Apostila - Inferência baseada na Verossimilhança

Professor: Wagner Hugo Bonat - LEG/UFPR

1 Introdução

Este texto se refere a inferência estatística baseada na função de Verossimilhança. Pretende ser um guia inicial a alunos de graduação em estatística. O texto apresenta os principais fundamentos da inferência, e alguns exemplos de aplicações. Todo o texto vem acompanhado de figuras e códigos computacionais desenvolvidos em linguagem **R** para facilitar o entendimento.

1.1 Algumas definições

- **Estatística** - a ciência que extrai informações de um conjunto de dados.
- **Amostra** - os dados à serem analisados.
- **População** - o universo do qual as amostras foram retiradas.
- **Inferência** - o processo de usar a amostra para dizer algo sobre a população.
- **Amostra aleatória** - amostra obtida por um mecanismo aleatório bem definido.Denota-se os membros da população de 1 a N, diz-se então que foi retirada uma amostra aleatória de tamanho n, escolhendo n de N possíveis valores de tal forma que todas as $\binom{N}{n}$ amostras possíveis que são igualmente prováveis.
- **Amostra pseudo-aleatória** - algum tipo de mecanismo que comporta-se como uma amostra aleatória.
- **Modelo** - uma representação matemática de um mecanismo que assume-se como gerador do conjunto de dados.
- **Modelo probabilístico** - um modelo em que o mecanismo gerador de dados incorpora um elemento aleatório.

- **Parâmetro** - uma quantidade aparente do modelo, cujos valores não podem ser mensurados diretamente mas que podem ser estimados através dos dados. Denota-se por θ.
- **Espaço paramétrico** - o conjunto Θ em que θ assume valores.
- **Verossimilhança** - um princípio feral que pode ser usado para estimar parâmetros eficientemente, e saber o quanto boa são as estimativas.

1.2 Exemplo motivacional

Considere o problema de inferir sobre o parâmetro θ de uma distribuição Binomial. Este problema aparece sempre que deseja-se estimar a proporção de individuos de uma dada população com alguma característica de interesse. Exemplos comuns são: proporção de pessoas que tem intenção de votar em um determinado canditado, proporção de peças defeituosas em um lote, proporção de individuos com um trauma dentário ou qualquer outro tipo de doenças, entre outros.

Para começar vamos sortear um determinado θ que será o **parâmetro** populacional, nesta situação hipotética conhecido. Sabemos antes de qualquer coisa que o parâmetro θ da Binomial só pode tomar valores no intervalo 0 a 1 o que configura o seu **espaço paramétrico**. Isso equivale a simular um valor de uma distribuição Uniforme $U(a, b)$ com $a = 0$ e $b = 1$.

```
> set.seed(100)
> theta <- runif(1, min = 0, max = 1)
```

Uma vez conhecido o **parâmetro** da distribuição, podemos simular amostras aleatórias desta distribuição. Para melhor explorar as idéias vamos tomar amostras de diferentes tamanhos. Para simular de uma distribuição Binomial, usamos os comandos abaixo:

```
> amostra10 <- rbinom(10, size = 1, prob = theta)
> amostra30 <- rbinom(30, size = 1, prob = theta)
> amostra50 <- rbinom(50, size = 1, prob = theta)
> amostra100 <- rbinom(100, size = 1, prob = theta)
> amostra1000 <- rbinom(1000, size = 1, prob = theta)
```

O problema geral da inferência é encontrar um ou mais valores plausíveis para θ baseado na amostra observada. O passo inicial para isto é lembrar do conceito de distribuição conjunta. No caso $B(1, \theta)$ que estamos considerando é equivalente a n ensaios de Bernoulli, sendo n o tamanho da amostra. Temos que a distribuição conjunta fica dado por:

$$p(\underline{x}) = \prod_{i=1}^{n} \binom{n}{x_i} \theta^{x_i} (1-\theta)^{1-x_i}, \quad x = 0, 1 \quad 0 \leq \theta \leq 1 \tag{1}$$

Podemos escrever uma função em R para calcular os valores desta distribuição conjunta.

```
> conjunta <- function(amostra, theta) {
+     saida = prod(theta^amostra * (1 - theta)^(1 -
+         amostra))
+     return(saida)
+ }
```

Esta mesma função pode ser escrita em logaritmo neperiano, isto é muito importante para evitar problemas numéricos e em muitos casos é a única forma de conseguir avaliar a função em todo o espaço paramétrico. Passando o logaritmo na expressão 1 temos,

$$\log(p(\underline{x})) = \sum_{i=1}^{n} x_i \log(\theta) + \sum_{i=1}^{n} (1 - x_i) \log(1 - \theta) \tag{2}$$

Novamente escrevemos uma função R para avaliar a distribuição conjunta em logaritmo.

```
> log.conjunta <- function(amostra, theta) {
+     saida <- sum(amostra * log(theta)) + sum((1 -
+         amostra) * log(1 - theta))
+     return(saida)
+ }
```

Para verificar que as funções foram escritas corretamente

```
> conjunta(amostra10, theta = 0.5)
```

```
[1] 0.0009765625
```

```
> log.conjunta(amostra10, theta = 0.5)

[1] -6.931472

> log(conjunta(amostra10, theta = 0.5))

[1] -6.931472

> exp(log.conjunta(amostra10, theta = 0.5))

[1] 0.0009765625
```

Podemos interpretar estes valores como a plausibilidade ou possibilidade de um determinado θ que neste caso foi usado arbitrariamente $\theta = 0.5$ ter gerado a amostra observada. Observe que apartir deste momento a amostra é fixa uma vez que já foi observada, e que neste caso podemos avaliar a plausibilidade de diversos valores de θ a fim de encontrar o mais plausível de ter gerado a amostra observada.

Note que a distribuição conjunta é uma função da amostra, porém a estamos olhando em função de θ e neste caso a nova função ganha o nome de **função de verossimilhança**. Encontrar o valor que maximiza a função de verossimilhança é o objetivo fundamental da inferência baseada em verossimilhança, ou seja, o objetivo é encontrar o valor de θ mais plausível de ter gerado a amostra observada. O gráfico da figura 1 mostra a função de verossimilhança para o nosso exemplo.

