.A. Fisher, sobre qual o melhor estimador para σ. Fisher argumentava que um múltiplo do desvio padrão amostral

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

deveria ser utilizado, enquanto que Eddington propunha que um múltiplo do desvio médio amostral

$$\tilde{\sigma} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}|.$$

Os múltiplos naturais a serem considerados são aqueles que dão um estimador não viciado para σ. Sejam, $\hat{\sigma}_1 = a\hat{\sigma}$ e $\tilde{\sigma}_1 = c\tilde{\sigma}$. Vamos mostrar que $\hat{\sigma}_1$ é ENVUMV. Portanto, $\hat{\sigma}_1$ é sempre melhor que $\tilde{\sigma}_1$ e a aposta foi ganha por Fisher.

Na próxima subseção iremos mostrar que, ao procurarmos estimadores não viciados, devemos nos ater aqueles que sejam funções somente da estatísticas suficientes. Caso o estimador encontrado seja função somente de estatísticas suficiente, que tenham a propriedade de completude, propriedade esta que será definida a seguir, teremos encontrado um ENVUMV.

Definição 3.7: Completitude: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $f(\cdot, \theta)$ com espaço paramétrico Θ, e seja $T = t(X_1, \ldots, X_n)$ uma estatística. A família de distribuição de T é dita ser *completa* se, e somente se, a única função g que satisfaz

$$\mathbf{E}_\theta[g(T)] = 0, \text{para todo } \theta,$$

é a função $g(T) = 0$.

Um outro modo de dizer que T é completa é dizer o seguinte: T é completa se, e somente se, o único estimador não viciado de 0 que é uma função de T é a estatística que é identicamente zero.

Exemplo 3.10: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma densidade Bernoulli. A estatística $T = X_1 - X_2$ não é completa porque $E_\theta[X_1 - X_2] = 0$ e $X_1 - X_2$ não é igual a zero com probabilidade 1. Considere agora a estatística $T = \sum_{i=1}^{n} X_i$. Seja $g(T)$ qualquer estatística que é função de T e para a qual $E_\theta[g(T)] = 0$ para todo $\theta \in \Theta$, isto é, para $0 \leq \theta \leq 1$. Para mostrar através da definição que T é completa precisamos mostrar que $g(t) = 0$ para um conjunto de valores de T com probabilidade igual a 1, isto é, para $t = 0, 1, \ldots, n$. Mas

$$\begin{aligned} E_\theta[g(T)] &= \sum_{i=0}^{n} g(t) \begin{pmatrix} n \\ t \end{pmatrix} \theta^t (1-\theta)^{n-t} \\ &= (1-\theta)^n \sum_{i=0}^{n} g(t) \begin{pmatrix} n \\ t \end{pmatrix} \left(\frac{\theta}{1-\theta}\right)^t \end{aligned}$$

como $E_\theta[g(T)] \equiv 0$ para todo $0 \leq \theta \leq 1$ isto implica que

$$\sum_{i=0}^{n} g(t) \begin{pmatrix} n \\ t \end{pmatrix} \left(\frac{\theta}{1-\theta}\right)^t \equiv 0 \quad \text{isto é} \quad \sum_{i=0}^{n} g(t) \begin{pmatrix} n \\ t \end{pmatrix} \alpha^t \equiv 0$$

para todo α, onde $\alpha = \theta/(1-\theta)$. Para que o polinômio em α seja identicamente zero cada coeficiente de α^t tem que ser zero, isto é, $g(t) \begin{pmatrix} n \\ t \end{pmatrix} = 0$ para $t = 0, \ldots, n$; mas como $\begin{pmatrix} n \\ t \end{pmatrix} \neq 0$ temos que $g(t) = 0$ para $t = 0, \ldots, n$.

