Inferência Bayesiana

Notas de Aula

Luís Gustavo Esteves, Rafael Izbicki e Rafael Bassi Stern

Última revisão: 16 de Dezembro de 2019

Por favor, enviem comentários, typos e erros para rbstern@gmail.com

Agradecimentos: Gratos pelas sugestões de Michelangelo dos Anjos, Yuri Benites, Ian Danilevicz, Andressa Dantas, Thales Egydio, Luís Esteves, Jeremias Leão, Tarcísio Lobato, Rafael Paixão, Carlos Pereira, João Poloniato, Aimée Shirozono, João Silva, Julio Stern, Aline Tonon, Sergio Wechsler e Victor Zoré.

"Teaching is giving opportunities to students to discover things by themselves."

George Pólya

Conteúdo

```
opts_chunk$set(cache=TRUE,fig.path="./figures/")
```

1 Revisão

1.1 Teoria dos conjuntos

Teoria dos Conjuntos é o fundamento para a definição da matemática moderna. Em particular, ela é usada para definir a Teoria da Probabilidade. Conjuntos são usados para definir eventos. Esta seção faz uma revisão rápida e focada de Teoria dos Conjuntos.

Um conjunto é uma coleção de objetos. Se um conjunto é composto por um número finito de objetos, $w_1, w_2, \ldots, w_n$, denotamos este conjunto por $\{w_1, w_2, \ldots, w_n\}$. Alguns conjuntos são usados com tanta frequência que recebem símbolos especiais para designá-los:

- $\mathbb{N}$: Os números naturais, $\{0, 1, 2, 3, \ldots\}$.
- $\mathbb{Z}$: Os números inteiros, $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
- $\mathbb{R}$: Os números reais.

Exemplo 1.1 (Conjuntos)**.**

- O conjunto de resultados em um dado de 6 faces: $\{1, 2, 3, 4, 5, 6\}$.
- O conjunto de resultados em um lançamento de moeda: $\{T, H\}$.
- O conjunto de resultados em dois lançamentos de moeda: $\{(T, T), (T, H), (H, T), (H, H)\}$.
- O conjunto de números ímpares: $\{2n + 1 : n \in \mathbb{N}\}$ ou $\{1, 3, 5, 7, \ldots\}$.
- O conjunto de números reais não negativos: $\{x \in \mathbb{R} : x \geq 0\}$.
- Um círculo de raio 1: $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$.

Definição 1.2 ($\in$ e $\notin$)**.** Escrevemos $w \in S$ se o objeto w é um elemento do conjunto S e $w \notin S$, caso contrário.

Exemplo 1.3 ($\in$ e $\notin$)**.**

- $T \in \{T, H\}$.
- $7 \notin \{1, 2, 3, 4, 5, 6\}$.
- $7 \in \{2n + 1 : n \in \mathbb{N}\}$.

Definição 1.4 (**Conjunto vazio** - $\emptyset$)**.** $\emptyset$ é o único conjunto sem elementos. Isto é, para todo objeto w, $w \notin \emptyset$.

Definição 1.5 (**Conjuntos disjuntos**)**.**

- Dois conjuntos A e B são **disjuntos** se, para todo $w \in A$, temos $w \notin B$ e para todo $w \in B$, $w \notin A$.
- Uma sequência de conjuntos $(A_n)_{n\in\mathbb{N}}$ é disjunta se, para todo $i \neq j$, A_i é disjunto de A_j.

Exemplo 1.6 (Conjuntos disjuntos)**.**

- $\{1,2\}$ e $\{3,4\}$ são disjuntos.
- $\{1,2\}$ e $\{2,3\}$ não são disjuntos pois $2 \in \{1,2\}$ e $2 \in \{2,3\}$.

Definição 1.7 ($\subset$ e $=$)**.** Sejam A e B dois conjuntos. Dizemos que:

- $A \subset B$ se, para todo $w \in A$, $w \in B$.
- $A = B$ se $A \subset B$ e $B \subset A$.

Exemplo 1.8 ($\subset$ e $=$)**.**

- $\{1,2\} \subset \{1,2,3,4\}$.
- $\{n \in \mathbb{Z} : n \geq 1\} \subset \mathbb{N}$.
- $\{n \in \mathbb{Z} : n \geq 0\} = \mathbb{N}$.

Reservamos o símbolo Ω para o conjunto de todos os objetos considerados em um dado modelo. Ω é chamado em Teoria da Probabilidade de **Espaço Amostral**. Isto é, para todo conjunto A considerado no modelo, $A \subset \Omega$.

1.1.1 Operações sobre conjuntos

Definição 1.9 (complemento - ^{c})**.** Seja A um conjunto. w é um elemento de A^c se e somente se $w \notin A$. Isto é, o complemento de A é definido como $A^c = \{w \in \Omega : w \notin A\}$. Note que, para determinar A^c, é necessário conhecer Ω.

Exemplo 1.10 (^{c})**.**

- Seja $\Omega = \{T, H\}$, $\{T\}^c = \{H\}$.
- Seja $\Omega = \{1,2,3,4,5,6\}$, $\{1,2\}^c = \{3,4,5,6\}$.
- Seja $\Omega = \mathbb{N}$, $\{n \in \mathbb{N} : n > 0\}^c = \{0\}$.

Definição 1.11 (união - $\cup$)**.**

- Sejam A e B dois conjuntos, $w \in \Omega$ é um elemento da união entre A e B, se e somente se w é elemento de A **ou** w é elemento de B. Isto é, $A \cup B = \{w \in \Omega : w \in A \textbf{ ou } \in B\}$.
- Seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de conjuntos. $w \in \Omega$ é um elemento da união de $(A_n)_{n\in\mathbb{N}}$, $\cup_{n\in\mathbb{N}} A_n$, se e somente se existe $n \in \mathbb{N}$ tal que $w \in A_n$. Isto é, $\cup_{n\in\mathbb{N}} A_n = \{w \in \Omega : \textbf{existe } n \in \mathbb{N} \text{ tal que } w \in A_n\}$

Exemplo 1.12 ($\cup$)**.**

- $\{T\} \cup \{H\} = \{T, H\}$.
- $\{1,2\} \cup \{2,3\} = \{1,2,3\}$.
- $\{1\} \cup \{3\} \cup \{5\} = \{1,3,5\}$.
- $\{n \in \mathbb{Z} : n > 0\} \cup \{n \in \mathbb{Z} : n < 0\} = \{n \in \mathbb{Z} : n \neq 0\}$.
- $\cup_{n\in\mathbb{N}}\{n\} = \mathbb{N}$.

- $\cup_{n\in\mathbb{N}}\{x \in \mathbb{R} : x \geq n\} = \{x \in \mathbb{R} : x \geq 0\}$.
- $\cup_{n\in\mathbb{N}}\{x \in \mathbb{R} : x \geq 1/(n+1)\} = \{x \in \mathbb{R} : x > 0\}$.

Definição 1.13 (intersecção - $\cap$)**.**

- Sejam A e B dois conjuntos. ω é elemento da intersecção entre A e B, $A \cap B$, se e somente se $w \in \Omega$ é um elemento de A **e** w é um elemento de B. Isto é, $A \cap B = \{w \in \Omega : w \in A \textbf{ e } w \in B\}$.
- Seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de conjuntos, $w \in \Omega$ é elemento da intersecção de $(A_n)_{n\in\mathbb{N}}$, $\cap_{n\in\mathbb{N}}A_n$, se e somente se para todo $n \in \mathbb{N}$, $w \in A_n$. Isto é, $\cap_{n\in\mathbb{N}}A_n = \{w \in \Omega : \textbf{para todo } n \in \mathbb{N}, w \in A_n\}$

Exemplo 1.14 ($\cap$)**.**

- $\{T\} \cap \{H\} = \emptyset$.
- $\{1, 2\} \cap \{2, 3\} = \{2\}$.
- $(\{1, 2\} \cap \{2, 3\}) \cup \{5\} = \{2, 5\}$.
- $\{n \in \mathbb{Z} : n \geq 0\} \cap \{n \in \mathbb{Z} : n \leq 0\} = \{0\}$.
- $\cap_{n\in\mathbb{N}}\{i \in \mathbb{N} : i \geq n\} = \emptyset$.
- $\cap_{n\in\mathbb{N}}\{x \in \mathbb{R} : x \leq n\} = \{x \in \mathbb{R} : x \leq 0\}$.

Teorema 1.15 (Lei de De Morgan)**.** *Seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de subconjuntos de Ω. Para todo $n \in \mathbb{N}$,*

- $(\cup_{i=1}^{n} A_i)^c = \cap_{i=1}^{n} A_i^c$
- $(\cap_{i=1}^{n} A_i)^c = \cup_{i=1}^{n} A_i^c$

Ademais,

- $(\cup_{i\in\mathbb{N}} A_i)^c = \cap_{i\in\mathbb{N}} A_i^c$
- $(\cap_{i\in\mathbb{N}} A_i)^c = \cup_{i\in\mathbb{N}} A_i^c$

Definição 1.16 (Partição)**.** Seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de conjuntos. Dizemos que $(A_n)_{n\in\mathbb{N}}$ particiona Ω se:

- Para todo $i, j \in \mathbb{N}$ tais que $i \neq j$, A_i e A_j são disjuntos.
- $\cup_{n\in\mathbb{N}} A_n = \Omega$.

1.2 Teoria da probabilidade

1.2.1 Probabilidade

Definição 1.17 (Axiomas da Probabilidade)**.** $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ é uma probabilidade se:

1. (Não-negatividade) Para todo $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$.
2. (Aditividade) Se $(A_n)_{n\in\mathbb{N}}$ é uma sequência de conjuntos disjuntos em $\mathcal{F}$, $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$.
3. (Normalização) $\mathbb{P}(\Omega) = 1$.

Lema 1.18. $\mathbb{P}(\emptyset) = 0$.

Lema 1.19. *Se $A_1, A_2, \ldots, A_n$ são disjuntos, então $\mathbb{P}(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.*

Lema 1.20. *Para todo A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.*

Lema 1.21. *Para quaisquer eventos A e B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.*

Lema 1.22. *Se $A \subset B$, então $\mathbb{P}(B) \geq \mathbb{P}(A)$.*

Definição 1.23 (Axioma da Probabilidade Condicional)**.**

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$$

Definição 1.24. Dois eventos A e B são independentes se $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Lema 1.25. *Dois eventos A e B são independentes se e somente se $\mathbb{P}(A|B) = \mathbb{P}(A)$.*

Teorema 1.26 (Regra da multiplicação)**.** *Sejam $A_1, A_2, \ldots A_n$ eventos. Então*

$$\mathbb{P}(\cap_{i=1}^{n} A_i) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1 \cap A_2)\ldots\mathbb{P}(A_n|\cap_{i=1}^{n-1} A_i)$$

Teorema 1.27 (Lei da Probabilidade Total)**.** *Seja $(A_n)_{n\in\mathbb{N}}$ uma partição de Ω e B um evento.*

$$\mathbb{P}(B) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)\mathbb{P}(B|A_n)$$

Lema 1.28. *Se $A_1, \ldots, A_n$ particiona Ω e B é um evento, $\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)$.*

Teorema 1.29 (Teorema de Bayes)**.** *Seja $(A_i)_{i\in\mathbb{N}}$ uma partição de Ω e B um evento. Para todo $n \in \mathbb{N}$,*

$$\mathbb{P}(A_n|B) = \frac{\mathbb{P}(A_n)\mathbb{P}(B|A_n)}{\sum_{i\in\mathbb{N}} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}$$

Lema 1.30. *Seja $A_1, \ldots, A_n$ uma partição de Ω e B um evento.*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_{i=1}^{n} \mathbb{P}(A_i)\mathbb{P}(B|A_i)}$$

1.2.2 Variáveis aleatórias

Definição 1.31. Uma variável aleatória X é uma função tal que $X : \Omega \to \mathbb{R}$.

Definição 1.32 (Função indicadora)**.** A função indicadora é um tipo especial de variável aleatória. Considere um evento $A \in \mathcal{F}$. A função indicadora de A é denotada por $I_A : \Omega \to \mathbb{R}$ e definida da seguinte forma:

$$I_A(w) = \begin{cases} 1 & \text{, se } w \in A \\ 0 & \text{, caso contrário} \end{cases}$$

Definição 1.33. Seja X uma variável aleatória discreta. Para $x \in \mathbb{R}$, define-se $p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{w \in \Omega : X(w) = x\})$. A função $p_X : \mathbb{R} \to [0, 1]$ é chamada de função de massa de probabilidade de X.

Lema 1.34. *Seja X uma variável aleatória discreta e p_X sua função de massa de probabilidade. Seja χ os possíveis valores de X.*

- *Para todo $x \in \chi$, $0 \leq p_X(x) \leq 1$.*
- $\sum_{x \in \chi} p_X(x) = 1$.

Definição 1.35. Seja X uma variável aleatória contínua. Denotamos a função densidade de probabilidade de X por $f_X : \mathbb{R} \to \mathbb{R}$. Esta função satisfaz as seguintes propriedades:

1. $f_X(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
3. $\int_a^b f_X(x)dx = \mathbb{P}(a \leq X \leq b)$.

Definição 1.36. A função de distribuição acumulada de X é uma função $F_X : \mathbb{R} \to \mathbb{R}$,

$$F(x) = \mathbb{P}(X \leq x)$$

Lema 1.37.

$$P(X = x) = F_X(x) - \lim_{y \to x^-} F_X(y)$$

Lema 1.38. *F_X satisfaz as seguintes propriedades:*

1. *F_X é não-decrescente.*
2. $\lim_{x \to -\infty} F_X(x) = 0$.
3. $\lim_{x \to \infty} F_X(x) = 1$.
4. *F_X é contínua à direita.*

Lema 1.39. *Seja X uma variável aleatória. Para todo $b \geq a$, $F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b)$.*

Lema 1.40. *Se X é uma variável aleatória contínua, então:*

$$\frac{\partial F_X(x)}{\partial x} = f_X(x)$$

Teorema 1.41. *Se X e Y tem densidade conjunta $f(x, y)$, então:*

$$f(y_0|x_0) = \frac{f(y_0)f(x_0|y_0)}{\int f(y)f(x_0|y)dy}$$

Definição 1.42. Dizemos que $X_1, \ldots, X_n$ são independentes se, para todo $A_1, \ldots, A_n \subset \mathbb{R}$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) := \mathbb{P}(\cap_{i=1}^n X_i \in A_i) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i)$$

Portanto, para todo $A_1 \ldots A_n$, $\{\omega \in \Omega : X_i(\omega) \in A_i\}$ são conjuntamente independentes.

Definição 1.43. Seja X uma variável aleatória discreta e A um evento. O valor esperado de X dado A é denotado por $\mathbb{E}[X|A]$ e

$$\mathbb{E}[X|A] = \sum_{w \in \Omega} X(w)\mathbb{P}(\{w\}|A)$$

A esperança condicional de uma variável contínua pode ser definida similarmente.

Definição 1.44. Seja X uma variável aleatória discreta. O valor esperado de X é denotado por $\mathbb{E}[X]$ e é igual a $\mathbb{E}[X|\Omega]$. Isto é,

$$\mathbb{E}[X] = \sum_{w \in \Omega} X(w)\mathbb{P}(\{w\})$$

Caso X seja uma variável aleatória contínua,

$$\mathbb{E}[X] = \int_{\Omega} x f_X(x) dx$$

Lema 1.45 (Lei do estatístico inconsciente)**.** *Seja X uma variável aleatória discreta e que assume valores em χ:*

$$\mathbb{E}[f(X)] = \sum_{x \in \chi} f(x) \cdot p_X(x)$$

Lema 1.46. *Seja X uma variável aleatória discreta tal que $X \in \mathbb{N}$,*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i)$$

Lema 1.47 (Linearidade da esperança)**.**

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i \middle| A\right] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i|A]$$

Lema 1.48 (Lei da esperança total)**.** *Seja $A_1, \ldots, A_n$ uma partição de Ω e X uma variável aleatória,*

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i] \cdot \mathbb{P}(A_i)$$

Definição 1.49 (Variância)**.** A variância de uma variável aleatória X é dada por $\mathbb{E}[(X - \mathbb{E}[X])^2]$ e denotada por $Var[X]$.

Lema 1.50.

$$Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Lema 1.51. *Se X e Y são independentes, então $E[XY] = E[X]E[Y]$.*

Lema 1.52. *Se X e Y são independentes, então $Var[X + Y] = Var[X] + Var[Y]$.*

Lema 1.53. *$Var[X] = 0$ se e somente se X é uma constante (existe uma constante $c \in \mathbb{R}$ tal que $\mathbb{P}(X = c) = 1$).*

Definição 1.54 (Covariância)**.** Sejam X e Y duas variáveis aleatórias.

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Lema 1.55. $Cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Lema 1.56 (Propriedades da Covariância)**.** *Se X e Y são variáveis aleatórias, então*

a. $Cov[X,X] = Var[X] \geq 0$. *Portanto,* $Cov[X,X] = 0$ *apenas se X é uma constante.*

b. $Cov[X,Y] = Cov[Y,X]$.

c. $Cov[aX + bY, Z] = aCov[X,Z] + bCov[Y,Z]$.

Lema 1.57 (Cauchy-Schwarz para variáveis aleatórias)**.**

$$|Cov[X,Y]| \leq \sqrt{Var[X]}\sqrt{Var[Y]}.$$

A igualdade ocorre se e somente se existem $a, b \in \mathbb{R}$ tais que $Y = aX + b$.

Lema 1.58 (Teorema de Pitágoras para variáveis aleatórias)**.**

$$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$$

Portanto, se $Cov[X,Y] = 0$, $Var[X+Y] = Var[X] + Var[Y]$.

Definição 1.59 (Correlação)**.**

$$Corr[X,Y] = \frac{Cov[X,Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$$

1.2.3 Distribuições importantes

Variável aleatória (Y)	massa: $p_Y(y) = P(Y = y)$	$E[Y]$	$Var[Y]$
Binomial(n,p)	$\binom{n}{y}p^y(1-p)^{n-y}$ $y \in \{0,1,2,\ldots,n\}$	np	$np(1-p)$
Hipergeométrica(N,n,k)	$\frac{\binom{k}{y}\binom{N-k}{n-y}}{\binom{N}{n}}$ $\max(0, n-N+k) \leq y \leq \min(n,k)$	$n \cdot \frac{k}{N}$	$n \cdot \frac{k}{N} \cdot (1 - \frac{k}{N}) \cdot \frac{N-n}{N-1}$
Geométrica(p)	$p(1-p)^{y-1}$ $y \in \{1,2,3,\ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial Negativa(r,p)	$\binom{y-1}{r-1}p^r(1-p)^{y-r}$ $y \in \{r, r+1, r+2, \ldots\}$	$r \cdot \frac{1}{p}$	$r \cdot \frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^y}{y!}$ $y \in \mathbb{N}$	λ	λ

Tabela 1: Distribuições discretas.

2 Aprendizado e atualização de incertezas

2.1 O Teorema de Bayes

A probabilidade é a medida usada por você para quantificar sua incerteza sobre proposições. Em outras palavras, se A indica a proposição "Choverá no dia 10/05", então $\mathbb{P}(A)$ indica o quanto você acredita nesta afirmação. Em

Variável aleatória (Y)	densidade: $f_Y(y)$	$\mathbb{E}[Y]$	$Var[Y]$
Uniforme(a,b)	$\frac{1}{b-a}$ $y \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial(λ)	$\lambda e^{-\lambda y}$ $y \in \mathbb{R}^+$	λ^{-1}	λ^{-2}
Gamma(k,λ)	$\frac{\lambda^k}{\Gamma(k)} y^{k-1} e^{-y\lambda}$ $y \in \mathbb{R}^+$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$
Beta(α,β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $y \in (0,1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Normal(μ,σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ $y \in \mathbb{R}$	μ	σ^2

Tabela 2: Distribuições contínuas.

particular, se $\mathbb{P}(A) = 1$, então você tem certeza que choverá no dia 10/05. Similarmente, se $\mathbb{P}(A) = 0$, então você tem certeza que não choverá neste dia.

Como a probabilidade reflete as suas incertezas, para obtê-la, você deve analisar o quanto acredita em diferentes proposições. Contudo, nem todas as proposições são igualmente fáceis de avaliar. Em particular, considere a situação em que você considera várias hipóteses para explicar como um conjunto de dados foi gerado. Nesta situação, você está comumente interessada em determinar qual hipótese é a mais razoável dado que você observou certos dados. Em particular, seja "Hipótese" uma hipótese e "Dados" os dados observados. Você deseja determinar a probabilidade da hipótese dado os dados que foram observados, isto é,

$$\mathbb{P}(\text{Hipótese}|\text{Dados})$$

Em outras palavras, você deseja utilizar os dados para aprender sobre as hipóteses consideradas. Por exemplo, considere que pacientes submetidos a um tratamento recuperam-se de uma doença com uma frequência desconhecida. Se você observar que um paciente submetido ao tratamento recuperou-se da doença, de que forma sua incerteza sobre a frequência de recuperação seria atualizada? Os teoremas da Probabilidade fornecem uma ferramenta para guiar o seu aprendizado. Lembre-se que:

$$\mathbb{P}(\text{Hipótese}|\text{Dados}) = \frac{\mathbb{P}(\text{Hipótese})\mathbb{P}(\text{Dados}|\text{Hipótese})}{\mathbb{P}(\text{Dados})}$$

$$\mathbb{P}(\text{Dados}) = \int_{\text{Hipótese}^*} \mathbb{P}(\text{Hipótese}^*)\mathbb{P}(\text{Dados}|\text{Hipótese}^*)$$

Em outras palavras, se você conhece a probabilidade das hipóteses antes de observar os dados, $\mathbb{P}$(Hipótese), e a probabilidade de obter os dados considerando cada uma das hipóteses, $\mathbb{P}$(Dados|Hipótese), então o aprendizado sobre a hipótese a partir dos dados, $\mathbb{P}$(Hipótese|Dados), está unicamente determinado. A seguir, ilustramos este procedimento com alguns exemplos:

Exemplo 2.1. Antes de realizar um exame, um médico acredita que há probabilidade de 0.05 de que seu paciente tenha câncer. O médico realiza o exame e ele indica que o paciente tem câncer. A probabilidade do exame indicar que o paciente tem câncer quando ele o tem é 0.8. A probabilidade do exame indicar que o paciente tem câncer quando ele não o tem é 0.1. Qual é a probabilidade de que o paciente tenha câncer dado que o exame indica que ele tem câncer?

Considere os seguintes eventos:

- C: o paciente tem câncer.
- E: o exame indica que o paciente tem câncer.

Note que C é a hipótese levantada pelo médico e E são os dados obtidos por ele. Desejamos determinar $\mathbb{P}(C|E)$. Para tal, podemos utilizar o Teorema de Bayes:

$$\begin{aligned}\mathbb{P}(C|E) &= \frac{\mathbb{P}(C)\mathbb{P}(E|C)}{\mathbb{P}(C)\mathbb{P}(E|C) + \mathbb{P}(C^c)\mathbb{P}(E|C^c)} && \textit{Teorema 1.29}\\ &= \frac{0.05 \cdot 0.8}{0.05 \cdot 0.8 + 0.95 \cdot 0.1}\\ &= \frac{8}{27}\end{aligned}$$

Note que, mesmo após o exame indicar que o paciente tem câncer, a probabilidade de que o paciente de fato o tenha ainda é inferior a 50%. Em outras palavras, o resultado do exame aumenta a probabilidade da hipótese de câncer, mas o aumento não é tão grande quanto poderia se esperar. Isso ocorre pois o médico acreditava, a princípio, que a probabilidade do paciente ter câncer era baixa. Também o exame pode ter falsos positivos e falsos negativos. Como resposta ao resultado observado, o médico poderá receitar novos exames e assim reduzir a sua incerteza sobre o diagnóstico.

Exemplo 2.2. Dois candidatos participam do segundo turno de uma eleição presidencial, A e D. Um analista está interessado em inferir a respeito da proporção, θ, de eleitores que votarão em D. A priori, o analista acredita que os dois candidatos estão empatados e atribui $\theta \sim \text{Beta}(5,5)$. O analista amostra 30 indivíduos da população e anota o número, X, destes que declararam o voto em D. O resultado obtido foi que 20 indivíduos declaram que votarão em D e 10 que votarão em A. O analista acredita que, caso soubesse o valor de θ, então $X \sim \text{Binomial}(30,\theta)$. Qual a incerteza do analista sobre θ após observar a amostra?

$$\begin{aligned}f(\theta_0|x) &= \frac{f(\theta_0)\cdot f(x|\theta_0)}{\int_0^1 f(\theta)\cdot f(x|\theta)d\theta} && \textit{Teorema 1.41}\\ &= \frac{\beta^{-1}(5,5)\theta_0^4(1-\theta_0)^4 \cdot \binom{30}{20}\theta_0^{20}(1-\theta_0)^{10}\mathbb{I}(\theta_0)_{(0,1)}}{\int_{[0,1]}\beta^{-1}(5,5)\theta^4(1-\theta)^4\cdot\binom{30}{20}\theta^{20}(1-\theta)^{10}d\theta}\\ &= \frac{\theta_0^{24}(1-\theta_0)^{14}\mathbb{I}_{(0,1)}(\theta_0)}{\int_{[0,1]}\theta^{24}(1-\theta)^{14}d\theta}\\ &= \beta^{-1}(25,15)\theta_0^{25-1}(1-\theta_0)^{15-1}\mathbb{I}_{(0,1)}(\theta_0)\end{aligned}$$

Portanto, $\theta|X=20$, a distribuição *a posteriori* do analista para θ após observar $X=20$, é uma $\text{Beta}(25,15)$. A partir desta distribuição, pode-se calcular, por exemplo, a probabilidade do candidato D vencer a eleição. Esta é dada por $\mathbb{P}(\theta > 0.5|X=20) = \mathbb{P}(\text{Beta}(25,15) > 0.5) \approx 94,59\%$.

Por conveniência, podemos atribuir alguns nomes a conceitos estudados neste curso. Chamamos de distribuição *a priori* aquela que é atribuída a θ antes de observar a amostra. Chamamos de distribuição *a posteriori* aquela que é atribuída a θ após observar a amostra. Denotamos por $L_x(\theta_0)$ a função de verossimilhança de θ para o dado x e definimos $L_x(\theta_0) = f(x|\theta_0)$. Note que $L_x(\theta_0)$ é uma função de θ_0.

Podemos utilizar os conceitos definidos acima para estudar graficamente de que forma a distribuição a posteriori é obtida. A fig. 1 ilustra a relação entre os conceitos neste exemplo. Observe que a média da posteriori está entre

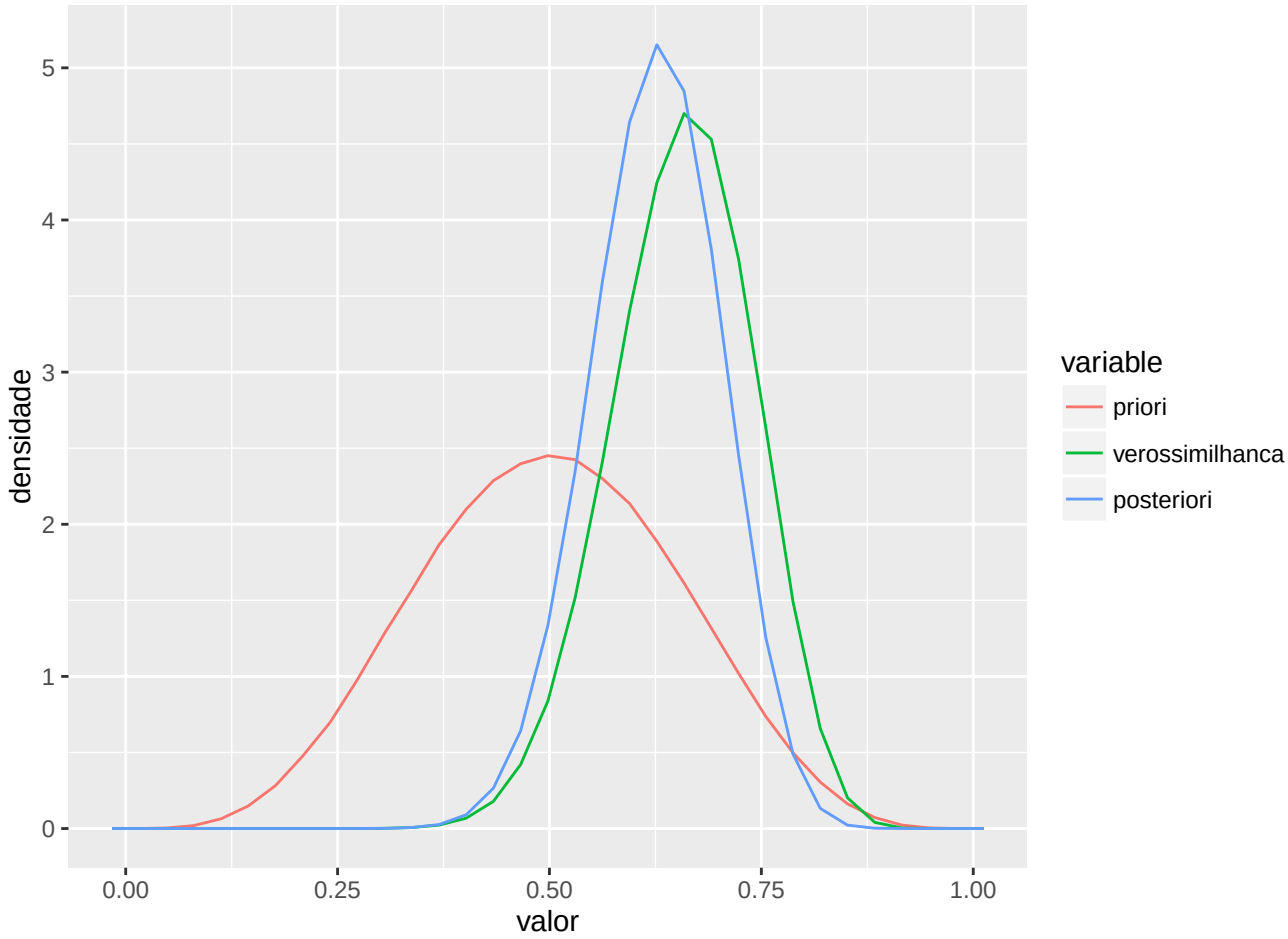

Figura 1: Comparação entre a distribuição a priori, a verossimilhança e a distribuição a posteriori no Exemplo 2.2.

a média da priori e da verossimilhança. Em outras palavras, a sua opinião sobre o parâmetro após observar os dados está entre a sua opinião anterior e a verossimilhança dos dados.

Exercícios

Exercício 2.3. No Exemplo 2.1, qual é a probabilidade de o exame indicar que não há câncer? (antes de saber o resultado do exame)

Exercício 2.4. Um homicídio foi cometido em uma cidade de aproximadamente 100.000 habitantes adultos. Antes de observar as evidências do caso, o juiz acreditava que qualquer um dos habitantes poderia ter cometido o crime com mesma probabilidade. Contudo, a promotoria traz um exame forense como evidência de que o réu cometeu o crime. Por um lado, caso o réu seja culpado, os resultados do exame seriam observados com certeza. Por outro lado, caso o réu fosse inocente, os resultados do exame somente seriam observados com uma probabilidade de 1 em 10.000. O promotor alega que há uma baixa probabilidade de que o réu seja inocente, e que ele deve portanto ser condenado. O juiz contrata você como uma perita judicial para ajudá-lo a entender o argumento do promotor. Qual é a sua análise?

Ressalvadas algumas adaptações, este argumento jurídico é real e foi apresentado nos Estados Unidos no caso *The People v. Collins*.

Exercício 2.5 (Monty Hall)**.** Marilyn vos Savant por muitos anos foi considerada a pessoa com o maior QI do mundo (228) pelo livro Guiness. Ela é a autora da coluna "Ask Marilyn", em que responde a perguntas de seus

leitores. Em um episódio célebre, sua resposta a uma pergunta de probabilidade gerou polêmica na época. A pergunta era a seguinte:

> Suponha que você está em um programa de televisão. O apresentador te permite escolher entre três portas. Atrás de uma das portas há uma carro e atrás das outras duas há cabras. Você ganhará o carro caso abra a sua respectiva porta. Considere que, a priori, você acredita ser equiprovável que o carro está atrás de cada uma das portas. Após você escolher a porta 1, o apresentador sempre abrirá aleatoriamente (dentre as portas que você não escolheu) uma das portas com uma cabra e te dará a oportunidade de trocar de porta. Digamos que o apresentador abriu a porta 2. É vantajoso trocar de porta?

Marylin vos Savant respondeu que era vantajoso. Você concorda com ela?

Exercício 2.6. Sejam $A_1, \ldots, A_n$ eventos disjuntos tais que $B = \cup_{i=1}^n A_i$. Prove que

$$\mathbb{P}(A_1|B) = \frac{\mathbb{P}(A_1)\mathbb{P}(B|A_1)}{\sum_{i=1}^n \mathbb{P}(A_i)\mathbb{P}(B|A_i)}$$

2.2 Modelo estatístico

O modelo estatístico oferece uma tradução padronizada de problemas envolvendo incerteza para um espaço de probabilidade. Ele é suficientemente geral para que muitos problemas que envolvem incerteza possam ser traduzidos por meio dele.

Simplificadamente, os elementos do modelo estatístico são os dados, denotados por X (que pertencem ao espaço amostral $\mathcal{X}$), e uma quantidade desconhecida de interesse, denotada por θ. θ é comumente chamado de parâmetro do modelo. Assumimos que $\theta \in \Theta$ (Θ é chamado de espaço paramétrico). Nosso interesse é em fazer inferência (isto é, fazer afirmações) sobre θ após observar os dados $X = x$.

A essência da abordagem Bayesiana para a inferência estatística é tratar tanto X quanto θ como quantidades aleatórias. Isso é feito pois, antes de observar X, ambas são desconhecidas e portanto temos incerteza sobre elas.

Veremos agora como essa perspectiva permite fazer inferência sobre θ. Como θ e X são aleatórios, eles possuem uma função de probabilidade conjunta. Na prática Bayesiana, esta probabilidade conjunta é especificada por dois elementos: (i) a probabilidade marginal de θ, denotada por $f(\theta_0)$ (a distribuição *a priori* dos parâmetros, que codifica matematicamente informações anteriores à realização do experimento) e (ii) a probabilidade dos dados condicional a θ, denotada por $f(x|\theta_0)$ (a função de verossimilhança, que codifica a informação sobre θ presente nos dados).[1] Estes dois elementos podem ser combinados através do Teorema de Bayes de modo a se obter $f(\theta_0|x)$, que é chamada de distribuição *a posteriori* de θ e que codifica a incerteza sobre θ após a realização do experimento:

$$f(\theta_0|x) = \frac{f(\theta_0)f(x|\theta_0)}{\int_\Theta f(\theta)f(x|\theta)d\theta}. \tag{1}$$

Assim, o Teorema de Bayes é a ferramenta matemática que permite combinar informações anteriores à realização do experimento com os dados obtidos, de modo a se atualizar a incerteza sobre θ após se observar x. A distribuição *a posteriori* $f(\theta_0|x)$ contém toda a incerteza que temos sobre θ após a realização do experimento.

[1] Note que o modelo Bayesiano está especificando uma função de probabilidade (ou densidade) conjunta para θ e X, $f(x, \theta_0)$. Em particular, $f(\theta_0) := \int f(x, \theta_0)dx$. Ainda que essa distribuição é usualmente especificada como *priori*+verossimilhança, ela poderia ser feita de outras maneiras.

Observe que $f(\theta_0|x)$ é uma função de θ_0. Também, $\int_\Theta f(\theta)f(x|\theta)d\theta$ não depende de θ_0 ou, em outras palavras, esta integral é a constante que faz com que $f(\theta_0|x)$ integre 1. Em alguns casos, é possível determinar a distribuição de $\theta|X$ sem calcular diretamente o valor da constante $\int_\Theta f(\theta)f(x|\theta)d\theta$. Nestes casos, a seguinte notação é útil,

Definição 2.7. Dizemos que $f(y) \propto g(y)$ se existe alguma constante, $C \in \mathbb{R}$, tal que $f(y) = Cg(y)$.

Lema 2.8.

$$f(\theta_0|x) \propto f(\theta_0)f(x|\theta_0)$$

Note que $f(\theta_0|x)$ é uma função de θ_0 e que x é uma constante. Isto ocorre pois, no paradigma Bayesiano, os dados, x, são conhecidos e a incerteza existe a respeito de θ. Por isso, $f(x) = \int f(\theta)f(x|\theta)d\theta$ é uma constante. O "θ" que aparece na integral é uma variável de integração.

Exemplo 2.9 ("é proporcional a" no Exemplo 2.2)**.**

$$\begin{aligned} f(\theta_0|x) &\propto f(\theta_0)f(x|\theta_0) && \text{Lema 2.8} \\ &= \beta^{-1}(5,5)\theta_0^4(1-\theta_0)^4 \cdot \binom{30}{20}\theta_0^{20}(1-\theta_0)^{10}\mathbb{I}(\theta_0)_{(0,1)} \\ &\propto \theta_0^{24}(1-\theta_0)^{14}I(\theta_0)_{(0,1)} \end{aligned}$$

Observe que $\theta_0^{24}(1-\theta_0)^{14}I(\theta_0)_{(0,1)}$ é a forma funcional de uma distribuição Beta$(25, 15)$. Portanto, a única constante que faz com que essa expressão integre 1 é $\beta^{-1}(25, 15)$.

É comum que os dados sejam um vetor de valores observados, $X = (X_1, \ldots, X_n)$. Ademais, é comum supor-se que os elementos deste vetor sejam independentes e identicamente distribuídos quando θ é conhecido. Neste caso, podemos escrever

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

onde $f(x_i|\theta)$ é a distribuição marginal de cada elemento do vetor x. Neste caso, obtemos,

Lema 2.10. *Quando os dados são i.i.d. dado θ, obtemos:*

$$\begin{aligned} f(\theta_0|x) &= \frac{f(\theta_0)\prod_{i=1}^n f(x_i|\theta)}{\int_\Theta f(\theta)\prod_{i=1}^n f(x_i|\theta)d\theta} \\ &\propto f(\theta_0)\prod_{i=1}^{n} f(x_i|\theta) \end{aligned}$$

Exercícios

Exercício 2.11. Considere que, a priori, $\theta \sim$ Gamma$(2, 2)$. Também, $X|\theta \sim$ Poisson(θ). Qual é a distribuição a posteriori para θ após observar que $X = 2$? Use um software computacional para traçar (i) a distribuição à priori, (ii) a função de verossimilhança e (iii) a distribuição à posteriori em função de $\theta \in \mathbb{R}^+$. Interprete o gráfico resultante.

Exercício 2.12. Considere que uma urna tem 3 bolas. Cada uma das bolas pode ser branca ou azul. Você deseja determinar o número de bolas azuis na urna. A priori, você acredita que todos os valores em $\{0, 1, 2, 3\}$ são

equiprováveis para o número de bolas azuis. Para aprender mais sobre esta quantidade, você retira duas bolas com reposição da urna e observa que as duas são azuis. Qual é o modelo estatístico neste problema? Use o Lema 2.10 para obter sua probabilidade a posteriori.

Exercício 2.13. Considere a mesma situação descrita no Exercício 2.12. Contudo, considere que a retirada da urna foi feita **sem** reposição. Encontre a sua probabilidade a posteriori. (Você não pode usar o Lema 2.10 neste caso!)

Exercício 2.14. Considere que, a priori, a proporção de indivíduos com uma determinada doença em uma população é uma uniforme em $(0, 1)$. 100 indivíduos são selecionados com reposição e verifica-se que 30 deles tem a doença. Qual é a probabilidade a posteriori para a proporção de indivíduos com a doença?

Exercício 2.15. Considere que, a priori, σ^2 tem distribuição uniforme em $\{1, 2\}$. Se $X|\sigma^2 \sim N(0, \sigma^2)$, qual a posteriori de σ^2 dado X? Exiba a posteriori exata, não é o suficiente indicá-la até uma constante de proporcionalidade. Use um software computacional para traçar o valor de $\mathbb{P}(\sigma^2 = 1|X = x)$ em função de $x \in \mathbb{R}$. Interprete o gráfico resultante.

Exercício 2.16. O som inicial para uma palavra em um idioma pode pertencer à categoria consoante ou à categoria vogal. Considere que, se o som inicial de um idioma é conhecido, cada um dos seus descendentes terá o som inicial na mesma categoria do ancestral independentemente e com probabilidade 0.9. Considere que os idiomas B e C são descendentes imediatos do idioma A. Contudo, não conhecemos os sons iniciais para as palavras de A. Acreditamos que o som inicial para o sentido "cachorro" em A é uma consoante ou uma vogal com mesma probabilidade. Observamos que os sons iniciais dos idiomas B e C para "cachorro" são ambos consoantes. Estamos interessados em predizer a categoria do som inicial para a palavra "cachorro" no idioma A.

(a) Quais são os dados e os parâmetros neste problema?

(b) Encontre a posteriori para o parâmetro.

(c) Qual seria a posteriori se o som inicial para cada descendente estivesse em uma categoria diferente?

