Universidade Estadual de Campinas
Departamento de Estatística — IMECC
Disciplina: MI402 – Inferência Estatística
Professor: Dr. Caio Lucidius Naberezny Azevedo
Período: 2° semestre de 2019
Acadêmico: André Felipe B. Menezes

Notas de Aulas de Inferência Estatística

Este documento contém um resumo das notas de aulas da disciplina MI402 - Inferência Estatística do mestrado em estatística na Unicamp.

O professor utilizou slides nas suas aulas, porém todos, sem nenhuma exceção, continham erros de notação, ortográficos e principalmente gramática. Isso fazia com que ele interrompesse as aulas para corrigi-los o que fazia a aula não fluir.

Campinas
Dezembro de 2019

Sumário

1 Introdução

Nesta seção, apresento uma discussão filosófica, sem formalidades matemáticas, sobre o conceito de inferência estatística. Por fim, as principais famílias de distribuições de probabilidade são introduzidas e algumas propriedades são enfatizadas.

Importante destacar que estas notas de aula tem como referência os seguintes livros: Rohatgi (1976), Rohatgi (1984), Lindsey (1996), Azzalini (1996), Mukhopadhyay (2000), Casella e Berger (2002),

1.1 Probabilidade, Estatística e Inferência

Alegações probabilísticas são parte integrante do nosso vocabulário. É muito comum utilizarmos e/ou escutarmos palavras como aleatório, chance, risco, verossímil, provável, plausível, possível, ou até mesmo expressões como tão provável quanto, mais frequente do que, quase certo, provável mas não possível, entre outras. Todas essas palavras e frases são utilizadas para transmitir um certo grau de incerteza. O objetivo principal da probabilidade é mensurar numericamente a incerteza (o acaso). Na probabilidade utilizamos todo o ferramental matemático para descrever fenômenos físicos, artificiais ou antropogênicos que apresentam incerteza.

A importância da estatística nos dias de hoje é algo intrínseco. Embora, muitas pessoas associam a estatística com fatos numéricos (dados). Essa intuição não esta completamente errônea, pois a ciência estatística lida com a coleta e descrição dos dados. No entanto, o estatístico esta presente no planejamento do experimento, na coleta das informações e na decisão sobre a melhor forma de utilizar as informações coletadas afim de fornecer uma diretriz para tomada de decisão. A inferência estatística surge neste último aspecto, como sendo a arte de avaliar informações para extrair conclusões confiáveis sobre a natureza do fenômeno em estudo.

Os fenômenos naturais, artificiais ou antropogênicos são descritos por modelos, isto é, idealizações matemáticas utilizadas para aproximar um fenômeno que pode ser observável. Nessa idealização, certas suposições são feitas e então certos detalhes são ignorados como sendo não importantes. O sucesso do modelo vai depender se as suposições empregadas são válidas e se os detalhes ignorados são de fato não importantes. Para avaliar a adequação do modelo, isto é, se o modelo proposto descreve adequadamente o fenômeno em estudo devemos realizar um experimento e então observar o comportamento do fenômeno.

É importante distinguir a natureza do modelo que será adotado para descrever o fenômeno. Os modelos puramente matemático estipulam que as condições nas quais o experimento é executado determinam precisamente o resultado do experimento, tais modelos são denominados de modelos *determinístico*. Exemplos clássicos de modelos determinísticos são: equação da aceleração, as leis gravitacionais e as leis de Kepler.

Em contrapartida, os modelos *estocásticos* ou *probabilísticos* não estipulam uma regra para determinar exatamente o resultado final do experimento, uma vez que levam em consideração o caráter aleatório do fenômeno. Um modelo estocástico é caracterizado pelas seguintes componentes:

(i) Todos os resultados possíveis do experimento;

(ii) Todos os eventos de interesse;

(iii) Atribuição de probabilidades aos eventos de interesse.

A parte mais importante e complicada do modelo estocástico é atribuir probabilidades aos eventos de interesse. Neste curso de inferência iremos nos restringir em estudar modelos probabilísticos caracterizados por distribuições de probabilidade, especificamente as famílias de locação-escala e exponencial. Entretanto, vale mencionar que existem infinitos modelos probabilísticos mais complexos que não necessariamente são caracterizados por distribuições de probabilidade.

O conjunto de modelos sob consideração são os **modelos estatísticos paramétricos** definidos pela tripla $(\mathcal{X}, \mathcal{P}, \Theta)$, em que

- $\mathcal{X}$ é o espaço amostral, medido com base na variável aleatória;
- $\mathcal{P}$ é uma família de função de distribuição, $P_{\boldsymbol{\theta}}$, indexada pelo (vetor) parâmetro(s) $\boldsymbol{\theta}$;
- Θ é o espaço paramétrico definido conforme os parâmetros do modelo.

Por fim, é fundamental diferenciarmos entre estatística e probabilidade. Em probabilidade, fazemos certas suposições sobre a população e então concluímos algo sobre a amostra. Ou seja, o problema é: Dado um modelo estocástico (que representa um fenômeno), o que podemos dizer sobre seus resultados? Já na estatística, o processo é reverso. Dado uma amostra (conjunto de resultados), o que podemos dizer sobre a população (fenômeno que é representado por um modelo)? Mesmo com essas distinções vale notar que a estatística utiliza-se de um modelo probabilístico para tomar decisões acerca do fenômeno.

1.2 Famílias Comuns de Distribuições

Como o curso tem um viés totalmente paramétrico revisamos nas duas primeiras semanas de aula as famílias de distribuições exponencial e locação-escala. Uma atenção mais detalhada foi dada a família exponencial.

1.2.1 Família Exponencial

O conceito de família exponencial surgiu nos anos de 1935 e 1936. No entanto, foi na década de 1970 que com o surgimento dos modelos lineares generalizados que a família exponencial de distribuições obtiveram maior visibilidade. Nesta seção, apresentarei as principais características da família exponencial.

Seja X uma variável aleatória definida no espaço de probabilidade $(\Omega, \mathbb{A}, P)$. Então, a distribuiçao de X pertence a família exponencial k- paramétrica se sua função massa de probabilidade ou função densidade de probabilidade, puder ser expressa como

$$f(x \mid \boldsymbol{\theta}) = h(x) \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j(x) + d(\boldsymbol{\theta})\right\} I_A(x) \tag{1}$$

em que $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$; $h : \mathbb{R} \to \mathbb{R}^+$ é uma função que não depende de $\boldsymbol{\theta}$; $c_j : \mathbb{R}^k \to \mathbb{R}$ são funções que não dependem de x; $t_j : \mathbb{R} \to \mathbb{R}$ são funções que não dependem de $\boldsymbol{\theta}$; $d : \mathbb{R}^k \to \mathbb{R}$ é uma função que não depende de x; e $A \in \mathbb{R}$ não depende de $\boldsymbol{\theta}$.

A forma (1) permite generalizar resultados matemáticos e propriedades estatística úteis para a inferência acerca dos parâmetros bem como a caracterização da distribuição.