Exemplo 3.11: Suponha que $X_1, \ldots, X_n$ seja uma amostra aleatória de uma distribuição Poisson(λ). Sabemos que $T = \sum_{i=1}^{n} X_i$ é suficiente para λ e também que soma de v.a.'s independentes Poisson(λ) é Poisson($n\lambda$). Vamos verificar que T é completa. Suponha que g seja uma função tal que $\mathbf{E}_\lambda[T] = 0$ para todo $\lambda > 0$. Então:

$$e^{-n\lambda} \sum_{k=1}^{\infty} \frac{g(k)(n\lambda)^k}{k!} = 0$$

para todo $\lambda > 0$. Sabemos dos teoremas de cálculo que uma série de potência que é identicamente zero deve ter todos os seus coeficientes iguais a zero. Portanto, $g(k) = 0$, para todo $k = 0, 1, \ldots$,c.q.d.

Exemplo 3.12: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $U[0, \theta]$ onde $\Theta = (0, \infty)$. Mostre que a estatística $X_{(n)} = \max\{X_1, \ldots, X_n\}$ é completa.

Seja g uma função tal que $\mathbf{E}_\theta[X_{(n)}] = 0$ para todo $\theta > 0$. Sabemos a densidade de $X_{(n)}$ e portanto,

$$\mathbf{E}_\theta[X_{(n)}] = \int_0^\theta g(y) \theta^{-n} n y^{n-1} dy$$

e temos $\mathbf{E}_\theta[X_{(n)}] = 0$ para todo $\theta > 0$ se, e somente se,

$$\frac{n}{\theta^n} \int_0^\theta g(y) y^{n-1} dy = 0, \text{para todo } \theta > 0,$$

ou equivalentemente,

$$\int_0^\theta g(y)y^{n-1}dy = 0, \text{para todo } \theta > 0.$$

Quando derivamos ambos os lados desta igualdade com respeito à θ temos que $g(\theta)\theta^{n-1} = 0$ para todo $\theta > 0$ o que implica que $g(\theta) = 0$ para todo $\theta > 0$.

Como podemos ver descobrir se uma estatística é completa não é tarefa muito fácil, entretanto, para famílias exponenciais vimos no Teorema 3.5 como encontrar a estatística suficiente completa minimal.

Exercício 3.4: Utilize o Teorema 3.5 para verificar se as estatísticas encontradas nos Exemplos 3.3 e 3.4 são completas e minimais.

3.8 Métodos Para Encontrar ENVUMV

Nesta subseção serão dados alguns métodos para encontrar ENVUMV. O primeiro teorema mostra como a partir de qualquer estimador não viciado, que não seja função da estatística suficiente é possível encontrar outro estimador não viciado, funçã somente da estatística suficiente é e que tenha menor variância que o estimador inicial. O segundo teorema mostra que se a estatística suficiente utilizada for completa o estimador encontrado é um ENVUMV.

Teorema 3.8 Rao-Blackwell Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma distribuição $f(\cdot, \theta)$ e sejam $S_1 = s_1(X_1, \ldots, X_n), \ldots, S_k = s_k(X_1, \ldots, X_n)$ conjuntamente suficientes para θ. Seja a estatística $T = t(X_1, \ldots, X_n)$ um estimador não viciado de $\tau(\theta)$. Defina um outro estimador T^* por $T^* = \mathbf{E}[T|S_1, \ldots, S_k]$. Então,

(i) T^* é uma estatística e é uma função das estatísticas suficientes $S_1, \ldots, S_k$. Portanto, podemos escrever $T^* = t^*(S_1, \ldots, S_k)$.

(ii) $\mathbf{E}_\theta[T^*] = \tau(\theta)$; isto é, T^* é um estimador não viciado de $\tau(\theta)$.

(iii) $\text{Var}_\theta[T^*] \leq \text{Var}_\theta[T]$ para todo θ, e $\text{Var}_\theta[T^*] < \text{Var}_\theta[T]$ para algum θ a menos que $\mathbf{P}[T^* = T] = 1$.