Para isto primeiro vamos escrever a função de verossimilhança para o modelo de estamos propondo:

```
> verossimilhanca <- function(theta, amostra) {
+     vero <- prod(dbinom(amostra, size = 1, prob = theta))
+     return(vero)
+ }
```

Uma vez feita a função podemos avaliá-la em todo o espaço paramétrico e verificar através de uma gráfico qual o valor de θ que trna a função máxima. Vamos executar este procedimento para os diferentes tamanhos de amostras considerados, para verificar qual o impacto que este tem sobre a função de verossimilhança.

```
> grid.theta <- as.matrix(seq(0.001, 0.999, l = 1000))
> vero10 <- apply(grid.theta, 1, verossimilhanca, amostra = amostra10)
> vero30 <- apply(grid.theta, 1, verossimilhanca, amostra = amostra30)
> vero50 <- apply(grid.theta, 1, verossimilhanca, amostra = amostra50)
> vero100 <- apply(grid.theta, 1, verossimilhanca,
+     amostra = amostra100)
> vero1000 <- apply(grid.theta, 1, verossimilhanca,
+     amostra = amostra1000)
```

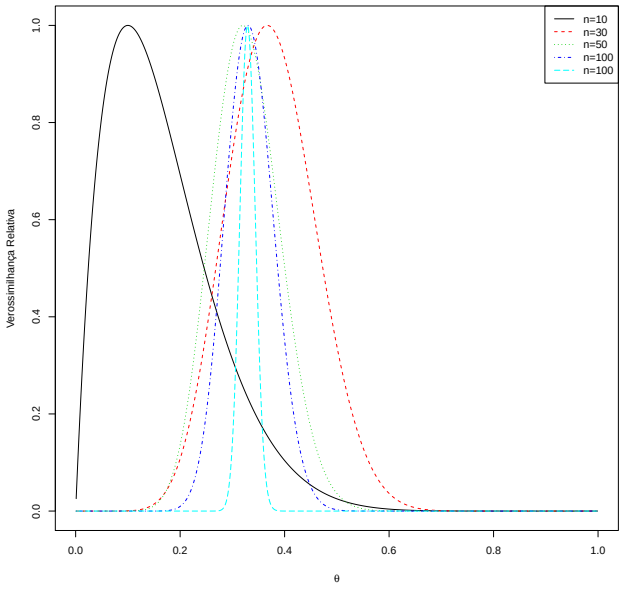

Figura 1: Função de verossimilhança por tamanho de amostra.

A figura 1 mostra que diferentes amostras levam a diferentes localizações do ponto de máximo, e também a diferentes graus de curvatura ao redor deste ponto. No decorrer do curso será visto como usar estas informações para estimar o valor de θ, obter intervalos de confiança e fazer teste de hipóteses sobre os parâmetros sendo estimados.

2 Revisão dos conceitos básicos de probabilidade

2.1 Distribuições discretas

Função de probabilidade denotada por $p(x) = P(X = x)$

- Distribuição Binomial $X \sim B(n, \theta)$

$$p(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n \quad \text{e} \quad 0 \leq \theta \leq 1$$

- Distribuição Poisson $X \sim P(\theta)$

$$p(x) = \frac{\exp^{-\theta}\theta^x}{x!} \quad x = 0, 1, \ldots, n \quad \text{e} \quad \theta \geq 0$$

.

2.2 Distribuições contínuas

Função distribuição de probabilidade $F(x) = P(X \leq x)$.
Função densidade de probabilidade $f(x) = F'(x) = P(a \leq X \leq b) = \int_a^b f(x)dx$.

- Distribuição Normal $X \sim N(\mu, \sigma^2)$

$$f(x) = (2\pi\sigma^2)^-\frac{1}{2}\exp^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x, \mu < \infty, \quad \text{e} \quad \sigma \geq 0$$

.

- Distribuição Gama $X \sim \Gamma(k, \theta)$

$$f(x) = \frac{1}{\Gamma(k)}\theta^k x^{k-1}\exp^{-\theta x}, \quad x \geq 0 \quad \text{e} \quad k, \theta > 0$$

.

- Distribuição Exponencial $X \sim \Gamma(1, \theta)$

$$f(x) = \theta\exp^{-\theta x}, \quad x > 0 \quad \text{e} \quad \theta > 0$$

.

- Distribuição Uniforme $X \sim U(a, b)$

$$f(x) = \frac{1}{b-1}, \quad a \leq x \leq b$$

.

2.3 Esperança

- Caso discreto - $E[h(X)] = \sum h(x)p(x)$.
- Caso contínuo - $E[h(X)] = \int h(x)f(x)dx$.
- $E[X] = \mu$, $V[X] = E[(X - \mu_x)^2] = \sigma^2$ e $SD[X] = \sqrt{V[X]} = \sigma$.

2.4 Variáveis aleatórias bivariadas

- Função de probabilidade conjunta $p(x, y) = P(X = x, Y = y)$.
- Função de densidade de probabilidade conjunta

$$P((x, y) \in A) = \int_A f(x, y)dxdy$$

.

- Covariância $COV[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$.
- Correlação $\rho = \frac{COV[X,Y]}{\sqrt{V[X]V[Y]}}$.
- Independência - (X, Y) são independentes $\iff p(x, y) = p_x(x)p_y(y), \forall(x, y)$.

2.5 Teorema Central do Limite

Seja $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, onde X_i são independentes com $E[X_i] = \mu$ e $V[X_i] = \sigma^2$ e $Z_n = \frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}$, então $E[\bar{x}] = \mu$ e $V[\bar{x}] = \frac{\sigma^2}{n}$ e $Z_n \sim N(0, 1)$ quando $n \to \infty$.

Uma outra forma deste teorema é

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

Este teorema nos diz que podemos usar a distribuição Normal para aproximar a distribuição de $\bar{x}$, independente de qual seja a distribuição de $X - i$. Como será visto no decorrer do curso, este resultado servirá para provar muitos dos resultados mais poderosos sobre estimação de parâmetros em modelos estatísticos.

3 Estimação de parâmetros

Considere n variáveis aleatórias independentes X_i com função densidade ou probabilidade dada por

$$p(x;\theta) = P(X_i = x)$$

onde θ é um parâmetro.