Exercício 2.17. Considere que um sistema é composto por 3 peças e que o sistema falha quando qualquer uma das peças falhar. Considere que cada peça falha independentemente em um tempo (em dias) determinado pela mesma taxa de falha. Assuma que, dada a taxa de falha, θ, o tempo para falha de uma peça é uma Exponencial(θ).

(a) Qual é a taxa de falha do sistema?

(b) Considere que, a priori, você acreditava que a taxa de falha de cada componente seguia uma distribuição Gamma$(1, 1)$. Se você observou um sistema falhar em 3 dias, qual é a sua posteriori para a taxa de falha dos componentes? Identifique o nome e os hiperparâmetros da distribuição a posteriori.

Exercício 2.18. Dado θ, $X_1, \ldots X_n$ são independentes. Mostre que

$$f(\theta|x_1, \ldots, x_n) \propto f(\theta|x_1, \ldots, x_{n-1}) f(x_n|\theta)$$

Em outras palavras, a posteriori obtida após condicionar nas observações anteriores pode ser usada como priori na aplicação do Teorema de Bayes para a próxima observação.

	$\mathbb{I}_A - \Pr(A)$
A ocorre	$1 - \Pr(A)$
A não ocorre	$-\Pr(A)$

Tabela 3: Ganho obtido comprando uma aposta em A.

3 Coerência e Probabilidade: como evitar prejuízos certos?

3.1 Probabilidade marginal

Em cursos anteriores, você usou a probabilidade como uma forma de representar a sua incerteza. Contudo, possivelmente você não viu o porquê dessa representação ser razoável. Por que é razoável assumir que a representação da sua incerteza satisfaz os axiomas da probabilidade (Definição 1.17)?

1. (Não-negatividade) Para todo A, $\mathbb{P}(A) \geq 0$.
2. (Aditividade) Se $(A_n)_{n\in\mathbb{N}}$ é uma sequência de conjuntos disjuntos, então $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$.
3. (Normalização) $\mathbb{P}(\Omega) = 1$.

Neste Capítulo, discutiremos a questão acima. Para tal, faremos uso de um experimento mental. Considere que A é uma proposição sobre a qual você tem incerteza. Por exemplo, A pode designar o evento de que choverá amanhã. Também considere uma aposta tal que o vendedor deve pagar $R\$1$ ao comprador se A ocorrer e $R\$0$, caso contrário. Qual seria o preço tal que você aceitaria tanto comprar quanto vender esta aposta? Chamaremos este preço de $\Pr(A)$.

Note que Pr é uma representação da sua incerteza. Quanto maior o valor de $\Pr(A)$, mais você acredita que A irá ocorrer. Também, o seu valor de Pr não necessariamente é o mesmo de seus colegas, cada um podendo coerentemente ter os seus próprios preços. Isto significa que você pode coerentemente designar qualquer valor a $\Pr(A)$?

Para responder a esta pergunta, é necessário definir coerência. Dizemos que uma designação de preços é incoerente se ela é tal que você pode perder dinheiro com certeza. Observar que você está incorrendo em perda certa provavelmente te daria incentivo suficiente para rever seus preços e a incerteza das proposições ligadas a estes. Portanto, investigaremos sob quais condições você incorre em perda certa.

Para tal, faremos uso da Definição 1.32. Lembre-se que $\mathbb{I}_A$ é a função indicadora da proposição A, uma variável aleatória que assume o valor 1, quando A ocorre, e 0, caso contrário. Assim, o seu lucro após comprar a aposta em A pode ser escrito como $\mathbb{I}_A - \Pr(A)$. O resultado desta variável aleatória é resumido na Tabela 3. Note que, se $\Pr(A) > 1$, então o seu ganho obtido é sempre menor do que 0. Em outras palavras, você está tendo prejuízo certo. Assim, para que $\Pr(A)$ seja uma representação coerente da sua incerteza, é necessário que $\Pr(A) \leq 1$.

Similarmente, podemos considerar o caso em que você vende a aposta em A. Neste caso, o seu lucro é dado por $-(I_A - \Pr(A))$. Esta variável é resumida na tabela 4. Observe que, se $\Pr(A) < 0$, então $-(\mathbb{I}_A - \Pr(A))$ é sempre menor que 0. Novamente, este caso te levaria a um prejuízo certo. Portanto, também podemos concluir que, se Pr representa coerentemente a sua incerteza, então $\Pr(A) \geq 0$. As considerações apresentadas acima nos levam ao resultado no Lema 3.1.

Lema 3.1. *Se existe A tal que $Pr(A) < 0$ ou $Pr(A) > 1$, então você pode ser levada a prejuízo certo.*

Para prosseguir na análise, consideraremos a composição de apostas. Em outras palavras, avaliaremos o resultado de você compor apostas, comprando e vendendo apostas em eventos individuais. Formalmente, para

	$-(\mathbb{I}_A - \Pr(A))$
A ocorre	$\Pr(A) - 1$
A não ocorre	$\Pr(A)$

Tabela 4: Ganho obtido vendendo uma aposta em A.

quaisquer números, $\alpha_1, \ldots, \alpha_n$, e eventos, $A_1, \ldots, A_n$, denotamos um portfólio de apostas por $(\alpha_i, A_i)_{1 \leq i \leq n}$. Este portfólio é tal que, se $\alpha_i > 0$, então você comprará α_i unidades da aposta em A_i e, se $\alpha_i < 0$, então você venderá α_i unidades desta aposta. Note que o seu balanço dado pelo portfólio $(\alpha_i, A_i)_{1 \leq i \leq n}$ é $\sum_{i=1}^n \alpha_i(\mathbb{I}_{A_i} - \Pr(A_i))$. Em outras palavras, $\sum_{i=1}^n \alpha_i(\mathbb{I}_{A_i} - \Pr(A_i))$ é a soma do resultado de todas as apostas em que você participou.

Definição 3.2. Você pode ser levada a prejuízo certo se existe um portfólio de apostas, $(\alpha_i, A_i)_{1 \leq i \leq n}$, tal que você perde dinheiro com certeza. Isto é, para todo $\omega \in \Omega$, $\sum_{i=1}^n \alpha_i(\mathbb{I}_{A_i}(w) - \Pr(A_i)) < 0$.

O próximo lema será uma ferramenta importante para provar que certas atribuições de preços levam a prejuízo certo. Ele mostra que, como você está disposta tanto a comprar quanto a vender quaisquer apostas, então um portfólio com balanço constante e diferente de 0 a levará a prejuízo certo.

Lema 3.3. *Considere que para todo $\omega \in \Omega$, $\sum_{i=1}^n \alpha_i(\mathbb{I}_{A_i(w)} - Pr(A_i))$ assume um valor constante, c. Se $c \neq 0$, então você pode ser levada a prejuízo certo.*

Demonstração. Se $c < 0$, então o portfólio $(\alpha_i, A_i)_{1 \leq i \leq n}$ traz um balanço c e a leva a prejuízo certo. Se $c > 0$, então note que o balanço do portfólio $(-\alpha_i, A_i)_{1 \leq i \leq n}$ é:

$$\begin{aligned} &\sum_{i=1}^n -\alpha_i(\mathbb{I}_{A_i(w)} - \Pr(A_i)) \\ =& -\sum_{i=1}^n \alpha_i(\mathbb{I}_{A_i(w)} - \Pr(A_i)) = -c \end{aligned}$$

Portanto, existe um portfólio, $(-\alpha_i, A_i)_{1 \leq i \leq n}$, que a leva a prejuízo certo. □

O Lema 3.3 nos permite provar mais resultados a respeito da atribuição coerente de preços.

Lema 3.4. *Se $Pr(\Omega) \neq 1$, então você pode ser levada a prejuízo certo.*

Demonstração. Note que o balanço da compra de uma unidade de Ω, $((1, \Omega))$, é dado por $\mathbb{I}_\Omega - \Pr(\Omega)$. Também, $\mathbb{I}_\Omega$ é outra forma de escrever 1. Assim, o balanço de $((1, \Omega))$ é $1 - \Pr(\Omega)$, uma constante. Portanto, o Lema 3.3 garante que, se $1 - \Pr(\Omega) \neq 0$, então seus preços a levam a prejuízo certo. Em outras palavras, para evitar prejuízo certo é necessário que $\Pr(\Omega) = 1$. □

Lema 3.5. *Se existem A e B disjuntos tais que*

$$Pr(A \cup B) \neq Pr(A) + Pr(B)$$

então você pode ser levada a prejuízo certo.

Demonstração. Como A e B são disjuntos, existem três possíveis ocorrências: A, B ou $(A \cup B)^c$ (como A e B são disjuntos, $A \cap B$ é impossível). Portanto, o balanço de comprar A, comprar B e vender $A \cup B$ é resumido pela tabela 5. Note que o balanço do portfólio $((1, A), (1, B), (-1, A \cup B))$ é constante e igual a $\Pr(A \cup B) - \Pr(A) - \Pr(B)$. Portanto, pelo Lema 3.3, se $\Pr(A \cup B) - \Pr(A) - \Pr(B) \neq 0$, você pode ser levada a prejuízo certo. □

	$\mathbb{I}_A(w) - \Pr(A)$	$\mathbb{I}_B(w) - \Pr(B)$	$-(\mathbb{I}_{(A\cup B)} - \Pr(A \cup B))$	portfólio (soma)
$\omega \in A$	$1 - \Pr(A)$	$-\Pr(B)$	$\Pr(A \cup B) - 1$	$\Pr(A \cup B) - \Pr(A) - \Pr(B)$
$\omega \in B$	$-\Pr(A)$	$1 - \Pr(B)$	$\Pr(A \cup B) - 1$	$\Pr(A \cup B) - \Pr(A) - \Pr(B)$
$\omega \in (A \cup B)^c$	$-\Pr(A)$	$-\Pr(B)$	$\Pr(A \cup B)$	$\Pr(A \cup B) - \Pr(A) - \Pr(B)$

Tabela 5: Ganho obtido por um portfólio que compra A, compra B e vende $A \cup B$, quando A e B são disjuntos.

Resumindo os Lemas 3.1, 3.4 e 3.5, obtemos condições que Pr necessariamente deve obedecer para que você não possa ser levada a prejuízo certo.

Lema 3.6. *Se Pr é tal que você não pode ser levada a prejuízo certo, então é necessário que:*

1. *(Não-negatividade) Para todo A, $0 \leq Pr(A) \leq 1$.*
2. *(Aditividade) Se A e B são disjuntos, então $Pr(A \cup B) = Pr(A) + Pr(B)$.*
3. *(Normalização) $Pr(\Omega) = 1$.*

Demonstração. Decorre dos Lemas 3.1, 3.4 e 3.5. □

Exercícios

Exercício 3.7 (Kadane (2011; p.8))**.** Considere os seguintes eventos:

- A_1: Choverá amanhã e a temperatura máxima será superior a 25^o C.
- A_2: Choverá amanhã e a temperatura máxima não será superior a 25^o C.
- A_3: Não choverá amanhã e a temperatura máxima será superior a 25^o C.
- A_4: Não choverá amanhã e a temperatura máxima não será superior a 25^o C.

Quais seriam os preços que você definiria para cada proposição e qual a sua linha de raciocínio? Os seus preços são coerentes? Se não, você estaria disposta a revê-los?

Exercício 3.8. Mostre que se $\Pr(\emptyset) \neq 0$, então você pode ser levada a prejuízo certo.

Exercício 3.9. Se o prêmio das apostas fosse modificado de $R\$1,00$ para $R\$2,00$, alguma das condições para a coerência de preços seria modificada? De que forma a sua resposta afeta a sua interpretação do quanto é razoável a analogia de apostas?

Exercício 3.10. Mostre que, se $Pr(A \cup B) \neq Pr(A) + Pr(B) - Pr(A \cap B)$, então você pode ser levada a prejuízo certo.

3.2 Probabilidade condicional

Na seção passada, discutimos uma analogia entre os axiomas da probabilidade e apostas. Contudo, esta analogia não nos permite obter o axioma da probabilidade condicional. Em outras palavras, o sistema de apostas que estudamos não permite estudar de que forma a incerteza é alterada com o aprendizado de novos fatos. Nesta seção, discutiremos uma extensão da analogia de apostas que permite obter o axioma da probabilidade condicional na Definição 1.23.

	$\mathbb{I}_{A\cap B} - \Pr(A\|B)\mathbb{I}_B$
A ocorre e B ocorre	$1 - \Pr(A\|B)$
A não ocorre e B ocorre	$-\Pr(A\|B)$
B não ocorre	0

Tabela 6: Balanço da compra da aposta de A condicional a B.

	$(1, A, B)$	$(1, A^c \cap B, \Omega)$	$(1 - \Pr(A\|B), B^c, \Omega)$
A ocorre e B ocorre	$1 - \Pr(A\|B)$	$-\Pr(A^c \cap B)$	$-(1 - \Pr(A\|B))\Pr(B^c)$
A não ocorre e B ocorre	$-\Pr(A\|B)$	$1 - \Pr(A^c \cap B)$	$-(1 - \Pr(A\|B))\Pr(B^c)$
B não ocorre	0	$-\Pr(A^c \cap B)$	$(1 - \Pr(A\|B))(1 - \Pr(B^c))$

Tabela 7: Balanço da compra de diversas apostas condicionais.

1. $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B)$

Que tipo de aposta seria avaliada de acordo com a sua incerteza em A dado que B ocorreu? Para responder a essa pergunta, é importante considerar qual o resultado da aposta quando B não acontece. Uma saída é dizer que o comprador (e o vendedor) tem um balanço de 0 quando B não acontece. Assim, a aposta somente tem efeito sobre os apostadores quando B ocorre. Finalmente, para completar a aposta, quando B ocorre, consideramos uma aposta em que o comprador ganha $R\$1$ quando A acontece e $R\$0$ quando A não acontece. Dizemos que a aposta definida acima é de A condicional a B e o seu preço para comprá-la é $\Pr(A|B)$. Observe que a aposta de A condicional a B é similar à aposta da seção passada, com a exceção de que ela só surte efeitos quando B ocorre. O balanço de comprar esta aposta pode ser resumido por $\mathbb{I}_{A\cap B} - \Pr(A|B)\mathbb{I}_B$. A tabela 6 ilustra este balanço.

Para obter um resultado a partir da aposta condicional, também generalizaremos a definição anterior de portfólio de apostas. Um portfólio de apostas será definido como $(\alpha_i, A_i, B_i)_{1\leq i\leq n}$, onde α_i são números reais e A_i e B_i são eventos. Ao contrário da seção anterior, cada aposta em um portfólio é definida a partir de dois eventos. Isto ocorre pois estamos comprando apostas de A_i condicional a B_i. Também é permitido que α_i seja qualquer número real. A interpretação é similar ao caso anterior. Por exemplo, se $\alpha_i = 0.5$, compra-se 0.5 unidade de uma aposta de A_i condicional a B_i. Similarmente, se $\alpha_i = -0.37$, vende-se 0.37 de uma unidade de uma aposta de A_i condicional a B_i. O balanço de um portfólio é dado pela soma das apostas individuais, isto é, ele é igual a $\sum_{i=1}^{n} \alpha_i \left(\mathbb{I}_{A_i\cap B_i} - \Pr(A_i|B_i)\mathbb{I}_{B_i}\right)$.

Antes de prosseguirmos, note que uma aposta de A condicional a Ω tem exatamente o mesmo balanço que uma aposta em A. Este é o caso pois Ω sempre ocorre. Desta forma, para toda proposição A, definimos que $\Pr(A) := \Pr(A|\Omega)$. Com essa definição, poderemos mostrar que $\Pr(A \cap B) = \Pr(B)\Pr(A|B)$ é uma condição necessária para que você não possa ser levada a uma perda certa.

Para tal, considere um portfólio que consiste em comprar uma unidade de A condicional a B, comprar uma unidade de B^c (condicional a Ω) e comprar $(1 + \Pr(A|B))$ unidades de $A^c \cap B$ (condicional a Ω). A tabela 7 resume os balanços para cada uma destas apostas. Somando as colunas em cada uma das linhas, observamos que o balanço deste portfólio é constante e igual a

$$1 - \Pr(A|B) - \Pr(B^c) - \Pr(A^c \cap B) + \Pr(A|B)\Pr(B^c)$$

Utilizando o Lema 3.3, observe que, como o portfólio tem balanço constante, para que você possa evitar prejuízo

certo é necessário que

$$0 = 1 - \Pr(A|B) - \Pr(B^c) - \Pr(A^c \cap B) + \Pr(A|B)\Pr(B^c) \tag{2}$$

Também, decorre do Lema 3.6 que, para evitar prejuízo certo, é necessário que $\Pr(B^c) = 1 - \Pr(B)$ e $\Pr(A^c \cap B) = \Pr(B) - \Pr(A \cap B)$. Realizando estas substituições na eq. (2), obtemos:

$$\begin{aligned} 0 &= 1 - \Pr(A|B) - (1 - \Pr(B)) - (\Pr(B) - \Pr(A \cap B)) + \Pr(A|B)(1 - \Pr(B)) \\ 0 &= \Pr(A \cap B) - \Pr(B)\Pr(A|B) \\ \Pr(A \cap B) &= \Pr(B)\Pr(A|B) \end{aligned} \tag{3}$$

Assim, para que você não possa ser levada a prejuízo certo, é necessário que a eq. (3) seja satisfeita.

Lema 3.11. *Se Pr é tal que você não pode ser levada a prejuízo certo, então:*

$$Pr(A \cap B) = Pr(B)Pr(A|B)$$

Exercícios

Exercício 3.12 (DeGroot (1986)[p.63])**.** Se A e B são disjuntos e $\mathbb{P}(B) > 0$, qual o valor de $\mathbb{P}(A|B)$?

Exercício 3.13 ((DeGroot, 1986)(p.63))**.** Se A e B são independentes, $\mathbb{P}(A) = 0.3$ e $\mathbb{P}(B^c) > 0$, qual o valor de $\mathbb{P}(A^c|B^c)$?

Exercício 3.14. Suponha que $\Pr(B) = 0$, $\Pr(A \cap B) = 0$ e $\Pr(A|B) = 0.9$. Estes preços satisfazem a propriedade do Lema 3.11? Note que $\Pr(A|B)$ estaria indefinido se usássemos a fórmula $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$.

Exercício 3.15 ((Kadane, 2011)(p.33))**.** Exiba um exemplo tal que $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.

Exercício 3.16. Suponha que $\Pr(B) > 0$. Mostre que, se $\Pr(\emptyset|B) \neq 0$, então você pode ser levada a prejuízo certo.

Exercício 3.17 ((Kadane, 2011)(p.34))**.** Seja A um evento tal que $\mathbb{P}(A) > 0$. Prove que:

1. Para todo B, $0 \leq \mathbb{P}(B|A) \leq 1$.
2. $\mathbb{P}(\Omega|A) = 1$.
3. Se B e C são disjuntos, então $\mathbb{P}(B \cup C|A) = \mathbb{P}(B|A) + \mathbb{P}(C|A)$.

Em outras palavras, $P(\cdot|A)$ satisfaz todos os axiomas da probabilidade marginal descritos na Definição 1.17.

3.3 Caracterização da coerência

Note que as condições expressas no Lema 3.6 e no Lema 3.11 são necessárias para que se evite prejuízo certo. Também, estas condições são muito semelhantes àquelas expressas na Definição 1.17. Portanto, é razoável perguntar se estas condições também são suficientes para que você evite o prejuízo certo. A resposta para esta pergunta é dada a seguir.

Definição 3.18. Dizemos que um conjunto de subconjuntos de Ω, $\mathcal{A}$, é uma álgebra se:

1. $\Omega \in \mathcal{A}$.

2. Se $A \in \mathcal{A}$ e $B \in \mathcal{A}$, então $A \cup B \in \mathcal{A}$.

3. Se $A \in \mathcal{A}$, então $A^c \in \mathcal{A}$.

Lema 3.19. *Considere que os eventos aos quais você atribuiu preços formam uma álgebra. Se seus preços satisfazem as condições no Lema 3.6 e no Lema 3.11, então você não pode ser levada a prejuízo certo.*

Demonstração. Por hipótese, você atribuiu preços, Pr, sobre eventos que formam uma álgebra e seus preços satisfazem a condição no Lema 3.6. Portanto, podemos definir uma probabilidade, P, tal que para todos os eventos A e B sobre os quais você atribuiu preços, $\mathbb{P}(A|B) = \Pr(A|B)$. Fixe um portfólio arbitrário, $(\alpha_i, A_i, B_i)_{1\leq i\leq n}$. A esperança (segundo P) do balanço deste portfólio é:

$$\begin{aligned}
\mathbb{E}_P\left[\sum_{i=1}^{n} \alpha_i(\mathbb{I}_{A_i\cap B_i} - \Pr(A_i|B_i)\mathbb{I}_{B_i})\right] &= \sum_{i=1}^{n} \alpha_i(\mathbb{E}_P\left[\mathbb{I}_{A_i\cap B_i}\right] - \Pr(A_i|B_i)\mathbb{E}_P[\mathbb{I}_{B_i}]) && \text{linearidade}\\
&= \sum_{i=1}^{n} \alpha_i(\mathbb{P}(A_i \cap B_i) - \Pr(A_i|B_i)\mathbb{P}(B_i))\\
&= \sum_{i=1}^{n} \alpha_i(\Pr(A_i \cap B_i) - \Pr(A_i|B_i)\Pr(B_i)) && \Pr \equiv P\\
&= \sum_{i=1}^{n} \alpha_i(\Pr(A_i \cap B_i) - \Pr(A_i \cap B_i)) && \text{Lema 3.11}\\
&= 0
\end{aligned}$$

Portanto (Kadane, 2011; p.24), existe $\omega \in \Omega$, tal que $\sum_{i=1}^{n} \alpha_i \mathbb{I}_{B_i(\omega)}(\mathbb{I}_{A_i(\omega)} - \Pr(A_i|B_i)) \geq 0$ e $(\alpha_i, A_i, B_i)_{1\leq i\leq n}$ não leva a prejuízo certo. Como $(\alpha_i, A_i, B_i)_{1\leq i\leq n}$ era arbitrário, conclua que você não pode ser levada a prejuízo certo. □

Teorema 3.20 (de Finetti (1931))**.** *Considere que os eventos aos quais você atribuiu preços formam uma álgebra. Você não pode ser levada a prejuízo certo se e somente se:*

1. *(Não-negatividade) Para todo A, $0 \leq Pr(A) \leq 1$.*

2. *(Aditividade) Se A e B são disjuntos, então $Pr(A \cup B) = Pr(A) + Pr(B)$.*

3. *(Normalização) $Pr(\Omega) = 1$.*

4. *(Multiplicação) $Pr(A \cap B) = Pr(A)Pr(B|A)$.*

Demonstração. Decorre dos Lemas 3.6, 3.11 e 3.19. □

Com base na analogia entre apostas e representação de incerteza, o Teorema 3.20 nos leva a uma possível justificativa para os axiomas da probabilidade. Incidentalmente, esta analogia também pode ser usada para que você avalie quais são suas probabilidades (a partir dos preços que você aceitaria para cada aposta).

Exercícios

Exercício 3.21. O conjunto de todas as possíveis ocorrências em um sorteio é $\Omega = \{1, 2, 3, 4\}$. O Sr. Falido está vendendo apostas no evento $\{1, 2\}$ por R\$0.25, no evento $\{2, 3\}$ por R\$0.65 e no evento $\{4\}$ por R\$0.05. Se você compra uma aposta em um evento e ele ocorre o seu ganho é R\$1.

(a) Mostre que é possível levar o Sr. Falido a prejuízo certo.

(b) Existe uma probabilidade tal que $\mathbb{P}(\{1, 2\}) = 0.25$, $\mathbb{P}(\{2, 3\}) = 0.65$ e $\mathbb{P}(\{4\}) = 0.05$?

Exercício 3.22 (Desafio)**.** No Lema 3.19 usamos a condição de que os eventos sobre os quais você atribuiu preços formam uma álgebra. Exiba um exemplo em que esta condição não é satisfeita, que os preços, Pr, satisfazem todos os axiomas da probabilidade, e que você pode ser levada a prejuízo certo.

3.3.1 Infinitas apostas*

No Teorema 3.20, obtemos que, para evitar prejuízo certo, é necessário que para quaisquer A e B disjuntos, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. Esta condição é chamada de aditividade finita e é mais fraca do que a aditividade enumerável, descrita a seguir: se $A_1, \ldots, A_n, \ldots$ é uma sequência de eventos disjuntos, então $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. Note que, nos axiomas da probabilidade, temos a aditividade enumerável. Assim, o Teorema 3.20 obtém uma caracterização de coerência que exige propriedades mais fracas dos axiomas da probabilidade.

Neste sentido, uma possível linha de pesquisa consiste em investigar outras apostas que caracterizam os axiomas da probabilidade como condições necessárias e suficientes para a coerência. Neste sentido, pode-se estudar os efeitos de permitir um número enumerável de apostas (Stern and Kadane, 2015). Intuitivamente, se um número finito de apostas permite obter aditividade finita, então é razoável que um número enumerável de apostas permita obter aditividade enumerável.

Uma dificuldade inicial consiste em definir quais apostas são permitidas quando se considera um número enumerável de apostas. Por exemplo, tome um evento, A, tal que $\mathbb{P}(A) = 0.5$. Podemos considerar um portfólio de apostas que consiste em sucessivamente vender e comprar A, infinitamente. Assim, obteríamos o portfólio $((-1)^i, A_i)_{i\in\mathbb{N}}$. Se $\omega \in A$, então o balanço deste portfólio é $\sum_{i=1}^{\infty} (-0.5)^i$. Se $\omega \notin A$, então o balanço deste portfólio é $\sum_{i=1}^{\infty} (-0.5)^{i+1}$. Em ambos os casos, o balanço não converge. Assim, é razoável restringir a análise ao menos aos portfólios de aposta cujos balanços convergem.

Stern and Kadane (2015) considera várias possíveis restrições aos portfólios de apostas válidos. Nesta seção, replicaremos a análise quando consideramos apenas os potfolios de aposta, $(\alpha_i, A_{1,i}, A_{2,i})_{i\in\mathbb{N}}$, tais que o preço da aposta satisfaz $\sum_{i=1}^{\infty} |\alpha_i| \mathbb{P}(A_{1,i}|A_{2,i}) < \infty$ e $\sum_{i=1}^{\infty} \alpha_i \mathbb{I}_{A_{2,i}} (\mathbb{I}_{A_{1,i}} - \mathbb{P}(A_{1,i}|A_{2,i}))$ converge para todo $\omega \in \Omega$. Com base neste espaço de apostas, obteremos o seguinte teorema:

Teorema 3.23. *Considere que um preço, P, é definido sobre uma álgebra de eventos. P é coerente se e somente se ele satisfaz todos os axiomas da probabilidade.*

Demonstração. Decorre dos Lemas 3.24 e 3.25, a seguir. □

Lema 3.24. *Se um preço, $\mathbb{P}$, é coerente, então ele satisfaz todos os axiomas da probabilidade.*

Demonstração. Note que o espaço de apostas enumeráveis inclui o espaço de apostas finitas. Portanto, decorre dos Lemas 3.6 e 3.11 que, se $\mathbb{P}$ é coerente, então $\mathbb{P}$ é finitamente aditivo.

Portanto, basta provar que, se $\mathbb{P}$ é coerente, então $\mathbb{P}$ também é enumeravalmente aditivo. Considere eventos disjuntos arbitrários, $A_1, \ldots, A_n, \ldots$, Defina um portfólio, $(\beta_i, B_i, \Omega)_{i\in\mathbb{N}}$, tal que $\beta_1 = 1$ e $B_1 = \cup_{i=1}^{\infty} A_i$ e, para todo

$i > 1$, $\beta_i = -1$ e $B_i = A_{i-1}$. Em outras palavras, o portfólio compra $\cup_{i=1}^{\infty} A_i$ e vende todos os A_i separadamente. Mostraremos que este é um portfólio válido e, com ele, provaremos que P é enumeravelmente aditivo.

Primeiramente, mostraremos que o portfólio construído é válido. Como $\mathbb{P}$ é finitamente aditivo, para todo n, $\sum_{i=1}^{n} \mathbb{P}(A_i) \leq 1$. Portanto, $\sum_{i=1}^{\infty} \mathbb{P}(A_i) < 1$. Conclua que $\sum_{i=1}^{\infty} |\beta_i| \mathbb{P}(B_i) \leq 1 + \mathbb{P}(B_1) < \infty$. Também, para todo $\omega \in \Omega$, $\sum_{i=1}^{\infty} \beta_i \mathbb{I}_{B_i} = 0$. Portanto, para todo $\omega \in \Omega$, $\sum_{i=1}^{\infty} \beta_i (\mathbb{I}_{B_i}(w) - \mathbb{P}(B_i))$ converge.

A seguir, note que o balanço do portfólio considerado é $-\mathbb{P}(\cup_{i=1}^{\infty} A_i) + \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, uma constante. Assim, decorre do Lema 3.3 que, se $\mathbb{P}$ é coerente, então $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \mathbb{P}(\cup_{i=1}^{\infty} A_i)$. Como os A_i eram arbitrários, conclua que, se $\mathbb{P}$ é coerente, então $\mathbb{P}$ é enumeravelmente aditivo. □

Lema 3.25. *Considere que um preço, $\mathbb{P}$, é definido sobre uma álgebra de eventos. Se $\mathbb{P}$ satisfaz os axiomas da probabilidade, então ele é coerente.*

Demonstração. Como $\mathbb{P}$ satisfaz todos os axiomas da probabilidade e está definido sobre uma álgebra de eventos, então decorre do Teorema de Carathéodory (Billingsley, 1986) que $\mathbb{P}$ admite uma extensão para a sigma-álgebra gerada por esta álgebra. Defina esta extensão por $\mathbb{P}^*$.

Tome um portfólio de apostas válido arbitrário, $(\alpha_i, A_{1,i}, A_{2,i})_{i \in \mathbb{N}}$. Considere que $(\beta_i, B_{1,i}, B_{2,i})_{i \in \mathbb{N}}$ é a subsequência de $(\alpha_i, A_{1,i}, A_{2,i})_{i \in \mathbb{N}}$ tal que $\alpha_i > 0$. Similarmente, $(\gamma_i, C_{1,i}, C_{2,i})_{i \in \mathbb{N}}$ é a subsequência tal que $\alpha_i < 0$. Decorre do Teorema da Convergência Monotônica (Billingsley, 1986; p.211) que

$$\begin{aligned}
\sum_{i=1}^{\infty} \mathbb{E}_{P^*}\left[\beta_i \mathbb{I}_{B_{1,i} \cap B_{2,i}}\right] &= \sum_{i=1}^{\infty} \beta_i \mathbb{P}(B_{1,i} \cap B_{2,i}) \leq \sum_{i=1}^{\infty} |\alpha_i| \mathbb{P}(A_{1,i} | A_{2,i}) < \infty \\
\sum_{i=1}^{\infty} -\mathbb{E}_{P^*}\left[\gamma_i \mathbb{I}_{C_{1,i} \cap C_{2,i}}\right] &= \sum_{i=1}^{\infty} -\gamma_i \mathbb{P}(C_{1,i} \cap C_{2,i}) \leq \sum_{i=1}^{\infty} |\alpha_i| \mathbb{P}(A_{1,i} | A_{2,i}) < \infty \\
\mathbb{E}_{P^*}\left[\sum_{i=1}^{\infty} \beta_i \mathbb{I}_{B_{2,i}} \mathbb{P}(B_{1,i} | B_{2,i})\right] &= \sum_{i=1}^{\infty} \beta_i \mathbb{P}(B_{1,i} | B_{2,i}) \mathbb{E}_{P^*}[\mathbb{I}_{B_{2,i}}] \leq \sum_{i=1}^{\infty} |\alpha_i| \mathbb{P}(A_{1,i} | A_{2,i}) < \infty \\
\mathbb{E}_{P^*}\left[\sum_{i=1}^{\infty} -\gamma_i \mathbb{I}_{C_{2,i}} \mathbb{P}(C_{1,i} | C_{2,i})\right] &= \sum_{i=1}^{\infty} -\gamma_i \mathbb{P}(C_{1,i} | C_{2,i}) \mathbb{E}_{P^*}[\mathbb{I}_{C_{2,i}}] \leq \sum_{i=1}^{\infty} |\alpha_i| \mathbb{P}(A_{1,i} | A_{2,i}) < \infty
\end{aligned}$$

Portanto, se chamarmos $X_1 = \sum_{i=1}^{\infty} \beta_i \mathbb{I}_{B_{1,i} \cap B_{2,i}}$, $X_2 = \sum_{i=1}^{\infty} -\gamma_i \mathbb{I}_{C_{1,i} \cap C_{2,i}}$, $X_3 = \sum_{i=1}^{\infty} \beta_i \mathbb{I}_{B_{2,i}} \mathbb{P}(B_{1,i} | B_{2,i})$ e $X_4 = \sum_{i=1}^{\infty} -\gamma_i \mathbb{I}_{C_{2,i}} \mathbb{P}(C_{1,i} | C_{2,i})$, estas variáveis estão definidas a.s. P^*. Portanto,

$$\begin{aligned}
\mathbb{E}_{P^*}\left[\sum_{i=1}^{\infty} \alpha_i \mathbb{I}_{A_{2,i}} (\mathbb{I}_{A_{1,i}} - \mathbb{P}(A_{1,i} | A_{2,i}))\right] &= \mathbb{E}_{P^*}[X_1 - X_2 - X_3 + X_4] \\
&= \mathbb{E}_{P^*}[X_1] - \mathbb{E}_{P^*}[X_2] - \mathbb{E}_{P^*}[X_3] + \mathbb{E}_{P^*}[X_4] \\
&= \sum_{i=1}^{\infty} \beta_i \mathbb{P}(B_{1,i} \cap B_{2,i}) + \sum_{i=1}^{\infty} \gamma_i \mathbb{P}(C_{1,i} \cap C_{2,i}) \\
&- \sum_{i=1}^{\infty} \beta_i \mathbb{P}(B_{1,i} | B_{2,i}) \mathbb{P}(B_{2,i}) - \sum_{i=1}^{\infty} \gamma_i \mathbb{P}(C_{1,i} | C_{2,i}) \mathbb{P}(C_{2,i}) \\
&= \sum_{i=1}^{\infty} \beta_i (\mathbb{P}(B_{1,i} \cap B_{2,i}) - \mathbb{P}(B_{1,i} | B_{2,i}) \mathbb{P}(B_{2,i})) + \\
&\sum_{i=1}^{\infty} \gamma_i (\mathbb{P}(C_{1,i} \cap C_{2,i}) - \mathbb{P}(C_{1,i} | B_{2,i}) \mathbb{P}(C_{2,i})) = 0
\end{aligned}$$

Portanto, existe $\omega \in \Omega$ tal que $\sum_{i=1}^{\infty} \alpha_i I_{A_{2,i}}(\omega)(\mathbb{I}_{A_{1,i}}(w) - \mathbb{P}(A_{1,i}|A_{2,i})) > 0$ e o portfólio $(\alpha_i, A_{1,i}, A_{2,i})$ não acarreta prejuízo certo. Como o portfólio $(\alpha_i, A_{1,i}, A_{2,i})$ era arbitrário, conclua que $\mathbb{P}$ é coerente. □

Exercícios

Exercício 3.26 (Desafio)**.** Considere que só são aceitos portfólios de apostas, $(\alpha_i, A_{1,i}, A_{2,i})_{i\in\mathbb{N}}$, tais que, para todo $\omega \in \Omega$, $\sum_{i=1}^{\infty} |\alpha_i \mathbb{I}_{A_{2,i}}(\mathbb{I}_{A_{1,i}} - \mathbb{P}(A_{1,i}|A_{2,i}))| < \infty$. Esta condição foi estudada em Adams (1962). Mostre que $\mathbb{P}$ é coerente se e somente se satisfaz todos os axiomas da probabilidade.

Exercício 3.27 (Desafio)**.** Considere que são aceitos todos os portfólios de apostas, $(\alpha_i, A_{1,i}, A_{2,i})_{i\in\mathbb{N}}$, tais que, para todo $\omega \in \Omega$, $\sum_{i=1}^{\infty} \alpha_i \mathbb{I}_{A_{2,i}}(\mathbb{I}_{A_{1,i}} - \mathbb{P}(A_{1,i}|A_{2,i})) < \infty$. Mostre que, se $\Omega = \{1, 2, \ldots\}$, então $\mathbb{P}(\{i\}) = 2^{-i}$ não é coerente. Se muitos portfólios de apostas são permitidos, então os axiomas da probabilidade não são suficientes para evitar prejuízo certo.

Exercício 3.28 (Desafio)**.** Usando as mesmas condições para apostas do Exercício 3.27, ache condições necessárias e suficientes para que $\mathbb{P}$ seja coerente.

4 Explorando o modelo estatístico

4.1 Predições usando o modelo estatístico

A seção 2.2 introduziu o modelo estatístico usado em estatística bayesiana. Este modelo tem características diferentes daquele que é tipicamente usado na estatística frequentista. Nesta seção desenvolveremos a sua compreensão do modelo estatístico bayesiano pela exploração de algumas de suas propriedades. O Exemplo 4.1 ilustra uma destas propriedades.

Exemplo 4.1. A primeira vez que vi este exemplo, ele foi apresentado oralmente pelo Professor Carlos Alberto de Bragança Pereira, um dos precursores da Inferência Bayesiana no Brasil (e no mundo).

Considere que, dado θ, X_1 e X_2 são i.i.d. e $X_i \sim \text{Uniforme}(\theta - 0.5, \theta + 0.5)$. A princípio, você acredita que é plausível que θ seja qualquer número real, $\theta \in \mathbb{R}$. Assim, como $\theta - 0.5 \leq X_2 \leq \theta + 0.5$ e $\theta \in \mathbb{R}$, a princípio, $X_2 \in \mathbb{R}$.

Agora, suponha que você observa $X_1 = 0$. Como $\theta - 0.5 \leq X_1 \leq \theta + 0.5$, deduza que $X_1 - 0.5 \leq \theta \leq X_1 + 0.5$. Também, como $X_1 = 0$, deduza que $-0.5 \leq \theta \leq 0.5$. Assim, como $\theta - 0.5 \leq X_2 \leq \theta + 0.5$, conclua que $X_2 \in [-1, 1]$.

Assim, antes de observar $X_1 = 0$, você acreditava que X_2 poderia ser qualquer número real. Contudo, após observar $X_1 = 0$, você acredita que X_2 é um número entre -1 e 1. Portanto, observar o valor de X_1 traz informação a respeito de X_2.

A seguir, você verá como o modelo Bayesiano leva em conta esta informação, uma vez que induz dependência entre X_1 e X_2. Assim, permite calcular, usando os axiomas da probabilidade, de que forma a observação de X_1 altera a sua incerteza a respeito de X_2.

A seguir, consideraremos o caso em que os dados são condicionalmente i.i.d. dado que θ é conhecido. Neste caso, temos a seguinte igualdade:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{4}$$

Em particular, temos que

$$\begin{aligned} f(x_2|x_1,\theta) &= \frac{f(x_1,x_2|\theta)}{f(x_1|\theta)} && \text{Definição 1.23} \\ &= \frac{f(x_1|\theta)f(x_2|\theta)}{f(x_1|\theta)} && \text{equação 4} \\ &= f(x_2|\theta) \end{aligned}$$

Em palavras, quando θ é conhecido, X_1 não traz informação a respeito de X_2. E se θ é desconhecido? Ainda é verdade que X_1 não traz informação a respeito de X_2? No modelo estatístico Bayesiano, θ é uma variável aleatória. Portanto, podemos obter a densidade marginal dos dados diretamente do Teorema 1.27.

$$\begin{aligned} f(x_1,\ldots,x_n) &= \int_\Theta f(x_1,\ldots,x_n|\theta)\pi(\theta)d\theta \\ &= \int_\Theta \prod_{i=1}^n f(x_i|\theta)f(\theta)d\theta && \text{eq. (4)} \end{aligned} \tag{5}$$

Em particular, a densidade marginal de X_2 é dada por

$$f(x_2) = \int_\Theta f(x_2|\theta)f(\theta)d\theta \tag{6}$$

Também, quando θ é desconhecido, temos que a densidade de X_2 dado X_1 é dada por

$$\begin{aligned} f(x_2|x_1) &= \frac{f(x_1,x_2)}{f(x_1)} \\ &= \frac{\int_\Theta f(x_1|\theta)f(x_2|\theta)f(\theta)d\theta}{\int_\Theta f(x_1|\theta)f(\theta)d\theta} && \text{eq. (4)} \\ &= \int_\Theta f(x_2|\theta)\left(\frac{f(\theta)f(x_1|\theta)}{\int_\theta f(\theta)f(x_1|\theta)}\right)d\theta \\ &= \int_\Theta f(x_2|\theta)f(\theta|x_1)d\theta && \text{eq. (1)} \end{aligned} \tag{7}$$

Comparando as eqs. (6) e (7), podemos observar de que forma X_1 traz informação a respeito de X_2. Enquanto que a distribuição marginal de X_2 é obtida integrando a densidade de X_2 dado θ com respeito à priori para θ, a distribuição de X_2 dado X_1 é obtida integrando a densidade de X_2 dado θ com respeito à posteriori para θ dado X_1. Especificamente, θ traz informação a respeito de X_2. Assim, quando X_1 traz informação a respeito de θ, também traz informação a respeito de X_2.