A seguir apresento dois resultados bastante proveitoso sob o ponto de vista inferencial relacionado a familia exponencial. Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes e identicamente com distribuição pertencente a família exponencial k-paramétrica. Então, a distribuição do vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_n)$ também pertence a família exponencial k-paramétrica. Além disso, a distribuição de $T = t(\boldsymbol{X}) = \sum_{i=1}^{n} t(X_i)$ pertence a familia exponencial. Importante mencionar que o conceito de familia exponencial bem como os resultados enunciados são passíveis de generalizações para de vetores aleatórios.

É comum utilizar uma parametrização canônica ou natural para distribuições de probabilidade pertencentes a família exponencial. Definindo $\eta_j = c_j(\boldsymbol{\theta}), j = 1, \ldots, k$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^\top$ e $d_0(\boldsymbol{\eta}) = d(\boldsymbol{\eta})$, podemos escrever (1) da seguinte forma

$$f(x \mid \boldsymbol{\eta}) = h(x) \exp\left\{\sum_{j=1}^{k} \eta_j t_j(x) + d_0(\boldsymbol{\eta})\right\} I_A(x). \tag{2}$$

Se para todo j, η_j for uma transformação bijetora (1 a 1) então $d_0(\boldsymbol{\eta}) = d_0(c^{-1}(\boldsymbol{\eta}))$, em que $c^{-1}(\cdot) = (c_1^{-1}(\cdot), \ldots, c_j^{-1}(\cdot))$.

Seja X uma variável aleatória com distribuição pertence a familia exponencial uniparamétrica. Então os seguintes resultados são válidos para a variável aleatória $T = t(X)$.

- A função geradora de momentos de T é dada por

$$M_T(s) = e^{d_0(\eta) - d_0(s+\eta)}, \forall\, s \in (-h, h) \text{ para algum } h > 0$$

- Utilizando a $M_T(s)$, determinamos os seguintes momentos:

$$\mathbb{E}(T) = -\frac{d}{d\eta}\, d_0(\eta) \qquad \text{e} \qquad \mathrm{Var}(T) = -\frac{d^2}{d\eta^2}\, d_0(\eta)$$

Tais resultados podem ser generalizados para o caso geral k-paramétrica bem como para uma amostra aleatória.

1.2.2 Família Locação-Escala

Estas famílias de distribuições considera somente variáveis aleatórias contínuas. De fato, a partir de uma função densidade de probabilidade, $f_X(x)$, associada a uma variável aleatória X qualquer, obtemos a família de locação cuja f.d.p. é dada por

$$f_Z(z \mid \mu) = f_X(z - \mu)$$

em que $\mu \in \mathbb{R}$ é o parâmetro de locação. Observe que a nova densidade esta associada a variável aleatória $Z = X - \mu$.

Por outro lado, considerando f.d.p. $f_X(x)$ temos a família de escola, cuja a f.d.p. é dada por

$$f_W(w \mid \sigma) = \frac{1}{\sigma} f_X\left(\frac{w}{\sigma}\right)$$

em que $\sigma > 0$ é o parâmetro de escala. Observe que a nova densidade esta associada a variável aleatória $W = \sigma X$.

Por fim, têm-se a família locação-escala cuja f.d.p. é definida por

$$f_Y(y \mid \mu, \sigma) = \frac{1}{\sigma} f_X\left(\frac{y - \mu}{\sigma}\right)$$

em que $\mu \in \mathbb{R}$ e $\sigma > 0$ são os parâmetros de locação e escala, respectivamente. Observe que a nova densidade esta associada a variável aleatória $Y = \dfrac{X - \mu}{\sigma}$.

O efeito de introduzir um parâmetro de locação (μ) não altera o formato da função densidade de probabilidade, porém altera a localização em que a f.d.p. esta situada. Já o parâmetro escala (σ) afeta o formato da f.d.p., sendo que quanto maior seu valor mais esparsa a densidade.

Por fim, enunciamos dois teoremas que formalizam resultados importantes na familia de distribuições locação-escala. A prova dos teoremas pode ser encontrada em Casella e Berger (2002).

Teorema 1.1. *Seja $f(\cdot)$ uma função densidade de probabilidade (f.d.p.) e seja $\mu \in \mathbb{R}$ e $\sigma > 0$. Então, X é uma variável aleatória com f.d.p. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ se, e somente se, existir uma variável aleatória Z com f.d.p. $f(z)$ e $X = \sigma Z + \mu$*

Teorema 1.2. *Seja Z uma variável aleatória com f.d.p. $f(z)$ e $X = \sigma Z + \mu$. Suponha que $\mathbb{E}(Z)$ e $\mathrm{Var}(Z)$ existam. Se X é uma variável aleatória com f.d.p. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, então*

$$\mathbb{E}(X) = \sigma\mathbb{E}(Z) + \mu \qquad e \qquad \mathrm{Var}(Z) = \sigma^2 \mathrm{Var}(Z).$$

2 Princípios da Redução de Dados

Em resumo, a redução de dados consiste em resumir toda a informação disponível na amostra coletada ($\boldsymbol{x}$) de tal maneira que não perdemos a essência do mecanismo gerador dos dados.

Seja $\boldsymbol{X} = (X_1, \ldots, X_n)$ um vetor aleatório composto por variáveis aleatórias independente e identicamente distribuídas e seja $\boldsymbol{x} = (x_1, \ldots, x_n)$ uma amostra aleatória de $\boldsymbol{X}$. Então temos a seguinte definição

Definição 2.1. Qualquer função da amostra $T = t(\boldsymbol{X})$ é uma estatística se $t(\cdot)$ é uma função que não depende dos parâmetros $\boldsymbol{\theta}$.

Observe que qualquer estatística $T = t(\boldsymbol{X})$ é também uma variável aleatória, pois $\boldsymbol{X}$ é um vetor aleatório. A redução de dados utilizando um estatística em particular pode se considerada um partição do espaço amostral. Seja $\mathcal{T} = \{t : t = t(\boldsymbol{x}) \text{ para algum } \boldsymbol{x} \in \mathcal{X}\}$ a imagem de $\mathcal{X}$ em $t(\boldsymbol{x})$. Então, $t(\boldsymbol{x})$ realiza a partição do espaço amostral em conjuntos $A_t, t \in \mathcal{T}$, definidos por $A_t = \{\boldsymbol{x} : t(\boldsymbol{x}) = t\}$.

Como o objeto de estudo deste curso são os modelos paramétricos, ou seja, estamos supondo que o mecanismo gerador dos dados pode ser bem descrito por uma distribuição de probabilidade indexada por um vetor de parâmetros ($\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$). Então, buscaremos métodos para redução dos dados que não desconsideram informações importante sobre $\boldsymbol{\theta}$.

2.1 Princípio da Suficiência

Este princípio é baseado na definição de estatística suficiente.

Definição 2.2. Uma estatística $T = t(\boldsymbol{X})$ é uma estatística suficiente para $\boldsymbol{\theta}$ se a distribuição condicional da amostra $\boldsymbol{X}$, dado o valor de T, não depende de $\boldsymbol{\theta}$.