Os dados são um conjunto de vaores realizados x_i, ou seja, x_i é uma amostra aleatória proveniente da distribuição com fp ou fdp $p(x;\theta)$. Denota-se isto,

$$\underline{X} = (X_1, X_2, \ldots, X_n), \quad \underline{x} = (x_1, x_2, \ldots, x_n)$$

.

Definição 1 (Estatística) *Uma **estatística** é uma variável aleatória $T = t(\underline{X})$, onde a função $t(.)$ não depende de θ.*

Definição 2 (Estimador) *Uma estatística T é um **estimador** para θ se o valor realizado $t = t(\underline{x})$ é usado como uma estimativa para o valor de θ.*

Definição 3 (Distribuição amostral) *A distribuição de probabilidade de T é chamada de **distribuição amostral** do estimador $t(\underline{X})$.*

Exemplo 1 *Seja $p(x;\theta) = \frac{\exp^{-\theta}\theta^x}{x!}, \quad x = 0, 1, 2, \ldots$. Considere as seguintes estatísticas*

1. $t_1(\underline{X}) = n^{-1}\sum_{i=1}^n X_i$
2. $t_2(\underline{X}) = X_1$
3. $t_3(\underline{X}) = \sum_{i=1}^n w_i X_i$, *para w_i números reais pre-especificados*
4. $t_4(\underline{X}) = n^{-1}\sum_{i=1}^n X_i^2$

Quais são bons estimadores para θ ?

Definição 4 (Viés) *O **viés** de um estimador T é a quantidade $B(T) = E(T-\theta)$. O estimador T é dito não-viciado para θ se $B(T) = 0$, tal que $E(T) = \theta$. O estimador T é assintóticamente não-viciado se $E(T) \to \theta$ quando $n \to \infty$.*

Definição 5 (Eficiência relativa) *A **eficiência relativa** de dois estimadores não-viciados T_1 e T_2 é a razão $re = \frac{V(T_2)}{V(T_1)}$.*

Definição 6 (Erro quadrático médio) *O **erro quadrático médio** de um estimador T é a quantidade*

$$EQM(T) = E[(T-\theta)^2] = V(T) + B(T)^2$$

Definição 7 (Consistência) *Um estimador T é **médio quadrático consistente** para θ se o $EQM(T) \to 0$ quando $n \to \infty$. O estimador T é **consistente em probabilidade** se $\forall \epsilon > 0, P(|T-\theta|) > \epsilon) \to 0, \quad quando \quad n \to \infty$.*

- Fracamente falando, se T é um estimador consistente, então $T \to \theta$ quando $n \to \infty$.
- Uma necessidade minima para um bom estimador é que $E(T) \to \theta$ e $V(T) \to 0$ quando $n \to \infty$. Ou seja, que o estimador seja médio quadrático consistente. Máxima eficiência, é atingida quando a $V(T)$ é tão pequena quanto possível, sujeita a condições de não-viés assintótico.
- O ideal é poder encontrar um estimador assintóticamente não-viciado que tenha a menor variância possível.

Exemplo 2 *Seja $X \sim B(1,\theta)$, proponha um estimador para θ, este estimador é não-viciado ?*

3.1 O método dos momentos

Para uma amostra aleatóri X cuja distribuição é indexada por apenas um simples parâmetro θ, a **esperança** de X vai usualmente ser uma função de θ. Escreva isto como

$$E(X) = \mu(\theta)$$

Para uma amostra de n independentes realizações de X, digamos $x_i : i = 1, 2, \ldots, n$ a **média amostral** é

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$$

Se nos não conhecemos θ, nos podemos usar $\bar{x}$ como guia para **estimar** um de μ. Isto também sugere que podemos estimar θ pelo valor $\hat{\theta}$, que é a solução da equação $\bar{x} = \mu(\hat{\theta})$. Isto é o **método dos momentos**.

Se existem dois parâmetros, θ_1 e θ_2, então é usual que

$$E(X) = \mu(\theta_1, \theta_2)$$

e

$$V(X) = \sigma^2(\theta_1, \theta_2)$$

Se nos definirmos a variância amostral, $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$, o método dos momentos estima θ_1 e θ_2 pelos valores $\hat{\theta}_1$ e $\hat{\theta}_2$, que resolve o seguinte sistema de equações,

$$\bar{x} = \mu(\hat{\theta}_1, \hat{\theta}_2)$$

$$s^2 = \sigma^2(\hat{\theta}_1, \hat{\theta}_2)$$

Exemplo 3 *O seguinte conjunto de dados é uma amostra de* 10 *independentes realizações de uma variável aleatória exponencial com função densidade de probabilidade dada por*

$$f(x; \theta) = \theta \exp^{-\theta x} : \quad x \geq 0$$

Estime θ *com os seguintes dados:*
$x_i = 1,2; 0,5; 3,1; 0,8; 1,5; 5,3; 1,6; 0,2; 1,6.$

Exemplo 4 *Se* $X \sim \Gamma(k, \theta)$ *tal que sua função densidade de probabilidade é*

$$f(x; k, \theta) = \Gamma(k, \theta)^{-1} \theta^k x^{k-1} \exp^{-\theta x} : \quad x \geq 0$$

, mostre que $E(X) = \frac{k}{\theta}$ *e* $V(X) = \frac{k}{\theta^2}$.

Exemplo 5 *Usando os mesmos dados do exemplo anterior, se assumimos o modelo* $X \sim \Gamma(k, \theta)$ *com* k *e* θ *desconhecidos, encontre os estimadores dos momentos* $\hat{k}$ *e* $\hat{\theta}$.

3.2 Estimação intervalar

Dado um estimador $T = t(\underline{X})$ para um parâmetro θ, o valor observado $t(\underline{x})$ é chamado de **estimativa pontual** de θ. Como T é uma variável aleatória, o valor realizado $t(\underline{x})$ muito provavelmente não será o verdadeiro valor de θ, sendo assim, qual é o real valor de usar apenas uma estimativa pontual ?

Uma melhor estratégia é construir um intervalo, digamos $t_1(\underline{x}), t_2(\underline{x})$, que apresente uma certa confiança de conter o verdadeiro valor de θ.