Prova: (i) $S_1, \ldots, S_k$ são estatísticas suficientes; portanto a distribuição condicional conjunta da amostra , e consequentemente a distribuição condicional de T, dado $S_1, \ldots, S_k$ é independente de θ. Portanto, T^* é independente de θ e como é uma função de $S_1, \ldots, S_k$ é uma estatística.
(ii) Pela definição de T^* e propriedades de esperança condicional temos:

$$\mathbf{E}_\theta[T^*] = \mathbf{E}_\theta[\mathbf{E}[T|S_1, \ldots, S_k]] = \mathbf{E}_\theta[T] = \tau(\theta)$$

(iii) Pelas propriedades de variância e esperança condicional temos que:

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

Entretanto,

$$\mathbf{E}[(T - T^*)(T^* - \mathbf{E}_\theta[T^*])] = \mathbf{E}_\theta[\mathbf{E}[(T - T^*)(T^* - \mathbf{E}_\theta[T^*])|S_1, \ldots, S_k]] = 0$$

e portanto,
$$\text{Var}_\theta[T] = \mathbf{E}[(T-T^*)^2] + \text{Var}[T^*] \geq \text{Var}_\theta[T^*]$$

Geralmente, temos um ou dois parâmetros desconhecidos e a distribuição condicional é possível de ser encontrada. Pode ou não ser fácil de ser encontrada.

Exemplo 3.13: Considere uma amostra de tamanho n de uma Bernoulli com parâmetro θ. No teorema anterior podemos tomar $T = X_1$ porque sabemos que X_1 é um estimador não viciado de θ. Já vimos que $S = \sum_{i=1}^n X_i$ é uma estatística suficiente para a Bernoulli. Para aplicar o teorema de Rao-Blackwell precisamos calcular $T^* = E(X_1 | \sum_{i=1}^n X_i = s)$ Como X_1 é uma função indicadora temos que $T^* = P(X_1 = 1 | \sum_{i=1}^n X_i = s)$. Dado a independência sabemos que os "sucessos" estão uniformemente distribuídos, e como existem s "sucessos" em n ensaios temos que $T^* = s/n$. É fácil verificar que este estimador é não viciado, tem variância menor do que o estimador inicial e é função somente da estatística completa. Pode-se verificar que atinge o LICR e portanto temos um ENVUMV. Resultados posteriores permitem dizer que ele é o único ENVUMV.
Verifique no livro do Mood it et al. na página 322-3 como se chega a este resultado calculando a distribuição de probabilidade de X_1 condicionada à estatística suficiente.

Obs.:
(1) Se um estimador T não viciado de $\tau(\theta)$ já é uma função das estatísticas suficientes então $T^* = \mathbf{E}[T|S_1, \ldots, S_k]$ é igual a T e não podemos esperar uma diminuição da variância.
(2) Para aplicar o teorema acima, podemos utilizar qualquer conjunto de estatísticas suficientes, mas é natural utilizar um conjunto de estatísticas suficientes minimal.

Se aplicássemos o Teorema de Rao-Blackwell aos nossos estimadores $\hat{\sigma}_1$ e $\tilde{\sigma}_1$, sabemos que $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ são as estatísticas suficientes e como $\hat{\sigma}_1$ é função de $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$, mas $\tilde{\sigma}_1$ não é, temos que podemos melhorar $\tilde{\sigma}_1$ mas não $\hat{\sigma}_1$. Mas ainda não podemos concluir que $\hat{\sigma}_1$ é ENVUMV, para isso precisamos do conceito de completitude definido na subseção anterior e o próximo teorema.

Teorema 3.9: Lehmann-Scheffé Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $f(\cdot, \theta)$, $\theta \in \Theta$. Se $S = s(X_1, \ldots, X_n)$ é uma estatística suficiente e completa e se $T^* = t^*(S)$ é um estimador não viciado de $\tau(\theta)$, então T^* é um ENVUMV de $\tau(\theta)$.

Prova: Seja T' qualquer outro estimador não viciado de $\tau(\theta)$ que seja função de S, digamos $T' = t'(S)$. Então
$$\mathbf{E}_\theta[T^* - T'] = 0, \text{para todo } \theta \in \Theta$$
Mas, $T^* - T'$ é uma função de S e como S é completa temos que $T^* = T'$. Assim, há um único estimador não viciado de $\tau(\theta)$ que é função de S. Sabemos que se T é qualquer outro estimador não viciado de $\tau(\theta)$ então $\mathbf{E}[T|S] = T^*$ e daí, pelo teorema de Rao-Blackwell temos que
$$\text{Var}_\theta[T^*] \leq \text{Var}_\theta[T]$$
para todo $\theta \in \Theta$ e portanto um ENVUMV para $\tau(\theta)$.

