Estatística Inferencial com apoio computacional

PhD Wagner Hugo Bonat
PhD Paulo Justiniano Ribeiro Junior
Dr Walmes Marques Zeviani

Contents

Prefácio

A ideia de verossimilhança como instrumento para avaliar a evidência contida nos dados está no centro dos fundamentos da metodologia estatística. Embora formalizada nos trabalhos de Sir Ronald Aylmer Fisher nos anos 20, apenas muitas décadas depois e principalmente com as possibilidades abertas pela computação estatística, esta pôde ser explorada, investigada, aplicada, modificada e expandida nas mais diferentes formas.

A necessidade de computação em estatística está presente desde sua origem, seja de forma manual ou, na agora onipresente, forma eletrônica com o uso de computadores. O desenvolvimento de linguagens e aplicativos para computação estatística ampliam rápida e largamente as possibilidades de geração e tratamento de dados. Linguagens interpretadas, direcionadas e/ou adaptáveis para computação estatística diminuem dramaticamente a distância entre programação e uso de aplicativos permitindo usuários investigar possibilidades e conceitos, experimentar ideias, adaptar códigos, implementar protótipos com grande flexibilidade ainda que sem um investimento proibitivo no domínio dos recursos utilizados.

Em particular os projetos de software livre cumprem tal papel sem impor obstáculos ao usuário. Neste contexto o Projeto `R` de Computação Estatística iniciado na década de 90 e com a primeira versão lançada no ano 2000, tem uma impactante contribuição e larga abrangência que, em muito, ultrapassa os limites da área de estatística. A linguagem já imprimiu uma marca indelével no conjunto de recursos disponíveis para interessados em computação e métodos estatísticos.

O presente texto situa-se na interface entre métodos de inferência estatística baseada em verossimilhança e métodos computacionais (com implementações em ambiente `R`). Sem nos aprofundarmos em nenhuma das duas áreas, procuramos ilustrar suas conexões por meio de diversos exemplos básicos de modelagem estatística. Nossa expectativa é a de que o texto possa servir como material introdutório ao leitor e facilitar seu caminho para construir

suas próprias implementações em modelagens de seu interesse.

O material foi motivado por nossa experiência em grupos de estudos e disciplinas conduzidas no âmbito do LEG/UFPR (Laboratório de Estatística e Geoinformação da Universidade Federal do Paraná) nos últimos anos. Procuramos mesclar a discussão de princípios básicos de inferência estatística, com ênfase em métodos baseados na função de verossimilhança, com a implementação computacional. Nossa estratégia usual é a de escrever nossas funções, na forma de protótipos, para melhor desenvolver a intuição sobre as características dos modelos e métodos estatísticos em discussão. Desta forma as funções e códigos apresentados são predominantemente ilustrativos, privilegiando a facilidade de leitura e entendimento. Os códigos não devem ser vistos como implementações definitivas nem tampouco tentam explorar o uso eficiente da linguagem, ainda que alguns cuidados para evitar problemas numéricos sejam tomados na definição de certas operações. Por vezes os resultados são comparados com os fornecidos por funções do `R` e alguns de seus pacotes. Seguimos a sugestão de que *''programming is the best way to debug your ideias"*.

Nosso público alvo são alunos de final de graduação, com alguma exposição anterior a conceitos de inferência estatística e ao uso do ambiente `R`. Outros potenciais interessados são alunos em início de pós-graduação e/ou profissionais que tenham interesse em se familiarizar com elementos de programação em `R` para inferência estatística. Incentivamos os leitores do material a nos enviar comentários, sugestões e correções.

O texto é permeado de códigos em linguagem `R` que são identificados pelo uso de fonte estilo `VERBATIM como esta`. Um tratamento especial é dado as funções em `R` que são definidas dentro de caixas em destaque. Tipicamente estas definem funções implementando alguma metodologia ou alguma função de verossimilhança a ser chamada por funções otimizadoras.

Todo o material é produzido utilizando *software* livre. As implementações de métodos e algoritmos é toda feita no ambiente `R` de computação estatística. O texto é escrito utilizando *latex* e a integração com o `R` pelo mecanismo `rmarkdown`. Os recursos são utilizados em sistema operacional `LINUX`.

W.H.B, P.J.R. Jr e W.M.Z.

Curitiba, Junho, 2021.

Chapter 1

Introdução

A abordagem estatística para análise e resumo de informações contidas em um conjunto de dados, consiste na suposição de que existe um mecanismo estocástico gerador do processo em análise. Este mecanismo é descrito por meio de um modelo probabilístico, representado por uma distribuição de probabilidade. Em situações reais a verdadeira distribuição de probabilidade geradora do processo é desconhecida, sendo assim, distribuições de probabilidade adequadas devem ser escolhidas de acordo com o tipo de fenômeno em análise. Por exemplo, se o fenômeno em estudo consiste em medir uma característica numérica de um grupo de indivíduos em uma escala contínua, distribuições com este **suporte** devem ser escolhidas. O **suporte** de uma distribuição de probabilidade informa qual o domínio da função, ou seja, quais são os valores que a variável aleatória pode assumir. Considere o caso da distribuição gaussiana, o **suporte** é a reta real, no caso da distribuição gama o suporte é apenas os reais positivos. Um cuidado adicional deve ser dado quando a variável de interesse é discreta, por exemplo contagens, onde é comum atribuir uma distribuição de Poisson que tem **suporte** nos naturais positivos.

Em quase todos os problemas de modelagem estatística não existe uma única distribuição de probabilidade que pode representar o fenômeno. Porém, na maioria das situações assume-se que a distribuição de probabilidade geradora do processo é conhecida, com exceção dos valores de um ou mais **parâmetros** que a indexam. Por exemplo, considere que o tempo de vida de um tipo de componente eletrônico tem distribuição exponencial com **parâmetro** λ, mas o valor exato de λ é desconhecido. Se o tempo de vida de vários componentes de mesmo tipo são observados, baseado nestas observações e qualquer outra fonte relevante de informação que esteja disponível, é possível

fazer **inferência** sobre o valor desconhecido do parâmetro λ. O processo de **inferência** consiste em encontrar um valor mais plausível para λ, bem como, informar um intervalo para o qual acredita-se conter o verdadeiro valor de λ, além de decidir ou opinar se λ é igual, maior ou menor que algum valor previamente especificado. Em alguns problemas há ainda interesse em fazer previsões sobre possíveis valores do processo, por exemplo, em outros tempos ou locais.

Em implementações computacionais para inferência estatística, deve-se sempre estar atento ao **espaço paramétrico** (Θ) de um modelo probabilístico. No caso do tempo de vida de componentes eletrônicos, assumindo que a distribuição exponencial é adequada e está sendo indexada pelo parâmetro λ. De acordo com a construção do modelo exponencial, tem-se que o **espaço paramétrico** de λ é dado pelo conjunto dos reais positivos. Em um modelo gaussiano, com média μ e variância σ^2, o espaço paramétrico é $\Re \times \Re_+$ ou seja, todo o conjunto dos reais para média μ enquanto que para σ^2 o espaço paramétrico restringe-se aos reais positivos. Outro caso comum são modelos em que o parâmetro representa alguma proporção p e tem como espaço paramétrico o intervalo $(0, 1)$. Estas restrições precisam ser levadas em consideração no processo de inferência e são de fundamental importância para o sucesso de muitos algoritmos de maximização numérica. Não raramente nas implementações computacionais são feitas reparametrizações com novos parâmetros para os quais os valores são projetados na reta real, com resultados transformados de volta ao espaço original. Por exemplo, pode-se adotar $\psi = \log \sigma$ para variância do modelo normal e $\psi = \log p/(1-p)$ para a proporção de sucesso em um experimento binomial.