Indo além desta definição podemos afirmar que uma estatística suficiente para $\boldsymbol{\theta}$ contém todas as informações sobre $\boldsymbol{\theta}$ que estão na amostra.

O princípio da suficiência afirma que se $T = t(\boldsymbol{X})$ é uma estatística suficiente para $\boldsymbol{\theta}$, então qualquer inferência sobre $\boldsymbol{\theta}$ deverá depender da amostra $\boldsymbol{x}$ somente pelo valor $t(\boldsymbol{x})$. Isto é, se $\boldsymbol{x}$ e $\boldsymbol{y}$ são amostras distintas de modo que $t(\boldsymbol{x}) = t(\boldsymbol{y})$, então a inferência sobre $\boldsymbol{\theta}$ deverá ser a mesma , se $\boldsymbol{X} = \boldsymbol{x}$ ou $\boldsymbol{X} = \boldsymbol{y}$ for observado.

Teorema 2.1. *Se $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ é a f.d.p.[1] ou f.m.p.[2] conjunta de $\boldsymbol{X}$ e $q(t \mid \boldsymbol{\theta})$ a f.d.p. ou f.m.p. de $t(\boldsymbol{X})$, então $t(\boldsymbol{X})$ é uma estatística suficiente para $\boldsymbol{\theta}$ se para todo $\boldsymbol{x}$ no espaço amostral a razão*

$$\frac{f(\boldsymbol{x} \mid \boldsymbol{\theta})}{q(t \mid \boldsymbol{\theta})}$$

não depender de $\boldsymbol{\theta}$.

Este teorema é útil quando conhecemos a distribuição de $t(\boldsymbol{X})$, em particular a forma funcional da f.d.p. ou f.m.p. $q(t \mid \boldsymbol{\theta})$. No entanto, isso nem sempre é factível. Para enfrentar este problema, temos o Teorema (critério) da Fatoração introduzido pelo estatístico sueco Jerzy Neyman.

Teorema 2.2. *Seja $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ a f.d.p. ou f.m.p. de uma amostra $\boldsymbol{X}$. Uma estatística $t(\boldsymbol{X})$ é a estatística suficiente para $\boldsymbol{\theta}$ se, e somente se,*

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = g(t(\boldsymbol{x}) \mid \boldsymbol{\theta})\, h(\boldsymbol{x}).$$

em que $g(t \mid \boldsymbol{\theta})$ é uma função que depende de $\boldsymbol{x}$ apenas por meio de t e $h(\boldsymbol{x})$ é uma função que não depende de $\boldsymbol{\theta}$.

Portanto, além da definição apresentada existem dois teoremas que apresentam condições suficientes para mostrar se uma estatística $t(\boldsymbol{X})$ é suficiente.

Em particular, para a família exponencial temos o seguinte teorema.

Teorema 2.3. *Seja $\boldsymbol{X} = (X_1, \ldots, X_n)$ um vetor aleatório composto por variáveis aleatórias i.i.d. de uma distribuição pertencente a família exponencial, com f.d.p. ou f.m.p. dada por*

$$f(x \mid \boldsymbol{\theta}) = h(x)\, \exp\left\{ \sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j(x) + d(\boldsymbol{\theta}) \right\} I_A(x)$$

em que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p), p \leq k$. Então,

$$t(\boldsymbol{X}) = \left(\sum_{j=1}^{k} t_1(X_j), \ldots, \sum_{j=1}^{k} t_k(X_j) \right)$$

é uma estatística suficiente para $\boldsymbol{\theta}$

[1] função densidade de probabilidade
[2] função massa de probabilidade

Notemos pelo critério da fatoração que a própria amostra $\boldsymbol{x}$ é uma estatística suficiente para $\boldsymbol{\theta}$. Pelo fato que existem diversas estatísticas suficientes surge o questionamento se uma estatística, de algum modo, é melhor que outra.

Definição 2.3. Uma estatística suficiente $T = t(\boldsymbol{X})$ é denominada estatística suficiente miníma se, para qualquer outra estatística suficiente T^*, T é uma função de T^*

Dizer que $t(\boldsymbol{x})$ é função de $t^*(\boldsymbol{x})$ significa que se $t^*(\boldsymbol{x}) = t^*(\boldsymbol{y})$, então $t(\boldsymbol{x}) = t(\boldsymbol{y})$. Enunciaremos agora um teorema que irá auxiliar na busca e identificação de estatísticas suficientes mínimas.

Teorema 2.4. *Seja $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ a f.d.p. ou f.m.p. de $\boldsymbol{X} = (X_1, \ldots, X_n)$. Suponha que existe uma função $t(\boldsymbol{x})$ de modo que, a razão $\dfrac{f(\boldsymbol{x} \mid \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{\theta})}$ não dependa de $\boldsymbol{\theta}$ se, e somente se, $t(\boldsymbol{x}) = t(\boldsymbol{y})$. Então, $t(\boldsymbol{X})$ é uma estatística suficiente mínima para $\boldsymbol{\theta}$.*

Ressaltamos que uma estatística suficiente mínima não é única, pois qualquer função um a um (bijetora) de uma estatística suficiente mínima também é uma estatística suficiente mínima.

Apresentaremos agora um tipo de estatística complementar as estatísticas suficientes, ou seja, estatísticas que não carregam informações sobre o parâmetro de interesse.

Definição 2.4. Uma estatística $S = s(\boldsymbol{X})$ cuja distribuição não depende do parâmetro é denominada estatística ancilar.

Discutimos alguns exemplos os quais é possível encontrar funções de estatísticas suficientes que são estatísticas ancilares. Isso indica que apesar de um estatística suficiente conter toda informação relevante sobre o parâmetro, ela também pode conter informações não relevantes. Nesse sentido, chegamos a última definição desta seção.

Definição 2.5. Seja $f_T(t \mid \boldsymbol{\theta})$ uma família de f.d.p. ou f.m.p. para um estatística $T = t(\boldsymbol{X})$. A família de distribuições é denominada completa se $\mathbb{E}_{\boldsymbol{\theta}} g(T) = 0$ para todo $\boldsymbol{\theta}$, o que implica $P_{\boldsymbol{\theta}}(g(T) = 0) = 1$ para todo $\boldsymbol{\theta}$. De forma equivalente, $t(\boldsymbol{X})$ é uma estatística completa.

Importante enunciarmos dois resultados de cálculo que são úteis para verificação se uma estatística é completa.