Para ilustrar estas relações de dependência, considere o caso de um diagnóstico médico.

Exemplo 4.2. Sabemos que são sintomas frequentes da dengue a dor atrás dos olhos e a perda do paladar. Considere as seguintes variáveis aleatórias:

- θ: A indicadora de que a paciente está infectada pela dengue.
- X_1: A indicadora de que a paciente sente dor atrás dos olhos.
- X_2: A indicadora de que a paciente teve perda do paladar.

Considere que a probabilidade de cada sintoma é aumentada, se soubermos que a paciente está infectada pela

dengue. Também, se soubermos que a paciente está infectada ou não pela dengue, então os sintomas ocorrem independentemente. Especificamente, considere que:

- $\mathbb{P}(X_1 = x_1, X_2 = x_2|\theta = t) = \mathbb{P}(X_1 = x_1|\theta = t)\mathbb{P}(X_2 = x_2|\theta = t)$
- $\mathbb{P}(X_i = 1|\theta = 1) = 0.9$
- $\mathbb{P}(X_i = 1|\theta = 0) = 0.01$
- $\mathbb{P}(\theta = 1) = 0.01$

Observe que

$$\begin{aligned}\mathbb{P}(X_i = 1) &= \mathbb{P}(X_i = 1|\theta = 1)\mathbb{P}(\theta = 1) + \mathbb{P}(X_i = 1|\theta = 0)P(\theta = 0)\\ &= 0.9 \cdot 0.01 + 0.01 \cdot 0.99 \approx 0.019\end{aligned}$$

Portanto, a priori, a probabilidade de uma paciente sofrer um dos sintomas é relativamente baixa. Continuando, o raciocínio, podemos calcular de que forma a paciente ter dor atrás dos olhos afeta o diagnóstico da médica em relação à dengue.

$$\begin{aligned}\mathbb{P}(\theta = 1|X_1 = 1) &= \frac{\mathbb{P}(\theta = 1)\mathbb{P}(X_1 = 1|\theta = 1)}{\mathbb{P}(X_1 = 1)} && \text{Teorema 1.29}\\ &\approx \frac{0.01 \cdot 0.9}{0.019} \approx 0.47\end{aligned}$$

Observamos que, após observar que a paciente tem dor atrás dos olhos, a probabilidade da paciente ter dengue aumenta de 0.01 para 0.47. Assim, é razoável acreditar que a médica passará a acreditar que o sintoma de perda de paladar é mais provável após observar a dor atrás dos olhos. De fato, como X_1 e X_2 são i.i.d. dado θ, obtemos

$$\begin{aligned}\mathbb{P}(X_2 = 1|X_1 = 1) &= \mathbb{P}(X_2 = 1|\theta = 1)\mathbb{P}(\theta = 1|X_1 = 1) + \mathbb{P}(X_2 = 1|\theta = 0)\mathbb{P}(\theta = 0|X_1 = 1) && \text{eq. (7)}\\ &\approx 0.9 \cdot 0.47 + 0.01 \cdot 0.53 \approx 0.42\end{aligned}$$

Assim, dado que uma pessoa está infectada pela dengue, saber que ela sente dor atrás dos olhos não traz informação a respeito de ela ter perdido o paladar. Contudo, se não soubermos que uma pessoa está infectada pela dengue, observar um dos sintomas da dengue aumenta a probabilidade dos demais sintomas. Mais especificamente, observar um sintoma da dengue aumenta a probabilidade de haver um caso de dengue e, assim, aumenta a probabilidade dos demais sintomas da dengue.

A equação 7 pode ser generalizada para o seguinte resultado

Teorema 4.3. *Se $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m}$ são independentes dado θ, então:*

$$f(x_n, \ldots, x_{n+m}|x_1, \ldots, x_{n-1}) = \int_\Theta f(x_n, \ldots, x_{n+m}|\theta) f(\theta|x_1, \ldots, x_{n-1}) d\theta$$

Demonstração.

$$
\begin{aligned}
f(x_n, \ldots, x_{n+m}|x_1, \ldots, x_{n-1}) &= \frac{f(x_1, \ldots, x_{n+m})}{f(x_1, \ldots, x_{n-1})} && \text{Definição 1.23} \\
&= \frac{\int_\Theta f(x_1, \ldots, x_{n+m}|\theta) f(\theta) d\theta}{f(x_1, \ldots, x_{n-1})} && \text{Teorema 1.27} \\
&= \frac{\int_\Theta f(x_1, \ldots, x_{n-1}|\theta) f(x_n, \ldots, x_{n+m}|\theta) f(\theta) d\theta}{f(x_1, \ldots, x_{n-1})} && \text{independência dado } \theta \\
&= \int_\Theta f(x_n, \ldots, x_{n+m}|\theta) \left(\frac{f(x_1, \ldots, x_{n-1}|\theta) f(\theta)}{f(x_1, \ldots, x_{n-1})} \right) d\theta \\
&= \int_\Theta f(x_n, \ldots, x_{n+m}|\theta) f(\theta|x_1, \ldots, x_{n-1}) d\theta && \text{Teorema 1.41}
\end{aligned}
$$

□

Exercícios

Exercício 4.4. Considere que, dado θ, $X \sim \text{Bernoulli}(\theta)$. Também, $\theta \sim \text{Uniforme}(0, 1)$.

(a) Calcule $f(x)$.

(b) Calcule $\mathbb{E}[X]$.

Exercício 4.5. Considere o caso da eleição no Exemplo 2.2. Neste caso, cada indivíduo pode ter a intenção de votar em A ou D. A priori, um analista acreditava que a proporção de indivíduos votando em D seguia uma distribuição $\text{Beta}(a, b)$. O analista escolhe n indivíduos usando uma amostragem simples com reposição e observa que n_D deles tem a intenção de votar em D e $n - n_D$ deles tem a intenção de votar em A.

(a) Calcule a probabilidade a posteriori para θ.

(b) Se o analista sortear (com mesma probabilidade) mais um indivíduo da população, qual a probabilidade de que ele tenha a intenção de votar em D?

(c) Se, em uma amostragem simples com reposição, o analista sortear mais 2 indivíduos da população, qual é a probabilidade de que nenhum deles tenha a intenção de votar em D?

Exercício 4.6. Considere que uma urna grande tem 10 bolas azuis e 10 bolas verdes. 4 bolas são retiradas com reposição da urna grande. A seguir, colocam-se em uma urna média 4 bolas de mesmas cores que aquelas obtidas na amostra com reposição. 2 bolas são retiradas sem reposição da urna média.

(a) Qual é a distribuição para o total de bolas azuis na urna média?

(b) Qual a distribuição marginal da amostra obtida a partir da urna média?

(c) Se a primeira bola obtida na amostragem realizada na urna média for azul, qual é a probabilidade de que a segunda bola também o seja?

Exercício 4.7. Em uma mesa há 2 caixas. A primeira caixa tem 3 bolas azuis e 2 bolas vermelhas. A segunda caixa tem 3 bolas azuis e 4 bolas vermelhas. Considere que uma das caixas é escolhida aleatoriamente (com igual probabilidade entre elas) e duas bolas são retiradas desta com reposição. Defina X_i como a indicadora de que a i-ésima bola retirada era azul.

(a) Ache a distribuição marginal de (X_1, X_2).

(b) Ache $Cov(X_1, X_2)$. X_1 e X_2 são independentes? Este resultado é compatível com a amostragem ter sido feita com reposição?

(c) Dado que as duas bolas retiradas foram azuis, qual a probabilidade a posteriori delas terem sido retiradadas da primeira caixa?

Exercício 4.8. Considere que $X_1, \ldots, X_n, X_{n+1}$ são i.i.d. dado θ e $X_i|\theta \sim \text{Poisson}(\theta)$. Também, a distribuição a priori para θ é dada por $\theta \sim \text{Gamma}(1, 1)$. Ache $\mathbb{P}(x_{n+1}|x_1, \ldots, x_n)$.

Exercício 4.9. Considere que $X_1, \ldots X_{n+1}$ são i.i.d. dado θ, $X_i|\theta \sim \text{Binomial}(n, \theta)$ e $\theta \sim \text{Beta}(a, b)$. Determine $\mathbb{P}(x_{n+1}|x_1, \ldots, x_n)$.

4.2 Permutabilidade*

Nesta subseção, você estudará os fundamentos e a interpretação dos componentes do Modelo Estatístico. Na seção 2.2, você viu que o Modelo Estatístico tem 2 objetos desconhecidos: os dados, X, e uma quantidade não observada associada, θ. Uma vez que os dados são observáveis, em geral é fácil descrever e interpretá-los. Contudo, o mesmo pode não ser verdadeiro de θ. Considere o Exemplo 4.10.

Exemplo 4.10. Uma moeda é lançada 100 vezes. Defina X_i como a indicadora de que o i-ésimo lançamento da moeda seja cara. Ao analisar os dados obtidos, um estatístico supõe que, dado θ, $X_1, \ldots, X_{100}$ são i.i.d. Bernoulli(θ).

O que é θ? Note que θ não é a probabilidade de que um lançamento da moeda seja cara. De fato, enquanto que a probabilidade de um evento deve ser um número, θ é uma variável aleatória. Neste modelo, θ não é uma característica da moeda que possa ser medida diretamente. É necessário usar os dados para aprender a respeito de θ. Mas se você não for capaz de interpretar θ, como poderá colocar uma distribuição sobre θ que reflete a sua incerteza sobre esta quantidade? Nesta subseção você verá ao menos uma forma de interpretar θ no modelo do Exemplo 4.10.

Para tal, considere a definição de permutabilidade.

Definição 4.11 (Permutabilidade finita)**.** $X_1, \ldots, X_n$ são permutáveis se, para qualquer permutação, π, dos índices $\{1, \ldots, n\}$ e $A_i \subset \mathbb{R}$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_{\pi(1)} \in A_1, \ldots, X_{\pi(n)} \in A_n)$$

Note que a suposição de permutabilidade diz respeito à distribuição conjunta dos dados. Ela não se refere a θ ou qualquer objeto não-observável.

Lema 4.12. *Se $X_1, \ldots, X_n$ são i.i.d., então também são permutáveis.*

Demonstração. Seja π uma permutação arbitrária de $\{1, \ldots, n\}$. Temos

$$\begin{aligned}
\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) &= \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) && \text{independência} \\
&= \prod_{i=1}^{n} \mathbb{P}(X_{\pi(i)} \in A_i) && \text{i.d.} \\
&= \mathbb{P}(X_{\pi(1)} \in A_1, \ldots, X_{\pi(n)} \in A_n) && \text{independência}
\end{aligned}$$

X_1 \ X_2	0	1
0	$\frac{1}{6}$	$\frac{1}{3}$
1	$\frac{1}{3}$	$\frac{1}{6}$

Tabela 8: Distribuição conjunta de X_1 e X_2 no Exemplo 4.14.

□

Lema 4.13. *Se $X_1, \ldots, X_n$ são permutáveis, então são identicamente distribuídos.*

Combinando os Lemas 4.12 e 4.13, pode-se concluir que permutabilidade é uma propriedade mais forte que i.d. e mais fraca que i.i.d.

Demonstração. Para todo $n \geq 2$ e $A \subset \mathbb{R}$, temos

$$\begin{aligned}\mathbb{P}(X_1 \in A) &= \mathbb{P}(X_1 \in A, X_2 \in \mathbb{R}, \ldots, X_n \in \mathbb{R}) \\ &= \mathbb{P}(X_1 \in \mathbb{R}, X_2 \in \mathbb{R}, \ldots, X_n \in A) = \mathbb{P}(X_n \in A)\end{aligned}$$

□

Exemplo 4.14. Considere que duas bolas são retiradas sem reposição de uma urna com 2 bolas azuis e 2 bolas brancas. Defina X_i como a indicadora de que a i-ésima bola retirada é azul. Usando a tabela 8, é possível mostrar que X_1 e X_2 são permutáveis mas não são independentes. Como X_1 e X_2 são variáveis Bernoulli e $\mathbb{P}(X_1 = 1, X_2 = 0) = \mathbb{P}(X_1 = 0, X_2 = 1) = \frac{1}{3}$, temos que X_1 e X_2 são permutáveis. Também, como $\mathbb{P}(X_1 = 1, X_2 = 1) = \frac{1}{6} \neq \frac{1}{4} = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1)$, X_1 e X_2 não são independentes.

Combinando-se este exemplo ao Lema 4.12, conclua que i.i.d. é uma propriedade mais forte que permutabilidade. Isto é, se $X_1, \ldots, X_n$ são i.i.d., então são permutáveis. Contudo, a relação inversa não necessariamente é verdadeira.

Definição 4.15 (Permutabilidade infinita)**.** $X_1, \ldots, X_n, \ldots$ são infinitamente permutáveis se, para todo $m \in \mathbb{N}$, $X_1, \ldots, X_m$ são permutáveis.

A seguir, a definição de permutabilidade infinita é conectada ao modelo do Exemplo 4.10. Esta conexão se dá pelo Teorema 4.16.

Teorema 4.16 (De Finetti (1931))**.** *Considere que $X_1, \ldots, X_n, \ldots$ são tais que $X_i \in \{0, 1\}$. $X_1, \ldots, X_n, \ldots$ são infinitamente permutáveis se e somente se existe uma variável aleatória, $\theta \in [0, 1]$ tal que, dado θ, $X_1, \ldots, X_n, \ldots$ são i.i.d. Ber(θ). Isto é, para todo $n \geq 1$ e $x_1, \ldots, x_n \in \{0, 1\}^n$,*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \int_{[0,1]} \theta^{n\bar{x}}(1 - \theta)^{n(1-\bar{x})} P_\theta(d\theta)$$

Ademais $\theta = \lim_{n\to\infty} \bar{X}_n$. Chama-se $\lim_{n\to\infty} \bar{X}_n$ de média assintótica da sequência. Ela corresponde ao limite das frequências de 1 observadas.

Em palavras, dada a média assintótica de uma sequência infinitamente permutável de Bernoullis, esta sequência é i.i.d. com média igual à média assintótica. Em geral, a média assintótica é desconhecida e, assim, é uma variável aleatória. Esta variável aleatória é precisamente o parâmetro do modelo do Exemplo 4.10!

O Teorema 4.16 será provado em duas etapas. Primeiramente, será mostrado que existe uma subsequência de $\bar{X}_n$ que converge quase certamente. A seguir, este fato e uma aproximação da distribuição hipergeométrica pela distribuição binomial serão usados para completar o Teorema de Representação. Esta demonstração é baseada naquelas que se encontram em Schervish (2012; pp.34-38) e Heath and Sudderth (1976).

Teorema 4.17. *Considere que $X_1, \ldots, X_n, \ldots$ é uma sequência infinitamente permutável tal que $-\infty < E[X_1 X_2] = m_2 < \infty$ e $E[X_1^2] = \mu_2 < \infty$. $\bar{X}_{8^k} = \frac{\sum_{i=1}^{8^k} X_i}{8^k}$ converge quase certamente.*

Demonstração. Note que

$$\begin{aligned}
\mathbb{P}(|\bar{X}_{8^k} - \bar{X}_{8^{k+1}}| > 2^{-k}) &= \mathbb{P}(|\bar{X}_{8^k} - \bar{X}_{8^{k+1}}|^2 > 4^{-k}) \\
&\leq \frac{\mathbb{E}[|\bar{X}_{8^k} - \bar{X}_{8^{k+1}}|^2]}{4^{-k}} && \text{Markov} \\
&= \frac{\mathbb{E}[\bar{X}_{8^k}^2] + \mathbb{E}[\bar{X}_{8^{k+1}}^2] - 2\mathbb{E}[\bar{X}_{8^k}\bar{X}_{8^{k+1}}]}{4^{-k}} \\
&= \frac{8^{-2k}(8^k\mu_2 + 8^k(8^k-1)m_2)}{4^{-k}} + \frac{8^{-2(k+1)}(8^{k+1}\mu_2 + 8^{k+1}(8^{k+1}-1)m_2)}{4^{-k}} \\
&- 2\frac{8^{-(2k+1)}(8^k\mu_2 + 8^k(8^k-1)m_2 + 8^{2k}m_2)}{4^{-k}} \\
&= \frac{(8^{-k} - 8^{-(k+1)})(\mu_2 + m_2)}{4^{-k}} < 2^{-k}(\mu_2 + m_2)
\end{aligned}$$

Defina $A_k = \{w \in \Omega : |\bar{X}_{8^k} - \bar{X}_{8^{k+1}}| > 2^{-k}\}$. Defina $A = \{A_k \ i.v.\} = \cap_{i=1}^{\infty} \cup_{j=i}^{\infty} A_j$. Como

$$\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \sum_{i=1}^{\infty} 2^{-i}(\mu_2 + m_2) = (\mu_2 + m_2) < \infty$$

conclua por Borel-Cantelli que $\mathbb{P}(A) = 0$. Finalmente, para mostrar que $\bar{X}_{8^k}$ converge quase certamente, mostraremos que para todo $\omega \in A^c$, $\bar{X}_{8^k}(w)$ é uma sequência de Cauchy. Isto é, para todo $\epsilon > 0$, existe N tal que, para todo $n, m > N$, $|\bar{X}_{8^n}(\omega) - \bar{X}_{8^m}(\omega)| < \epsilon$. Considere um $\epsilon > 0$ arbitrário. Tome $\omega \in A^c$. Por definição, existe um l_ω tal que, para todo $n > l_\omega$, $|\bar{X}_{8^n}(\omega) - \bar{X}_{8^{n+1}}(\omega)| < 2^{-k}$. Para todo $m > n > l_\omega$, temos

$$\begin{aligned}
|\bar{X}_{8^n}(\omega) - \bar{X}_{8^m}(\omega)| &\leq \sum_{k=n}^{m} |\bar{X}_{8^k}(\omega) - \bar{X}_{8^{k+1}}(\omega)| \\
&\leq \sum_{k=n}^{\infty} |\bar{X}_{8^k}(\omega) - \bar{X}_{8^{k+1}}(\omega)| \\
&\leq \sum_{k=n}^{\infty} 2^{-k} < 2^{-n+1}
\end{aligned}$$

Portanto, para todo $n, m > \max(l_\omega, \log_2 \epsilon) = N$, obtemos $|\bar{X}_{8^n}(\omega) - \bar{X}_{8^m}(\omega)| < \epsilon$. Conclua que $\bar{X}_{8^k}(\omega)$ é uma sequência de Cauchy e, assim, uma sequência convergente. Como $\omega \in A^c$ era arbitrário e $\mathbb{P}(A^c) = 1$, conclua que $\bar{X}_{8^k}$ converge quase certamente. □

Prova do Teorema 4.16. Defina $N = 8^k$ e note que

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= \mathbb{P}(X_1 = 1, \ldots, X_{n\bar{x}} = 1, X_{n\bar{x}+1} = 0, \ldots, X_n = 0) \\
&= \sum_{i=0}^{N} \mathbb{P}(X_1 = 1, \ldots, X_{n\bar{x}} = 1, X_{n\bar{x}+1} = 0, \ldots, X_n = 0 | N\bar{X}_N = i)\mathbb{P}(N\bar{X}_N = i) \\
&= \sum_{i=0}^{N} \frac{\binom{N-n}{i-n\bar{x}_n}}{\binom{N}{i}} \mathbb{P}(N\bar{X}_N = i) \\
&= \mathbb{E}\left[\frac{\binom{N-n}{N\bar{X}_N - n\bar{x}_n}}{\binom{N}{N\bar{X}_N}}\right] \\
&= \mathbb{E}\left[\frac{\frac{(N-n)!}{(N\bar{X}_N - n\bar{x}_n)!(N(1-\bar{X}_N)-n(1-\bar{x}_n))!}}{\frac{N!}{(N\bar{X}_N)!(N(1-\bar{X}_N))!}}\right] \\
&= \mathbb{E}\left[\frac{\frac{(N\bar{X}_N)!}{(N\bar{X}_N - n\bar{x}_n)!}\frac{(N(1-\bar{X}_N))!}{(N(1-\bar{X}_N)-n(1-\bar{x}_n))!}}{\frac{N!}{(N-n)!}}\right]
\end{aligned}$$

Como a igualdade vale para todo N, obtemos

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= \lim_N \mathbb{E}\left[\frac{\frac{(N\bar{X}_N)!}{(N\bar{X}_N - n\bar{x}_n)!}\frac{(N(1-\bar{X}_N))!}{(N(1-\bar{X}_N)-n(1-\bar{x}_n))!}}{\frac{N!}{(N-n)!}}\right] \\
&= \mathbb{E}\left[\lim_N \frac{\frac{(N\bar{X}_N)!}{(N\bar{X}_N - n\bar{x}_n)!}\frac{(N(1-\bar{X}_N))!}{(N(1-\bar{X}_N)-n(1-\bar{x}_n))!}}{\frac{N!}{(N-n)!}}\right] \\
&= \mathbb{E}\left[\lim_N \frac{(N\bar{X}_N)^{n\bar{x}_n} \cdot (N(1-\bar{X}_N))^{n(1-\bar{x}_n)}}{N^n}\right] \\
&= \mathbb{E}\left[\lim_N \bar{X}_N^{n\bar{x}_n}(1-\bar{X}_N)^{n(1-\bar{x}_n)}\right] \\
&= \mathbb{E}\left[\theta^{n\bar{x}_n}(1-\theta)^{n(1-\bar{x}_n)}\right] \qquad \text{Teorema 4.17} \\
&= \int_{[0,1]} \theta^{n\bar{x}_n}(1-\theta)^{n(1-\bar{x}_n)} P_\theta(d\theta)
\end{aligned}$$

□

O Teorema 4.16 mostra que, para toda sequência infinitamente permutável de quantidades aleatórias em {0,1}, existe uma quantidade aleatória, θ, tal que, dado θ, a sequência é i.i.d. Poderíamos perguntar se o mesmo resultado vale para quantidades aleatórias em $\mathbb{R}$. A resposta afirmativa é provada em De Finetti (1973) e enunciada a seguir.

Teorema 4.18 (De Finetti (1973))**.** *Considere que $X_1, \ldots, X_n, \ldots$ são tais que $X_i \in \mathbb{R}$. $X_1, \ldots, X_n, \ldots$ são infinitamente permutáveis se e somente se existe uma medida de probabilidade, μ, sobre as funções de distribuição acumulada tal que para todo $n \geq 1$ e $x_1, \ldots, x_n$*

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq, x_n) = \int \left(\prod_{i=1}^{n} F(x_i)\right) d\mu(F)$$

Isto é, existe uma função de distribuição acumulada aleatória, F, dada a qual $X_1, \ldots, X_n$ são i.i.d. com distri-

buição F.

Além destes teoremas, muitas outras extensões do Teorema de Bruno de Finetti foram provadas. Por exemplo, Kingman et al. (1978), Aldous (1985) revisam a bibliografia sobre o tema.

Exercícios

Exercício 4.19 (Urna de Pólya)**.** Considere que, inicialmente, uma urna tem 1 bola branca e 1 bola azul. A cada iteração, retira-se uma bola da urna e, em seguida, recoloca-se 2 bolas da mesma cor na urna. Assim, por exemplo, se a primeira bola retirada for azul, na segunda iteração haverá 1 bola branca e 2 bolas azuis na urna. Considere que X_i é a indicadora de que a i-ésima bola retirada é azul.

(a) Para cada $n \in \mathbb{N}$, ache a distribuição conjunta de $(X_1, \ldots, X_n)$.

(b) Mostre que $X_1, \ldots, X_n, \ldots$ é infinitamente permutável.

(c) De acordo com a Definição 1.23, dado $\lim_n \bar{X}_n$, $X_1, \ldots, X_n$ são i.i.d. e $X_1 \sim \text{Bernoulli}(\lim_n \bar{X}_n)$. Use os itens anteriores para determinar a distribuição de $\lim_n \bar{X}_n$.

Exercício 4.20. Refaça o Exercício 4.19 considerando que, inicialmente, a urna tem "a" bolas azuis e "b" bolas brancas e que, a cada iteração, recoloca-se na urna "c" bolas da mesma cor da bola retirada.

Exercício 4.21 (Rodrigues and Wechsler (1993))**.** Considere que $X_1, \ldots, X_n$ é uma sequência infinitamente permutável de variáveis aleatórias em $\{0, 1\}$. Defina $r(i) = \mathbb{P}(X_i = 1 | X_{i-1} = 0, \ldots, X_1 = 0)$ e $r(1) = P(X_1 = 1)$.

(a) Mostre que $r(1) \geq r(2)$.

(b) Mostre que $r(i)$ é decrescente.

Exercício 4.22 (O'Neill and Puza (2005))**.** Considere que, dado θ, $X_1, \ldots, X_n$ são i.i.d. e $X_i \sim \text{Bernoulli}(\theta)$. Também, θ e $(1 - \theta)$ são permutáveis, ou seja, a distribuição de θ é simétrica em relação a 0.5. Mostre que,

$$\mathbb{P}(X_{n+1} = 1 | X_1 = x_1, \ldots, X_n = x_n) > 0.5 \text{ se e somente se } \sum_{i=1}^{n} x_i > \frac{n}{2}.$$

Exercício 4.23 (Bonassi et al. (2015))**.** Dizemos que $Y \sim \text{CMP-Binomial}(n, p, \nu)$ (Kadane et al., 2016) se

$$\mathbb{P}(Y = y) \propto \binom{n}{y}^{\nu} p^y (1 - p)^{n-y}$$

Considere que $X_1, \ldots, X_{n+1}$ são permutáveis e $\sum_{i=1}^{n+1} X_i \sim \text{CMP-Binomial}(n + 1, 0.5, \nu)$.

(a) Mostre que, se $\nu > 1$ e $\sum_{i=1}^{n} x_i > \frac{n}{2}$, então $\mathbb{P}(X_{n+1} = 1 | X_1 = x_1, \ldots, X_n = x_n) > 0.5$.

(b) Mostre que, se $\nu < 1$ e $\sum_{i=1}^{n} x_i > \frac{n}{2}$, então $\mathbb{P}(X_{n+1} = 1 | X_1 = x_1, \ldots, X_n = x_n) < 0.5$.

Exercício 4.24 (Permutabilidade e covariância)**.**

(a) Considere que $X_1, \ldots, X_n, \ldots$ é uma sequência infinitamente permutável tal que $X_i \in \{0, 1\}$. Mostre que, para qualquer $i \neq j$, $Cov(X_i, X_j) \geq 0$.

(b) Considere que $X_1, \ldots, X_n$ é uma sequência permutável. Mostre que

$$Corr(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var[X_i]Var[X_j]}} \geq -(n-1)^{-1}.$$

Para provar este resultado, defina $Y = \sum_{i=1}^{n} X_i$ e tome como ponto de partida a desigualdade $\mathbb{V}[Y] \geq 0$.

(c) Para todo $n \in \mathbb{N}$, construa um exemplo tal que $X_1, \ldots, X_n$ é permutável e, para todo $i \neq j$, $Cov(X_i, X_j) < 0$.

4.3 Famílias conjugadas

Em muitos dos exemplos que vimos nas seções anteriores, a priori e a posteriori pertenciam a uma mesma "família" de distribuições. Por exemplo, no Exemplo 2.2, consideramos a priori como sendo uma Beta$(5, 5)$. Após observar que uma Binomial$(30, \theta)$ assume o valor 20, nossa posteriori segue a distribuição Beta$(25, 15)$. Assim, tanto a priori quanto a posteriori pertenciam à família Beta.

O fato de que a priori e a posteriori pertenciam à mesma família nos exemplos que consideramos não é uma coincidência. Pelo contrário, os modelos estatísticos que apresentam esta propriedade são convenientes computacionalmente e é justamente por isso que começamos nossos estudos por eles. Para entender o motivo desta facilidade computacional, lembre-se que:

$$f(\theta_0|x) = \frac{f(\theta_0)f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta}$$

Em geral, a expressão $\int f(\theta)f(x|\theta)d\theta$ pode ser difícil de calcular. Nestes casos, será difícil obter $f(\theta_0|x)$ diretamente a partir do Teorema de Bayes (capítulos futuros, analisarão o que pode ser feito nesta situação). Contudo, vimos que, para alguns modelos estatísticos, pode ser conveniente escrever

$$f(\theta_0|x) \propto f(\theta_0)f(x|\theta_0)$$

Note que existe uma única constante K, tal que $\int Kf(\theta_0)f(x|\theta_0) = 1$. Assim, como $\int f(\theta_0|x)d\theta = 1$, se identificarmos que $f(\theta_0)f(x|\theta_0)$ é proporcional a uma distribuição, $g_{a,b}(\theta_0)$, então podemos concluir que, $f(\theta_0|x) = g_{a,b}(\theta_0)$. Observe que, neste caso, sabemos calcular $\int f(\theta)f(x|\theta)d\theta$:

$$\begin{aligned} f(\theta_0|x) &= \frac{\pi(\theta_0)f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta} \\ g_{a,b}(\theta_0) &= \frac{f(\theta_0)f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta} \qquad\qquad f(\theta_0|x) = g_{a,b}(\theta_0) \\ \int f(\theta)f(x|\theta)d\theta &= \frac{f(\theta_0)f(x|\theta_0)}{g_{a,b}(\theta_0)} \end{aligned}$$

No Exemplo 2.2, observamos que $f(\theta_0)f(x|\theta_0)$ era proporcional a uma distribuição Beta(25,15). Assim, $\int f(\theta)f(x|\theta)d\theta$ é a constante que, dividida por $f(\theta_0)f(x|\theta_0)$, resulta na distribuição Beta$(25, 15)$.

É possível formalizar a propriedade que faz com que saibamos determinar uma distribuição proporcional a $f(\theta_0)f(x|\theta_0)$. Para tal, considere a seguinte definição

Definição 4.25. Considere $\mathcal{P}$ como sendo uma família de funções sobre Θ não-negativas e integráveis. Também, seja $f(x|\theta)$ a densidade condicional de x dado θ. Dizemos que $\mathcal{P}$ é conjugada para $f(x|\theta)$ se, para todo $x \in \chi$, $f(x|\theta_0)f(\theta_0) \in \mathcal{P}$.

Em palavras, se os dados seguem a distribuição $f(x|\theta)$, a sua priori para θ está em $\mathcal{P}$ e $\mathcal{P}$ é conjugada para $f(x|\theta)$, então sua posteriori também estará em $\mathcal{P}$.

Exemplo 4.26. Considere $f(x|\theta) = \theta^x(1-\theta)^{1-x}$, a densidade da Bernoulli(θ). Defina

$$\mathcal{P} = \{f(\theta_0) : f(\theta_0) = K \cdot \theta_0^{a-1}(1-\theta_0)^{b-1}, (a, b, K) > 0\}$$

Note que, para todo $f(\theta_0) \in \mathcal{P}$,

$$\begin{aligned} f(\theta_0)f(x|\theta_0) &= K \cdot \theta_0^{a-1}(1-\theta_0)^{b-1}\theta_0^x(1-\theta_0)^{1-x} \\ &= K \cdot \theta_0^{a+x-1}(1-\theta_0)^{b+1-x-1} \in \mathcal{P} \end{aligned}$$

Portanto, $\mathcal{P}$ é conjugada para $f(x|\theta)$.

Em particular, considere que $\mathcal{P}$ é um conjunto de funções que sabemos integrar. Se $\mathcal{P}$ é conjugada para $f(x|\theta)$, então para todo x, $f(\theta_0)f(x|\theta_0) \in \mathcal{P}$. Portanto, sabemos integrar $f(\theta_0)f(x|\theta_0) \in \mathcal{P}$, ou seja, sabemos calcular $\int f(\theta)f(x|\theta)d\theta$. Assim, podemos calcular diretamente a posteriori a partir da equação

$$f(\theta_0|x) = \frac{f(\theta_0)f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta}$$

Exemplo 4.27. Considere os $\mathcal{P}$ e $f(x|\theta)$ definidos no Exemplo 4.26. Sabemos que

$$\int K\theta^{a-1}(1-\theta)^{b-1}d\theta = K \cdot \beta(a,b)$$

Portanto,

$$\begin{aligned} f(\theta_0|x) &= \frac{f(\theta_0)f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta} \\ &= \frac{K\theta_0^{a+x-1}(1-\theta_0)^{b+1-x-1}}{\int K\theta^{a+x-1}(1-\theta)^{b+1-x-1}d\theta} \\ &= \frac{K\theta_0^{a+x-1}(1-\theta_0)^{b+1-x-1}}{K \cdot \beta(a+x, b+1-x)} \\ &= \beta^{-1}(a+x, b+1-x)\theta_0^{a+x-1}(1-\theta_0)^{b+1-x-1} \end{aligned}$$

Em resumo, se $\mathcal{P}$ é conjugada para $f(x|\theta)$ e sabemos calcular a integral das funções em $\mathcal{P}$, então podemos calcular diretamente a posteriori quando usamos uma priori em $\mathcal{P}$. Esta é uma das razões pela qual $\mathcal{P}$ é computacionalmente conveniente. Também, lembre-se que a densidade preditiva dos dados é

$$f(x) = \int f(\theta)f(x|\theta)d\theta$$

Saber calcular $\int f(\theta)f(x|\theta)d\theta$ é equivalente a saber determinar a densidade preditiva dos dados.

Lema 4.28. *Se, dado θ, $X_1, \ldots, X_n$ são i.i.d. com densidade dada por $f(x_1|\theta)$ e $\mathcal{P}$ é conjugada para $f(x_1|\theta)$, então $\mathcal{P}$ é conjugada para $f(x_1, \ldots, x_n|\theta)$.*

Demonstração. Provaremos o resultado por indução. Por definição, $\mathcal{P}$ é conjugada para $f(x_1|\theta)$. Assuma que $\mathcal{P}$ é conjugada para $f(x_1, \ldots, x_{n-1}|\theta)$. Note que

$$f(\theta)f(x_1, \ldots, x_n|\theta) = f(\theta)f(x_1, \ldots, x_{n-1}|\theta)f(x_n|\theta) \qquad \text{independência condicional}$$

Por hipótese de indução $f^*(\theta) = f(\theta)f(x_1, \ldots, x_{n-1}|\theta) \in \mathcal{P}$. Como $X_1|\theta$ tem mesma distribuição que $X_n|\theta$ e $f^*(\theta) \in \mathcal{P}$, conclua que $f^*(\theta)f(x_n|\theta) \in \mathcal{P}$. Assim, $f(\theta)f(x_1, \ldots, x_n|\theta) \in \mathcal{P}$. □

A seguir, estudaremos algumas famílias conjugadas que são tradicionalmente utilizadas. Para cada uma delas,

derivaremos a forma da posteriori e a densidade preditiva dos dados.

4.3.1 O modelo beta-binomial (Dirichlet-multinomial)

Lema 4.29. *Se $X|\theta \sim Binomial(n,\theta)$, então $\mathcal{P} = \{f(\theta) : f(\theta) = K\theta^{a-1}(1-\theta)^{b-1}, (a,b,K) > 0\}$ é conjugada de $f(x|\theta)$. Note que todas as densidades em $\mathcal{P}$ são da forma $Beta(a,b)$.*

Demonstração.

$$\begin{aligned} f(\theta)f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x}K\theta^{a-1}(1-\theta)^{b-1} \\ &= \binom{n}{x}K\cdot\theta^{a+x-1}(1-\theta)^{b+n-x-1} \in \mathcal{P} \end{aligned}$$

□

Lema 4.30. *Se $X|\theta \sim Binomial(n,\theta)$ e $\theta \sim Beta(a,b)$, então $\theta|X = x \sim Beta(a+x, b+n-x)$.*

Demonstração.

$$\begin{aligned} f(\theta_0|x) &= \frac{f(\theta_0)f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta} \\ &= \frac{\beta(a,b)\theta_0^{a-1}(1-\theta_0)^{b-1}\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}}{\int \beta(a,b)\theta^{a-1}(1-\theta)^{b-1}\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta} \\ &= \frac{\theta_0^{a+x-1}(1-\theta_0)^{b+n-x-1}}{\int \theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta} \\ &= \beta^{-1}(a+x,b+n-x)\theta_0^{a+x-1}(1-\theta_0)^{b+n-x-1} \end{aligned}$$

□

Lema 4.31. *Se $X|\theta \sim Binomial(n,\theta)$ e $\theta \sim Beta(a,b)$, então:*

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\binom{n}{x}\frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)}$$

dizemos que $f(x)$ é a densidade da distribuição Beta-Binomial$(n, a+x, b+n-x)$.

Demonstração.

$$\begin{aligned} f(x) &= \int f(x|\theta)f(\theta)d\theta \\ &= \int \binom{n}{x}\theta^x(1-\theta)^{n-x}\beta^{-1}(a,b)\theta^{a-1}(1-\theta)^{b-1}d\theta \\ &= \binom{n}{x}\beta^{-1}(a,b)\int \theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta \\ &= \binom{n}{x}\beta^{-1}(a,b)\beta(a+x,b+n-x) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\binom{n}{x}\frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} \end{aligned}$$

□

Definição 4.32. Dizemos que $\mathbf{X} \in \{0,1\}^d$ tem distribuição multinomial com parâmetro (m, θ), para $\theta \in \mathbb{R}^d$ tal que $\theta_i \geq 0$ e $\sum_{i=1}^d \theta_i = 1$, e escrevemos $\mathbf{X} \sim \text{Mult}(m, \theta)$ se

$$\mathbb{P}(X_1 = x_1, \dots X_d = x_d | \theta) = \left(\frac{m!}{\prod_{j=1}^d x_i!} \prod_{i=1}^d \theta_i^{x_i} \right) \mathbb{I}\left(\sum_{i=1}^d x_i = m \right)$$

Definição 4.33. Dizemos que θ tem distribuição Dirichlet com parâmetro $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$, onde $\alpha_i \geq 0$, e escrevemos $\theta \sim \text{Dirichlet}(\alpha)$ se

$$f(\theta) = \beta^{-1}(\alpha) \left(\prod_{i=1}^d \theta_i^{\alpha_i - 1} \right) \mathbb{I}\left(\sum_{i=1}^d \theta_i = 1 \right) \qquad \text{, onde } \beta(\alpha) = \frac{\prod_{i=1}^d \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^d \alpha_i \right)}$$

Lema 4.34. *Se $\mathbf{X}|\theta \sim$ Mult(m, θ), então $\mathcal{P} = \left\{ f(\theta) : f(\theta) = K \prod_{i=1}^d \theta_i^{\alpha_i - 1}, \alpha \in \mathbb{R}_+^d \right\}$ é conjugada para $f(\mathbf{x}|\theta)$. Ademais, se $\theta \sim$ Dirichlet(α), então $\theta|\mathbf{X} \sim$ Dirichlet$(\alpha + \mathbf{X})$.*

Demonstração.

$$\begin{aligned} f(\theta) f(\mathbf{x}|\theta) &= \left(\prod_{i=1}^d \theta_i^{\alpha_i - 1} \right) \mathbb{I}\left(\sum_{i=1}^d \theta_i = 1 \right) \left(\frac{m!}{\prod_{j=1}^d x_i!} \prod_{i=1}^d \theta_i^{x_i} \right) \mathbb{I}\left(\sum_{i=1}^d x_i = m \right) \\ &\propto \left(\prod_{i=1}^d \theta_i^{\alpha_i + x_i - 1} \right) \mathbb{I}\left(\sum_{i=1}^d \theta_i = 1 \right) \in \mathcal{P} \end{aligned}$$

□

Exercícios

Exercício 4.35. Considere que, dado θ, $X_1, \dots, X_n$ são i.i.d Binomial(m, θ). Também, $\theta \sim \text{Beta}(a, b)$.

(a) Ache a posteriori para θ após observar $X_1, \dots, X_n$.

(b) Ache a densidade preditiva para $X_1, \dots, X_n$.

(c) Derive $\lim_{n \to \infty} \mathbb{E}[\theta | X_1, \dots, X_n]$.

(d) Derive $\lim_{n \to \infty} \mathbb{V}[\theta | X_1, \dots, X_n]$.

(e) Interprete, em suas palavras, os dois itens anteriores.

Exercício 4.36. Se $\mathbf{X}|\theta \sim \text{Mult}(m, \theta)$ e $\theta \sim \text{Dirichlet}(\alpha)$, determine $f(\mathbf{x})$.

4.3.2 O modelo normal-normal

O modelo normal é um dos mais usados em Estatística. Lembre-se que se $X \sim N(\mu, \sigma^2)$, então:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

Para nossa conveniência, usaremos uma parametrização do modelo normal diferente da usual. Definimos $\tau^2 = \sigma^{-2}$. Assim, quanto maior o valor de τ^2, menor o valor de σ^2. Dada esta propriedade, é comum chamarmos τ^2 de

precisão da distribuição. Realizando a substituição adequada, obtemos:

$$f(x|\mu,\tau^2) = \frac{\tau}{\sqrt{2\pi}} \exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right)$$

A princípio, a distribuição normal pode ter dois parâmetros, μ e τ^2. Contudo, como uma simplificação inicial, consideraremos que τ^2 é conhecido e μ é desconhecido. Como candidato para uma família conjugada neste caso, considere,

$$\mathcal{P} = \left\{ f : \mathbb{R} \to \mathbb{R}^+ : f(\mu) = K \cdot \exp\left(-\frac{\tau_0^2(\mu-\mu_0)^2}{2}\right) \right\}$$

Em outras palavras, $\mathcal{P}$ é o conjunto de funções proporcionais à densidade de uma distribuição normal. Podemos mostrar que $\mathcal{P}$ é conjugada para a família normal.