Partindo destes conceitos, um fenômeno aleatório ou estocástico é descrito minimamente por uma **distribuição de probabilidade**, que por sua vez é indexada por seus parâmetros e respectivos campos de variação **espaço paramétrico**, além do campo de variação da própria variável aleatória que deve ser compatível com o **suporte** da distribuição atribuída ao fenômeno. Por exemplo, não é correto atribuir uma distribuição de Poisson para a altura (medida contínua) de trabalhadores, uma vez que o campo de variação da variável de interesse (resposta) não é compatível com o suporte da distribuição de probabilidade.

Considere o caso onde deseja-se fazer uma pesquisa a respeito da intenção de voto em um plebiscito. Suponha que n eleitores selecionados aleatoriamente são questionados sobre a sua intenção em votar a favor (1) ou contra (0) uma determinada mudança na legislação. Deseja-se estimar a proporção θ de eleitores favoráveis à mudança. Assume-se que o modelo Bernoulli seja adequado para a intenção de voto de cada eleitor e portanto o número de favoráveis em uma amostra aleatória de n eleitores tem distribuição

binomial B(n, θ). Este modelo tem como possíveis respostas para cada indivíduo da amostra os valores 0 e 1 e como parâmetro indexador θ que representa a proporção de favoráveis e tem o intervalo unitário como seu espaço paramétrico. Com este conjunto de suposições e conhecimentos a respeito do modelo probabilístico, tem-se total condições de fazer inferência sobre o parâmetro θ a partir dos dados de uma amostra.

A Figura **??**(A) representa a região definida pelo modelo para uma amostra aleatória de tamanho 100. A superfície é obtida calculando-se os valores das probabilidades de se observar y favoráveis em uma amostra para cada um dos possíveis valores do parâmetro. Para visualização omitimos os valores próximos às bordas $[0, 1]$. Um corte da superfície de probabilidades em um particular valor do parâmetro fornece uma **distribuição de probabilidades** para as possíveis respostas como ilustrado na **??**(B) para $\theta = 0,65$. Um corte para um valor de $y = 60$, que poderia ser o obtido em uma determinada amostra, fornece a **função de verossimilhança** denotada por L$(\theta|y)$ que é apresentada na Figura **??**(C). Tal função fornece uma medida de proximidade entre cada possível valor do parâmetro e a amostra observada.

Figure 1.1: Superfície de probabilidades (esquerda), distribuição de probabilidades (centro) e função de verossimilhança (direita) para um modelo binomial.

Em outras palavras, no gráfico da função de verossimilhança L$(\theta|y)$, a ordenada de cada ponto da curva é dada pela probabilidade do valor y observado na amostra, ter sido gerado por cada um dos possíveis valores de θ. Desta forma é intuitivo pensar que a melhor estimativa para o parâmetro, baseada na amostra, é o valor do parâmetro que tem maior probabilidade de gerar o resultado visto na amostra, portanto o valor que maximiza a função de verossimilhança. Isto define o **estimador de máxima verossimilhança** $\hat{\theta}$. Também é intuitivo pensar que podemos definir uma "faixa" de valores que possuem uma probabilidade "não muito distante e aceitável" da máxima probabilidade de gerar o resultado visto na amostra. Tal faixa define um **estimador por intervalo** com valores inferior e superior $(\hat{\theta}_I, \hat{\theta}_S)$ que delimitam uma região no espaço paramétrico que possui valores de verossimilhança que não estejam abaixo de um percentual pré-definido do máximo possível valor da verossimilhança. Finalmente, pode-se verificar se um determinado valor de interesse, tal como $\theta_0 = 0,5$ no exemplo considerado, é **compatível** com a amostra comparando-se sua verossimilhança com a máxima possível. Este último caso permite definir um **teste de hipótese** de interesse para guiar uma tomada de decisão. Na Figura **??** redesenhamos o gráfico da função de verossimilhança agora com a escala vertical com valores relativos

ao máximo valor e os elementos no gráfico ilustram os três objetivos centrais da inferência estatística.

Figure 1.2: Visualização da estimativa de máxima verossimilhança, estimativa intervalar e valor correspondente a uma hipótese na função de verossimilhança relativa (RL) para o modelo binomial.

Neste texto será dada ênfase na inferência baseada na verossimilhança, que fornece **estimadores** e procedimentos com propriedades desejáveis para os parâmetros desconhecidos de um modelo probabilístico. A **função de verossimilhança** fornece todos os elementos necessários para a obtenção de estimativas pontuais e intervalares, além da construção de testes de hipóteses. Toda a metodologia será descrita através de exemplos abordando diversos aspectos teóricos, com ênfase na implementação computacional para estimação de parâmetros desconhecidos cobrindo desde modelos simples até modelos altamente estruturados. A abordagem de inferência pela verossimilhança não é, entretanto, a solução de todo e qualquer problema, podendo tornar-se intratável analítica e/ou computacionalmente em certos modelos. Entretanto, os princípios permanecem válidos e diversos procedimentos estendem, aproximam ou substituem a verossimilhança quando necessário. Tais extensões estão fora do escopo deste texto que concentra-se na obtenção da verossimilhança com ênfase em procedimento numéricos. Nossa intenção é reforçar a intuição para aspectos fundamentais e básicos de inferência.

Exercícios

1. Para cada uma das distribuições de probabilidade abaixo escreva a função de probabilidade ou densidade probabilidade, identifique o suporte, a esperança, a variância, os parâmetros e o espaço paramétrico.
   a) Distribuição Poisson de parâmetro λ.
   b) Distribuição binomial de parâmetros n e p.
   c) Distribuição exponencial de parâmetro λ.
   d) Distribuição normal de parâmetros μ e σ^2.
   e) Distribuição gama de parâmetros α e β.
   f) Distribuição uniforme de parâmetros a e b.
   g) Distribuição binomial negativa de parâmetros μ e ϕ.
   h) Distribuição log-normal de parâmetros μ e σ^2.
   i) Distribuição inversa Gaussiana de parâmetros μ e σ^2.
   j) Distribuição Tweeedie de parâmetros μ, ϕ e p.
2. Para cada uma das situações abaixo proponha uma distribuição de probabilidade adequada e justifique sua escolha baseado em aspectos do fenômeno aleatório e características da distribuição. Descreva quais

aspectos da inferência estatística podem estar associados com cada uma das situações mencionadas.

a) Itens em uma linha de produção são classificados quanto a sua adequação aos padrões de produção. Apenas as condições conforme ou não-conforme são possíveis.
b) Uma pesquisa de mercado visa identificar o potencial de um novo negócio em uma cidade. Para isto um questionário com perguntas em uma escala likert de cinco níveis foi construído e aplicado a uma amostra de tamanho n da população de interesse.
c) Número de carros que chegam a um caixa automático de um banco durante um período de uma hora nas manhãs de fins de semana.
d) Ocorrência de defeitos relevantes em uma rodovia um mês após sua construção.
e) Medidas antropométricas (peso e altura) são tomadas em crianças do nono ano de escolas públicas brasileiras. Deseja-se caracterizar tais medidas para auxiliar na construção de equipamentos escolares de tamanho adequado.
f) Deseja-se estudar a distribuição do número de horas que um equipamento eletrônico funciona antes de apresentar defeitos com o objetivo de estabelecer um prazo razoável de garantia.
g) Número de quilômetros rodados que um novo pneu é capaz de rodar antes de apresentar defeitos.

Chapter 2

Definições e propriedades

Neste Capítulo apresentamos os três objetivos fundamentais da inferência estatística, estimação pontual, intervalar e teste de hipóteses baseado na função de verossimilhança.