Definição 8 *Um intervalo de confiança* $100(1-p)\%$ *para* θ *é uma realização de um intervalo aleatório* $t_1(\underline{X}), t_2(\underline{X})$ *com a propriedade que, qualquer que seja o valor atual de* θ*,*

$$P(t_1(\underline{X}) \leq \theta \leq t_2(\underline{X})) = 1 - p$$

Comentários

1. Um bom intervalo é aquele tão pequeno quanto possível mantendo o mesmo valor de p próximo de zero.

2. A parte aleatória da definição está em $\underline{X}$, não em θ. É sem sentido falar sobre a probabilidae que o intervalo atual $t_1(\underline{x}), t_2(\underline{x})$ contém θ.

3. Convencionalmente, as pessoas tendem a usar $p = 0,1$, $p = 0,05$ ou $p = 0,01$, para dar intervalos com $90\%, 95\%$ e 99% de confiança, respectivamente.

Exemplo 6 *Suponha que temos uma amostra aleatória* $X_i \sim N(\theta, 1)$*, tal que*

$$f(x;\theta) = (2\pi)^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x-\theta)^2}$$

e que os dados são $n = 5$*;* $x_i = 2,8; 1,5; 3,0; 2,3; 3,2$*; sendo que* $\sum_{i=1}^{5} x_i = 12,8$*. Proponha um estimador para* θ *e construa um intervalo de confiança.*

3.3 Quantidade pivotal

O ponto chave do exemplo anterior é que a variável aleatória Z é uma função de T e θ cuja distribuição de probabilidade não depende de θ. Uma variável aleatória deste tipo é chamada

de **pivotal**. Uma vez encontrada a quantidade pivotal, e sua distribuição de probabilidade, podemos trabalhar como os números reais a e b tal que

$$P(a \leq Z \leq b) = 1 - p$$

onde p é escolhido para um dado nível de confiança. Então, manipulações algébricas convertem o par de desigualdades

$$a \leq Z$$

$$Z \leq b$$

para o par de desigualdades

$$T_1 \leq \theta$$

$$\theta \leq T_2$$

Finalmente, plugando os dados observados para obter os valores realizados de t_1 e t_2 no lugar de T_1 e T_2 da o intervalo de confiança. Note que neste processo, o par (a, b) não é único.

Exemplo 7 *Seja* $f(x;\theta) = \theta^{-1} : \quad 0 \leq x \leq \theta$. *Encontre a distribuição de* $T = max(X_i)$. *Assim, entre uma quantidade pivotal,* Z, *e calcule a função distribuição de* Z. *Derive um método para obter uma intervalo de confiança para* θ, *e aplique para o seguinte conjunto de dados:*$x_i = 0,7; 1,3; 1,1; 0,4; 0,8$.

O método da quantidade pivotal é um tanto *ad hoc* no sentido de que ele não te diz como encontrar uma quantidade pivotal, você tem que usar a sua intuição para encontrar Z para um dado problema. Na próxima seção nos vamos introduzir o conceito de **verossimilhança** e mostrar uma abordagem unificada para estimação pontual e intervalar, com excelentes propriedades.

4 Verossimilhança

Nesta seção nos vamos apresentar um método geral de inferência que vai nos trazer bons estimadores pontuais e intervalares para uma grande dimensão de problemas práticos. O métodos dos momentos faz isto ?

Definição 9 (Verossimilhança) *Suponha que os dados $\underline{x}$ são uma realização de um vetor aleatório $\underline{X}$ com fp ou fdp $p(\underline{x};\theta)$. A* ***verossimilhança*** *para θ. dado dos valores observados $\underline{x}$, é a função $L(\theta) = p(\underline{x};\theta)$.*

Comentários

- Pense em $L(\theta)$ com uma medida do quanto compatível os dados $\underline{x}$ são com o valor de θ.
- Se X_i são independentes com fdp ou fp comum, a $L(\theta) = p(x_1, \theta) \ldots, p(x_2, \theta)$.
- $L(\theta)$ pode ser pequena para todos os possíveis valores de θ. Porém o que interessa para inferência são os valores **relativos** de $L(\theta)$ para diferentes valores de θ.

Exemplo 8 *Seja $X \sim B(N, \theta)$, considere $n = 10$ e $x = 8$, escreva a função de verossimilhança e desenhe seu gráfico.*

O primeiro passo é escrever a função de verossimilhança, que neste caso é muito simples

$$L(\theta) = \binom{10}{8} \theta^8 (1-\theta)^{10-8}$$

Podemos escrever esta função no R seguindo o código abaixo

```
> vero.binomial <- function(theta, x, n) {
+     ll = choose(n, x) * (theta^x) * (1 - theta)^(n -
+         x)
+     return(ll)
+ }
```

Com base na função é trivial desenhar o gráfico, como mostrado abaixo.

Exemplo 9 *Seja $X_i \sim \Gamma(1, \theta)$ independentes, escreva a função de verossimilhança e desenhe seu gráfico. Considere o seguinte conjunto de dados,*
$x_i = 1,5; 0,3; 0,6; 1,8; 4,0; 1,5; 0,7; 1,1; 4,1; 1,7$.

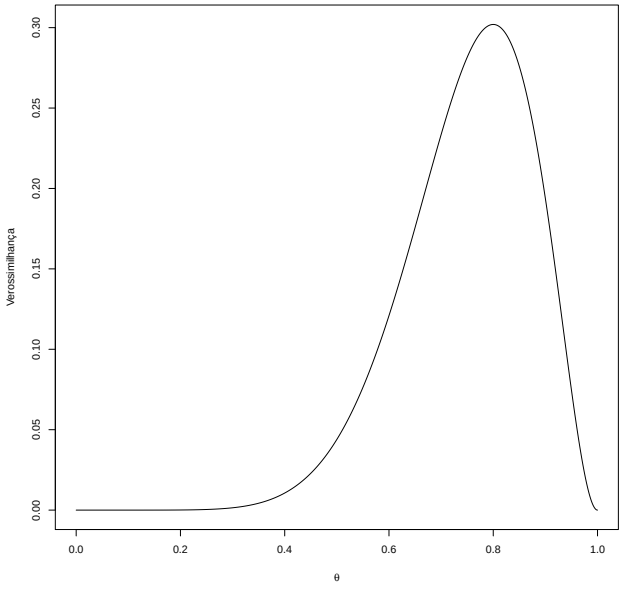

Figura 2: Função de verossimilhança - Binomial $n = 10$ e $x = 8$.