Lema 4.37. *Se τ^2 é conhecido e $X|\mu \sim N(\mu,\tau^2)$, então*

$$\mathcal{P} = \left\{ f : \mathbb{R} \to \mathbb{R}^+ : f(\mu) = K \cdot \exp\left(-\frac{\tau_0^2(\mu-\mu_0)^2}{2}\right) \right\}$$

é conjugada de $f(x|\mu)$. Note que todas as densidades em $\mathcal{P}$ são da forma $N(\mu_0,\tau_0^2)$. Também, se $\mu \sim N(\mu_0,\tau_0^2)$, então

$$\mu|X \sim N\left(\frac{\tau_0^2\mu_0 + \tau^2 x}{\tau_0^2+\tau^2}, \tau_0^2+\tau^2\right)$$

Demonstração.

$$\begin{aligned}
f(\mu)\cdot f(x|\mu) &= K \cdot \exp\left(-\frac{\tau_0^2(\mu-\mu_0)^2}{2}\right) \cdot \frac{\tau}{\sqrt{2\pi}} \cdot \exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\
&\propto \exp\left(-\frac{\tau_0^2(\mu-\mu_0)^2}{2}\right) \cdot \exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\
&= \exp\left(-\frac{\tau_0^2\mu^2 - 2\tau_0^2\mu\mu_0 + \tau^2\mu^2 - 2\tau^2\mu x}{2}\right) \cdot \exp\left(-\frac{\tau_0^2\mu_0^2 + \tau^2x^2}{2}\right) \\
&\propto \exp\left(-\frac{\tau_0^2\mu^2 - 2\tau_0^2\mu\mu_0 + \tau^2\mu^2 - 2\tau^2\mu x}{2}\right) \\
&= \exp\left(-\frac{(\tau_0^2+\tau^2)\mu^2 - (2\tau_0^2\mu_0 + 2\tau^2x)\mu}{2}\right) \qquad (8)
\end{aligned}$$

Desejamos transformar a expressão da eq. (8) em um quadrado perfeito. Para tal, observe que:

$$\exp\left(-\frac{(a\mu-b)^2}{2}\right) = \exp\left(-\frac{a^2\mu^2 - 2ab\mu + b^2}{2}\right) \qquad (9)$$

Para que exista a chance da eq. (8) ser colocada na forma da eq. (9), é necessário que os coeficientes de μ sejam equivalentes. Assim,

$$\begin{cases} a^2 &= \tau_0^2 + \tau^2 \\ 2ab &= 2\tau_0^2\mu_0 + 2\tau^2 x \end{cases}$$

e obtemos

$$\begin{cases} a &= \sqrt{\tau_0^2 + \tau^2} \\ b &= \frac{\tau_0^2\mu_0 + \tau^2 x}{\sqrt{\tau_0^2+\tau^2}} \end{cases} \tag{10}$$

Realizando esta substituição na eq. (8), obtemos:

$$\begin{aligned} f(\mu)\cdot f(x|\mu) &\propto \exp\left(-\frac{a^2\mu^2 - 2ab\mu}{2}\right) \\ &\propto \exp\left(-\frac{a^2\mu^2 - 2ab\mu + b^2}{2}\right) && \exp(b^2/2) \text{ é constante} \\ &= \exp\left(-\frac{(a\mu - b)^2}{2}\right) && \textit{eq. (9)} \\ &= \exp\left(-\frac{a^2\left(\mu - \frac{b}{a}\right)^2}{2}\right) \in \mathcal{P} \end{aligned}$$

Também observe que $f(\mu)\cdot f(x|\mu)$ é proporcional à densidade de uma normal com $\mu = \frac{b}{a}$ e $\tau^2 = a^2$. Substituindo os valores obtidos na eq. (10) temos que

$$\mu|X \sim N\left(\frac{\tau_0^2\mu_0 + \tau^2 x}{\tau_0^2 + \tau^2}, \tau_0^2 + \tau^2\right)$$

□

Os parâmetros da posteriori no Lema 4.37 ilustram como a posteriori combina a priori e o dado. A precisão da posteriori é $\tau_0^2 + \tau^2$, ou seja, a soma das precisões da priori e do dado. Também, a média da posteriori pode ser reescrita da seguinte forma:

$$\begin{aligned} \mathbb{E}[\mu|x] &= \frac{\tau_0^2\mu_0 + \tau^2 x}{\tau_0^2 + \tau^2} \\ &= \mu_0 \cdot \frac{\tau_0^2}{\tau_0^2 + \tau^2} + x \cdot \frac{\tau^2}{\tau_0^2 + \tau^2} \\ &= \mu_0 \cdot \frac{\tau_0^2}{\tau_0^2 + \tau^2} + x \cdot \left(1 - \frac{\tau_0^2}{\tau_0^2 + \tau^2}\right) \end{aligned}$$

Assim, a média da posteriori pode ser escrita como uma média ponderada (por suas respectivas precisões) da priori e do dado.

Exemplo 4.38. Considere que desejamos saber μ, a altura média de um adulto brasileiro. Para tanto, vamos assumir que $\mu \sim N(165, 1/100)$. Medimos então X, a altura de um brasileiro selecionado ao acaso. Se $X|\mu \sim N(\mu, 1/25)$ e observamos um indivíduo com $1,8m$ (i.e., $x = 180$), então $\mu|x \sim N(177, 0.05)$. A Figura 2 mostra a distribuição a priori e a distribuição a posteriori. Note que a posteriori tem a maior parte de sua massa para valores de μ mais próximos do dado observado.

```
library(ggplot2)
grid_mu <- seq(100,200,length.out = 1000)
```

```r
# priori para mu (altura média de um brasileiro)
mu0 <- 165
sigma0 <- 10
tau0_2 <- 1/(sigma0)^2
data_plot <- data.frame(mu=grid_mu,
                        density=dnorm(grid_mu,mu0,sigma0),
                        distribution="Priori")

# dados (altura medida de um indivíduo - normal com média conhecida)
x <- 180
sigma <- 5
tau_2 <- 1/(sigma)^2

# posteriori
tau_linha_2 <- tau0_2 + tau_2
mu_linha <- tau0_2/tau_linha_2*mu0+
 tau_2/tau_linha_2*x

data_plot <- rbind(data_plot,
                   data.frame(mu=grid_mu,
                              density=dnorm(grid_mu,mu_linha,1/sqrt(tau_linha_2)),
                                            distribution="Posteriori"))

ggplot(data_plot)+
 geom_line(aes(x=mu,y=density,color=distribution),size=1.2)+
 theme_minimal(base_size = 14)+xlab(expression(mu))+ylab("Densidade")+
 theme(legend.title=element_blank())
```

A seguir, consideraremos o caso em que μ é conhecido e τ^2 é desconhecido.

Lema 4.39. *Se μ é conhecido e $X|\tau^2 \sim N(\mu, \tau^2)$, então*

$$\mathcal{P} = \left\{ f : \mathbb{R} \to \mathbb{R}^+ : f(\tau^2) = K \cdot \left(\tau^2\right)^{\alpha-1} \cdot \exp(-\beta\tau^2) \right\}$$

é conjugada de $f(x|\tau^2)$. Note que todas as densidades em $\mathcal{P}$ são da forma $Gamma(\alpha, \beta)$. Também, se $\tau^2 \sim Gamma(\alpha, \beta)$, então

$$\tau^2|X \sim Gamma\left(\alpha + \frac{1}{2}, \beta + \frac{(X-\mu)^2}{2}\right)$$

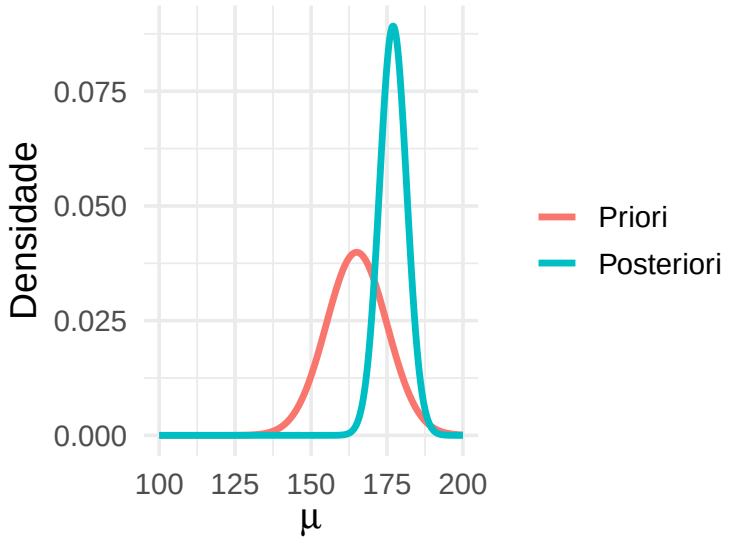

Figura 2: Distribuições a priori e a posteriori (depois de observamos um indivíduo com 1,8m) para a altura média de um brasileiro

Demonstração.

$$\begin{aligned} f(\tau^2)\cdot f(x|\mu) &= K\cdot\left(\tau^2\right)^{\alpha-1}\cdot\exp(-\beta\tau^2)\cdot\frac{\tau}{\sqrt{2\pi}}\cdot\exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\ &\propto \left(\tau^2\right)^{\alpha-1}\cdot\tau\cdot\exp(-\beta\tau^2)\cdot\exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\ &= \left(\tau^2\right)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{(x-\mu)^2}{2}\right)\tau^2\right)\in\mathcal{P} \end{aligned}$$

Também observe que $f(\tau^2)\cdot f(x|\tau^2)$ é proporcional à densidade de uma Gamma com parâmetros $\alpha+\frac{1}{2}$ e $\left(\beta+\frac{(x-\mu)^2}{2}\right)$. Portanto,

$$\tau^2|X\sim\text{Gamma}\left(\alpha+\frac{1}{2},\beta+\frac{(x-\mu)^2}{2}\right).$$

□

Finalmente, prosseguimos ao caso em que tanto μ quanto τ^2 são desconhecidos

Lema 4.40. *Se $X|\mu,\tau^2\sim N(\mu,\tau^2)$, então*

$$\mathcal{P}=\left\{f:\mathbb{R}\to\mathbb{R}^+ : f(\mu)=K\cdot(\tau^2)^{\alpha-1}\exp(-\beta\tau^2)\exp\left(-\frac{\lambda\tau^2(\mu-\mu_0)^2}{2}\right)\right\}$$

é conjugada de $f(x|\mu,\tau^2)$. As densidades em $\mathcal{P}$ pertencem à família bivariada Normal-Gamma com parâmetros $(\mu_0,\lambda,\alpha,\beta)$. Note que se (μ,τ^2) tem distribuição Normal-Gamma, então μ e τ^2 não são independentes. De fato, se $(\mu,\tau^2)\sim$ Normal-Gamma$(\mu_0,\lambda,\alpha,\beta)$, então temos:

$$\begin{aligned} \tau^2 &\sim \textit{Gamma}(\alpha,\beta) \\ \mu|\tau^2 &\sim \textit{Normal}(\mu_0,\lambda\tau^2) \end{aligned}$$

Também, se $(\mu, \tau^2) \sim$ *Normal-Gamma*$(\alpha, \beta, \mu_0, \lambda)$*, então*

$$(\mu, \tau^2)|X \sim \textit{Normal-Gamma}\left(\frac{(\lambda\mu_0 + X)}{\lambda+1}, \lambda+1, \alpha+\frac{1}{2}, \beta+\frac{\lambda(\mu_0-X)^2}{2(\lambda+1)}\right)$$

Demonstração. Note que, nos passos a seguir, tanto μ quanto τ^2 são desconhecidos. Assim, expressões que dependam de quaisquer destes parâmetros não são constantes.

$$\begin{aligned}
f(\mu, \tau^2|x) &\propto f(\mu, \tau^2) f(x|\mu, \tau^2) \\
&= K \cdot (\tau^2)^{\alpha-1} \exp(-\beta\tau^2) \exp\left(-\frac{\lambda\tau^2(\mu-\mu_0)^2}{2}\right) \cdot \frac{\tau}{\sqrt{2\pi}} \exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\
&\propto (\tau^2)^{\alpha+\frac{1}{2}-1} \cdot \exp(-\beta\tau^2) \cdot \exp\left(-\frac{\lambda\tau^2(\mu^2-2\mu\mu_0+\mu_0^2)}{2}\right) \cdot \exp\left(-\frac{\tau^2(x^2-2x\mu+\mu^2)}{2}\right) \\
&= (\tau^2)^{\alpha+\frac{1}{2}-1} \cdot \exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2\right) \cdot \exp\left(-\frac{\lambda\tau^2\mu^2-2\lambda\tau^2\mu\mu_0+\tau^2\mu^2-2\tau^2x\mu}{2}\right) \\
&= (\tau^2)^{\alpha+\frac{1}{2}-1} \cdot \exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2\right) \cdot \exp\left(-\frac{(\lambda+1)\tau^2\mu^2-2(\lambda\tau^2\mu_0+\tau^2x)\mu}{2}\right) \qquad (11)
\end{aligned}$$

Similarmente ao caso em que τ^2 é conhecido, desejamos completar o quadrado em função de μ para obter o formato da distribuição normal. Usamos novamente a eq. (9) para obter

$$\begin{cases} a^2 &= (\lambda+1)\tau^2 \\ 2ab &= 2(\lambda\tau^2\mu_0+\tau^2x) \end{cases}$$

e obtemos

$$\begin{cases} a &= \sqrt{\lambda+1}\cdot\tau \\ b &= \frac{\tau(\lambda\mu_0+x)}{\sqrt{\lambda+1}} \end{cases} \qquad (12)$$

Substituindo a eq. (12) em eq. (11), obtemos:

$$
\begin{aligned}
f(\mu,\tau^2|x) &\propto (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2\right)\cdot\exp\left(-\frac{a^2\mu^2-2ab\mu}{2}\right)\\
&= (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2\right)\cdot\exp\left(-\frac{a^2\mu^2-2ab\mu+b^2}{2}\right)\cdot\exp\left(\frac{b^2}{2}\right)\\
&= (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2+\frac{b^2}{2}\right)\cdot\exp\left(-\frac{a^2\left(\mu-\frac{b}{a}\right)^2}{2}\right)\\
&= (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2+\frac{\tau^2(\lambda\mu_0+x)^2}{2(\lambda+1)}\right)\cdot\exp\left(-\frac{a^2\left(\mu-\frac{b}{a}\right)^2}{2}\right)\\
&= (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{x^2}{2}+\frac{\lambda\mu_0^2}{2}\right)\tau^2+\frac{\tau^2(\lambda\mu_0+x)^2}{2(\lambda+1)}\right)\cdot\exp\left(-\frac{a^2\left(\mu-\frac{b}{a}\right)^2}{2}\right)\\
&= (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{\lambda(\mu_0-x)^2}{2(\lambda+1)}\right)\tau^2\right)\cdot\exp\left(-\frac{a^2\left(\mu-\frac{b}{a}\right)^2}{2}\right)\\
&= (\tau^2)^{\alpha+\frac{1}{2}-1}\cdot\exp\left(-\left(\beta+\frac{\lambda(\mu_0-x)^2}{2(\lambda+1)}\right)\tau^2\right)\cdot\exp\left(-\frac{(\lambda+1)\tau^2\left(\mu-\frac{(\lambda\mu_0+x)}{\lambda+1}\right)^2}{2}\right)\in\mathcal{P}
\end{aligned}
$$

Do resultado acima, podemos concluir que

$$
(\mu,\tau^2)|X\sim\text{Normal-Gamma}\left(\frac{(\lambda\mu_0+X)}{\lambda+1},\lambda+1,\alpha+\frac{1}{2},\beta+\frac{\lambda(\mu_0-X)^2}{2(\lambda+1)}\right)
$$

□

Exemplo 4.41. Considere novamente o Exemplo 4.38, mas vamos agora assumir que a altura de um brasileiro selecionado ao acaso, X, é tal que $X|\mu\sim N(\mu,\tau^2)$, com τ^2 desconhecido. Agora precisamos fazer a inferência simultaneamente para μ e τ^2 Para tanto, vamos assumir que $(\mu,\tau^2)\sim\text{Normal-Gamma}(1.65,1,6,0.05)$. $X=1.8$, então $(\mu,\tau^2)|x\sim\text{Normal-Gamma}(1.725,2,6.5,0.055625)$. A Figura **??** mostra a distribuição a priori e a distribuição a posteriori para essa observação.

```
library(ggplot2)

grid_bivariado <- expand.grid(mu=seq(1.4,1.8,length.out = 300),
                              tau2=seq(60,200,length.out = 300))
```

Exercícios

Exercício 4.42. Considere que, dado μ, $X_1,\ldots,X_n$ são i.i.d. e $X_i\sim N(\mu,\tau^2)$, com τ^2 conhecido. Se $\mu\sim N(\mu_0,\tau_0^2)$, então:

(a) Ache a posteriori para $\mu|X_1,\ldots,X_n$.

(b) Derive $\lim_{n\to\infty}\mathbb{E}[\mu|X_1,\ldots,X_n]$.

(c) Derive $\lim_{n\to\infty}\mathbb{V}[\mu|X_1,\ldots,X_n]$.

(d) Interprete em suas próprias palavras os resultados obtidos nos itens anteriores.

Exercício 4.43. Considere que, dado τ^2, $X_1, \ldots, X_n$ são i.i.d. e $X_i \sim N(\mu, \tau^2)$, com μ conhecido. Considere que $\tau^2 \sim \text{Gamma}(\alpha, \beta)$.

(a) Ache a posteriori para $\tau^2 | X_1, \ldots, X_n$.

(b) Derive $\lim_{n\to\infty} \mathbb{E}[\tau^2 | X_1, \ldots, X_n]$.

(c) Interprete em suas próprias palavras os resultados obtidos nos itens anteriores.

4.3.3 A família exponencial

A família exponencial é uma generalização de diversas famílias de distribuições.

Definição 4.44. Dizemos que, dado um vetor de parâmetros θ, um vetor de dados, X, tem distribuição pertencente à família exponencial se

$$f(x|\theta) = h(x) \exp\left(g(\theta) \cdot T(x) - A(\theta)\right)$$

onde $g(\theta)$ e $T(x)$ são funções multivariadas dos parâmetros e dos dados.

Note que, para cada valor de θ, $A(\theta)$ é o valor que faz com que $f(x|\theta)$ integre 1. Assim, $A(\theta)$ é completamente especificado em função dos demais elementos da família exponencial.

Exemplo 4.45. Considere que $X|\theta \sim \text{Binomial}(n, \theta)$.

$$\begin{aligned} f(x|\theta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &= \binom{n}{x} \exp\left(x \log(\theta) + (n-x)\log(1-\theta)\right) \\ &= \binom{n}{x} \exp\left(x \cdot \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right) \end{aligned}$$

Portanto, $f(x|\theta)$ pertence à família exponencial, tomando-se

$$\begin{cases} h(x) & = \binom{n}{x} \\ T(X) & = x \\ g(\theta) & = \log\left(\frac{\theta}{1-\theta}\right) \\ A(\theta) & = -n\log(1-\theta) \end{cases}$$

Exemplo 4.46. Considere que $\theta = (\mu, \tau^2)$ e $X|\theta \sim \text{Normal}(\mu, \tau^2)$.

$$\begin{aligned} f(x|\mu,\tau^2) &= \frac{\tau}{\sqrt{2\pi}} \exp\left(-\frac{\tau^2(x-\mu)^2}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\tau^2x^2 - 2\tau^2x\mu + \tau^2\mu^2}{2} + \log(\tau)\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left((x, -x^2/2) \cdot (\tau^2\mu, \tau^2) - \left(\frac{\tau^2\mu^2}{2} - \log(\tau)\right)\right) \end{aligned}$$

Portanto, $f(x|\theta)$ pertence à família exponencial, tomando-se

$$\begin{cases} h(x) & = \frac{1}{\sqrt{2\pi}} \\ T(X) & = (x, -x^2/2) \\ g(\theta) & = (\tau^2\mu, \tau^2) \\ A(\theta) & = \left(\frac{\tau^2\mu^2}{2} - \log(\tau)\right) \end{cases}$$

Muitas outras distribuições pertencem à família exponencial. Por exemplo, a Binomial-Negativa, a Poisson, a Multinomial, a Hipergeométrica, a Geométrica, a log-Normal, a Exponencial, a Gamma, a Gamma-Inversa, a Beta, a Weibull e a Laplace.

Além de incluir diversas distribuições, a família exponencial também apresenta propriedades importantes. No contexto desta seção, podemos derivar a família conjugada para um membro da família exponencial.

Lema 4.47. *Se $f(x|\theta)$ faz parte da família exponencial, isto é, $f(x|\theta) = h(x)\exp\left(g(\theta)\cdot T(x) - A(\theta)\right)$, então*

$$\mathcal{P} = \{f(\theta) : f(\theta) = K\cdot\exp\left(-\alpha A(\theta) + g(\theta)\cdot\beta\right)\}$$

é conjugada a $f(x|\theta)$.

Demonstração. Note que

$$\begin{aligned} f(\theta)f(x|\mu) &\propto \exp\left(-\alpha A(\theta) + g(\theta)\cdot\beta\right)\cdot h(x)\cdot\exp\left(g(\theta)\cdot T(x) - A(\theta)\right) \\ &\propto \exp\left(-(\alpha+1)A(\theta) + g(\theta)\cdot(\beta + T(x))\right) \in \mathcal{P} \end{aligned}$$

Portanto, $\mathcal{P}$ é conjugada de $f(x|\theta)$. Similarmente, para o caso em que $X_1, \ldots, X_n$ são i.i.d. dado θ, obtemos

$$\begin{aligned} f(\theta)f(x_1,\ldots,x_n|\mu) &= f(\theta)\prod_{i=1}^{n} f(x_i|\theta) \\ &\propto \exp\left(-\alpha A(\theta) + g(\theta)\cdot\beta\right)\cdot\prod_{i=1}^{n}\exp\left(g(\theta)\cdot T(x_i) - A(\theta)\right) \\ &\propto \exp\left(-(\alpha+n)A(\theta) + g(\theta)\cdot\left(\beta + \sum_{i=1}^{n} T(x_i)\right)\right) \in \mathcal{P} \end{aligned}$$

□

Exemplo 4.48. Considere que $X|\theta \sim \text{Binomial}(n,\theta)$. Vimos no Exemplo 4.45 que $f(x|\theta)$ pertence à família exponencial. Portanto, temos que

$$\mathcal{P} = \{f(\theta) = K\cdot\exp\left(-\alpha A(\theta) + g(\theta)\cdot\beta\right)\}$$

é conjugada de $f(x|\theta)$. Substituindo os termos encontrados no Exemplo 4.45, obtemos

$$\begin{aligned} \mathcal{P} &= \left\{f(\theta) = K\cdot\exp\left(\alpha n\log(1-\theta) + \log\left(\frac{\theta}{1-\theta}\right)\cdot\beta\right)\right\} \\ &= \{f(\theta) = K(1-\theta)^{n\alpha-\beta}\theta^{\beta}\} \end{aligned}$$

Assim encontramos que a família Beta é conjugada da Binomial, assim como na seção 4.3.1. Observe que, para que seja possível obter $\int f(\theta)d\theta = 1$, é necessário que $n\alpha > \beta$ e $\beta > 0$.

Ao aplicar o Lema 4.47, nem sempre é imediato quais valores de α e β são tais que $f(\theta)$ é integrável. No Exemplo 4.48 verificamos que, para obter esta condição, era necessário que $n\alpha > \beta$ e $\beta > 0$. Esta análise é generalizada por um teorema em Diaconis and Ylvisaker (1979) que é descrito a seguir.

Teorema 4.49. *Considere que $X \in \chi$, $\theta \in \mathbb{R}^d$ e $f(x|\theta)$ está na forma canônica da família exponencial, isto é, $f(x|\theta) = h(x)\exp(\theta \cdot T(x) - A(\theta))$. A função $f(\theta) = \exp(-\alpha A(\theta) + \theta \cdot \beta)$ é integrável se e somente se $\alpha > 0$ e $\frac{\beta}{\alpha} \in Interior[Conv[T[\chi]]]$.*

Exemplo 4.50. Considere que $X|\theta \sim \text{Poisson}(\theta)$. Obtemos,

$$\begin{aligned} f(x|\theta) &= \frac{\exp(-\theta)\theta^x}{x!} \\ &= (x!)^{-1}\exp(\log(\theta)\cdot x - \theta) \end{aligned}$$

Assim, definindo $\eta = \log(\theta)$, obtemos:

$$f(x|\eta) = (x!)^{-1}\exp(\eta \cdot x - \exp(\eta))$$

Conclua que $f(x|\eta)$ está na forma canônica da familia exponencial, tomando

$$\begin{cases} h(x) &= (x!)^{-1} \\ T(x) &= x \\ A(\eta) &= \exp(\eta) \end{cases}$$

Portanto, decorre do Lema 4.47 que a seguinte familia é conjugada para $f(x|\eta)$

$$\mathcal{P} = \{f(\eta) : f(\eta) \propto \exp(-\alpha\exp(\eta) + \eta\cdot\beta)\}$$

Ademais, podemos aplicar o Teorema 4.49 para obter os valores de α e β tais que $\exp(-\alpha\exp(\eta) + \eta\cdot\beta)$ é integrável. Obtemos que $\exp(-\alpha\exp(\eta) + \eta\cdot\beta)$ é integrável se e somente se $\alpha > 0$ e $\frac{\beta}{\alpha} \in \text{Interior}[\text{Conv}[T[\chi]]]$. Como $X|\theta \sim \text{Poisson}(\theta)$, obtemos $\chi = \mathbb{N}$. Assim, como $T(x) = x$, $T[\mathbb{N}] = \mathbb{N}$. Ademais, $\text{Conv}[\mathbb{N}] = \mathbb{R}^+$. Finalmente, $\text{Interior}[\mathbb{R}^+] = \mathbb{R}^+_*$. Portanto, $\text{Interior}[\text{Conv}[T[\chi]]] = \mathbb{R}^+_*$. Note que, se $\alpha > 0$, então $\frac{\beta}{\alpha} \in \mathbb{R}^+_*$ se e somente se $\beta > 0$. Portanto, conclua do Teorema 4.49 que $\exp(-\alpha\exp(\eta)+\eta\cdot\beta)$ é integrável se e somente $\alpha > 0$ e $\beta > 0$. Tomando a transformação $\log(\theta) = \eta$, obtemos que $\left|\frac{d\log(\theta)}{d\theta}\right|\exp(-\alpha\theta + \log(\theta)\cdot\beta)$ é integrável se e somente se $\alpha > 0$ e $\beta > 0$. Assim,

$$\mathcal{P}^* = \left\{f(\theta) : f(\theta) \propto \theta^{\beta-1}\exp(-\alpha\theta), \alpha > 0, \beta > 0\right\}$$

é uma familia de distribuições integráveis. Ademais, decorre do Lema 4.47 que $\mathcal{P}$ é conjugada para $f(x|\theta)$. Note que as densidades em $\mathcal{P}^*$ correspondem à família de distribuições Gamma.

Exercícios

Exercício 4.51. Escolha duas de suas famílias de distribuições favoritas e descubra se elas pertencem ou não à família exponencial.

Exercício 4.52. Se $X|\theta \sim \text{Geom}(\theta)$:

(a) Mostre que $f(x|\theta)$ pertence à família exponencial.

(b) Ache a família conjugada para $f(x|\theta)$.

(c) Ache a posteriori para θ quando a priori é conjugada.

Exercício 4.53. Se $X|\theta \sim \text{Exponencial}(\theta)$:

(a) Mostre que $f(x|\theta)$ pertence à família exponencial.

(b) Ache a família conjugada para $f(x|\theta)$.

(c) Ache a posteriori para θ quando a priori é conjugada.

Exercício 4.54. Em uma população, a proporção de indivíduos com uma determinada doença é θ. Considere que uma amostra de 100 indivíduos é tomada da população. Para cada indivíduo, i, defina Z_i como sendo a indicadora de que o indivíduo tem a doença. Um teste é realizado em cada um dos indivíduos da amostra. Este teste é tal que, se o indivíduo for doente, o teste acusa afirmativo certamente. Contudo, se o indivíduo não for doente, há uma probabilidade 0.1 de um falso positivo. Para cada indivíduo, i, defina X_i como sendo a indicadora de que o resultado do exame para o indivíduo i foi positivo. Observou-se que 30 indivíduos obtiveram o resultado positivo no teste. Note que as variáveis Z_i não foram observadas.

(a) Note que $X_1|\theta$ é uma Bernoulli. Use a lei da probabilidade total para mostrar que

$$X_1|\theta \sim \text{Bernoulli}(0.1 + 0.9\theta)$$

(b) Ache a distribuição de $\sum_{i=1}^{100} X_i|\theta$ e prove que ela pertence à família exponencial.

(c) Ache a família conjugada para $\sum_{i=1}^{100} X_i|\theta$.

(d) Tome uma priori na família conjugada para θ e ache a distribuição de $\theta|\sum_{i=1}^{100} X_i = 30$.

Exercício 4.55. Considere que $X|\theta \sim \text{Bernoulli}(\theta)$.

(a) Reparametrize esta distribuição de $X|\theta$ para que se enquadra na forma canônica da família exponencial.

(b) Determine uma família conjugada para a reparametrização utilizando o Lema 4.47.

(c) Utilize o Teorema 4.49 para determinar as densidades na família encontrada no item acima.

4.3.4 O processo de Dirichlet*

Nas seções anteriores, consideramos que, dado um conjunto de parâmetros, θ, $X_1, \ldots, X_n$ eram i.i.d. com uma distribuição determinada por θ. Em outras palavras, para cada θ havia uma densidade f_θ tal que

$$f(x_1, \ldots, x_n | f_\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

Neste tipo de modelo, você atribuiu prioris sobre θ (e, portanto, sobre f_θ) que eram convenientes computacionalmente. Contudo, as possíveis distribuições que f_θ podia assumir eram limitadas. Por exemplo, na seção 4.3.2 f_θ era necessariamente uma distribuição normal com média e variância dadas por θ.

Contudo, em alguns casos você poderá querer que f_θ tenha como possíveis valores uma classe mais geral. Este é o tipo de problema que é estudado pela estatística não-paramétrica. A dificuldade da estatística não-paramétrica é obter uma classe de f_θ que seja geral e, ainda assim, conveniente computacionalmente. Uma maneira de obter este resultado é pelo processo de Dirichlet, que veremos a seguir.

Para definir o processo de Dirichlet, ao invés de f_θ, consideraremos uma função de probabilidade aleatória sobre $\mathbb{R}$, P_θ. Note que P_θ é uma função aleatória, isto é, para cada $A \subset \mathbb{R}$, $P_\theta(A)$ é uma variável aleatória. Consideraremos também que, dado P_θ, $(X_1, \ldots, X_n)$ são i.i.d. com distribuição dada por P_θ, isto é,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n | P_\theta) = \prod_{i=1}^{n} P_\theta(A_i) \tag{13}$$

O processo de Dirichlet é uma distribuição sobre P_θ que faz com o suporte de P_θ possa ser a familia de todas as distribuições univariadas. O processo de Dirichlet tem como parâmetros uma função de probabilidade sobre $\mathbb{R}$, P_0, e $\alpha \in \mathbb{R}_+$. Ele é definido da seguinte forma:

Definição 4.56. Dizemos que P_θ tem distribuição dada pelo processo de Dirichlet com parâmetros P_0 e α e escrevemos $P_\theta \sim \text{PD}(P_0, \alpha)$ se P_θ é uma função de probabilidade com probabilidade 1 e, também, para todo $d \in \mathbb{N}$ e partição finita de $\mathbb{R}$, $(B_i)_{1 \leq i \leq d}$,

$$(P_\theta(B_1), \ldots, P_\theta(B_d)) \sim \text{Dirichlet}(\alpha(P_0(B_1), \ldots, P_0(B_d)))$$

Frente a uma definição de processo estocástico como a Definição 4.56, existem algumas perguntas comumente feitas. Por exemplo, existe de fato um processo estocástico que satisfaça a Definição 4.56? Também, qual é a probabilidade de que P_θ seja efetivamente uma função de probabilidade sobre $\mathbb{R}$? Dentre outras referências, por exemplo (Ferguson, 1973, Sethuraman, 1994) mostram que existe um processo estocástico que satisfaz a Definição 4.56 e que, com probabilidade 1, P_θ é uma probabilidade sobre $\mathbb{R}$.

Para compreender o processo de Dirichlet, podemos calcular algumas de suas propriedades relevantes. Por exemplo, para cada $x \in \mathbb{R}$, podemos definir $F_\theta(x) := \mathbb{P}_\theta((-\infty, x])$ e $F_0(x) := \mathbb{P}_0((-\infty, x])$. Assim, $F_\theta(x)$ é uma variável aleatória e $F_0(x)$ é uma constante. Pela definição do processo de Dirichlet,

$$(P_\theta((-\infty, x]), P_\theta((x, \infty))) \sim \text{Dirichlet}(\alpha P_0((-\infty, x]), \alpha P_0((x, \infty)))$$

Dada a relação entra as distribuições Beta e a Dirichlet, decorre diretamente que

$$F_\theta(x) \sim \text{Beta}(\alpha P_0((-\infty, x]), \alpha P_0((x, \infty)))$$

Portanto, obtemos que, para cada $x \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[F_\theta(x)] &= F_0(x)\\ \mathbb{V}[F_\theta(x)] &= \frac{F_0(x)(1-F_0(x))}{1+\alpha}\end{aligned}$$

Assim, enquanto que P_0 representa a tendência central do processo de Dirichlet, α indica o quanto o processo está concentrado em torno de P_0. De fato, a distribuição marginal de X é dada por P_0

Lema 4.57. *Se $X|P_\theta \sim P_\theta$ e $P_\theta \sim DP(P_0, \alpha)$, então $X \sim P_0$.*

Demonstração.

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{E}[\mathbb{P}(X \leq x|P_\theta)]\\ &= \mathbb{E}[\mathbb{P}_\theta((-\infty, x])]\\ &= P_0((\infty, x])\end{aligned}$$

□

Além de ser uma priori abrangente para P_θ, o processo de Dirichlet também é conveniente computacionalmente. Esta propriedade é apresentada a seguir e acompanha a demonstração em (Ferguson, 1973).

Definição 4.58 (δ de Dirac)**.** Para cada $x \in \mathbb{R}$, defina δ_x tal que $\delta_x = \mathbb{I}(x \in A)$.

Teorema 4.59. *Se $P_\theta \sim DP(P_0, \alpha)$ e, dado P_θ, X tem distribuição P_θ, então $P_\theta|X \sim DP\left(\frac{\alpha P_0 + \delta_x}{\alpha+1}, \alpha+1\right)$.*

Para provar o Teorema 4.59, alguns resultados sobre a distribuição Dirichlet serão úteis. Estes podem ser provados usando técnicas comumente usadas para vetores de variáveis aleatórias e são enunciados a seguir.

Definição 4.60. Considere que $(Y_1, \ldots, Y_n) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_n)$. Denotamos a função de densidade acumulada de $(Y_1, \ldots, Y_n)$ por $\mathcal{D}(y_1, \ldots, y_n|\alpha_1, \ldots, \alpha_n)$.

Lema 4.61. *Se $(X_1, \ldots, X_n)$ são independentes, $X_i \sim Gamma(\alpha_i)$ e $S = \sum_{i=1}^n X_i$, então*

$$\frac{(X_1, \ldots, X_n)}{S} \sim Dirichlet(\alpha_1, \ldots, \alpha_n)$$

Lema 4.62. *Se $Y_1, \ldots, Y_n \sim Dirichlet$, então $(Y_1, \ldots, Y_{n-2}, Y_{n-1} + Y_n) \sim Dirichlet(\alpha_1, \ldots, \alpha_{n-2}, \alpha_{n-1} + \alpha_n)$.*

Lema 4.63. *Considere que $(Y_1, \ldots, Y_n) \sim Dirichlet(\alpha_1, \ldots, \alpha_n)$.*

$$\mathbb{E}\left[Y_n \mathbb{I}(Y_{n-1} + Y_n \leq y_{n-1}) \prod_{i<n-1} \mathbb{I}(Y_i \leq y_i)\right] = \frac{\alpha_n}{\sum_{i=1}^n \alpha_i} \mathcal{D}(y_1, \ldots, y_{n-1}|\alpha_1, \ldots, \alpha_{n-2}, \alpha_{n-1} + \alpha_n + 1)$$

Demonstração. Defina $C = \bigcap_{i<n-1}\{Y_i \leq y_i\} \cap \{Y_{n-1} + Y_n < y_{n-1}\} \cap \{\sum_{i=1}^n Y_i = 1\}$, $\alpha_i^* = \alpha_i + \mathbb{I}(i = n)$, e

$(Y_1^*, \ldots, Y_n^*) \sim \text{Dirichlet}(\alpha_1^*, \ldots, \alpha_n^*)$.

$$\begin{aligned}
\mathbb{E}\left[Y_n \mathbb{I}(Y_{n-1} + Y_n \leq y_{n-1}) \prod_{i<n-1} \mathbb{I}(Y_i \leq y_i)\right] &= \int_C y_n \Gamma\left(\sum_{i=1}^n \alpha_i\right) \prod_{i=1}^n \left(\frac{y_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}\right) d\mathbf{y} \\
&= \int_C \Gamma\left(\sum_{i=1}^n \alpha_i\right) \prod_{i=1}^n \left(\frac{y_i^{\alpha_i^* - 1}}{\Gamma(\alpha_i)}\right) d\mathbf{y} \\
&= \frac{\alpha_n}{\sum_{i=1}^n \alpha_i} \int_C \Gamma\left(\sum_{i=1}^n \alpha_i^*\right) \prod_{i=1}^n \left(\frac{y_i^{\alpha_i^* - 1}}{\Gamma(\alpha*_i)}\right) d\mathbf{y} \\
&= \frac{\alpha_n}{\sum_{i=1}^n \alpha_i} \mathbb{E}\left[\mathbb{I}(Y_{n-1}^* + Y_n^* \leq y_{n-1}) \prod_{i<n} \mathbb{I}(Y_i \leq y_i)\right] \\
&= \frac{\alpha_n}{\sum_{i=1}^n \alpha_i} \mathcal{D}(y_1, \ldots, y_{n-1} | \alpha_1, \ldots, \alpha_{n-2}, \alpha_{n-1} + \alpha_n + 1) \quad \text{Lema 4.62}
\end{aligned}$$

□

Lema 4.64. *Se $Y_1, \ldots, Y_n \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_n)$ e π é uma permutação de $\{1, \ldots, n\}$, então obtemos que $Y_{\pi(1)}, \ldots, Y_{\pi(n)} \sim$ Dirichlet$(\alpha_{\pi(1)}, \ldots, \alpha_{\pi(n)})$.*

Teorema 4.59. Seja $(B_i)_{1 \leq i \leq d}$ uma partição de $\mathbb{R}$ e $A \subset \mathbb{R}$. Defina $B_{i,0} = A^c \cap B_i$, $B_{i,1} = A \cap B_i$ e $\mathcal{I} = \{1, \ldots, d\} \times \{0, 1\}$. Note que $(B_{i,j})_{(i,j) \in \mathcal{I}}$ particiona $\mathbb{R}$. Também,

$$\begin{aligned}
\mathbb{E}[I(X_1 \in A) | P_\theta(B_{i,j})_{(i,j) \in \mathcal{I}}] &= \mathbb{E}\left[\sum_{k=1}^d I(X_1 \in B_{k,1}) | P_\theta(B_{i,j})_{(i,j) \in \mathcal{I}}\right] \\
&= \sum_{k=1}^d P_\theta(B_{k,1}) \qquad (14)
\end{aligned}$$

Assim,

$$\begin{aligned}
\mathbb{P}(X_1 \in A, P_\theta(B_i) \leq y_i, 1 \leq i \leq d) &= \mathbb{E}\left[\mathbb{I}(X \in A) \prod_{i=1}^d \mathbb{I}(P_\theta(B_i) \leq y_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(X \in A) \prod_{i=1}^d \mathbb{I}(P_\theta(B_i) \leq y_i) \middle| P_\theta(B_{i,j})_{(i,j) \in \mathcal{I}}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(X \in A) \middle| P_\theta(B_{i,j})_{(i,j) \in \mathcal{I}}\right] \prod_{i=1}^d \mathbb{I}(P_\theta(B_i) \leq y_i)\right] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^d P_\theta(B_{k,1})\right) \prod_{i=1}^d \mathbb{I}(P_\theta(B_i) \leq y_i)\right] \qquad \text{eq. (14)} \\
&= \sum_{k=1}^d \mathbb{E}\left[P_\theta(B_{k,1}) \prod_{i=1}^d \mathbb{I}(P_\theta(B_i) \leq y_i)\right] \qquad (15)
\end{aligned}$$

Para facilitar o raciocínio, defina $Y_i := P_\theta(B_i)$, $Y_{i,j} = P_\theta(B_{i,j})$ e $\alpha_i^* = \alpha_i + \mathbb{I}(i = k)$. Note que $Y_k = Y_{k,0} + Y_{k,1}$.

$$\begin{aligned}
\mathbb{E}\left[P_\theta(B_{k,1})\prod_{i=1}^{d}\mathbb{I}(P_\theta(B_i) \le y_i)\right] &= \mathbb{E}\left[Y_{k,1}\prod_{i=1}^{d}\mathbb{I}(Y_i \le y_i)\right] \\
&= \mathbb{E}\left[Y_{k,1}\mathbb{I}(Y_{k,0} + Y_{k,1} \le y_k)\prod_{i\neq k}\mathbb{I}(Y_i \le y_i)\right] \\
&= \frac{\alpha P_0(B_{k,1})}{\alpha(P_0(B_{k,0}) + P_0(B_{k,1}) + \sum_{i\neq k} P_0(B_i))}\mathcal{D}(y_1,\dots,y_d|\alpha_1^*,\dots,\alpha_d^*) \quad \text{Lema 4.63} \\
&= P_0(A\cap B_k)\mathcal{D}(y_1,\dots,y_d|\alpha_1^*,\dots,\alpha_d^*) \qquad (16)
\end{aligned}$$

Juntando-se eqs. (15) e (16), obtem-se:

$$\mathbb{P}(X_1 \in A, P_\theta(B_i) \le y_i, 1 \le i \le d) = \sum_{i=1}^{d} P_0(A\cap B_i)\mathcal{D}(y_1,\dots,y_d|\alpha_1 + \mathbb{I}(i=1),\dots,\alpha_d + \mathbb{I}(i=d)) \qquad (17)$$

Note que $\mathbb{P}(P_\theta(B_1) \le y_1,\dots,P_\theta(B_d) \le y_d|X)$ é definido como a função tal que, para todo A,

$$\int_A \mathbb{P}(P_\theta(B_1) \le y_1,\dots,P_\theta(B_d) \le y_d|x)dF_X(x) = \mathbb{P}(X_1 \in A, P_\theta(B_1) \le y_1,\dots P_\theta(B_d) \le y_d) \qquad (18)$$

Portanto, para completar a demonstração, basta utilizar a $DP\left(\frac{\alpha P_0+\delta_x}{\alpha+1}, \alpha+1\right)$ no lado esquerdo de eq. (18) e chegar à expressão para $\mathbb{P}(X_1 \in A, P_\theta(B_1) \le y_1,\dots P_\theta(B_d) \le y_d)$ obtida em eq. (17).

$$\begin{aligned}
&= \int_A \mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \delta_x(B_1),\dots,\alpha P_0(B_d) + \delta_x(B_d))dF_X(x) \\
&= \int_A \mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \delta_x(B_1),\dots,\alpha P_0(B_d) + \delta_x(B_d))dP_0(x) \qquad \text{Lema 4.57} \\
&= \sum_{i=1}^{d}\int_{B_{i,1}} \mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \delta_x(B_1),\dots,\alpha P_0(B_d) + \delta_x(B_d))dP_0(x) \qquad \cup_{i=1}^{d} B_{i,1} = A \\
&= \sum_{i=1}^{d}\int_{B_{i,1}} \mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \mathbb{I}(i=1),\dots,\alpha P_0(B_d) + \mathbb{I}(i=d))dP_0(x) \\
&= \sum_{i=1}^{d} \mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \mathbb{I}(i=1),\dots,\alpha P_0(B_d) + \mathbb{I}(i=d))\int_{B_{i,1}} dP_0(x) \\
&= \sum_{i=1}^{d} \mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \mathbb{I}(i=1),\dots,\alpha P_0(B_d) + \mathbb{I}(i=d))P_0(B_{i,1}) \\
&= \sum_{k=1}^{d} P_0(A\cap B_i)\mathcal{D}(y_1,\dots,y_d|\alpha P_0(B_1) + \mathbb{I}(i=1),\dots,\alpha P_0(B_d) + \mathbb{I}(i=d))
\end{aligned}$$

A demonstração decorre diretamente de eq. (17). □

Teorema 4.65. *Se $P_\theta \sim DP(P_0, \alpha)$ e, dado P_θ, $X_1,\dots,X_n$ são i.i.d. e tem distribuição P_θ, então temos que $P_\theta|X_1,\dots,X_n \sim DP\left(\frac{\alpha P_0+\sum_{i=1}^{n}\delta_{x_i}}{\alpha+n}, \alpha+n\right)$.*

Demonstração. Basta utilizar o Teorema 4.59 iterativamente. □

Dada a complexidade do Processo de Dirichlet, é essencial saber simular deste. Uma forma de obter este objetivo é introduzida por Sethuraman (1994) e discutida a seguir.