Definição 2.1. Função de verossimilhança - Seja $\mathbf{y}$ um vetor $n \times 1$ representando uma realização de um vetor aleatório $\mathbf{Y}$ com função de probabilidade ou densidade probabilidade $\mathrm{f}(\mathbf{Y}, \boldsymbol{\theta})$, onde $\boldsymbol{\theta}$ denota um vetor $p \times 1$ de parâmetros, com $\boldsymbol{\theta} \in \Theta$, sendo Θ o respectivo espaço paramétrico. A função de verossimilhança ou simplesmente verossimilhança para $\boldsymbol{\theta}$ dado os valores observados $\mathbf{y}$ é a função $\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) \equiv \mathrm{f}(\mathbf{Y}, \boldsymbol{\theta})$.

A função de verossimilhança é dada pela expressão da distribuição conjunta de todas as variáveis aleatórias envolvidas no modelo, porém vista como função dos parâmetros, uma vez que tendo os dados sido observados, são quantidades fixas. Para cada particular valor do parâmetro que pode ser escalar ou vetor, a verossimilhança é uma medida de compatibilidade, plausibilidade ou similaridade do modelo com a amostra observada medida pela probabilidade ou densidade conjunta dos valores observados. Fracamente falando, pode-se dizer que a verossimilhança nos fornece a probabilidade de observar o que foi realmente observado, dado o modelo assumido para os dados.

A expressão da verossimilhança $\mathrm{L}(\boldsymbol{\theta}|\mathbf{y})$ pode ser mais cuidadosamente definida considerando a natureza das variáveis aleatórias. Para modelos discretos não há ambiguidade e o valor da função de verossimilhança é a probabilidade de ocorrer o dado observado, ou seja

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) \equiv \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{Y} = \mathbf{y}).$$

No caso de modelos contínuos a probabilidade de um particular conjunto de valores ser observado é nula. Entretanto, na prática medidas contínuas são tomadas com algum grau de precisão em um intervalo, digamos, $y_{iI} \leq y_i \leq y_{iS}$ e a verossimilhança para um conjunto de observações é:

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{P}_{\boldsymbol{\theta}}(y_{1I} \leq y_1 \leq y_{1S}, y_{2I} \leq y_2 \leq y_{2S}, \ldots, y_{nI} \leq y_n \leq y_{nS}). \quad (2.1)$$

Esta definição é geral e requer a especificação da distribuição conjunta de Y. Fazendo a suposição de observações independentes tem-se que:

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) = \mathrm{P}_{\boldsymbol{\theta}}(y_{1I} \leq y_1 \leq y_{1S}) \cdot \mathrm{P}_{\boldsymbol{\theta}}(y_{2I} \leq y_2 \leq y_{2S}), \ldots, \mathrm{P}_{\boldsymbol{\theta}}(y_{nI} \leq y_n \leq y_{nS}). \quad (2.2)$$

Até este ponto a definição pode ser utilizada tanto para dados considerados pontuais quanto para dados intervalares, como no caso de dados censurados. Vamos supor agora uma situação mais simples e comum na qual todos os dados são medidos a um grau de precisão comum. Neste caso, cada dado é medido em um intervalo $(y_i - \delta/2 \leq Y_i \leq y_i + \delta/2)$ e a verossimilhança é dada por:

$$\begin{aligned}\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) &= \prod_{i=1}^{n} \mathrm{P}_{\boldsymbol{\theta}}(y_i - \delta/2 \leq Y_i \leq y_i + \delta/2) \\ &= \prod_{i=1}^{n} \int_{y_i-\delta/2}^{y_i+\delta/2} \mathrm{f}(y_i, \boldsymbol{\theta}) \mathrm{d}y_i.\end{aligned}$$

Se o grau de precisão é alto, ou seja δ é pequeno em relação a variabilidade dos dados a expressão se reduz a

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) \approx \left(\prod_{i=1}^{n} \mathrm{f}(y_i, \boldsymbol{\theta}) \right) \delta^n,$$

e se δ não depende dos valores dos parâmetros temos a verossimilhança como produto das densidades individuais,

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) \approx \prod_{i=1}^{n} \mathrm{f}(y_i, \boldsymbol{\theta}), \quad (2.3)$$

e de forma mais geral para observações não independentes com a densidade multivariada:

$$\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) \approx \mathrm{f}(\mathbf{y}, \boldsymbol{\theta}). \quad (2.4)$$

No caso onde os elementos de $\mathbf{y}$ são independentes a verossimilhança é simplesmente um produto das distribuições de cada variável aleatória Y_i individualmente, ou seja, $\mathrm{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \mathrm{f}(y_i, \boldsymbol{\theta})$. Neste caso, o procedimento

de inferência pode ser bastante facilitado tanto analítica como computacionalmente. Porém, cabe ressaltar que isso não é uma exigência, e situações onde as amostras não são independentes são tratadas da mesma forma, escrevendo a verossimilhança de uma forma adequada, considerando a distribuição conjunta do vetor $\mathbf{Y}$.

Esta parte do texto concentra-se exclusivamente no uso da função de verossimilhança como base para explicar os aspectos envolvidos na inferência estatística, seja na obtenção de estimativas pontuais, intervalares ou testes de hipóteses. Começamos revisando conceitos de estimação e suas relações com a função de verossimilhança.

2.1 Estimação pontual

Seja $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias com forma conhecida da função probabilidade no caso de variáveis aleatórias discretas ou da função densidade de probabilidade para variáveis aleatórias contínuas, em ambos os casos denotadas por $\mathrm{f}(\mathbf{Y}, \boldsymbol{\theta})$. O vetor $\boldsymbol{\theta}$ denota os parâmetros desconhecidos, sendo que um único elemento de $\boldsymbol{\theta}$ será denotado por θ, o qual queremos estimar através de uma amostra $y_1, y_2, \ldots, y_n$, de realizações das variáveis aleatórias $Y_1, Y_2, \ldots, Y_n$. Denota-se de forma simplificada, $Y_i \sim \mathrm{f}(\boldsymbol{\theta})$ com $i = 1, \ldots, n$. Esta notação deve ser lida da seguinte forma: a variável aleatória Y_i segue uma distribuição $\mathrm{f}(\cdot)$ que por sua vez é indexada, descrita ou governada por um vetor de parâmetros $\boldsymbol{\theta}$. A seguir apresentamos algumas definições importantes para o decorrer do texto.

Definição 2.2. Estatística - Uma estatística é uma variável aleatória $\mathrm{T} = \mathrm{t}(\mathbf{Y})$, onde a função $\mathrm{t}(\cdot)$ não depende de $\boldsymbol{\theta}$.

Definição 2.3. Estimador - Uma estatística T é um estimador para θ se o valor realizado $\mathrm{t} = \mathrm{t}(\boldsymbol{y})$ é usado como uma estimativa para o valor de θ.

Definição 2.4. Distribuição amostral - A distribuição de probabilidade de T é chamada de distribuição amostral do estimador $\mathrm{t}(\mathbf{Y})$.

Definição 2.5. Viés - O viés de um estimador T é a quantidade

$$\mathrm{B(T)} = \mathrm{E(T} - \theta).$$

O estimador T é dito não viciado para θ se $\mathrm{B(T)} = 0$, tal que $\mathrm{E(T)} = \theta$. O estimador T é assintoticamente não viciado para θ se $\mathrm{E(T)} \to \theta$ quando $n \to \infty$.

Definição 2.6. Eficiência relativa - A eficiência relativa entre dois estimadores $\mathrm{T_1}$ e $\mathrm{T_2}$ é a razão $\mathrm{er} = \frac{\mathrm{V(T_1)}}{\mathrm{V(T_2)}}$ em que $\mathrm{V}(\cdot)$ denota a variância do respectivo estimador.

