Inferência para CS
Tópico 11 - Otimalidade do MLE

Renato Martins Assunção

DCC - UFMG

Uma abordagem de estimação

- Considere a classe $\mathcal{C}$ de *todos* os estimadores não-viciados de $\boldsymbol{\theta}$.
- Por exemplo, se os dados $Y_1, \ldots, Y_n$ forem uma amostra aleatória de uma $N(\mu, \sigma^2)$ e se $n = 2k + 1$ é um ímpar, então:
  - a média amostral $\overline{Y}_n$ é não-viciada para estimar μ e portanto pertence a $\mathcal{C}$
  - a mediana amostral $M = Y_{(r+1)}$ (a estatística de ordem $r + 1$) é não-viciada para estimar μ e portanto pertence a $\mathcal{C}$
  - Se $w \in (0, 1)$, qualquer combinação linear da forma $w\overline{Y}_n + (1 - w)M$ é não-viciada para estimar μ e portanto também pertence a $\mathcal{C}$
  - Existem infinitos outros estimadores não viciados de μ que pertencerão à classe $\mathcal{C}$
- Estratégia: procurar dentre os estimadores não-viciados em $\mathcal{C}$ por um estimador que tenha a variância mínima: estimador *ótimo* para $\boldsymbol{\theta}$ na classe dos estimadores não-viciados.

Cota de Cramér-Rao

- Como podemos saber que um estimador tem variância mínima?
- Usando a Desigualdade da Informação (Cramér-Rao).
- Fixado o tamanho da amostra n,
  - existe um limite na variância de qualquer $\hat{\theta}$ não-viciado.
  - É a cota inferior de Cramér-Rao.
  - Isto fornece um limite inferior para a precisão (ou MSE) de um estimador não-viaciado de θ.
  - NADA pode ser mais preciso que esta cota de Cramér-Rao (dentre os não–viciados).

Teorema

- *Sejam* $Y_1, \ldots, Y_n$ *variáveis i.i.d. com densidade conjunta* $f(\mathbf{y}; \theta)$.
- *Se* $\hat{\theta}$ *é qualquer estimador não viciado de* θ, *então*

$$Var(\hat{\theta}) \geq \frac{1}{I(\theta)} .$$

onde $I(\theta)$ *é a Informação de Fisher e é dada por*

$$I(\theta) = E\left[\frac{\partial}{\partial\theta} \log f(\mathbf{Y}; \theta)\right]^2 .$$

Exemplo:

- Suponha que $Y_1, \ldots, Y_n$ são i.i.d. Poisson(θ).
- Então
$$p(\mathbf{y};\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{\theta^{\sum_i y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!} .$$
- Portanto
$$\log p(\mathbf{y};\theta) = \left(\sum_{i=1}^{n} y_i \right) \log\theta - n\theta - \log\left(\prod_{i=1}^{n} y_i! \right) .$$
- Derivando com relação a θ temos que
$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \log p(\mathbf{y};\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - n$$
- A quantidade $\frac{\partial \ell}{\partial \theta}$ é muito importante e é chamada de função escore (ou *score function*, em inglês).

Exemplo: (continuação)

- No caso de v.a.'s i.i.d. Poisson(θ), a função escore é

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - n$$

- Esta função depende dos dados observados.
- Por exemplo, se $n = 4$ e $\boldsymbol{y} = (3, 1, 0, 3)$ então

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - n = \frac{3+1+0+3}{\theta} - 4 = \frac{7}{\theta} - 4$$

- Note que $\partial \ell / \partial \theta$ é uma função de θ.
- É esta função que usamos para obter o MLE ao igualar o escore a zero e resolver para θ:

$$0 = \frac{\partial \ell}{\partial \theta} = \frac{7}{\theta} - 4$$

Exemplo: (continuação)

- Neste exemplo, com $n = 4$ e $\boldsymbol{y} = (3, 1, 0, 3)$ o escore

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - n = \frac{7}{\theta} - 4$$

  é uma função matemática de θ, não é uma variável aleatória.
- Os dados $\boldsymbol{y} = (3, 1, 0, 3)$ são considerados fixos, são as instâncias observadas no experimento.
- Entretanto, para estudar as propriedades do MLE, vamos transformar este escore numa VARIÁVEL ALEATÓRIA.
- Para isto, vamos substituir o vetor de instâncias $\boldsymbol{y} = (3, 1, 0, 3)$ pelas variáveis aleatórias $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)$:

$$\frac{\partial \log p(\mathbf{Y}; \theta)}{\partial \theta} = \frac{\sum_{i=1}^{4} Y_i}{\theta} - 4$$

- O que mudou? O escore $\partial \ell / \partial \theta$ é agora uma v.a.: possui lista de valores possíveis e probabilidades associadas.