Começamos escrevendo a fdo de uma observação

$$f(x;\theta) = \theta \exp^{-\theta x} : \quad x \geq 0$$

Como são independentes a distribuição conjunta de $\underline{X}$ é simplesmente o produto,

$$L(\theta) = \prod_{i=1}^{n} \theta \exp^{-\theta x_i}$$

Trabalhando um pouco esta equação chegamos a

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

Escrevendo esta função no R temos

```
> vero.exponencial <- function(theta, x) {
+     n <- length(x)
+     ll = theta^n * exp(-theta * sum(x))
+     return(ll)
+ }
```

Entrando com os dados,

```
> x <- c(1.5, 0.3, 0.6, 1.8, 4, 1.5, 0.7, 1.1, 4.1,
+     1.7)
```

Com a função de verossimilhança e os dados podemos facilmente plotar o gráfico como mostra o código abaixo.

```
> grid.theta <- as.matrix(seq(0, 2, l = 1000))
> vero <- apply(grid.theta, 1, vero.exponencial, x = x)
> plot(vero ~ grid.theta, type = "l", ylab = "Verossimilhança",
+     xlab = expression(theta))
```

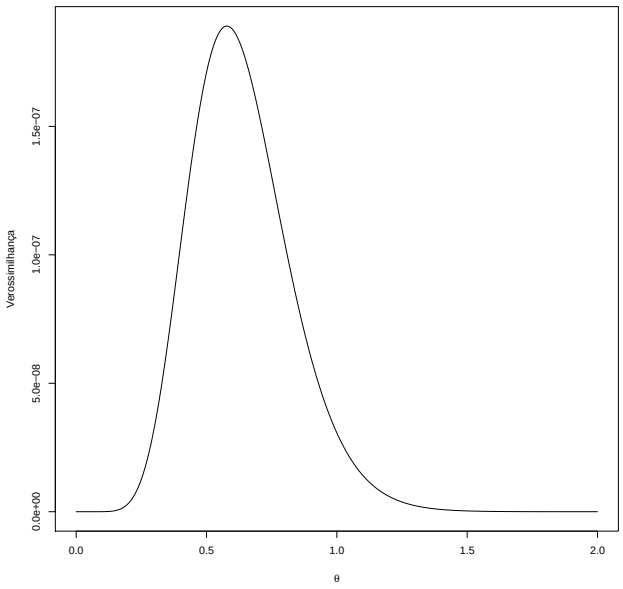

Figura 3: Função de verossimilhança - Exponencial.

Exemplo 10 *Seja* $X_i \sim U(0, \theta)$: $0 \leq x \leq \theta$. *Escreva a função de verossimilhança e desenhe seu gráfico, considere o seguinte conjunto de dados.*
$x_i = 7,1; 2,6; 0,5; 1,3; 9,2; 3,0; 0,7; 1,1; 5,0; 1,3.$

A verossimilhança é dada por

$$l(\theta) = \theta^{-n} \quad : \theta > max(x_i) \quad \text{e} \quad 0 \quad \text{cc}$$

Escrevendo esta função em R

```
> vero.uniforme <- function(theta, x) {
+     ll <- ifelse(theta > max(x), prod(dunif(x, min = 0,
+         max = theta)), 0)
+     return(ll)
+ }
```

Entrando com os dados e plotando o gráfico,

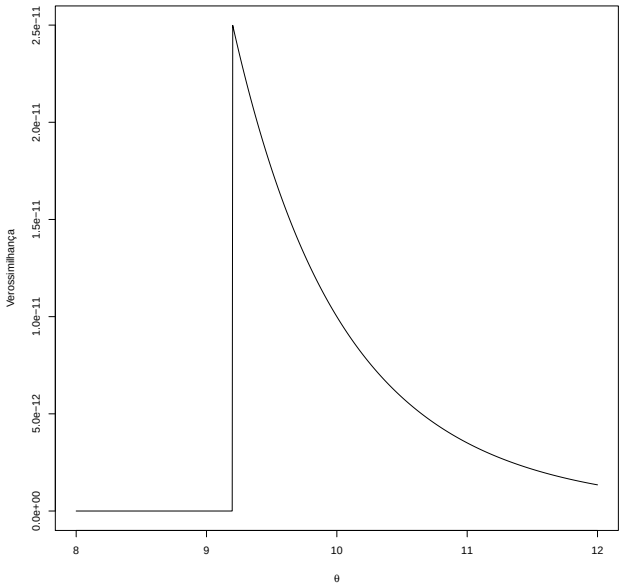

Figura 4: Função de verossimilhança - Uniforme.

4.1 Estimação pontual usando a função de verossimilhança

A função de verossimilhança traz uma abordagem unificada para estimação pontual e intervalar, e pode ser mostrado que têm boas propriedades teóricas em uma grande variedade de situações práticas.

Definição 10 *Seja $L(\theta)$ a função de verossimilhança. O valor $\hat{\theta} = \hat{\theta}(\underline{x})$ é a estimativa de máxima verossimilhança para θ se $L(\hat{\theta}) \geq L(\theta), \forall\theta$.*

Definição 11 *Se $\hat{\theta}(\underline{x})$ é a estimativa de máxima verossimilhança, então $\hat{\theta}(\underline{X})$ é o estimador de máxima verossimilhança.*

Comentários

- $\hat{\theta}(\underline{x})$ é um número real, uma particular função dos dados, $\underline{x}$.
- $\hat{\theta}(\underline{X})$ é uma variável aleatória.
- Nos vamos usar a abreviação $\hat{\theta}$ para ambos, o contexto vai indicar o real sentido de $\hat{\theta}$.

Definição 12 *Se $L(\theta)$ é a função de verossimilhança, então $l(\theta) = \log(L(\theta))$ é chamada de função de log-verossimilhança.*

Teorema 1 *Uma estimativa de máxima verossimilhança também maximiza $l(\theta)$.*

Prova: Segue do fato da função log(.) ser monotona crescente.
Comentário: Apesar de matematicamente trivial, este teorema é muito importante operacionalmente porque $l(\theta)$ é usualmente uma função mais fácil de maximizar que $L(\theta)$, e porque poderosos teoremas sobre as propriedades de $\hat{\theta}(\underline{X})$ envolvem $l(\theta)$ ao invés de $L(\theta)$.