Definição 2.7. Erro quadrático médio - O erro quadrático médio de um estimador T é a quantidade

$$\text{EQM(T)} = \text{E}((\text{T} - \theta)^2) = \text{V(T)} + \text{B(T)}^2.$$

Definição 2.8. Consistência - Um estimador T é **médio quadrático consistente** para θ se o EQM(T) $\to 0$ quando $n \to \infty$. O estimador T é **consistente em probabilidade** se $\forall \epsilon > 0$, $\text{P}(|\text{T} - \theta| > \epsilon) \to 0$, quando $n \to \infty$.

Estas definições introduzem conceitos e propriedades básicas para uma estatística ser um estimador adequado para um determinado parâmetro. Fracamente falando, o desejo é obter um estimador que seja assintóticamente não-viciado, ou seja, conforme o tamanho da amostra aumenta ele se aproxima cada vez mais do verdadeiro valor do parâmetro. Além disso, é interessante que ele seja eficiente, ou seja, apresente a menor variância possível entre todos os estimadores de θ. Esta definição de eficiência, introduz o conceito de variância minima. Sendo assim, para saber se um estimador é eficiente é necessário conhecer um limite inferior para a variância de um estimador, uma vez que tal quantidade exista e seja passível de calcular, ao propor um estimador para θ, basta calcular a sua variância e comparar com a menor possível, se ele atingir este limite será eficiente. Além disso, tomando sua esperança pode-se concluir sobre o seu viés dependendo da situação em termos assintóticos. O Teorema (limite inferior de Cramér-Rao) **??**, ajuda a responder sobre a eficiência de um estimador qualquer. Mas antes precisamos de mais algumas definições.

Como dito, a verossimilhança é uma medida de compatibilidade da amostra observada com um particular vetor de parâmetros, desta forma é natural definir como estimador para o vetor de parâmetros $\boldsymbol{\theta}$, aquele particular vetor digamos, $\hat{\boldsymbol{\theta}}$, que tenha a maior compatibilidade com a amostra, ou em outras palavras o vetor que maximiza a função de verossimilhança ou compatibilidade. O particular valor assumido pela função de verossimilhança neste caso não é importante, o que interessa para **inferência** são os valores relativos de $\text{L}(\boldsymbol{\theta}|\text{y})$ para diferentes conjuntos de $\boldsymbol{\theta}$.

Definição 2.9. Estimativa de máxima verossimilhança - Seja $\text{L}(\boldsymbol{\theta}|\mathbf{y})$ a função de verossimilhança. O valor $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y})$ é a estimativa de máxima verossimilhança para $\boldsymbol{\theta}$ se $\text{L}(\hat{\boldsymbol{\theta}}) \geq \text{L}(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta} \in \Theta$.

Definição 2.10. Estimador de máxima verossimilhança - Se $\hat{\boldsymbol{\theta}}(\mathbf{y})$ é a estimativa de máxima verossimilhança, então $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ é o estimador de máxima verossimilhança. Em geral vamos usar a abreviação EMV para nos referirmos ao estimador de máxima verossimilhança.

Nesta etapa é preciso ter cuidado com a notação. Veja que $\hat{\boldsymbol{\theta}}(\mathbf{y})$ é um vetor de escalares, por outro lado $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ é um vetor de variáveis aleatórias. Daqui em diante usaremos apenas $\hat{\boldsymbol{\theta}}$, para ambos os casos sendo que o contexto indicará o real sentido de $\hat{\boldsymbol{\theta}}$. A função de verossimilhança contêm toda a informação proveniente dos dados sobre o vetor de parâmetros $\boldsymbol{\theta}$. Apesar disso, a $\mathrm{L}(\boldsymbol{\theta}|\mathbf{y})$ é computacionalmente incoveniente, uma vez que esta função apresentará valores muito próximos de zero, conforme o tamanho da amostra aumenta. Por razões meramente computacionais é mais comum usar a função de log-verossimilhança.

Definição 2.11. Log-verossimilhança - Se $\mathrm{L}(\boldsymbol{\theta}|\mathbf{y})$ é a função de verossimilhança, então $\mathrm{l}(\boldsymbol{\theta}|\mathbf{y}) = \log \mathrm{L}(\boldsymbol{\theta}|\mathbf{y})$ é a função de log-verossimilhança.

Segue do fato da função logaritmo ser monótona crescente que maximizar $\mathrm{L}(\boldsymbol{\theta}|\mathbf{y})$ e $\mathrm{l}(\boldsymbol{\theta}|\mathbf{y})$ levam ao mesmo ponto de máximo. Neste ponto estamos habilitados a enunciar um dos teoremas mais fortes da inferência estatística que permitirá concluir sobre a eficiência de um estimador. Neste texto vamos enunciar o Teorema apenas para um o caso em que θ é um escalar, porém o caso multiparâmetros segue de forma analoga.

Teorema 2.1. ***Limite inferior de Cramer-Rao*** - *Se* T *é um estimador não-viciado para* θ *e* $\mathrm{l}(\theta|\mathbf{Y})$ *é duas vezes diferenciável com respeito a* θ*, então*

$$\mathrm{V}(\mathrm{T}) \geq \frac{1}{\mathrm{E}(-\mathrm{l}''(\theta|\mathbf{Y}))}.$$

Este teorema informa o limite inferior para a variância de um estimador $\hat{\mathrm{T}}$ qualquer. O estimador de máxima verossimilhança apresenta propriedades ótimas e uma delas é a eficiência, ou seja, assintóticamente o EMV atinge o limite inferior de Cramer-Rao. Na sequência discutimos como expressar a incerteza com relação ao vetor de parâmetros $\boldsymbol{\theta}$ através da construção de intervalos de confiança.

2.2 Intervalos de confiança

Definição 2.12. Intervalo de confiança - Um intervalo de verossimilhança para θ é um intervalo da forma $\theta : \mathrm{L}(\theta|\mathbf{y}) \geq r\mathrm{L}(\hat{\theta}|\mathbf{y})$ ou equivalentemente, $\theta : \mathrm{D}(\theta) \leq c^*$, com $\mathrm{D}(\theta) = -2[\mathrm{l}(\theta) - \mathrm{l}(\hat{\theta})]$ e $c^* = -2\log(r)$.

Esta definição é bastante geral para o caso uniparamétrico, para o caso multiparâmetros os princípios se mantêm e trocamos o intervalo de confiança por uma região de confiança, o que será abordado mais adiante. Nesta definição o valor de r precisa ser especificado entre 0 e 1, para intervalos não vazios, logo $c^* > 0$. Quanto maior o valor de c^* mais largo será o intervalo,

algumas vezes o intervalo pode ser a união de sub-intervalos disjuntos, porém este caso não é usual. Apesar do valor de c^* ser necessário para a construção dos intervalos ainda não temos elementos suficientes para especificá-lo.

Usando esta definição pode-se pensar ao menos duas formas de construção de intervalos de confiança. A primeira é considerar a quantidade $\frac{\mathrm{L}(\theta)}{\mathrm{L}(\hat{\theta})} \geq r$ que é a **verossimilhança relativa**, ou seja, compara cada valor de θ com o máximo. Nestas condições a verossimilhança relativa toma sempre valores entre 0 e 1 e o intervalo é a região do espaço paramétrico para qual os valores associados de verossimilhança sejam uma fração não menor que r do máximo valor. Por exemplo, definindo $r = 0.8$ estamos deixando que faça parte do intervalo de confiança valores que tenham até 80% de compatibilidade com a amostra observada, da mesma forma poderíamos definir $r = 0.20$ ou 0.50, dependendo de nosso critério. Royall (1997) propõe que este valor seja definido por analogias com resultados considerados aceitáveis em experimentos simples como lançamento de uma moeda. Porém, em grande parte dos problemas práticos uma interpretação probabilística baseada em idéias frequentistas é usada. Voltaremos a escolha do ponto de corte mais adiante quando apresentarmos as propriedades assintóticas do EMV. Uma forma equivalente é utilizar a função *deviance* definindo o intervalo pelos valores que satisfazem $\mathrm{D}(\theta) = -2[\mathrm{l}(\theta) - \mathrm{l}(\hat{\theta})] \leq -2\log(r)$. Esta é uma outra forma de considerar a verossimilhança relativa, agora em termos de diferença em log-verossimilhança. Neste caso a região de confiança pode ser definida como anteriormente ou valendo-se de propriedades frequentistas desta quantidade conforme veremos na sequência.