- Seja $h(x \mid \theta)$ uma função diferenciável em relação a θ e $\dfrac{\partial h(x \mid \theta)}{\partial \theta}$ contínua em função de x e θ, então

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{a(\theta)}^{b(\theta)} h(x \mid \theta)\, \mathrm{d}x = h\left(b(\theta) \mid \theta\right) \frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta} - h\left(a(\theta) \mid \theta\right) \frac{\mathrm{d}a(\theta)}{\mathrm{d}\theta} + \int_{a(\theta)}^{b(\theta)} \frac{\partial h(x \mid \theta)}{\partial \theta}\, \mathrm{d}x$$

  Em particular

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_0^{\theta} g(x)\, \mathrm{d}x = g(\theta)$$

- Se

$$\sum_{k=0}^{\infty} c_k \theta^k = 0 \quad \forall\, \theta,$$

então $c_k = 0 \,\forall\, k$.

Finalizamos esta seção com o Teorema de Basu.

Teorema 2.5. *Se $t(\boldsymbol{X})$ é uma estatística suficiente mínima e completa, então $t(\boldsymbol{X})$ é independente de toda estatística ancilar.*

Teorema 2.6. *Seja $\boldsymbol{X} = (X_1, \ldots, X_n)$ um vetor aleatório composto por variáveis aleatórias i.i.d. de uma distribuição pertencente a família exponencial com f.d.p. ou f.m.p. dada por*

$$f(x \mid \boldsymbol{\theta}) = h(x) \, \exp\left\{ \sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j(x) + d(\boldsymbol{\theta}) \right\} I_A(x)$$

em que $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p), p \leq k$. Então,

$$t(\boldsymbol{X}) = \left(\sum_{j=1}^{k} t_1(X_j), \ldots, \sum_{j=1}^{k} t_k(X_j) \right)$$

é uma estatística suficiente e completa para $\boldsymbol{\theta}$ desde que o espaço paramétrico $\boldsymbol{\Theta}$ seja um subconjunto aberto em $\mathbb{R}^k$

2.2 Princípio da Verossimilhança

Considere o modelo estatístico definido por $(\mathcal{X}, \mathcal{P}, \boldsymbol{\Theta})$. Suponha que uma amostra $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ do modelo $\mathcal{P}$ foi observada. Então a função de distribuição $F_{X_i}(x_i \mid \boldsymbol{\theta})$ e a função densidade ou probabilidade $f_{X_i}(x_i \mid \boldsymbol{\theta})$, que caracterizam o modelo $\mathcal{P}$, são quantidades fixas em relação a X_i dependendo somente de $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

Definição 2.6. A função de verossimilhança, ou simplesmente verossimilhança, do vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_n)$ é definida como

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \boldsymbol{\theta}) \tag{3}$$

em que $f(\boldsymbol{x} \mid \boldsymbol{\theta})$ é a função densidade ou probabilidade do vetor aleatório $\boldsymbol{X}$

A verossimilhança é definida como sendo igual a função do modelo, embora seja vista como função dos parâmetros, pois os dados observados são quantidades fixas. Para cada valor do parâmetro $\boldsymbol{\theta}$, a verossimilhança fornece uma medida de compatibilidade, plausibilidade ou similaridade com a amostra observada.

Importante ressaltar que a forma funcional da verossimilhança depende das suposições adotadas pelo modelo probabilístico. Por exemplo, assumindo que as variáveis

aleatórias são independentes, então a distribuição conjunta é dada pelo produto das funções densidade ou probabilidade. Por outro lado, no caso de variáveis aleatórias dependentes as densidades ou probabilidades condicionais são utilizadas.

A função de verossimilhança reflete informações do fenômeno expressas pelo modelo e conforme as observações.

Definição 2.7. Para o modelo estatístico especificado por $f(\cdot \mid \boldsymbol{\theta})$, tal que $\boldsymbol{\theta} \in \Theta$, se dois pontos $\boldsymbol{x}$ e $\boldsymbol{y} \in \mathcal{X}$ produzem verossimilhanças proporcionais, isto é, se $L(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto L(\boldsymbol{\theta} \mid \boldsymbol{y})$, então as verossimilhanças devem levar as mesmas conclusões inferências.

Essa definição é conhecida como princípio fraco da verossimilhança. A versão forte afirma que as mesmas conclusões podem ser obtidas para diferentes modelos e pontos amostrais.

Definição 2.8. Sejam $\boldsymbol{x}$ e $\boldsymbol{y}$ observações amostrais provenientes dos modelos $f(\cdot \mid \boldsymbol{\theta})$ e $g(\cdot \mid \boldsymbol{\theta})$, respectivamente, em que $\boldsymbol{\theta} \in \Theta$. Se $L_f(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto L_g(\boldsymbol{\theta} \mid \boldsymbol{y})$, então as verossimilhanças devem levar as mesmas conclusões inferências.

3 Estimação Pontual

Na classe de modelos estatísticos paramétrico, foco de estudo deste curso, a forma mais simples de realizar inferência sobre o fenômeno é estimando pontualmente os parâmetros, isto é, determinando um valor para o parâmetro(s) do modelo estatístico em questão.

Definição 3.1. Qualquer estatística $T = t(\boldsymbol{X})$ que assuma valores no espaço paramétrico $\boldsymbol{\Theta}$ de θ é um estimador para θ.

Note que um estimador é uma estatística e consequentemente uma variável aleatória. Denominamos de estimativa o valor numérico que do estimador para uma dada amostra observada. Usualmente tem-se o interesse em funções de θ, por exemplo $\boldsymbol{\tau}(\theta)$. Nessas situações temos uma definição análoga.

Definição 3.2. Qualquer estatística que assuma valores somente no conjunto de possíveis valores de $\boldsymbol{\tau}(\theta)$ é um estimador para $\boldsymbol{\tau}(\theta)$.

O problema da estimação pontual é determinar um estimador T para o parâmetro desconhecido que satisfaça algumas propriedades.

3.1 Propriedades dos Estimadores

Como vimos a definição de estimador é muito ampla no sentido que podemos propor diferentes estimadores para um mesmo parâmetro. Então, buscamos estimadores que desfrutem de certas propriedades. Tais propriedades foram estabelecidas no enfoque da inferência clássica, isto é, tratando o parâmetro como uma quantidade desconhecida e fixa. Embora, mesmo no enfoque Bayesiano, onde o parâmetro é uma variável aleatória, tais propriedades também são úteis.

Definição 3.3. O viés de um estimador T é a quantidade

$$\mathcal{B}(T) = \mathbb{E}(T) - \theta$$

em que $\mathbb{E}(\cdot)$ denota a esperança com relação a distribuição amostral do estimador. Um estimador cujo viés é igual a zero, $\mathcal{B}(T) = 0$, é denominado de estimador não viesado para θ. Analogamente, o estimador T é assintoticamente não viesado se $\mathbb{E}(T) \to 0$ quando $n \to \infty$.

Portanto, que o viés de um estimador mensura quão afastado o estimador esta, em média, do verdadeiro valor do parâmetro. É natural pensarmos em uma medida para avaliar a variabilidade do estimador. Podemos pensar na variância, porém se o estimador for tendencioso, então a variância não será a medida mais apropriada.