Exemplo: (continuação)

- Vamos entender o que é o escore como v.a.

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \log p(\mathbf{Y};\theta)}{\partial \theta} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{\theta} - 4$$

- O que torna esta expressão uma v.a. é a presença da soma das v.a.'s Y_i no numerador.
- Resultado de probabilidade: Se $Y_1, \ldots Y_n$ são independentes com distribuição Poisson(λ_i) então a sua soma é uma outra v.a. Poisson(λ) com valor esperado $\lambda = \lambda_1 + \ldots \lambda_n$.
- Note a presença da soma $Y_1 + Y_2 + Y_3 + Y_4$ no numerador do escore: esta soma é uma v.a. com distribuição Poisson(4θ).

Exemplo: (continuação)

- Assim, o escore $\partial \ell / \partial \theta$ tem uma distribuição associada com uma Poisson(4θ):

$$\frac{\partial \ell}{\partial \theta} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{\theta} - 4 \sim \frac{\text{Poisson}(4\theta)}{\theta} - 4$$

- Os valores possíveis e probabilidades associadas de $\partial \ell / \partial \theta$ são:

valores	$\frac{0}{\theta} - 4$	$\frac{1}{\theta} - 4$	$\frac{2}{\theta} - 4$	$\frac{3}{\theta} - 4$	...
probabs	$e^{-4\theta}$	$e^{-4\theta} 4\theta$	$e^{-4\theta}(4\theta)^2/2$	$e^{-4\theta}(4\theta)^3/3!$	...

- As probabilidades são obtidas a partir da fórmula das probabilidades de uma Poisson(4θ).

Exemplo: (continuação)

- Assim, transformamos o escore numa v.a. substituindo as instâncias $\boldsymbol{y}$ pelas v.a.'s $\boldsymbol{Y}$

$$\frac{\partial \ell}{\partial \theta} = \frac{\partial \log p(\mathbf{Y};\theta)}{\partial \theta} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{\theta} - 4$$

- Sendo agora uma v.a., podemos calcular sua esperança e sua variância.
- Por exemplo, a esperança da função escore:

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right) &= \mathbb{E}\left(\frac{Y_1 + Y_2 + Y_3 + Y_4}{\theta} - 4\right) \\
&= \frac{\mathbb{E}(Y_1 + Y_2 + Y_3 + Y_4)}{\theta} - 4 \\
&= \frac{\mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \mathbb{E}(Y_3) + \mathbb{E}(Y_4)}{\theta} - 4 \\
&= \frac{\theta + \theta + \theta + \theta}{\theta} - 4 = 0
\end{aligned}$$

Exemplo: (continuação)

- Mais importante é a variância do escore.
- Para qualquer v.a. X temos $\mathbb{V}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$
- Assim, como a esperança do escore é igual a zero, temos

$$\mathbb{V}\left(\frac{\partial \ell}{\partial \theta}\right) = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] + \left[\mathbb{E}\left(\frac{\partial \ell}{\partial \theta}\right)\right]^2 = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right]$$

- Assim, usando a definição de $I(\theta)$, temos:

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\left(\frac{\partial}{\partial \theta}\log f(\mathbf{Y};\theta)\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] \\ &= \mathbb{V}\left(\frac{\partial \ell}{\partial \theta}\right) = \mathbb{V}\left(\frac{\text{Poisson}(4\theta)}{\theta} - 4\right) \\ &= \frac{\mathbb{V}(\text{Poisson}(4\theta))}{\theta^2} = \frac{4\theta}{\theta^2} = \frac{4}{\theta} \end{aligned}$$

Exemplo: (continuação)

- Pela desigualdade de Cramér-Rao, se $\hat{\theta}$ é não viciado para estimar θ numa amostra de tamanaho $n = 4$ de v.a.'s i.i.d. Poisson(θ), então

$$MSE(\hat{\theta}) = \mathbb{V}(\hat{\theta}) \geq \frac{\theta}{n} \, .$$

- Considere o estimador $\hat{\theta} = \overline{Y}$.
- Faça as contas para verificar que $\overline{Y}$ é não-viciado para θ. Além disso,

$$MSE(\overline{Y}) = \mathbb{V}(\overline{Y}) = \frac{\theta}{n} = \frac{1}{I(\theta)} \, .$$

- Assim, se as v.a.'s são i.i.d. Poisson(θ), ninguém pode ser melhor do que o bom e velho $\overline{Y}$ para estimar θ na classe dos estimadores não-viciados.