Em ambas abordagens surge o problema de que após definir o valor $c^* = -2\log(r)$, é necessário encontrar as raízes da função de verossimilhança relativa ou da *deviance* que fornecem os limites do intervalo de confiança para um c^* especificado. Em geral vamos chamar o valor c^* ou equivalentemente r de ponto de corte. Encontrar as raízes da função comumente envolve métodos numéricos, uma vez que na maioria das situações práticas não é possível obter expressões fechadas para os limites do intervalo.

Dado esta restrição é comum fazer uma expansão em séries de Taylor para a $\mathrm{l}(\theta)$ em torno de $\hat{\theta}$ de forma a facilitar a obtenção do intervalo de confiança. Expandindo $\mathrm{l}(\theta)$ em série de Taylor até segunda ordem em torno de $\hat{\theta}$, resulta na seguinte equação,

$$\mathrm{D}(\theta) = -2[\mathrm{l}(\theta) - \mathrm{l}(\hat{\theta})] = 2\left\{\mathrm{l}(\hat{\theta}) - [\mathrm{l}(\hat{\theta}) + (\theta - \hat{\theta})\mathrm{l}'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2\mathrm{l}''(\hat{\theta})]\right\}.$$

Como por definição do EMV $\mathrm{l}'(\hat{\theta}) = 0$, eliminando termos a aproximação em série de Taylor, toma a seguinte forma quadrática e define a região

$$\mathrm{D}(\theta) = -(\theta - \hat{\theta})^2 \mathrm{l}''(\hat{\theta}) \leq c^*.$$

que por sua vez, define intervalos de confiança da forma,

$$\hat{\theta} \pm \sqrt{\frac{c^*}{-\mathrm{l}''(\hat{\theta})}}.$$

Isto corresponde a fazer uma aproximação quadrática da função *deviance*, que torna o intervalo fácil de ser obtido. Estendendo para o caso de multi-parâmetros, tem-se que uma região de confiança para $\boldsymbol{\theta}$ é dada pelo conjunto $\boldsymbol{\theta} \in \Theta : \mathrm{D}(\boldsymbol{\theta}) \leq c^*$. Portanto, as duas formas de interpretar o intervalo de confiança discutidas no caso uniparamétrico podem ser estendidas para o caso multiparamétrico, sem problemas. Novamente a questão que surge é a definição de um valor para c^*. Pela abordagem frequentista é desejável que o intervalo tenha uma interpretação em termos de probabilidades ou frequência e isto é atingido através das propriedades assintóticas dos estimadores de máxima verossimilhança que serão apresentadas adiante, mas antes vamos materializar apresentar o terceiro objetivo da inferência estatística, ou seja, testes de hipóteses.

2.3 Testes de hipóteses

Nesta seção mostramos como diversos testes de hipóteses surgem naturalmente a partir da interpretação da função de verossimilhança.

Definição 2.13. Hipótese estatística - Chamamos de hipótese estatística qualquer afirmação acerca da distribuição de probabilidade de uma ou mais variáveis aleatórias.

Definição 2.14. Teste de hipótese - Chamamos de teste de uma hipótese estatística a função de decisão $\chi \to \{a_0, a_1\}$, em que a_0 corresponde à ação de considerar a hipótese H_0, como verdadeira e a_1 corresponde à ação de considerar a hipótese H_1 como verdadeira.

Na definição acima, χ denota o espaço amostral associado à amostra $y_1, y_2, \ldots, y_n$. A função de decisão d divide o espaço amostral χ em dois conjuntos,

$$A_0 = \{(y_1, \ldots, y_n \in \chi; d(y_1, \ldots, y_n) = a_0\}$$

e

$$A_1 = \{(y_1, \ldots, y_n \in \chi; d(y_1, \ldots, y_n) = a_1\}$$

onde $A_0 \cup A_1 = \chi$ e $A_0 \cap A_1 = \emptyset$. Como em A_0 temos os pontos amostrais que levam à não rejeição de H_0, vamos chamar de A_0 de região de não rejeição

e, por analogia, A_1 de região de rejeição de H_0, também chamada de região crítica.

Um teste de hipótese pode resultar em um de dois tipos de erros. Tradicionalmente, esses dois tipos de erros recebem os nomes de erro Tipo I (α) e erro Tipo II (β). O erro tipo I ocorre quando rejeitamos H_0 e esta é verdadeira. O erro Tipo II ocorre quando não rejeitamos H_0 e esta é falsa. Em termos de probabilidade temos,

$$\alpha = \mathrm{P}(Y \in A_1|\theta_0) \quad \text{e} \quad \beta = \mathrm{P}(Y \in A_0|\theta_1).$$

Definição 2.15. O poder do teste com região crítica A_1 para testar $\mathrm{H}_0 : \theta = \theta_0$ contra $\mathrm{H}_1 : \theta = \theta_1$ é dado por

$$\pi(\theta_1) = \mathrm{P}(Y \in A_1|\theta_1).$$

Note que $\pi(\theta_1) = 1 - \beta$, e β é a probabilidade do erro Tipo II.

Estas definições tratam o teste de hipótese diretamente como uma função de decisão que quantifica a incerteza associada com cada possível decisão.

Uma forma mais intuitiva de construir um teste de hipótese é definir uma estratégia para definir se um particular valor, digamos, θ_0 é plausível para o parâmetro θ. Como já discutido a função de verossimilhança nos fornece exatamente a plausibilidade de um determinado valor do parâmetro ser o gerador da amostra realizada dada uma função de probabilidade ou densidade probabilidade especificada. Assim, usar a função de verossimilhança para decidir sobre a plausibilidade de θ_0 é natural.

Considere o gráfico da função de log-verossimilhança apresentado na Figura **??**. Note que decidir sobre a plausibilidade do valor θ_0, consiste basicamente em medir o quanto longe ele está do valor mais plausível, ou seja, do EMV $\hat{\theta}$. Baseado na Figura **??** fica claro que tal distância pode ser medida de pelo menos três formas diferentes: no eixo das ordenadas, no eixo das abscissas ou verificando se a inclinação da reta tangente a θ_0 é diferente de zero. Estas três formas de medir a distância entre θ_0 e $\hat{\theta}$ levam a construção de três tipos de testes de hipóteses que serão discutidos nas próximos subseções.

Figure 2.1: Diferentes formas de construir teste de hipótese baseado em verossimilhança.

2.3.1 Teste da razão de verossimilhança

Se decidirmos pela distância medida pelo eixo das ordenadas, somos levados ao teste da razão de verossimilhança.

Definição 2.16. A estatística do teste da razão de verossimilhança para testar $\mathrm{H}_0 : \theta \in \Theta_0$ versus $\mathrm{H}_1 : \theta \in \Theta_0^c$ é

$$\lambda(\mathbf{y}) = \frac{\sup_{\Theta_0} \mathrm{L}(\theta|\mathbf{y})}{\sup_{\Theta} \mathrm{L}(\theta|\mathbf{y})},$$

onde sup denota o supremo de $\mathrm{L}(\theta|\mathbf{y})$ restrito ao conjunto Θ. O teste da razão de verossimilhança (TRV) é qualquer teste que tenha uma região de rejeição da forma $\mathbf{y} : \lambda(\mathbf{y}) \leq r$ onde r é qualquer número que satisfaça $0 \leq r \leq 1$.

