Mariella Ananias Bogoni

O Teorema de Rao-Cramér para Estimadores de Máxima Verossimilhança

Volta Redonda, RJ

2019

Mariella Ananias Bogoni

O Teorema de Rao-Cramér para Estimadores de Máxima Verossimilhança

Trabalho de Conclusão de Curso submetido ao Curso de Matemática da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Matemática.

Universidade Federal Fluminense

Instituto de Ciências Exatas

Curso de Matemática

Orientador: Alan Prata de Paula

Coorientador: Marina Sequeiros Dias de Freitas

Volta Redonda, RJ

2019

Ficha catalográfica automática - SDC/BAVR
Gerada com informações fornecidas pelo autor

B674t Bogoni, Mariella Ananias
O Teorema de Rao-Cramér para Estimadores de Máxima Verossimilhança / Mariella Ananias Bogoni ; Alan Prata de Paula, orientador ; Marina Sequeiros Dias de Freitas, coorientador. Volta Redonda, 2019.
52 f. : il.

Trabalho de Conclusão de Curso (Graduação em Matemática)- Universidade Federal Fluminense, Instituto de Ciências Exatas, Volta Redonda, 2019.

1. Máxima Verossimilhança. 2. Estatística. 3. Probabilidade. 4. Regressão Logística. 5. Produção intelectual. I. Paula, Alan Prata de, orientador. II. Freitas, Marina Sequeiros Dias de, coorientador. III. Universidade Federal Fluminense. Instituto de Ciências Exatas. IV. Título.

CDD -

Bibliotecária responsável: Ana Claudia Felipe da Silva - CRB7/4794

Mariella Ananias Bogoni

O Teorema de Rao-Cramér para Estimadores de Máxima Verossimilhança

Trabalho de Conclusão de Curso submetido ao Curso de Matemática da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Bacharel em Matemática.

Trabalho aprovado. Volta Redonda, RJ, 06 de Dezembro de 2019:

Prof. Dr. Alan Prata de Paula – UFF
Orientador

Prof[a]. Dra. Marina Sequeiros Dias de Freitas – UFF
Coorientador

Prof. Dr. Leandro Gines Egea – UFF

Prof[a]. Dra. Marina Ribeiro Barros Dias – UFF

Volta Redonda, RJ
2019

À minha família, amigos e à todos aqueles que me deram força, suporte e amor.

Agradecimentos

Agradeço Àquele que sempre se manteve presente, em sua infinita bondade e amor.

Agradeço muito à minha família, minha mãe pelo exemplo de mulher forte, persistente e por mostrar através de sua vida, a nunca desistir dos sonhos. Ao meu irmão, por acreditar em mim todo o tempo e por todo carinho e suporte, até mesmo financeiro, dado a mim com amor. Aos meus avós, pelas orações e pelo colo nos momentos difíceis.

Agradeço à todos os amigos que percorreram esta trajetória junto comigo, que me deram força e muitos motivos para sorrir. Em especial, agradeço a minha amiga e irmã de coração Daiana Gomes, por estar e lutar comigo o tempo todo.

Um agradecimento especial ao meu orientador, Alan Prata, por sua ótima orientação, empenho e preocupação em me preparar para o melhor. Obrigada pela confiança e pelo valioso incentivo que me deu em cada conversa que tivemos. Com certeza a "culpa"do meu interesse nesta área é dele, pela forma em que me apresentou todos os objetos desde o início. Espero poder aprender mais com você novamente.

Não poderia deixar de agradecer a minha coorientadora, Marina Sequeiros Dias de Freitas, pelo imenso apoio e contribuição neste trabalho, a orientação de monitoria e nos outros projetos acadêmicos e disciplinas. Sem dúvida, seu toque agregou muito a tudo isso.

Outro agradecimento especial para o professor Rodrigo Amorim, que me apresentou e muito me ensinou sobre Computação de Alto Desempenho. Obrigada pela paciência, dedicação e suporte nesta reta final.

Por fim, agradeço ao corpo docente do Departamento de Matemática do Instituto de Ciências Exatas da UFF pelo envolvimento, paciência e carinho com os alunos durante a graduação.

Resumo

Em muitos problemas de inferência busca-se encontrar métodos para a estimação de parâmetros desconhecidos. Dentre os métodos existentes para estimação, o Método de Máxima Verossimilhança se mostra eficaz, pois produz estimadores com ótimas propriedades. Neste trabalho apresentaremos o Método de Máxima Verossimilhança e as principais propriedades dos estimadores obtidos, dentre elas: consistência, eficiência, caracterizada pelo Limite Inferior de Rao-Cramér, e distribuição assintótica. A partir da distribuição assintótica destes estimadores, abordaremos os procedimentos de inferência utilizados no contexto da máxima verossimilhança, fornecendo ao leitor uma aplicação prática do método em Regressão Logística. Com o método, obtemos uma maneira simples e satisfatória de encontrar estimadores para a resolução de problemas paramétricos. Tal solução é apropriada e vantajosa do ponto de vista matemático e estatístico.

Palavras-chave: Máxima Verossimilhança. Estimadores. Regressão Logística.

Abstract

In many inference problems, the goal is to find methods to estimate unknown parameters. Among the existing methods for estimation, the Maximum Likelihood Method is effective because it produces estimators with great properties. In this work, we present the Maximum Likelihood Method and the main properties of the estimators obtained, among them: consistency, efficiency, characterized by Rao-Cramér Lower Bound, and asymptotic distribution. From asymptotic distribution of that estimators, we will approach the inference procedures used in the context of maximumm likelihood, providing the reader a practical aplication of the method in Logistic Regression. With this method, we have a simple and satisfactory way to find estimators for the resolution of parametric problems. Such a solution is appropriate and advantageous from mathematical and statistical point of view.

Keywords: Maximum Likelihood. Estimators. Logístic Regression.

Sumário

1 Introdução

Em Estatística, um procedimento de estimação é geralmente caracterizado por um modelo estatístico, onde há um parâmetro desconhecido, e um conjunto de dados, que será uma das ferramentas utilizadas para a estimação. A partir disso, há alguns métodos que podem ser aplicados para obter o estimador e o desafio é obter estimadores com boas propriedades. Este trabalho apresenta o Método de Máxima Verossimilhança, construindo a teoria do método com base em [3].

Sendo θ um parâmetro desconhecido, o método se baseia em encontrar um estimador $\hat{\theta}$ de θ maximizando a função de probabilidade conjunta de uma amostra aleatória da população. Isto é, seja $X_1, \ldots, X_n$ uma amostra aleatória da variável aleatória X com função densidade de probabilidade $f(x, \theta)$, onde θ é o parâmetro desconhecido. A função de probabilidade conjunta da amostra é

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta), \tag{1.1}$$

onde $\mathbf{x} = (x_i, \ldots, x_n)$ representa os valores que a amostra $X_1, \ldots, X_n$ assume. Assim, o ponto $\hat{\theta}$ que maximiza a função $L(\theta; \mathbf{x})$ será o estimador de Máxima Verossimilhança de θ.

O Método de Máxima Verossimilhança começou a ser introduzido muito antes do século XX, embora ainda não existisse nenhuma defesa fundamentada para o desempenho do método. Somente a partir do século XX que Ronald Fisher buscou formas de provar alguma propriedade para os Estimadores de Máxima Verossimilhança [12], [13]. A primeira delas envolvia o conceito de suficiência. Um estimador suficiente é aquele que captura toda a informação do parâmetro contida na amostra. Formalmente, $\hat{\theta}$ é uma estatística suficiente para θ se

$$\mathbb{P}[X_1, \ldots, X_n = x_1, \ldots, x_n \mid \hat{\theta} = \theta_0] \tag{1.2}$$

não depende de θ, ou seja, a probabilidade condicional em 1.2 é a mesma para todo valor de θ. Fisher descobriu o que estimador para desvio padrão amostral é suficiente e isso o motivou a tentar definir o mesmo resultado para aos Estimadores de Máxima Verossimilhança.

Em 1922, Fisher apresentou uma justificativa matemática de que o método de máxima verossimilhança sempre levava a uma estatística suficiente, quando combinado a hipótese de consistência e distribuição aproximadamente normal para o estimador. Mas Fisher percebeu que não se podia garantir suficiência num caso geral o que o levou a reformular seu argumento, agora envolvendo eficiência, que está relacionada ao quão preciso é o estimador. Um estimador $\hat{\theta}$ é eficiênte se sua variância atinge o limite

$$Var(\hat{\theta}) \geq \frac{1}{nI(\theta)}, \tag{1.3}$$

onde $I(\theta)$ é chamada Informação de Fisher e n o tamanho da amostra.

A segunda prova de Fisher, baseada na análise da variância, afirmava que o método de máxima verossimilhança sempre fornecerá uma estatística que, se normalmente distribuída, possuirá variância mínima, atingindo a cota dada em 1.3, chamada por Fisher de Desigualdade da Informação, estabelecendo eficiência. A terceira e última prova foi publicada em 1930 quando, em parceria com o matemático Harold Hotteling, Fisher finalmente forneceu uma prova satisfatória em termos de rigor matemático, desta vez, impondo mais algumas hipóteses, as Condições de Regularidade.

A Desigualdade da Informação citada na demonstração de Fisher é hoje conhecida como o Limite Inferior de Rao-Cramér, utilizada como critério para a eficiência de estimadores, ponto principal abordado neste trabalho. Além disso, Hotteling apresentou uma prova reformulada para a distribuição assintótica dos estimadores de máxima verossimilhança, o que viabilizou todos os procedimentos de inferência estatística. Tudo isso faz com que o método seja ótimo, pois fornece uma maneira simples para se obter uma estimativa ideal em problemas paramétricos.

Neste trabalho será apresentada a construção teórica deste método, a justificativa matemática para considerá-lo, bem como as principais propriedades dos estimadores. A principal delas é dada pelo Limite Inferior de Rao-Cramér, estabelecendo a eficiência assintótica desses estimadores. Além disso, veremos que os Estimadores de Máxima Verossimilhança possuem distribuição normal assintoticamente, um resultado muito importante que possibilita realizar inferência utilizando essa estatística, ou seja, podemos construir Teste de Hipóteses e Intervalo de Confiança.

No segundo Capítulo, introduziremos o leitor aos conceitos básicos e fundamentais de Probabilidade e Estatística. No terceiro Capítulo definimos a noção de convergência de variáveis aleatórias e seus principais resultados, como o Teorema Central do Limite. Estenderemos as definições de convergência para o caso multivariado, finalizando com o Teorema Central do Limite Multivariado. Nos capítulos 4 e 5, apresentamos o Método de Máxima Verossimilhança, seus estimadores e suas propriedades, discutindo também Teste de Hipóteses e Intervalo de Confiança. Finalmente, apresentamos no Capítulo 6 ao leitor uma aplicação do método para a estimação dos parâmetros de um modelo de Regressão Logística.

2 Noções de Probabilidade e Estatística

Neste capítulo começamos a construir nosso estudo a partir de algumas noções e resultados básicos em Probabilidade e Estatística. O objetivo é fornecer ao leitor os conceitos necessários para estudarmos o tema central deste trabalho. Podemos começar entendendo o que é um Espaço de Probabilidade. Para mais detalhes e demonstrações o leitor por consultar [1] e [2].

Um Espaço de Probabilidade é um trio $(\Omega, \mathcal{F}, \mathbb{P})$ composto pelo Espaço Amostral, que é o conjunto Ω não vazio cujos elementos representam todos os resultados possíveis de um determinado experimento; o subconjunto $\mathcal{F} \subset \Omega$ de eventos aleatórios [1] e uma medida de probabilidade, ou seja, uma função $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ que a cada elemento $A \in \mathcal{F}$ atribui um valor real que chamamos de probabilidade. Esta função satisfaz três propriedades:

P1 $\mathbb{P}(A) \geq 0 \ \forall A \in \mathcal{F}$;

P2 $\mathbb{P}(\Omega) = 1$;

P3 Dados $A_1, A_2, \ldots \in \mathcal{F}$ com $A_i \cap A_j = \varnothing \ \forall i \neq j$ então $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

A partir de um experimento do Espaço Amostral, podemos estar interessados em alguma função deste experimento. Por exemplo, ao jogarmos um dado talvez seja mais interessante analisar quantos números pares que pode aparecer do que o evento "aparecer um número par". Chamamos a função de Variável Aleatória. Note que como os valores que a variável aleatória assume dependem do resultado do experimento, podemos adicionar probabilidade a esses valores.

Definição 2.1 (Variável Aleatória[2])**.** Um variável aleatória é uma função $X : \Omega \to \mathbb{R}$ que a cada elemento $w \in \Omega$ associa um número real $X(w)$ tal que $\{w \in \Omega; X(w) \leq x\} \in \mathcal{F}$ para todo $x \in \mathbb{R}$.

Os dois principais tipos de variáveis aleatórias são as discretas e as contínuas. As variáveis aleatórias discretas são aquelas que assumem apenas valores em um conjunto enumerável e podemos calcular probabilidades atráves de sua função de probabilidade.

Definição 2.2 (Função de Probabilidade)**.** Seja X uma variável aleatória discreta. Definimos a função de probabilidade de X como,

$$p(a) = \mathbb{P}(X = a),$$

[1] De modo preciso, $\mathcal{F}$ é σ-álgebra de subconjuntos de Ω.

[2] Precisamente, uma variável aleatória é uma função mensurável de $(\Omega, \mathcal{F})$ em $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

onde a pertence a um conjunto enumerável de valores que X assume.

Assim, devido as propriedades P1 e P2, se X assume os valores $x_1, x_2, \ldots$ então para cada $i = 1, 2$, vale que $p(x_i) \geq 0$ e $\sum_{i=1}^{\infty} p(x_i) = 1$.

Abaixo veremos os principais distribuições discretas.

Definição 2.3 (Distribuição de Bernoulli)**.** Dizemos que X tem Distribuição Bernoulli com parâmetro $p \in [0, 1]$, e denotamos por $X \sim Be(p)$, se X assume os valores 0 e 1 com função de probabilidade dada por:

$$P(X = 1) = p \tag{2.1}$$

$$P(X = 0) = 1 - p \tag{2.2}$$

onde o evento $X = 1$ é associado como "sucesso" e $X = 0$ como "fracasso".

Definição 2.4 (Distribuição Binomial)**.** Considere n ensaios de Bernoulli independentes e com mesmo parâmetro p, e seja X o número de sucessos obtidos. Dizemos que X segue o modelo Binomial com parâmetros n e p, e denotamos por $X \sim b(n, p)$, se a função de probabilidade é dada por,

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

Ao contrário das variáveis aleatórias discretas, se X é variável aleatória contínua $\mathbb{P}(X = a) = 0$, $\forall a \in \mathbb{R}$. Desse modo, calculamos probabilidades de variáveis aleatórias contínuas atribuindo chance diretamente a cada evento aleatório da reta, através da função densidade de probabilidade.