Recordando probab

- É muito útil recordar uma fórmula de probabilidade.
- Seja $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ um vetor aleatório com densidade de probabilidade $f(\boldsymbol{y}) = f(y_1, \ldots, y_n)$.
- Seja $g(\boldsymbol{Y})$ uma nova v.a. obtida através de uma função matemática qualquer aplicada ao vetor $\boldsymbol{Y}$.
- Por exemplo, $g(\boldsymbol{Y})$ poderia ser $g(\boldsymbol{Y}) = \overline{Y}$ ou $g(\boldsymbol{Y}) = \sum 1/\log(Y_i) - \pi^2$
- Como calcular a esperança desta nova v.a. $g(\boldsymbol{Y})$?

Recordando probab

- Temos
$$\mathbb{E}(g(\boldsymbol{Y})) = \int \cdots \int g(\mathbf{y})\ f(\mathbf{y}) d\mathbf{y}$$
onde a integral é tomada sobre todos os valores possíveis do vetor $\boldsymbol{y}$.
- Por exemplo, se $g(\boldsymbol{Y}) = \sum 1/\log(Y_i) - \pi^2$ então

$$\begin{aligned}
\mathbb{E}(g(\boldsymbol{Y})) &= \mathbb{E}\left(\sum_i \frac{1}{\log(Y_i)} - \pi^2\right) \\
&= \int \cdots \int \left(\sum_i \frac{1}{\log(y_i)} - \pi^2\right)\ f(\mathbf{y}) d\mathbf{y} \\
&= \int \cdots \int \left(\sum_i \frac{1}{\log(y_i)} - \pi^2\right)\ f(y_1, \ldots, y_n) d\mathbf{y}
\end{aligned}$$

- Não se preocupe. Não teremos de calcular esta integral explicitamente...

Prova da desigualdade de Cramér-Rao

- Vamos considerar três lemas auxiliares para provar a desigualdade de Cramér-Rao.
- No caso particular de uma amostra de v.a.'s i.i.d. Poisson, verificamos que a esperança do escore é zero.
- Isto é verdade em qualquer modelo estatístico, não apenas neste exemplo particular.
- Este é o primeiro lema: a esperança da função escore é sempre igual a zero.

Lema 1

Lema

$$\mathbb{E}\left[\frac{\partial \ell}{\partial \theta}\right] = \mathbb{E}\left[\frac{\partial}{\partial \theta} \log f(\mathbf{Y}; \theta)\right] = 0$$

Prova:

- Como $f(\mathbf{y}; \theta)$ é uma densidade de probabilidade, sua integral sobre todos os valores possíveis de $\boldsymbol{y}$ é igual a 1:

$$1 = \int \cdots \int f(\mathbf{y}; \theta) d\mathbf{y}$$

Derivando com respeito a θ dos dois lados desta igualdade, temos

$$0 = \frac{\partial (1)}{\partial \theta} = \frac{\partial}{\partial \theta} \int \cdots \int f(\mathbf{y}; \theta) d\mathbf{y} = \int \cdots \int \frac{\partial}{\partial \theta} f(\mathbf{y}; \theta) d\mathbf{y}$$

- Multiplicando e dividindo o integrando por $f(\mathbf{y};\theta)$ obtemos

$$\begin{aligned} 0 &= \int \cdots \int \frac{\partial}{\partial\theta} f(\mathbf{y};\theta) d\mathbf{y} = \int \cdots \int \frac{\frac{\partial}{\partial\theta} f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} f(\mathbf{y};\theta) d\mathbf{y} \\ &= \int \cdots \int \frac{\partial \ell}{\partial\theta} \, f(\mathbf{y};\theta) d\mathbf{y} \end{aligned}$$

já que

$$\frac{\partial \ell}{\partial\theta} = \frac{\partial}{\partial\theta} \log f(\mathbf{y};\theta) = \frac{\frac{\partial}{\partial\theta} f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}$$