Definição 3.4. O erro quadrático médio de um estimador T é a quantidade

$$\mathrm{EQM}\,(T) = \mathbb{E}\left[(T-\theta)^2\right] = \mathrm{Var}\,(T) - \left[\mathcal{B}\,(T)\right]^2.$$

Quando o estimador é não viesado o EQM coincide com a variância do estimador. Como um estimador é uma variável aleatória, então propriedades probabilísticas podem ser de interesse.

Definição 3.5. O estimador $T_n = t(X_1, \ldots, X_n)$ é consistente para θ se

$$T_n \xrightarrow{P} \theta \quad \text{quando} \quad n \to \infty$$

para cada $\theta \in \Theta$ fixo.

Em outras palavras, dizemos que um estimador é consistente se ele converge em probabilidade para o verdadeiro valor do parâmetro. Conforme o tipo de convergência podemos definir tipos de estimadores consistentes. Por exemplo, se $T_n \xrightarrow{q.c.} \theta$ então T_n é um estimador consistente com probabilidade 1.

Teorema 3.1. *Se $T_n = t(X_1, \ldots, X_n)$ é um estimador para θ tal que $\mathbb{E}(T_n) \to \theta$ e $\mathrm{Var}(T_n) \to 0$ quando $n \to \infty$, então T_n é consistente para θ.*

A desigualdade de Tchebychev pode ser utilizada para provar este teorema.

Definição 3.6. Seja $\tau(\theta)$ qualquer função do parâmetro θ, podemos pensar inclusive que $\tau(x) = x$. Seja $C_\tau = \{\widehat{\tau}(\theta) : \mathbb{E}\left[\widehat{\tau}(\theta)\right] = \tau(\theta)\}$ a classe dos estimadores não viesados de $\tau(\theta)$ e suponha que $C_\tau \neq \emptyset$. Um estimador $\widehat{\tau}^*(\theta)$ pertencente a classe C_τ que satisfaz

$$\mathrm{Var}\left[\widehat{\tau}^*(\theta)\right] \leq \mathrm{Var}\left[\widehat{\tau}(\theta)\right], \quad \forall\, \theta \in \Theta$$

é denominado de Estimador Não Viesado de Variância Uniformemente Mínima (ENVVUM).

Seja $(\mathcal{X}, \mathcal{P}, \mathbf{\Theta})$um modelo paramétrico estatístico especificado pela função densidade $f_X(\cdot \mid \theta)$, em que $\theta \in \Theta \subset \mathbb{R}$. As seguintes suposições definem as condições de regularidades.

(i) O espaço paramétrico Θ tem dimensão finita, é fechado e o verdadeiro valor do parâmetro pertence a Θ;

(ii) A função de distribuição definida para dois diferentes valores de θ é distinta, isto é, o modelo é identificável;

(iii) A diferenciação em relação a θ e integração em relação a x pode ser intercambiável duas vezes, isto é,

$$\int_{\mathcal{X}} \frac{\partial}{\partial\theta} f_X(x \mid \theta)\mathrm{d}x = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f_X(x \mid \theta)\mathrm{d}x$$

e

$$\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta^2} f_X(x \mid \theta)\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\int_{\mathcal{X}} f_X(x \mid \theta)\mathrm{d}x$$

Para $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ temos

$$\int_{\mathcal{X}} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} f_X(x \mid \boldsymbol{\theta})\mathrm{d}x = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\int_{\mathcal{X}} f_X(x \mid \theta)\mathrm{d}x$$

Tais condições também devem ser válidas para modelos discretos, substituindo a integral por somatório.

Sob as condições de regularidades podemos mostrar que

$$\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f_X(X \mid \boldsymbol{\theta})\right] = 0$$

e

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_X(X \mid \boldsymbol{\theta})\right)^2\right] = \mathbb{E}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f_X(X \mid \boldsymbol{\theta})\right].$$

Na busca pelo melhor estimador não viesado enunciamos o teorema de Cramér-Rao, seguramente um dos principais teoremas da inferência estatística.

Teorema 3.2. *Seja $\mathbf{X} = (X_1, \ldots, X_n)^\top$ uma amostra aleatória cuja a função densidade ou probabilidade $f_{X_i}(x_i \mid \theta)$ satisfaz as condições de regularidades. Seja $T = t(\mathbf{X})$ um estimador não viesado para $\tau(\theta)$, então*

$$\mathrm{Var}\,(T) \geq \frac{\left[\dfrac{\mathrm{d}\,\tau(\theta)}{\mathrm{d}\,\theta}\right]^2}{\mathbb{E}\left[\left(\dfrac{\partial}{\partial\boldsymbol{\theta}}\log f_{\mathbf{X}}\,(\mathbf{X} \mid \theta)\right)^2\right]}.$$

Este teorema é também conhecido como desigualdade de Cramér-Rao. Podemos afirma que se a variância de um estimador não viesado atinge o limite da desigualdade de Cramér-Rao, então tal estimador é um ENVVUM. Importante mencionar que a prova deste teorema é uma aplicação inteligente da desigualdade de Cauchy-Schwarz.

Corolário 3.1. *Sejam $X_1, \ldots, X_n$ uma amostra aleatória de uma função densidade ou probabilidade $f_{X_i}(x_i \mid \theta)$ que satisfaz as condições de regularidades. Seja $L(\theta \mid \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \mid \theta)$ a verossimilhança e $T = t(\mathbf{X})$ um estimador não viesado para $\tau(\theta)$, então T*

atinge o limite inferior de Cramér-Rao se, e somente se,

$$a(\theta)\left[T-\tau(\theta)\right]=\frac{\partial}{\partial\theta}\log L(\theta\mid\boldsymbol{x})$$

para alguma função $a(\theta)$*.*

Entretanto, em algumas situações tal limite pode não ser atingível ou então a família de distribuições não satisfaz as condições de regularidades estipuladas. Nestes casos surge o Teorema de Rao-Blackwell.

Teorema 3.3. *Seja* $T=t(\boldsymbol{X})$ *um estimador não viesado para* $\tau(\theta)$*, e seja* $S=s(\boldsymbol{X})$ *uma estatística suficiente para* θ*. Defina* $\phi(T)=\mathbb{E}\left(T\mid S\right)$*, então*

$$\mathbb{E}_\theta\left[\phi(T)\right]=\tau(\theta)\quad e\quad \mathrm{Var}_\theta\left[\phi(T)\right]\leq\mathrm{Var}_\theta\left(T\right)$$

para todo $\theta\in\Theta$*. Assim,* $\phi(T)$ *é um* ENVVUM *para* $\tau(\theta)$*.*

Conclui-se também que ao condicionar qualquer estimador não viesado em uma estatística suficiente resultará em um estimador melhorado.

A seguir apresento dois outros teoremas que garantem a unicidade e existência do melhor estimador não viesado, isto é, o ENVVUM

Teorema 3.4. *Se* $T=t(\boldsymbol{X})$ *é um Estimador Não Viesado de Variância Uniformemente Mínima para* $\tau(\theta)$*, então* T *é único.*

Teorema 3.5. *Seja* $T=t(\boldsymbol{X})$ *um estimador não viesado para* $\tau(\theta)$*, então* T *será o* ENVVUM *de* $\tau(\theta)$ *se, e somente se,* T *não estiver correlacionado com todos os outros estimadores não viesados de 0.*

Note que existem infinitos estimadores não viesados para 0, assim é inviável verificar se um determinar estimador não esta correlacionado com todos os estimadores não viesados de 0, afim de caracteriza-lo como o ENVVUM.