Definição 2.5 (Variável Aleatória Contínua)**.** Dizemos que X é uma variável aleatória contínua se existir uma função não negativa f, definida para todo real $x \in (-\infty, \infty)$, que tenha a propriedade de que, para qualquer evento aleatório $B \in \mathcal{F}$,

$$P(X \in B) = \int_B f(x)dx.$$

A função f é chamada de *função densidade de probabilidade* da variável aleatória X.

A Definição 2.5 diz que a probabilidade de X pertencer a um conjunto B é a integral de sua função densidade f sob o conjunto B. Neste caso, sendo $B = [a, b]$ então

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

Da propriedade P2 temos que

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Outro objeto importante para o cálculo de probabilidades é a função de distribuição acumulada.

Definição 2.6 (Função Distribuição Acumulada)**.** Seja X uma variável aleatória. A função de distribuição acumulada de X é definida por

$$F_X(a) = \mathbb{P}(X \leq a) = \begin{cases} \sum\limits_{x \leq a} p(x), & \text{se X é discreta} \\ \int_{-\infty}^{a} f(x)dx, & \text{se X é contínua} \end{cases}. \tag{2.3}$$

Vejamos algumas distribuições contínuas conhecidas.

Definição 2.7 (Distribuição Normal)**.** Dizemos que a variável aleatória X tem distribuição normal com parâmetros $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, denotado por $X \sim N(\mu, \sigma^2)$, se X tem como densidade,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.4}$$

A distribuição $Z \sim N(0, 1)$ é chamada normal padrão. Denotamos a função de distribuição acumulada de uma normal padrão Z por,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}. \tag{2.5}$$

Definição 2.8 (Distribuição Qui-Quadrado)**.** Sejam $Z_1, Z_2, \ldots, Z_n$ variáveis aleatórias independentes com distribuição Normal Padrão. Então a variável aleatória $Q = Z_1^2 + Z_2^2 + \ldots + Z_n^2$ tem distribuição Qui-Quadrado com n graus de liberdade, denotada por $Q \sim \chi_n^2$.

Definição 2.9 (Distribuição t-Student)**.** Sejam Z variável aleatória Normal Padrão e Q_n^2 variável aleatória Qui-Quadrado com n graus de liberdade e independentes. Então a variável aleatória $T = \dfrac{Z}{\sqrt{\frac{Q_n^2}{n}}}$ tem distribuição t-Student com n graus de liberdade e denotamos por $T \sim t_n$.

Outro conceito importante envolvendo variáveis aleatória é o de Valor Esperado ou Esperança, como definimos abaixo.

Definição 2.10 (Esperança para Váriáveis Aleatórias Discretas)**.** Seja X uma variável aleatória discreta com função de probabilidade $p(x)$. Então o valor esperado de X é dado por

$$\mathrm{E}(X) = \sum_x x \cdot p(x). \tag{2.6}$$

Definição 2.11 (Esperança para Váriáveis Aleatória Contínua)**.** Seja X uma variável aleatória contínua com função densidade de probabilidade f, definida para todo real $x \in (-\infty, \infty)$. Então o valor esperado de X é dado por

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)dx. \tag{2.7}$$

Em outras palavras, o valor esperado de uma variável aleatória X é a média ponderada dos valores que X pode assumir, sendo cada valor ponderado pela probabilidade de X assumir tal valor.

Assim como o valor esperado, outra medida muito importante para uma variável aletória é sua variância. Sendo X uma variável aleatória a variância de X nos informa o quão dispersos da média estão os valores que X pode assumir.

Definição 2.12 (Variância)**.** Seja X uma variável aleatória com $\mathrm{E}(X) = \mu$. Então, a variância de X é dada por

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \mathrm{E}(X^2) - \mathrm{E}(X)^2. \tag{2.8}$$

Uma vez que já sabemos a definição de valor esperado, podemos definir a Função Geradora de Momentos. Esta função tem propriedades muito importantes como veremos adiante.

Definição 2.13 (Função Geradora de Momentos)**.** Seja X uma variável aleatória. A função geradora de momentos de X é a função $M_X : \mathbb{R} \to [0, \infty)$ definida por

$$M_X(t) = \mathrm{E}[e^{tX}].$$

A partir das definições (2.10) e (2.11) temos que

$$M_X(t) = \begin{cases} \sum\limits_x e^{tx} p(x), & \text{se X é discreta} \\ \int_{-\infty}^{\infty} e^{tx} f(x)dx, & \text{se X é contínua} \end{cases}. \tag{2.9}$$

A primeira grande propriedade da Função Geradora de Momentos é que ela determina, de modo único, a distribuição de uma variável aleatória, como vemos no teorema a seguir.

Teorema 2.14. *Sejam X e Y variáveis aleatórias com função geradora de momentos $M_X(t)$ e $M_Y(t)$, respectivamente. Então $F_X(z) = F_Y(z)$ para todo $z \in \mathbb{R}$ se, e somente se, $M_X(t) = M_Y(t)$ para todo $t \in (-h, h)$, para algum $h > 0$.*

A demonstração deste teorema pode ser vista em [11] e uma de suas conclusões é que a distribuição da variável aleatória é completamente determinada pela função geradora de momentos e, através dela, podemos obter propriedades que descrevem a distribuição. A existência de $M_X(t)$ implica que existe sua derivada de todas ordens em $t = 0$ no qual vale que $M'_X(0) = E(X)$, $M''_X(0) = E(X^2), \ldots, M_X^{(m)}(0) = E(X^m)$. Chamamos $E(X^m)$ de m-ésimo momento da distribuição e por isso $M_X(t)$ é chamada Função Geradora de Momentos.

Dizemos que duas variáveis aleatórias X e Y são independentes se, e somente se, $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$, ou seja, a ocorrência de X não interfere na ocorrência de Y e vice versa. Neste sentido, temos outra propriedade da função geradora de momentos.

Teorema 2.15. *Sejam X e Y variáveis aleatórias independentes com função geradora de momentos $M_X(t)$ e $M_Y(t)$, respectivamente. Então, $M_{X+Y}(t) = M_X(t)M_Y(t)$.*

Demonstração. De fato,

$$\begin{aligned} M_{X+Y}(t) &= \mathrm{E}[e^{t(X+Y)}] \\ &= \mathrm{E}[e^{t(X)}e^{t(Y)}] \\ &= {}^{3}\,\mathrm{E}[e^{t(X)}]\,\mathrm{E}[e^{t(Y)}] \\ &= M_X(t)M_Y(t). \end{aligned}$$

□

Existem algumas desigualdades que serão importantes para este trabalho. A primeira delas é a Desigualdade de Chebyshev.

Proposição 2.16 (Desigualdade de Chebyshev)**.** *Se X é uma variável aleatória com média finita μ e variância $\sigma^2 < \infty$. Então, para qualquer valor $k > 0$,*

$$\mathbb{P}[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}.$$

A demonstração pode ser vista em [1]. A importância deste resultado está no fato de que conseguimos encontrar uma cota para a probabilidade do desvio da média através da variância.

A Lei Forte dos Grandes Números é, sem dúvida, um dos resultados mais importantes de probabilidade, pois garante que a média amostral converge para a media populacional.

Proposição 2.17 (A Lei Forte dos Grandes Números)**.** *Seja $X_1, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, cada uma com média finita $\mu = \mathrm{E}[X_i]$. Então, com probabilidade 1,*

$$\frac{X_1 + \ldots + X_n}{n} \to \mu$$

quando $n \to \infty$.

Novamente, a demonstração pode ser vista em [1]. O próximo resultado é chamado Desigualdade de Jensen, cuja demonstração pode ser vista em [1] e [3].

[3] Isso é consequência da independência e do Teorema de Fubini.

Proposição 2.18 (Desigualdade de Jensen)**.** *Se f é uma função convexa num intervalo aberto I e X uma variável aleatória com suporte em I e valor esperado finito então*

$$f[E(X)] = \mathrm{E}[f(X)]\cdot$$

Se f é estritamente convexa, a desigualdade é estrita, a menos que X seja uma variável aleatória constante.

A partir de agora, podemos estudar os conceitos relacionados a Estatística que precisaremos mais adiante.

Quando precisamos realizar algum tipo de inferência sobre uma população, é sempre natural obter informações a partir de uma amostra. O mesmo acontece quando temos um parâmetro desconhecido, o estimador para este parâmetro é construído a partir da amostra.

Definição 2.19 (Amostra Aleatória)**.** As variáveis aleatórias $X_1, \ldots, X_n$ são chamadas amostra aleatória se $X_1, \ldots, X_n$ são independentes e possuem a mesma distribuição.

Definição 2.20 (Estimador Pontual ou Estatística)**.** Um estimador pontual T para um parâmetro θ é qualquer função $T = T(X_1, \ldots, X_n)$ da amostra.

Uma vez conhecido o estimador, chamamos de *estimativa* o valor que o estimador fornece ao ser avaliado na amostra. A seguir, veremos duas propriedades interessantes para estimadores, que nos ajudam a decidir se o estimador é bom ou não.

Definição 2.21 (Estimador não-viesado)**.** O estimador T é não viesado para θ se $E(T) = \theta$, para todo θ.

Definição 2.22 (Estimador Consistente)**.** O estimador $T = T(X_1, \ldots, X_n)$ de θ é consistente se

$$\lim_{n\to\infty} \mathrm{E}(T) = \theta$$

e

$$\lim_{n\to\infty} \mathrm{Var}(T) = 0\cdot$$

O exemplo 2.23 mostra que a média amostral é um estimador consistente para a média populacional.

Exemplo 2.23. *Sejam X uma variável aleatória com média μ e variância σ^2 e $X_1, \ldots, X_n$ uma amostra aleatória de X. Definindo a média amostral $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$ temos que,*

$$\mathrm{E}[\overline{X}] = \mathrm{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{\mathrm{E}[X_1] + \ldots + \mathrm{E}[X_n]}{n} = \frac{n\,\mathrm{E}[X_1]}{n} = \mu\cdot$$

Além disso,

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)}{n^2} = \frac{n\,\mathrm{Var}(X_1)}{n^2} = \frac{\sigma^2}{n}.$$

Como $\mathrm{Var}(\overline{X}) \to 0$ *quando* $n \to \infty$ *então* $\overline{X}$ *é um estimador consistente para* μ.

Há outros métodos usados para encontrar o estimador além do método que estudaremos neste trabalho. Informações sobre eles podem ser encontradas em [6].

Como o estimador é uma função da amostra, e esta é obtida aleatoriamente, o estimador sempre será uma variável aleatória. Isso nos leva a sempre buscar qual é a distribuição do estimador. É a partir da distribuição que podemos realizar outros métodos de inferência, como Intervalo de Confiança e Teste de Hipóteses. Esses dois métodos de inferência podem ser vistos com mais detalhes em [7] e [9].

O Intervalo de Confiança é um intervalo numérico que construímos a partir da distribuição do estimador a um nível de confiança $\alpha \in [0, 1]$. Sendo $Q(X; \theta)$ uma variável aleatória cuja distribuição não depende de θ, para cada α fixado encontramos k_1 e k_2 tais que

$$\mathbb{P}(k_1 < Q(X; \theta) < k_2) = \alpha. \tag{2.10}$$

Note que $Q(X; \theta)$ não é uma estatística, pois depende de θ, mas sua distribuição independe de θ. Desse modo, se podemos encontrar $t_1(X), t_2(X)$ tais que

$$\mathbb{P}(t_1(X) < \theta < t_2(X)) = \alpha, \tag{2.11}$$

o intervalo de confiança para θ com nível de confiança α, será dado por

$$IC[\theta, \alpha] = [t_1(X); t_2(X)]. \tag{2.12}$$

E dizemos que, com confiança α, a estimativa de T pertence a $[t_1(X); t_2(X)]$.

Exemplo 2.24. *Suponha que a média amostral* $\overline{X}$ *é tal que* $\overline{X} \sim N(\mu, \sigma^2)$. *Então,* $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. *Logo,*

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = \mathbb{P}\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = \alpha,$$

onde $Z \sim N(0, 1)$ *e* $z_{\alpha/2}$ *é tal que* $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$. *Isso implica que,*

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \alpha.$$

Portanto, $IC[\overline{X}, \alpha] = \left[\overline{X} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}; \overline{X} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right]$.

Outra maneira de realizar inferência sobre o estimador é construíndo Teste de Hipóteses.

Uma hipótese é uma afirmação que fazemos sobre parâmetro e, no teste, assumimos sempre duas hipóteses, a hipótese nula e a hipótese alternativa, denotadas por H_0 e H_a, respectivamente.

Dado um parâmetro θ, a forma mais simples do teste é definir $H_0 : \theta = \theta_0$ e $H_a : \theta \neq \theta_0$, onde θ_0 é um valor real qualquer. Ao tomarmos uma decisão a respeito dessas hipóteses, podemos cometer dois tipos de erros:

Erro do tipo I Rejeitar H_0 quando H_0 é verdade. Neste caso, definimos

$$\alpha = \mathbb{P}(rejeitar\ H_0 | H_0 \text{ é } verdade)\cdot$$

Erro do tipo II Não rejeitar H_0 quando H_0 é falsa. Neste caso, definimos

$$\beta = \mathbb{P}(\text{não } rejeitar\ H_0 | H_0 \text{ é } falsa)\cdot$$

O objetivo do Teste de Hipóteses é decidir, a partir do valor observado de T, se a hipótese nula é ou não aceita, controlando o erro do tipo I. Para isso, precisamos de um critério de decisão, que leva em consideração o que chamamos de região crítica. Se a estimativa de T pertence a região crítica, rejeitamos H_0.

A região crítica, denotada por RC, é a região onde rejeitamos H_0 e é construída de modo que $\alpha = \mathbb{P}(T \in RC | H_0 \text{ é } verdade)$, utilizando a distribuição de T.

Tendo a hipótese nula definida, podemos obter a região crítica a partir da distribuição de T, calcular a estimativa e verificar se pertence ou não a região crítica, concluindo o teste.

3 Convergência de Variáveis Aleatórias

Quando falamos em convergência de sequências, é natural pensar e associar tal conceito com a noção de proximidade. A ideia é a mesma quando estamos interessados em convergência de variáveis aleatórias. Em Probabilidade há mais de um tipo de convergência e neste capítulo estamos interessados em: Convergência em Probabilidade e Convergência em Distribuição. Veremos algumas propriedades que cada uma possui e sua influência em resultados muito importantes em Probabilidade e Estatística, como a Lei dos Grandes Números e o Teorema Central do Limite.