Pela regra de cálculo de $\mathbb{E}(g(\boldsymbol{Y}))$ em probabilidade temos que

$$\int \cdots \int \frac{\partial \ell}{\partial\theta} \, f(\mathbf{y};\theta) d\mathbf{y} = \mathbb{E}\left[\frac{\partial \ell}{\partial\theta}\right] .$$

- Concluímos assim que

$$\mathbb{E}\left[\frac{\partial \ell}{\partial\theta}\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta} \log f(\mathbf{Y};\theta)\right] = 0 .$$

Lema 2

- Recordando de probab: Se W e V são v.a.'s então $|\text{Corr}(W, V)| \leq 1$.
- Mas como $\text{Corr}(W, V) = \frac{\text{Cov}(W,V)}{\sqrt{\mathbb{V}(W)}\sqrt{\mathbb{V}(V)}}$, nós temos que

$$\text{Cov}^2(W, V) \leq \mathbb{V}(W)\mathbb{V}(V)$$

Tomando $W = \hat{\theta}$, temos que para qualquer v.a. V,

$$\text{Cov}^2(\hat{\theta}, V) \leq \mathbb{V}(\hat{\theta})\mathbb{V}(V)$$

- Rearranjando a ordem, podemos concluir que

Lema

Para qualquer v.a. V

$$Var(\hat{\theta}) \geq \frac{Cov^2(\hat{\theta}, V)}{Var(V)} .$$

Lema 3

- Recordando de probab: Se W e V são v.a.'s então $\mathrm{Cov}(W, V) = \mathbb{E}(W\ V) - \mathbb{E}(W)\ \mathbb{E}(V)$.
- Tome $V = \partial\ell/\partial\theta$, o escore e $W = \hat{\theta}$, um estimador QUALQUER de θ (não precisa ser o MLE, é um estimador arbitrário).
- Pelo Lema 1, temos $\mathbb{E}(\partial\ell/\partial\theta) = 0$. Portanto, temos

Lema

$$\begin{aligned} \mathbb{V}\left(\frac{\partial\ell}{\partial\theta}\right) &= \mathbb{E}\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right] + \left(\mathbb{E}\frac{\partial\ell}{\partial\theta}\right)^2 \\ &= \mathbb{E}\left[\left(\frac{\partial\ell}{\partial\theta}\right)^2\right] + 0 = I(\theta) \end{aligned}$$

e

$$Cov\left(\hat{\theta}, \frac{\partial\ell}{\partial\theta}\right) = \mathbb{E}\left(\hat{\theta}\ \frac{\partial\ell}{\partial\theta}\right) - \mathbb{E}(\hat{\theta})\ E\left(\frac{\partial\ell}{\partial\theta}\right) = \mathbb{E}\left(\hat{\theta}\ \frac{\partial\ell}{\partial\theta}\right)$$

Final da prova da desigualdade de Cramer-Rao

- Pelos três lemas anteriores temos que

$$\mathbb{V}(\hat{\theta}) \geq \frac{\left(\text{Cov}(\hat{\theta},\ \frac{\partial \ell}{\partial \theta})\right)^2}{I(\theta)} = \frac{\left[\mathbb{E}\left(\hat{\theta}\ \frac{\partial \ell}{\partial \theta}\right)\right]^2}{I(\theta)} .$$

- Resta apenas mostrar que o numerador é igual a um.
- Como $\hat{\theta}$ é não-viciado para θ temos que

$$\theta = E(\hat{\theta}) = \int \cdots \int \hat{\theta}(\mathbf{y}) f(\mathbf{y}; \theta) d\mathbf{y} .$$

- Vamos derivar dos dois lados em relação a θ. Pelo lado esquerdo ficamos com

$$\frac{\partial}{\partial \theta}\theta = 1 .$$

- Pelo lado direito

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int\cdots\int \hat{\theta}(\mathbf{y}) f(\mathbf{y};\theta) d\mathbf{y} &= \int\cdots\int \hat{\theta}(\mathbf{y}) \frac{\partial f(\mathbf{y};\theta)}{\partial\theta} d\mathbf{y} \\
&= \int\cdots\int \hat{\theta}(\mathbf{y}) \frac{\partial f(\mathbf{y};\theta)}{\partial\theta} \; \frac{f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)} \; d\mathbf{y} \\
&= \int\cdots\int \hat{\theta}(\mathbf{y}) \frac{\partial \log f(\mathbf{y};\theta)}{\partial\theta} \; f(\mathbf{y};\theta) \; d\mathbf{y} \\
&= \mathbb{E}\left[\hat{\theta}(\mathbf{Y}) \frac{\partial \log f(\mathbf{Y};\theta)}{\partial\theta}\right] \\
&= \text{Cov}\left(\hat{\theta}(\mathbf{Y}), \frac{\partial \log f(\mathbf{Y};\theta)}{\partial\theta}\right)
\end{aligned}$$