Teorema 4.66 ("Stick-breaking process")**.** *Considere que $(Y_n)_{n\in N}$ são i.i.d., $Y_i \sim P_0$, $(\theta_n)_{n\in N}$ são i.i.d., $\theta_i \sim Beta(1,\alpha)$, e $\mathbf{Y}$ e θ são independentes. Defina $p_i = \theta_i \prod_{j<i}(1-\theta_j)$. Se $P_\theta = \sum_{i=1}^{\infty} p_i \delta_{Y_i}$, então $P_\theta \sim DP(P_0,\alpha)$.*

Lema 4.67. *Considere que $U \in \mathbb{R}^d$, $V \in R^d$, $W \in (-1,1)$ e (U,W) é independente de V. Existe uma única distribuição de V tal que $V \sim U + WV$, isto é, V e $U+WV$ tem mesma distribuição.*

Demonstração. Considere que F_{V_1} e F_{V_2} são duas distribuições e V_1 e V_2 são variáveis independentes tais que $V_i \sim F_{V_i}$ e $V_i \sim U + WV_i$. Defina $(U_n, W_n)_{n\in N}$ como i.i.d. e tais que (U_i, W_i) tem mesma distribuição de (U,W). Defina $V_{i,1} = V_i$ e $V_{i,n} = U_{n-1} + W_{n-1}V_{i,n-1}$. Decorre da relação proposta que $V_{i,n} \sim V_i$, para todo n. Note que

$$\begin{aligned}
|V_{1,n+1} - V_{2,n+1}| &= |U_n + W_n V_{1,n} - U_n - W_n V_{2,n}| \\
&= |W_n||V_{1,n} - V_{2,n}| \\
&= \prod_{i=1}^{n} |W_i||V_1 - V_2| \xrightarrow{a.s.} 0
\end{aligned}$$

Conclua que $F_{V_1} = F_{V_2}$. □

Lema 4.68. *Considere que P_θ é tal qual definida no Teorema 4.66. Para toda partição de $\mathbb{R}$, $B_1,\ldots,B_d$, se $V = (P_\theta(B_1),\ldots,P_\theta(B_d))$, $Y_0 \sim P_0$, $W \sim Beta(1,\alpha)$, $X = (\delta_{Y_0}(B_1),\ldots\delta_{Y_0}(B_d))$, $U = (1-W)X$, e (Y_0,V,W) são independentes então $V \sim U + WV$.*

Demonstração. Defina $\theta_1^* = W$, $\theta_n^* = \theta_{n-1}$. Por construção $(\theta_n^*)_{n\in N}$ são i.i.d. e $\theta_i^* \sim \text{Beta}(1,\alpha)$. Defina $Y_1^* = Y_0$, $Y_n^* = Y_{n-1}$, $p_1^* = W$, e $p_n^* = (1-W)p_{n-1}$. Note que

$$\begin{aligned}
W\delta_Y + (1-W)P_\theta &= \theta_1^*\delta_{Y_1^*} + (1-W)\sum_{n=1}^{\infty} p_n \delta_{Y_n} \\
&= \theta_1^*\delta_{Y_1^*} + \sum_{n=2}^{\infty} p_n^*\delta_{Y_n^*} \\
&= \sum_{n=1}^{\infty} p_n^*\delta_{Y_n^*}
\end{aligned}$$

onde $p_n^* = \theta_n^* \prod_{i=1}^{n-1}(1-\theta_i^*)$ e $(Y_n)_{n\in N}$ são i.i.d. P_0. Portanto, $P_\theta \sim W\delta_Y + (1-W)P_\theta$. □

Lema 4.69. *Considere que $B_1,\ldots,B_d$ é uma partição de $\mathbb{R}$, $W \sim Beta(1,\alpha)$, $Y_0 \sim P_0$, $X = (\delta_{Y_0}(B_1),\ldots\delta_{Y_0}(B_d))$, e $U = (1-W)X$. Se $V = (P_\theta(B_1),\ldots,P_\theta(B_d))$, $V \sim Dirichlet(\alpha P_0(B_1),\ldots\alpha P_0(B_d))$ e e (Y_0,V,W) são independentes, então então $V \sim U + (1-W)V$.*

Demonstração. □

Exercícios

Exercício 4.70 (Distribuição Dirichlet)**.**

(a) Prove o Lema 4.61.

(b) Prove o Lema 4.62 usando o Lema 4.61.

(c) Prove o Lema 4.64.

Exercício 4.71. Prove que Definição 4.56 é uma caracterização de P_θ. Isto é, se $P_{1,\theta}$ e $P_{2,\theta}$ satisfazem Definição 4.56, então eles tem a mesma distribuição.

Exercício 4.72. Seja $A_{n,x} = \{y : |y - x| < n^{-1}\}$, $X|P_\theta \sim P_\theta$ e $P_\theta \sim DP(P_0, \alpha)$. Argumente informalmente que $\mathbb{P}(P_\theta(B_1), \ldots, P_\theta(B_d)|X \in A_{n,x})$ converge para $\mathbb{P}(P_\theta(B_1), \ldots, P_\theta(B_d)|X)$ quando $n \to \infty$. Este exercício nos dá uma ideia de como (Ferguson, 1973) pode ter obtido a intuição de qual era a posteriori correta no Teorema 4.59.

Exercício 4.73. Se $V_{i,n} \sim F_i$ e $|V_{1,n} - V_{2,n}| \xrightarrow{a.s.} 0$, mostre que $F_1 = F_2$.

5 Revisão sobre o teorema de Bayes e o modelo estatístico

Exercício 5.1. A proporção de mulheres em um país é aproximadamente 50%. Considere que a distribuição da altura das mulheres e dos homens seguem distribuições normais com médias conhecidas, respectivamente, 165 e 170 e variâncias iguais a 9. Se a altura de uma pessoa selecionada aleatoriamente é 167, qual é a probabilidade de que ela seja uma mulher?

Exercício 5.2. Considere que $\mu \in \{\mu_1, \mu_2\}$, onde $\mu_1, \mu_2 \in \mathbb{R}^n$ e Σ^2 é uma matriz de variância conhecida. Também, $X|\mu \sim N(\mu, \Sigma^2)$ e $\mathbb{P}(\mu = \mu_1) = 0.5$.

(a) Ache $\mathbb{P}(\mu = \mu_1|X = x)$.

(b) Mostre que $\{x : P(\mu = \mu_1|X = x) = 0.5\}$ é um hiperplano.

Exercício 5.3. A proporção de deputados aliados ao governo em um determinado país é aproximadamente 60%. Para cada projeto de lei, o deputado pode votar contra o projeto, a seu favor ou se abster da votação. Se um deputado é aliado ao governo, a probabilidade de que ele vote a favor de cada lei é 70%, de que ele se abstenha é 20% e de que vote contra é 10%. Similarmente, se o deputado não é aliado ao governo, a probabilidade de que ele vote a favor de cada lei é 40%, de que ele se abstenha é 10% e de que vote contra é 50%. Se um deputado selecionado aleatoriamente votou a favor de 2 projetos, se absteve de 1 projeto e votou contra 1 projeto, qual é a probabilidade de que ele seja aliado ao governo?

Exercício 5.4. Considere que dado θ, $X_1 \ldots X_n$ são i.i.d. e $X_i|\theta \sim \text{Uniforme}(0, \theta)$. Considere que, a priori, temos que, para $\alpha, \beta > 0$,

$$f(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}\mathbb{I}(\theta)_{(\beta,\infty)}.$$

(a) Ache a posteriori para $\theta|X_1, \ldots, X_n$. Ache a forma exata da posteriori, não basta indicá-la até uma constante de proporcionalidade.

(b) Calcule $\lim_{n\to\infty} \mathbb{E}[\theta|X_1, \ldots, X_n]$.

(c) Ache $f(x_n|x_1, \ldots, x_{n-1})$. Lembre-se que esta é a distribuição preditiva, que não depende de θ.

Exercício 5.5. Considere que $\beta > 0$ é conhecido e, dado α, $X_1, \ldots, X_n$ são i.i.d. e $X_1 \sim \text{Pareto}(\alpha, \beta)$. Ou seja, $f(x_1|\alpha) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}\mathbb{I}_{(\beta,\infty)}(x)$.

(a) Mostre que $f(x_1|\alpha)$ faz parte da família exponencial.

(b) Se α fosse conhecido e β desconhecido, $f(x_1|\beta)$ faria parte da família exponencial?

(c) Se $\alpha \sim \text{Gamma}(\gamma, \delta)$, identifique o nome e os hiperparâmetros da posteriori, $f(\alpha|x_1, \ldots, x_n)$.

(d) Ache $\lim_{n\to\infty} \mathbb{E}[\alpha|X_1, \ldots, X_n]$.

(e) Ache uma família conjugada para $f(x_1|\alpha)$.

Exercício 5.6. Considere que foram colocadas 3 bolas em uma urna, sendo todas elas azuis ou verdes. Você está interessada em determinar o número de bolas azuis na urna e, a priori, este número pode assumir qualquer um dos valores possíveis com mesma probabilidade. Também considere que esta é uma urna de Polya, ou seja, a cada vez que uma bola é retirada, duas bolas da mesma cor da bola retirada são repostas na urna. Por exemplo, se uma urna de Polya tem 2 bolas azuis e 2 verdes e uma bola azul é retirada, então a composição passará a ser 3 bolas azuis e 2 verdes. Duas bolas foram retiradas da urna, sendo que suas cores foram, respectivamente: azul e azul.

(a) Identifique os elementos do modelo estatístico e os valores que estes podem assumir.

(b) Dado cada possível valor para o parâmetro, calcule a covariância entre a indicadora de que a primeira bola retirada é azul e a indicadora de que a segunda é azul. As observações são independentes dado o parâmetro?

(c) Calcule a distribuição a posteriori para o parâmetro do modelo.

(d) Calcule a probabilidade preditiva de que a próxima bola retirada seja azul.

(e) Calcule a probabilidade marginal de observar os dados.

Exercício 5.7. Considere que $Y_i = \theta x_i + \epsilon_i$, onde ϵ_i são i.i.d. e $\epsilon_1 \sim N(0, 1)$. Ou seja, dado θ, $Y_i \sim N(\theta x_i, 1)$. Considere que, a priori, $\theta \sim N(0, 1)$.

(a) Ache a posteriori para $\theta|(x_1, y_1), \ldots, (x_n, y_n)$.

(b) Ache $\lim_{n\to\infty} \mathbb{E}[\theta|(x_1, y_1), \ldots, (x_n, y_n)]$.

Exercício 5.8. Considere que, dado ν, $X \sim \text{N}_n(\mu, \nu)$, ou seja, X tem distribuição normal multivariada com média μ (conhecida) e precisão ν. Se, a priori, $\nu \sim \text{Wishart}(V, a)$, ache o nome da distribuição de $\nu|X$ e seus hiperparâmetros.

Exercício 5.9 (Gelman et al. (2014))**.** X_1 e X_2 são condicionalmente i.i.d. dado θ. Mostre que, se $\mathbb{V}[\mathbb{E}[X_1|\theta]] > 0$, então $Cov(X_1, X_2) > 0$.

Exercício 5.10. Seja $T(X)$ uma estatística suficiente para θ. Mostre que $f(\theta|T = t(x)) = f(\theta|X = x)$. Interprete esse resultado.

6 Tomando decisões conscientemente

Você está constantemente tomando decisões. Ir ou não ir à próxima aula de Estatística Bayesiana? O que comer no jantar? Trabalhar ou se divertir? Se você pensar com cuidado, geralmente será capaz de identificar diversas alternativas para cada uma de suas ações. A sua ação foi uma decisão dentre todas as alternativas possíveis.

Mesmo assim, apesar de sua ação ter alternativas, você nem sempre pensa conscientemente sobre qual será sua decisão. Seja por falta de tempo, seja por não conhecer outra forma de fazê-lo, muitas vezes você toma decisões intuitivamente, alheia aos motivos que tornam a sua escolha melhor ou pior do que as alternativas.

Contudo, nossa intuição muitas vezes não passaria por uma inspeção mais cuidadosa. Por exemplo, considere a decisão entre comer ou não comer uma barra de chocolate. Nós evoluímos de macacos em uma guerra permanente para sobreviver à fome. Destes mesmos macacos herdamos o impulso de avançar sobre a barra de chocolate e comê-la inteira. Contudo, a nossa situação é consideravelmente diferente da dos macacos. Temos maneiras seguras de reservar alimentos e ingerí-los mais tarde. Para muitos humanos, a guerra não é mais sobreviver à fome, mas sim à obesidade. Portanto, ainda que nosso primeiro impulso seja comer a barra de chocolate inteira, em muitas situações é possível elaborar argumentos que justificariam reservar parte ou a totalidade da barra de chocolate para ser comida mais tarde. Pensando com cuidado, às vezes você concluirá que a melhor decisão para você será diferente de sua intuição inicial.

Neste Capítulo você estudará um processo consciente de tomada de decisões. Chamaremos este processo de Teoria da Decisão. A Teoria da Decisão é dividida em etapas, que indicam questões relevantes na tomada de qualquer decisão. Quando você completar todas as etapas indicadas, a Teoria da Decisão sugerirá qual ação é a mais proveitosa para você.

6.1 Elementos da tomada de decisões

A Teoria da decisão lhe indicará qual é a melhor ação dentre aquelas que você tem disponíveis. Para tal, você terá de especificar determinados elementos, detalhados a seguir.

Denotaremos por $\mathcal{A}$ o conjunto de ações ou alternativas que você tem disponíveis. Essas alternativas deverão ser expressas de tal forma que sejam mutuamente exclusivas, ou seja, você somente poderá escolher uma única alternativa. Contudo, esta não é uma grande limitação. Por exemplo, se você pensar em alternativas A e B que não sejam mutuamente exclusivas, poderá obter "A e B", "A e não B" , "não A e B", e "não A e não B" como alternativas mutuamente exclusivas. Por exemplo, considere que você pode levar um guarda-chuva e uma calculadora à aula de Estatística Bayesiana. Ainda que "levar o guarda-chuva" e "levar a calculadora" não sejam alternativas mutuamente exclusivas, "levar o guarda-chuva e levar a calculadora", "levar o guarda-chuva e não levar a calculadora", "não levar o guarda-chuva e levar a calculadora", e "não levar o guarda-chuva e não levar a calculadora" o são. Sucintamente, denotaremos as alternativas neste exemplo por $\mathcal{A} = \{(gc, ca), (gc, \bar{ca}), (\bar{gc}, ca), (\bar{gc}, \bar{ca})\}$.

É importante que você inclua em $\mathcal{A}$ todas as alternativas relevantes. Se você esquecer de incluir a melhor decisão, então o procedimento descrito neste Capítulo não será capaz de indicá-la como sendo a melhor decisão. De fato, muitas vezes o segredo do protagonista de uma história de sucesso foi a capacidade deste de considerar uma alternativa que outros não consideraram. A Teoria da Decisão não enuncia diretamente qual é $\mathcal{A}$, mas reforça a importância de conscientemente analisar com cuidado este aspecto.

Denotamos por Θ o conjunto de possíveis ocorrências que são relevantes para a tomada da sua decisão. Similarmente às alternativas possíveis, as possibilidades também devem ser mutuamente exclusivas e exaustivas. Por exemplo, no caso em que você decide sobre levar o guarda-chuva e a calculadora, uma variável importante pode ser a ocorrência de chuva no dia. Não queremos molhar a calculadora, mas por outro lado levar o guarda-chuva quando não chove também não é agradável. Assim, Θ poderia ser o conjunto com os elementos "chove no dia", e "não chove no dia". Denotaremos as possibilidades neste exemplo por $\Theta = \{ch, \bar{ch}\}$.

Associaremos uma medida de probabilidade, $\mathbb{P}$, a Θ. Esta medida indica a plausibilidade que você atribui a cada elemento de Θ no momento da tomada da decisão. Note que, caso você tenha observado dados antes de tomar a sua decisão, então $\mathbb{P}$ será a probabilidade condicionada aos dados, a posteriori. Suponhamos que você confia no

$\mathcal{A}\backslash\Theta$	ch	$\bar{ch}$
(gc, ca)	0.5	0.9
$(gc, \bar{ca})$	0.6	0.7
$(\bar{gc}, ca)$	0	1
$(\bar{gc}, \bar{ca})$	0.1	0.8

Tabela 9: Utilidade para cada combinação de alternativa em $\mathcal{A}$ e possibilidade em Θ.

relato metereológico, segundo o qual a probabilidade de chuva é de 10%. Assim, você obteria $\mathbb{P}(\{ch\}) = 10\%$ e $\mathbb{P}(\{\bar{ch}\}) = 90\%$.

Finalmente, você atribuirá uma utilidade para cada par composto por uma alternativa, $a \in \mathcal{A}$ e uma possibilidade, $\theta_0 \in \Theta$. Esta utilidade representa o quão é desejável para você obter a possibilidade θ_0 quando você escolheu a. A utilidade é representada por uma função, $U : \mathcal{A} \times \Theta \to \mathbb{R}$, sendo que $U(a, \theta_0)$, indica a utilidade da ocorrência θ_0 tendo decidido por a. Por exemplo, considere que seu par predileto é aquele em que você leva a calculadora e não leva o guarda-chuva em um dia sem chuva. Também, considere que seu par menos desejado é aquele em que você leva a calculadora e não leva o guarda-chuva em um dia com chuva. Para simplicidade, considere que a utilidade do seu par favorito é 1 e a do seu par menos desejado é 0. Assim, para cada um dos demais pares, a utilidade será um número entre 0 e 1. Neste caso, se a sua utilidade para um determinado par é p, então obter este par será tão desejável quanto obter o melhor par com probabilidade p e o pior par com probabilidade $(1 - p)$. Quando $\mathcal{A}$ e Θ são conjuntos finitos, U pode ser representado como uma tabela. A tabela 9 ilustra a utilidade no exemplo envolvendo a chuva, o guarda-chuva e a calculadora.

Definição 6.1. Os elementos de um problema de decisão são:

- $\mathcal{A}$: o conjunto das alternativas disponíveis. Você deve escolher exatamente uma destas alternativas.
- Θ: o conjunto de possibilidades que podem ocorrer. Você não escolhe qual destas possibilidades ocorrerá.
- P: uma medida de probabilidade sobre Θ. Uma medida do quão plausível é cada possibilidade em Θ.
- U: uma função de $\mathcal{A} \times \Theta$ a $\mathbb{R}$ que indica a utilidade de cada par.

6.2 Avaliando alternativas

Na tabela 9 você encontrará a utilidade para cada alternativa e possibilidade em um exemplo. Observe que o par favorito é $(\bar{gc}, ca)$ e $\bar{ch}$, ou seja, não trazer o guarda-chuva e trazer a calculadora quando não há chuva. Assim, se você pudesse garantir que não haveria chuva, sua satisfação máxima seria obtida não trazendo o guarda-chuva e trazendo a calculadora. Contudo, em geral, não é possível ter certeza sobre qual possibilidade em Θ ocorrerá. Assim, é necessário avaliar qual a alternativa em $\mathcal{A}$ que é mais desejável sem fixar a possibilidade em Θ que ocorrerá.

A Teoria da decisão fornece uma forma de avaliar a utilidade de uma decisão.

Definição 6.2. Considere que, para cada $a \in \mathcal{A}$, $U_a : \Theta \to \mathbb{R}$ é uma variável aleatória tal que $U_a(\theta) = U(a, \theta)$. A utilidade de uma alternativa, $a \in \mathcal{A}$, é $\mathbb{E}[U_a]$.

A Definição 6.2 permite que você avalie a utilidade de cada alternativa disponível sem fixar qual possibilidade em Θ ocorrerá. Assim, para achar qual a melhor alternativa, basta calcular a utilidade de cada uma delas e escolher aquela que atinge a maior utilidade. A melhor alternativa é chamada de *alternativa de Bayes*.

Exemplo 6.3. Considere o Exemplo da seção anterior. Podemos calcular a utilidade de cada alternativa em $\mathcal{A}$ usando a Definição 6.2.

$$\begin{cases} U_{(gc,ca)} &= \mathbb{P}(\{ch\})U((gc,ca),ch) + \mathbb{P}(\bar{ch})U((gc,ca),\bar{ch}) = 0.1 \cdot 0.5 + 0.9 \cdot 0.9 = 0.86 \\ U_{(gc,\bar{ca})} &= \mathbb{P}(\{ch\})U((gc,\bar{ca}),ch) + \mathbb{P}(\bar{ch})U((gc,\bar{ca}),\bar{ch}) = 0.1 \cdot 0.6 + 0.9 \cdot 0.7 = 0.69 \\ U_{(\bar{gc},ca)} &= \mathbb{P}(\{ch\})U((\bar{gc},ca),ch) + \mathbb{P}(\bar{ch})U((\bar{gc},ca),\bar{ch}) = 0 \cdot 0.1 + 1 \cdot 0.9 = 0.9 \\ U_{(\bar{gc},\bar{ca})} &= \mathbb{P}(\{ch\})U((\bar{gc},\bar{ca}),ch) + \mathbb{P}(\bar{ch})U((\bar{gc},\bar{ca}),\bar{ch}) = 0.1 \cdot 0.1 + 0.9 \cdot 0.8 = 0.73 \end{cases}$$

Portanto, no cenário descrito, a melhor decisão é não levar o guarda-chuva e levar a calculadora.

Em outros livros (DeGroot, 2005), a Definição 6.2 não é apresentada como uma definição. É possível obtê-la como um teorema a partir de outras propriedades mais elementares envolvendo utilidade. Esta técnica é similar àquela que aplicamos para obter os axiomas da probabilidade a partir de propriedades sobre apostas. Para efeitos deste curso, tomaremos a Definição 6.2 como um ponto de partida, deixando investigações sobre a sua razoabilidade por sua conta.

Exercícios

Exercício 6.4. Modele o problema de decisão na situação do guarda-chuva, calculadora e chuva usando a sua própria probabilidade e utilidade. Qual a decisão ótima para você?

Exercício 6.5. Considere que, no Exemplo 6.3, $\mathbb{P}(\{ch\}) = p$. Para cada $p \in [0,1]$, encontre a decisão ótima.

Exercício 6.6. Considere que $\theta \sim$ Bernoulli(0.5). Suas alternativas são escolher um número real em $[0,1]$. Seja a a sua decisão, sua utilidade é $-(a-\theta)^2$. Ou seja, quanto mais próximo de θ, melhor será sua escolha.

(a) Indique os elementos do problema de decisão.

(b) Ache a decisão ótima.

(c) Se o parâmetro da Bernoulli fosse p ao invés de 0.5, qual seria a decisão ótima?

(d) No caso acima, se sua utilidade fosse $-|a-\theta|$, qual seria a decisão ótima?

(e) Interprete o resultado anterior.

Exercício 6.7 ((Hacking, 1972))**.** Em um dos primeiros usos documentados da Teoria da Decisão, Blaise Pascal argumentou sobre o porquê uma pessoa deve acreditar em Deus. Segundo Pascal, caso Deus exista e você acredite nele, sua recompensa será infinita. Similarmente, Pascal argumenta que, se Deus existe e você não acredita nele, sua punição será infinita. Finalmente, Pascal completa que, caso Deus não exista, sua utilidade será finita. Destas premissas Pascal argumenta que, por menor que seja a plausibilidade da existência de Deus, sua melhor alternativa é acreditar nele.

(a) Identifique os elementos de um problema de decisão na aposta de Pascal.

(b) Acompanhe o argumento de Pascal. Supondo que suas premissas sejam verdadeiras, a sua conclusão está de acordo com a Teoria da Decisão?

(c) Você acha que as premissas de Pascal são razoáveis?

Exercício 6.8 (Kadane (2011))**.** [p.290] Considere que você tem $R\$kx$. Sua utilidade para ter $R\$f$ é $\log(f)$. O custo de uma aposta em A é x. Se você comprar a aposta, caso A ocorra, você ganha $R\$1$ e, caso A não ocorra, você ganha $R\$0$. Você acredita que $P(A) = p$. Se você pode comprar quantas unidades da aposta você quiser, qual valor lhe traria maior satisfação?

Exercício 6.9 (Kadane (2011))**.** [p.275] Considere que você deve decidir entre participar ou não de uma aposta. Para participar da aposta, você deve pagar $R\$a$. A seguir, uma moeda honesta é lançada até a primeira ocorrência de cara. Seja X o número de lançamentos da moeda, sua recompensa é $R\$2^X$. Considere que sua utilidade é o seu ganho monetário, caso você participe da aposta, e 0, caso contrário.

(a) Qual é o maior valor de $R\$a$ tal que a melhor opção é participar da aposta?

(b) A conclusão acima é razoável? Você concorda com todos os elementos do problema de decisão descritos no problema?

(c) Como a sua resposta para o item (a) seria alterada caso sua utilidade para um ganho monetário, m, seja

$$U(m) = \begin{cases} \log(m) & , \text{se } m > 0 \\ 0 & , \text{se } m = 0 \\ -\log(-m) & , \text{se } m < 0 \end{cases}$$

6.3 Usando dados para avaliar alternativas

Na seção 6.1, analisamos os elementos de um problema de decisão. Em todos os exemplos que estudamos nesta seção, uma decisão era tomada sem consultar dados. Em outras palavras, a decisão era tomada utilizando apenas a informação que era conhecida *a priori* a respeito de θ.

Agora estudaremos o problema de tomada de decisões a partir de dados. Veremos que este tipo de problema pode ser tratado igualmente a um problema de decisão sem dados. Assim, as conclusões obtidas nas seções anteriores ainda serão válidas neste tipo de problema. Iniciamos nossa análise definindo os elementos de um problema de decisão com dados.

- $X \in \chi$: uma quantidade desconhecida que expressa os dados que serão observados. Os dados assumem valores em χ.
- $\theta \in \Theta$: uma quantidade desconhecida que é relevante para a tomada de decisões.
- $\mathbb{P}$: uma medida de probabilidade conjunta sobre (X, θ).
- $\mathcal{A}$: o conjunto de alternativas disponíveis para o tomador de decisão. No caso de um problema de decisão com dados, as alternativas são funções que, para cada dado observado, indicam qual decisão será tomada. Mais formalmente, existe um conjunto de alternativas simples, $\mathcal{A}_*$, e $\mathcal{A} = \{f(x) : \chi \to \mathcal{A}_*\}$.
- $U : \mathcal{A}_* \times \Theta \to \mathbb{R}$: a utilidade de cada alternativa em $\mathcal{A}_*$ combinada a cada possibilidade em Θ.

Exemplo 6.10. Todo dia antes de sair de casa, eu decido se colocarei meu guarda-chuva em minha mochila. A princípio, eu acho que a probabilidade de chuva é de 20%. Contudo, antes de tomar uma decisão, eu consulto a previsão do tempo. Eu acredito que a probabilidade de a previsão do tempo estar correta é de 90%. Caso não chova e eu não leve meu guarda-chuva, eu ficarei maximamente satisfeito, atribuindo utilidade 1 a esse resultado.

alternativa	$X=0$	$X=1$
δ_1	$\bar{g}$	$\bar{g}$
δ_2	$\bar{g}$	g
δ_3	g	$\bar{g}$
δ_4	g	g

Tabela 10: Descrição de cada alternativa em $\mathcal{A}$, ou seja, $\delta_1, \delta_2, \delta_3$ e δ_4. Cada alternativa é descrita indicando a decisão simples que é tomada para cada possível valor observado de X.

Caso chova e eu não leve meu guarda-chuva, eu ficarei minimamente satisfeito, atribuindo utilidade 0 a esse resultado. Caso chova e eu leve meu guarda-chuva, a utilidade será 0.6. Finalmente, caso não chova e eu leve meu guarda-chuva, a utilidade é 0.9.

Posso traduzir a descrição acima em um problema de decisão com dados:

- $X \in \{0,1\}$: a indicadora de que há uma previsão de chuva.
- $\theta \in \{0,1\}$: a indicadora de ocorrência de chuva no dia.
- P: a função de probabilidade conjunta de (θ, X) obtida da seguinte forma:

$$\begin{cases} \mathbb{P}(\theta=0, X=0) & = \mathbb{P}(\theta=0)\mathbb{P}(X=0|\theta=0) = 0.8 \cdot 0.9 = 0.74 \\ \mathbb{P}(\theta=0, X=1) & = \mathbb{P}(\theta=0)\mathbb{P}(X=1|\theta=0) = 0.8 \cdot 0.1 = 0.08 \\ \mathbb{P}(\theta=1, X=0) & = \mathbb{P}(\theta=1)\mathbb{P}(X=0|\theta=1) = 0.2 \cdot 0.1 = 0.02 \\ P(\theta=1, X=1) & = \mathbb{P}(\theta=1)\mathbb{P}(X=1|\theta=1) = 0.2 \cdot 0.9 = 0.18 \end{cases}$$

- $\mathcal{A}$: O conjunto de funções que tomam a decisão de levar ou não levar o guarda-chuva para cada possível previsão do tempo. Denote "levar o guarda-chuva" por "g" e "não levar o guarda-chuva" por "$\bar{g}$". Como existem 2 previsões possíveis e 2 decisões para cada previsão, existe um total de 4 destas funções. Mais precisamente, $\mathcal{A} = \{\delta : \{0,1\} \to \{g, \bar{g}\}\}$. O conjunto de alternativas é descrito na tabela 10. Estas alternativas podem ser interpretadas em linguagem comum.

  δ_1 nunca leva o guarda-chuva.

  δ_2 leva o guarda-chuva se é previsto chuva e não leva, caso contrário.

  δ_3 leva o guarda-chuva se é previsto não haver chuva e não leva, caso contrário.

  δ_4 sempre leva o guarda-chuva.

- $U(0,0) = 1$, $U(0,1) = 0$, $U(1,0) = 0.9$, $U(1,1) = 0.6$.

Os elementos de um problema de decisão com dados são semelhantes àqueles existentes em um problema de decisão sem dados. As únicas diferenças são a adição de um novo elemento representando os dados, X, e a extensão da probabilidade para também incluir os dados. Assim, no contexto em que temos dados, a melhor alternativa ainda é aquela que apresenta a melhor utilidade esperada.

Exemplo 6.11 (Continuação do Exemplo 6.10)**.** Existem quatro alternativas em $\mathcal{A}$, conforme descrito na tabela

10. Assim, para determinar qual a melhor delas, podemos calcular a utilidade esperada de cada uma delas.

$$\begin{aligned}\mathbb{E}[U_{\delta_i}] &= \mathbb{E}[U(\delta_i(X), \theta)] \\ &= \sum_{\theta_0=0}^{1}\sum_{x=0}^{1} U(\delta_i(x), \theta_0)P(\theta = \theta_0, X = x)\end{aligned}$$

$$\begin{aligned}\mathbb{E}[U_{\delta_1}] =& U(\delta_1(0), 0)\mathbb{P}(\theta = 0, X = 0) + U(\delta_1(0), 1)\mathbb{P}(\theta = 0, X = 1) + \\ & U(\delta_1(1), 0)\mathbb{P}(\theta = 1, X = 0) + U(\delta_1(1), 1)\mathbb{P}(\theta = 1, X = 1) \\ =& U(\bar{g}, 0)\mathbb{P}(\theta = 0, X = 0) + U(\bar{g}, 0)\mathbb{P}(\theta = 0, X = 1) + \\ & U(\bar{g}, 1)\mathbb{P}(\theta = 1, X = 0) + U(\bar{g}, 1)\mathbb{P}(\theta = 1, X = 1) \\ =& 1 \cdot 0.72 + 1 \cdot 0.08 + 0 \cdot 0.02 + 0 \cdot 0.18 = 0.8\end{aligned}$$

$$\begin{aligned}\mathbb{E}[U_{\delta_2}] =& U(\delta_2(0), 0)\mathbb{P}(\theta = 0, X = 0) + U(\delta_2(0), 1)\mathbb{P}(\theta = 0, X = 1) + \\ & U(\delta_2(1), 0)\mathbb{P}(\theta = 1, X = 0) + U(\delta_2(1), 1)\mathbb{P}(\theta = 1, X = 1) \\ =& U(\bar{g}, 0)\mathbb{P}(\theta = 0, X = 0) + U(g, 0)\mathbb{P}(\theta = 0, X = 1) + \\ & U(\bar{g}, 1)\mathbb{P}(\theta = 1, X = 0) + U(g, 1)\mathbb{P}(\theta = 1, X = 1) \\ =& 1 \cdot 0.72 + 0.9 \cdot 0.08 + 0 \cdot 0.02 + 0.6 \cdot 0.18 = 0.9\end{aligned}$$

$$\begin{aligned}\mathbb{E}[U_{\delta_3}] =& U(\delta_3(0), 0)\mathbb{P}(\theta = 0, X = 0) + U(\delta_3(0), 1)\mathbb{P}(\theta = 0, X = 1) + \\ & U(\delta_3(1), 0)\mathbb{P}(\theta = 1, X = 0) + U(\delta_3(1), 1)\mathbb{P}(\theta = 1, X = 1) \\ =& U(g, 0)\mathbb{P}(\theta = 0, X = 0) + U(\bar{g}, 0)\mathbb{P}(\theta = 0, X = 1) + \\ & U(g, 1)\mathbb{P}(\theta = 1, X = 0) + U(\bar{g}, 1)\mathbb{P}(\theta = 1, X = 1) \\ =& 0.9 \cdot 0.72 + 1 \cdot 0.08 + 0.6 \cdot 0.02 + 0 \cdot 0.18 = 0.72\end{aligned}$$

$$\begin{aligned}\mathbb{E}[U_{\delta_4}] =& U(\delta_4(0), 0)\mathbb{P}(\theta = 0, X = 0) + U(\delta_4(0), 1)\mathbb{P}(\theta = 0, X = 1) + \\ & U(\delta_4(1), 0)\mathbb{P}(\theta = 1, X = 0) + U(\delta_4(1), 1)\mathbb{P}(\theta = 1, X = 1) \\ =& U(g, 0)\mathbb{P}(\theta = 0, X = 0) + U(g, 0)\mathbb{P}(\theta = 0, X = 1) + \\ & U(g, 1)\mathbb{P}(\theta = 1, X = 0) + U(g, 1)\mathbb{P}(\theta = 1, X = 1) \\ =& 0.9 \cdot 0.72 + 0.9 \cdot 0.08 + 0.6 \cdot 0.02 + 0.6 \cdot 0.18 = 0.84\end{aligned}$$

Como $\mathbb{E}[U_{\delta_2}] > \mathbb{E}[U_{\delta_4}] > \mathbb{E}[U_{\delta_1}] > \mathbb{E}[U_{\delta_3}]$, decorre da Definição 6.2 que as alternativas são ordenadas de melhor para pior da seguinte forma: $\delta_2, \delta_4, \delta_1$ e δ_3. Esta ordenação segue nossa intuição. Como a previsão do tempo é precisa, a melhor alternativa é tomar a decisão de acordo com a previsão do tempo. Em seguida, o melhor é sempre levar o guarda-chuva. Isto ocorre pois existe uma penalidade pequena em levar o guarda-chuva quando não chove. A seguir, temos a alternativa de nunca levar o guarda-chuva. Finalmente, a pior alternativa é tomar a decisão em linha oposta à previsão do tempo.