3.1 Convergência em Probabilidade

Definição 3.1. Seja $\{X_n\}$ uma sequência de variáveis aleatórias e X uma variável aleatória definidas num mesmo espaço amostral Ω. Dizemos que $\{X_n\}$ converge em probabilidade para X se $\forall \epsilon > 0$ vale que $\lim_{n\to\infty} \mathbb{P}[|X_n - X| \geq \epsilon] = 0$. E denotamos por $X_n \xrightarrow{P} X$.

Um dos exemplos mais importantes de convergência em probabilidade é a Lei Fraca dos Grandes Números, que garante que a média amostral $\overline{X}_n$ converge para a média populacional μ em probabilidade.

Teorema 3.2 (Lei Fraca dos Grandes Números)**.** *Seja $\{X_n\}$ uma sequência de variáveis aleatórias independentes e identicamente distribuidas com média μ e variância $\sigma^2 < \infty$ e considere a variável aleatória $\overline{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$. Então, $\overline{X}_n \xrightarrow{P} \mu$.*

Demonstração. Como $\mathrm{E}[\overline{X}_n] = \mu$ e $\mathrm{Var}[\overline{X}_n] = \dfrac{\sigma^2}{n}$, então da Proposição 2.16 obtemos que:

$$\mathbb{P}[|\overline{X}_n - \mu| \geq \epsilon] = \mathbb{P}\left[|\overline{X}_n - \mu| \geq \left(\frac{\epsilon\sqrt{n}}{\sigma}\right)\left(\frac{\sigma}{\sqrt{n}}\right)\right] \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ quando } n \to \infty.$$

Logo, $\overline{X}_n \xrightarrow{P} \mu$. □

A partir deste teorema, conseguimos afirmar que à medida que o tamanho da amostra cresce, a média amostral se aproxima da média populacional. Posteriormente, o mesmo resultado foi provado em um sentido mais forte, assumindo apenas que a média μ seja finita. Algumas propriedades sobre a convergência em probabilidade serão mostradas a seguir.

Teorema 3.3. *Sejam $\{X_n\}$ e $\{Y_n\}$ sequências de variáveis aleatórias. Se $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$ então $X_n + Y_n \xrightarrow{P} X + Y$.*

Demonstração. Considere $\epsilon > 0$. Pela desigualdade triangular temos que:

$$|X_n - X| + |Y_n - Y| \geq |(X_n + Y_n) - (X + Y)|.$$

Assim,

$$\begin{aligned}\mathbb{P}[|(X_n + Y_n) - (X + Y)| \geq \epsilon] &\leq \mathbb{P}[|X_n - X| + |Y_n - Y| \geq \epsilon] \\ &\leq \mathbb{P}[|X_n - X| \geq \tfrac{\epsilon}{2}] + \mathbb{P}[|Y_n - Y| \geq \tfrac{\epsilon}{2}]\end{aligned}$$

Como por hipótese $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$, então $X_n + Y_n \xrightarrow{P} X + Y$. □

De modo análogo, podemos mostrar que se $X_n \xrightarrow{P} X$ então $aX_n \xrightarrow{P} aX$, para qualquer constante $a \in \mathbb{R}$.

Teorema 3.4. *Seja $\{X_n\}$ uma sequência de variáveis aleatórias tal que $X_n \xrightarrow{P} a$ e g uma função contínua em $a \in \mathbb{R}$. Então $g(X_n) \xrightarrow{P} g(a)$.*

Demonstração. Seja $\epsilon > 0$. Como g é contínua em a, então existe $\delta > 0$ tal que se $|x-a| < \delta$ então $|g(x) - g(a)| < \epsilon$. Assim, $|g(x) - g(a)| \geq \epsilon$ implica $|x - a| \geq \delta$.

Substituindo x por X_n obtemos da desigualdade anterior que,

$$\mathbb{P}[|g(X_n) - g(a)| \geq \epsilon] \leq \mathbb{P}[|X_n - a| \geq \delta].$$

Como por hipótese, $X_n \xrightarrow{P} a$ concluímos que $g(X_n) \xrightarrow{P} g(a)$. □

O resultado em 3.4 pode ser provado para variáveis aleatórias: se $X_n \xrightarrow{P} X$ e g é uma função contínua então $g(X_n) \xrightarrow{P} g(X)$. Este resultado será usado na demonstração da próxima propriedade.

Teorema 3.5. *Sejam $\{X_n\}$ e $\{Y_n\}$ sequências de variáveis aleatórias. Suponha que $X_n \xrightarrow{P} X$ e $Y_n \xrightarrow{P} Y$. Então $X_nY_n \xrightarrow{P} XY$.*

Demonstração. Note que podemos escrever X_nY_n como $X_nY_n = \frac{1}{2}X_n^2 + \frac{1}{2}Y_n^2 - \frac{1}{2}(X_n - Y_n)^2$. Como $g(y) = y^2$ é uma função contínua $\forall y \in \mathbb{R}$ então:

$$X_nY_n = \tfrac{1}{2}X_n^2 + \tfrac{1}{2}Y_n^2 - \tfrac{1}{2}(X_n - Y_n)^2 \xrightarrow{P} \tfrac{1}{2}X^2 + \tfrac{1}{2}Y^2 - \tfrac{1}{2}(X - Y)^2 = XY.$$ □

Assim como a Lei Fraca dos Grandes Números, há outro resultado importante em Estatística que é reescrito a partir da definição de convergência em probabilidade. Quando estamos numa situação em que temos uma variável aleatória X cuja distribuição tem um parâmetro θ desconhecido, frequentemente estamos interessados em encontrar um bom estimador para tal parâmetro. Neste cenário, definimos consistência.

Definição 3.6. Seja X uma variável aleatória com função distribuição acumulada $F(x, \theta)$, $\theta \in I \subset \mathbb{R}$. Considere $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de X e T_n uma estatística. Dizemos que T_n é um estimador consistente para θ se $T_n \xrightarrow{P} \theta$.

A partir desta definição, vemos pela Lei Fraca dos Grandes Números que $\overline{X}_n$ é um estimador consistente para a média populacional μ.

3.2 Convergência em Distribuição

Diferente da convergência em probabilidade, a convergência em distribuição está diretamente ligada a proximidade da função distribuição acumulada de uma sequência de variáveis aleatórias com a de uma variável aleatória. Nesta seção, veremos que em geral a convergência em distribuição não garante a convergência em probabilidade, mas há um caso especial onde tal propriedade acontece.

Definição 3.7. Considere $\{X_n\}$ sequência de variáveis aleatórias e X uma variável aleatória onde F_{X_n} e F_X são as funções distribuição acumulada de cada X_n e X, respectivamente. Dizemos que $\{X_n\}$ converge em distribuição para X se $\lim_{n\to\infty} F_{X_n} = F_X$, para todo x no conjunto dos pontos onde F_X é contínua. Denotamos $X_n \xrightarrow{D} X$

Enquanto a convergência em probabilidade diz que $\{X_n\}$ está próximo de X com chance alta, a convergência em distribuição diz que F_{X_n} está próxima de F_X. A convergência em distribuição pode acontecer com variáveis aleatórias definidas em diferentes espaços de probabilidade, pois só estamos interessados no contra-domínio onde está definida F_X e F_{X_n}. O próximo exemplo nos mostra que a convergência em distribuição não garante a convergência em probabilidade.

Exemplo 3.8. *Seja X uma variável aleatória com função densidade f_X simétrica, ou seja, $f_X(-x) = f_X(x)$. Neste caso, as variáveis aleatórias X e $-X$ têm mesma distribuição. Considere a sequência de variáveis aleatórias* $X_n = \begin{cases} X, & \text{se } n \text{ é ímpar} \\ -X, & \text{se } n \text{ é par} \end{cases}$.
Claramente $F_{X_n} = F_X$, $\forall x$ ponto de continuidade, ou seja, $X_n \xrightarrow{D} X$. Mas $X_n \xrightarrow{P} X$, pois

$$\mathbb{P}[|X_n - X| \geq \epsilon] = \begin{cases} 0, & \text{se } n \text{ é ímpar.} \\ \mathbb{P}[|-2X| \geq \epsilon], & \text{se } n \text{ é par.} \end{cases}$$

E neste caso, o limite $\lim_{n\to\infty} \mathbb{P}[|X_n - X| \geq \epsilon]$ não existe.

Formalmente, o próximo teorema mostra o contrário do exemplo anterior, que convergência em probabilidade é mais forte que convergência em distribuição.

Teorema 3.9. *Seja $\{X_n\}$ uma sequência de variáveis aleatórias. Se $X_n \xrightarrow{P} X$ então $X_n \xrightarrow{D} X$.*

Demonstração. Seja x um ponto de continuidade de $F_X(x)$. Para todo $\epsilon > 0$, temos que:

$$\begin{aligned} F_{X_n}(x) &= \mathbb{P}[X_n \leq x] \\ &= \mathbb{P}[\{X_n \leq x\} \cap \{|X_n - X| < \epsilon\}] + \mathbb{P}[\{X_n \leq x\} \cap \{|X_n - X| \geq \epsilon\}] \\ &\leq \mathbb{P}[X \leq x] + \mathbb{P}[|X_n - X| \geq \epsilon] \end{aligned}$$

Como $X_n \xrightarrow{P} X$ temos que $\limsup_{n\to\infty} F_{X_n} \leq F_X(x+\epsilon)$.

Procedendo de modo análogo ao passo anterio, mas tomando o complementar obtemos que $\liminf_{n\to\infty} F_{X_n}(x) \geq F_x(x-\epsilon)$. Da relação $\liminf_{n\to\infty} F_{X_n} \leq \limsup_{n\to\infty} F_{X_n}$, obtemos que:

$$F_X(x-\epsilon) \leq \liminf_{n\to\infty} F_{X_n}(x) \leq \limsup_{n\to\infty} F_{X_n}(x) \leq F_X(x+\epsilon)$$

Fazendo $\epsilon \to 0$ e aplicando o teorema do sanduíche, obtemos que $F_{X_n} \to F_X$, ou seja, $X_n \xrightarrow{D} X$. □

Acabamos de ver que, em geral, convergência em distribuição não garante convergência em probabilidade. Veremos no próximo teorema o que acontece quando X é uma variável aleatória degenerada[1].

Teorema 3.10. *Seja $\{X_n\}$ uma sequência de variáveis aleatórias. Se $X_n \xrightarrow{D} b$, então $X_n \xrightarrow{P} b$, onde b é uma constante.*

Demonstração. Dado $\epsilon > 0$, temos que:

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}[|X_n - b| \leq \epsilon] &= \lim_{n\to\infty} \mathbb{P}[b-\epsilon \leq X_n \leq b+\epsilon] \\ &= \lim_{n\to\infty} F_{X_n}(b+\epsilon) - F_{X_n}(b-\epsilon) \\ &= \lim_{n\to\infty} F_b(b+\epsilon) - F_b(b-\epsilon) = 1 - 0 = 1 \end{aligned}$$

Logo, $X_n \xrightarrow{P} b$. □

Além das já citadas anteriormente, a convergência em distribuição possui mais algumas propriedades. Uma delas une as duas definições de convergência vistas até aqui, o Teorema de Slutsky.

Teorema 3.11. *Sejam $\{X_n\}$ e $\{Y_n\}$ sequências de variáveis aleatórias. Suponha que $X_n \xrightarrow{D} X$ e $Y_n \xrightarrow{P} 0$. Então, $X_n + Y_n \xrightarrow{D} X$.*

Teorema 3.12. *Suponha que $X_n \xrightarrow{D} X$ e seja g uma função contínua. Então, $g(X_n) \xrightarrow{D} g(X)$.*

Teorema 3.13 (Teorema de Slutsky)**.** *Sejam $\{X_n\}$, $\{A_n\}$, $\{B_n\}$ sequências de variáveis aleatórias, X variável aleatória e a,b constantes. Se $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$, $B_n \xrightarrow{P} b$ então $A_n + B_n X_n \xrightarrow{D} a + bX$.*

[1] Uma variável aleatória que pode tomar um só valor tem uma distribuição degenerada.

Intuitivamente, podemos ver que a demonstração do Teorema 3.13 vem das propriedades de convergência em probabilidade e convergência em distribuição obtidas dos Teoremas 3.11, 3.12. A demonstração formal pode ser vista em [8].

3.2.1 Convergência via Função Geradora de Momentos

Vimos anteriormente que convergência em distribuição é totalmente caracterizada pela função de distribuição acumulada, mas quando $X_n \xrightarrow{D} X$ é impossível obter a função limite F_X sem conhecer F_{X_n}. Além disso, explicitar o formato de F_{X_n} pode ser muito difícil. Casos como esse são resolvidos utilizando o que chamamos de técnica da função geradora de momentos. Quando existir, a função geradora de momentos correspondente a F_{X_n} é uma maneira conveniente de determinar a função limite F_X. O teorema a seguir, mostra como obter convergência em distribuição através da função geradora de momentos.

Teorema 3.14. *Seja $\{X_n\}$ uma sequência de variáveis aleatórias com função geradora de momentos $M_{X_n}(t)$ que existe para $-h < t < h$, $\forall n$. Seja X uma variável aleatória com função geradora de momentos $M_X(t)$ que existe para $|t| \leq h_1 \leq h$.*

Se $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ para $|t| \leq h_1$ então $X_n \xrightarrow{D} X$.

A demonstração do Teorema 3.14 pode ser visto em [10] na versão para Função Característica. Uma aplicação deste teorema é vista no exemplo abaixo, aproximando a distribuição Poisson pela distribuição de Binomial.

Exemplo 3.15. *Seja $Y_n \sim Bi(n,p)$ com média $\mu = np$ $\forall n \in \mathbb{N}$, ou seja, $p = \frac{\mu}{n}$ onde μ é constante. Queremos encontrar a distribuição limite de Y_n quando $p = \frac{\mu}{n}$ através de $M_{Y_n}(t)$.*

Como $Y_n = X_1 + \ldots + X_n$ onde cada X_i tem distribuição Bernoulli, pelo Teorema 2.15 sabemos que $M_{Y_n}(t) = M_{X_1}(t).\ldots.M_{X_n}(t)$, onde $M_{X_i}(t) = pe^t + (1-p)$ $\forall i = 1, \ldots, n$.

Logo, $M_{Y_n}(t) = [pe^t + (1-p)]^n = \left[1 + \dfrac{\mu(e^t - 1)}{n}\right]^n$.