Resultados análogos (que não serão enunciados aqui) valem para famílias multiparamétricas. Pode-se mostrar que existe um único ENVUMV caso a variância do estimador seja finito para todo o espaço paramétrico.

Métodos para encontrar ENVUMV: A interpretação dos teoremas de Rao-Blackwell e Lehmann-Scheffé nos dá os seguintes métodos para encontrar estimadores NVUMV:

Método 1: Aplicação direta do teorema de Lehmann-Scheffé. Se temos um estimador não viciado função somente da estatística suficiente e completa, então ele é um ENVUMV.

Método 2: Se T é um estimador não viciado S é uma estatística suficiente e completa então $T^* = E[T|S]$ pelos 2 teoremas é um ENVUMV. Este método é bastante utilizado quando queremos encontrar um ENVUMV de $P[X \in A]$. Neste caso sabemos que a função indicadora $I_A(X)$ é um estimador nao viciado da probabilidade procurada.

Método 3: Dado um estimador função somente da estatística suficiente e completa podemos encontrar a sua esperança. Se o estimador for não viciado já temos um ENVUMV. Caso o estimador tenha um vício que pode ser corrigido continuando um estimador função somente da estatística suficiente e completa temos um ENVUMV.

Método 4: Às vezes queremos encontrar o ENVUMV de $\tau(\theta)$ e temos um estimador T função somente da estatística suficiente e completa. Neste caso podemos verificar se o estimador $\tau(T)$. Claramente este estimador é função somente da estatística suficiente e completa e podemos tentar aplicar o método anterior.

Método 5: Este método não é aplicação dos 2 teoremas, mas aplicação do limite inferior de Cramér-Rao. Caso um estimador não viciado atinja o LICR ele é o ENVUMV.

Nota: Observe que sempre colocamos **um** ENVUMV. No entanto, se a variância for finita para todo o espaço paramétrico temos **o** ENVUMV.

Exemplo 3.14: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $N(\mu, \sigma^2)$. Como já vimos no *Exercício 3.4* que a distribuição $f(\cdot; \mu, \sigma^2)$ pertence à família exponencial biparamétrica e que $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ é uma estatística suficiente, completa e minimal.

$\bar{X}$ é uma função de T e estimador não viciado de μ. Portanto, $\bar{X}$ é o ENVUMV de μ. Sempre procuraremos adotamos o artigo definido quando a variância do estimador for finito para todo o espaço paramétrico.

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ é uma função de T e estimador não viciado de σ^2; portanto S^2 é um ENVUMV de σ^2. Aplicamos aqui o **Método 1**. Ele é o único ENVUMV porque sabemos que a variância de uma qui-quadrada é finita.

Exemplo 3.15: No problema anterior encontre um ENVUMV do desvio padrão.
Sabemos que S^2 é um estimador de σ^2 função somente da estatística completa. Sabemos que S é um estimador viciado de σ, mas se pudermos corrigir o vício conforme sugerido antes teremos um ENVUMV de σ. Estamos aqui tentando aplicar o **Método 4**. Temos que

$$E(S) = \frac{\Gamma(n/2)}{\Gamma(\frac{n-1}{2})\sqrt{\frac{n-1}{2}}}\sigma;$$

logo,

$$\frac{\Gamma(\frac{n-1}{2})\sqrt{\frac{n-1}{2}}}{\Gamma(n/2)}S$$

é um ENVUMV de σ como sugeria Fisher.

Exemplo 3.16 *Estimando a probabilidade de falha precoce.* Algumas vezes é razoável se pensar que o tempo de vida (tempo para o primeiro conserto) de um equipamento é uma variável aleatória exponencialmente distribuída com parâmetro λ o qual é desconhecido.

