FGV EMAp

João Pedro Jerônimo

Inferência Estatística

Revisão para A1

Rio de Janeiro

2025

Conteúdo

Definições

Primeiro de tudo, precisamos entender o que estuda a Inferência Estatística. Isso nada mais é que um nome bonitinho para “chutar”. Na probabilidade, tinhamos uma variável aleatória com uma distribuição e um parâmetro bem-definidos, porém, na vida real, não é muito bem o que ocorre. Imagine que queremos saber o tempo de vida que as lâmpadas da minha fábrica vivem, a única coisa que vou ter como me basear são as lâmpadas que já tenho, é dessas informações que eu tenho que fazer uma inferência, seja ela qual for, como por exemplo, qual sua distribuição e qual é o parâmetro a ela associado

Definição 1.1 (Modelo Estatístico): Um modelo estatístico consiste em:

1. Identificar variáveis de interesse (Sejam elas observáveis ou hipoteticamente observáveis, como um parâmetro de distribuição)
2. Especificar a distribuição conjunta (Ou uma família de distribuições) para variáveis observáveis
3. Identificar os parâmetros de interesse em (2)
4. (Se desejado) especificar uma distribuição para os parâmetros descritos

Definição 1.2 (Inferência Estatística): É um procedimento que produz afirmações probabilísticas sobre algumas ou todas as partes de um modelo estatístico

Definição 1.3 (Parâmetro e Espaço Paramétrico): Em um problema de inferência estatística, uma característica (ou combinações de características) que determina(m) a distribuição conjunta da(s) variável(eis) de interesse é chamada de parâmetro. O conjunto Θ de todos os possíveis valores de um parâmetro θ (Ou vetor paramétrico $(\theta_1, ..., \theta_k)$) é chamado de **espaço paramétrico**

Dentro da estatística, podemos dividir os problemas que encontramos em algumas categorias:

- **Predição**: Podemos tentar prever o resultado de uma variável aleatória com base nas observações anteriores. Quando a variável é um parâmetro, chamamos de Estimação.
- **Problemas de Decisão Estatística**: Depois que dados experimentais foram analisados, podemos querer tomar decisões com base nos resultados do experimento. As consequências da decisão dependem dos resultados.
- **Design Experimental**: Em alguns problemas de inferência estatística, temos controle sobre o tipo de dados ou quantidade de dados experimentais coletados.

Definição 1.4 (Estatística): Suponha que as variáveis aleatórias observáveis de interesse são $X_1, ..., X_n$. Seja $r : \mathbb{R}^n \to \mathbb{R}$. Então a variável aleatória $T = r(X_1, ..., X_n)$ é chamada de estatística.

Há também uma discussão sobre se os parâmetros que estamos procurando serem variáveis aleatórias ou valores fixos. Por enquanto, assumiremos que os parâmetros são variáveis aleatórias. Essa discussão está mais bem detalhada no livro do **DeGroot**

Estatística Bayesiana

2.1 Introdução

Nesse primeiro momento, iremos fazer experimentos assumindo que, ao fazer experimentos e obter resultados X_j eles estão saindo de uma distribuição com parâmetro (ou vetor paramétrico) θ, e esse θ é uma variável aleatória da qual desconhecemos.

2.2 Distribuições Priori e Posteriori

Quando fazemos um experimento em que θ é uma V.A., é interessante chutar uma distribuição para ele antes de observar qualquer dado.

Definição 2.2.1 (Distribuição a priori): Dado um modelo estatístico com parâmetro θ, se θ for uma variável aleatória, a distribuição de θ antes de qualquer dado é chamada de distribuição a priori (Podemos denotar $\xi(\theta)$ ou $f_\theta(\theta)$).

Quando estamos trabalhando com observações $X_1, ..., X_k$, denotamos a distribuição a priori dos dados como $X_1, ..., X_k|\theta \sim \mathrm{Dist}(\theta)$ onde **Dist** representa qualquer distribuição, Ué, como que condicionamos X_j em θ? Qual o sentido disso? Imagine que cada experimento é o output de uma máquina industrial, porém, para que essa máquina funcione, alguém precisa passar algumas informações para ela, porém, o seu chefe não mandou você colocar as informações, então você não sabe quais são elas, mas você está vendo os resultados da máquina, e sabe que aqueles resultados só estão acontecendo porque aquela configuração foi colocada, então por mais que não sabemos o θ, ele os valores de $X_1, ..., X_k$ só sairam como estamos vendo porque o parâmetro da distribuição é θ (Que ainda queremos descobrir)

Assim como especificamos uma distribuição para θ antes de qualquer dado ser observado, podemos atualizar a distribuição conforme observamos dados.

Definição 2.2.2 (Função de verossimilhança): A função de verossimilhança $\mathbb{L}(\theta)$ é definida por

$$\mathbb{L}(\theta) = f_{X|\theta}(x_1, ..., x_k \mid \theta) \tag{1}$$

De forma que $f_{X|\theta}(\underline{x}|\theta)$ é a f.d.p de $X_1, ..., X_k$

Definição 2.2.3 (Distribuição a posteriori): Dado um modelo estatístico com variáveis aleatórias observáveis $X_1, ..., X_n$, a distribuição de $X_1, ..., X_n|\theta$ é chamada de distribuição a posteriori

E agora, com o teorema de bayes, podemos relacionar essas nossas definições

Teorema 2.2.1 (Bayes): Suponha que $X_1, ..., x_k$ são amostras de uma população com distribuição conhecida de parâmetro θ tal que sua f.d.p é $f_{X|\theta}(x_1, ..., x_k|\theta)$. Suponha também que θ é desconhecido e a distribuição a priori de θ é tal que sua f.d.p é $f_\theta(\theta)$, então a posteriori de θ é tal que:

$$f_\theta(\theta|x_1, ..., x_k) = \frac{f_X(x_1, ..., x_k|\theta)f_\theta(\theta)}{f_X(x_1, ..., x_k)} \tag{2}$$

ou

$$f_\theta(\theta|x_1, ..., x_k) = \frac{\mathbb{L}(\theta)\xi(\theta)}{\int \mathbb{L}(\theta)\xi(\theta)d\theta} \tag{3}$$