Fazendo $n \to \infty$ obtemos $M_{Y_n}(t) = e^{\mu(e^t-1)}$, que é a função geradora de momentos de uma variável aleatória $Y \sim Po(\mu)$ onde μ é constante. Pelo Teorema 3.14, $Y_n \xrightarrow{D} Y$.

3.2.2 Teorema Central do Limite

Já sabemos que dada uma sequência de variáveis aleatórias $X_1, \ldots, X_n$ independentes e identicamente distribuídas com distribuição Normal com média μ e variância σ^2, a variável aleatória $Y = \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma/\sqrt{n}}$ é normalmente distribuída com $\mu = 0$ e $\sigma = 1$ para todo valor de n, mas o que podemos afirmar quando não sabemos a distribuição da sequência? Ou se sabemos, podemos aplicar este resultado para outras distribuições?

Em probabilidade há um teorema robusto e elegante chamado Teorema Central do Limite, que garante que dado uma sequência de variáveis aleatórias de qualquer distribuição com média μ e variância σ^2 finita, a variável aleatória Y converge em distribuição para a variável aleatória Z com distribuição normal padrão.

Um outro ponto de vista deste teorema é encontrar a distribuição da média amostral: para n suficientemente grande a variável aleatória $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ tem distribuição aproximadamente normal padrão. Desse modo, podemos utilizar a distribuição normal para calcular probabilidades envolvendo a média amostral $\overline{X}_n$, sendo muito importante ao realizar inferência sobre a média populacional.

Teorema 3.16 (Teorema Central do Limite)**.** *Considere $X_1, \ldots, X_n$ uma amostra aleatória de uma distribuição com média μ e variância σ^2 finita. Então, a variável aleatória $Y_n = \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$ converge em distribuição para uma variável aleatória com distribuição normal padrão.*

Demonstração. Para esta prova, vamos assumir que a função geradora de momentos $M(t) = \mathrm{E}[e^{tX}]$ existe para $-h < t < h$. Neste caso, a função $m(t) = \mathrm{E}[e^{t(X-\mu)}] = e^{-\mu t}M(t)$ também existe para $-h < t < h$. Note que $m(t)$ é a função geradora de momentos da variável aleatória $(X - \mu)$, assim $m(0) = 1$, $m'(0) = \mathrm{E}[(X-\mu)e^0] = \mathrm{E}[X - \mu] = 0$, $m''(0) = \mathrm{E}[(X-\mu)^2] = \sigma^2$. Pela fórmula de Taylor, existe $0 < \epsilon < t$ tal que:

$$m(t) = m(0) + m'(0)t + \frac{m''(\epsilon)t^2}{2} = 1 + \frac{m''(\epsilon)t^2}{2}. \tag{3.1}$$

Somando e subtraindo $\dfrac{\sigma^2 t^2}{2}$ em 3.1, obtemos:

$$m(t) = 1 + \frac{\sigma^2 t^2}{2} + \frac{[m''(\epsilon) - \sigma^2]t^2}{2}. \tag{3.2}$$

Agora, vamos calcular a função geradora de momentos para a variável aleatória $Y_n = \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$.

$$\begin{aligned}
M(t;n) &= \mathrm{E}\left[\exp\left(\frac{t\left(\sum_{i=1}^n X_i - n\mu\right)}{\sigma\sqrt{n}}\right)\right] \\
&= \mathrm{E}\left[\exp\left(\frac{tX_1 - \mu}{\sigma\sqrt{n}}\right) \ldots\ldots \exp\left(\frac{tX_n - \mu}{\sigma\sqrt{n}}\right)\right] \\
&= \left(\mathrm{E}\left[\exp\left(\frac{tX - \mu}{\sigma\sqrt{n}}\right)\right]\right)^n \\
&= \left[m\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.
\end{aligned}$$

Calculando $m\left(\frac{t}{\sigma\sqrt{n}}\right)$, obtemos: $m\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + \frac{[m''(\epsilon) - \sigma^2]t^2}{2\sigma^2 n}$, onde $0 < \epsilon < \frac{t}{\sigma\sqrt{n}}$ e $-h\sigma\sqrt{n} < t < h\sigma\sqrt{n}$.

Logo,

$$M(t;n) = \left(1 + \frac{t^2}{2n} + \frac{[m''(\epsilon) - \sigma^2]t^2}{2\sigma^2 n}\right)^n \tag{3.3}$$

Note que $\epsilon \to 0$ quando $n \to \infty$. Além disso, $m''(t)$ é contínua em $t = 0$. Assim, $\lim_{n\to\infty}[m''(\epsilon) - \sigma^2] = 0$ e, neste caso, $\lim_{n\to\infty} M(t;n) = e^{\frac{t^2}{2}}$.

Portanto, pelo Teorema 3.14, Y_n converge em distribuição para a variável aleatória Z normal padrão. □

Nota histórica: A primeira versão do Teorema Central do Limite foi feita pelo matemático Abraham de Moivre em seu artigo publicado em 1733. Neste artigo, Abraham utilizou a Distribuição Normal para aproximar o número de caras obtidas em muitos lançamentos de uma moeda não viciada. Infelizmente, naquela época a descoberta de Abraham não foi notada. Mais tarde, o matemático Pierre Simon de Laplace tornou um pouco notável tal descoberta através de seu artigo publicado em 1812, onde Laplace encontrou a aproximação da Distribuição Binomial a partir da Distribuição Normal. Apesar disso, somente no final do século XIX, em 1912, o Teorema Central do Limite foi definido e provado matematicamente pelo matemático russo Aleksandr Lyapunov.

3.3 Extensão para Distribuição Multivariada

Nesta seção, vamos generalizar os conceitos de convergência em probabilidade e convergência em distribuição do caso univariado para o caso multivariado, em que $\{\mathbf{X}_n\}$ é uma sequência de vetores p-dimensionais.

Definição 3.17. Sejam $\{\mathbf{X}_n\}$ uma sequência de vetores p-dimensionais e $\mathbf{X}$ um vetor aleatório, todos definidos num mesmo espaço amostral. Dizemos que $\{\mathbf{X}_n\}$ converge em probabilidade para $\mathbf{X}$ se para todo $\epsilon > 0$ vale que $\lim_{n\to\infty} \mathbb{P}[\| \mathbf{X}_n - \mathbf{X} \|] \geq \epsilon] = 0$, e denotamos por $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$.

O próximo teorema mostra que convergência em probabilidade para vetores é equivalente a convergência em probabilidade coordenada a coordenada. A demonstração de tal resultado pode ser vista em [3].

Teorema 3.18. *Sejam $\{\mathbf{X}_n\}$ uma sequência de vetores p-dimensionais e $\mathbf{X}$ um vetor aleatório, todos definidos num mesmo espaço amostral. Então, $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ se, e somente se, $X_{nj} \xrightarrow{P} X_j$, $\forall j = 1, \ldots, p$.*

Através desse teorema, muitos resultados envolvendo convergência em probabilidade podem ser extendidos para o caso multivariado. Por exemplo, sendo $\{\mathbf{X}_n\}$ uma sequência de vetores aleatórios independentes e identicamente distribuídos com média $\boldsymbol{\mu}$ e matriz de variância-covariância[2] dada por $\boldsymbol{\Sigma}$, então o vetor média amostral é definido como $\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i = \left(\overline{\mathbf{X}}_1, \ldots, \overline{\mathbf{X}}_p\right)$. Pela Lei Fraca dos Grandes Números 3.2 e o Teorema 3.18, $\overline{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}$.

Definição 3.19. Considere $\{\mathbf{X}_n\}$ uma sequência de vetores aleatórios e $\mathbf{X}$ um vetor aleatório onde $F_n(\mathbf{x})$ e $F(\mathbf{x})$ são as funções distribuição acumulada de cada $\mathbf{X}_n$ e $\mathbf{X}$, respectivamente. Dizemos que $\{\mathbf{X}_n\}$ converge em distribuição para $\mathbf{X}$ se $\lim_{n\to\infty} F_n(\mathbf{x}) = F(\mathbf{x})$. Denotamos $\mathbf{X}_n \xrightarrow{D} \mathbf{X}$.

Geralmente, é difícil determinar convergência em distribuição pela definição, mas assim como na discussão do caso univariado, podemos utilizar a Técnica da Função Geradora de Momentos no caso multivariado. O teorema abaixo é a versão do Teorema 3.14 para o caso multivariado.

Teorema 3.20. *Seja $\{\mathbf{X}_n\}$ uma sequência de vetores aleatórios com função geradora de momentos $M_n(\mathbf{t})$. Seja $\mathbf{X}$ um vetor aleatório com função geradora de momentos $M(\mathbf{t})$. Então $\{\mathbf{X}_n\}$ converge em distribuição para $\mathbf{X}$ se, e somente se, para algum $h > 0$, $\lim_{n\to\infty} M_n(\mathbf{t}) = M(\mathbf{t})$, para todo $\mathbf{t}$ tal que $\| \mathbf{t} \| < h$.*

A partir dos resultados vistos neste Capítulo, é possível formalizar o Teorema Central do Limite para o caso multivariado.

Teorema 3.21 (Teorema Central do Limite Multivariado)**.** *Considere $\{\mathbf{X}_n\}$ uma sequência de vetores aleatórios independentes e identicamente distribuídos de uma distribuição com média $\boldsymbol{\mu}$ e matriz de variância-covariância $\boldsymbol{\Sigma}$ definida positiva. Assuma que a função geradora de momentos $M(\mathbf{t})$ existe em uma vizinhança aberta de $\mathbf{0}$. Então a variável aleatória $\mathbf{Y}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i - \boldsymbol{\mu}$ converge em distribuição para uma distribuição $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.*

Demonstração. Seja $\mathbf{t} \in \mathbb{R}^p$ um vetor numa vizinhança de $\mathbf{0}$. A função geradora de momentos de $\mathbf{Y}_n$ é:

$$\begin{aligned} M_n(\mathbf{t}) &= \mathrm{E}\left[\exp\left(\mathbf{t}'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i - \boldsymbol{\mu}\right)\right] \\ &= \mathrm{E}\left[\exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{t}'(\mathbf{X}_i - \boldsymbol{\mu})\right)\right] \\ &= \mathrm{E}\left[\exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_i\right)\right] \end{aligned} \tag{3.4}$$

[2] Matriz cujas entradas a_{ii} representam a variância de X_i e a_{ij} para $i \neq j$ a covariância de X_i e X_j.

onde $W_i = \mathbf{t}'(\mathbf{X}_i - \boldsymbol{\mu})$. Mas, note que cada W_i é variável aleatória com média $\boldsymbol{\mu}$ e variância $\text{Var}(W_i) = \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}$, para todo $i = 1, \ldots, n$. Pelo Teorema Central do Limite temos que,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \xrightarrow{D} N(0, \mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}). \tag{3.5}$$

Além disso, a equação 3.4 é a função geradora de momentos de $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i$ em $\mathbf{t} = 1$. Da equação 3.5 concluímos que,

$$M_n(\mathbf{t}) = \text{E}\left[\exp\left((1)\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i\right)\right] \xrightarrow{D} e^{1^2\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2} = e^{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2}$$

onde $e^{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2}$ é a função geradora de momentos da distribuição $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, o que completa a prova.

□

Até aqui, definimos convergência de variáveis aleatórias e vimos um dos teoremas mais importantes em Probabilidade. Extendemos essas definições para o caso multivariado e construímos a base essencial para estudarmos o tema deste trabalho.

Nos próximos capítulos, discutiremos o Método de Máxima Verossimilhança utilizando tudo o que vimos até agora.

4 O Método de Máxima Verossimilhança

Agora que já estudamos os conceitos a cerca da convergência de variáveis aleatórias e os principais teoremas resultantes de tais conceitos, estamos prontos para estudar o tema central deste trabalho. Vamos introduzir o Método de Máxima Verossimilhança, que consiste em encontrar estimadores para um determinado parâmetro de uma distribuição. Após a construção do método, veremos que os Estimadores de Máxima Verossimilhança possuem ótimas propriedades, como consistência e eficiência. Além disso, veremos uma justificativa teórica para se considerar os Estimadores de Máxima Verossimilhança. A formalização do método será feita para o caso contínuo, com a função densidade de probabilidade, mas é análogo para o caso discreto com a função distribuição de probabilidade como veremos em alguns exemplos.

Considere uma variável aleatória X com função densidade de probabilidade $f(x;\theta)$, onde $\theta \in I \subset \mathbb{R}$ é um parâmetro desconhecido. O objetivo é realizar inferência sobre o parâmetro θ, ou seja, queremos encontrar um estimador para tal parâmetro, para isso, considere $X_1, X_2, \ldots, X_n$ uma amostra aleatória de X. Então a função de probabilidade conjunta desta amostra aleatória é dada por:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{4.1}$$

A função 4.1 é a chamada função de verossimilhança, e será a função utilizada para a estimação do parâmetro θ. Observe que na função de verossimilhança, fixamos o valor da amostra e a função de probabilidade conjunta se torna uma função de θ e não mais uma densidade.

Definição 4.1 (Função de Verossimilhança)**.** Considere a variável aleatória X com função densidade de probabilidade $f(x,\theta)$, onde $\theta \in I \subset \mathbb{R}$ é um parâmetro desconhecido. Suponha que $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de X. Então, a função de verossimilhança $L(\theta, \mathbf{x})$, ou apenas $L(\theta)$, é dada por:

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) \tag{4.2}$$

onde $\theta \in I \subset \mathbb{R}$ e $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

O ponto $\hat{\theta}$ que maximiza a função de verossimilhança será o estimador de máxima verossimilhança para θ. Posteriormente, justificaremos através de um teorema o motivo de se tomar o máximo da função de verossimilhança.

Matematicamente, será conveniente tomar o log de L para cálculos e análises futuras, obtendo:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta) \cdot \tag{4.3}$$

Observe que não há problema em utilizar $l(\theta)$, pois $\log(x)$ é uma função bijetora e crescente, portanto terá um máximo. Dito isto, podemos definir o Estimador de Máxima Verossimilhança.

Definição 4.2 (Estimador de Máxima Verossimilhança)**.** Dizemos que $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X})$ é um estimador de máxima verossimilhança de θ se $\hat{\theta}_n = \arg\max L(\theta, \mathbf{X})$.