Suponha que n equipamentos idênticos são selecionados e seus tempos de falha $X_1, \ldots, X_n$ são observados. Desejamos estimar a probabilidade de uma falha precoce, isto é, $\mathbf{P}_\lambda[X_1 \leq x] = 1 - e^{-\lambda x}$ para algum valor x pré-fixado.

Observe que estamos tentando estimar a probabilidade de um evento A definido como falha até o tempo x. Como comentado antes podemos tentar o **Método 2**, a partir da função indicadora $S(X_1) = I_{[X_1 \leq x]}$, que é um estimador não viciado de $\tau(\lambda)$, mas não é função da estatística suficiente. Já vimos que a distribuição exponencial pertence à família exponencial e que $T = \sum_{i=1}^n X_i$ é suficiente e completa para λ. Um ENVUMV pode ser encontrando calculando

$$T^* = \mathbf{E}_\theta[S(X_1)|T]$$

$$\begin{aligned} E[S|T] &= E[I_{[X_1 \leq x]}|T = t] = P(X_1 \leq x|T = t) \\ &= \int_0^x f_{(X_1|T)}(y|t)dy \end{aligned}$$

mas,

$$\begin{aligned} f_{(X_1|T)}(y|t) &= \frac{f_{(X_1,T)}(y,t)}{f_T(t)} \\ f_{(X_1,T)}(y,t) &= f_{X_1}(y) f_{(T|X_1)}(t|y) \end{aligned}$$

Como

$$\left[\sum_{i=1}^n X_i \leq t | X_1 = x\right] = \left[\sum_{i=2}^n X_i \leq t - x\right]$$

e a soma de $(n-1)$ $\exp(\theta)$ é uma $Gama(n-1, \theta)$ temos que

$$\begin{aligned} f_{(T|X_1)}(t|x) &= \frac{\theta}{\Gamma(n-1)}[\theta(t-x)]^{n-2} e^{-\theta(t-x)} I_{[0,\infty)}(t-x) \\ f_{X_1}(x) &= \theta e^{-\theta x} I_{[0,\infty)}(x) \\ f_{(T,X_1)}(t,x) &= \frac{\theta^n}{\Gamma(n-1)}(t-x)^{n-2} e^{-\theta t} I_{[x,\infty)}(t) I_{[0,\infty)}(x) \\ f_T(t) &= \frac{\theta}{\Gamma(n)}(\theta t)^{n-1} e^{-\theta t} I_{[0,\infty)}(t) \end{aligned}$$

Portanto,

$$f_{(X_1|T)}(y|t) = \frac{n-1}{t}\left(1 - \frac{y}{t}\right)^{n-2} I_{[0,t]}(y)$$

Como

$$P(X_1 \leq x|T = t) = \int_0^x f_{(X_1|T}(u|t)du$$

e obtemos

$$T^* = \begin{cases} 1 - (1 - \frac{x}{\sum_{i=1}^n X_i})^{n-1}, & \text{se } \sum_{i=1}^n X_i \geq x \\ 1, & \text{caso contrário} \end{cases}$$

Esta estimativa difere muito da obtida pelo método de máxima verossimilhança onde o estimador de $\Gamma(\lambda)$ é $1 - e^{-\hat{\lambda}x}$ onde $\hat{\lambda} = (1/\bar{X})$. O EMV também é função da estatística suficiente e como não é igual ao ENVUMV T^*, isto significa que o EMV é viciado.

Exemplo 3.17: No exemplo anterior encontre o ENVUMV da taxa de falha, isto é, de θ.
Utilizando o **Método 4** vamos "chutar" que o estimador é da forma $\frac{c}{\sum X_i}$.
Como T tem distribuição $Gama(n, \theta)$ temos que

$$\begin{aligned} E[T^{-1}] &= \int_0^\infty t^{-1} \frac{1}{\Gamma(n)} \theta^n t^{n-1} e^{-\theta t} dt \\ &= \frac{1}{\Gamma(n)} \int_0^\infty \theta^n t^{n-2} e^{-\theta t} dt \end{aligned}$$

utilizando a trasformação $u = \theta t$ temos $du = \theta dt$ e

$$\begin{aligned} E[T^{-1}] &= \frac{\theta}{\Gamma(n)} \int_0^\infty u^{n-2} e^{-u} du \\ &= \frac{\theta}{\Gamma(n)} \Gamma(n-1) = \frac{\theta}{n-1}. \end{aligned}$$

logo $\frac{n-1}{\sum X_i}$ é o ENVUMV de θ.