Perceba porém, que o termo do denominador não depende de θ, ou seja, podemos reescrever isso tudo como:

$$f_\theta(\theta|x_1, ..., x_k) \propto \mathbb{L}(\theta)\xi(\theta) \tag{4}$$

Demonstração: Usar teorema de Bayes □

Por conta do teorema acima, todos os termos constantes que encontramos em nossa distribuição nós podemos pegar e jogar fora e, ao final, encontramos um termo constante geral, já que para descobrir essa constante C basta fazer:

$$\frac{1}{C} = \int_{|\Theta|} \mathbb{L}(\theta)\xi(\theta)d\theta \tag{5}$$

2.3 Observações sequenciais e predições

Porém, perceba que, até agora, eu vi o caso em que eu tenho todas as amostras de uma vez, porém se, por exemplo, eu quero descobrir se uma vacina é eficaz, isso é inviável, faz muito mas sentido eu ir atualizando minha distribuição conforme recebo mais informações, mas será que isso vai dar a mesma coisa?

Vamos fazer com duas observações condicionalmente independentes, para generalizar se faz analogamente. Como vimos, a posteriori de θ após eu observar o dado x_1 se dá como:

$$f_\theta(\theta|x_1) \propto f_{X|\theta}(x_1|\theta)f_\theta(\theta) \tag{6}$$

Agora queremos obter $f_\theta(\theta|x_1, x_2)$. Pelo teorema de bayes para várias condicionais, temos que:

$$f_\theta(\theta|x_1, x_2) \propto f_\theta(\theta|x_1)f_X(x_2|x_1, \theta) \tag{7}$$

Porém, estamos assumindo que eles são condicionalmente independentes, ou seja:

$$\begin{aligned} f_X(x_2|\theta, x_1) &= f_X(x_2|\theta) \\ \Rightarrow f_\theta(\theta|x_1, x_2) & \\ \propto f_\theta(\theta|x_1)f_X(x_2|\theta) & \\ \propto f_\theta(\theta)f_{X|\theta}(x_1|\theta)f_X(x_2|\theta) & \\ \propto f_\theta(\theta)f_{X|\theta}(x_1, x_2|\theta) & \end{aligned} \tag{8}$$

Ou seja, independentemente se eu estou recebendo dado após o outro ou se eu tenho todos de uma vez para trabalhar, o resultado final deve ser o mesmo.

Porém, se voltarmos na equação (5), podemos perceber algo interessante. Lembra da **Lei da Probabilidade Total**?

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(B_i)\mathbb{P}(A|B_i) \tag{9}$$

Com B_i sendo disjuntos. Porém, temos também a versão contínua do teorema:

$$f_X(x) = \int_{\Omega} f_{X|Y}(x|y) f_Y(y) dy \tag{10}$$

Porém, se fizermos algumas substituições, nós obtemos:

$$f(x_k|x_1, ..., x_{k-1}) = \int_{|\Theta|} f(x_k|\theta)\xi(\theta|x_1, ..., x_{k-1}) d\theta \tag{11}$$

Ou seja, podemos utilizar essa equação caso tenhamos n observações e estamos interessados em prever o resultado da próxima observação.

2.4 Distribuições à Priori Conjugadas

São famílias de distribuições de tal forma que, quando selecionamos elas como distribuições para um modelo estatístico, a posteriori também será daquela distribuição

> **Definição 2.4.1** (Famílias/Hiperparâmetros Conjugados): Seja $X_1, X_2, ...|\theta$ serem **i.i.d** com mesma f.d.p ou f.m.p $f(x|\theta)$. Seja Ψ uma família de distribuições no espaço paramétrico Θ. Suponha que, não importa qual seja a distribuição à priori ξ que eu escolher de Ψ, não importa quantas observações $\underline{X} = (X_1, ..., X_n)$ nós registramos e não importa seus valores observados $\underline{x} = (x_1, ..., x_n)$, a distribuição à posteriori $\xi(\theta|\underline{x})$ está em Ψ. Então Ψ é chamada de uma **família de distribuições à priori conjugadas** para amostras de com distribuições $f(x|\theta)$. Finalmente, se as distribuições em Ψ possuem parâmetros associados, estes são chamados de **hiperparâmetros à priori** e os associados à distribuição posteriori são **hiperparâmetros à posteriori**

Vamos ver as principais famílias de distribuições conjugadas

> **Teorema 2.4.1**: Suponha que $X_1, ..., X_n|\theta$ são uma amostra aleatória de variáveis de Bernoulli com parâmetro θ (Desconhecido). Suponha também que a distribuição a priori de θ é uma **beta** com parâmetros $\alpha > 0$ e $\beta > 0$. Então a distribuição a posteriori de $\theta|x_1, ..., x_n$ é a distribuição beta com parâmetros $\alpha + \sum_{i=1}^{n} x_i$ e $\beta + n - \sum_{i=1}^{n} x_i$

Demonstração:

$$f(\theta|x_1, ..., x_n) \propto \xi(\theta) f(x_1, ..., x_n|\theta) \tag{12}$$

$$\Leftrightarrow f(\theta|x_1, ..., x_n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{13}$$

$$\Leftrightarrow f(\theta|x_1,...,x_n) \propto \theta^{\alpha-1+\sum_{i=1}^n x_i}(1-\theta)^{\beta-1+n-\sum_{i=1}^n x_i} \tag{14}$$

Ou seja, $\theta|x_1,...,x_n \sim \text{Beta}\Big(\alpha+\sum_{i=1}^n x_i,\ \beta+n-\sum_{i=1}^n x_i\Big)$ □