Dizer que $\hat{\theta}_n$ satisfaz $\hat{\theta}_n = \arg\max L(\theta, \mathbf{X})$ significa que $L(\theta, \mathbf{X})$ atinge seu máximo em $\hat{\theta}_n$, ou seja, $\hat{\theta}_n$ resolve a equação,

$$\frac{\partial l(\theta)}{\partial \theta} = 0 \cdot \tag{4.4}$$

A Equação (4.4) é chamada equação de estimação e do ponto de vista matemático, encontrar a solução desta equação é apenas um problema de Cálculo Diferencial. Mas há casos em que no modelo em questão não conseguimos explicitar soluções analíticas para esta equação, ou seja, não conseguimos encontrar analiticamente os estimadores de máxima verossimilhança. Quando isso acontece, recorremos a métodos computacionais para a aproximação.

No exemplo abaixo, encontramos a equação de estimação para a Distribuição Binomial e explicitamos o estimador de Máxima Verossimilhança.

Exemplo 4.3 (Distribuição Binomial)**.** *Sejam X uma variável aleatória com distribuição Binomial e θ, com $0 < \theta < 1$, a probabilidade de sucesso. A função distribuição de probabilidade de X é dada por $p(x; \theta) = \theta^x (1-\theta)^{1-x}$, onde $x = 0$ ou $x = 1$.*

Se $X_1, \ldots, X_n$ é uma amostra aleatória de X então a função de máxima verossimilhança é,

$$L(\theta) = \prod_{i=1}^{n} p(x_i; \theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}.$$

Aplicando o log, obtemos:

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

Derivando $l(\theta)$ obtemos a equação de estimação:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta}. \tag{4.5}$$

Igualando a derivada em (4.5) a zero e resolvendo para θ obtemos $\hat{\theta}(x_1, \ldots, x_n) = n^{-1}\sum_{i=1}^{n} x_i$, ou seja, o estimador de máxima verossimilhança para X, dado por $\hat{\theta}(X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^{n} X_i$, é a proporção de sucessos em n tentativas. Como $\mathrm{E}[\hat{\theta}] = \theta$, $\hat{\theta}$ é um estimador não viesado para θ.

O estimador de máxima verossimilhança nem sempre existe, e quando existe nem sempre é único. O exemplo abaixo mostra que o Estimador de Máxima Verossimilhança para a Distribuição Logística existe e é único.

Exemplo 4.4 (Distribuição Logística)**.** *Sejam $X_1, \ldots, X_n$ independentes e identicamente distribuídas com função densidade de probabilidade dada por :*

$$f(x;\theta) = \frac{\exp\{-(x-\theta)\}}{(1+\exp\{-(x-\theta)\})^2}, \quad -\infty < \theta < \infty, -\infty < x < \infty \tag{4.6}$$

O log da função de verossimilhança nos dá,

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta) = n\theta - n\overline{X} - 2\sum_{i=1}^{n} log(1+\exp\{-(x_i-\theta)\}) \tag{4.7}$$

e derivando esta equação em relação a θ obtemos:

$$l'(\theta) = n - 2\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{1+\exp\{-(x_i-\theta)\}}. \tag{4.8}$$

Observe que a segunda derivada de $l(\theta)$ será estritamente negativa, isso significa que a solução da equação de estimação $l'(\theta) = 0$ é um ponto de máximo. Resta ver que a solução é unica. Igualando a equação 4.8 a zero e reorganizando os termos, obtemos:

$$\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{1+\exp\{-(x_i-\theta)\}} = \frac{n}{2} \tag{4.9}$$

Derivando em relação a θ o lado esquerdo da equação 4.9 obtemos,

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{1+\exp\{-(x_i-\theta)\}} = \sum_{i=1}^{n} \frac{\exp\{-(x_i-\theta)\}}{(1+\exp\{-(x_i-\theta)\})^2} > 0 \tag{4.10}$$

Assim, o lado esquerdo da equação 4.9 é estritamente crescente, se aproxima de 0 a medida que $\theta \to -\infty$ e se aproxima de n a medida que $\theta \to \infty$. Logo, a equação 4.9 tem solução única.

O próximo teorema justifica o motivo de se considerar como estimador de máxima verossimilhança o ponto de máximo da função de verossimilhança, mas antes de estudarmos o teorema precisamos definir algumas hipóteses, chamadas Condições de Regularidade. Para a definição a seguir, considere θ_0 como o valor verdadeiro de θ.

Definição 4.5 (Condições de Regularidade)**.** As condições de Regularidade $R0 - R2$ são dadas a seguir.

R0 As funções densidade de probabilidade são distintas; ou seja, $\theta \neq \theta' \Rightarrow f(x,\theta) \neq f(x,\theta')$.

R1 As funções densidade de probabilidade têm mesmo conjunto suporte para todo θ;

R2 O ponto θ_0 é ponto interior de $I \subset \mathbb{R}$;

A condição (**R0**) diz que o parâmetro identifica a função densidade de probabilidade. A condição (**R1**) diz que o conjunto suporte não depende do parâmetro, e a última condição diz que podemos aproximar θ_0 por pontos em I.

Teorema 4.6. *Seja θ_0 o valor verdadeiro de θ. Sob as condições $R0 - R2$ vale que,*

$$\lim_{n\to\infty} \mathbb{P}[L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})] = 1 \quad \forall\theta \neq \theta_0 \ .$$

Demonstração. Considere a desigualdade $L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})$. Aplicando log nesta desigualdade, obtemos:

$$\begin{aligned} L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X}) &\Leftrightarrow \log L(\theta_0, \mathbf{X}) > \log L(\theta, \mathbf{X}) \\ &\Leftrightarrow \log \prod_{i=1}^{n} f(X_i, \theta_0) > \prod_{i=1}^{n} f(X_i, \theta). \\ &\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta_0) > \frac{1}{n}\sum_{i=1}^{n} \log f(X_i, \theta) \\ &\Leftrightarrow \frac{1}{n}\left[\sum_{i=1}^{n} \log f(X_i, \theta) - \sum_{i=1}^{n} \log f(X_i, \theta_0)\right] < 0 \\ &\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \theta)}{f(X_i, \theta_0)} < 0. \end{aligned}$$

Assim, $\mathbb{P}[L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})] = \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \theta)}{f(X_i, \theta_0)} < 0\right]$. Como $X_1, \ldots, X_n$ são independentes e identicamente distribuídas, segue da Lei dos Grandes Números que, quando θ_0 é o valor verdadeiro do parâmetro,

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i, \theta)}{f(X_i, \theta_0)} \xrightarrow{P} E_{\theta_0}\left[\log \frac{f(X_1, \theta)}{f(X_1, \theta_0)}\right]. \tag{4.11}$$

Além disso, pela Proposição 2.18, temos que:

$$E_{\theta_0}\left[\log \frac{f(X_1, \theta)}{f(X_1, \theta_0)}\right] < \log E_{\theta_0}\left[\frac{f(X_1, \theta)}{f(X_1, \theta_0)}\right].$$

Mas, $E_{\theta_0}\left[\frac{f(X_1, \theta)}{f(X_1, \theta_0)}\right] = \int \frac{f(x, \theta)}{f(x, \theta_0)} f(x, \theta_0) dx = \int f(x, \theta) dx = 1.$

Como $\log 1 = 0$, segue da Equação (4.11), $\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i,\theta_0)} \xrightarrow{P} a < 0$. Isso significa que, sendo $Y_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i,\theta_0)}$ e ϵ fixado,

$$1 = \lim_{n\to\infty} \mathbb{P}[|Y_n - a| < \epsilon] \leq \lim_{n\to\infty} \mathbb{P}[Y_n < a - \epsilon] = \mathbb{P}[Y_n < 0].$$

Logo,

$$\lim_{n\to\infty} \mathbb{P}[L(\theta_0, \mathbf{X}) > L(\theta, \mathbf{X})] = \lim_{n\to\infty} \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i,\theta)}{f(X_i,\theta_0)} < 0\right] = 1.$$

□

O teorema provado anteriormente diz que para qualquer $\theta \neq \theta_0$ a curva da função de verossimilhança em θ_0 está acima da curva da função de verosimilhança para qualquer outro θ com chance alta, isso significa que ao buscar um estimador para θ seria natural considerar o que maximiza a função de verossimilhança, pois o máximo da função de verossimilhança é a estimativa mais próxima do valor verdadeiro do parâmetro.

Outro resultado interessante envolvendo os estimadores de máxima verossimilhança é dado a seguir. Nele veremos que estes estimadores são consistentes.

Teorema 4.7. *Suponha que $X_1, X_2, \ldots, X_n$ satisfazem as condições de regularidade $R0 - R2$ com $f(x,\theta)$ duas vezes diferenciável com respeito a $\theta \in I \subset \mathbb{R}$ e considere θ_0 o valor verdadeiro de θ. Então, a equação de verossimilhança*

$$\frac{\partial L(\theta)}{\partial \theta} = 0,$$

ou equivalentemente

$$\frac{\partial l(\theta)}{\partial \theta} = 0,$$

tem solução $\hat{\theta}_n$ de modo que $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Demonstração. Como θ_0 é um ponto interior de I então para algum $a > 0$, $(\theta_0 - a, \theta_0 + a) \subset I$. Considere o evento,

$$S_n = \{X : l(\theta_0, X) > l(\theta_0 - a, X)\} \cap \{X : l(\theta_0, X) > l(\theta_0 + a, X)\}.$$

Pelo Teorema 4.6, temos que $\mathbb{P}(S_n) \to 1$ a medida que $n \to \infty$. Além disso, em S_n a função $l(\theta)$ tem máximo local, que chamaremos de $\hat{\theta}_n$ e $\theta_0 - a < \hat{\theta}_n < \theta_0 + a$ e $l'(\hat{\theta}_n) = 0$.

Assim, $S_n \subset \left\{X : |\hat{\theta}_n - \theta_0| < a\right\} \cap \left\{X : l'(\hat{\theta}_n(X)) = 0\right\}$.

Logo,

$$1 = \lim_{n\to\infty} \mathbb{P}(S_n) \leq \limsup_{n\to\infty} \mathbb{P}\left[\left\{X : |\hat{\theta}_n - \theta_0| < a\right\} \cap \left\{X : l'(\hat{\theta}_n(X)) = 0\right\}\right] \leq 1.$$

Portanto, a medida que $n \to \infty$, $\mathbb{P}\left[|\hat{\theta}_n - \theta_0| < a\right] \to 1$, como queríamos demonstrar.

□

Suponha agora que o parâmetro que se tem interesse em estimar não é θ, mas uma função de θ. O que se pode esperar do estimador de máxima verossimilhança para uma função do parâmetro? Além da consistência, há uma outra propriedade muito interessante que ocorre quando o parâmetro a ser estimado é uma função de θ.

Teorema 4.8. *Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas com função densidade de probabilidade $f(x, \theta)$ com $\theta \in I \subset \mathbb{R}$. Para uma função g bijetiva, seja $\eta = g(\theta)$ o parâmetro de interesse. Suponha que $\hat{\theta}$ é o estimador de máxima verossimilhança de θ. Então, $g(\hat{\theta})$ é o estimador de máxima verissimilhança para $\eta = g(\theta)$.*

Demonstração. Suponha $g : I \to B \subset \mathbb{R}^n$ uma função bijetiva. Sendo $\hat{\theta}$ o estimador de Máxima Verossimilhança para θ temos que $\hat{\theta} \in I$ e $L(\hat{\theta}, \mathbf{X}) > L(\theta, \mathbf{X})\ \forall \theta \in I$.

Como g é bijetora, temos que $\hat{\eta} = g(\hat{\theta}) \in B$ e $L(g^{-1}(\hat{\eta}), \mathbf{X}) > L(g^{-1}(\eta), \mathbf{X})\ \forall \eta \in B$.

Logo, $\hat{\eta} = g(\hat{\theta})$ é o estimador de máxima verissimilhança para $\eta = g(\theta)$.

□

Até aqui, estudamos todo o Método de Máxima Verossimilhança, bem como suas propriedades e justificativa teórica para tal abordagem. Mas existem outras informações necessárias ao realizar inferência sobre algum parâmetro, como variância, valor esperado e distribuição desse estimador. Essas informações são muito importantes para realizar inferência e modelar problemas envolvendo estimadores e variáveis aleatórias. No capítulo seguinte, discutiremos propriedades assintóticas, distribuição dos Estimadores de Máxima Verossimilhança e o principal resultado deste trabalho, o Limite Inferior de Rao-Cramér.

5 O Limite Inferior de Rao-Cramér

Nesta seção, trataremos um dos principais resultados em Inferência que envolve diretamente os Estimadores de Máxima Verosssimilhança, o Limite Inferior de Rao-Cramér, utilizando como principal referência [3].

A origem do Limite Inferior de Rao-Cramér foi publicada pelo jornal Times da Índia, numa matéria intitulada "As dez maiores contribuições para a ciência indiana" e conta que, aos 24 anos de idade Calyampudi Radhakrishna Rao (atualmente com 98 anos) estava lecionando sobre Estimadores aos alunos do mestrado da Universidade de Calcutá quando provou o limite inferior para um caso particular. Mas um de seus alunos o questionou dizendo: "por que você não prova o mesmo resultado para o caso geral?". Neste mesmo dia, Rao foi pra casa e trabalhou a noite toda, e no dia seguinte estava provado o que hoje conhecemos como O Limite Inferior de Rao-Cramér. O mais interessante é que nesta mesma época, H. Cramér (1893-1985) também estava publicando este mesmo resultado, em 1946. Por isso o resultado se chama O Limite Inferior de Rao-Cramér. Além de Cramér e Rao, haviam outros dois matemáticos francêses Georges Darmois(1888- 1960) e Maurice René Fréchet (1878 - 1973) que também reportaram este resultado para os casos uniparamétrico e multiparamétrico, respectivamente. Desse modo, este resultado as vezes é chamado de O Limite Inferior de Cramér-Rao-Fréchet-Darmois.

Para discutir o resultado acima, vamos começar adicionando mais duas condições de regularidade àquelas vistas no Capítulo 4.

Considere X uma variável aleatória com função densidade de probabilidade $f(x;\theta)$, com $\theta \in I \subset \mathbb{R}$ e seja o espaço do parâmetro θ um intervalo aberto.

Definição 5.1 (Condições de Regularidade)**.** As condições de Regularidade $R3 - R4$ são dadas a seguir.

R3 A função densidade de probabilidade é duas vezes diferenciável como função de θ.

R4 A integral $\int f(x;\theta)dx$ pode ser duas vezes diferenciável sobre o sinal da integral como função de θ.