Para testar $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta \neq \theta_0$, suponha $Y_1, \ldots, Y_n$ sejam iid $f(\mathbf{y}|\theta)$, $\hat{\theta}$ seja o EMV de θ, e $f(\mathbf{y}|\theta)$ satisfaça as condições de regularidade. Desse modo, de acordo com H_0, pelo Teorema **??** à medida que $n \to \infty$

$$-2 \log \lambda(\underline{y}) \to \chi^2_1.$$

2.3.2 Teste de Wald

Por outro lado, se optamos pela distância no eixo das abscissas temos o chamado teste de Wald. Suponha que deseja-se testar a hipótese bilateral $\mathrm{H}_0 : \theta = \theta_0$ versus $\mathrm{H}_1 : \theta \neq \theta_0$.

Definição 2.17. A estatística de Wald é dada por

$$\mathrm{Z_n} = (\hat{\theta} - \theta_0)/\sqrt{\mathrm{V}(\hat{\theta})}$$

e rejeita H_0 se, e somente se, $\mathrm{Z_n} < -z_{\alpha/2}$.

Se H_0 for verdadeira, então $\theta = \theta_0$ e Z_n convergem em distribuição para $Z \sim N(0,1)$. Portanto, a probabilidade do Erro Tipo I, $\mathrm{P}_{\theta_0}(\mathrm{Z_n} < -z_{\alpha/2} \quad \text{ou} \quad \mathrm{Z_n} > z_{\alpha/2}) \to \mathrm{P}(Z < -z_{\alpha/2} \quad \text{ou} \quad Z > z_{\alpha/2}) = \alpha$, e este é, assintoticamente, um teste de tamanho α. Em geral, um teste de Wald é um teste com base em uma estatística da forma,

$$\mathrm{Z_n} = \frac{\mathrm{W_n} - \theta_0}{\mathrm{S_n}}$$

onde θ_0 é um valor hipotético do parâmetro θ, $\mathrm{W_n}$ é um estimador de θ e $\mathrm{S_n}$ é o erro padrão de $\mathrm{W_n}$, uma estimativa do desvio padrão de $\mathrm{W_n}$. Se $\mathrm{W_n}$ for o EMV para θ, então, $\sqrt{\mathrm{I_O}(\hat{\theta})}$ é o erro padrão de $\mathrm{W_n}$.

2.3.3 Teste escore

Por fim, podemos olhar para a inclinação da reta tangente no ponto θ_0 o que leva ao chamado teste escore. Lembre-se que a estatística escore é definida como

$$\mathrm{U}(\theta) = \frac{\partial}{\partial \theta} \mathrm{l}(\theta|\boldsymbol{Y}).$$

Sabemos (ver **??**) que para todo θ, $\mathrm{E}_\theta(\mathrm{U}(\theta)) = 0$. Em particular, se estivermos testando $\mathrm{H}_0 : \theta = \theta_0$ e se H_0 for verdadeira, então $\mathrm{U}(\theta)$ tem média 0. Além disso,

$$\mathrm{V}_\theta(\mathrm{U}(\theta)) = -\mathrm{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\mathrm{l}(\theta|\boldsymbol{Y})\right) = \mathrm{I}_\mathrm{E}(\theta)$$

ou seja, o número de informações é a variância da estatística escore.

Definição 2.18. A estatística de teste escore é

$$\mathrm{Z}_\mathrm{S} = \mathrm{U}(\theta_0)/\sqrt{\mathrm{I}_\mathrm{E}(\theta_0)}.$$

Se H_0 for verdadeira, Z_S tem distribuição normal com média 0 e variância 1.

2.4 Propriedades do EMV

Apesar de definirmos a função de verossimilhança como uma quantidade fixa avaliada em $\mathbf{y}$, devemos lembrar que ela é baseada em apenas uma realização do vetor aleatório $\mathbf{Y}$, sendo assim, estudar o comportamento probabilístico dos estimadores de máxima verossimilhança é de fundamental importância para definir suas propriedades probabilística e baseado nisto obter intervalos de confiança e testes de hipóteses com interpretações probabilísticas. Para isto, vamos precisar de mais algumas definições.

Definição 2.19. Função escore - Sendo $\mathrm{l}(\boldsymbol{\theta}|\mathbf{y})$ a função de log-verossimilhança, o vetor *escore* é definido por

$$\mathrm{U}(\boldsymbol{\theta}) = \left(\frac{\partial \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_p}\right)^\top.$$

Note que a função escore nada mais é que o vetor gradiente da função de log-verossimilhança. Definimos as matrizes de informação **observada** e **esperada**, também chamada de matriz de informação de Fisher.

Definição 2.20. Matriz de informação observada - Sendo $\mathrm{l}(\boldsymbol{\theta}|\mathbf{y})$ a função de log-verossimilhança, a matriz de informação observada é definida por

$$\mathrm{I}_O(\boldsymbol{\theta}) = \begin{bmatrix} -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_1^2} & \cdots & \cdots & -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_1\partial\theta_p} \\ \vdots & \ddots & -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} & \vdots \\ \vdots & -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_i} & \ddots & \vdots \\ -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_p\partial\theta_1} & \cdots & \cdots & -\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial\theta_p^2} \end{bmatrix}.$$

Definição 2.21. Matriz de informação esperada - Sendo $\mathrm{l}(\boldsymbol{\theta}|\mathbf{y})$ a função de log-verossimilhança, a matriz de informação esperada é definida por

$$\mathrm{I}_E(\boldsymbol{\theta}) = \begin{bmatrix} \mathrm{E}\left[-\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial \theta_1^2}\right] & \cdots & \cdots & \mathrm{E}\left[-\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p}\right] \\ \vdots & \ddots & \mathrm{E}\left[-\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right] & \vdots \\ \vdots & \mathrm{E}\left[-\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_i}\right] & \ddots & \vdots \\ \mathrm{E}\left[-\frac{\partial^2 \mathrm{l}(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1}\right] & \cdots & \cdots & \mathrm{E}\left[-\frac{\partial^2 \mathrm{l}(\theta)}{\partial \theta_p^2}\right] \end{bmatrix}.$$

Duas propriedades importantes da função escore são apresentadas nos Teoremas a seguir.

Teorema 2.2. ***Primeira igualdade de Bartlett*** - *Sendo* $\mathrm{U}(\boldsymbol{\theta})$ *a função escore, então*

$$\mathrm{E}(\mathrm{U}(\boldsymbol{\theta})) = 0.$$

Teorema 2.3. ***Segunda igualdade de Bartlett*** - *Sendo* $\mathrm{U}(\boldsymbol{\theta})$ *a função escore, então*

$$\mathrm{V}(\mathrm{U}(\boldsymbol{\theta})) = \mathrm{E}(\mathrm{I}_O(\boldsymbol{\theta})) = \mathrm{I}_E(\boldsymbol{\theta}).$$