Contudo, quando uma família de distribuições é completa, não existe nenhuma função da amostra cuja esperança seja igual a 0, com exceção da própria função nula. Porém, qualquer estimador não viesado T satisfaz $\mathrm{Cov}(T,0)=0$, assim a procura pelo ENVVUM esta encerrada.

Finalizamos esta seção enunciando o Teorema de Lehmann-Scheffé o qual estabelece condições para encontrar o ENVVUM quando existe um estatística suficiente e completa.

Teorema 3.6. *Seja* $T=t(\boldsymbol{X})$ *um estimador não viesado para* $\tau(\theta)$*, e seja* $S=s(\boldsymbol{X})$ *uma estatística suficiente e completa para* θ*. Então* $\phi(T)=\mathbb{E}(T\mid S)$ *é o melhor estimador não viesado de* $\tau(\theta)$*, isto é,* $\phi(T)$ *é o* ENVVUM *de* $\tau(\theta)$*.*

3.2 Métodos de Estimação

Exitem na literatura diversos procedimentos para encontrar estimadores para os parâmetros de uma dada distribuição de probabilidade. Neste curso, o método do momentos e o método da máxima verossimilhança serão discutidos.

Procedimentos baseados na minimização da distância entre as funções de distribuição empírica e teórica são alternativas interessantes. Além disso, métodos computacionais como Bootstrap e simulação Monte Carlo também podem ser úteis para estimação de parâmetros.

3.2.1 Métodos dos Momentos

Seja $\{X_n\}_{n\geq 1}$ uma sequência de variáveis aleatórias independente e identicamente distribuídas tal que $\mathbb{E}\left(X_n^k\right) < \infty$, isto é, com k-ésimo momento finito. Então, a Lei Forte dos Grandes Números afirma que

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{q.c.} \mathbb{E}\left(X_n^k\right) \quad \text{quando} \quad n \to \infty$$

Utilizando este resultado probabilístico podemos definir estimadores para o vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ da função densidade ou probabilidade $f_X(x \mid \boldsymbol{\theta})$, em que $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$.

Define-se as funções reais $h_1(\boldsymbol{\theta}), \ldots, h_p(\boldsymbol{\theta})$ como

$$h_k(\boldsymbol{\theta}) = \mathbb{E}\left(X^k\right)$$

e sejam

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

para $k = 1, \ldots, p$. As funções $h_1(\boldsymbol{\theta}), \ldots, h_p(\boldsymbol{\theta})$ são os momentos teóricos da variável aleatória X, enquanto que $\mu_1, \ldots, \mu_p$ são os correspondentes momentos amostrais.

Definição 3.7. O estimador de momentos $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ é dado pela solução do sistema de equações

$$h_k(\widehat{\boldsymbol{\theta}}) = \mu_k$$

para $k = 1, \ldots, p$.

Importante observar que este método não é aplicável a situações onde os momentos teóricos não existem.

Teorema 3.7. *Seja* $h(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_p(\boldsymbol{\theta}))^\top$ *e seja* $\mathbf{H}(\boldsymbol{\theta})$ *uma matriz de posto* p *cujos*

elementos são $H_{ij} = \dfrac{\partial h_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, i, j = 1, \ldots, p$, *sendo estas funções contínuas em* $\boldsymbol{\theta}$. *Então,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{D} N_p\left(\mathbf{0}, \mathbf{H}^{-1}\,\boldsymbol{\Sigma}\,\left[\mathbf{H}^{-1}\right]^{\top}\right),$$

em que $\boldsymbol{\Sigma}$ *é uma matriz* $p \times p$ *cujos elementos são* $\Sigma_{ij} = h_{i+j} - h_i h_j$

3.2.2 Método da Máxima Verossimilhança

Sem perdas generalidades vimos que a verossimilhança é definida como sendo função de distribuição conjunta da amostra observada visto como função do vetor de parâmetros parâmetro $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Podemos definir mais especificamente a verossimilhança conforme a natureza do fenômeno.

Para o caso de variáveis discretas não há ambiguidade e o valor da verossimilhança é a probabilidade do valor observado, isto é,

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{X} = \boldsymbol{x}) \tag{4}$$

em que o subscrito enfatiza que a distribuição depende do parâmetro $\boldsymbol{\theta}$.

Para os modelos contínuos a probabilidade de um conjunto de valores $\boldsymbol{x}$ é nula. No entanto, na prática variáveis contínuas são medidas com um certo grau de precisão no intervalo $x_{1i} < x_i \leq x_{2i}$. Assim, a verossimilhança é dada por

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \mathrm{P}_{\boldsymbol{\theta}}(x_{11} < X_1 \leq x_{21}, \ldots, x_{1n} < X_n \leq x_{2n}). \tag{5}$$

Note que essa definição é bem geral e considera diversas situações, por exemplo quando os dados são censurados. Vamos supor que as variáveis aleatórias $X_1, \ldots, X_n$ são independentes, então a verossimilhança é

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{i=1}^{n} \mathrm{P}_{\boldsymbol{\theta}}(x_{1i} < X_i \leq x_{2i}) = \prod_{i=1}^{n} \int_{x_{1i}}^{x_{2i}} f_{X_i}(t_i \mid \boldsymbol{\theta}) \mathrm{d}t_i. \tag{6}$$

Se o grau de precisão das observações é alto em relação a variabilidade dos dados, então a verossimilhança pode ser aproximada por

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) \approx \prod_{i=1}^{n} f_{X_i}(t_i \mid \boldsymbol{\theta}). \tag{7}$$

A expressão (7) é a mais difundida na literatura para se referir a verossimilhança de modelos contínuos. No entanto, temos que ter em mente que tal expressão é válida somente quando as observações forem medidas com alto grau de precisão.

Podemos dizer que a função verossimilhança informa a ordem natural de preferên-

cia entre as diversas possibilidades de $\boldsymbol{\theta}$. Assim, se o objetivo é estimar o verdadeiro valor de $\boldsymbol{\theta}$, sob o paradigma da verossimilhança, o valor mais plausível para $\boldsymbol{\theta}$ é aquele de maior verossimilhança.

Definição 3.8. Dada a função de verossimilhança $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ para determinada amostra observada $\boldsymbol{x}$. A estimativa de máxima verossimilhança de $\boldsymbol{\theta}$ é dada por

$$\widehat{\boldsymbol{\theta}} = \arg\left[\sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta} \mid \boldsymbol{x})\right]. \tag{8}$$

Quando a verossimilhança é duplamente diferenciável em relação a $\boldsymbol{\theta}$ é comum trabalharmos com o logaritmo natural, ou seja, a função log-verossimilhança $\ell(\boldsymbol{\theta} \mid \boldsymbol{x})$. Uma vez que o $\log$ é uma função estritamente crescente, ambas as funções $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ e $\ell(\boldsymbol{\theta} \mid \boldsymbol{x})$ levam ao mesmo ponto de máximo.