- Como os dois lados devem ser iguais, concluímos que o numerador é igual a 1. Isto é, $1 = \text{Cov}\left(\hat{\theta}, \frac{\partial\ell}{\partial\theta}\right)$ e portanto

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{I(\theta)} \, .$$

Uma forma mais fácil de calcular a $I(\theta)$

Lema

Sob condições de regularidade temos que

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{Y};\theta)\right]$$

Prova:

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{\partial^2}{\partial \theta^2} \log f(\mathbf{y};\theta) = \frac{\partial}{\partial \theta}\left(\frac{\partial}{\partial \theta} \log f(\mathbf{y};\theta)\right) = \frac{\partial}{\partial \theta}\left(\frac{\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\right)$$

$$= \frac{f(\mathbf{y};\theta)\left(\frac{\partial^2}{\partial \theta^2} f(\mathbf{y};\theta)\right) - \left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)\left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)}{f^2(\mathbf{y};\theta)}$$

$$= \frac{f(\mathbf{y};\theta)\frac{\partial^2 f(\mathbf{y};\theta)}{\theta^2} - \left(\frac{\partial}{\partial \theta} f(\mathbf{y};\theta)\right)^2}{f^2(\mathbf{y};\theta)}.$$

- Tomando a esperança ficamos com

$$
\begin{aligned}
-E\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{y};\theta)\right] &= -\int\cdots\int \frac{f(\mathbf{y};\theta)\frac{\partial^2}{\partial\theta^2}f(\mathbf{y};\theta)-\left(\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)\right)^2}{f^2(\mathbf{y};\theta)} f(\mathbf{y};\theta)d\mathbf{y}\,. \\
&= -\int\cdots\int \frac{\partial^2 f(\mathbf{y};\theta)}{\partial\theta^2}d\mathbf{y} + \int\cdots\int \frac{\left(\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)\right)^2}{f(\mathbf{y};\theta)}d\mathbf{y} \\
&= -\frac{\partial^2}{\partial\theta^2}\int\cdots\int f(\mathbf{y};\theta)d\mathbf{y} + \int\cdots\int \left(\frac{\frac{\partial}{\partial\theta}f(\mathbf{y};\theta)}{f(\mathbf{y};\theta)}\right)^2 f(\mathbf{y};\theta)d\mathbf{y} \\
&= -\frac{\partial^2}{\partial\theta^2}(1) + \int\cdots\int \left(\frac{\partial}{\partial\theta}\log f(\mathbf{y};\theta)\right)^2 f(\mathbf{y};\theta)d\mathbf{y} \\
&= 0 + E\left[\frac{\partial}{\partial\theta}\log f(\mathbf{y};\theta)\right]^2 = I(\theta)\,.
\end{aligned}
$$

$I(\theta)$ com n v.a.'s i.i.d.

- Podemos calcular a informação de Fisher com uma única observação $n = 1$.
- Podemos calcular a informação de Fisher com $n > 1$.
- Vamos denotar $I_n(\theta)$ a informação de Fisher com n dados.
- Qual a relação entre $I_n(\theta)$ e $I_1(\theta)$.
- $I_n(\theta)$ aumenta com n? A que taxa?
- Resultado: Se $Y_1, \ldots Y_n$ são i.i.d. então $I_n(\theta) = nI_1(\theta)$.
- A informação sobre θ numa amostra de tamanaho n é igual a n vezes a informação numa amostra de tamanho 1.
- A informação sobre θ aumenta linearmente com n.

Resultado principal

- O estimador de máxima de verossimilhança apresenta a seguinte distribuição assintótica:

$$\hat{\theta}_{EMV} \approx N\left(\theta, \frac{1}{I_n(\theta)}\right)$$

- Isso significa que, para n grande e de forma aproximada, o MLE apresenta as seguintes características:
  - Tem distribuição normal;
  - É não viciado;
  - Atinge a cota de Cramér-Rao, ou seja, apresenta a menor variância possível.
- Este resultado é universal, não interessa o modelo de probabilidade para os dados!!