Exemplo 11 $X \sim B(n, \theta)$. *Encontre o estimador de máxima verossimilhança para θ.*

Exemplo 12 $X_i \sim \Gamma(1, \theta)$. *Encontre o estimador de máxima verossimilhança para θ.*

Exemplo 13 $X_i \sim U(0, \theta)$. *Encontre o estimador de máxima verossimilhança para θ.*

Ao aplicar o método de máxima verossimilhança, você pode usar as ferramentas rotineiras de cálculo para maximizar a log-verossimilhança sempre que $L(\theta)$ for uma função diferenciável em θ. Uma situação geral em que a log-verossimilhança não é diferenciável é quando o parâmetro θ define a dimensão da distribuição.

Agora vamos olhar uma outra idéia geral na estimação de parâmetros, chamada **suficiência**. Fracamente falando, a idéia por traz de suficiência é que em muitas aplicações, **toda** a informação sobre θ em uma amostra de n observações x_i pode ser resumida por uma única estatística $t(\underline{x})$.

Definição 13 (Suficiência) *Seja* $\underline{X} = (X_1, \ldots, X_n)$ *tem distribuição conjunta* $p(\underline{x}; \theta)$ *e seja* $S = S(\underline{X})$ *uma estatística. Então,* S *é* ***suficiente*** *para* θ *se existe uma função* $g(.)$ *e* $h(.)$ *tal que*

$$p(\underline{x}; \theta) = g(\underline{x})hs(\underline{x}); \theta$$

ou equivalentemente

$$\log p(\underline{x}; \theta) = G(\underline{x}) + H(\underline{x}; \theta)$$

onde $G(.) = \log g(.)$ *e* $H(.) = \log h(.)$.

Teorema 2 *Se* S *é suficiente para* θ*, então* $\hat{\theta}(\underline{X}) = \hat{\theta}(S)$.

Comentários

- Estatísticas suficientes nem sempre existem, mas quando existem tornam a vida mais fácil.
- Em um exemplo, você pode usar a definição ou o teorema para encontrar uma estatśitica suficiente, ou para mostrar que uma dada estatística é suficiente.

Exemplo 14 *Seja* $X_i \sim N(\theta, 1)$*. Encontre uma estatística suficiente e o estimador de máxima verossimilhança para* θ*.*

Exemplo 15 $X_i \sim U(-\theta, \theta)$*. Encontre uma estatística suficiente para* θ*.*

4.2 Estimação intervalar usando a função de verossimilhança

Definição 14 *Um* ***intervalo de verossimilhança*** *para* θ *é um intervalo da forma* $\theta : l(\theta) \geq rl(\hat{\theta})$ *ou equivalentemente,* $\theta : D(\theta) \leq c$*, onde* $D(\theta) = 2(l(\hat{\theta}) - l(\theta))$ *e* $c = -\log r$.

Comentários

- O valor de r precisa estar no intervalo 0 a 1 para intervalos não-vazios, assim $c > 0$.
- Grandes valores de c implicam em intervalos largos.
- Algumas vezes (mas não sempre), o intervalo é a união de sub-intervalos disjuntos.

Exemplo 16 *Seja $X_i \sim P(\theta)$. Encontre uma estatística suficiente e o estimador de máxima verossimilhança para θ. Para $n = 10, 25, 100$, e faça um gráfico da Deviance. Discuta como você pode usar este gráfico para encontrar um intervalo de confiança baseado na Deviance.*

Primeira vamos lembrar que $p(x;\theta) = \frac{\exp^{-\theta}\theta^x}{x!}: \quad x = 0, 1, \ldots$. Assim a função de verossimilhança é dada por

$$L(\theta) = \prod_{i=1}^{n} \frac{\exp^{-\theta}\theta^{x_i}}{x_i!}$$

Para facilitar a derivação tomamos a log-verossimilhança,

$$l(\theta) = -n\theta + sum_{i=1}^{n} x_i \log\theta - \sum_{i=1}^{n} \log x_i!$$

Com isso é fácil ver que $S = \sum_{i=1}^{n} X_i$ é suficiente para θ. Também

$$l'(\theta) = -n + \frac{\sum_{i=1}^{n} x_i}{\theta}$$

O que nos leva ao estimador de Máxima Verossimilhança

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Observando a segunda derivada

$$l''(\theta) = -\frac{\sum_{i=1}^{n} x_i}{\theta^2}, \quad < 0 \quad \forall\theta$$

Mostrando que o ponto é realmente de máximo. Para a construção do intervalo baseado na Deviance, temos que a deviance é dada por

$$D(\theta) = 2n(\theta + \bar{x}(\log\bar{x} - \log\theta - 1))$$

Vamos fazer o gráfico desta função para diferentes tamanhos de amostras provenientes de uma $P(\theta = 2)$. O primeiro passo é construir uma função R com a deviance do modelo.

```
> deviance.poisson <- function(theta, media, n) {
+     dv = 2 * n * (theta + media * (log(media) - log(theta) -
+         1))
```

```
+     return(dv)
+ }
```

Segundo passo, vamos avaliar a deviance em uma grade de valores de θ. Vamos fixar $\bar{x} = 2$.

```
> grid.theta <- as.matrix(seq(1.5, 2.5, l = 1000))
> dev10 <- apply(grid.theta, 1, deviance.poisson, media = 2,
+     n = 10)
> dev25 <- apply(grid.theta, 1, deviance.poisson, media = 2,
+     n = 25)
> dev100 <- apply(grid.theta, 1, deviance.poisson,
+     media = 2, n = 100)
```

Terceiro passo, desenhamos o gráfico.