Teorema 2.4.2: Suponha que $X_1,...,X_n|\theta$ são uma amostra aleatória de variáveis com distribuição Poisson com parâmetro θ (Desconhecido). Suponha também que a distribuição a priori de θ é uma **Gamma** com parâmetros $\alpha > 0$ e $\beta > 0$. Então a distribuição a posteriori de $\theta|x_1,...,x_n$ é a distribuição Gamma com parâmetros $\alpha + \sum_{i=1}^n x_i$ e $\beta + n$

Demonstração: Seja $y = \sum_{i=1}^n x_i$, então a função de verossimilhança de $\mathbb{L}(\theta)$ satisfaz:

$$\mathbb{P}(\underline{x}|\theta) \propto e^{-n\theta}\theta^y \tag{15}$$

A priori $\xi(\theta)$ se estrutura assim:

$$\xi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta} \text{ para } \theta > 0 \tag{16}$$

Temos então que:

$$\begin{aligned} &f(\theta|\underline{x}) \propto e^{-n\theta}\theta^y\theta^{\alpha-1}e^{-\beta\theta} \\ \Leftrightarrow\ &f(\theta|\underline{x}) \propto \theta^{\alpha+y-1}e^{-(n+\beta)\theta} \end{aligned} \tag{17}$$

Ou seja, $\theta|\underline{x} \sim \text{Gamma}(\alpha+y, n+\beta)$ □

Teorema 2.4.3: Suponha que $X_1,...,X_n|\theta$ são uma amostra aleatória de variáveis com distribuição Normal com média θ (Desconhecido) e variância $\sigma^2 > 0$ conhecido. Suponha também que a distribuição a priori de θ é uma **Normal** com média μ_0 e variância v_0^2. Então a distribuição a posteriori de $\theta|x_1,...,x_n$ é a distribuição normal com média μ_1 e variância v_1^2 onde:

$$\mu_1 = \frac{\sigma^2\mu_0 + nv_0^2\tilde{x}_n}{\sigma^2 + nv_0^2} \tag{18}$$

e

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + nv_0^2} \tag{19}$$

Demonstração: Temos que:

$$\mathbb{L}(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2\right) \tag{20}$$

Temos que:

$$\sum_{i=1}^n (x_i-\theta)^2 = \sum_{i=1}^n x_i^2 - 2x_i\theta + \theta^2 \tag{21}$$

Definimos então $\tilde{x}_n := \frac{1}{n}\sum_{i=1}^n x_i$ e assim temos que:

$$\sum_{i=1}^{n} x_i^2 - 2x_i\theta + \theta^2 = n\theta^2 - 2n\tilde{x}_n\theta + \sum_{i=1}^{n} x_i^2$$

$$= n(\theta^2 - 2\theta\tilde{x}_n) + \sum_{i=1}^{n} x_i^2 = n(\theta^2 - 2\theta\tilde{x}_n + \tilde{x}_n^2) - n\tilde{x}_n + \sum_{i=1}^{n} x_i^2 \tag{22}$$

$$= n(\theta - \tilde{x}_n)^2 + \sum_{i=1}^{n} (x_i - \tilde{x}_n)^2$$

Temos então:

$$\begin{gathered} \mathbb{L}(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right) \\ \Leftrightarrow \mathbb{L}(\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\left(n(\theta - \tilde{x}_n)^2 + \sum_{i=1}^{n}(x_i - \tilde{x}_n)^2\right)\right) \end{gathered} \tag{23}$$

Temos que $\sum_{i=1}^{n}(x_i - \tilde{x}_n)^2$ não depende de θ então pode ir para a constante de proporcionalidade. De forma que

$$\mathbb{L}(\theta) \propto \exp\left(-\frac{n}{2\sigma^2}(\theta - \tilde{x}_n)^2\right) \tag{24}$$

Sabemos que a priori de θ segue a forma:

$$\xi(\theta) \propto \exp\left(-\frac{1}{2v_0^2}(\theta - \mu_0)^2\right) \tag{25}$$

Então temos que

$$f(\theta|\underline{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\theta - \tilde{x}_n)^2 + \frac{1}{v_0^2}(\theta - \mu_0)^2\right]\right\} \tag{26}$$

Se abrirmos os termos em quadrado, retirar as constantes, e completar os quadrados, chegamos nos resultados das equações (18) e (19), de forma que:

$$f(\theta|\underline{x}) \propto \exp\left[-\frac{1}{2v_1^2}(\theta - \mu_1)^2\right] \tag{27}$$

Ou seja, $f(\theta|\underline{x}) \sim N(\mu_1, v_1^2)$ □

Conseguimos dividir μ_1 da seguinte forma:

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + nv_0^2}\mu_0 + \frac{nv_0^2}{\sigma^2 + nv_0^2}\tilde{x}_n \tag{28}$$

Isso nos mostra que, conforme nossa amostra vai aumentando, o termo da direita referente à média amostral vai dominando. Mas o que isso quer dizer? Quer dizer que, independente do quanto você acredita que μ_0 seja a média verdadeira de θ, mais a média após a observação dos dados vai se aproximando de $\tilde{x}_n$, de forma que acabamos mudando de ideia aos poucos

Teorema 2.4.4: Suponha que $X_1, ..., X_n|\theta$ são uma amostra aleatória de variáveis com distribuição Exponencial com parâmetro $\theta > 0$ (Desconhecido). Suponha também que a distribuição a priori de θ é uma **Gamma** com parâmetros $\alpha > 0$ e $\beta > 0$. Então a distribuição a posteriori de $\theta|x_1, ..., x_n$ é a distribuição Gamma com parâmetros $\alpha + n$ e $\beta + \sum_{i=1}^{n} x_i$

Demonstração: Novamente vamos chamar $y := \sum_{i=1}^{n} x_i$. Então temos que a função de verossimilhança é:

$$\mathbb{L}(\theta) = \theta^n e^{-\theta y} \tag{29}$$

E a priori tem a forma:

$$\xi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \text{ para } \theta > 1 \tag{30}$$