Observe que todas as condições de regularidade definidas até aqui, mostra-nos que o parâmetro θ nunca está nas extremidades onde $f(x;\theta) = 0$. O próximo passo agora é encontrar uma expressão para a variância do estimador de máxima verossimilhança. Vamos começar nossa análise com a identidade abaixo:

$$1 = \int_{-\infty}^{\infty} f(x;\theta)dx.$$

Derivando com relação a θ obtemos a expressão

$$0 = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta} dx.$$

A expressão anterior pode ser reescrita como

$$0 = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)/\partial\theta}{f(x;\theta)} f(x;\theta)dx = \int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx. \tag{5.1}$$

Observe que o lado direito da Equação 5.1 pode ser visto como o valor esperado da variável aleatória $\dfrac{\partial \log f(x;\theta)}{\partial \theta}$ e pelo lado esquerdo da igualdade concluímos que

$$\mathrm{E}\left[\frac{\partial \log f(x;\theta)}{\partial \theta}\right] = 0. \tag{5.2}$$

Derivando a Equação 5.1 novamente em relação a θ obtemos

$$0 = \int_{-\infty}^{\infty} \left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta} \cdot \frac{\partial f(x;\theta)}{\partial \theta}\right] dx \tag{5.3}$$

$$= \int_{-\infty}^{\infty} \left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta) + \frac{\partial \log f(x;\theta)}{\partial \theta} \cdot \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\right] dx \tag{5.4}$$

$$= \int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)dx + \int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial \theta} \cdot \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx, \tag{5.5}$$

ou equivalentemente,

$$-\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)dx = \int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial \theta} \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx. \tag{5.6}$$

Novamente, podemos notar que o segundo termo do lado direito da equação 5.5 pode ser visto como um valor esperado, desta vez da variável aleatória $\left(\dfrac{\partial \log f(x;\theta)}{\partial \theta}\right)^2$, no qual chamamos de Informação de Fisher e denotamos por $I(\theta)$. Logo,

$$I(\theta) = \int_{-\infty}^{\infty} \frac{\partial \log f(x;\theta)}{\partial \theta} \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx = \mathrm{E}\left[\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2\right]. \tag{5.7}$$

Além disso, pela Equação 5.6 há outra forma de definir a Informação de Fisher,

$$I(\theta) = -\int_{-\infty}^{\infty} \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)dx = -\mathrm{E}\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right]. \tag{5.8}$$

Portanto, da equação 5.2, podemos concluir que a variância da variável aleatória $\dfrac{\partial \log f(x;\theta)}{\partial \theta}$ que é usada na estimação pelo método de máxima verossimilhança é a Informação de Fisher, ou seja,

$$I(\theta) = \mathrm{Var}\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right).$$

Exemplo 5.2. *Seja X uma variável aleatória com distribuição Bernoulli. Então a função de distribuição de probabilidade é $f(x;\theta) = p(x;\theta) = \theta^x(1-\theta)^{1-x}$, onde θ é o parâmetro que representa a chance de sucesso. Como $\log f(x;\theta) = x\log\theta + (1-x)\log(1-\theta)$, então*

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

Logo, como $\mathrm{E}(X) = \theta$,

$$\begin{aligned} I(\theta) &= -\,\mathrm{E}\left[\frac{-X}{\theta^2} - \frac{(1-X)}{(1-\theta)^2}\right] = -\frac{1}{\theta^2}\,\mathrm{E}(-X) + \frac{1}{(1-\theta)^2}\,\mathrm{E}(1-X) \\ &= \frac{\theta}{\theta^2} + \frac{(1-\theta)}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{(1-\theta)} = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

Portanto, a Informação de Fisher para uma variável aleatória com distribuição Bernoulli é $I(\theta) = \dfrac{1}{\theta(1-\theta)}$, que é grande para valores de θ próximos de zero.

Tanto na explicação anterior quanto no Exemplo 5.2, vimos a Informação de Fisher para o caso onde o tamanho da amostra era $n = 1$. Como podemos interpretar a Informação de Fisher para uma amostra de tamanho $n > 1$?

Suponha $X_1, \ldots, X_n$ independentes e identicamente distribuídas com distribuição $f(x;\theta)$. Então, a variável aleatória cuja variância é a Informação de Fisher da amostra é,

$$\frac{\partial \log L(\mathbf{X};\theta)}{\partial\theta} = \frac{\partial \log \prod_{i=1}^{n} f(X_i,\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta)}{\partial\theta},$$

onde, $L(\theta)$ é a função de verossimilhança.

Como as variáveis aleatórias são independentes e têm mesma variância, podemos escrever a Informação de Fisher da amostra como,

$$\mathrm{Var}\left(\frac{\partial \log L(\mathbf{X};\theta)}{\partial\theta}\right) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta)}{\partial\theta}\right) = nI(\theta).$$

Logo, a Informação de Fisher para uma amostra de variáveis aleatórias independentes e identicamente distribuídas de tamanho $n > 1$ é n vezes a Informação de Fisher de uma amostra de tamanho $n = 1$.

Desse modo, a partir do Exemplo 5.2 podemos concluir que se X tem distribuição Binomial, a Informação de Fisher é $I(\theta) = \dfrac{n}{\theta(1-\theta)}$.

Uma vez caracterizada a Informação de Fisher, podemos discutir o principal teorema deste trabalho. Este teorema nos garante que, sob certas condições, a variância de uma estatística é limitada inferiormente e este limite inferior depende apenas do valor esperado da estatística.

Teorema 5.3 (Limite Inferior de Rao-Cramér)**.** *Sejam $X_1, \ldots, X_n$ independentes e identicamente distribuídas com distribuição $f(x;\theta)$ para $\theta \in I \subset \mathbb{R}$. Assumimos que as condições de regularidade $R0 - R4$ são válidas. Considere $Y = \mu(X_1, \ldots, X_n)$ uma estatística com média $\mathrm{E}(Y) = \mathrm{E}[\mu(X_1, \ldots, X_n)] = k(\theta)$. Então*

$$\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

Demonstração. Vamos provar o caso contínuo, mas a prova para o caso discreto segue os mesmos passos. Sendo Y a estatística, podemos escrever seu valor esperado $k(\theta)$ como

$$k(\theta) = \int_{\mathbb{R}^n} \mu(x_i, \ldots, x_n) f(x_1;\theta) \ldots f(x_n;\theta) dx_1 \ldots dx_n.$$

Derivando em relação a θ, obtemos:

$$\begin{aligned} k(\theta) &= \int_{\mathbb{R}^n} \mu(x_1, \ldots, x_n) \left[f(x_1;\theta) \ldots f(x_n;\theta)\right]' dx_1 \ldots dx_n \\ &= \int_{\mathbb{R}^n} \mu(x_i, \ldots, x_n) \left[\sum_{i=1}^{n} \frac{\partial f(x_i;\theta)}{\partial\theta} \frac{1}{f(x;\theta)}\right] \times f(x_1;\theta) \ldots f(x_n;\theta) dx_1 \ldots dx_n \\ &= \int_{\mathbb{R}^n} \mu(x_i, \ldots, x_n) \left[\sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial\theta}\right] \times f(x_1;\theta) \ldots f(x_n;\theta) dx_1 \ldots dx_n. \end{aligned} \tag{5.9}$$

Considere a variável aleatória $Z = \sum_{i=1}^{n} \frac{\partial \log f(x_i;\theta)}{\partial\theta}$. Vimos que $\mathrm{E}(Z) = 0$ e $\mathrm{Var}(Z) = nI(\theta)$. Assim, a equação 5.9 pode ser reescrita como o valor esperado

$$k'(\theta) = \mathrm{E}(YZ) = \mathrm{E}(Y)\,\mathrm{E}(Z) + \rho\sigma_Y\sigma_Z, \tag{5.10}$$

onde $\rho\sigma_Y\sigma_Z = \mathrm{Cov}(Y, Z)$, $\sigma_Z = \sqrt{nI(\theta)}$ e ρ é o coeficiente de correlação entre Y e Z. Como $\mathrm{E}(Z) = 0$ a Equação 5.10 pode ser reescrita como,

$$k'(\theta) = \rho\sigma_Y\sigma_Z \Rightarrow \rho = \frac{k'(\theta)}{\sigma_Y\sqrt{nI(\theta)}}.$$

Além disso, $-1 < \rho < 1$, ou seja, $\rho^2 \leq 1$. Logo,

$$\frac{[k'(\theta)]^2}{\sigma_Y^2 nI(\theta)} \leq 1 \Rightarrow \sigma_Y^2 \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

Portanto, $\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}$ como queríamos mostrar.

□

Analisando o Exemplo 5.2, vemos que se X tem distribuição Bernoulli onde θ é o parâmetro que representa a chance de sucesso, então $\frac{1}{nI(\theta)} = \frac{(\theta(1-\theta))}{n}$. Além disso, no Exemplo 4.3 vimos que o estimador de máxima verossimilhança para θ era $\overline{X}$. Como

$\mathrm{E}(X) = \theta$ e $\mathrm{Var}(X) = \theta(1-\theta)$ então o $\mathrm{E}(\overline{X}) = \theta$ e $\mathrm{Var}(\overline{X}) = \dfrac{\theta(1-\theta)}{n}$ e portanto, o estimador de máxima verossimilhança para θ na distribuição Bernoulli satisfaz o Limite Inferior de Rao-Cramér, pois $\dfrac{\theta(1-\theta)}{n} \geq \dfrac{1}{nI(\theta)}$.

Existe um corolário do Teorema 5.3 para o caso onde o estimador é não-viesado.

Corolário 5.4. *Sob as condições* $(R0) - (R4)$*, se* $Y = \mu(X_1, \cdots, X_n)$ *é um estimador não-viesado para* θ *então o Limite Inferior de rao-Cramér se torna*

$$Var(Y) \geq \frac{1}{nI(\theta)}. \tag{5.11}$$

A partir do Teorema 5.3, encontramos uma forma de verificar se o estimador de máxima verossimilhaça é eficiente ou não, uma propriedade muito importante de estimadores não viesados. Além disso, ainda podemos caracterizar a eficiência do estimador a partir do Limite Inferior de Rao-Cramér. Mais uma vez, isso mostra o quão importante é tal resultado.

Definição 5.5 (Eficiência)**.** Nas condições onde se pode diferenciar com respeito a um parâmetro sob o sinal da integral ou de uma soma, dizemos que a razão entre o limite inferior de Rao-Cramér e a variância real de qualquer estimador não viesado é a eficiência do estimador.

Definição 5.6 (Estimador Eficiente)**.** Seja Y um estimador não viesado para θ. Dizemos que Y é um estimador eficiente se, e somente se, a variância de Y atinge o limite inferior de Rao-Cramér, isto é,

$$\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)},$$

onde $k(\theta)$ é o valor esperado de Y.

Os estimadores de máxima verossimilhança têm distribuição Normal assintoticamente, com variância dada pela Informação de Fisher. Para mostrarmos este grande e importante resultado, precisamos definir mais uma condição de regularidade e dois outros resultados envolvendo convergência em probabilidade, como veremos a seguir.

Definição 5.7 (Condições de Regularidade)**.** A condição de Regularidade R5 é dada a seguir.

R5 A função densidade de probabilidade é três vezes diferenciável como função de θ. Além disso, $\forall\theta \in I \subset \mathbb{R}$ existe uma constante c e uma função $M(x)$ tal que,

$$\frac{\partial^3 \log f(x;\theta)}{\partial\theta^3} \leq M(x),$$

com $E[M(X)] < \infty$, $\forall\theta$ com $|\theta - \theta_0| < c$ e $\forall x$ no suporte de X.

Definição 5.8. Dizemos que a sequência de variáveis aleatórias $\{X_n\}$ é limitada em probabilidade se, $\forall \epsilon > 0$, existe umas constante $B_\epsilon > 0$ e um inteiro N_ϵ tal que,

$$n \geq N_\epsilon \Rightarrow \mathbb{P}[|X_n| \leq B_\epsilon] \geq 1 - \epsilon.$$

Teorema 5.9. *Sejam $\{X_n\}$ uma sequência de variáveis aleatórias limitada em probabilidade e $\{Y_n\}$ uma sequência de variáveis aleatórias que converge em probabilidade para zero. Então,*

$$X_n Y_n \xrightarrow{P} 0.$$

A demonstração deste teorema pode ser vista em [3].

Finalmente, apresentaremos o teorema que fornece a distribuição assintótica dos estimadores de máxima verossimilhança.

Teorema 5.10. *Considere $X_1, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas com função densidade de probabilidade $f(x;\theta)$, $\theta \in I \subset \mathbb{R}$ tais que as condições $R0 - R5$ são válidas. Suponha que a informação de Fisher satisfaz $0 < I(\theta) < \infty$. Então, qualquer sequência de soluções da equação de estimação satisfaz,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \frac{1}{I(\theta)}\right).$$

Demonstração. Vamos expandir a função $l'(\theta)$ em Série de Taylor de ordem 2 em torno de θ_0 e avaliar em $\hat{\theta}$.

$$l'(\hat{\theta}) = l'(\theta_0) + (\hat{\theta} - \theta_0) l''(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2 l'''(\theta_n^*), \tag{5.12}$$

onde $\theta_0 < \theta_n^* < \hat{\theta}$. Como $l'(\hat{\theta}) = 0$, então reorganizando os termos da Equação (5.12) obtemos:

$$\begin{aligned} l'(\theta_0) &= -(\hat{\theta} - \theta_0) l''(\theta_0) - \frac{1}{2}(\hat{\theta} - \theta_0)^2 l'''(\theta_n^*) \\ &= (\hat{\theta} - \theta_0)\left[-l''(\theta_0) - \frac{1}{2} l'''(\theta_n^*)(\hat{\theta} - \theta_0)\right] \end{aligned} \tag{5.13}$$

Logo,

$$(\hat{\theta} - \theta_0) = \frac{l'(\theta_0)}{-l''(\theta_0) - \frac{1}{2} l'''(\theta_n^*)(\hat{\theta} - \theta_0)}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{n^{-1/2} l'(\theta_0)}{-n^{-1} l''(\theta_0) - (2n)^{-1} l'''(\theta_n^*)(\hat{\theta} - \theta_0)}. \tag{5.14}$$

Como $X_1, \ldots, X_n$ são independentes, pelo Teorema Central do Limite 3.16, temos que

$$n^{-1/2} l'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta_0)}{\partial \theta} \xrightarrow{D} N\left(0, nI(\theta_0)\right).$$

Além disso, pela Lei dos Grandes Números,

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta_0)}{\partial\theta^2} \xrightarrow{P} I(\theta_0).$$

Para completar a prova, só precisamos mostrar que $-(2n)^{-1}l'''(\theta_0)(\hat{\theta}-\theta_0) \xrightarrow{P} 0$.

Note que como $(\hat{\theta}-\theta_0) \xrightarrow{P} 0$, só precisamos mostrar que $n^{-1}l'''(\theta_n^*)$ é limitado em probabilidade e aplicar o Teorema 5.9.