Como isto é demonstrado?

- Vamos lembrar de algumas propriedades importantes:

$$E\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \quad \text{e} \quad E\left(\frac{\partial \ell}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = I_n(\theta) = nI_1(\theta)$$

- Vamos exemplificar estas propriedades no caso de v.a.'s iid Poisson com parâmetro θ.

Caso particular: Poisson

- $X_1, X_2, ..., X_n$ v.a.'s iid com ditribuição Poisson(θ).
- Conjunta:

$$p(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

- Log-verossimilhança:

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

- Denotando $\ell_i = x_i \log(\theta) - \theta - \log(x_i!)$, podemos escrever $l(\theta) = \sum_{i=1}^{n} l_i(\theta)$.
- Derivando em relação a θ podemos obter a função escore

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta} = \sum_{i=1}^{n} \left(\frac{x_i}{\theta} - 1 \right)$$

Função escore

- Vista como *variável aleatória*, a função escore é dada por

$$\sum_{i=1}^{n} \left(\frac{X_i}{\theta} - 1 \right) = \sum_{i=1}^{n} Y_i$$

- Como $E(X_i) = \theta$, sabemos que

$$E(Y_i) = E\left(\frac{X_i}{\theta} - 1 \right) = E\left(\frac{X_i}{\theta} \right) - 1 = 0$$

- Teremos então que

$$Var(Y_i) = E(Y_i^2) = I_1(\theta)$$

- As variáveis aleatórias $Y_1, Y_2, ..., Y_n$ são i.i.d com média zero e variância dada por $I_1(\theta)$.

Mais uma derivada

- Derivando pela segunda vez em relação a θ:

$$\frac{\partial^2 \ell}{\partial \theta^2} = \sum_{i=1}^{n} \frac{\partial^2 \ell_i}{\partial \theta^2} = \sum_{i=1}^{n} \frac{-x_i}{\theta^2}$$

- Olhando para essa função como uma variável aleatória temos que

$$E\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = \sum_{i=1}^{n} E\left(\frac{-X_i}{\theta^2}\right) = -\frac{n}{\theta}$$

- Mas é fácil perceber que

$$E\left(\frac{\partial l}{\partial \theta}\right)^2 = \sum_{i} E\left(\frac{x_i^2}{\theta} - 2\frac{x_i}{\theta} + 1\right) = \frac{n}{\theta}$$

- Ou seja, verificamos o resultado $I_n(\theta) = -E(\frac{\partial^2 l}{\partial \theta^2}) = E(\frac{\partial l}{\partial \theta})^2 = nI_1(\theta)$.

Dois teoremas limite

- Vamos mostrar que o MLE é assitoticamente normal: precisamos relembrar dois teoremas importantes.
- **Lei Forte dos Grandes Números** Se $Y_1, Y_2, ..., Y_n$ são variáveis aletórias i.i.d com esperanaça finnita então $\bar{Y}_n \to E(Y)$ quando $n \to \infty$.
- **Teorema Central do Limite**: Se $Y_1, Y_2, ..., Y_n$ são variáveis aletórias i.i.d com $E(Y) = 0$ e $Var(Y) = v$. Então
    - $\sqrt{n}(\bar{Y}_n)$ tende em distribuição para uma $N(0, v)$ ou
    - $\sqrt{n}\frac{\bar{Y}_n}{\sqrt{v}}$ tende em distribuição para uma $N(0, 1)$ ou ainda
    - $\frac{1}{\sqrt{n}}(Y_1 + Y_2 + ... + Y_n)$ tende em distribuição para uma $N(0, v)$.

MLE é aproximadamente normal

- Vamos fazer a expnsão de $\frac{\partial l}{\partial \theta}$ em torno do verdadeiro valor do parâmetro θ, que denotaremos por θ^*.
- $\frac{\partial l}{\partial \theta}(\theta^*)$ o valor de $\frac{\partial l}{\partial \theta}$ avaliado no ponto θ^*.
- Como o EMV maximiza a função de verossimilhança:

$$\frac{\partial l}{\partial \theta}(\hat{\theta}_{EMV}) = 0$$

- Expandindo a derivada:

$$0 = \frac{\partial l}{\partial \theta}(\hat{\theta}_{EMV}) \approx \frac{\partial l}{\partial \theta}(\theta^*) + \frac{\partial^2 l}{\partial \theta^2}(\theta^*)(\hat{\theta}_{EMV} - \theta^*)$$