Exemplo 3.18: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma $\mathrm{U}(0, \theta)$. Encontre o ENVUMV de θ.
Vimos no exemplo que o máximo amostral, $X_{(n)}$, que é o estimador de máxima verossimilhança de θ é uma estatística suficiente e completa. Vamos calcular $E[X_n]$ para ver se é possível aplicar o **Método 3**.

$$\begin{aligned} E[X_{(n)}] &= \int_0^\theta y \theta^{-n} n y^{n-1} dy \\ &= \frac{n}{\theta^n} \int_0^\theta y^n dy \\ &= \frac{n\theta}{n+1} \end{aligned}$$

ou seja, $X_{(n)}$ é um estimador viciado. Porém o seu vício pode ser corrigido e aplicado o método 3. Note que

$$T^* = \frac{n+1}{n} X_{(n)}$$

função somente da estatística suficiente e completa é não viciado e portanto o ENVUMV. Ele é único porque qualquer que seja θ sua variância é limitada.

Exercício 3.5: Mostre que o estimador "ridículo" encontrado no Exemplo 2.3 é o ENVUMV de $e^{-3\lambda}$.

Exercício 3.6: No Exemplo 18 encontramos o ENVUMV de θ na $U(0, \theta)$. Encontre o estimador $T_1 = aX_{(n)}$ que minimiza o erro quadrático médio. Compare o valor encontrado com os erros quadráticos médios do ENVUMV e do estimador de máxima verrossimilhança e mostre que os dois últimos estimadores são inadmissíveis.

Exercício 3.7: Assuma que lâmpadas à vácuo têm tempos de vida que são exponencialmente distribuidas com parâmetro λ, isto é, tempo médio de vida $1/\lambda$. Se nós tomamos uma amostra aleatória de n desses tubos denotando X_i = tempo de vida do i-ésimo tubo, $i = 1, 2, \ldots, n$, (a) Se nosso interesse é estimar a mediana do tempo de vida, isto é, quero o valor de c tal que $P(X > c) = P(X < c) = 0.5$. Encontre o EMV e o ENVUMV de c.

Exercício 3.8: O raio de um círculo é medido com um erro aleatório o qual tem distribuição $N(0, \sigma^2)$, σ desconhecido. Dadas n medidas independentes do raio $(R_1, \ldots, R_n)$, ache um estimador não viciado para a área do círculo. Esse estimador é o ENVUMV?

Exercício 3.9: Para uma amostra aleatória de tamanho n de uma população com distribuição de Poisson com parâmetro λ, encontre um estimador não viciado de $\tau(\lambda) = (1 + \lambda)e^{-\lambda}$. Encontre o EMV de $\tau(\lambda)$. Encontre o ENVUMV de $\tau(\lambda)$.

Exercício 3.10: Deseja-se estimar a proporção de moradores de Campinas que são a favor do programa de reciclagem de lixo. Para isso entrevista-se 500 pessoas e para cada pessoa anota-se se ela é contra ou a favor do programa. Com base nesta amostra aleatória:
(a) Qual seria o ENVUMV para a proporção de pessoas a favor? Chame este estimador de T_1.
(b) Qual o ENVUMV para a variância do estimador T_1 obtido em (a)? Chame o estimador da variância de T_2.
(c) Qual a cota inferior de Cramér-Rao para estimadores da variância de T_1? Esta cota é atingida por T_2?

Exercício 3.11 Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição $N(\theta, \theta)$.
(a) Qual o espaço paramétrico?
(b) Ache a estatística suficiente e completa.
(c) Argumente que $\bar{X}$ não é ENVUMV para θ.
(d) Como voce acharia o ENVUMV para θ. Se voce não conseguir achar explicitamente, pelo menos indique como este poderia ser encontrado.