Note que a variância do vetor $\mathrm{U}(\boldsymbol{\theta})$ é a matriz com entradas

$$\begin{bmatrix} \mathrm{Cov}(\mathrm{U}_1, \mathrm{U}_1) & \cdots & \cdots & \mathrm{Cov}(\mathrm{U}_1, \mathrm{U}_\mathrm{d}) \\ \vdots & \ddots & \mathrm{Cov}(\mathrm{U}_\mathrm{i}, \mathrm{U}_\mathrm{j}) & \vdots \\ \vdots & \mathrm{Cov}(\mathrm{U}_\mathrm{j}, \mathrm{U}_\mathrm{i}) & \ddots & \vdots \\ \mathrm{Cov}(\mathrm{U}_\mathrm{d}, \mathrm{U}_1) & \cdots & \cdots & \mathrm{Cov}(\mathrm{U}_\mathrm{d}, \mathrm{U}_\mathrm{d}) \end{bmatrix}.$$

onde $\mathrm{Cov}(\mathrm{U}_\mathrm{i}, \mathrm{U}_\mathrm{i}) = \mathrm{V}(\mathrm{U}_\mathrm{i})$. Uma propriedade importante de $\mathrm{I}_O(\hat{\boldsymbol{\theta}})$ e $\mathrm{I}_E(\hat{\boldsymbol{\theta}})$ é que elas são matrizes definida positiva, as quais mensuram a curvatura observada/esperada da superfície de log-verossimilhança. Com estas definições, pode-se escrever a função *deviance* aproximada para um vetor de parâmetros da seguinte forma:

$$\mathrm{D}(\boldsymbol{\theta}) \approx (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathrm{I}_\mathrm{O}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Assim $\mathrm{D}(\boldsymbol{\theta})$ é não negativa uma vez que $\mathrm{I}_\mathrm{O}(\hat{\boldsymbol{\theta}})$ é uma matriz positiva definida. Uma vez definidas as quantidades envolvidas, estamos aptos a enunciar o teorema a seguir.

Teorema 2.4. ***Distribuição assintótica do EMV*** *- Para um problema de estimação regular, no limite com $n \to \infty$, se $\boldsymbol{\theta}$ é o verdadeiro vetor de parâmetros, então*

$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, \mathrm{I_E}(\boldsymbol{\theta})^{-1}),$$

ou seja, a distribuição assintótica de $\hat{\boldsymbol{\theta}}$ é uma normal multivariada com matriz de variância/covariância dada pela inversa da matriz de informação esperada.

Corolário 2.1. *Qualquer termo assintóticamente equivalente a $\mathrm{I_E}(\boldsymbol{\theta})$ pode ser usado no Teorema* **??***. Assim,*

$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, {\mathrm{I_E}}^{-1}(\hat{\boldsymbol{\theta}}))$$

$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, {\mathrm{I_O}}^{-1}(\boldsymbol{\theta}))$$

$$\hat{\boldsymbol{\theta}} \sim NM_p(\boldsymbol{\theta}, {\mathrm{I_O}}^{-1}(\hat{\boldsymbol{\theta}})).$$

Teorema 2.5. ***Distribuição assintótica da deviance*** *- Para um problema de estimação regular, no limite com $n \to \infty$, se $\boldsymbol{\theta}$ é o verdadeiro valor do parâmetro, então*

$$\mathrm{D}(\boldsymbol{\theta}) = -2[\mathrm{l}(\boldsymbol{\theta}) - \mathrm{l}(\hat{\boldsymbol{\theta}})] \sim \chi^2_d$$

ou seja, a função deviance segue uma distribuição qui-Quadrado com p graus de liberdade, onde p é a dimensão do vetor $\boldsymbol{\theta}$.

De acordo com os teoremas apresentados, podemos chegar a algumas das principais propriedades dos estimadores de máxima verossimilhança:

- O estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ é assintóticamente não-viciado, isto é, $\mathrm{E}(\hat{\boldsymbol{\theta}}) \to \boldsymbol{\theta}$.
- Assintóticamente $\mathrm{V}(\hat{\boldsymbol{\theta}}) \to {\mathrm{I_E}}^{-1}(\boldsymbol{\theta})$, o qual por uma versão multivariada do limite de Cramér-Rao é o melhor possível, mostrando que o EMV é eficiente para o vetor $\boldsymbol{\theta}$, ao menos para grandes amostras.
- Denote $\mathrm{J} = {\mathrm{I_E}}^{-1}(\boldsymbol{\theta})$, então $\mathrm{V}(\hat{\boldsymbol{\theta}}) = \mathrm{J}$, sendo que, J é uma matriz simétrica e definida positiva, com elementos $\mathrm{J_{ij}} = \mathrm{Cov}(\hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\theta}}_j)$ então $\mathrm{J_{ii}}$ é a variância de $\hat{\boldsymbol{\theta}}_i$. Denota-se ${\mathrm{J_{ii}}}^{\frac{1}{2}}$ o desvio padrão de $\hat{\boldsymbol{\theta}}_i$.
- Podemos construir intervalos de $100(1-\alpha)\%$ de confiança para θ_i na forma $\hat{\theta}_i \pm z_{\frac{\alpha}{2}} {\mathrm{J_{ii}}}^{\frac{1}{2}}$. Intervalos desta forma serão denominados, intervalos de Wald ou baseados em aproximação quadrática da verossimilhança. Importante notar que estes intervalos coincidem com os obtidos baseado em aproximação por Série de Taylor de segunda ordem da função *deviance*.

- Para regiões de confiança baseados na *deviance* considera-se $[\boldsymbol{\theta} \in \Theta : \mathrm{D}(\boldsymbol{\theta}) \leq c^*]$, *para algum valor* c^* a ser especificado. Pode-se escolher c^* baseado em justificativas assintóticas de que $\mathrm{D}(\boldsymbol{\theta}) \sim \chi^2_p$ é uma escolha razoável para $c^* = c_\alpha$ com $\mathrm{P}(\chi^2_d \geq c_\alpha) = \alpha$, por exemplo se $\alpha = 0.05$, então $c_\alpha = 3.84$. Isto gera uma região de $100(1-\alpha)\%$ de confiança. Estes intervalos serão denominados de intervalos *deviance*.

De acordo com as propriedades apresentadas tem-se duas formas básicas de construir intervalos de confiança. A primeira mais simples é baseada na aproximação quadrática da log-verossimilhança e a segunda utilizando diretamente a função *deviance* obtida com os dados. A segunda opção é em geral mais trabalhosa computacionalmente, uma vez que usualmente gera uma equação não linear que precisa ser resolvida numericamente. A primeira opção é bastante direta, uma vez obtida a matriz de segundas derivadas basta invertê-la e tirar a raiz dos termos da diagonal para se obter o intervalo de confiança para cada parâmetro, marginalmente. Esta abordagem é muito simples mas apresenta limitações. Restrições naturais no espaço paramétrico como, por exemplo, para parâmetros de variância e correlação não são respeitadas e podem resultar em limites absurdos, com limite(s) do intervalo fora do espaço paramétrico. Os intervalos serão sempre simétricos ao aproximar a verossimilhança por uma forma quadrática, o que normalmente não produz resultados adequados para parâmetros de variância e correlação. Em modelos com efeitos aleatórios há um interesse natural nos parâmetros de variância, precisão e correlação. Testar a significância de tais efeitos utilizando as variâncias associadas às estimativas que indexam o modelo podem produzir resultados imprecisos. Logo, esta abordagem é limitida em classes mais gerais de modelos estatísticos.

A segunda opção resulta em uma região conjunta para o caso de dois ou mais parâmetros, enquanto que pela aproximação é possível obter um intervalo marginal para cada parâmetro, porém baseado em uma aproximação quadrática da superfície de log-verossimilhança. Este tipo de representação é a mais desejável para inferência, porém não pode ser obtida diretamente apenas com o Teorema **??**. Por exemplo, suponha que tem-se interesse em um determinado componente do vetor de parâmetros, digamos θ_i. A partir da aproximação quadrática podemos facilmente construir um intervalo de confiança, tendo como $\hat{\theta}_I$ e $\hat{\theta}_S$ o seu limite inferior e superior, respectivamente. Pelo Teorema **??** para o caso em que a dimensão de $\boldsymbol{\theta}$ é maior que um, não temos um intervalo desta forma mas sim uma região, o que apesar de mais informativa tem menor apelo prático e apresenta dificuldades de interpretação. Uma forma intuitiva de obter um intervalo da forma $\hat{\theta}_I$ e $\hat{\theta}_S$ é fixar o restante do vetor de parâmetros nas suas estimativas de máxima verossimilhança e obter os limites em uma direção de cada vez. Esta abordagem tem uma clara

restrição que é não levar em consideração a incerteza associada ao restante do vetor de parâmetros para a construção do intervalo.