- Então

$$\hat{\theta}_{EMV} - \theta^* \approx \left[-\frac{\partial^2 l}{\partial \theta^2}(\theta^*) \right]^{-1} \frac{\partial l}{\partial \theta}(\theta^*)$$

MLE

- Logo

$$\begin{aligned}\sqrt{n}(\hat{\theta}_{EMV} - \theta^*) &\approx \sqrt{n}\left[-\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1} \frac{\partial l}{\partial \theta}(\theta^*) \\ &= \left[-\frac{1}{n}\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1} \frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}(\theta^*)\end{aligned}$$

- Chegando então a

$$\underbrace{\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)}_{\sqrt{n}\ *\ \text{erro de estimação}} \approx \underbrace{\left[-\frac{1}{n}\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right]^{-1}}_{A} \underbrace{\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}(\theta^*)}_{B}$$

- Desenvolvendo o termo em (A) temos

$$\left[-\frac{1}{n}\frac{\partial^2 l}{\partial \theta^2}(\theta^*)\right] = -\left(\frac{1}{n}\sum_i \frac{\partial^2 l_i}{\partial \theta^2}(\theta^*)\right)$$

MLE

- Olhando para $\frac{\partial^2 l_i}{\partial \theta^2}(\theta^*)$ como uma variável aleatória chega-se a

$$-\left(\frac{1}{n}\sum_i \frac{\partial^2 l_i}{\partial \theta^2}(\theta^*)\right) = -\left(\frac{1}{n}\sum_i Z_i\right)$$

- onde $Z_1, Z_2, ..., Z_n$ são variáveis aleatórias i.i.d. com $E_\theta(Z_i) = -I_1(\theta^*)$
- Pela Lei Forte dos Grandes Números, $\overline{Z}_n \to -I_1(\theta^*)$ quando $n \to \infty$
- Portanto o termo (A) se aproxima de $I_1(\theta^*)^{-1}$ para n suficientemente grande.

..

- Vamos agora desenvolver o termo em (B)

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}(\theta^*) = \frac{1}{\sqrt{n}}\sum_i \frac{\partial l_i}{\partial \theta}(\theta^*)$$

- Novamente olhando para $\frac{\partial l_i}{\partial \theta^2}(\theta^*)$ como uma variável aleatória temos que

$$\frac{1}{\sqrt{n}}\sum_i \frac{\partial l_i}{\partial \theta}(\theta^*) = \frac{1}{\sqrt{n}}\sum_i W_i$$

- onde $W_1, W_2, ..., W_n$ são variáveis aleatórias i.i.d. com $E_{\theta^*}(W_i) = 0$ e $Var_{\theta^*}(W_i) = E_{\theta^*}(W_i^2) = I_1(\theta^*)$
- Pelo Teorema Central do Limite temos que

$$\frac{1}{\sqrt{n}}(W_1 + W_2 + ... + W_n) \xrightarrow{d} N(0, I_1(\theta^*))$$

onde a notação d significa tender em distribuição.

..

- Portanto, como $\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)$ é aproximadamente o termo em (B) multiplicado por $I_1(\theta)$,
- o erro de estimação é normalmente distribuído com

$$E(\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)) = 0 \text{ e } Var(\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)) = \frac{I_1(\theta^*)}{I_1^2(\theta^*)} = \frac{1}{I_1(\theta^*)}$$

- Em outras palavras, o erro de estimação $\sqrt{n}(\hat{\theta}_{EMV} - \theta^*)$ converge em distribuição para uma $N(0, I_1(\theta^*)^{-1})$.
- Logo se n é grande o suficiente o MLE tem distribuição aproximadamente normal com média θ^* e variância $I_1(\theta^*)^{-1}$, ou seja, é aproximadamente não viciado e atinge a cota de Cramer-Rao. Dessa forma, provamos as três principais propriedades do MLE.

..

- Podemos concluir que

$$\sqrt{n}(\hat{\theta}_{EMV} - \theta^*) \approx N(0, I_1(\theta^*)^{-1})$$

- Isto é,

$$(\hat{\theta}_{EMV} - \theta^*) \approx N\left(0, \frac{1}{nI_1(\theta^*)}\right)$$

- ou ainda

$$\hat{\theta}_{EMV} \approx N\left(\theta^*, \frac{1}{nI_1(\theta^*)}\right)$$