Exercício 3.12: Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição com densidade

$$f(x, \theta) = e^{-(x-\theta)}\mathbf{I}_{[\theta,\infty)}(x)$$

para $-\infty < \theta < \infty$.
(a) Ache uma estatística suficiente para θ.
(b) A estatística obtida em (a) é completa?
(c) Ache o ENVUMV de θ se tal existir.

Exercício 3.13: Encontre o LICR dos estimadores não viciados da média populacional da Poisson, Binomial, Normal com variância conhecida. Mostre que em todos os casos as médias amostrais atingem o LICR. Como estes estimadores são não viciados eles são ENVUMV. Neste caso estamos adotando o **Método 5** para encontrar o ENVUMV.

4 Propriedades Ótimas dos Estimadores de Máxima Verossimilhança

Apesar da apresentação de vários métodos o estimador mais utilizado, na prática, é o de máxima verossimilhança. Esta ênfase será parcialmente justificada nesta seção ao considerarmos algumas propriedades ótimas dos estimadores de máxima verossimilhança.

Embora não seja necessário termos uma amostra aleatória para aplicarmos o método, apenas por simplicidade, vamos considerar que temos uma amostra aleatória de uma densidade $f(\cdot;\theta)$, onde θ é um número real. Também por simplicidade vamos considerar que estamos interessados em estimar o próprio parâmetro e denote o estimador de máxima verossimilhanca por $\hat{\Theta} = \theta(X_1, \ldots, X_n)$. Algumas das propriedades dos estimadores definidos anteriormente como não tendenciosidade e uniformemente de mínima variância são válidas para qualquer tamanho de amostra finita, e são chamadas de propriedades de pequenas amostras. Este nome pode trazer confusão já que as propriedades são válidas para qualquer tamanho de amostra fixa, pequena ou não. Este nome, na verdade, vem em contrapartida a algumas propriedades assintóticas como consistência e ótimo entre os assintoticamente normais, propriedades estas chamadas de propriedades para grandes amostras ou de propriedades assintóticas. Observe, no entanto, que estas mesmas propriedades podem ser válidas, ou aproximadamente válidas, mesmo para pequenas amostras. Porém ao contrário das propriedades para pequenas amostras, que são válidas para qualquer tamanho da amostra, as propriedades assintóticas podem não ser, nem aproximadamente, válidas para pequenas amostras.

Já vimos que os estimadores de máxima verossimilhança podem ser viciadas ou não, e que as não viciadas podem ser ou não ENVUMV. O teorema a seguir dá as propriedades assintóticas dos estimadores de máxima verossimilhança quando $f(\cdot;\theta)$ satisfaz certas condições de regularidade.

Teorema 4.1:. Se a densidade $f(\cdot;\theta)$ satisfaz certas condições de regularidade e se $\hat{\Theta} = \theta(X_1, \ldots, X_n)$ é o estimador de máxima verossimilhança de θ para uma amostra aleatória de tamanho n de $f(x;\theta)$, então:
(i) $\hat{\Theta}$ tem distribuição assintótica normal com média θ e variância igual ao LICR, isto é,

$$\left\{ nE_\theta\{[\frac{\partial}{\partial\theta} \log f(X;\theta)]^2\} \right\}^{-1}$$

(ii) A sequência de estimadores de máxima verossimilhança $\{\hat{\Theta}_n\}$ é o "melhor" entre os assintoticamente normais (BAN).

Nota: O teorema garante que:
1. A grosso modo, para grandes amostras, não existe estimador melhor do que o estimador de máxima verossimilhança, se utilizarmos o critério de vício e variância porque ele é assintoticamente não viciado e atinge o limite inferior de Cramér-Rao.