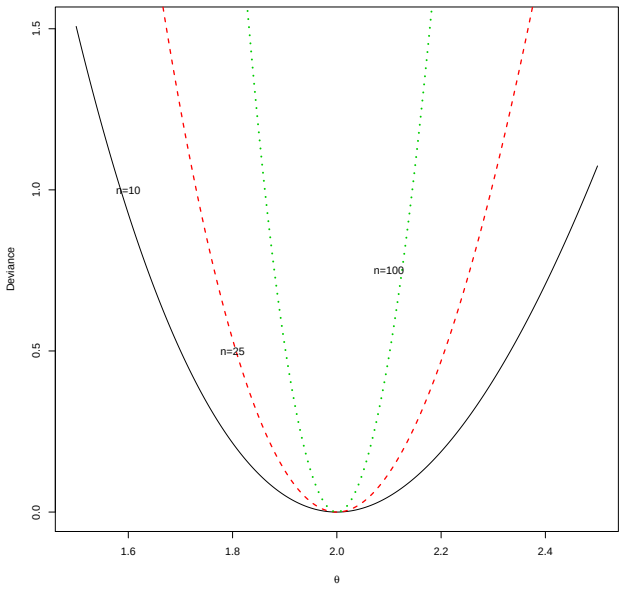

Figura 5: Função Deviance - Poisson.

Note que se você quiser construir um intervalo baseado na Deviance para uma particular aplicação, você não precisa trabalhar algébricamente para encontrar os limites. Basta desenhar o gráfico de $D(\theta)$ contra θ e traçar uma linha horizontal em uma certa altura c e irá obter os limites desejados. Porém qual c escolher é ainda uma pergunta a ser respondida.

Antes de considerar esta pergunta, vamos apresentar o conceito de invariância do estimador de máxima verossimilhança. Esta é uma importante propriedade deste tipo de estimador. Em estatística aplicada, nos sempre temos a escolha de trabalhar em diferentes parametrizações de um mesmo modelo. Por exemplo, podemos trabalhar com $N(0, \sigma^2)$ ou $N(0, \phi)$. É interessante que possamos estimar os parâmetros em uma parametrização adequada, porém podemos ter interesse em uma outra parametrização, o princípio da invariância torna possível, estimar parâmetros em uma determinada parametrização e passar para outras que possam ser de interesse.

Teorema 3 *Estimativas e intervalos baseados na verossimilhança são invariantes a reparametrizações.*

Exemplo 17 *Considere o seguinte conjunto de dados:* $x_i = -0,08; -0,16; -0,03; -0,07; -0,21; 0,13; -0,01; 0,16; 0,04; 0,24.$
Se $X \sim N(0, \sigma^2)$ *tal que* σ *representa o desvio padrão de* X*, então*

$$f(x, \sigma) = (2\pi)^{-0.5} \exp^{-0.5x^2/\sigma^2}$$

e a log-verossimilhança é

$$l(\theta) = const - \frac{\sum_{i=1}^{n} x_i^2}{\sigma^2}$$

O gráfico da log-verossimilhança é mostrado na figura 6, usando $c = 4$*, o intervalo para* σ *é* $0,093 \leq \theta \leq 0,22$.
Se $X \sim N(0, \phi)$*, tal que* ϕ *representa a variância de* X*, então*

$$f(x; \phi) = (2\pi)^{0.5} \exp^{-0.5x^2/\phi}$$

e a log-verossimilhança é

$$l(\phi) = const - \frac{\sum_{i=1}^{n} x_i^2}{\phi}$$

O gráfico da log.verossimilhança é mostrado na figura 6, usando $c = 4$*, o intervalo para* ϕ *é* $0,0087 \leq \phi \leq 0,0492$

Relembre que $\phi = \theta^2$. O que acontece se você tirar a raiz dos pontos finais do intervalo de verossimilhança de ϕ ?

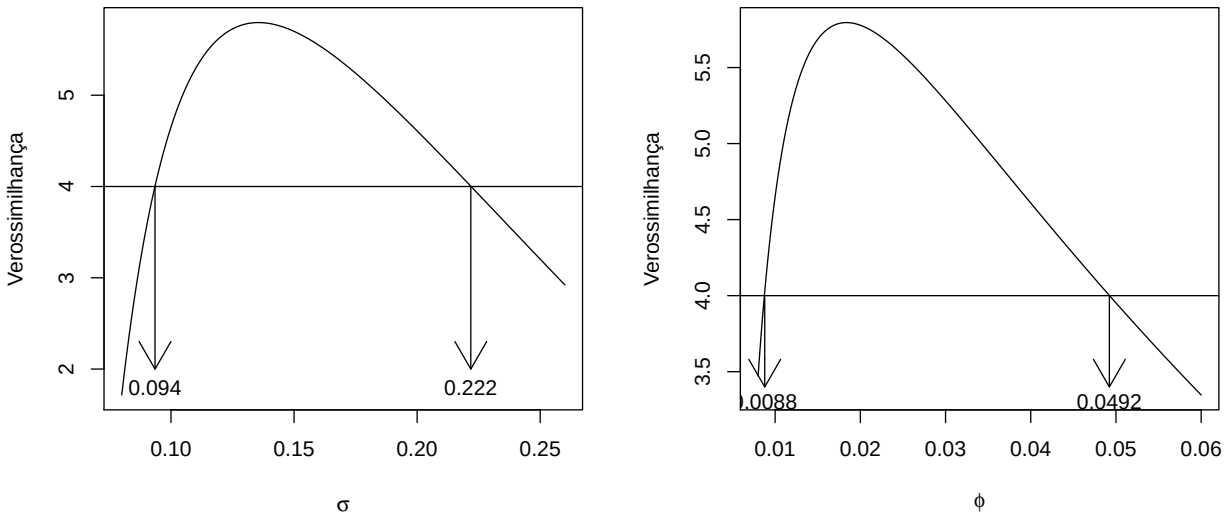

Figura 6: Invariância caso Normal.

4.3 Estimador não-viciado de variância mínima

Relembre que $T = t(\underline{X})$ denota um estimador para θ, o parâmetro da distribuição de $\underline{X}$, e que T é não-viciado se $E(T) = \theta$, $\forall\theta$.

Questão: Se nos queremos usar um estimador não-viciado T, existe um limite inferior para sua variância, $V(T) = v(\theta)$? Caso exista nos podemos atingir este limite na prática ?