Então temos que:

$$\begin{aligned} f(\theta|\underline{x}) &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^n e^{-\theta y} \\ \Leftrightarrow f(\theta|\underline{x}) &\propto \theta^{n+\alpha-1} e^{-(\beta+y)\theta} \end{aligned} \tag{31}$$

Ou seja, $f(\theta|\underline{x}) \sim \text{Gamma}(n + \alpha,\ \beta + y)$ □

2.5 Distribuições Impróprias

Definição 2.5.1 (Distribuição Imprópria): Seja $\xi : C \to \mathbb{R}$ uma função não-negativa cujo domínio inclui o espaço paramétrico ($\Omega \subset C$) de um modelo estatístico. Suponha também que:

$$\int_C \xi(\theta)\, \mathrm{d}\theta = \infty \tag{32}$$

Se nós imaginarmos que ξ é a f.d.p à priori de θ, então ξ é uma **distribuição imprória** de θ

Um bom exemplo é utilizar a distribuição **beta** assumindo que $\alpha = \beta = 0$. Mesmo que isso viole a condição da distribuição beta, o resultado da posteriori ainda sim é uma distribuição beta. Porém, existem diversos métodos para se escolher uma distribuição imprópria para θ. O mais comum é se utilizar de uma família de conjugados para o modelo estatístico, e forma a adaptarmos seus parâmetros para obter uma distribuição imprópria.

Estimadores de Bayes

3.1 Estimador e Estimativa

Com estimadores, queremos, a partir, puramente, de nossas observações dos dados gerar uma função que, ao longo prazo, converge para uma medida de nosso interesse (Um parâmetro de distribuição, por exemplo)

Definição 3.1.1 (Estimador/Estimativa): Seja $X_1, ..., X_n$ os dados observados que a distribuição conjunta é indexada por um parâmetro θ e assume valores em um conjunto Ω na reta real (Cada observação X_i). Um estimador do parâmetro θ é uma função $\delta : \Omega^n \to \mathbb{R}$ ($\delta(X_1, ..., X_n)$). Se $X_1 = x_1, ..., X_n = x_n$ são observados, então $\delta(x_1, ..., x_n)$ é uma estimativa de θ

Vale ressaltar a diferença entre **estimador** e **estimativa**. O **estimador** é uma função das variáveis aleatórias, ou seja, ele também é uma variável aleatória e pode ter sua distribuição derivada da distribuição conjunta de $X_1, ..., X_n$. Já uma **estimativa** é o resultado de $\delta(\underline{X})$ após serem observado os valores $x_1, ..., x_n$

3.2 Função de Perda

Muito comumente, criamos um estimador δ com o objetivo de aproximar um parâmetro θ, ou seja, um bom estimador é aquele que $\delta(\underline{x}) - \theta \approx 0$

Definição 3.2.1 (Função de perca): A função de perca é uma função real de duas variáveis $L(\theta, a)$, onde $\theta \in \Omega$ e $a \in \mathbb{R}$. A interpretação é que $L(\theta, a)$ decai conforme $a \to \theta$

Queremos estimar θ apenas com nossos valores observados, porém, vamos supor que não vimos nenhum ainda, então se escolhermos a como uma estimativa, vamos ter:

$$\mathbb{E}[L(\theta, a)] = \int_\Omega L(\theta, a)\xi(\theta)\,\mathrm{d}\theta \qquad (\text{LOTUS}) \tag{33}$$

3.3 Estimador de Bayes

Supondo agora que nós temos acesso as observações $\underline{x}$. Então também temos acesso à distribuição posteriori $\xi(\theta|x_1, ..., x_n)$, então podemos escolher uma estimativa a tal que ela minimize:

$$\mathbb{E}[L(\theta, a)|\underline{x}] = \int_\Omega L(\theta, a)\xi(\theta|\underline{x})\,\mathrm{d}\theta \tag{34}$$

Definição 3.3.1 (Estimador de Bayes): Seja $L(\theta, a)$ uma função de perca. Para cada valor possível $\underline{x}$ de $\underline{X}$, deixe que $\delta^*(\underline{x})$ ser o valor de a que minimiza $\mathbb{E}[L(\theta, a)]$ é minimizado. Então δ^* é chamado de **Estimador de Bayes** de θ. Uma vez que $\underline{X} = \underline{x}$ é observado, chamamos $\delta^*(\underline{x})$ de **estimativa bayesiana** de θ

Podemos também descrever como, para todos os valores possíveis de $\underline{x}$, queremos:

$$\mathbb{E}(L(\theta, \delta^*(\underline{x})|\underline{x})) = \min_{\text{Todos } a} \mathbb{E}(L(\theta, a)|\underline{x}) \tag{35}$$

Algumas percas de função muito comum são:

$$L(\theta, a) = (\theta - a)^2 \tag{36}$$

$$L(\theta, a) = |\theta - a| \tag{37}$$

Teorema 3.3.1: Seja $L(\theta, a) = (\theta - a)^2$, então $\delta^*(\underline{x}) = \mathbb{E}(\theta|\underline{x})$

Demonstração: Queremos provar que

$$\mathbb{E}\big[(X - \mu)^2\big] \leq \mathbb{E}\big[(X - d)^2\big] \qquad \forall d \in \mathbb{R} \tag{38}$$

E a igualdade só vale quando $\mu = d$. Ou seja:

$$\mu = \text{argmin}_{d \in \mathbb{R}} \mathbb{E}\big[(X - d)^2\big] \tag{39}$$

Então:

$$\begin{aligned} \mathbb{E}\big[(X - d)^2\big] &= \mathbb{E}\big[X^2 - 2Xd + d^2\big] \\ \mathbb{E}\big[X^2\big] - 2d\mathbb{E}[X] + d^2 &= \mathbb{E}\big[X^2\big] - 2d\mu + d^2 \end{aligned} \tag{40}$$