2. A distribuição assintótica depende apenas da densidade da população, isto é, não é necessário encontrar a forma analítica do estimador de máxima verossimilhança para saber sua distribuição. Esta é uma das grandes vantagens do EMV sobre o ENVUMV. Enquanto este último em muitos casos não é encontrado no caso do EMV para encontrar a estimativa de máxima verossimilhança basta utilizar um algoritmo numérico confiável que encontre o máximo de uma função. Tome cuidado porque o máximo pode nã existir. Nestes algoritmos normalmente é necessário utilizar um valor inicial. Muitas vezes a estimativa pelo método dos momentos é utilizado como valor inicial.

3. Uma das dificuldades poderia ser encontrar o LICR. No entanto, neste caso pode-se aproximar a esperança da segunda derivada pela valor da derivada no ponto da estimativa de máxima verossimilhança (a justificativa para se utilizar a segunda derivada dada pelo *Teorema 4.3*).

Exemplo 4.1: Em muitos casos o estimador de máxima verossimilhança da média populacional, como visto para o caso da Poisson, normal, exponencial, binomial , etc é a média amostral. Nestes casos já é possível verificar que a aplicação do teorema central do limite garante a distribuição assintótica normal e para qualquer tamanho amostral o EMV não é viciado e suas variâncias atingem o LICR. Portanto, para estes, e em muitos outros casos é fácil verificar que o *Teorema 4.1* se aplica.

Os resultados valem também para o caso de termos parâmetros n-dimensionais. Neste caso basta utilizar as derivadas em relação ao vetor $\theta = (\theta_1, \ldots, \theta_r)$.

Teorema 4.2: Sob certas condições de regularidade que englobam a existência de segundas derivadas da f.d.p. e a validade de se trocar as ordens de certas derivadas e integrais temos:

$$E_\theta\{[\frac{\partial}{\partial\theta}\log f(X;\theta)]^2\} = -E_\theta[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)]$$

Teorema 4.3: Vimos anteriormente que o estimador de máxima verossimilhança de $\tau(\theta)$ é dada por $\tau(\hat{\Theta})$. Se $\tau(\cdot)$ for diferenciável então temos que $\tau(\hat{\Theta})$ tem assintoticamente uma distribuição normal com média $\tau(\theta)$ e variância

$$\frac{[\tau^{'}(\hat{\theta})]^2}{nE_\theta\{[\frac{\partial}{\partial\theta}\log f(X;\theta)]^2\}}$$

que é o LICR.

Exemplo 4.2: Considere a estimativa da taxa de falha da distribuição exponencial que foi tratada no *Exemplo 3.17*. Caso você tenha feito os exercícios anteriores já saberia que o estimador de máxima verossimilhança da taxa de falha é dada pela estatística $T = n/\bar{X}$. Uma aplicação imediata dos resultados do *Exemplo 3.17* nos mostra que $E(T) = [n/(n-1)]\theta$ que tem vício assintótico igual a zero. Vamos calcular para tamanho da amostra igual a 20 e $\theta = 0,5$ a probabilidade de que o erro da estimativa seja menor do que 0,05.

No *Exemplo 3.9* já havíamos calculado através do teorema inicial que o denominador do LICR é igual a θ^{-2}. Vamos agora calculá-lo utilizando o *Teorema 4.2*.
Do *Exemplo 3.9* temos

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = 1/\theta - x$$

ou seja,

$$-\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{1}{\theta^2},$$

que é análogo ao resultado anterior.
Pelo *Teorema 4.1* temos que T tem distribuição assintótica com média igual a 0,5 e variância igual a $0,5^2/20 = 0,0125$, e

$$\begin{aligned} P[|T-\theta| < 0,05] &\approx P[|N(0,5;0,0125) - 0,5| < 0,05] = \\ &\approx P[|Z| < 0,447] = 0,345 \end{aligned}$$

Exercício 4.1 No caso anterior estime a probabilidade exata utilizando simulação e compare com o resultado encontrado.

Exercício 4.2 No caso da distribuição exponencial dê a distribuição aproximada do estimador de máxima verossimilhança de falha precoce quando os tempos são exponenciais independentes.

Exercício 4.3 Encontre a distribuição aproximada do estimador de máxima verossimilhança de $e^{-3\lambda}$ no caso da distribuição de Poisson.