Denote por c_0 a constante definida na condição $(R5)$. Observe que $|\hat{\theta}-\theta_0| < c_0 \Rightarrow |\theta_n^*-\theta_0| < c_0$, que por sua vez, pela condição $(R5)$, implica que

$$\left|-\frac{1}{n}l'''(\theta_n^*)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}\right| \leq \frac{1}{n}\sum_{i=1}^{n}M(X_i) \tag{5.15}$$

e, como $E_{\theta_0}[M(X_i)] < \infty \ \forall i = 1,\ldots,n$, então pela Lei dos Grandes Números,

$$\frac{1}{n}\sum_{i=1}^{n}M(X_i) \xrightarrow{P} E_{\theta_0}[M(X)].$$

Seja $\epsilon > 0$ e escolha N_1, N_2 tais que:

$$n \geq N_1 \Rightarrow \mathbb{P}\left[|\hat{\theta}-\theta_0| < c_0\right] \geq 1-\frac{\epsilon}{2}, \tag{5.16}$$

$$n \geq N_2 \Rightarrow \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n}M(X_i) - E_{\theta_0}[M(X)]\right| < 1\right] \geq 1-\frac{\epsilon}{2}. \tag{5.17}$$

Segue das Equações 5.15 e 5.17 que,

$$n \geq \max\{N_1,N_2\} \Rightarrow \mathbb{P}\left[\left|-\frac{1}{n}l'''(\theta_n^*)\right| \leq 1 + E_{\theta_0}[M(X)]\right] \geq 1-\frac{\epsilon}{2},$$

e isso é justamente a Definição 5.8.

Logo, pelo Teorema 5.9, $-(2n)^{-1}l'''(\theta_0)(\hat{\theta}-\theta_0) \xrightarrow{P} 0$ e, portanto, concluímos da Equação 5.14 que $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{P} N\left(0, \frac{1}{I(\theta)}\right)$.

□

Com este último resultado, estamos aptos a generalizar as Definições 5.5 e 5.6 para o caso assintótico.

Definição 5.11. Sejam $X_1,\ldots,X_n$ variáveis aleatórias independentes e identicamente distribuídas com função densidade de probabilidade $f(x;\theta)$. Suponha que $\hat{\theta}_{1n} = \hat{\theta}_{1n}(X_1,\ldots,X_n)$ é um estimador de θ_0 tal que $\sqrt{n}(\hat{\theta}_{1n}-\theta_0) \xrightarrow{P} N\left(0,\sigma^2_{\hat{\theta}_{1n}}\right)$.

(a) A eficiência assintótica de $\hat{\theta}_{1n}$ é definida por $e(\hat{\theta}_{1n}) = \dfrac{1/I(\theta_0)}{\sigma^2_{\hat{\theta}_{1n}}}$;

(b) O estimador $\hat{\theta}_{1n}$ é dito eficiente assintoticamente se $e(\hat{\theta}_{1n}) = 1$.

Portanto, pelo Teorema 5.10 e sob as condições de regularidade, os estimadores de máxima verossimilhança são estimadores eficientes assintoticamente.

Munidos de todas as informações e propriedades dos estimadores de máxima verossimilhança, podemos agora estudar como fazemos inferência para esses estimadores, aplicando os conceitos vistos no Capítulo 2.

5.1 Teste de Hipóteses e Intervalo de Confiança

Uma vez que o Teorema 5.10 nos fornece a distribuição dos estimadores de máxima verossimilhança e com os conceitos abordados no Capítulo 2 deste trabalho, podemos definir o Intervalo de Confiança para θ. Mas antes, observe que na nossa discussão vimos que $\hat{\theta}_n \xrightarrow{P} \theta_0$. Sendo $I(\theta)$ uma função contínua, podemos afirmar que $I(\hat{\theta}_n) \xrightarrow{P} I(\theta_0)$, ou seja, $I(\hat{\theta}_n)$ é um estimador consistente para a variância.

Logo, com um nível de confiança $0 \leq \alpha \leq 1$, o Intervalo de Confiança para θ é,

$$IC(\hat{\theta}_n; \alpha) = \left[\hat{\theta}_n - z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta}_n)}}; \hat{\theta}_n + z_{\alpha/2}\frac{1}{\sqrt{nI(\hat{\theta}_n)}}\right]. \tag{5.18}$$

A seguir, derivaremos três Testes de Hipóteses possíveis para os estimadores de verossimilhança. Ao final veremos que todos são assintoticamente equivalentes.

Considere $X_1, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas com $f(x;\theta)$, onde $\theta \in I \subset \mathbb{R}$ e $\hat{\theta}$ é o estimador de máxima verossimilhança. O objetivo é testar as hipóteses,

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta \neq \theta_0,$$

onde θ_0 é o valor verdadeiro de θ. Vimos no Teorema 4.6 que $L(\theta_0) > L(\theta)\ \forall \theta \neq \theta_0$. Assim, sendo $\Delta = \dfrac{L(\theta_0)}{L(\hat{\theta})}$ temos que se H_0 for verdade, Δ valerá 1 e se H_1 for verdade, a razão Δ ficará pequena, menor que 1. Isso nos leva a um critério de decisão.

Ao nível de confiança $0 \leq \alpha \leq 1$, rejeitamos H_0 se $\Delta = \dfrac{L(\theta_0)}{L(\hat{\theta})} \leq c$, onde c é tal que $\alpha = \mathbb{P}[\Delta \leq c]$. Este teste é chamado **Teste da Razão de Verossimilhança**.

Exemplo 5.12. *Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas com distribuição Normal, onde o parâmetro θ a ser estimado é a média μ e $\sigma^2 > 0$ é conhecido. Considere o teste,*

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_1 : \theta \neq \theta_0.$$

Temos que,

$$L(\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-(2\sigma^2)^{-1}\sum_{i=1}^{n}(x_i - \overline{X})^2\right) \exp\left(-(2\sigma^2)^{-1}n(\hat{x} - \theta)^2\right).$$

E, neste caso, $\hat{\theta} = \overline{X}$ *e* $\Delta = \dfrac{L(\theta_0)}{L(\hat{\theta})} = \exp\left(-(2\sigma^2)^{-1}n(\overline{X} - \theta_0)^2\right)$. *Mas,* $\Delta \leq c$ *é equivalente a* $-2\log\Delta \geq -2\log c$. *Assim,*

$$-2\log\Delta = \left(\frac{\overline{X} - \theta_0}{\sigma/\sqrt{n}}\right)^2$$

tem distribuição Qui-Quadrado com 1 grau de liberdade sob H_0. *Logo, com nível de confiança* $0 \leq \alpha \leq 1$, *rejeitamos* H_0 *quando* $-2\log\Delta = \left(\dfrac{\overline{X} - \theta_0}{\sigma/\sqrt{n}}\right)^2 \geq \chi^2(1)$.

O teorema a seguir nos dá a distribuição da estatística do teste da razão de verossimilhança. Mas antes, vamos definir um resultado que será importante para a prova do teorema.

Corolário 5.13. *Sob as condições* $R0 - R5$ *temos que,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}I(\theta_0)}\sum_{i=0}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} + R_n,$$

onde $R_n \xrightarrow{P} 0$.

A demonstração deste Corolário será omitida, mas segue diretamente do Teorema 5.10.

Teorema 5.14. *Considerando as condições de regularidade* $R0 - R5$ *temos que sob a hipótese nula* $H_0 : \theta = \theta_0$ *vale que*

$$-2\log\Delta \xrightarrow{D} \chi^2(1).$$

Demonstração. Vamos expandir a função $l(\theta)$ em série de Taylor sob θ_0 e avaliar em $\hat{\theta}$.

$$l(\hat{\theta}) = l(\theta_0) + (\hat{\theta} - \theta_0)l'(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2 l''(\theta_n^*),$$

onde $\hat{\theta} \leq \theta_n \leq \theta_0$. Como $\hat{\theta}_n \xrightarrow{P} \theta_0$ então $\hat{\theta}_n^* \xrightarrow{P} \theta_0$. Usando isso, e o fato de que $l''(\theta)$ é contínua então

$$-\frac{1}{n}l''(\theta_n^*) \xrightarrow{P} I(\theta_0). \tag{5.19}$$

Pelo Corolário 5.13,

$$-\frac{1}{n}l'(\theta) = \sqrt{n}(\hat{\theta}_n - \theta_0)I(\theta_0) + R_n, \tag{5.20}$$

onde $R_n \xrightarrow{P} 0$. Aplicando as Equações 5.19 e 5.20 na expansão de Taylor, obtemos:

$$-2\log\Delta = 2(l(\hat{\theta}) - l(\theta_0)) = \{\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)\}^2 + R_n^*,$$

onde $R_n^* \xrightarrow{P} 0$. Como $\{\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)\}$ converge em distribuição para a distribição Normal, então $-2\log\Delta \xrightarrow{D} \chi^2(1)$. □

Do Teorema 5.14 vimos que a estatística a ser considerada no Teste da Razão de Verossimilhança é,

$$\chi_L^2 = -2\log\Delta \tag{5.21}$$

que tem distribuição Qui-Quadrado sob H_0.

Assim, a um nível de confiança $0 \leq \alpha \leq 1$ rejeitamos H_0 se $\chi_L^2 = -2\log\Delta \geq \chi^2(1)$.

O próximo teste é conhecido como **Teste de Wald** e é semelhante ao Teste da Razão de Verossimilhança, se diferindo apenas no fato de que utilizaremos a distribuição assintótica de $\hat{\theta}$, para os casos onde não obtemos a distribuição de forma fechada como no Exemplo 5.12. Neste teste, a estatística a ser considerada é,

$$\chi_W^2 = \{\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)\}^2. \tag{5.22}$$

Sabemos que sob H_0, $I(\hat{\theta}) \xrightarrow{P} I(\theta_0)$. Desse modo, sob H_0, a estatística converge assintoticamente para a distribuição Qui-Quadrado com 1 grau de liberdade. Portanto, ao nível de confiança $0 \leq \alpha \leq 1$, rejeitamos H_0 se $\chi_W^2 \geq \chi^2(1)$.

Por último, temos o **Teste Score** que se difere dos outros testes por não utilizar informações acerca do estimador de máxima verossimilhança $\hat{\theta}$.

Neste teste, a estatística a ser considerada é,

$$\chi_R^2 = \left(\frac{l'(\theta_0)}{\sqrt{nI(\theta_0)}}\right)^2. \tag{5.23}$$

Observe que das Equações 5.20 e 5.22 temos que $\chi_R^2 = \chi_W^2 + R_{0n}$, onde $R_{0n} \xrightarrow{P} 0$.

Portanto, ao nível de confiança $0 \leq \alpha \leq 1$, rejeitamos H_0 se $\chi_R^2 \geq \chi^2(1)$.

Até agora, podemos claramente observar que em todos os testes, a distribuição da estatística é Qui-Quadrado com 1 grau de liberdade, tornando a regra de decisão a mesma em todos os casos. Mas então, como decidir qual o melhor teste?

Bom, a diferença entre eles está no quão conveniente o teste será na aplicação em questão, e também na facilidade de interpretação. Tanto o Teste da Razão de Verossimilhança quando o Teste de Wald requer informações a cerca do estimador encontrado para θ, ao contrário do Teste Score. Mas, por outro lado, o Teste de Wald é conveniente em termos de interpretação, pois se torna mais natural considerar a estatística $\chi_W^2 = \{\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0)\}^2$. Mas há uma desvantagem: o Teste de Wald não é invariante por reparametrização, veja em [5]. Já nos casos onde não obtemos solução analítica para o estimador, o Teste Score é mais conveniente, pois bastaria calcular a derivada de $l(\theta)$.

De qualquer forma, ainda é difícil definir se há um teste que possa ser dito como o melhor. O ponto importante é que todos são eficientes e equivalentes assintoticamente, pois possuem mesma distribuição. O leitor pode encontrar um estudo mais detalhado em [5].

6 Aplicação

Nesta seção, veremos uma aplicação do Método de Máxima Verossimilhança em Regressão Logística. O objetivo é analisar o funcionamento do método na prática, pontuando as partes teóricas que vimos neste trabalho.

A aplicação foi feita em risco de crédito e o objetivo é criar um modelo de Regressão Logística que forneça a chance de inadimplência de um indivíduo, analisando suas informações pessoais. A estimação por Máxima Verossimilhança foi feita para mais de um parâmetro, e o leitor pode encontrar os detalhes do método para o caso multiparamétrico em [3].

Neste primeiro momento, definiremos brevemente o modelo de Regressão Logística e logo após, discutiremos os objetos utilizados na aplicação. Toda a parte computacional envolvida nesta aplicação foi feita pelo sofwtare estatístico R e não abordaremos aqui assuntos relacionados ao código. Para saber mais sobre Regressão Logística o leitor pode consultar [4].

6.1 Regressão Logística

O modelo de Regressão Logística é utilizado quando a variável resposta Y é do tipo binária, ou seja, a variável assume apenas dois valores e geralmente denotamos esses valores como 1 para sucesso (ocorrência do evento) e 0 para fracasso (não ocorrência do evento). O evento de interesse é a variável assumir valor 1, ou seja, o sucesso. Desse modo, a variável resposta tem uma distribuição que já conhecemos, a distribuição Bernoulli. Sendo assim, $\mathbb{P}(Y = 1) = \pi(X)$ e $\mathbb{P}(Y = 0) = 1 - \pi(X)$, com $\mathrm{E}(Y) = \pi(X)$ e $X = (1, x_1, \ldots, x_n)$ representando a amostra das variáveis explicativas conhecidas.