Para responder a esta questão, nos precisamos de alguns teoremas sobre a log-verossimilhança e suas derivadas com relação a θ. Até agora, nos temos tratado $l(\theta)$ com uma função de θ, com os elementos do vetor $\underline{x}$ fixados nos valores observados nos dados. Para estudar as propriedades teóricas da log-verossimilhança, nos precisamos relembrar que $\underline{x}$ é uma realização de um vetor de variáveis aleatórias $\underline{X}$, e considerar $l(\theta)$ como uma familia de variáveis aleatórias indexadas por um parâmetro θ. Para enfatizar isto, vamos usar uma notação aumentada $l(\theta; \underline{X})$. Então, o principal resultado desta seção é o seguinte:

Teorema 4 (Cramér-Rao) *Se T é um estimador não-viciado para θ e $l(\theta; \underline{X}$ é duas vezes diferenciável com respeito a θ, então*

$$V(T) \geq 1/E[-l''(\theta; \underline{X}))]$$

Este resultado foi primeiro provado por Harold Cramér e C R Rao, e é chamado de **limite de Cramér-Rao**. Para provar este teorema precisamos de uma definição e uma dupla de lemas.

Definição 15 *A **função escore** é uma familia de variáveis aleatórias definidas, para cada valor do parâmetro θ, como*

$$U(\theta) = \frac{d}{d\theta} l(\theta; \underline{X})$$

Lema 1 *Seja $X_i : i = 1, \ldots, n$ variáveis aleatórias independentes e identicamente distribuídas com função densidade ou probabilidae comum $p(x; \theta)$. Então*

$$U(\theta) = \sum_{i=1}^{n} U_i(\theta)$$

onde $U_i(\theta)$ são variáveis aleatórias mutuamente independentes e identicamente distribuídas, definidas como

$$U_i(\theta) = \frac{d}{d\theta} \log p(X_i; \theta)$$

.

Lema 2 *(i) $E[(U(\theta)] = 0$, (ii) $V[U(\theta)] = E\left[-\frac{d^2}{d\theta^2} l(\underline{X}; \theta)\right]$*

Exemplo 18 *Seja X_1 e X_2 independentes, cada uma com função densidade de probabilidade dada por*

$$f(x) = \theta \exp^{-\frac{x}{\theta}} : \quad x \geq 0$$

Seja $T = 2min(X_1, X_2)$.Mostre que T é não-viciado. Verifique se T atinge o limite de Cramér-Rao. Encontre um estimador não-viciado que atinja o limite.

4.4 Propriedades do estimador e de intervalos baseados na verossimilhança

Primeiro, lembre-se que se $X \sim N(0, \sigma^2)$ então $Z = (X - \mu)/\sigma \sim N(0, 1)$.

Teorema 5 *Se $Z_i : i = 1, \ldots, k$ são variáveis aleatórias independentes $N(0, 1)$ então $Y = Z_1^2 + Z_2^2 \ldots Z_n^2$ tem distribuição **Qui-Quadrado** com n graus de liberdade. Denotamos isso como $Y \sim \chi_n^2$.*

Teorema 6 *Se $Y_i : i = 1, \ldots, k$ são variáveis aleatórias independentes Qui-Quadrado com $Y_i \sim \chi^2_{n_i}$, então $X = Y_1 + \ldots + Y_k$ tem distribuição Qui-Quadrado com $\sum_{i=1}^{k} n_i$ graus de liberdade.*

4.4.1 Distribuição do estimador de máxima verossimilhança em grandes amostras

Para provar os resultados gerais sobre as propriedades de $\hat{\theta}(\underline{X})$ em grandes amostras nos precisamos assumir que as seguintes condições de regularidade estão satisfeitas:

1. O verdadeiro valor de θ é um ponto interior do espaço paramétrico.
2. O valor do máximo da função só é atingido por um θ.
3. As primeiras três derivadas de $l(\theta)$ existem em uma vizinhança do valor de θ.

Definição 16 *A **informação de Fisher** para θ contida em $\underline{X}$ é*

$$J(\theta) = E\left[-\frac{d^2}{d\theta^2} l(\underline{X}; \theta)\right] = E[U'(\theta)]$$

Lema 3 *Se $X_i : i = 1, \ldots, n$ são iid com fdp $p(x; \theta)$ e $j(\theta) = E\left[-\frac{d^2}{d\theta^2} p(X_i; \theta)\right]$, então $J(\theta) = nj(\theta)$.*

Comentário Nos chamamos $j(\theta)$ **informação de Fisher para uma única amostra**, $j(\theta) = E(-U_i'(\theta))$. Nos agora apresentamos um teorema sem demonstração que é muito útil, e intuitivo, porém muito dificil de provar.

Teorema 7 *O estimador de máxima verossimilhança $\hat{\theta}(\underline{X})$ é um estimador consistente para θ.*

Nos agora temos todos os ingredientes necessários para provar dois teoremas mais poderosos sobre as propriedades dos estimadores de máxima verossimilhança. As condições gerais são que $X_i : i = 1, \ldots, n$ são variáveis aleatórias independentes com função densidade ou probabilidade comum $p(x; \theta)$ indexada por um parâmetro escalar θ e satisfazendo as condições de regularidade.

Teorema 8 *No limite quando $n \to \infty$,*

$$(\hat{\theta}(\underline{X}) - \theta)\sqrt{J(\theta)} \sim N(0, 1)$$

Este é um teorema muito poderoso. Ele nos diz que em grandes amostras,

- $\hat{\theta}$ é aproximadamente não-viciado para θ.
- $\hat{\theta}$ atinge o limite de Cramér-Rao.
- Nos podemos construir intervalos da forma $\hat{\theta} \pm c[J(\theta)]^{-\frac{1}{2}}$, onde c é obtido vindo de uma tabela da distribuição Normal.

Teorema 9 *No limite quando* $n \to \infty$,

$$D(\theta) = 2(l(\hat{\theta}) - l(\theta)) \sim \chi^2_1$$

Exemplo 19 *Seja* $f(x;\theta) = \theta \exp^{-\theta x} : \quad x \geq 0$. *Escreva a verossimilhança, encontre o estimador de máxima verossimilhança, verifique se ele atinge o limite de Cramér-Rao e construa um intervalo de confiança.*

Exemplo 20 *Seja* $f(x;\theta) = \frac{\exp^{-\theta}\theta^x}{x!} : \quad x \geq 0$. *Escreva a verossimilhança, encontre o estimador de máxima verossimilhança, verifique se ele atinge o limite de Cramér-Rao e construa um intervalo de confiança.*