Como queremos minimizar isso, com relação a d, vamos derivar:

$$\frac{\partial}{\partial d}(\mathbb{E}\big[X^2\big] - 2d\mu + d^2) = -2\mu + 2d \tag{41}$$

E isso é igual a 0 quando $d = \mu$ $\square$

Teorema 3.3.2: Seja $L(\theta, a) = |\theta - a|$, então $\delta^*(\underline{x})$ é a mediana de $\theta|\underline{x}$

Demonstração:

$$\mathbb{E}|X - a| \geq \mathbb{E}|X - m| \qquad \forall a \in \mathbb{R} \tag{42}$$

Então queremos provar que

$$\mathbb{E}|X - a| - \mathbb{E}|X - m| \geq 0 \tag{43}$$

Vamos assumir que $m < a$ ($m > a$ é análogo)

Se $X \leq m$, então: $|X - a| - |X - m| = a - X - (m - X) = a - m$

Se $X > m$, então: $|X - a| - |X - m| = X - a - X + m = m - a$

Defina então $Y = |X - a| - |X - m|$. Defina também:

$$\mathbb{I}_X = \begin{cases} 1 \text{ se } X \leq m \\ 0 \text{ se } X > m \end{cases} \tag{44}$$

Então teremos que:

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}(Y \cdot \mathbb{I}_X) + \mathbb{E}(Y \cdot (1 - \mathbb{I}_X)) \\ &\geq (a - m)\mathbb{E}(\mathbb{I}_X) + (m - a)\mathbb{E}(1 - \mathbb{I}_X) \\ &= (a - m)\mathbb{P}(X \leq m) + (m - a)\mathbb{P}(X > m) \\ &= (a - m)\mathbb{P}(X \leq m) - (a - m)(1 - \mathbb{P}(X \leq m)) \\ &= (a - m)(2\mathbb{P}(X \leq m) - 1) \geq 0 \end{aligned} \tag{45}$$

Porém essa equação final é satisfeita pela definição de mediana! □

Quando estamos tentando tentando estimar um parâmetro θ, queremos que, quanto mais amostras tivermos, ou seja, quando $n \to \infty$, o nosso estimador vai convergindo para θ

Definição 3.3.2 (Consistência): Quando uma sequência $(\delta_n)_{n\geq 1}$ converge para o valor verdadeiro do parâmetro θ, dizemos que $(\delta_n)_{n\geq 1}$ é consistente para θ

Ou seja, com grandes quantidades de dados, a probabilidade do estimador $\hat{\theta}$ estar **muito** próximo de θ é alta

3.4 Estimadores para Parâmetros mais gerais

Até agora nós vimos estimadores para os parâmetros em si, porém, as vezes podemos estar interessados em outras generalizações. Um exemplo de generalização é para estimar, por exemplo, dois parâmetros de uma só vez, como estimar uma média e uma variância (Saída multivariada) ou uma função do parâmetro em si, por exemplo, se θ é a taxa de falha, então podemos querer estimar $1/\theta$ que é a média de falhas

Definição 3.4.1 (Estimador/Estimativa): Seja $X_1, ..., X_n$ serem dados observados em que a distribuição conjunta é dado um parâmetro $\theta \in \Omega \subset \mathbb{R}^k$. Defina $h : \Omega \to \mathbb{R}^d$. Defina $\psi = h(\theta)$. Um **estimador** de ψ é a função $\delta(X_1, ..., X_n) : \mathbb{R}^n \to \mathbb{R}^d$. Se $X_1 = x_1, ..., X_n = x_n$ são observados, então $\delta(x_1, ..., x_n)$ é uma **estimativa** de ψ

Estatística Frequentista

Até agora, tratamos o parâmetro θ como uma variável com distribuição, onde ele poderia ter vários valores possíveis. Porém, agora vamos fazer uma abordagem diferente. Vamos imaginar que nosso valor de θ é **fixo**, ele é um número pronto que **simplesmente não conhecemos**, então ele não tem distribuição nem nada do gênero

4.1 Estimadores/Estimações de Máxima Verossimilhança

Já que estamos trabalhando com um valor fixo, sem distribuições nem nada do gênero, os métodos que vimos para a estimação de parâmetros não vai funcionar. Porém, tem algum método para que eu consiga fazer essa estimação sem ter uma distribuição priori/posteriori? **SIM**. Vamos pensar de uma maneira **intuitiva**, o que vou falar aqui é apenas uma aproximação intuitiva de como funciona o método, mas depois eu vou explicar o porquê de não ser exatamente o que estou dizendo

Quando jogamos uma moeda, e observamos uma proporção de **caras** de $1/20$, por exemplo, obviamente pensamos: "Essa moeda ta muito viciada". E então pensamos que a probabilidade de cair cara seja de **aproximadamente** $1/20$. Essa intuição que temos é isso (Não exatamente) o que o método da **Estimação por Máxima Verossimilhança** faz

Definição 4.1.1 (Estimadores/Estimação de Máxima Verossimilhança): Seja $f(\underline{X}|\theta)$ a p.f ou a p.d.f de $X_1, ..., X_n|\theta$ e $\theta \in \Omega \subset \mathbb{R}^m$, a função $\delta(\underline{X}) = \max_\theta f(\underline{X}|\theta)$ é chamada de **Estimador de Máxima Verossimilhança**. Quando os valores $\underline{x} = (x_1, ..., x_n)$ são observados, então $\delta(\underline{x}) = \max_\theta f(\underline{x}|\theta)$ é chamado de **Estimativa de Máxima Verossimilhança**

Essa definição parece ser bem intuitiva, correto? Vamos ver um exemplo:

Exemplo 4.1.1: Suponha que você tem n variáveis de tal forma que:

$$X_1, ..., X_n \mid \theta \sim \text{Bern}(\theta) \tag{46}$$

De forma que θ é desconhecido. Para todos os valores $x_1, ..., x_n$ observados, temos que a p.m.f conjunta é:

$$f(\underline{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \tag{47}$$

Em vez de maximizar esse troço, vamos maximizar o $\log$ dele, já que, como ela é crescente, o máximo do $\log$ também é o máximo da função em si. Tirando o log, temos:

$$\begin{aligned} \ln(f(\underline{x}|\theta)) &= \sum_{i=1}^{n} x_i \ln(\theta) + (1-x_i)\ln(1-\theta) \\ &= \left(\sum_{i=1}^{n} x_i\right)\ln(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-\theta) \end{aligned} \tag{48}$$

Derivando isso em relação a θ e igualando a 0, vamos ter que:

$$\begin{aligned} 0 &= \left(\sum_{i=1}^{n} x_i\right)\frac{1}{\theta} - \left(n - \sum_{i=1}^{n} x_i\right)\frac{1}{1-\theta} \\ 0 &= (n\bar{x}_n)\frac{1}{\theta} - n(1-\bar{x}_n)\frac{1}{1-\theta} \\ 0 &= \frac{\bar{x}_n}{\theta} - \frac{1-\bar{x}_n}{1-\theta} \end{aligned} \tag{49}$$

$$
\begin{aligned}
0 &= (1-\theta)\bar{x}_n - (1-\bar{x}_n)\theta \\
0 &= \bar{x}_n - \theta\bar{x}_n - \theta + \theta\bar{x}_n \\
\theta &= \bar{x}_n
\end{aligned} \tag{50}
$$

Ou seja, chegamos a uma conclusão razoável que o estimador que minimiza a verossimilhança é a média das variáveis de Bernoulli

Porém, nem sempre essa abordagem é viável, vamos ver um exemplo:

Exemplo 4.1.2: Suponha que temos uma amostra $X_1, ..., X_n|\theta \sim \text{Unif}[0, \theta]$, de forma que a distribuição é levemente alterada para ser da forma:

$$
f(x|\theta) = \begin{cases} \frac{1}{\theta} \text{ para } 0 < x < \theta \\ 0 \text{ caso contrário} \end{cases} \tag{51}
$$

Temos então que a máxima verossimilhança é:

$$
f(\underline{x}|\theta) = \begin{cases} \frac{1}{\theta^n} \text{ para } 0 < x_i < \theta \ (i = 1, ..., n) \\ 0 \text{ caso contrário} \end{cases} \tag{52}
$$

Como $1/\theta^n$ é uma função decrescente, o valor de θ que maximiza a função e ainda se encaixa na restrição $\theta > \max\{x_1, ..., x_n\}$ seria $\theta = \max\{x_1, ..., x_n\}$, porém, não podemos usar esse valor por conta da desigualdade **estrita** na função de verossimilhança. O que isso quer dizer? Que esse caso **não possui um Estimador de Máxima Verossimilhança**

O livro também aborda outros casos em que um mesmo caso pode ter **vários** estimadores ou casos em que um estimador, mesmo existindo, não demonstra ser um valor interessante/desejado.

Mas eu falei antes que esses estimadores formalizavam a ideia de "O parâmetro θ parece ser esse daqui", mas não é bem assim que funciona. A gente viu que ele é o parâmetro que **maximiza** a probabilidade daquela observação ocorrer. E qual é a diferença disso pro que falei antes? Simples, quando vemos os valores de observações, o fato de eles aparecerem daquela forma, não significa que o valor que θ mais aparenta ser seja o real. Ué, como assim? Muitos fatores podem estar envolvidos, fatores que, logo de cara, não conseguimos observar apenas nos dados. Para que isso ocorresse, os dados deveriam conter muito mais informação do que tinhamos à priori (Antes das observações)

4.2 Propriedades

Teorema 4.2.1: Se $\hat{\theta} \in \Omega$ é o Estimador de Máxima Verossimilhança (EMV) de θ e g é uma função bijetiva, então $g\left(\hat{\theta}\right)$ é o EMV de $g(\theta)$

Demonstração: Seja Γ o novo espaço paramétrico, ou seja, $g : \Omega \to \Gamma$. Vamos definir h como sendo a função inversa, ou seja $\theta = h(\psi)$. Se expressarmos a p.d.f em função de ψ, vamos obter $f(x|h(\psi))$ e a função de verossimilhança será $f(\underline{x}|h(\psi))$.

Sabemos que o EMV $\hat{\psi}$ de ψ é vai ser o valor de ψ que maximiza $f(\underline{x}|h(x))$. Como $f(x|\theta)$ é maximizada quando $\theta = \hat{\theta}$, então $h(\psi) = \hat{\theta}$ maximiza a verossimilhança. Porém, aplicando g em ambos os lados, obtemos que:

$$
\hat{\psi} = g\left(\hat{\theta}\right) \tag{53}
$$

□

Essa propriedade é algo ótimo! Tendo em vista que no método anterior, o estimador de Bayes de $1/\theta$ podia ser diferente de $1/\hat{\theta}$. Porém, podemos estender esse teorema para casos em que a função g não é bijetiva. Vamos então definir o **estimador de uma função**