Calcular a utilidade esperada para cada uma das alternativas em $\mathcal{A}$ pode exigir um número considerável de cálculos. O número de alternativas disponíveis é $|\mathcal{A}| = |\mathcal{A}_*|^{|\chi|}$. A seguir, veremos que é possível descobrir a melhor alternativa sem realizar estas contas. Esta conclusão é dada pelo seguinte Teorema:

Teorema 6.12. *Seja $\delta^* \in \mathcal{A}$ a alternativa com a maior utilidade esperada em um problema de decisão com dados. Ou seja, $\delta^* = \arg\max_{\delta \in \mathcal{A}} \mathbb{E}[U_\delta]$. δ^* é tal que, para cada $x \in \chi$,*

$$\begin{aligned}\delta^*(x) &= \arg\max_{a \in \mathcal{A}_*} \mathbb{E}[U_a | X = x] \\ &= \arg\max_{a \in \mathcal{A}_*} \int_\Theta U(a, \theta) f(\theta | x) d\theta\end{aligned}$$

Demonstração. Seja δ^* tal que, $\delta^*(x) = \arg\max_{a \in \mathcal{A}_*} \mathbb{E}[U_a | X = x]$. Assim, para todo $\delta \in \mathcal{A}^*$,

$$\begin{aligned}\mathbb{E}[U(\delta, \theta) | X = x] &\leq \mathbb{E}[U(\delta^*, \theta) | X = x] \\ \mathbb{E}[U(\delta, \theta) | X] &\leq \mathbb{E}[U(\delta^*, \theta) | X] \\ \mathbb{E}[\mathbb{E}[U(\delta, \theta) | X]] &\leq \mathbb{E}[\mathbb{E}[U(\delta^*, \theta) | X]] \\ \mathbb{E}[U(\delta, \theta)] &\leq \mathbb{E}[U(\delta^*, \theta)] \\ U(\delta) &\leq U(\delta^*)\end{aligned}$$

□

Em palavras, o Teorema 6.12 indica que a melhor alternativa em um problema com dados é tal que, para cada dado observado, ela escolhe a alternativa simples com a melhor utilidade esperada *a posteriori*. Observe que, para determinar a decisão ótima, o Teorema 6.12 exige que seja calculada uma utilidade esperada para cada combinação de alternativa simples e possível dado observado. Assim, enquanto o cálculo direto realizado no Exemplo 6.10 exige $|\mathcal{A}_*|^{|\chi|}$ utilidades esperadas, o cálculo proposto no Teorema 6.12 exige $|\mathcal{A}_*||\chi|$ utilidades esperadas. A redução de um fator exponencial em $|\chi|$ para um fator multiplicativo pode tornar a determinação da alternativa ótima mais simples. A seguir, usaremos o Exemplo 6.10 para ilustrar uma aplicação do Teorema 6.12.

Exemplo 6.13 (Continuação do Exemplo 6.10)**.** Podemos calcular a probabilidade a posteriori para cada possível observação:

$$\begin{aligned}\mathbb{P}(\theta = 1 | X = 0) &= \frac{\mathbb{P}(\theta = 1, X = 0)}{\mathbb{P}(\theta = 1, X = 0) + \mathbb{P}(\theta = 0, X = 0)} = \frac{0.02}{0.02 + 0.72} \approx 0.03 \\ \mathbb{P}(\theta = 1 | X = 1) &= \frac{\mathbb{P}(\theta = 1, X = 1)}{\mathbb{P}(\theta = 1, X = 1) + \mathbb{P}(\theta = 0, X = 1)} = \frac{0.18}{0.18 + 0.08} \approx 0.69\end{aligned}$$

A seguir, podemos calcular a decisão ótima para cada possível valor de X. Para $X = 0$, temos:

$$\begin{aligned}\mathbb{E}[U_{\bar{g}} | X = 0] &= U(\bar{g}, 0)\mathbb{P}(\theta = 0 | X = 0) + U(\bar{g}, 1)\mathbb{P}(\theta = 1 | X = 0) \\ &\approx 1 \cdot 0.97 + 0 \cdot 0.03 = 0.97 \\ \mathbb{E}[U_g | X = 0] &= U(g, 0)\mathbb{P}(\theta = 0 | X = 0) + U(g, 1)\mathbb{P}(\theta = 1 | X = 0) \\ &\approx 0.9 \cdot 0.97 + 0.6 \cdot 0.03 \approx 0.89\end{aligned}$$

Portanto, como $\mathbb{E}[U_{\bar{g}}|X=0] > \mathbb{E}[U_g|X=0]$,

$$\delta^*(0) = \bar{g}$$

Para $X = 1$, temos:

$$\begin{aligned}
\mathbb{E}[U_{\bar{g}}|X=1] &= U(\bar{g},0)\mathbb{P}(\theta=0|X=1) + U(\bar{g},1)\mathbb{P}(\theta=1|X=1) \\
&\approx 1 \cdot 0.31 + 0 \cdot 0.69 = 0.31 \\
\mathbb{E}[U_g|X=1] &= U(g,0)\mathbb{P}(\theta=0|X=1) + U(g,1)\mathbb{P}(\theta=1|X=1) \\
&\approx 0.9 \cdot 0.31 + 0.6 \cdot 0.69 \approx 0.69
\end{aligned}$$

Portanto, como $\mathbb{E}[U_g|X=1] > \mathbb{E}[U_{\bar{g}}|X=1]$,

$$\delta^*(1) = g$$

Juntando as conclusões obtidas, temos que $\delta^*(0) = \bar{g}$ e $\delta^*(1) = g$, ou seja, $\delta^* = \delta_2$. Obtivemos a mesma alternativa ótima encontrada no Exemplo 6.11, tal qual preconizado pelo Teorema 6.12.

Nas próximas Seções reescreveremos problemas tradicionais da Teoria Estatística a partir da Teoria da Decisão. Os problemas que discutiremos serão: Estimação, Intervalos de Confiança e Testes de Hipótese. Veremos que todos estes problemas podem ser descritos como um problema de decisão. Assim, as suas respectivas análises estatísticas podem ser obtidas diretamente a partir dos resultados que obtivemos nesta Seção.

Exercícios

Exercício 6.14. Modele o problema do Exemplo 6.10 utilizando suas próprias probabilidades e utilidades. Qual a melhor alternativa para você?

Exercício 6.15. Considere que no Exemplo 6.10, a probabilidade *a priori* de chuva é $p \in (0,1)$. Ache a melhor alternativa para cada possível valor de p.

Exercício 6.16. Considere que $\theta \sim$ Bernoulli(0.5). Você observará um dado $X \in \{0,1\}$ tal que, $X|\theta \sim$ Bernoulli $\left(\frac{2\theta+1}{4}\right)$. Para cada possível dado observado, você deve escolher um número real em $[0,1]$. Seja a a sua decisão, sua utilidade é $-(a-\theta)^2$. Ou seja, quanto mais próximo de θ, melhor será sua escolha.

(a) Indique os elementos do problema de decisão.

(b) Ache a decisão ótima.

(c) Se o parâmetro da Bernoulli fosse p ao invés de 0.5, qual seria a decisão ótima?

(d) No caso acima, se sua utilidade fosse $-|\theta - a|$, qual seria a decisão ótima?

(e) Interprete os resultados anteriores.

7 Inferência Bayesiana

Neste capítulo, veremos como a Teoria da Decisão pode ser usada para criar procedimentos para resumir a informação a posteriori disponível em um problema. Neste capítulo, Θ, o conjunto de possíveis ocorrências que são relevantes para a tomada da sua decisão estudado no Capítulo 6, é justamente o espaço paramétrico. Assim, $f(\theta|x)$ representa a distribuição a posteriori.

7.1 Estimação Pontual

O problema de estimação consiste em escolher, a partir dos dados, um valor próximo ao parâmetro do modelo estatístico. O valor escolhido é chamado de estimador do parâmetro e será denotado por $\hat{\theta}$. Nesta seção, tomaremos que $\mathcal{A}_* = \Theta \subset \mathbb{R}^k$. Para capturar a ideia de que o valor escolhido deve estar próximo ao parâmetro, estudaremos utilidades do tipo $U(a, \theta) = -w(\theta)d(a, \theta)$, onde $w(\theta)$ é uma função não-negativa que indica a importância de acertar o valor θ e $d(a, \theta)$ é uma distância entre a e θ.

Para algumas utilidades deste tipo é possível derivar precisamente qual o melhor estimador (isto é, o estimador que maximiza a utilidade esperada). A seguir, estudaremos alguns destes casos.

7.1.1 Distância quadrática

A distância quadrática, d_2, é tal que $d_2(a, \theta) = (a - \theta)^2$. Esta é uma das funções mais frequentemente usadas em Teoria Estatística, estando ligada à técnica dos mínimos quadrados.

O seguinte lema é útil para provar resultados envolvendo a distância quadrática.

Lema 7.1. *Sejam θ e X duas variáveis aleatórias,*

$$\mathbb{E}[(\theta - f(X))^2|X] = \mathbb{V}[\theta|X] + (\mathbb{E}[\theta|X] - f(X))^2$$

Demonstração.

$$\begin{aligned}
\mathbb{E}[(\theta - f(X))^2|X] &= \mathbb{E}[\theta^2|X] - 2\mathbb{E}[\theta f(X)|X] + \mathbb{E}[f(X)^2|X] \\
&= \mathbb{E}[\theta^2|X] - 2f(X)\mathbb{E}[\theta|X] + f(X)^2 \\
&= \mathbb{E}[\theta^2|X] - \mathbb{E}[\theta|X]^2 + \mathbb{E}[\theta|X]^2 - 2f(X)\mathbb{E}[\theta|X] + f(X)^2 \\
&= \mathbb{V}[\theta|X] + (\mathbb{E}[\theta|X] - f(X))^2
\end{aligned}$$

□

Usando o lema acima, podemos achar o melhor estimador para a distância quadrática.

Teorema 7.2. *Seja $\hat{\theta}$ um estimador arbitrário. Se $U(\hat{\theta}, \theta) = -(\hat{\theta} - \theta)^2$ e existe $\hat{\theta}$ tal que $\mathbb{E}[U(\hat{\theta}, \theta)] > -\infty$, então o melhor estimador, $\hat{\theta}_*$, é tal que*

$$\hat{\theta}_* = \mathbb{E}[\theta|X]$$

Demonstração.

$$\begin{aligned}
\hat{\theta}_*(X) &= \arg\max_{\hat{\theta} \in \mathcal{A}} \mathbb{E}[U(\hat{\theta}, \theta)|X] && \text{Teorema 6.12} \\
&= \arg\max_{\hat{\theta} \in \mathcal{A}} \mathbb{E}[-(\hat{\theta} - \theta)^2|X] \\
&= \arg\max_{\hat{\theta} \in \mathcal{A}} -\mathbb{V}[\hat{\theta} - \theta|X] - |\mathbb{E}[\theta|X] - \hat{\theta}|^2 && \text{Lema 7.1} \\
&= \arg\max_{\hat{\theta} \in \mathcal{A}} -\mathbb{V}[\theta|X] - |\mathbb{E}[\theta|X] - \hat{\theta}|^2 && \hat{\theta} \text{ é constante dado X} \\
&= \arg\max_{\hat{\theta} \in \mathcal{A}} -|\mathbb{E}[\theta|X] - \hat{\theta}|^2 = \mathbb{E}[\theta|X] && \mathbb{V}[\theta|X] \text{ não depende de } \hat{\theta}
\end{aligned}$$

□

Assim, o melhor estimador segundo a distância quadrática é a média da distribuição *a posteriori* do parâmetro.

Podemos generalizar o resultado acima para o caso multivariado. Considere que θ é um vetor de parâmetros reais. Neste caso, a distância quadrática é generalizada em uma forma quadrática. Ou seja, para alguma matriz positiva definida, A, $d_2(\hat{\theta}, \theta)$ é definida como $(\hat{\theta}-\theta)^T A(\hat{\theta}-\theta)$ e $U(\hat{\theta}, \theta) = -d_2(\hat{\theta}, \theta)$. Neste caso, obtemos resultado semelhante ao Lema 7.1

Lema 7.3.

$$\mathbb{E}[(\hat{\theta}-\theta)^T A(\hat{\theta}-\theta)|X] = \mathbb{E}[(\theta - E[\theta|X])^T A(\theta - \mathbb{E}[\theta|X])] + (\hat{\theta} - \mathbb{E}[\theta|X])^T A(\hat{\theta} - \mathbb{E}[\theta|X])$$

Demonstração.

$$\begin{aligned}
\mathbb{E}[(\hat{\theta}-\theta)^T A(\hat{\theta}-\theta)|X] &= \mathbb{E}[\theta^T A\theta|X] - \mathbb{E}[\hat{\theta}^T A\theta|X] - \mathbb{E}[\theta^T A\hat{\theta}|X] + \mathbb{E}[\hat{\theta}^T A\hat{\theta}|X] \\
&= \mathbb{E}[\theta^T A\theta|X] - \hat{\theta}^T A\mathbb{E}[\theta|X] - \mathbb{E}[\theta^T|X]A\hat{\theta} + \hat{\theta}^T A\hat{\theta} \\
&= \mathbb{E}[\theta^T A\theta|X] - \mathbb{E}[\theta|X]^T AE[\theta|X] \\
&+ \mathbb{E}[\theta|X]^T A\mathbb{E}[\theta|X] - \hat{\theta}^T A\mathbb{E}[\theta|X] - \mathbb{E}[\theta^T|X]A\hat{\theta} + \hat{\theta}^T A\hat{\theta} \\
&= \mathbb{E}[\theta^T A\theta|X] - \mathbb{E}[\theta|X]^T AE[\theta|X] \\
&+ (\mathbb{E}[\theta|X] - \hat{\theta})^T A(\mathbb{E}[\theta|X] - \hat{\theta}) \\
&= \mathbb{E}[(\theta - \mathbb{E}[\theta|X])^T A(\theta - \mathbb{E}[\theta|X])|X] + (\mathbb{E}[\theta|X] - \hat{\theta})^T A(\mathbb{E}[\theta|X] - \hat{\theta})
\end{aligned}$$

□

A partir do Lema 7.3, podemos provar

Teorema 7.4. *Seja $\hat{\theta}$ um estimador arbitrário. Se $U(\hat{\theta}, \theta) = -(\hat{\theta}-\theta)^T A(\hat{\theta}-\theta)$, e existe $\hat{\theta}$ tal que $\mathbb{E}[U(\hat{\theta}, \theta)] > -\infty$, então o melhor estimador, $\hat{\theta}_*$, é tal que*

$$\hat{\theta}_* = \mathbb{E}[\theta|X]$$

Demonstração.

$$\begin{aligned}
\hat{\theta}_* &= \arg\max_{\hat{\theta}\in\mathcal{A}} \mathbb{E}[U(\hat{\theta}, \theta)|X] && \text{Teorema 6.12} \\
&= \arg\max_{\hat{\theta}\in\mathcal{A}} -\mathbb{E}[(\theta - \hat{\theta})^T A(\theta - \hat{\theta})|X] \\
&= \arg\max_{\hat{\theta}\in\mathcal{A}} -\mathbb{E}[(\theta - \mathbb{E}[\theta|X])^T A(\theta - \mathbb{E}[\theta|X])|X] - (\mathbb{E}[\theta|X] - \hat{\theta})^T A(\mathbb{E}[\theta|X] - \hat{\theta}) && \text{Lema 7.3} \\
&= \arg\max_{\hat{\theta}\in\mathcal{A}} -(\mathbb{E}[\theta|X] - \hat{\theta})^T A(\mathbb{E}[\theta|X] - \hat{\theta}) = \mathbb{E}[\theta|X]
\end{aligned}$$

□

7.1.2 Desvio absoluto

O desvio absoluto, d_1, é tal que $d_1(a, \theta) = |a - \theta|$. Esta distância, historicamente, foi menos estudada em estatística devido à maior dificuldade de obter resultados analíticos. O desvio absoluto está tipicamente associado

a estimadores robustos, ou seja, que não são fortemente influenciados por *outliers*. Isto ocorre pois, em relação à distância quadrática, o desvio penaliza menos grandes desvios do estimador em relação a θ. Para o desvio absoluto, obtemos o seguinte resultado:

Teorema 7.5. *Seja $\hat{\theta}$ um estimador arbitrário. Se $U(\hat{\theta}, \theta) = -|\hat{\theta} - \theta|$ e existe $\hat{\theta}$ tal que $\mathbb{E}[U(\hat{\theta}, \theta)] > -\infty$, então o melhor estimador, $\hat{\theta}_*$ é tal que*

$$\hat{\theta}_* = Med[\theta|X]$$

Lema 7.6. *Defina $A_X^- = \{\omega \in \Omega : \theta < M_X\}$, $A_X^+ = \{\omega \in \Omega : \theta > M_X\}$, $A_X^= = \{\omega \in \Omega : \theta = M_X\}$ e $M_X = Med[\theta|X]$. Obtém-se*

$$\begin{aligned}
\mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^-}|X] &\geq (\hat{\theta} - M_X)\mathbb{P}(A_X^-|X) \\
\mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^+}|X] &\geq -(\hat{\theta} - M_X)\mathbb{P}(A_X^+|X) \\
\mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^=}|X] &\geq |\hat{\theta} - M_X|\mathbb{P}(\mathbb{I}_{A_X^=}|X)
\end{aligned}$$

Demonstração. Note que

$$\begin{aligned}
|\hat{\theta} - \theta| &= \max(\hat{\theta} - \theta, \theta - \hat{\theta}) \geq \hat{\theta} - \theta \\
|\hat{\theta} - \theta| &= \max(\hat{\theta} - \theta, \theta - \hat{\theta}) \geq \theta - \hat{\theta}
\end{aligned} \tag{19}$$

Portanto,

$$\begin{aligned}
\mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^-}|X] &= \mathbb{E}[(|\hat{\theta} - \theta| - M_X + \theta)\mathbb{I}_{A_X^-}|X] \\
&\geq \mathbb{E}[(\hat{\theta} - \theta - M_X + \theta)\mathbb{I}_{A_X^-}|X] && \text{eq. (19)} \\
&= (\hat{\theta} - M_X)\mathbb{E}[\mathbb{I}_{A_X^-}|X] \\
&= (\hat{\theta} - M_X)\mathbb{P}(A_X^-|X)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^+}|X] &= \mathbb{E}[(|\hat{\theta} - \theta| + M_X - \theta)\mathbb{I}_{A_X^+}|X] \\
&\geq \mathbb{E}[(-\hat{\theta} + \theta + M_X - \theta)\mathbb{I}_{A_X^+}|X] && \text{eq. (19)} \\
&= (-\hat{\theta} + M_X)\mathbb{E}[\mathbb{I}_{A_X^+}|X] \\
&= -(\hat{\theta} - M_X)\mathbb{P}(A_X^+|X)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^=}|X] &= \mathbb{E}[(|\hat{\theta} - \theta|)\mathbb{I}_{A_X^=}|X] \\
&= \mathbb{E}[(|\hat{\theta} - M_X|)\mathbb{I}_{A_X^=}|X] \\
&= |\hat{\theta} - M_X|\mathbb{E}[\mathbb{I}_{A_X^=}|X] \\
&= |\hat{\theta} - M_X|\mathbb{P}(\mathbb{I}_{A_X^=}|X)
\end{aligned}$$

□

Demonstração do Teorema 7.5. Defina $M_X := Med[\theta|X]$, $A_X^- = \{\omega \in \Omega : \theta < M_X\}$, $A_X^+ = \{\omega \in \Omega : \theta > M_X\}$ e

$A_X^= = \{\omega \in \Omega : \theta = M_X\}$. Note que

$$\mathbb{P}(A_X^=|X) - |\mathbb{P}(A_X^+|X) - \mathbb{P}(A_X^-|X)| \geq 0 \qquad Med[\theta|X] \text{ é a mediana a posteriori de } \theta. \tag{20}$$

Também, como $\{A_X^-, A_X^+, A_X^=\}$ particiona Ω, $\mathbb{I}_{A_X^-} + \mathbb{I}_{A_X^+} + \mathbb{I}_{A_X^=} = 1$. Portanto,

$$\begin{aligned}
\mathbb{E}[U(M_X,\theta)|X] - \mathbb{E}[U(\hat{\theta},\theta)|X] =& \mathbb{E}[|M_X - \theta||X] + \mathbb{E}[|\hat{\theta} - \theta||X] \\
=& \mathbb{E}[|\hat{\theta} - \theta| - |M_X - \theta||X] \\
=& \mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)(\mathbb{I}_{A_X^-} + \mathbb{I}_{A_X^+} + \mathbb{I}_{A_X^=})|X] \\
=& \mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^-}|X] + \mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^+}|X] \\
&+ \mathbb{E}[(|\hat{\theta} - \theta| - |M_X - \theta|)\mathbb{I}_{A_X^=}|X] \\
&\geq (\hat{\theta} - M_X)(\mathbb{P}(A_X^-|X) - \mathbb{P}(A_X^+|X) + |\hat{\theta} - M_X|\mathbb{P}(A_X^=|X) && \text{Lema 7.6} \\
&\geq -|\hat{\theta} - M_X||(\mathbb{P}(A_X^-|X) - \mathbb{P}(A_X^+|X)| + |\hat{\theta} - M_X|\mathbb{P}(A_X^=|X) \\
&= |\hat{\theta} - M_X|(\mathbb{P}(A_X^=|X) - |(\mathbb{P}(A_X^-|X) - \mathbb{P}(A_X^+|X)|) \geq 0 && \text{eq. (20)}
\end{aligned}$$

Portanto, como $\mathbb{E}[U(M_X,\theta)|X] \geq \mathbb{E}[U(\hat{\theta},\theta)|X]$, decorre do Teorema 6.12 que Med$[\theta|X]$ é o melhor estimador para θ sob a utilidade $-d_1(\hat{\theta},\theta)$. □

Exercícios

Exercício 7.7. Defina $M_X := Med[\theta|X]$, $A_X^- = \{\omega \in \Omega : \theta < M_X\}$, $A_X^+ = \{\omega \in \Omega : \theta > M_X\}$ e $A_X^= = \{\omega \in \Omega : \theta = M_X\}$. Prove que

$$\mathbb{P}(A_X^=|X) - |\mathbb{P}(A_X^-|X) - \mathbb{P}(A_X^+|X)| \geq 0$$

Exercício 7.8. Considere que, dado θ, $X_1, \ldots, X_n$ são i.i.d. e $X_i \sim$ Binomial(m,θ). Se *a priori*, $\theta \sim$ Beta(α,β),

(a) Ache $\hat{\theta}_2^*$, o melhor estimador para θ sob $U(\hat{\theta},\theta) = -d_2(\hat{\theta},\theta)$.

(b) Ache $\lim_n \hat{\theta}_2^*$.

(c) Compare a utilidade esperada a posteriori de $\hat{\theta}_2^*$ e $m^{-1}\bar{X}$ sob $U(\hat{\theta},\theta) = -d_2(\hat{\theta},\theta)$.

Exercício 7.9. Considere que, dado θ, $X_1, \ldots, X_n$ são i.i.d. e $X_i \sim N(\theta,1)$. Se *a priori* $\theta \sim N(0,1)$,

(a) Ache $\hat{\theta}_2^*$, o melhor estimador para θ sob $U(\hat{\theta},\theta) = -d_2(\hat{\theta},\theta)$.

(b) Ache $\lim_n \hat{\theta}_2^*$.

(c) Compare a utilidade esperada a posteriori de $\hat{\theta}_2^*$ e $\bar{X}$ sob $U(\hat{\theta},\theta) = -d_2(\hat{\theta},\theta)$.

(d) Ache $\hat{\theta}_1^*$, o melhor estimador para θ sob $U(\hat{\theta},\theta) = -d_1(\hat{\theta},\theta)$.

Exercício 7.10. Considere que a proporção de indivíduos doentes em uma determinada população é um valor desconhecido, θ. Uma amostra de n indivíduos é retirada independentemente da população. Para cada indivíduo a probabilidade de que o exame seja positivo dado que o indivíduo está doente é α. A probabilidade de que o exame seja negativo dado que o indivíduo não está doente é β. Dos n indivíduos testados, $n\bar{x}$ tiveram o teste positivo. Considere que, *a priori*, $\theta \sim U(0,1)$. Ache o melhor estimador para θ de acordo com $U(\hat{\theta},\theta) = -d_2(\hat{\theta},\theta)$.

Exercício 7.11 (Schervish (2012; p.309))**.** Seja $0 < c < 1$. Considere um problema de estimação sob a utilidade:

$$U(\hat{\theta}, \theta) = \begin{cases} -c(\hat{\theta} - \theta) & , \text{se } \hat{\theta} \geq \theta \\ -(1-c)(\theta - \hat{\theta}) & , \text{se } \hat{\theta} < \theta \end{cases}$$

Determine o estimador ótimo.

Exercício 7.12. Considere que Θ é discreto. Qual é o estimador pontual de Bayes para θ com relação à utilidade $U(\hat{\theta}, \theta) = -\mathbb{I}(\hat{\theta} \neq \theta)$? Interprete essa utilidade.

7.2 Regiões de credibilidade

Na Seção passada estudamos o problema de estimação, em que desejamos escolher um número próximo ao parâmetro. Como resposta, obtivemos que medidas de centralidade da distribuição a posteriori são obtidas como estimadores ótimos. Por exemplo, a média a posteriori é o estimador ótimo para a utilidade quadrática e a mediana a posteriori é o estimador ótimo para a utilidade obtida pelo desvio absoluto. Pelo fato de a resposta para um problema de estimação ser um único número, ela não indica o quanto temos certeza de que o parâmetro está próximo desse número.

Nesta Seção, estudaremos o problema de decisão que consiste em obter regiões de credibilidade. Em outras palavras, o problema de obter um subconjunto do espaço paramétrico no qual o parâmetro provavelmente está contido. Contudo, também desejamos que o intervalo seja o menor possível, para que tenhamos mais precisão a respeito de qual é o valor do parâmetro. Assim como no problema de estimação, existem várias funções de perda que buscam capturar estas idéias. A seguir, estudaremos algumas destas funções.

7.2.1 Intervalos de credibilidade

Iniciaremos a análise buscando intervalos que provavelmente contém o parâmetro. Um intervalo é definido como $[a, b]$, onde $a \in \mathbb{R}$, $b \in \mathbb{R}$ e $a \leq b$. Assim, o conjunto de alternativas simples é $\mathcal{A}_* = \{(a, b) \in \mathbb{R}^2 : a \leq b\}$.

Uma possível função de utilidade para este problema é dada por

$$U((a, b), \theta) = \mathbb{I}(\theta \in [a, b]) - k(b - a) \qquad\qquad k > 0$$

O elemento $\mathbb{I}(\theta \in [a, b])$ indica que é desejável que θ esteja no intervalo. O elemento $-k(b - a)$ indica que é desejável que o intervalo seja pequeno.

Teorema 7.13. *Se $U((a, b), \theta) = \mathbb{I}(\theta \in [a, b]) - k(b - a)$, então a decisão ótima, $(a^*(X), b^*(X))$ satisfaz*

$$f_{\theta|X}(a^*|X) = f_{\theta|X}(b^*|X) = k$$

Demonstração. Podemos deduzir o intervalo ótimo neste caso usando o Teorema 6.12.

$$\begin{aligned} \mathbb{E}[U((a, b), \theta)|X] &= \mathbb{E}[\mathbb{I}(\theta \in [a, b]) - k(b - a)|X] \\ &= \mathbb{P}(\theta \in [a, b]|X) - k(b - a) \\ &= \int_a^b f(\theta|X)d\theta - k(b - a) \\ &= \left(ka - \int_{-\infty}^a f(\theta|X)d\theta\right) + \left(\int_{-\infty}^b f(\theta|X)d\theta - kb\right) \end{aligned}$$

Portanto, temos que

$$\nabla \mathbb{E}[U((a,b),\theta)|X] = (k - f_{\theta|X}(a|X), f_{\theta|X}(b|X) - k)$$

□

Assim, para que (a,b) seja um ponto de máximo, é necessário que $f_{\theta|X}(a|X) = f_{\theta|X}(b|X) = k$. Em geral, podem existir vários pontos que satisfazem esta propriedade. Neste caso, devemos testar todas as possíveis combinações de pontos e escolher aquela que maximiza a utilidade esperada. Em particular, se $f_{\theta|X}(\cdot|X)$ é unimodal, então apenas existem dois pontos $(a,b) \in \mathbb{R}^2$ tais que $f_{\theta|X}(a|X) = f_{\theta|X}(b|X) = k$ e temos que $[a,b] = \{\theta : f(\theta|X) \geq k\}$. Note que é possível tomar k de tal forma que $\mathbb{P}(\theta \in [a,b]|X) = 1 - \alpha$. Neste caso, dizemos que construímos um intervalo de credibilidade com credibilidade $1-\alpha$ (veja Seção 7.2.3).

Outra possível função utilidade é

$$U((a,b),\theta) = -k_1((a-\theta)_+ + (\theta - b)_+) - k_2(b-a) \qquad \text{, onde } (x-y)_+ = \max(0, x-y)$$
$$k_1, k_2 > 0$$

Similarmente à utilidade anterior, o elemento $-k_2(b-a)$ indica que é desejável que o intervalo seja pequeno. Como contraponto, o elemento $-k_1((a-\theta)_+ + (\theta-b)_+) - k_2(b-a)$ indica que é desejável que a distância de θ a $[a,b]$ seja baixa e, em particular, que θ esteja dentro de $[a,b]$.

Teorema 7.14. *Se $U((a,b),\theta) = -k_1((a-\theta)_+ + (\theta-b)_+) - k_2(b-a)$, então a melhor decisão (a,b) é tal que a e b são, respectivamente, o $\frac{k_2}{k_1}$-ésimo e $1 - \frac{k_2}{k_1}$-ésimo percentis da* posteriori *para θ dado X.*

Demonstração. Podemos deduzir o intervalo ótimo neste caso usando o Teorema 6.12.

$$\begin{aligned}
\mathbb{E}[U((a,b),\theta)|X] &= \mathbb{E}[-k_1((a-\theta)_+ + (\theta-b)_+) - k_2(b-a)|X] \\
&= -k_1(\mathbb{E}[(a-\theta)_+|X] + \mathbb{E}[(\theta-b)_+|X]) - k_2(b-a) \\
&= -k_1\left(\int_{-\infty}^{a}(a-\theta)f(\theta|X)d\theta + \int_{b}^{\infty}(\theta-b)f(\theta|X)d\theta\right) - k_2(b-a) \\
&= \underbrace{\left(-k_1\int_{-\infty}^{a}(a-\theta)f(\theta|X)d\theta + k_2 a\right)}_{g(a)} + \underbrace{\left(-k_1\int_{b}^{\infty}(\theta-b)f(\theta|X)d\theta - k_2 b\right)}_{h(b)}
\end{aligned}$$

Portanto, podemos tomar a maximizando $g(a)$ e b maximizando $h(b)$. Note que

$$\begin{aligned}
g(a) &= -k_1\int_{-\infty}^{a}(a-\theta)f(\theta|X)d\theta + k_2 a \\
&= -k_1 a\int_{-\infty}^{a}f(\theta|X)d\theta + k_1\int_{-\infty}^{a}\theta f(\theta|X)d\theta + k_2 a
\end{aligned}$$

Portanto, pelo Teorema Fundamental do Cálculo,

$$\begin{aligned}
g'(a) &= -k_1\int_{-\infty}^{a}f(\theta|X)d\theta - k_1 a f(a|X) + k_1 a f(a|X) + k_2 \\
&= -k_1\int_{-\infty}^{a}f(\theta|X)d\theta + k_2
\end{aligned}$$

Assim, para que $g'(a) = 0$, é necessário que

$$\int_{-\infty}^{a} f(\theta|X)d\theta = \frac{k_2}{k_1}$$
$$\mathbb{P}(\theta \leq a|X) = \frac{k_2}{k_1}$$

Ou seja, a deve ser o $\frac{k_2}{k_1}$-ésimo percentil da *posteriori* para θ dado X. Similarmente, para b, obtemos

$$\begin{aligned} h(b) &= -k_1 \int_b^{\infty} (\theta - b) f(\theta|X)d\theta - k_2 b \\ &= -k_1 \int_b^{\infty} \theta f(\theta|X)d\theta + k_1 b \int_b^{\infty} f(\theta|X)d\theta - k_2 \end{aligned}$$

Portanto, pelo Teorema Fundamental do Cálculo,

$$\begin{aligned} h'(b) &= k_1 b f(b|X) + k_1 \int_b^{\infty} f(\theta|X)d\theta - k_1 b f(b|X) - k_2 b \\ &= k_1 \int_b^{\infty} f(\theta|X)d\theta - k_2 \end{aligned}$$

Assim, para que $h'(b) = 0$, é necessário que

$$\int_b^{\infty} f(\theta|X)d\theta = \frac{k_2}{k_1}$$
$$\mathbb{P}(\theta \geq b|X) = \frac{k_2}{k_1}$$
$$\mathbb{P}(\theta < b|X) = 1 - \frac{k_2}{k_1}$$

Ou seja, b deve ser o $1 - \frac{k_2}{k_1}$-ésimo percentil da *posteriori* para θ dado X. □

Note que apenas os valores relativos entre k_1 e k_2 são relevantes para a decisão tomada. Por exemplo, o mesmo intervalo será obtido se $k_1 = 1$ e $k_2 = 0.5$ ou se $k_1 = 2$ e $k_2 = 1$. Também note que é possível tomar k_1 e k_2 de tal forma que $\mathbb{P}(\theta \in [a, b]|X) = 1 - \alpha$. Para tal, basta escolher $\frac{k_2}{k_1} = \frac{\alpha}{2}$ Neste caso, dizemos que construímos um intervalo de credibilidade com credibilidade $1 - \alpha$.

O último intervalo de credibilidade que veremos é baseado na utilidade

$$U((a, b), \theta) = -k_1(b - a) - \frac{k_2}{b - a}\left(\theta - \frac{a + b}{2}\right)^2 \qquad k_1, k_2 > 0$$

Similarmente às utilidades anteriores, $-k_1(b-a)$ indica que é desejável que o intevalo seja pequeno. Também, $\frac{a+b}{2}$ é o centro do intervalo. Assim, $-\frac{k_2}{b-a}\left(\theta - \frac{a+b}{2}\right)^2$ indica que é desejável que o centro do intervalo esteja próximo a θ.

Teorema 7.15. *Se* $U((a, b), \theta) = -k_1(b - a) - \frac{k_2}{b-a}\left(\theta - \frac{a+b}{2}\right)^2$, *então a melhor alternativa é tal que*

$$[a, b] = \left[E[\theta|X] - 2^{-1}\sqrt{\frac{k_2}{k_1}Var[\theta|X]}, E[\theta|X] + 2^{-1}\sqrt{\frac{k_2}{k_1}Var[\theta|X]}\right]$$

Demonstração. Podemos deduzir o intervalo ótimo neste caso usando o Teorema 6.12.

$$\begin{aligned}
\mathbb{E}[U((a,b),\theta)|X] &= \mathbb{E}\left[-k_1(b-a) - \frac{k_2}{b-a}\left(\theta - \frac{a+b}{2}\right)^2 |X\right] \\
&= -k_1(b-a) - \frac{k_2}{b-a}\mathbb{E}\left[\left(\theta - \frac{a+b}{2}\right)^2 |X\right] \\
&= -k_1 t - \frac{k_2}{t}\mathbb{E}\left[(\theta - c)^2|X\right] && t = b-a, c = \frac{a+b}{2} \\
&= -k_1 t - \frac{k_2}{t}\left(\mathbb{V}[\theta|X] + (\mathbb{E}[\theta|X] - c)^2\right) && \text{Lema 7.1} \qquad (21)
\end{aligned}$$

Note que, qualquer que seja o valor de t, a expressão em eq. (21) é maximizada tomando $c^* = E[\theta|X]$. Substituindo esse valor em eq. (21), obtemos

$$\mathbb{E}[U((a,b),\theta)|X] = -k_1 t - \frac{k_2}{t}\mathbb{V}[\theta|X]$$

Para achar o ponto de máximo dessa expressão, determinamos o seu ponto crítico

$$\frac{d\mathbb{E}[U((a,b),\theta)|X]}{dt} = -k_1 + k_2 t^{-2}\mathbb{V}[\theta|X]$$

Assim, $\frac{d\mathbb{E}[U((a,b),\theta)|X]}{dt} = 0$ para $t^* = \sqrt{\frac{k_2}{k_1}\mathbb{V}[\theta|X]}$. Como

$$\frac{d^2\mathbb{E}[U((a,b),\theta)|X]}{dt^2} = -2k_2 t^{-3}\mathbb{V}[\theta|X] < 0$$

t^* é um ponto de máximo de $\mathbb{E}[U((a,b),\theta)|X]$. O resultado final é obtido usando $c^* = \frac{a+b}{2}$ e $t^* = b-a$. □

7.2.2 Regiões de credibilidade

Na subseção anterior utilizamos intervalos para resumir a informação a respeito da distribuição a posteriori. Em geral, construímos os intervalos de credibilidade de tal forma que eles fossem pequenos e contivessem o parâmetro com alta probabilidade. Contudo, em algumas ocasiões, um intervalo pode ser inadequado para obter estes objetivos. Por exemplo, considere que

$$f(\theta|X) = \frac{1}{2\sqrt{2\pi}}\exp(-\theta^2) + \frac{1}{2\sqrt{2\pi}}\exp(-(\theta^2 - 100))$$

Esta é a distribuição obtida misturando-se duas normais com variâncias iguais a 1 e médias iguais a 0 e 100. A densidade $f(\theta|X)$ é apresentada na fig. 3. Note que $\mathbb{E}[\theta|X] = 50$. Assim, se usarmos a terceira função de utilidade da subseção passada, o intervalo de credibilidade obtido será da forma $[50-k, 50+k]$. Usando um software computacional, encontramos que $k = 51.7$ gera um intervalo de credibilidade próxima a 95%. Contudo, o intervalo obtido, $[-1.7, 101.7]$, não resume adequadamente $f(\theta|X)$. Isso ocorre pois o intervalo inclui os valores de baixa densidade próximos a 50.

Como uma alternativa a um intervalo de credibilidade, poderíamos combinar um intervalo de credibilidade para cada uma das normais na mistura. Sabemos que uma $N(0,1)$ tem probabilidade de 95% de estar em $[-1.96, 1.96]$. Similarmente, a $N(100,1)$ tem probabilidade de 95% de estar em $[98.04, 101.96]$. Assim, a mistura de uma $N(0,1)$ e uma $N(100,1)$ tem alta probabilidade de estar em $[-1.96, 1.96] \cup [98.04, 101.96]$. Neste exemplo, intervalos são

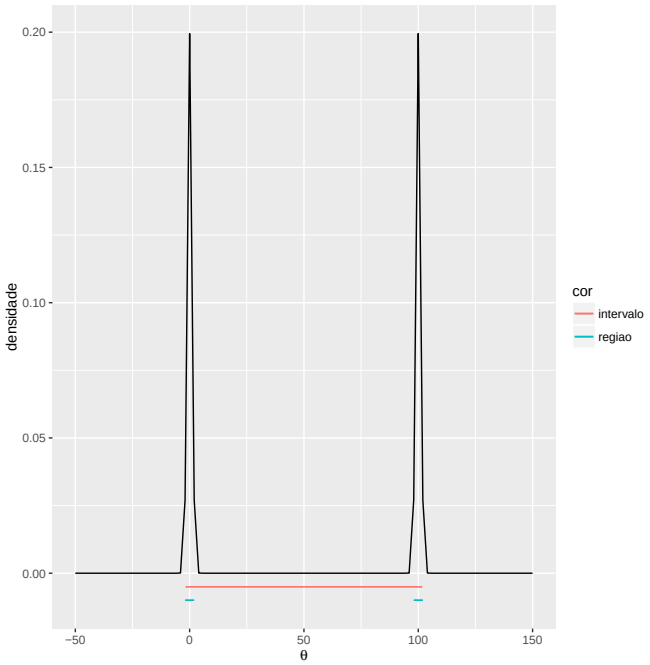

Figura 3: Densidade da mistura de uma $N(0,1)$ e uma $N(100,1)$ acompanhada de um intervalo de credibilidade e de uma região de credibilidade.

demasiadamente restritivos e não descrevem adequadamente a *posteriori* multimodal.

Assim, nesta subseção consideraremos um problema de decisão em que, ao invés de escolher um intervalo para descrever a posteriori, é necessário escolher uma região para fazê-lo. De forma geral, uma região pode ser qualquer subconjunto do espaço paramétrico. Assim, temos que $\mathcal{A}_* = \{R \subset \Theta\}$. Para estudar este problema de decisão, consideraremos a seguinte função de utilidade

$$U(R(X), \theta) = \mathbb{I}(\theta \in R(X)) - k \int_{x \in R(x)} 1 \cdot d\theta_0 \qquad\qquad k > 0$$

O elemento $\mathbb{I}(\theta \in R(X))$ indica que é desejável que θ esteja na região de credibilidade, $R(X)$. Por outro lado, $\int_{x \in R(x)} d\theta_0$ indica que é desejável que a região $R(x)$ seja pequena, ou seja, tenha um volume pequeno.

Teorema 7.16. *Se $U(R(X), \theta) = \mathbb{I}(\theta \in R(X)) - k \int_{x \in R(x)} 1 \cdot d\theta_0$, então a melhor região de credibilidade é tal que $R(X) = \{\theta \in \Theta : f(\theta|X) \geq k\}$. Dizemos que $R(X)$ é um HPD (highest posterior density) de $f(\theta|X)$. Entre todas as regiões que têm uma dada probabilidade, $R(X)$ também é a menor delas.*

7.2.3 Regiões de credibilidade com credibilidade especificada

Na prática, é comum criar uma região de credibilidade de modo que a probabilidade (a posteriori) do parâmetro estar nessa região seja um valor pré-especificado $1 - \alpha$. Este valor de cobertura é chamado de *credibilidade* dessa região:

Definição 7.17. Dizemos que uma região $R \subset \Theta$ tem credibilidade $1 - \alpha$ se $\mathbb{P}(\theta \in R|\mathbf{x}) = 1 - \alpha$.

Assim, é comum escolher o formato da região desejada usando alguma das funções de perda apresentadas nesta seção, e então buscar, dentre essas regiões, aquela que tem a credibilidade desejada.