Em particular, quando a log-verossimilhança é uma função côncava a estimativa de máxima verossimilhança pode ser obtida resolvendo o sistema de equações

$$\frac{\partial}{\partial\theta_j}\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = 0, \quad j = 1, \ldots, p. \tag{9}$$

Lembremos que uma função côncava possui uma propriedade interessante que seus pontos críticos são máximo globais. Em outras situações, podemos verificar se $\ell(\boldsymbol{\theta} \mid \boldsymbol{x})$ é localmente côncava no ponto estacionário, ou seja, se a matriz de segundas derivadas

$$H(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) \tag{10}$$

é negativa definida em $\widehat{\boldsymbol{\theta}}$.

Em geral, as estimativas de máxima verossimilhança não possuem forma analítica sendo assim, necessário procedimentos computacionais de maximização. Isso nos leva a impossibilidade de obter a distribuição exata dos estimadores de máxima verossimilhança.

- Distribuição assintótica dos EMVs
- Propriedade da invariância
- Propriedades assintóticas
- Método delta

4 Estimação Intervalar

A segunda forma de realizar inferência sobre o fenômeno, considerando a classe dos modelos paramétrico, é determinando um intervalo ou região de confiança para o parâmetro ou vetor de parâmetros, respectivamente. As definições e conceitos apresentados são para o caso em que θ é um escalar.

Definição 4.1. Sejam $L(\boldsymbol{X})$ e $U(\boldsymbol{X})$ duas estatísticas tais que $L(\boldsymbol{x}) \leq U(\boldsymbol{x})\ \forall\, \boldsymbol{x} \in \mathcal{X}$. Então $[L(\boldsymbol{X}); U(\boldsymbol{X})]$ é um estimador intervalar para θ.

Importante notar $L(\boldsymbol{X})$ e $U(\boldsymbol{X})$ são variáveis aleatórias. Assim, surge a seguinte definição.

Definição 4.2. Seja $[L(\boldsymbol{X}); U(\boldsymbol{X})]$ um estimador intervalar para θ. Então, a probabilidade de cobertura é definida como sendo

$$P_\theta\left(\theta \in [L(\boldsymbol{X}); U(\boldsymbol{X})]\right) = P(\theta \in [L(\boldsymbol{X}); U(\boldsymbol{X})] \mid \theta) = P(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X}) \mid \theta),$$

isto é, a probabilidade do intervalo aleatório $[L(\boldsymbol{X}); U(\boldsymbol{X})]$ abranger o verdadeiro valor de θ.

Uma vez que não conhecemos o verdadeiro valor do parâmetro então faremos declarações sobre o intervalo aleatória com um certo nível de confiança $\gamma = 1 - \alpha$. Daí surge o termo intervalo de confiança.

Definição 4.3. Seja $[L(\boldsymbol{X}); U(\boldsymbol{X})]$ um estimador intervalar para θ, então seu coeficiente de confiança é dado por

$$\gamma = \inf_\theta P_\theta\left(\theta \in [L(\boldsymbol{X}); U(\boldsymbol{X})]\right).$$

O método mais usual para encontrar um intervalo de confiança exato para um parâmetro é baseado em estatísticas que dependem do parâmetro de interesse, porém sua distribuição não depende.

Definição 4.4. A variável aleatória $Q(\boldsymbol{X}, \theta)$ é uma quantidade pivotal se sua distribuição independe do parâmetro θ.

Dessa forma, temos um método genérico para construção de intervalo de confiança baseado em uma quantidade pivotal.

Definição 4.5. Seja $Q = Q(\boldsymbol{X}, \theta)$ uma quantidade pivotal para θ. Dado um coeficiente de confiança $\gamma \in (0, 1)$ tal que $P(q_1 \leq Q \leq q_2) = \gamma$. Se existirem estatísticas $L(\boldsymbol{X})$ e $U(\boldsymbol{X})$ que satisfazem

$$q_1 \leq Q(\boldsymbol{X}, \theta) \leq q_2 \iff L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})$$

então $[L(\boldsymbol{X}); U(\boldsymbol{X})]$ é um estimador intervalar para θ com coeficiente de confiança γ.

5 Teste de Hipótese

A terceira e última forma, discutida neste curso, para realizar inferência sobre um fenômeno é conjecturando sobre a classe de distribuição de probabilidade que governa o mecanismo gerados dos dados por meio de hipóteses estatística que especificam completamente ou parcialmente a distribuição.

Definição 5.1. Uma hipótese estatística é qualquer afirmação acerca da distribuição de probabilidade de uma ou mais variáveis aleatórias.

A definição acima é bastante genérica e engloba casos da inferência não paramétrica, isto é, quando assumimos que o fenômeno de interesse pode ser especificado por modelos não paramétricos. Entretanto, neste curso assumimos que a amostra aleatória $\boldsymbol{X} = (X_1, \ldots, X_n)$ de uma distribuição F_θ com $\theta \in \Theta$, sendo que a forma funcional de F_θ é conhecido exceto pelo valor do parâmetro θ, que em geral é desconhecido.

Definição 5.2. Uma hipótese estatística paramétrica é qualquer afirmação sobre o parâmetro desconhecido θ.

A hipótese nula é especificado por $\mathcal{H}_0 : \theta \in \Theta_0 \subset \Theta$, ao passo que, a hipótese alternativa é $\mathcal{H}_1 : \theta \in \Theta_1 \subset \Theta$, sendo as hipóteses disjuntas. Em outras palavras, qualquer teste de hipótese divide o espaço paramétrico em duas regiões complementares.

Definição 5.3. Se Θ_0 (Θ_1) contém apenas um ponto, então $\mathcal{H}_0$ ($\mathcal{H}_1$) é uma hipótese simples, caso contrário é uma hipótese composta.

Observe que se a hipótese é simples, então a distribuição de probabilidade é completamente especificada. Enquanto que, se a hipótese é composta a família de distribuição é especificada.

Definição 5.4. Seja $\boldsymbol{X} \sim F_\theta$ com $\theta \in \Theta$. Um subconjunto $C \in \mathbb{R}^n$ tal que se $\boldsymbol{x} \in C$, então $\mathcal{H}_0$ é rejeitada com probabilidade 1 é denominado região de critica (rejeição).

Definição 5.5. Qualquer função φ definida em $\mathbb{R}^n \to [0, 1]$ é uma função teste.

Alguns exemplos de função teste são: $\varphi(\boldsymbol{x}) = 1$ para todo $\boldsymbol{x} \in \mathbb{R}^n$ e $\varphi(\boldsymbol{x}) = 0$ para todo $\boldsymbol{x} \notin \mathbb{R}^n$, ou $\varphi(\boldsymbol{x}) = \alpha$ para todo $\boldsymbol{x} \in \mathbb{R}^n$, em que $0 \leq \alpha \leq 1$.