Temos um método simples via aproximação quadrática, porém que não funciona bem quando a superfície de log-verossimilhança é assimétrica. Por outro lado, o método baseado na função *deviance* não apresenta esta restrição mas fornece regiões de confiança conjuntas, e não diretamente limites $\hat{\theta}_I$ e $\hat{\theta}_S$ para cada parâmetro. Duas abordagens básicas para este problema podem ser consideradas: a primeira é fazer uma reparametrização do modelo nos parâmetros que apresentam forte assimetria ou são restritos, para torná-los irrestritos e aproximadamente simétricos, obter a variância baseada na aproximação quadrática nesta reparametrização e depois converter para a escala original. Quando este procedimento é satisfatório o custo computacional é baixo. Uma outra opção é construir uma nova verossimilhança que não dependa dos outros parâmetros, de forma que podemos atuar como no caso uniparamétrico. Estas duas alternativas serão discutidas na próxima seção.

2.5 Reparametrização e verossimilhança perfilhada

Considere o problema de obter a estimativa pontual e intervalar para um parâmetro de interesse $\phi = \mathrm{g}(\boldsymbol{\theta})$, onde $\mathrm{g}(\cdot)$ é uma função e, desde que $\mathrm{L}(\phi) = \mathrm{L}(\mathrm{g}(\boldsymbol{\theta}))$, a função de verossimilhança para ϕ é obtida da função de verossimilhança de $\boldsymbol{\theta}$ por uma transformação de escala. Consequentemente, como $\hat{\phi} = \mathrm{g}(\hat{\boldsymbol{\theta}})$, quando o intervalo de confiança digamos $\hat{\theta}_I$ e $\hat{\theta}_S$ for obtido diretamente pela função de verossimilhança, log-verossimilhança ou *deviance*, o intervalo para ϕ pode ser obtido simplesmente transformando os limites obtidos para θ, no caso unidimensional. Esta propriedade é conhecida como **invariância** do estimador de máxima verossimilhança. Porém, quando o intervalo for obtido pela aproximação quadrática isso não é válido e um Teorema adicional é necessário para esta transformação.

Teorema 2.6. *Considere obter um intervalo de confiança para $\phi = \mathrm{g}(\boldsymbol{\theta})$ por invariância temos que $\hat{\phi} = \mathrm{g}(\hat{\boldsymbol{\theta}})$ e a variância de $\hat{\phi}$ é dada por*

$$\mathrm{V}(\hat{\phi}) = \mathrm{V}(\mathrm{g}(\hat{\boldsymbol{\theta}})) = \nabla \mathrm{g}(\hat{\boldsymbol{\theta}})^\top \mathrm{I_E}(\hat{\boldsymbol{\theta}})^{-1} \nabla \mathrm{g}(\hat{\boldsymbol{\theta}})$$

com

$$\nabla \mathrm{g}(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial \mathrm{g}(\hat{\boldsymbol{\theta}})}{\partial \theta_1}, \ldots, \frac{\partial \mathrm{g}(\hat{\boldsymbol{\theta}})}{\partial \theta_d} \right)^\top.$$

A partir do Teorema **??** é imediato o seguinte teorema.

Teorema 2.7. *Para um problema de estimação regular se $\phi = g(\boldsymbol{\theta})$ são os verdadeiros valores dos parâmetros, então quando $n \to \infty$ tem-se que*

$$\hat{\boldsymbol{\phi}} \sim \mathrm{NM_p}(\phi, \nabla \mathrm{g}(\boldsymbol{\theta})^\top \mathrm{I_E}(\boldsymbol{\theta})^{-1} \nabla \mathrm{g}(\boldsymbol{\theta})).$$

Pelo Teorema **??**, podemos construir intervalos de confiança da mesma forma anterior, porém usando a nova matriz de variância e covariância ponderada pelo gradiente da função $\mathrm{g}(\cdot)$, e assim passar de uma reparametrização para outra torna-se uma tarefa trivial. Apesar deste procedimento ser bastante útil, nem sempre é fácil encontrar uma transformação $\mathrm{g}(\cdot)$ que torne a log-verossimilhança simétrica. A forma mais efetiva de construir intervalos de confiança para parâmetros de difícil estimação é o intervalo baseado em **perfil de verossimilhança**.

Seja $\boldsymbol{\theta} = (\boldsymbol{\phi}^\top, \boldsymbol{\lambda}^\top)^\top$, o vetor de parâmetros particionado nos vetores $\boldsymbol{\phi}$ e $\boldsymbol{\lambda}$, vamos chamar a primeira componente de interesse e a segunda de incômodo, no sentido que desejamos intervalos ou regiões de confiança para $\boldsymbol{\phi}$, que pode ser apenas um escalar. Seja $\mathrm{L}(\boldsymbol{\phi}, \boldsymbol{\lambda})$ a verossimilhança para $\boldsymbol{\phi}$ e $\boldsymbol{\lambda}$. Denota-se $\hat{\boldsymbol{\lambda}}_\phi$ a estimativa de máxima verossimilhança de $\boldsymbol{\lambda}$ para dado um valor para $\boldsymbol{\phi}$.

Definição 2.22. Verossimilhança perfilhada - A verossimilhança perfilhada de $\boldsymbol{\phi}$ é definida por

$$\mathrm{L}(\boldsymbol{\phi}) = \mathrm{L}(\boldsymbol{\phi}, \hat{\boldsymbol{\lambda}}_\phi)$$

A forma apresentada na definição **??** sugere um procedimento de maximização em duas etapas. A primeira consiste em obter $\hat{\boldsymbol{\lambda}}_\phi$ que maximiza $\mathrm{l}(\boldsymbol{\phi}, \boldsymbol{\lambda}) = \log \mathrm{L}(\boldsymbol{\phi}, \boldsymbol{\lambda})$ com respeito a $\boldsymbol{\lambda}$ supondo $\boldsymbol{\phi}$ fixo. A seguir maximiza-se $\mathrm{l}(\boldsymbol{\phi})$. Assim, uma região ou intervalo de confiança para $\boldsymbol{\phi}$ pode ser obtida usando que

$$\mathrm{D}(\boldsymbol{\phi}) = -2[\mathrm{l}(\boldsymbol{\phi}) - \mathrm{l}(\hat{\boldsymbol{\phi}})] \sim \chi^2_d$$

onde d é a dimensão de $\boldsymbol{\phi}$. Note que esta forma de construção não usa a aproximação em séries de Taylor e portanto pode resultar em intervalos assimétricos. Porém, é cara computacionalmente, uma vez que precisamos resolver numéricamente uma equação não-linear que para cada avaliação necessita de um algoritmo numérico de maximização.

Neste Capítulo apresentamos uma série de definições e propriedades do procedimento de inferência baseado na função de verossimilhança. Os exemplos de ilustração foram o mais simples possível, porém mesmo nestes casos a necessidade do uso de métodos numéricos ficou bastante evidente. No próximo capítulo vamos discutir uma série de exemplos que em geral tem soluções analiticas, porém vamos resolvê-los também numéricamente para ilustrar as

principais ideias e desafios relacionados a obtenção de estimadores pontuais, intervalares e testes de hipóteses.

Royall, R. 1997. *Statistical Evidence*. Chapman; Hall.