Na equação do modelo de Regressão Logística escrevemos

$$\mathrm{E}(Y) = \pi(X) = \frac{\exp^{X'\beta}}{1 + \exp^{X'\beta}}, \tag{6.1}$$

onde $\beta = (\beta_0, \ldots, \beta_n)$ é o vetor dos parâmetros desconhecidos e $X'\beta = \beta_0 + \beta_1 x_1, \ldots, \beta_n x_n$. Repare que esta equação não é linear, mas há uma representação linear para a equação do modelo, chamada Transformação Logit [4]

$$\log\left(\frac{\pi(X)}{1 + \pi(X)}\right) = \beta_0 + \beta_1 x_1, \ldots, \beta_n x_n. \tag{6.2}$$

Dado uma observação Y_i com distribuição Bernoulli, a função distribuição de probabilidade é dada por

$$f(Y_i) = \pi(X)^{Y_i}(1 - \pi(X))^{1-Y_i}.$$

Assim, para uma amostra de observações $Y_1, \ldots, Y_n$ a função distribuição de probabilidade conjunta é da forma

$$g(Y_i, \ldots, Y_n) = \prod_{i=1}^{n} f(Y_i) = \prod_{i=1}^{n} \pi(X)^{Y_i}(1 - \pi(X))^{1-Y_i}. \tag{6.3}$$

Como o interesse é explicar Y_i através das variáveis explicativas $X = (1, x_1, \ldots, x_n)$, vamos substituir $\pi(X)$ na Equação 6.3 pela Equação 6.1, obtendo

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(X'\beta)}{1 + \exp(X'\beta)} \right)^{Y_i} \left(\frac{1}{1 + \exp(X'\beta)} \right)^{1-Y_i}. \tag{6.4}$$

Pelo exemplo 4.4 podemos afirmar que a função dada pela Equação 6.4 é uma densidade de probabilidade e vale as condições de regularidade. Por isso, podemos aplicar o Método de Máxima Verossimilhança sendo $L(\beta)$ a função de verossimilhança da Regressão Logística, onde $\beta_0, \ldots, \beta_n$ serão os parâmetros desconhecidos a serem estimados.

Novamente, é mais fácil encontrar os estimadores de máxima verossimilhança maximizando o log da função de verossimilhança dado por

$$l(\beta) = \sum_{i=1}^{n} Y_i(X'\beta) - \sum_{i=1}^{n} \log(1 + \exp(X'\beta))\cdot \tag{6.5}$$

6.2 Construção do Modelo

Na aplicação, utilizamos o conjunto de dados Default do pacote ISLR do software R, contendo informação de 10.000 clientes. O objetivo é prever quais clientes não irão pagar sua dívida do cartão de crédito, ou seja, desejamos prever se o cliente ficará inadimplente ou não. As variáveis explicativas são *student* - uma variável binária indicando se o cliente é ou não estudante e *balance* - o saldo médio que o cliente tem em seu cartão de crédito após fazer seu pagamento mensal. A variável aleatória Y representa o evento estar ou não indadimplente e continua tendo distribuição Bernoulli.

Logo, considerando a amostra obtida do conjunto de dados, já temos a função de verossimilhança para a estimação, obtida na Equação 6.5. A solução desta equação não foi obtida analiticamente, o software utilizou o método computacional Fisher Scoring. Este método inicia escolhendo um parâmetro como valor inicial e realiza sucessivos passos iterativos até obter a convergência.

Na Figura 1 mostra a saída do modelo pelo software R. Na segunda coluna vemos os valores estimados para cada parâmetro, assim como o valor da estatística de Wald para cada parâmetro, dados na coluna 4.

```
> logitmod<-glm(default~student+balance, family="binomial", data=trainData)
> summary(logitmod)

Call:
glm(formula = default ~ student + balance, family = "binomial",
    data = trainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4899  -0.1379  -0.0534  -0.0190   3.7769

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.095e+01  3.990e-01 -27.452  < 2e-16 ***
studentYes  -6.665e-01  1.553e-01  -4.292 1.77e-05 ***
balance      5.858e-03  2.499e-04  23.445  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2630.7  on 9000  degrees of freedom
Residual deviance: 1393.9  on 8998  degrees of freedom
AIC: 1399.9

Number of Fisher Scoring iterations: 8
```

Figura 1: Saída do modelo fornecida pelo software R.

Obtemos o vetor $\hat{\beta} = (-10.95, -0.6665, 0.005858)$ dos estimadores para os parâmetros, obtendo a equação do modelo

$$\log\left(\frac{\pi(X)}{1+\pi(X)}\right) = -10.95 - 0.6665 * student + 0.005858 * balance.$$

Portanto, tendo em mãos as duas informações do cliente, podemos obter explicitamente a chance $\pi(X)$ do cliente não pagar a dívida.

6.3 Inferência para os Parâmetros

O intervalo de confiança para os parâmetros é feito a partir do Teorema 5.10 que nos fornece a distribuição dos estimadores de máxima verossimilhança. Da Equação 5.18 o intervalo de confiança ao nível $\alpha = 95\%$ para os parâmetros pode ser visto na Figura 2.

```
> ICbeta <-confint.default(logitmod,level=0.95)
> ICbeta
                    2.5 %        97.5 %
(Intercept) -11.735459446 -10.171382444
studentYes   -0.970879368  -0.362115624
balance       0.005368533   0.006348012
```

Figura 2: Intervalo de Confiança para os parâmetros.

O Teorema 5.10 afirma que a variância destes estimadores é dada pela Informação de Fisher. O análogo para a Regressão Logística é a Matriz de Informação, cujas entradas são as derivadas parciais do log da função de verossimilhança em relação aos parâmetros, ou seja, a Matriz de Informação G é tal que

$$G = [g_{ij}] = \frac{\partial l(\beta)}{\partial \beta_i \partial \beta_j}.$$

Quando aplicamos o vetor $\hat{\beta}$ dos estimadores de máxima verossimilhança na matriz G e calculamos sua inversa, obtemos a matriz estimada de variância e covariância dos estimadores e denotamos por $s^2(\hat{\beta})$. A entrada onde $i = j$ representa a variância do estimador β_i. Esta matriz geralmente é obtida por método computacionais.

Para os parâmetros de Regressão Logística, podemos realizar Teste de Hipóteses de duas maneiras, testando apenas um parâmetro por vez, ou testando mais de um parâmetro por vez. Para o primeiro caso geralmente utilizamos o Teste de Wald e o Teorema 5.10. Assim, se queremos testar $H_0 : \beta_i = 0$ vs $H_a : \beta_i \neq 0$ para algum $i = 1, \ldots, n$ utilizamos como estatística

$$W_i = \frac{\hat{\beta}_i - \beta_i}{s^2[\beta_{ii}]} \sim N(0, 1),$$

e rejeitamos H_0 se $|W_i| > |z_{\alpha/2}|$, lembrando que $z_{\alpha/2}$ é tal que $\mathbb{P}(W_i \geq z_{\alpha/2}) = \alpha/2$.

No momento da criação do modelo, o software R forneceu os valores das estatísticas de Wald para cada parâmetro com nível $\alpha = 0.95$. Para $\hat{\beta}_0$ obtemos $|W_0| = |-27.452| > 1.96$ e neste caso, rejeitamos H_0. Para $\hat{\beta}_1$ obtemos $|W_1| = |-4.292| > 1.96$ e neste caso, rejeitamos H_0. Para $\hat{\beta}_2$ obtemos $|W_2| = |23.445| > 1.96$ e neste caso, também rejeitamos H_0.

Para testar se alguns parâmetros são nulos utilizamos o Teste da Razão de Verossimilhança com estatísitca dada pela Equação 5.21, onde Δ é a razão da verossimilhança do modelo sob H_0 e a verossimilhança do modelo sob H_a.

Para testar $H_0 : \beta_0 = \beta_1 = 0$ vs $H_a : \beta_0 \neq 0$ e $\beta_1 \neq 0$, basta gerar o modelo sob H_0 obtendo o vetor dos estimadores $\hat{\beta}_{H_0}$ e o modelo sob H_a obtendo o vetor dos estimadores $\hat{\beta}_{H_a}$. A estatística, neste caso, é

$$\chi^2_L = -2(\log L(\hat{\beta}_{H_0}) - \log L(\hat{\beta}_{H_a})) \sim \chi^2_{(2)},$$

e rejeitamos H_0 se $\chi^2_L > \chi^2_{(2)}$.

A partir dos dois modelos gerados, a estatística do teste com nível $\alpha = 0.95$ é $\chi^2_L = 1413.2 - 1393.9 = 19.3 > 5.99$. Logo, rejeitamos H_0.

Dos testes e intervalo de confiança, podemos concluir que a equação encontrada estima bem a chance de inadimplência de acordo com as variáveis explicativas utilizadas.

Para a intepretação do modelo, a Figura 3 mostra a curva da chance de inadimplência para estudantes e não estudantes. O eixo x representa os valores da variável balance e o eixo y representa a chance de inadimplência.

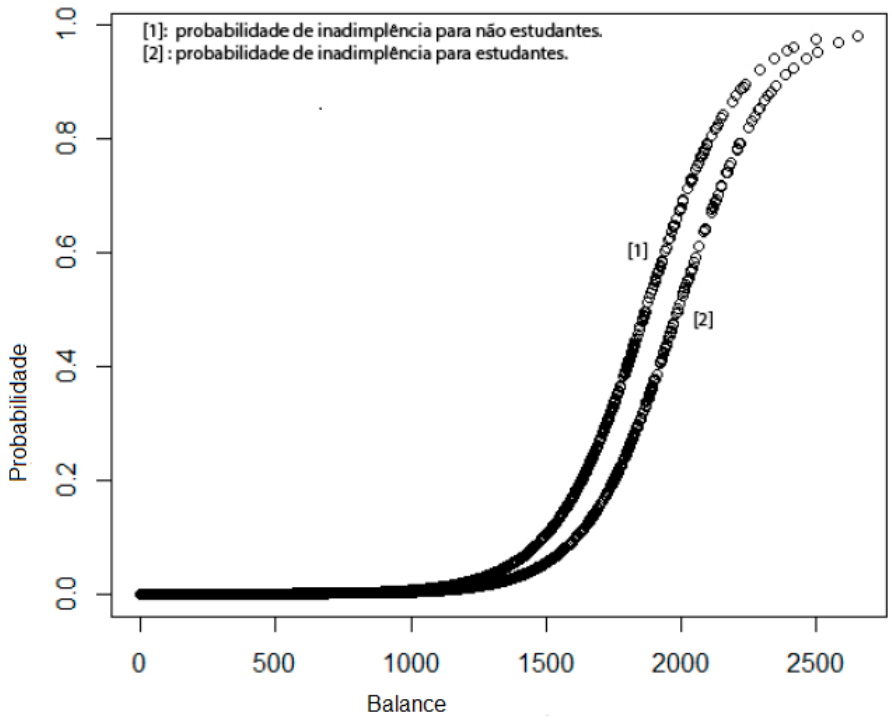

Figura 3: Gráfico do Modelo Ajustado

Observando o gráfico, verificamos que quando o saldo médio do cartão de crédito do cliente é maior do que $R\$1500$, ser ou não estudante passa a fazer alguma diferença na probabilidade de inadimplência. Mas para saldos maiores, a partir de $R\$2500$, a chance de inadimplência é alta independente de ser ou não estudante.

7 Considerações finais

Neste trabalho o objetivo foi desenvolver os aspectos teóricos do Método de Máxima Verossimilhança, seus estimadores e suas propriedades, apresentando o Limite Inferior de Rao-Cramér e sua importante contribuição para o método, vista a partir da distribuição assintótica dos estimadores.

Durante o desenvolvimento, fornecemos os pré requisitos necessários para o assunto, abordando os conceitos básicos em Probabilidade e Estatística e definindo convergência de variáveis aleatórias, chegando ao Teorema Central do Limite. Em seguida introduzimos o método matematicamente, enfatizando a sua justificativa teórica. Provamos as propriedades dos estimadores de máxima verossimilhança, sendo consistência a primeira delas. A partir das Condições de Regularidade definimos a Informação de Fisher e, consequentemente, o teorema central deste trabalho, o Limite Inferior de Rao-Cramér. Com este resultado caracterizamos e provamos a eficiência assintótica dos estimadores, discutimos a distribuição assintótica e os procedimentos para inferência. Após toda discussão teórica, vimos como o método funciona na prática estimando os coeficientes do modelo de Regressão Logística.

Vimos brevemente no Capítulo 4 que, para maximizar a função de máxima verossimilhança, pode ser necessário algum método computacional. Diante disso, um tópico que poderia agregar na continuação deste trabalho seria desenvolver a abordagem computacional do método, ou seja, estudar os algoritmos utilizados para maximizar a função de máxima verossimilhança. Discutindo os critérios de convergência destes algoritmos e adicionando este tópico à aplicação que foi desenvolvida.

Outro tópico interessante para trabalhos futuros seria mostrar a relação entre suficiência e os estimadores de máxima verossimilhança. É possível mostrar que se existe uma estatística suficiente para um parâmetro θ e o estimador de máxima verossimilhança $\hat{\theta}$ também existir e for único, então $\hat{\theta}$ é uma função desta estatística suficiente.

Um último tópico seria estender o Método de Máxima Verossimilhança para o caso em que se tem mais de um parâmetro a ser estimado, que foi o caso da aplicação fornecida neste trabalho. Em modelos de regressão há sempre mais de um parâmetro que precisa ser estimado e neste caso, a extensão do método para o caso multiparamétrico seria ideal.

Referências

1 ROSS, S. *A first course in Probability*. 8. ed. New York: Person, 2008. Citado 2 vezes nas páginas 3 e 7.

2 ROLLA, L. T. *Introdução à Probabilidade: Notas de Aula*. 1. ed. Rio de Janeiro: [s.n.], 2018. Citado na página 3.

3 HOGG J. W. MCKEAN, A. T. C. R. V. *Introduction to Mathematical Statistics*. 7. ed. Boston: Person, 2012. Citado 6 vezes nas páginas 1, 7, 17, 26, 31 e 36.

4 KUTNER, M.H. ; NACHTSHEIM, C. . N. J.; LI, W. *Applied Linear Statistical Models*. 5. ed. Georgia: McGraw-Hill Irwin, 2005. Citado na página 36.

5 LEHMANN, E.L. *Elements of Large-Sample Theory*. 2. ed. EUA: Thomson Learning, 2002. Citado na página 35.

6 CASELLA, G. ; BERGER, R.L. *Statistical Inference*. 8. ed. New York: Person, 2008. Citado na página 9.

7 BOLFARINE, H. ; SANDOVAL, M.C. *Introdução à Inferência Estatística*. 2. ed. Rio de Janeiro: SBM, 2010. Citado na página 9.

8 JAMES,B.R. *Probabilidade: um curso em nível intermediário*. 3. ed. Rio de Janeiro: IMPA, 2006. Citado na página 15.

9 BUSSAB, W. O. *Estatística Básica*. 7. ed. São Paulo: Editora Saraiva, 2010. Citado na página 9.

10 BREIMAN, L. *Probability*. 1. ed. Massachusetts, EUA: Addison-Wesley, 1968. Citado na página 15.

11 CHUNG, K.L. *A Course in Probability Theory*. 2. ed. New York: Academic Press, 1974. Citado na página 6.

12 STIGLER, S.M. *The Epic Story of Maximum Likelihood*. Statistical Science, v. 22, n. 4, p.598–620, 2007. Citado na página 1.

13 ALDRICH, J. *R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922*. Statistical Science, v. 12, n. 3, p.162-176, 1997. Citado na página 1.

—