Exemplo 7.18. Considere novamente o Exemplo 2.2. Vamos construir intervalos de credibilidade para θ. Para tanto, lembre-se que $\theta|X = 20 \sim \text{Beta}(25, 15)$. O seguinte código ilustra como encontrar os intervalos de credibilidade investigados nesta seção, com constantes na utilidade escolhidas de modo que a credibilidade dos intervalos seja 95%. Neste caso, como a distribuição é próxima de ser simétrica, os três intervalos são muito próximos (Figura **??**).

```
a <- 25 # hiperparâmetro a posteriori
b <- 15 # hiperparâmetro a posteriori

grid_theta <- seq(0,1,length.out = 1000)
plot(grid_theta,
     dbeta(grid_theta,a,b),
     type="l",lwd=3,xlab = expression(theta),cex.lab=1.2,
     ylab="Densidade a posteriori")

# função auxiliar para determinar a cobertura de um intervalo
cobertura <- function(limites,a,b)
{
  return(pbeta(limites[2],a,b)-pbeta(limites[1],a,b))
}

intervalo1 <- function(K,a,b,grid_theta)
{
  soma <- sum(dbeta(grid_theta,a,b))
  lim_inf <- min(grid_theta[dbeta(grid_theta,a,b)>K])
  lim_sup <- max(grid_theta[dbeta(grid_theta,a,b)>K])
  return(c(lim_inf,lim_sup))
}

# cobertura real menos 1-alpha
funcao_para_minimizar <- function(K,alpha,a,b,grid_theta)
{
  limites <- intervalo1(K,a,b,grid_theta)
  return(abs(cobertura(limites,a,b)-(1-alpha)))
}

melhor_k <- optim(par = max(dbeta(grid_theta,a,b))/2,funcao_para_minimizar,alpha=0.05,
      a=a,b=b,
      grid_theta=grid_theta,
      lower=0,
      upper=max(dbeta(grid_theta,a,b)),method = "L-BFGS-B")
K <- melhor_k$par
limites1 <- intervalo1(K,a,b,grid_theta)
```

```r
intervalo2 <- function(alpha,a,b)
{
  # esperança e variância a posteriori
  lim_inf <- qbeta(alpha/2,a,b)
  lim_sup <- qbeta(1-alpha/2,a,b)
  return(c(lim_inf,lim_sup))
}
limites2 <- intervalo2(0.05,a,b)

intervalo3 <- function(K,a,b)
{
  # esperança e variância a posteriori
  esp_theta <- a/(a+b)
  var_theta <- a*b/((a+b+1)*(a+b)^2)

  lim_inf <- esp_theta-K*sqrt(var_theta)
  lim_sup <- esp_theta+K*sqrt(var_theta)
  return(c(lim_inf,lim_sup))
}

# cobertura real menos 1-alpha
funcao_para_minimizar <- function(K,alpha,a,b)
{
  limites <- intervalo3(K,a,b)
  return(abs(cobertura(limites,a,b)-(1-alpha)))
}

melhor_k <- optim(par = max(dbeta(grid_theta,a,b))/2,funcao_para_minimizar,alpha=0.05,
      a=a,b=b,
      lower=0,
      upper=max(dbeta(grid_theta,a,b)),method = "L-BFGS-B")
K <- melhor_k$par
limites3 <- intervalo3(K,a,b)

lines(limites1,c(0.05,0.05),lwd=3)
lines(limites2,c(0.15,0.15),lwd=3,col=2)
lines(limites3,c(0.25,0.25),lwd=3,col=4)
legend(0,5,col=c(1,2,4),legend = c("Tipo 1","Tipo 2","Tipo 3"),lwd=3)
```

Exercícios

Exercício 7.19. Dado μ, $X_1, \ldots, X_n$ são i.i.d. e $X_1 \sim N(\mu, 1)$. *A priori*, $\mu \sim N(0, 1)$.

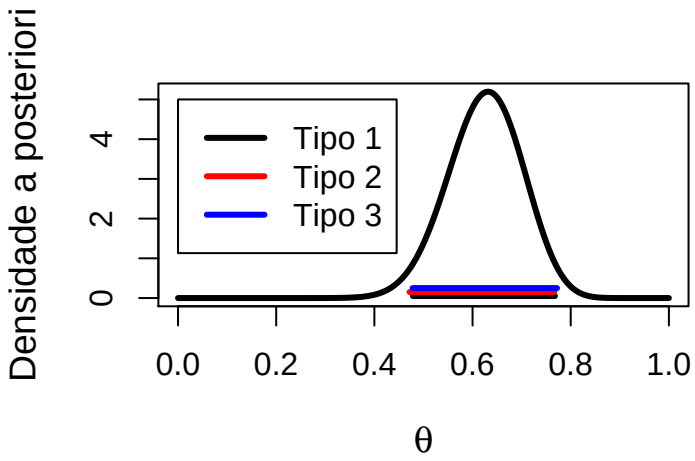

Figura 4: Intervalos de credibilidade

(a) Ache um estimador para μ usando cada utilidade que vimos em aula.

(b) Ache um intervalo para μ com credibilidade 95% para cada utilidade que vimos em aula.

(c) Ache o HPD para μ.

Exercício 7.20. Considere que $\theta|X \sim$ Beta(0.05,0.05).

(a) $f(\theta|X)$ é uma função côncava?

(b) Use um *software* estatístico para achar um intervalo de credibilidade para $\theta|X$.

(c) Use um *software* estatístico para achar um HPD para $\theta|X$.

Exercício 7.21 (Sugestão de Aline Tonon)**.** Considere o problema de determinar um intervalo de credibilidade, (a, b), com a função de utilidade, $U((a, b), \theta) = \frac{\mathbb{I}(\theta \in (a,b))}{b-a}$. Prove que, se (a^*, b^*) é um intervalo de credibilidade ótimo, então $f_{\theta|X}(a^*|X) = f_{\theta|X}(b^*|X)$.

Exercício 7.22. Considere o problema de determinar um intervalo de credibilidade, (a, b), com a função de utilidade, $U((a, b), \theta) = \frac{k-(a-\theta)_+-(b-\theta)_+}{b-a}$. Prove que, se (a^*, b^*) é um intervalo de credibilidade ótimo, então $F_{\theta|X}(a^*|X) = 1 - F_{\theta|X}(b^*|X)$.

7.3 Testes de hipótese

Um problema de teste de hipótese consiste em escolher uma entre duas proposições disjuntas e mutuamente exclusivas. Neste contexto, estas proposições recebem nomes especiais: uma delas é a hipótese nula e a outra é a hipótese alternativa. Denotaremos a hipótese nula por H_0 e a hipótese alternativa por H_1.

Exemplo 7.23. Uma máquina pode estar regulada ou desregulada. Caso a máquina esteja regulada, ela produz componentes com uma taxa de falha de 5%. Caso a máquina esteja desregulada, ela produz componentes com uma taxa de falha de 20%. Uma amostra de 100 produtos é selecionada e 9 deles são defeituosos.

Sejam $X_1, \ldots, X_n$ as indicadoras de que cada peça amostrada é defeituosa e θ é a taxa de falha atual da máquina, $\theta \in \{5\%, 20\%\}$. Consideramos que, dado θ, $X_1, \ldots, X_n$ são i.i.d. e $\sum_{i=1}^{100} X_i|\theta \sim$ Binomial$(100, \theta)$.

	$\theta \in H_0$	$\theta \in H_1$
d=0	0	-1
d=1	-c	0

Tabela 11: Descrição dos valores da utilidade $0-1-c$, $U(d,\theta)$.

Podemos estar interessados em testar a hipótese de que a máquina está regulada contra a hipótese de que ela está desregulada. Neste caso, $H_0 = \{\theta = 5\%\}$ e $H_1 = \{\theta = 20\%\}$.

Exemplo 7.24. Considere a mesma descrição do Exemplo 7.23. Seja X_{n+1} a indicadora de que uma nova peça produzida pela máquina é defeituosa. Podemos testar a hipótese de que esta peça é defeituosa. Neste caso, $H_0 = \{X_{n+1} = 0\}$ e $H_1 = \{X_{n+1} = 1\}$.

Exemplo 7.25. Considere o Exercício 5.1. Seja θ a indicadora de que a pessoa selecionada é uma mulher. Podemos testar $H_0 = \{\theta = 0\}$ contra $H_1 = \{\theta = 1\}$.

Exemplo 7.26. Em um experimento, larga-se uma pedra de uma determinada altura e mede-se a sua posição relativa ao ponto de largada a cada 1 segundo. Se Y_i denota a posição relativa da pedra no segundo i, assumimos que $Y_i = -g\frac{i^2}{2} + \epsilon_i$, onde ϵ_i são i.i.d. e $\epsilon_i \sim N(0,1)$, com τ^2 conhecido.

Podemos testar a hipótese $H_0 = \{g = 10\}$ contra a hipótese $H_1 = \{g \neq 10\}$. Semelhantemente, podemos testar a hipótese $H_0 = \{g \geq 10\}$ contra $H_1 = \{g < 10\}$.

Você deve escolher entre H_0 e H_1. Assim, ao tentar expressar um teste de hipótese como um problema de decisão, é comum definir $\mathcal{A}_* = \{0, 1\}$, onde 0 significa não rejeitar H_0 e 1 significa rejeitar H_0.

7.3.1 Hipóteses plenas

É comum o uso da função de utilidade $0-1-c$, descrita na tabela 11. Dizemos que rejeitar H_0 quando H_0 é verdadeiro é um erro do tipo I. Também, não rejeitar H_0 quando H_1 é verdadeiro é um erro do tipo II. Na tabela 11, o valor de c indica o quanto o erro de tipo I é mais grave que o erro de tipo II. Se $c > 1$, o erro de tipo I é mais grave que o erro de tipo II. Se $c < 1$, o erro de tipo I é menos grave que o erro de tipo II. Se $c = 1$, então ambos os erros são igualmente graves.

Segundo estas condições, podemos calcular uma regra ótima de decisões.

Teorema 7.27. *Em um problema de teste de hipótese com a utilidade dada pela tabela 11, as decisões ótimas, δ^*, são tais que*

$$\delta^*(x) = \begin{cases} 0, & \text{se } \mathbb{P}(\theta \in H_0|x) > (1+c)^{-1} \\ 1, & \text{se } \mathbb{P}(\theta \in H_0|x) < (1+c)^{-1} \end{cases}$$

Demonstração. Decorre do Teorema 6.12 que a decisão ótima, δ^*, é aquela que, para cada valor de X, maximiza $\mathbb{E}[U(d,\theta)|X]$. Para cada valor de X, somente existem 2 possíveis decisões (0 ou 1). Assim, a δ^* é tal que

$$\delta^*(x) = \begin{cases} 0, & \text{se } \mathbb{E}[U(0,\theta)|X] > \mathbb{E}[U(1,\theta)|X] \\ 1, & \text{se } \mathbb{E}[U(0,\theta)|X] < \mathbb{E}[U(1,\theta)|X] \end{cases} \tag{22}$$

Note que

$$\begin{aligned}\mathbb{E}[U(0,\theta)|X] &= \mathbb{E}[U(0,\theta)\mathbb{I}(\theta \in H_0)|X] + \mathbb{E}[U(0,\theta)\mathbb{I}(\theta \in H_1)|X] \\ &= \mathbb{E}[0 \cdot \mathbb{I}(\theta \in H_0)|X] + \mathbb{E}[-1 \cdot \mathbb{I}(\theta \in H_1)|X] \\ &= -1 \cdot \mathbb{E}[\mathbb{I}(\theta \in H_1)|X] = -\mathbb{P}(\theta \in H_1|X)\end{aligned}$$

$$\begin{aligned}\mathbb{E}[U(1,\theta)|X] &= \mathbb{E}[U(1,\theta)\mathbb{I}(\theta \in H_0)|X] + \mathbb{E}[U(1,\theta)\mathbb{I}(\theta \in H_1)|X] \\ &= \mathbb{E}[-c \cdot \mathbb{I}(\theta \in H_0)|X] + \mathbb{E}[0 \cdot \mathbb{I}(\theta \in H_1)|X] \\ &= -c \cdot \mathbb{E}[\mathbb{I}(\theta \in H_0)|X] = -c\mathbb{P}(\theta \in H_0|X)\end{aligned}$$

Utilizando as expressões derivadas acima na eq. (22), obtemos que $\delta^*(x) = 0$ quando

$$\begin{aligned}-\mathbb{P}(\theta \in H_1|X) &> -c\mathbb{P}(\theta \in H_0|X) \\ -(1 - \mathbb{P}(\theta \in H_0|X)) &> -c\mathbb{P}(\theta \in H_0|X) \\ -1 &> -(1+c)\mathbb{P}(\theta \in H_0|X) \\ \mathbb{P}(\theta \in H_0|X) &> (1+c)^{-1}\end{aligned}$$

Semelhantemente, a decisão ótima é 1 quando $\mathbb{P}(\theta \in H_0|X) < (1+c)^{-1}$. □

Em palavras, o Teorema 7.27 indica que a decisão ótima é não rejeitar H_0 quando sua probabilidade a posteriori é suficientemente grande. Em especial, se $c = 1$, temos que o erro tipo I é tão grave quanto o erro tipo II. Neste caso, a regra de decisão ótima segundo o Teorema 7.27 é não rejeitar H_0 quando sua probabilidade a posteriori é maior do que 0.5.

Exemplo 7.28. Considere o Exemplo 7.23. Também considere que, a priori, $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 0.5$, e o erro de tipo I é considerado 2 vezes menos grave que o erro tipo II. Assim $c = 0.5$. Neste caso,

$$\begin{aligned}\mathbb{P}\left(H_0|\sum_{i=1}^{100} X_i = 9\right) &= \frac{\mathbb{P}(H_0)P(\sum_{i=1}^{100} X_i = 9|H_0)}{\mathbb{P}(H_0)\mathbb{P}(\sum_{i=1}^{100} X_i = 9|H_0) + \mathbb{P}(H_1)\mathbb{P}(\sum_{i=1}^{100} X_i = 9|H_1)} \\ &= \frac{0.5\binom{100}{9}(0.05)^9(0.95)^{91}}{0.5\binom{100}{9}(0.05)^9(0.95)^{91} + 0.5\binom{100}{9}(0.2)^9(0.8)^{91}} \\ &\approx 0.96\end{aligned}$$

Como $\mathbb{P}\left(H_0|\sum_{i=1}^{100} X_i = 9\right) \approx 0.96 > 0.66 \approx (1+c)^{-1}$, a decisão ótima é não rejeitar H_0.

Exemplo 7.29. Considere o Exemplo 7.24. Também considere que, a priori, $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 0.5$, e o erro de

tipo I é considerado 9 vezes mais grave que o erro tipo II. Assim $c = 9$. Temos

$$\begin{aligned}\mathbb{P}\left(X_{n+1}=0|\sum_{i=1}^{n}X_i=9\right) =& \mathbb{P}\left(X_{n+1}=0,\theta=0.05|\sum_{i=1}^{n}X_i=9\right)+\mathbb{P}\left(X_{n+1}=0,\theta=0.2|\sum_{i=1}^{n}X_i=9\right)\\ =& \mathbb{P}(X_{n+1}=0|\theta=0.05)\mathbb{P}(\theta=0.05|\sum_{i=1}^{n}X_i=9)+\\ & \mathbb{P}(X_{n+1}=0|\theta=0.2)\mathbb{P}(\theta=0.2|\sum_{i=1}^{n}X_i=9)\\ \approx& 0.05\cdot 0.96+0.2\cdot 0.04=0.056\end{aligned}$$

Como $\mathbb{P}\left(X_{n+1}=0|\sum_{i=1}^{n}X_i=9\right)=0.056<10^{-1}=(1+c)^{-1}$, rejeitamos H_0.

Exemplo 7.30. Considere o Exemplo 7.25. Também considere que, a priori, $\mathbb{P}(H_0)=\mathbb{P}(H_1)=0.5$, e o erro de tipo I é considerado tão grave quanto o erro tipo II. Assim $c=1$. Pelo Exercício 5.1, sabemos que

$$\mathbb{P}(H_1|x)=\frac{1}{1+\exp\left(\frac{10x-1675}{18}\right)}$$

Pelo Teorema 7.27, rejeitamos H_0 quando $\mathbb{P}(H_1|x)>(1+c)^{-1}=0.5$. Assim, rejeitamos H_0 se

$$\begin{aligned}\frac{1}{1+\exp\left(\frac{10x-1675}{18}\right)}&>0.5\\ 1&>0.5+0.5\exp\left(\frac{10x-1675}{18}\right)\\ \exp\left(\frac{10x-1675}{18}\right)&<1\\ x&<167.5\end{aligned}$$

Dizemos que esse é um critério de decisão linear, dado que a regra de decisão pode ser explicada observando se uma reta em função dos dados é maior ou menor que um determinado valor.

Também note que, neste problema, a altura dos homens e mulheres seguem distribuições normais com média 170cm e 165cm e variâncias iguais. Assim, não é surpreendente que o critério de decisão é rejeitar que a pessoa é um homem se a sua altura for menor que 167.5 a média simples entre 170 e 165.

Corolário 7.31. *Em um problema de teste de hipótese com a utilidade dada pela tabela 11, as decisões ótimas, δ^*, são tais que*

$$\delta^*(x)=\begin{cases}0, & se\ \frac{\mathbb{P}(x|\theta\in H_0)}{\mathbb{P}(x|\theta\in H_1)}>\frac{\mathbb{P}(\theta\in H_1)}{c\mathbb{P}(\theta\in H_0)}\\ 1, & se\ \frac{\mathbb{P}(x|\theta\in H_0)}{\mathbb{P}(x|\theta\in H_1)}<\frac{\mathbb{P}(\theta\in H_1)}{c\mathbb{P}(\theta\in H_0)}\end{cases}$$

Demonstração. Podemos desenvolver a expressão $\mathbb{P}(\theta \in H_0|x) > (1+c)^{-1}$ da seguinte forma.

$$\begin{aligned}
\mathbb{P}(\theta \in H_0|x) &> (1+c)^{-1} \\
\frac{\mathbb{P}(\theta \in H_0)\mathbb{P}(x|\theta \in H_0)}{\mathbb{P}(\theta \in H_0)\mathbb{P}(x|\theta \in H_0) + \mathbb{P}(\theta \in H_1)\mathbb{P}(x|\theta \in H_1)} &> (1+c)^{-1} \\
\mathbb{P}(\theta \in H_0)\mathbb{P}(x|\theta \in H_0) &> (1+c)^{-1}(\mathbb{P}(\theta \in H_0)\mathbb{P}(x|\theta \in H_0) + \mathbb{P}(\theta \in H_1)\mathbb{P}(x|\theta \in H_1)) \\
c(1+c)^{-1}\mathbb{P}(\theta \in H_0)\mathbb{P}(x|\theta \in H_0) &> (1+c)^{-1}\mathbb{P}(\theta \in H_1)\mathbb{P}(x|\theta \in H_1) \\
\frac{\mathbb{P}(x|\theta \in H_0)}{\mathbb{P}(x|\theta \in H_1)} &> \frac{\mathbb{P}(\theta \in H_1)}{c\mathbb{P}(\theta \in H_0)}
\end{aligned}$$

Semelhantemente, $\mathbb{P}(\theta \in H_0|x) < (1+c)^{-1}$ se e somente se $\frac{\mathbb{P}(x|\theta \in H_0)}{\mathbb{P}(x|\theta \in H_1)} < \frac{\mathbb{P}(\theta \in H_1)}{c\mathbb{P}(\theta \in H_0)}$. O corolário está provado substituindo-se as expressões encontradas no Teorema 7.27. □

Exercícios

Exercício 7.32. Considere o Exemplo 7.23 e a função de utilidade dada pela tabela 11.

(a) Se $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 0.5$, como a decisão ótima varia de acordo com o valor de c?

(b) Se $c = 1$, como a decisão ótima varia em função de $\mathbb{P}(H_0)$?

(c) Como a decisão ótima varia conjuntamente em função de c e $\mathbb{P}(H_0)$?

Exercício 7.33. No 2^o turno de uma eleição presidencial, existem dois candidatos. Suponha que todo indivíduo dentre os eleitores votará em um dos dois candidatos. O vencedor é aquele que obtiver mais que 50% dos votos. Considere que a população é extremamente grande e n indivíduos foram selecionados com reposição. Dentre os indivíduos selecionados, a votarão no candidato 1 e $n-a$ votarão no candidato 2. Considere que, *a priori*, você acredita que todas as composições de votos são equiprováveis. Você deseja testar a hipótese de que o candidato 1 vencerá a eleição. Qual a sua regra de decisão se for tão grave cometer um erro do tipo I quanto um erro do tipo II?

Exercício 7.34. Para dois tipos de propaganda, M analisa o número de visualizações em uma página. Definimos $X_{i,j}$ como o número de visualizações da página no dia i a partir da propaganda j. Também definimos θ_j como o efeito da propaganda j no número de visualizações. Para este problema, consideramos que, dado θ_j, $X_{i,j}$ são i.i.d. e $X_{i,j} \sim \text{Poisson}(\theta_j)$. A priori, consideramos que θ_1 e θ_2 são i.i.d. e $\theta_j \sim \text{Gamma}(a_j, b_j)$. M deseja provar que a propaganda 2 é mais efetiva que a propaganda 1. M considera o erro tipo I tão grave quanto o erro tipo II. Como você analisaria esse caso?

Exercício 7.35. Existem 2 tipos de tratamento para evitar um tipo de contágio em uma plantação: 1 e 2. M acredita que o tratamento 2 é mais efetivo que o 1 para prevenir o contágio e, para testar esta hipótese, realiza um experimento. M submete n plantações a cada tipo de tratamento e coleta dados a respeito da taxa de contágio após o tratamento. Defina $Y_{i,j}$ como a taxa de contágio da i-ésima plantação submetida ao j-ésimo tratamento.

Para analisar os seus dados, M assume um modelo de ANOVA Bayesiano (Geinitz and Furrer, 2013). Considere que μ_j é a média da taxa de contágio de uma plantação submetida ao tratamento j. Assumimos que $Y_{i,j} = \mu_j + \epsilon_{i,j}$, onde $\epsilon_{i,j}$ são i.i.d. e $\epsilon_{i,j} \sim N(0, \tau_0^2)$. Também, μ é a média de todas as plantações submetidas a algum tratamento. Assumimos que, $\mu_j = \mu + \delta_j$, onde δ_j são i.i.d. e $\delta_j \sim N(0, \tau_1^2)$. Note que, até este ponto, o modelo especificado

é um modelo ANOVA tradicional com fatores aleatórios. Finalmente, M acredita *a priori* que $\mu \sim N(\nu, \tau_2^2)$. Considere que M conhece os valores de τ_0^2, τ_1^2 e τ_2^2.

Em termos do seu modelo, M deseja testar $H_0 : \mu_1 \geq \mu_2$ contra $H_1 : \mu_1 < \mu_2$.

7.3.2 Hipóteses precisas

Dizemos que uma hipótese é precisa quando ela tem dimensão menor que a do espaço paramétrico. Por exemplo, quando $\Theta = \mathbb{R}$, $H_0 : \theta = 0$ é uma hipótese precisa. Por outro lado, para o mesmo Θ: $H_0 : \theta \in (-\epsilon, \epsilon)$ não é uma hipótese precisa, uma vez que $(-\epsilon, \epsilon)$ tem mesma dimensão de Θ.

Um problema ligado a hipóteses precisas é o de que, para modelos estatísticos comumente utilizados, se H_0 é uma hipótese precisa, então $\mathbb{P}(H_0) = 0$. Ademais, $\mathbb{P}(H_0|x)$ também é 0 e, assim, se usarmos a função de utilidade dada pela tabela 11, então, para todo $c > 0$, $\mathbb{P}(H_0|x) < (1+c)^{-1}$. Portanto, de acordo com o Teorema 7.27, H_0 sempre será rejeitada.

Contudo, é comum que pesquisadores desejem testar uma hipótese precisa e não achem que é razoável sempre rejeitá-la. É possível justificar o seu raciocínio como coerente de acordo com a Inferencia Bayesiana? Existem duas respostas afirmativas frequentemente utilizadas para essa pergunta.

A primeira resposta consiste em utilizar um modelo tal que $\mathbb{P}(H_0) > 0$ ainda que H_0 tenha dimensão menor do que Θ. Por exemplo, considere que $H_0 : \theta = \theta_0$. É comum escolher a distribuição dos parâmetros e dados, $f(x, \theta)$, como:

$$f(x,\theta) = \begin{cases} p_0 f(x|\theta), & \text{se } \theta = \theta_0 \\ (1-p_0) f(\theta) f(x|\theta), & \text{caso contrário} \end{cases}$$

Em palavras, de acordo com esse modelo, $\mathbb{P}(H_0) = p_0 > 0$. Assim, obtemos que

$$\mathbb{P}(H_0|x) = \frac{f(x,\theta_0)}{f(x,\theta_0) + \int_{\theta \neq \theta_0} f(x,\theta) d\theta}$$
$$\frac{p_0 f(x|\theta_0)}{p_0 f(x|\theta_0) + (1-p_0) \int f(\theta) f(x|\theta) d\theta}$$

Portanto, se usarmos a utilidade dada pela tabela 11, decorre do teorema Teorema 7.27 que a decisão ótima é rejeitar H_0 quando $\frac{p_0 f(x|\theta_0)}{p_0 f(x|\theta_0) + (1-p_0) \int f(\theta) f(x|\theta) d\theta} < (1+c)^{-1}$. Similarmente, pelo Corolário 7.31, rejeita-se H_0 quando

$$\frac{f(x|\theta_0)}{\int f(\theta) f(x|\theta) d\theta} < \frac{1-p_0}{c p_0}$$

A expressão $\frac{f(x|\theta_0)}{\int f(\theta) f(x|\theta) d\theta}$ é comumente chamada de Fator de Bayes. Pelo critério de decisão encontrado, verificamos que, para valores pequenos do fator de Bayes, a hipótese nula é rejeitada. O quão pequeno deve ser o Fator de Bayes para que se rejeite H_0 depende tanto de p_0 quanto de c.

Corolário 7.36 (Fator de Bayes)**.** *Considere que θ segue uma distribuição contínua fora de H_0, que $H_0 : \theta = \theta_0$,*

que $\mathbb{P}(H_0) = p_0$ *e que usamos a função de perda dada pela tabela 11. Neste caso, rejeitamos* H_0 *se*

$$\frac{f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta} < \frac{1-p_0}{cp_0}$$

$\frac{f(x|\theta_0)}{\int f(\theta)f(x|\theta)d\theta}$ *é chamado de Fator de Bayes.*

Uma outra alternativa para testar hipóteses precisas consiste em considerar outra função de perda que não a da tabela 11 (Madruga et al., 2001). Esta é uma possível justificativa para o FBST (Full Bayesian Significance Test) (de Bragança Pereira and Stern, 1999). Ainda que a função de utilidade que gera esse teste não seja tratada nesse curso, é possível indicar a regra de decisão induzida por ela. Para obter a regra de decisão, constrói-se um HPD (Teorema 7.16) de probabilidade $1-\alpha$ para θ (α é um valor que é determinado pelo tomador de decisões. Quanto menor o valor de α, mais evidência é exigida para que não se rejeite H_0). Caso H_0 tenha interseção nula com o HPD, então rejeita-se H_0. Neste sentido, o FBST guarda relação com testes de hipótese obtidos por inversão de intervalo de confiança.

Teorema 7.37 (FBST)**.** *Existe uma função de utilidade (Madruga et al., 2001) tal que o melhor teste de hipótese é obtido da seguinte forma:*

1. *Constrói-se um HPD para* $\theta|X$ *com credibilidade* $1-\alpha$*.*
2. *Rejeita-se* H_0 *se e somente se nenhum ponto de* H_0 *está no HPD.*

Exercícios

Exercício 7.38. Considere que, dado μ, $X_1, \ldots, X_n$ são i.i.d., $X_i|\theta \sim N(\mu, \tau^2)$ e $\theta \sim N(0, \tau_0^2)$. Desejamos testar $H_0 : \mu = 0$ contra $H_1 : \mu \neq 0$.

(a) Obtenha um teste para H_0 usando o Fator de Bayes e tomando $p_0 = 95\%$ e $c = 1$.

(b) Obtenha um teste para H_0 usando o FBST a um nível $1 - \alpha = 95\%$.

(c) Compare os testes obtidos.

7.3.3 Coerência em testes de hipótese

Em algumas situações, realizamos simultaneamente diversos testes de hipótese a respeito de um parâmetro. Por simplicidade, uma hipótese do tipo $H_0 : \theta \in A$ será denotada nesta subseção apenas por A. No contexto de testes múltiplos é útil averiguar quais propriedades lógicas são satisfeitas pelos testes conjuntamente. Para estudar estas propriedades, Izbicki and Esteves (2015) considera uma notação capaz de expressar o resultado de vários testes de hipótese simultaneamente. $\mathcal{L}(A)(x)$ é a indicadora de que a hipótese A foi rejeitada a partir do dado x. A partir desta notação, Izbicki and Esteves (2015) define algumas propriedades lógicas que poderíamos esperar de testes de hipótese.

Definição 7.39 (Monotonicidade)**.** Se $A \subset B$, $\mathcal{L}(A)(x) \geq \mathcal{L}(B)(x)$.

Em palavras, se A é uma hipótese mais específica do que B, então, de acordo com a monotonicidade, a rejeição de B implica a rejeição de A. Intuitivamente, como A é mais específica do que B, acreditar em A implica acreditar em B. Assim, é estranho acreditar em A (não rejeitar A) mas não acreditar em B (rejeitar B).

Definição 7.40 (Invertibilidade)**.** Para todo A, $\mathcal{L}(A^c)(x) = 1 - \mathcal{L}(A)(x)$.

Assim, se A não é rejeitado, então A^c é rejeitado e vice-versa. Intuitivamente, como A e A^c são exaustivos e mutuamente exclusivos, então um e apenas um deles deveria ser aceito.

Definição 7.41 (Consonância com a intersecção)**.** Para todo A e B, se $\mathcal{L}(A) = 0$ e $\mathcal{L}(B) = 0$, então $\mathcal{L}(A \cap B) = 0$.

Portanto, se não rejeitamos A e não rejeitamos B, então não rejeitamos $A \cap B$. O raciocínio é análogo a: se $\theta \in A$ e $\theta \in B$, então $\theta \in A \cap B$.

Definição 7.42 (Consonância com a união)**.** Para todo A e B, se $\mathcal{L}(A) = 1$ e $\mathcal{L}(B) = 1$, então $\mathcal{L}(A \cup B) = 1$.

Portanto, se rejeitamos A e rejeitamos B, então rejeitamos $A \cup B$. Semelhantemente à consonância com a intersecção, o raciocínio é análogo a: se $\theta \notin A$ e $\theta \notin B$, então $\theta \notin A \cap B$.

Izbicki and Esteves (2015) ilustra que, ainda que essas propriedades sejam desejáveis, os testes usualmente utilizados falham uma ou mais delas. Por exemplo, em um ANOVA com efeitos α_1, α_2 e α_3, é possível não rejeitar $H_0 : \alpha_1 = \alpha_2$ e não rejeitar $H_0 : \alpha_2 = \alpha_3$ e, ainda assim, rejeitar $H_0 : \alpha_1 = \alpha_2 = \alpha_3$. Estas conclusões podem ser difíceis de explicar a um pesquisador. Também, para testes em que a não-rejeição da hipótese nula não pode ser interpretada como a aceitação desta, em geral a invertibilidade não é satisfeita. Por exemplo, quando os dados são pouco informativos para θ, então, nestes casos, não se rejeita nem A nem A^c.

De fato, estas ilustrações são consequências de um resultado ainda mais forte em Izbicki and Esteves (2015):

Teorema 7.43. *Sob algumas condições técnicas fracas, todo teste de hipótese que satisfaz as* 4 *propriedades indicadas é da seguinte forma:*

1. *Escolha um estimador pontual, $\hat{\theta}$.*
2. *Rejeite A se $\hat{\theta} \notin A$ e aceite A, caso contrário.*

O Teorema 7.43 nos indica três possíveis caminhos.

1. Aceitamos o teste de hipótese no Teorema 7.43 como o único que possa ser usado.
2. Deixamos de realizar testes de hipótese e substituímo-nos por procedimentos estatísticos que estejam mais próximos de nossos objetivos.
3. Revemos o juízo de razoabilidade das propriedades indicadas e achamos aquelas com as quais não concordamos. Como consequência devemos interpretar o resultado de um teste de hipótese de forma compatível com o fato de ele não satisfazer a propriedade escolhida.

Para efeitos práticos, a primeira posição é equivalente a aceitar que o valor do parâmetro é aquele assumido por uma estimativa pontual. Às vezes, esta pode ser uma boa escolha. Contudo, ela ignora a variabilidade do estimador pontual e, assim, potencialmente, pode não atender às demandas do pesquisador e até mesmo ser enganadora.

Por outro lado, ainda que aparentemente radical, a segunda posição tem sido sugerida com frequência. Testes de hipótese resumem toda a informação na amostra a uma única resposta: rejeitar ou não rejeitar a hipótese. Para muitos problemas, este resumo é insuficiente. Por exemplo, é possível argumentar que, além de rejeitar ou não rejeitar uma hipótese, um teste de hipótese deveria poder indicar outras respostas. Por exemplo, aceitar, não rejeitar ou não se posicionar em relação à hipótese. Assim, a não rejeição de uma hipótese poderia ser dividida

em dois casos: um que há evidência a favor da hipótese e outro em que não há evidência suficiente, nem para rejeitar, nem para aceitar a hipótese. Também, um intervalo de confiança ou de credibilidade indica os valores mais verossímeis para o parâmetro. Para muitos problemas, um intervalo traz mais informação relevante do que o resultado de um teste de hipótese.

A terceira posição pode envolver observar quais propriedades não são satisfeitas por um teste de hipótese proposto e questionar se esta falha prejudica os objetivos do pesquisador. A seguir, faremos esta análise para testes de hipótese obtidos pela tabela 11. Uma análise para o FBST e para fatores de Bayes pode ser encontrada em Izbicki and Esteves (2015).

Teorema 7.44. *Em geral, o teste de hipótese obtido pela tabela 11 satisfaz monotonicidade, satisfaz invertibilidade se e somente se $c = 1$, e não satisfaz consonância com a intersecção ou com a união.*

Demonstração. Obtivemos pelo Teorema 7.27 que a hipótese A não é rejeitada se $\mathbb{P}(A|X) > (1+c)^{-1}$. Portanto, se não rejeitamos A e $A \subset B$, então $\mathbb{P}(B|X) \geq \mathbb{P}(A|X) > (1+c)^{-1}$. Portanto, se não rejeitamos A e $A \subset B$, então não rejeitamos B. Conclua que o teste obtido pela tabela 11 satisfaz monotonicidade.

Similarmente, se $c = 1$, então decorre do Teorema 7.27 que a hipótese A não é rejeitada se $\mathbb{P}(A|X) > 0.5$. Assim, se $c = 1$ e A não é rejeitada, então $\mathbb{P}(A^c|X) = 1 - \mathbb{P}(A|X) < 0.5$. Portanto, nestas condições A^c é rejeitada. Também, se $c = 1$ e A é rejeitada, então $\mathbb{P}(A|X) < 0.5$. Portanto, $\mathbb{P}(A^c|X) = 1 - \mathbb{P}(A|X) > 0.5$. Assim, nestas condições, A^c não é rejeitada. Decorre, das últimas sentenças que, se $c = 1$, o teste de hipótese derivado da tabela 11 satisfaz monotonicidade. Finalmente, se $c \neq 1$ e $\mathbb{P}(A|X) = \mathbb{P}(A^c|X) = 0.5$, então A e A^c são ambos rejeitados ou ambos não rejeitados. Portanto, se $c \neq 1$, o teste decorrente da tabela 11 não satisfaz invertibilidade. Conclua que o teste decorrente da tabela 11 satisfaz invertibilidade se e somente se $c = 1$.

Considere que $A_1, \ldots, A_n$ são disjuntos, tais que $\cup_{i=1}^n A_i = \Omega$ e $\mathbb{P}(A_i|X) = n^{-1}$. Se tomarmos n suficientemente grande, $\mathbb{P}(A_i|X) < (1+c)^{-1}$. Portanto, todos os A_i são rejeitados. Contudo, $\mathbb{P}(\Omega|X) = 1 > (1+c)^{-1}$. Assim, Ω não é rejeitado. Portanto, existem $A_1, \ldots, A_n$ tais que A_i é rejeitado para todo i, mas $\cup_{i=1}^n A_i$ não é rejeitado. Portanto, o teste decorrente da tabela 11 não satisfaz consonância com a união.

Finalmente, considere os mesmos $A_1, \ldots, A_n$ usados no parágrafo anterior. Defina $A_{-i} = \cup_{j \neq i} A_j$. Note que $\mathbb{P}(A_{-i}|X) = \frac{n-1}{n}$. Portanto, tomando n suficientemente grande, $\mathbb{P}(A_{-i}|X) > (1+c)^{-1}$ e A_{-i} não é rejeitado. Contudo, $\cap_{i=1}^n A_{-i} = \emptyset$ e $\mathbb{P}(\emptyset|X) = 0 < (1+c)^{-1}$. Portanto, para todo i, A_{-i} não é rejeitado mas $\cap_{i=1}^n A_{-i}$ é rejeitado. Conclua que o teste decorrente da tabela 11 não satisfaz consonância com a intersecção. □

A partir do Teorema 7.27, sabemos que o teste derivado da tabela 11 não rejeita uma hipótese se e somente se sua probabilidade a posteriori é superior a uma dada constante. Assim, este teste de hipótese pode ser visto como um resumo que separa as hipóteses cuja probabilidade a posteriori ultrapassa esta constante, daquelas em que a constante não é ultrapassada. Para avaliar se este teste de hipótese é útil ao pesquisador, é necessário verificar se este resumo é suficiente para responder às perguntas deste. Como intuição para esta pergunta, o Teorema 7.44 mostra que é possível que a união de uma coleção de conjuntos ultrapasse o corte mas nenhum deles o ultrapasse. Também, é possível achar uma coleção de conjuntos tais que cada um deles ultrapassa o corte, mas a intersecção de todos não ultrapassa. Assim, o resumo dado pelo teste de hipótese na tabela 11 não satisfaz as consonâncias com a união e com a intersecção.

7.4 Princípio da verossimilhança*

Esta seção resume um resultado descoberto em Birnbaum (1962). Birnbaum estudava princípios que orientam o nosso ganho de informação em experimentos. Para poder lidar formalmente com este conceito, Birnbaum

define a função $\text{Inf}(X, x, \theta)$, a quantidade de informação que é ganha sobre θ ao observar x em um experimento, X. Birnbaum estuda quais propriedades Inf deve satisfazer para que represente a nossa intuição sobre o que é informação.

Uma das propriedades estudadas por Birnbaum foi o princípio da Suficiência. Lembre-se que $T(X)$ é uma estatística suficiente para θ se X e θ são independentes dado T. Em outras palavras, a partir de T, é possível gerar X usando um método de aleatorização que não depende de θ. Birnbaum argumenta que, como um método de aleatoriazação que não depende de θ não traz informação sobre θ, então T resume toda a informação sobre θ contida em X. Formalmente, o princípio da Suficiência diz que,

Definição 7.45 (princípio da Suficiência)**.** Se $T(X)$ é uma estatística suficiente para θ e x e x' são tais que $T(x) = T(x')$, então

$$\text{Inf}(X, x, \theta) = \text{Inf}(X, x', \theta)$$

.

Em palavras, se T resume x e x' atribuindo a eles o mesmo valor, então estes dois pontos devem trazer a mesma informação sobre θ.

Uma outra propriedade estudada por Birnbaum foi o princípio da Condicionalidade. Considere que X_0 e X_1 são dois experimentos e você decide qual deles irá realizar através do lançamento de uma moeda cujo resultado não depende de θ ou dos experimentos. Seja $Y \sim \text{Bernoulli}(p)$ o resultado do lançamento da moeda. Observe que o procedimento realizado pode ser denotado por X_Y, que é um experimento aleatorizado. O princípio da condicionalidade diz que o lançamento da moeda não deve trazer informação sobre θ e, mais especificamente, observar os valores da moeda e do experimento realizado no experimento aleatorizado deve trazer a mesma informação do que simplesmente observar o resultado do experimento realizado. Formalmente, o princípio da condicionalidade é definido da seguinte forma:

Definição 7.46 (princípio da Condicionalidade)**.** Se $Y \in \{0, 1\}$ é independente de (X_1, X_2, θ), então

$$\text{Inf}((Y, X_Y), (i, x), \theta) = \text{Inf}(X_i, x, \theta)$$

Birnbaum provou que ambos os princípios são satisfeitos se e somente se um terceiro princípio é satisfeito. Este é o princípio da Verossimilhança. O princípio da verossimilhança diz que, se x e y são dois pontos em experimentos, X e Y, com verossimilhanças proporcionais, $L_x(\theta) \propto L_y(\theta)$, então ambos os pontos devem trazer a mesma informação sobre θ. Formalmente,

Definição 7.47 (princípio da Verossimilhança)**.** Se X e Y são dois experimentos com possíveis observações x e y tais que $L_x(\theta) \propto L_y(\theta)$, então

$$\text{Inf}(X, x, \theta) = \text{Inf}(Y, y, \theta)$$

Em outras palavras, a informação trazida por um experimento é completamente resumida pela função de verossimilhança.