Definição 4.2.1 (MVE de uma Função): Seja $g(\theta)$ uma função arbitrária do parâmetro com $g : \Omega \to G$. Para cada $t \in G$, defina $G_t := \{\theta : g(\theta) = t\}$ e $L^*\big(\hat{t}\big) := \max_{\theta \in G_t} \log f(\underline{x}|\theta)$, defina então o EMV de $g(\theta)$ como $\hat{t}$ como:

$$L^*\big(\hat{t}\big) = \max_{t \in G_t} L^*(t) \tag{54}$$

Teorema 4.2.2: Dado $\hat{\theta}$ sendo o EMV de θ e $g : \Omega \to G$, então:

$$\widehat{g(\theta)} = g\Big(\hat{\theta}\Big) \tag{55}$$

Demonstração: Como $L^*(t)$ é o máximo de $\log f(\underline{x}|\theta)$ em θ num subconjunto de Ω, e como $\log f\Big(\underline{x}|\hat{\theta}\Big)$ é o máximo sob todos os θ, então sabemos que $L^*(t) \leq \log f(\underline{x}|\theta)\ \forall t \in G$. Denote $\hat{t} = g\Big(\hat{\theta}\Big)$. Perceba que $\hat{\theta} \in G_{\hat{t}}$. Como $\hat{\theta}$ maximiza $f(\underline{x}|\theta)$ em todos θ, então ele também maximiza $f(\underline{x}|\theta)$ sob todos os $\theta \in G_{\hat{t}}$. Por isso, $L^*\big(\hat{t}\big) = \log f\Big(\underline{x}|\hat{\theta}\Big)$ e $\hat{t} = g\Big(\hat{\theta}\Big)$ é um EVM de $g(\theta)$ □

4.3 Computação Numérica

Muitos problemas possuem um EVM $\hat{\theta}$ de um parâmetro θ, porém esses não podem ser computados com fórmulas fechadas. Nesses casos, precisamos utilizar de métodos numéricos para aproximações. Existem **inúmeros** métodos de aproximação numérica de funções, porém, aqui vamos abordar brevemente apenas um

Definição 4.3.1 (Método de Newton): Seja $f(\theta)$ uma função real de uma variável e suponha que nós desejamos resolver a equação $f(\theta) = 0$. Seja θ_0 um chute inicial da solução e θ_t o valor obtido na t-ésima iteração do programa. O método de Newton atualiza nossa resposta da seguinte forma:

$$\theta_{t+1} = \theta_t - \frac{f(\theta_t)}{f'(\theta_t)} \tag{56}$$

Se pararmos para interpretar, o que o algoritmo faz é checar se eu tenho que mexer θ_t para frente ou para trás dependendo do sinal e da inclinação de f. Quando $f(\theta_t)$ é negativo e $f'(\theta_t)$ é positivo, então eu preciso mover para a direita para poder chegar próximo a raíz, e aí vai

Método dos Momentos

Lembra que estimamos o primeiro momento $\mathbb{E}[X] \approx 1/n \cdot \sum_{i=1}^{n} X_i$? Será que da pra estimar os outros momentos de uma forma parecida? Na verdade sim! É bem intuitivo:

Definição 5.1 (Método dos Momentos): Suponha que $X_1, ..., X_n$ formam uma amostra aleatória de uma distribuição indexada por um parâmetro k-dimensional θ e que tem, pelo menos, k momentos finitos. Para $j = 1, ..., k$, deixe $\mu_j(\theta) = \mathbb{E}\left[X_1^j \mid \theta\right]$. Suponha que a função $\mu(\theta) = (\mu_1(\theta), ..., \mu_k(\theta))$ é uma função bijetiva de θ. Seja $M(\mu_1, ..., \mu_k)$ a função inversa, ou seja, para todo θ é válido que:

$$\theta = M(\mu_1(\theta), ..., \mu_k(\theta)) \tag{57}$$

Defina os *momentos amostrais* como:

$$m_j = \frac{1}{n} \sum_{i=1}^{n} (X_i)^j \tag{58}$$

Para $j = 1, ..., k$. O *método do estimador de momentos* de θ é $M\big(m_1, ..., m_j\big)$

O método mais usual de se implementar esse método é resolvendo todas as equações $m_j = \mu_j(\theta)$ e então resolver para θ

Teorema 5.1 (Consistência): Suponha que $X_1, X_2, ...$ são i.i.d com uma distribuição indexada por um parâmetro k-dimensional θ. Suponha também que o os primeiros k momentos da distribuição são finitos e existem para todo θ. Suponha também que a função inversa M é definida como na Definição 5.1 e é contínua. Então a sequência de estimadores pelo método dos momentos baseada em $X_1, ..., X_n$ é uma sequência consistente de estimadores de θ

Demonstração: A Lei dos Grandes Números diz que os momentos amostrais convergem em probabilidade para os momentos $\mu_1(\theta), \mu_2(\theta), ..., \mu_k(\theta)$. Isso implica que, ao generalizar isso para funções de k variáveis isso implica que M, nos momentos amostrais, converge em probabilidade para θ □

Estatística Suficiente

A gente viu alguns métodos para fazer estimativas que consistiam no uso de distribuições priori e posteriori. Porém, há alguns métodos que estimam dos parâmetros apenas utilizando de distribuições condicionais de funções dos dados.

A gente viu antes que nem sempre os métodos de estimativa que a gente tem vão ser interessantes ou vão ser boas estimativas (Para melhor detalhes, pode conferir os exemplos do livro). Nesses casos, precisamos desenvolver métodos novos de estimativa para nossos parâmetros.

Imagina que nós temos uma amostra elatória $X_1, ..., X_n$ e temos dois estatísticos, o A e o B. Vamos supor que A tem acesso a todos os valores de $X_1, ..., X_n$, enquanto B só pode saber sobre uma estimativa específica $T = r(X_1, ..., X_n)$. Qual deles vai poder fazer melhores estimativas para θ? Com certeza o mano estatístico A! Porém, entretudo, todavia, em alguns problemas, o estatístico B pode fazer estimativas tão bem quanto o estatístico A, pois a função T pode, de alguma forma, conter todas as informações relevantes e necessárias para que meu problema possa ser solucionado! Quando T tem essa característica, chamamos ela de **estatística suficiente**

> **Definição 6.1** (Estatística Suficiente): Seja $X_1, ..., X_n$ uma amostra aleatória de uma distribuição indexada pelo parâmetro θ e T uma estatística. Suponha também que, para todo θ e todo valor possível t de T, a distribuição conjunta de $X_1, ..., X_n | T = t, \theta$ depende apenas de t e não de θ. Ou seja, para cada t, a distribuição conjunta de $X_1, ..., X_n | T = t, \theta$ é a mesma **para todo** θ. Então dizemos que T é uma **estatística suficiente para o parâmetro θ**