A teoria clássica do teste de hipótese foi desenvolvida por Jerzy Neyman e Egon Pearson nos anos de 1920. Nesta teoria um teste de hipótese é completamente especificado pela função poder $\beta(\theta)$ que fornece a probabilidade do teste rejeitar $\mathcal{H}_0$ para cada valor de $\theta \in \Theta$.

Definição 5.6. Seja φ a função teste e $C \in \mathbb{R}^n$ a região de critica do teste, então a função poder é dada por

$$\beta_\varphi(\theta) = \mathbb{E}_\theta \varphi(\boldsymbol{X}) = P(\mathbf{X} \in C \mid \theta), \quad \theta \in \Theta.$$

Importante mencionar que alguns autores definem a função poder somente para $\theta \in \Theta_1$. Se a distribuição dos dados sob a hipótese nula é conhecida pode-se determinar a região crítica, sob $\mathcal{H}_0$, tal que a probabilidade de rejeita-la não exceda uma valor pré-estabelecido.

Definição 5.7. O nível de significância de um teste é dado por

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha. \tag{11}$$

Definição 5.8. O tamanho de um teste é dado por

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

Em um teste de hipótese dois tipos de erros podem ser cometidos. Ao rejeitar a hipótese nula quando ela é verdadeira comete-se o Erro do Tipo I. Note que o nível de significância do teste expressa que a probabilidade do Erro do Tipo I nunca excede α. Por outro lado, a não rejeição da hipótese nula quando ela é falsa é denominado de Erro do Tipo II.

Várias regiões críticas satisfazem a Equação (11). Qual delas é preferível? Este é um problema crucial na teoria dos testes de hipóteses. Na teoria clássica de Neyman-Pearson o nível de significância α é fixado e então o poder do teste é maximizado. Isso nos leva na seguinte definição.

Definição 5.9. Seja Φ_α a classe de testes de nível α para testar $\mathcal{H}_0 : \theta \in \Theta_0$ versus $\mathcal{H}_1 : \theta \in \Theta_1$. Um teste φ_0 é denominado de Uniformemente Mais Poderoso (UMP) se

$$\beta_{\varphi_0}(\theta) \geq \beta_\varphi(\theta) \quad \forall\, \varphi \in \Phi_\alpha, \forall\, \theta \in \Theta_1$$

em que $\beta_\varphi(\theta)$ é a função poder do teste φ.

Se a hipótese alternativa é simples, ou seja, se Θ_1 é um único valor, então o teste que satisfaz tais condições é denominado de Teste Mais Poderoso

Teorema 5.1. *Considere testar as hipóteses $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_1 : \theta = \theta_1$ com região crítica definida por*

$$\begin{aligned} \boldsymbol{x} \in \mathbb{R}^n \quad & \textit{se } L(\theta_1 \mid \boldsymbol{x}) > kL(\theta_0 \mid \boldsymbol{x}) \\ \boldsymbol{x} \notin \mathbb{R}^n \quad & \textit{se } L(\theta_1 \mid \boldsymbol{x}) < kL(\theta_0 \mid \boldsymbol{x}) \end{aligned}$$

ou equivalentemente, suponha que a função teste é dada por

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \textit{se } L(\theta_1 \mid \boldsymbol{x}) > kL(\theta_0 \mid \boldsymbol{x}) \\ 0 & \textit{se } L(\theta_1 \mid \boldsymbol{x}) < kL(\theta_0 \mid \boldsymbol{x}) \end{cases} \tag{12}$$

em que a constante $k > 0$ é determinada de tal forma que

$$\mathbb{E}_{\theta_0}\varphi(\boldsymbol{X}) = \alpha. \tag{13}$$

Qualquer teste satisfazendo (12) *e* (13) *é um teste MP de nível* α.

O Teorema enunciado acima é conhecido na literatura por Lema de Neyman-Pearson. Para encontrar um teste UMP utilizando o Lema de Neyman Pearson quando a hipótese alternativa é composta basta (i) reescrever a hipótese alternativa em termos de uma hipótese simples, (ii) invocar o Lema de Neyman para provar que existe um teste MP e por fim, (iii) verificar se a região critica obtida vale para qualquer valor do parâmetro especificado pela hipótese alternativa.

Definição 5.10. Uma família de distribuições $\{f(x \mid \theta) : \theta \in \Theta\}$ tem Razão de Verossimilhanças Monótona (RVM) em $t(\boldsymbol{x})$ se para $\theta_1 > \theta_2$

$$\frac{L(\theta_1)}{L(\theta_2)}$$

é uma função não decrescente (não crescente) em $t(\boldsymbol{x})$.

Definição 5.11. Suponha que $\{f(\boldsymbol{x} \mid \theta) : \theta \in \Theta\}$ denota a família exponencial com densidade dada por

$$f(\boldsymbol{x} \mid \theta) = h(x)\, \exp\{c(\theta)t(\boldsymbol{x}) + d(\theta)\}\, I_A(x)$$

e $c(\theta)$ é uma função não decrescente (não crescente) então $f(x \mid \theta)$ tem RVM não decrescente (não crescente) em $t(\boldsymbol{x})$.

Com base na definição de razão de verossimilhanças monótona Karlin e Rubin em 1956 enunciaram um teorema que permite genericamente encontrar um teste UMP.

Teorema 5.2. *Considere testar as hipótese* $\mathcal{H}_0 : \theta = \theta_0$ *versus* $\mathcal{H}_1 : \theta > \theta_0$. *Suponha que a família de distribuição* $\{f(\boldsymbol{x} \mid \theta) : \theta \in \Theta\}$ *tem RMV não decrescente em* $t(\boldsymbol{x})$. *Então a função teste*

$$\varphi(\boldsymbol{x}) = \begin{cases} 1 & \text{se } t(\boldsymbol{x}) > c \\ 0 & \text{se } t(\boldsymbol{x}) < c \end{cases} \tag{14}$$

corresponde ao teste UMP de nível α, *em que o valor de* c *é determinado por* $\mathbb{E}_{\theta_0}\varphi(\boldsymbol{X}) = \alpha$.

Referências

AZZALINI, A. **Statistical Inference: Based on the likelihood**. [S.l.]: Chapman & Hall, 1996.

CASELLA, G.; BERGER, R. L. **Statistical Inference**. 2nd. ed. [S.l.]: Duxbury Press, 2002.

LINDSEY, J. K. **Parametric Statistical Inference**. [S.l.]: Oxford, 1996.

MUKHOPADHYAY, N. **Probability and Statistical Inference**. [S.l.]: Marcel Dekker, 2000.

ROHATGI, V. K. **An Introduction to Probability and Mathematical Statistics**. [S.l.]: John Willey & Sons, Inc, 1976.

ROHATGI, V. K. **Statistical Inference**. [S.l.]: John Willey & Sons, Inc, 1984.