A principal característica que separa estatísticas suficientes de não-suficientes é a dependência no valor de θ. O livro traz um processo chamado **randomização auxiliar**, que consiste em simular variáveis $X'_1, ..., X'_n$ com mesma distribuição que $X_1, ..., X_n | \theta$, porém, essas variáveis simuladas são feitas baseando-se única e exclusivamente numa estatística suficiente T. Se a estatística T não fosse suficiente, não conseguiriamos nem fazer a randomização auxiliar, pois necessariamente precisariamos saber qual seria o valor de θ

Com esse processo fica fácil de ver agora o porquê de o estatístico B conseguir se sair tão bem quanto o estatístico A. Se A vai utilizar de um estimador $\delta(X_1, ..., X_n)$ para estimar θ, se B tiver acesso a estatística suficiente T, então ele pode criar variáveis auxiliares $X'_1, ..., X'_n$ que tem mesma distribuição que as originais, então de B utilizar o estimador δ porém aplicando as suas variáveis $\delta(X'_1, ..., X'_n)$, então a distribuição do estimador de A é a mesma distribuição do estimador de B

6.1 Critério de Fatoração

Agora nos é apresentado um método de identificar estatísticas suficientes (Um teorema bem interessante) desenvolvido por R. A. Fisher em 1922, J. Neyman em 1935 e P. R. Halmos e L. J. Savage em 1949

Teorema 6.1.1 (Critério da Fatoração): Sejam $X_1, ..., X_n$ uma amostra aleatória de uma distribuição contínua ou discreta para qual a p.d.f ou a p.f é $f(x|\theta)$, onde o valor de θ é desconhecido e pertence a um espaço paramétrico Ω. Uma estatística $T = r(X_1, ..., X_n)$ é **suficiente** para θ **se, e somente se,** a função de densidade ou de massa conjunta $f(\underline{x}|\theta)$ de $X_1, ...X_n$ pode ser fatorada, para todos os valores de $x = (x_1, ..., x_n) \in \mathbb{R}^n$ e para todos os valores de $\theta \in \Omega$, da seguinte forma:

$$f(\underline{x}|\theta) = u(\underline{x})v(r(\underline{x}), \theta) \tag{59}$$

Onde u e v são não-negativas. u pode depender em $\underline{x}$ mas não em θ e v pode depende em θ e na estatística t

Demonstração: Vamos fazer a prova só para o caso que X tem uma distribuição discreta, ou seja:

$$f(\underline{x}|\theta) = \mathbb{P}(X_1 = x_1, ..., X_n = x_n|\theta) \tag{60}$$

Primeiro vamos fazer a volta. Então vamos supor que $f(\underline{x}|\theta)$ pode ser fatorado daquela forma. Para cada valor possível t da estatística T, denote $A(t)$ como sendo o conjunto de todos os pontos $\underline{x} \in \mathbb{R}^n$ tal que $r(\underline{x}) = t$. Para cada valor de $\theta \in \Omega$, nós vamos determinar a distribuição condicional de $\underline{X}$ dado $T = t$, então para todo ponto $\underline{x} \in A(t)$:

$$\mathbb{P}(\underline{X} = \underline{x}|T = t, \theta) = \frac{\mathbb{P}(\underline{X} = \underline{x}|\theta)}{\mathbb{P}(T = t|\theta)} = \frac{f(\underline{x}|\theta)}{\sum_{\underline{y} \in A(t)} f\big(\underline{y}|\theta\big)} \tag{61}$$

Como $r\big(\underline{y}\big) = t$ para todo ponto $\underline{y} \in A(t)$ e como $\underline{x} \in A(t)$, então vamos ter que:

$$\mathbb{P}(\underline{X} = \underline{x}|T = t, \theta) = \frac{u(\underline{x})}{\sum_{\underline{y} \in A(t)} u\big(\underline{y}\big)} \tag{62}$$

Finalmente, para todo ponto $\underline{x} \notin A(t)$:

$$\mathbb{P}(\underline{X} = \underline{x}|T = t, \theta) = 0 \tag{63}$$

Conseguimos ver, então, que a distribuição conjunta de X não depende de θ, mas somente de T. Logo, por definição, T é uma estatística suficiente

Agora para fazer a ida, vamos supor que T é uma estatística suficiente. Então para todo valor dado t de T, todo ponto $\underline{x} \in A(t)$ e todo valor $\theta \in \Omega$ a probabiidade condicional $\mathbb{P}(\underline{X} = \underline{x}|T = t, \theta)$ não vai depender de θ (Como vimos anteriormente), então vai ser da forma:

$$\mathbb{P}(\underline{X} = \underline{x}|T = t, \theta) = u(\underline{x}) \tag{64}$$

Se chamarmos $\mathbb{P}(T = t|\theta) = v(t, \theta)$, então temos:

$$\begin{aligned} \mathbb{P}(\underline{X} = \underline{x}|\theta) &= \mathbb{P}(\underline{X} = \underline{x}|T = t, \theta)\mathbb{P}(T = t|\theta) \\ &= u(\underline{x})v(t, \theta) \end{aligned} \tag{65}$$

Assim, provamos (Para o caso discreto) que o teorema é válido (Também é válido para o caso contínuo, mas requer métodos diferentes e não será abordado) □

Nós vimos anteriormente que, ao calcular a posteriori, ela era proporcional única e excluisivamente ao valor de θ, de forma que tudo que não era relacionado a θ podia ser movido para a constante de proporcionalidade. Algo parecido ocorre aqui!

Corolário 6.1.1.1: Uma estatística $T = r(\underline{X})$ é suficiente **se, e somente se,** não importa qual seja a distribuição a priori, a distribuição a posteriori de θ depende dos dados única e exclusivamente através do valor de T