Teorema 7.48 (Teorema de Birnbaum)**.** *Inf satisfaz o princípio da Verossimilhança se e somente se Inf satisfaz os princípios da Suficiência e da Condicionalidade.*

Um argumento que pode ser articulado a partir do Teorema de Birnbaum é o seguinte: se você acha que os princípios da Suficiência e da Condicionalidade são razoáveis, então você deveria utilizar procedimentos inferenciais que satisfazem o princípio da Verossimilhança. Neste sentido, podemos mostrar que a probabilidade a posteriori satisfaz o princípio da verossimilhança.

Lema 7.49. *Se definirmos Inf*$(X, x, \theta) = f(\theta_0|X = x)$*, então Inf satisfaz o princípio da verossimilhança.*

Demonstração. Considere que $L_x(\theta_0) \propto L_y(\theta_0)$.

$$\begin{aligned} f(\theta_0|x) &\propto f(\theta_0)f(x|\theta_0) \\ &= f(\theta_0)L_x(\theta_0) \\ &\propto f(\theta_0)L_y(\theta_0) \\ &= f(\theta_0)f(y|\theta_0) \propto f(\theta_0|y) \end{aligned}$$

Como $f(\theta_0|x) \propto f(\theta_0|y)$ e ambas as funções integram 1, concluímos que $f(\theta_0|X = x) = f(\theta_0|Y = y)$. Graficamente, a fig. 1 é tal que a posteriori obtida depende apenas do formato da priori e da verossimilhança. □

Similarmente, provaremos em um exercício que a Estatística frequentista, em geral, não satisfaz o princípio da Verossimilhança. Contudo, isto não é um problema tão grave quanto se pode imaginar. De fato, vários estatísticos frequentistas indicaram razões pelas quais eles não acreditam que os princípios da Suficiência e da Condicionalidade, como descritos pelo Birnbaum, sejam razoáveis. Como consequência, o fato de não seguirem o princípio da Verossimilhança não os leva a uma contradição. Você acha os princípios razoáveis?

Exercícios

Exercício 7.50. Releia os três princípios descritos nesta Seção e tente descrevê-los em suas próprias palavras.

Exercício 7.51 (Wechsler et al. (2008))**.** Considere que $\theta \in (0, 1)$:

$$\begin{aligned} X_1|\theta &\sim \text{Binomial}(12, \theta) \\ X_2|\theta &\sim \text{Binomial-Negativa}(3, \theta) \end{aligned}$$

É verdade que $x_1 = 9$ e $x_2 = 9$ tem verossimilhanças proporcionais? Você está interessada na hipótese $H_0 : \theta \leq \frac{1}{2}$. Considere que Inf é o p-valor obtido no teste de hipótese. Calcule o p-valor obtido em cada um dos experimentos. Esta função de informação satisfaz o princípio da Verossimilhança?

Exercício 7.52 (DeGroot (1986)(p.353))**.** Defina Inf como sendo o estimador de máxima verossimilhança, isto é, se X é o experimento e θ a quantidade incerta de interesse, então $\text{Inf}(X, x, \theta) = \arg\max_{\theta_0} L_x(\theta_0)$. Inf satisfaz o princípio da verossimilhança?

8 Revisão sobre teoria da decisão e inferência bayesiana

Exercício 8.1. Considere que θ é a indicadora de que choverá hoje. A priori, você está indiferente entre a possibilidade de chover ou não chover. Para prever se choverá hoje, você toma uma medição da umidade relativa do ar, X. Considere que $X|\theta = 0 \sim N(0.2, 100)$ e $X|\theta = 1 \sim N(0.6, 100)$, onde 100 é a **precisão** da distribuição normal. Considere que sua utilidade é 1, caso sua previsão esteja correta, e 0, caso contrário.

(a) Determine $P(\theta = 1|X = x)$.

(b) Qual é a sua decisão ótima em função de X?

Exercício 8.2. Em uma corrida de cavalos, as apostas oferecidas são as seguintes:

- Se "Preguiça" vencer, é pago R\$10 para cada R\$1 que você apostar em "Preguiça".
- Se "Veloz" vencer, é pago R\$1.50 para cada R\$1 que você apostar em "Veloz".

(a) Você acredita que a probabilidade de "Veloz" vencer a corrida é 0.6 e a probabilidade de "Preguiça" vencer a corrida é 0.2. Se você pode realizar até R\$10 em apostas, qual é a melhor divisão de apostas para você?

(b) Como a resposta anterior mudaria se as probabilidades de "Veloz" e "Preguiça" vencer fossem p_v e p_p?

Exercício 8.3. O Sr. Bigode contratou você para ajudá-lo a determinar uma estratégia ótima para o seu negócio. O Sr. Bigode vende *hot dogs* por R\$5,00 durante partidas de futebol. Os *hot dogs* são perecíveis e, assim, caso o Sr. Bigode leve mais *hot dogs* do que a demanda, o excesso é desperdiçado. Os componentes necessários para fazer um *hot dog* custam R\$3,00. Considere que d é o número de *hot dogs* preparados pelo Sr. Bigode e θ é a demanda por *hot dogs* durante a partida de futebol. A utilidade do Sr. Bigode é dada por

$$U(d, \theta) = 5 \min(d, \theta) - 3d$$

(a) Mostre que

$$U(d) = \mathbb{E}[U(d, \theta)] = 5 \left(\sum_{i=0}^{d} i\mathbb{P}(\theta = i) + d\mathbb{P}(\theta > d) \right) - 3d$$

(b) Ache o valor de k tal que $U(d) > U(d-1)$ se e somente se $\mathbb{P}(\theta < d) < k$.

(c) Se $\theta \sim$ Geométrica(0.01), qual a decisão ótima para o Sr. Bigode?

Exercício 8.4. Considere um problema de estimação com dados em que $\theta \in \mathbb{R}^n$. Ache o melhor estimador quando

(a) $U(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_2^2 = (\hat{\theta} - \theta)^T(\hat{\theta} - \theta) = \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2$

(b) $U(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|_1 = \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|$

Exercício 8.5. Um estatístico deseja usar uma amostra para estimar um parâmetro, θ. Para tal, ele deve escolher o tamanho da amostra n e o estimador, $\hat{\theta}$. Considere que, uma vez escolhido n, a amostra, $X_1, \ldots, X_n$ é tal que dado θ as observações são i.i.d. e $X_1 \sim N(0, 1)$. Antes de obter a amostra, o estatístico acredita que $\theta \sim N(0, 1)$. Qual a escolha ótima do estatístico se $U((n, \hat{\theta}), \theta) = -cn - (\hat{\theta} - \theta)^2$?

Exercício 8.6. Considere que, dado θ, $X_1, \ldots, X_n$ são i.i.d. e $X_i \sim$ Uniforme$(0, \theta)$. A priori $\theta \sim$ Pareto(α, β). Se $\theta \sim$ Pareto(α, β), então $f(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}\mathbb{I}(\theta \geq \beta)$. Considere que $\alpha = \beta = 1$, $n = 10$ e $\max(x_1, \ldots, x_{10}) = 9$.

(a) Note que $\theta|X \sim$ Pareto(α^*, β^*). Ache os valores de α^* e β^*.

(b) Ache a distribuição acumulada da Pareto (α, β).

(c) Ache o estimador ótimo, $\hat{\theta}$, para θ segundo a utilidade

$$U(\hat{\theta}, \theta) = -|\hat{\theta} - \theta|$$

(d) Ache o estimador ótimo, $\hat{\theta}$, para θ segundo a utilidade

$$U(\hat{\theta}, \theta) = -(\hat{\theta} - \theta)^2$$

(e) Construa o intervalo de credibilidade ótimo, $[a, b]$, para θ segundo a utilidade

$$U([a, b], \theta) = -100((a - \theta)_+ + (\theta - b)_+) - 5(b - a)$$

(f) A distribuição a posteriori de θ é unimodal? Ache um HPD para θ com credibilidade 95%.

(g) Teste a hipótese $H_0 : \theta \geq 10$ usando a utilidade dada pela tabela 11 e tomando $c = 1$.

(h) Use o FBST ou o fator de bayes para testar $H_0 : \theta = 10$.

Exercício 8.7. Considere que, dado θ, $X_1, \ldots, X_n, X_{n+1}$ são i.i.d. e $X_i \sim \mathrm{N}(\theta, 1)$. A priori $\theta \sim \mathrm{N}(0, 1)$. Neste caso, $\theta | X_1, \ldots, X_n \sim \mathrm{N}(\frac{n\bar{X}}{n+1}, n + 1)$, onde $n + 1$ é a **precisão** da distribuição normal. Considere que $\bar{x} = 0.5$ e $n = 15$.

(a) Ache o estimador ótimo, $\hat{\theta}$, para θ segundo a utilidade

$$U(\hat{\theta}, \theta) = -|\hat{\theta} - \theta|$$

(b) Ache o estimador ótimo, $\hat{X}_{n+1}$, para X_{n+1} segundo a utilidade

$$U(\hat{X}_{n+1}, X_{n+1}) = -(\hat{X}_{n+1} - X_{n+1})^2$$

(c) Construa o intervalo de credibilidade ótimo, $[a, b]$, para θ segundo a utilidade

$$U([a, b], \theta) = -100((a - \theta)_+ + (\theta - b)_+) - 5(b - a)$$

(d) Ache o HPD para θ com credibilidade 95%.

(e) Teste a hipótese $H_0 : \theta \geq 0$ usando a utilidade $0/1/c$ e tomando $c = 1$.

(f) Use o FBST ($\alpha = 5\%$) para testar $H_0 : \theta = 0$.

Exercício 8.8. Em séries temporais, um modelo auto-regressivo AR(1) é tal que

$$X_n = \theta \cdot X_{n-1} + \epsilon_n,$$

ϵ_n é independente de $(X_1, \ldots, X_{n-1})$ e tal que $\epsilon_n \sim N(0, \tau^2)$. Considere que τ^2 é uma constante conhecida, que $X_0 = 0$, e que uma amostra, $x_1, \ldots, x_n$ foi observada.

(a) Ache $f(x_1, \ldots, x_n | \theta)$. Note que $(X_1, \ldots, X_n)$ não são i.i.d. dado θ.

(b) Se *a priori*, $\theta \sim N(\mu_0, \tau_0^2)$, qual a distribuição *a posteriori* de $\theta|x_1, \ldots, x_n$? Uma derivação similar pode ser encontrada no Exercício 5.7.

(c) Ache o estimador pontual para θ de acordo com a utilidade $U(\hat{\theta}, \theta) = -(\hat{\theta} - \theta)^2$.

(d) Construa um intervalo de credibilidade para θ com credibilidade 95%.

(e) Ache o estimador pontual para X_{n+1} de acordo com a utilidade $U(\hat{X}_{n+1}, X_{n+1}) = -(\hat{X}_{n+1} - X_{n+1})^2$.

Exercício 8.9 (Schervish (2012))**.** Você deseja testar $H_0 : \theta \geq k$ sob a utilidade $U(d, \theta) = d(\theta - k)_+ + (1-d)(k-\theta)_+$. Qual é a regra de decisão ótima?

Exercício 8.10. Se H_0 é uma hipótese precisa e θ segue uma distribuição contínua, descreva em palavras a razão de não testarmos H_0 pelo teste que vimos em classe para hipóteses plenas.

9 Estatística Bayesiana Computacional

Até o momento, trabalhamos com modelos tais que era possível calcular analiticamente a distribuição a posteriori para o parâmetro do modelo estatístico. Para tal, usamos a expressão obtida a partir do Teorema de Bayes:

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)d\theta}$$

Contudo, geralmente não é possível calcular diretamente $\int f(\theta)f(x|\theta)d\theta$ ou aproximar esta expressão por métodos determinísticos de integração (especialmente se o parâmetro tiver alta dimensionalidade). Assim, para realizar uma análise Bayesiana, é necessário desenvolver outras maneiras de avaliar a posteriori.

9.1 Método de Monte Carlo

Seja g uma função de θ, e assuma que desejamos aproximar $\mathbb{E}[g(\theta)|x]$. Por exemplo, podemos estar interessados em aproximar $\mathbb{E}[\theta|x]$. Considere que, de alguma forma, obtivemos uma amostra i.i.d. da posteriori de θ, $f(\theta|x)$. Denotaremos esta amostra por $T_1, \ldots, T_B$.

$$\begin{aligned} \mathbb{E}[g(T_i)] &= \int g(t)f(t|x)dt && T_i \text{ tem distribuição } f(t|x) \\ &= \int g(\theta)f(\theta|x)d\theta = \mathbb{E}[g(\theta)|x] \end{aligned}$$

Portanto, como $T_1, \ldots, T_B$ é uma amostra i.i.d., decorre da Lei dos Grandes Números que, se B for suficientemente grande, então

$$\hat{\mathbb{E}}[g(\theta)|x] := \frac{\sum_{i=1}^{B} g(T_i)}{B} \approx \mathbb{E}[g(\theta)|x]. \tag{23}$$

Assim, para estimar $\mathbb{E}[g(\theta)|x]$, basta (i) obter uma amostra i.i.d. da distribuição a posteriori e (ii) calcular a média dos $g(T_i)$'s.

Além disso, se $\mathbb{V}[g(\theta)|x] < \infty$, então decorre do Teorema do Limite Central que

$$\frac{\hat{E}[g(\theta)|x] - \mathbb{E}[g(\theta)|x]}{\sqrt{B}} \approx N(0, \sqrt{B}^{-1}\mathbb{V}[g(\theta)|x]) \tag{24}$$

Assim, se ψ é a função de distribuição acumulada da $N(0,1)$, então decorre das eqs. (24) e (26) que

$$\left[\hat{\mathbb{E}}[g(\theta)|x] + \psi^{-1}(0.5\alpha)\sqrt{B^{-1}\hat{\mathbb{V}}[g(\theta)|x]}, \hat{\mathbb{E}}[g(\theta)|x] + \psi^{-1}(1-0.5\alpha)\sqrt{B^{-1}\hat{\mathbb{V}}[g(\theta)|x]}\right] \tag{25}$$

é um intervalo de confiança aproximadamente $1-\alpha$ para $\mathbb{E}[g(\theta)|x]$. Note que decorre da eq. (23) que $\frac{\sum_{i=1}^{B} g(T_i)^2}{B} \approx \mathbb{E}[g(\theta)^2|x]$. Portanto, como podemos escrever $\mathbb{V}[g(\theta)|x]$ como $\mathbb{E}[g(\theta)^2|x] - \mathbb{E}[g(\theta)|x]^2$, então a variância de $g(\theta)$ pode ser aproximada como

$$\hat{\mathbb{V}}[g(\theta)|x] := \frac{\sum_{i=1}^{B} g(T_i)^2}{B} - \left(\frac{\sum_{i=1}^{B} g(T_i)}{B}\right)^2 \approx \mathbb{V}[g(\theta)|x]. \tag{26}$$

Assim,

Teorema 9.1 (Monte Carlo)**.** *Se $T_1, \ldots, T_B$ é uma amostra i.i.d. de $f(\theta|x)$ e g é uma função arbitrária, então $\hat{\mathbb{E}}[g(\theta)|x]$ aproxima $\mathbb{E}[g(\theta|x)]$ e um intervalo de confiança aproximadamente $1-\alpha$ para $\mathbb{E}[g(\theta)|x]$ é*

$$\left[\hat{\mathbb{E}}[g(\theta)|x] + \psi^{-1}(0.5\alpha)\sqrt{B^{-1}\hat{\mathbb{V}}[g(\theta)|x]}, \hat{\mathbb{E}}[g(\theta)|x] + \psi^{-1}(1-0.5\alpha)\sqrt{B^{-1}\hat{\mathbb{V}}[g(\theta)|x]}\right]$$

Observe que o Teorema 9.1 permite aproximar e avaliar o erro de aproximação para diversas quantidades importantes de $f(\theta|x)$. Por exemplo, tomando $g(\theta) = \theta^n$, é possível aproximar $\mathbb{E}[\theta^n|x]$. Similarmente, se $R \subset \theta$ e $g(\theta) = \mathbb{I}(\theta \in R)$, então é possível aproximar $\mathbb{E}[g(\theta)|x] = \mathbb{P}(\theta \in R|x)$. Podemos usar o seguinte código para implementar o Teorema 9.1 em R:

Algoritmo 9.2 (Monte Carlo)**.**

```
#########################################################################
## amostrador: funcao usada para obter a amostra do Monte Carlo,  ##
## deve retornar uma lista de amostras                            ##
## B: tamanho da amostra gerada.                                  ##
## g: funcao cujo valor esperado estamos interessados.            ##
## retorna: amostra de Monte Carlo e algumas de suas estatisticas ##
#########################################################################
monte_carlo <- function(amostrador, B, g, ...)
{
 amostra <- amostrador(B, ...)
 amostra_g <- sapply(amostra, g) #obtem g(x) para cada x em amostra
 est_g <- mean(amostra_g)
 var_g <- var(amostra_g)
 return(list(amostra=amostra,
             estimador=est_g,
             ic=c(est_g+qnorm(0.025)*sqrt(var_g/B),
                  est_g+qnorm(0.975)*sqrt(var_g/B))
       ))
}
```

Assim, utilizando o método de Monte Carlo, podemos obter diversas características relevantes da posteriori a partir de uma amostra desta. A pergunta que resta é: como obter uma amostra de $f(\theta|x)$ quando não conseguimos calcular esta quantidade analiticamente? Na sequência, veremos alguns métodos que têm essa finalidade.

9.1.1 O método da rejeição

O método da rejeição pode ser empregado para obter uma amostra da posteriori. Ele pode ser descrito nos seguintes passos:

1. Considere que $f(\theta)$ é a densidade da qual você quer simular e $\tilde{f}(\theta) \propto f(\theta)$.
2. Ache uma densidade, $h(\theta)$ e uma constante, M, tais que $\tilde{f}(\theta) \leq Mh(\theta)$.
3. Gere uma proposta, T, de $h(\theta)$.
4. Gere $U \sim$ Uniforme$(0, 1)$.
5. Se $U \leq \frac{\tilde{f}(T)}{Mh(T)}$, então retorne T como sua amostra. Caso contrário, retorne ao passo 3.

Código genérico para o método da rejeição é apresentado no Algoritmo 9.3, abaixo.

Algoritmo 9.3 (Método da rejeição)**.**

```
##############################################################################
## Código ilustrativo em R para o método da rejeição.                       ##
## B: tamanho da amostra a ser gerada                                       ##
## pf.avaliar: calcula o valor de uma função proporcional a f.              ##
## h.avaliar: calcula o valor da densidade h.                               ##
## h.gerar: gera uma variável aleatória com densidade h.                    ##
## M: a constante usada no método da rejeição.                              ##
## retorna: uma variável aleatória de densidade proporcional a pf.avaliar.  ##
##############################################################################
amostrador_rejeicao <- function(B, pf.avaliar, h.avaliar, h.gerar, M)
{
 amostra <- vector(mode="list", length=B)
 for(ii in 1:B)
 {
  T <- h.gerar()
  while(runif(1,0,1) > pf.avaliar(T)/(M*h.avaliar(T)))
  {
   T <- h.gerar()
  }
  amostra[[ii]] <- T
        }
 return(amostra)
}
```

Exemplo 9.4. Considere que você deseja simular de uma distribuição uniforme num círculo de raio centrado na origem. Os passos do método da rejeição são os seguintes:

1. Neste caso, $f(\theta) = \pi^{-1}(\theta_1^2 + \theta_1^2 \leq 1)$. Podemos tomar $\tilde{f}(\theta) = (\theta_1^2 + \theta_1^2 \leq 1)$, ou seja, $\tilde{f}(\theta) = \pi f(\theta)$.
2. Considere que você é capaz de simular de uma uniforme num quadrado de lado 2 centrado na origem. Neste caso, $h(\theta) = 4^{-1}\mathbb{I}(|\theta_1| \leq 1, |\theta_2| \leq 1)$. Note que $\tilde{f}(\theta) \leq 4h(\theta)$. Portanto, podemos tomar $M = 4$. Finalmente $\frac{\tilde{f}(\theta)}{4h(\theta)} = \tilde{f}(\theta)$.
3. Geramos T_1 e T_2 com densidade $h(\theta)$.
4. Geramos $U \sim$ Uniforme$(0, 1)$.
5. Se $T_1^2 + T_1^2 \leq 1$, então $\frac{\tilde{f}(\theta)}{4h(\theta)} = \tilde{f}(\theta) = 1$. Assim, para qualquer U gerado, retornaremos T. Se $T_1^2 + T_1^2 > 1$, então $\frac{\tilde{f}(\theta)}{4h(\theta)} = \tilde{f}(\theta) = 0$. Assim, para qualquer U gerado, retornaremos ao passo 3. Note que, neste caso, sequer precisaríamos ter gerado U.

Podemos descrever o algoritmo acima utilizando o seguinte código em R.

```
#########################################################
## retorna: uma variável aleatória de densidade      ##
## uniforme no círculo de raio 1 e centro na origem. ##
#########################################################
B = 100
h.gerar = function() runif(2,-1,1)
h.avaliar = function(x) ((x[1]^2 <= 1) & (x[2]^2 <= 1))/4
pf.avaliar = function(x) (x[1]^2 + x[2]^2 <= 1)/pi
M = 4/pi
dados = amostrador_rejeicao(B, pf.avaliar, h.avaliar, h.gerar, M)
```

O funcionamento do algoritmo acima é ilustrado na fig. 5. Os pontos indicam as 2000 propostas que foram geradas para obter uma amostra de tamanho 1586. Dentre estes pontos, os vermelhos foram aceitos e os azuis foram rejeitados. Neste caso, aproximadamente 80% dos pontos foram aceitos. Observe que pontos rejeitados são um desperdício computacional, dado que recursos são gastos para gerá-los, mas eles não fazem parte da amostra gerada. Neste sentido, gerar propostas de figuras geométricas mais próximas ao círculo diminuirá a rejeição e, portanto, a princípio, aumentará o rendimento do algoritmo. A dificuldade neste sentido é criar métodos para gerar propostas de outras figuras geométricas de forma tão eficiente quanto do quadrado.

Exemplo 9.5. Você deseja simular da densidade $f(\theta) = 6\theta(1-\theta)\mathbb{I}(\theta \in (0,1))$. Note que $f(\theta)$ é a densidade da Beta$(2, 2)$. Podemos tomar, por exemplo, $\tilde{f}(\theta) = \theta(1-\theta)\mathbb{I}(\theta \in (0,1))$. Assim, $\tilde{f}(\theta) \leq 0.25\mathbb{I}(\theta \in (0,1)) = 0.25 \cdot h(\theta)$, onde $h(\theta)$ é a densidade da Uniforme$(0, 1)$ e $M = 0.25$. Note que $\frac{\tilde{f}(\theta)}{Mh(\theta)} = 4\theta(1-\theta)$. Portanto, é possível simular de f pelo método da rejeição usando o seguinte código

```
#################################################################
## retorna: uma variável aleatória de distribuição Beta(2,2) ##
#################################################################
simula_beta_2_2_rejeicao <- function(B)
```

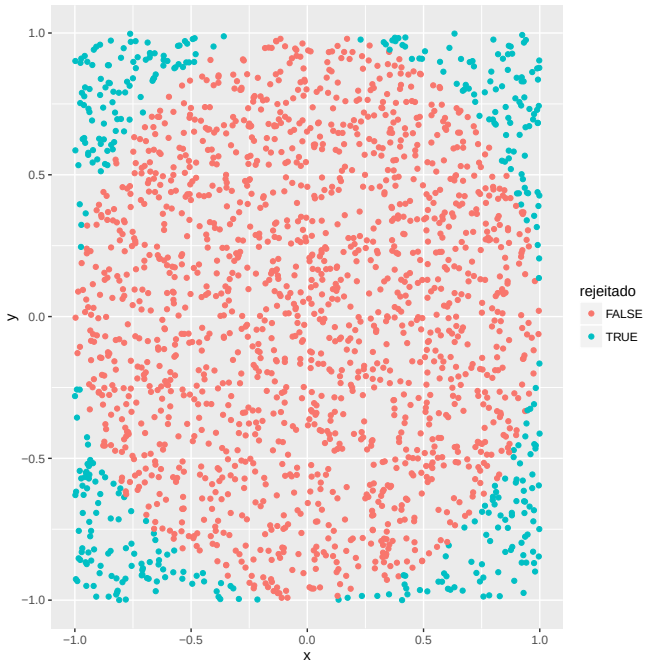

Figura 5: Amostra de 2000 propostas geradas pelo método da rejeição descrito no Exemplo 9.4.

```
{
 amostra <- vector(mode="list", length=B)
 for(ii in 1:B)
 {
  T <- runif(1)
  while(runif(1) > 4*T*(1-T)) T <- runif(1)
  amostra[[ii]] <- T
 }
 return(amostra)
}
```

O funcionamento do algoritmo acima é ilustrado na fig. 6. O eixo x dos pontos indicam as 2000 propostas que foram geradas para obter uma amostra de tamanho 1315. O eixo y dos pontos indicam as variáveis aleatórias uniformes geradas para determinar se as propostas eram rejeitadas. Dentre os pontos gerados, os vermelhos foram aceitos e os azuis foram rejeitados. Note que, como $\frac{\tilde{f}(\theta)}{h(\theta)} = 4\theta(1-\theta)$, T era rejeitado se $U > 4T(1-T)$. A fig. 6 também ilustra um outro aspecto do método da rejeição. Ao combinarmos T e U, obtemos uma distribuição uniforme em $[0,1]^2$. Ao remover os pontos azuis, os pontos vermelhos que restam distribuem-se de acordo com uma distribuição Beta(2,2) no eixo x.

Teorema 9.6. *Se θ foi gerado de acordo com o método da rejeição usando as funções $f(\theta)$, $\tilde{f}(\theta)$, $h(\theta)$ e M conforme a descrição no início deste capítulo, então θ tem distribuição com densidade f.*

Demonstração. Considere que N é o número de propostas geradas até a primeira aceitação, que $T_1, \ldots, T_i, \ldots$ é um conjunto de propostas e $U_1, \ldots, U_i, \ldots$ é um conjunto de uniformes. Note que $A_i = \mathbb{I}\left(U_i \leq \frac{\tilde{f}(T_i)}{Mh(T_i)}\right)$ são i.i.d.

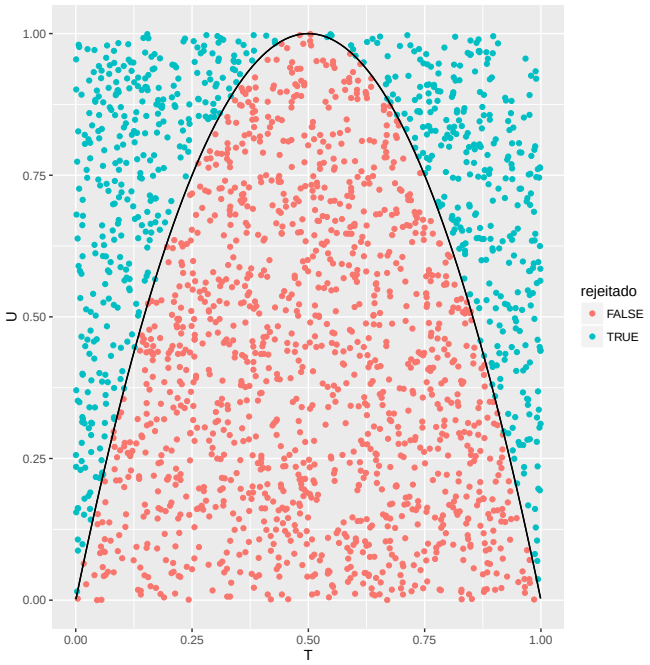

Figura 6: Amostra de 2000 propostas geradas pelo método da rejeição descrito no Exemplo 9.5.

e

$$\mathbb{P}\left(U_i \leq \frac{\tilde{f}(T_i)}{Mh(T_i)}\right) = \int_{-\infty}^{\infty} \frac{\tilde{f}(t)}{Mh(t)} h(t)dt$$
$$= M^{-1} \int_{-\infty}^{\infty} \tilde{f}(t)dt := p_1$$

Portanto, como $N = \min\left(i : \mathbb{I}(U_i \leq \frac{\tilde{f}(T_i)}{Mh(T_i)}) = 1\right)$, então $N \sim$ Geométrica(p_1). Assim, para todo $B \subset \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}(\theta \in B) &= \sum_n \mathbb{P}(\theta \in B, N = n) \\
&= \sum_n \mathbb{P}(\cap_{i=1}^{n-1} A_i^c \cap (A_n \cap T_n \in B)) \\
&= \sum_n \mathbb{P}(\cap_{i=1}^{n-1} A_i^c)\mathbb{P}(A_n \cap T_n \in B) \\
&= \sum_n (1-p_1)^{n-1} \int_B \frac{\tilde{f}(t)}{Mh(t)} h(t)dt \\
&= \int_B M^{-1}\tilde{f}(t)dt \sum_n (1-p_1)^{n-1} \\
&= \frac{\int_B M^{-1}\tilde{f}(t)dt}{p_1} \\
&= \frac{M^{-1}\int_B \tilde{f}(t)dt}{M^{-1}\int_{-\infty}^{\infty} \tilde{f}(t)dt} \\
&= \int_B \frac{\tilde{f}(t)}{\int_{-\infty}^{\infty} \tilde{f}(t)dt} dt \\
&= \int_B f(t)dt
\end{aligned}$$

□

Exercícios

Exercício 9.7. No Exemplo 9.4, usamos propostas de um quadrado de lado 2 e centro na origem. Poderíamos ter implementado o algoritmo da rejeição usando propostas de um quadrado de lado 1? E de um quadrado de lado 3? Esboce estas três figuras acompanhadas do círculo de raio 1 e centro na origem. Compare as propostas sugeridas em relação à sua eficiência computacional.

Exercício 9.8. Escreva um código para simular da distribuição uniforme em uma esfera de raio 1 e centro 0. Use o Método de Monte Carlo para estimar a distância média de um ponto amostrado à origem.

Exercício 9.9. Escreva um código para simular de uma distribuição uniforme em um círculo de raio r e centro c. Qual é a relação entre o raio do círculo e a distância média de um ponto amostrado ao seu centro? Exiba uma figura que ilustre essa relação.

Exercício 9.10. Escreva código para simular de uma Beta(a, b) a partir de uma Uniforme$(0, 1)$. Estime a média da Beta(a, b) pelo método de Monte Carlo. Compare as taxas de rejeição do algoritmo proposto para alguns valores de a e de b. Qual é a relação entre esses valores e a eficiência computacional do algoritmo proposto?

Exercício 9.11. Sob quais condições é possível usar o método da rejeição para simular de uma Beta(a, b) a partir de uma Beta(a^*, b^*)?

Exercício 9.12. Sob quais condições é possível usar o método da rejeição para simular de uma Gama(a, b) a partir de uma Exponencial(c)? Escreva o código para realizar esta simulação.

9.2 Método de Monte Carlo via cadeias de Markov

Foi ilustrado na seção passada que nem sempre é possível obter de forma eficiente uma sequência de variáveis aleatórias i.i.d. com distribuição dada pela posteriori. Assim, a aplicação do método de Monte Carlo é inviável.

Para contornar este problema, esta seção apresenta o método de Monte Carlo via cadeias de Markov. Neste método, ao invés de você gerar uma sequência i.i.d., você gerará uma cadeia de Markov. A seção 9.2.1 revisa cadeias de Markov e a seção 9.2.2 apresenta o algoritmo de Metropolis-Hastings, uma implementação geral do método de Monte Carlo via cadeias de Markov.

9.2.1 Cadeias de Markov

Definição 9.13 (Cadeia de Markov)**.** Seja $(T_n)_{n\in\mathbb{N}}$ uma sequência de variáveis discretas. Dizemos que $(T_n)_{n\in\mathbb{N}}$ é uma Cadeia de Markov se, para todo $n \in \mathbb{N}$,

$$\mathbb{P}(T_{n+1} = t_{n+1}|T_n = t_n, T_{n-1} = t_{n-1}, \ldots, T_1 = t_1) = \mathbb{P}(T_{n+1} = t_{n+1}|T_n = t_n)$$

Isto é, dado o passado inteiro da cadeia de Markov, $(T_n = t_n, T_{n-1} = t_{n-1}, \ldots, T_1 = t_1)$, apenas o estado imediatamente anterior, $(T_n = t_n)$, é usado para determinar o próximo estado (T_{n+1}).

Definição 9.14 (Cadeia homogênea)**.** Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov. Dizemos que $(T_n)_{n\in\mathbb{N}}$ é homogênea se, para todo n, $\mathbb{P}(T_{n+1} = t_{n+1}|T_n = t_n) = f(t_n, t_{n+1})$. Isto é, a distribuição do próximo estado da cadeia depende do estado anterior, mas não do índice atual. Neste caso, denotamos a função de transição da cadeia, $\mathbb{P}(T_{n+1} = t_{n+1}|T_n = t_n)$ por $p(t_{n+1}|t_n)$. A partir deste ponto, consideraremos apenas cadeias de Markov homogêneas.

Definição 9.15 (Matriz de transição)**.** Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov. $\mathbb{P}$ é a matriz de transição para $(T_n)_{n\in\mathbb{N}}$ se $\mathbb{P}_{i,j} = \mathbb{P}(T_{n+1} = j|Tn = i) = p(j|i)$. Isto é, a intersecção entre a i-ésima linha de $\mathbb{P}$ e a j-ésima coluna contém a probabilidade de a cadeia ir do estado i para o estado j.

Exemplo 9.16. Considere uma cadeia de Markov em $\{0, 1\}$ com a seguinte matriz de transição:

$$\mathbb{P} = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Isto é, $\mathbb{P}(T_1 = 0|T_0 = 0) = \frac{1}{3}$, $\mathbb{P}(T_1 = 1|T_0 = 0) = \frac{2}{3}$, $\mathbb{P}(T_1 = 0|T_0 = 1) = \frac{1}{2}$ e $\mathbb{P}(T_1 = 1|T_0 = 1) = \frac{1}{2}$.

Definição 9.17 (Distribuição estacionária)**.** Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov com matriz de transição, $\mathbb{P}$. Para cada estado, i, definimos

$$\mu(i) = \lim_{n\to\infty} \mathbb{P}(T_n = i|T_0 = t_0) = \lim_{n\to\infty} (\mathbb{P}^n)_{t_0,i}$$

Se o limite for bem definido, dizemos que μ é a distribuição estacionária da Cadeia. Podemos interpretar $\mu(i)$ como a probabilidade de que, após um tempo suficientemente grande ter se passado, a cadeia esteja no estado i.

A definição de distribuição estacionária envolve um limite de multiplicações de matrizes. Achar este limite por força bruta pode ser muito difícil. Assim, a seguir, definimos uma distribuição mais fácil de calcular e provamos que, sob certas circunstâncias, ela é equivalente à distribuição estacionária.

Definição 9.18 (Distribuição invariante)**.** Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov com matriz de transição, $\mathbb{P}$. Seja μ um vetor não-negativo tal que $\sum_i \mu(i) = 1$. μ é a distribuição invariante para $\mathbb{P}$ se

$$\mu\mathbb{P} = \mu$$

Isto é, se a sua informação sobre o estado atual da cadeia é μ, então sua informação para o próximo estado da cadeia também é μ. Desta forma, μ é um ponto fixo de $\mathbb{P}$.

Exemplo 9.19. Considere o Exemplo 9.16. A distribuição invariante deve satisfazer:

$$\mu = \mu \cdot \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Do que obtemos o sistema linear:

$$\begin{cases} \frac{\mu(0)}{3} + \frac{\mu(1)}{2} = \mu(0) \\ \frac{2\mu(0)}{3} + \frac{\mu(1)}{2} = \mu(1) \\ \mu(0) + \mu(1) = 1 \end{cases}$$

Portanto, $\mu(0) = \frac{3}{7}$ e $\mu(1) = \frac{4}{7}$.

Definição 9.20 (Distribuição reversível)**.** Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov com matriz de transição $\mathbb{P}$. Seja μ um vetor positivo tal que $\sum_i \mu(i) = 1$. μ é a distribuição reversível para $\mathbb{P}$ se, para todos estados i e j, $\mu(i)\mathbb{P}_{i,j} = \mu(j)\mathbb{P}_{j,i}$.

Exemplo 9.21. Considere o Exemplo 9.16. Seja $\mu(0) = \frac{3}{7}$ e $\mu(1) = \frac{4}{7}$.

$$\begin{aligned} \mu(0)\mathbb{P}_{0,1} &= \frac{3}{7} \cdot \frac{2}{3} = \frac{2}{7} \\ &= \frac{4}{7} \cdot \frac{1}{2} = \mu(1)\mathbb{P}_{1,0} \end{aligned}$$

Portanto, μ é reversível.

Lema 9.22. *Se μ é a distribuição reversível para $\mathbb{P}$, então μ é a distribuição invariante para $\mathbb{P}$.*

Demonstração. Considere que μ é a distribuição reversível para $\mathbb{P}$. Portanto, para todo j,

$$\begin{aligned} (\mu\mathbb{P})_j &= \sum_i \mu(i)\mathbb{P}(i,j) \\ &= \sum_i \mu(j)\mathbb{P}(j,i) && \text{Definição 9.20} \\ &= \mu(j)\sum_i \mathbb{P}(j,i) \\ &= \mu(j) && \text{Definição 9.15} \end{aligned}$$

Isto é, para todo j, $(\mu\mathbb{P})_j = \mu(j)$. Conclua que $\mu\mathbb{P} = \mu$, ou seja, μ é a distribuição invariante (Definição 9.18). □

Exemplo 9.23. Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov de matriz de transição $\mathbb{P}$. Considere, para quaisquer estados i e j, $\mathbb{P}_{i,j} = \mathbb{P}_{j,i}$. Defina $\mu(i) = \frac{1}{N}$, onde N é o número total de estados. Observe que $\sum_i \mu(i) = 1$ e também,

$$\begin{aligned}\mu(i)\mathbb{P}_{i,j} &= \frac{1}{N}\cdot \mathbb{P}_{i,j} \\ &= \frac{1}{N}\cdot \mathbb{P}_{j,i} = \mu(j)\mathbb{P}_{j,i}\end{aligned}$$

Portanto μ é reversível. Decorre do Lema 9.22 que μ é invariante. Portanto, sob a condição de que $\mathbb{P}_{i,j} = \mathbb{P}_{j,i}$, a distribuição uniforme é invariante.

Teorema 9.24. *Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov de matriz de transição $\mathbb{P}$. Sob algumas condições fracas de regularidade sobre $\mathbb{P}$, se μ é a distribuição invariante para $\mathbb{P}$, então μ é a distribuição estacionária para $\mathbb{P}$.*

Teorema 9.25 (Lei dos grandes números e Teorema do limite central para cadeias de Markov (Jones et al., 2004))**.** *Seja $(T_n)_{n\in\mathbb{N}}$ uma cadeia de Markov, μ a distribuição estacionária de $(T_n)_{n\in\mathbb{N}}$ e g uma função contínua. Sob algumas condições fracas de regularidade,*

$$\begin{aligned}\frac{\sum_{i=1}^B g(T_i)}{B} &\approx \mathbb{E}_\mu[g(T)] = \int g(\theta)\mu(\theta)d\theta \\ \mathbb{V}\left[\frac{\sum_{i=1}^B g(T_i)}{B}\right] &\approx \frac{\mathbb{V}[g(T_i)] + 2\sum_{k=1}^\infty Cov[g(T_i), g(T_{i+k})]}{B} \\ \frac{\frac{\sum_{i=1}^B g(T_i) - E_\mu[g(T)]}{B}}{\sqrt{Var\left[\frac{\sum_{i=1}^B g(T_i)}{B}\right]}} &\approx N(0,1)\end{aligned}$$

Portanto,

$$\left[\frac{\sum_{i=1}^B g(T_i)}{B} + \psi^{-1}(\alpha/2)\sqrt{\mathbb{V}\left[\frac{\sum_{i=1}^B g(T_i)}{B}\right]}, \frac{\sum_{i=1}^B g(T_i)}{B} - \psi^{-1}(\alpha/2)\sqrt{\mathbb{V}\left[\frac{\sum_{i=1}^B g(T_i)}{B}\right]}\right]$$

É um intervalo de confiança $1-\alpha$ para $\int g(\theta)\mu(\theta)d\theta$.

9.2.2 O algoritmo de Metropolis-Hastings

O Teorema 9.25 mostra que, se você construir uma cadeia de Markov, $T_1, \ldots, T_B$, com distribuição estacionária $f(\theta|x)$, então $\frac{\sum_{i=1}^B g(T_i)}{B}$ é um estimador consistente para $\int g(\theta)f(\theta|x)d\theta = \mathbb{E}[g(\theta)|X]$. O Teorema 9.25 também permite que você avalie a margem de erro para esse estimador a partir de $\mathbb{V}\left[\frac{\sum_{i=1}^B g(T_i)}{B}\right]$. Assim, se você obtiver uma cadeia de Markov com distribuição invariante igual à posteriori, $f(\theta|x)$, então poderá fazer Inferência Estatística. Nesta subseção, você estudará como obter esta cadeia.

Para tal, considere o algoritmo de Metropolis-Hastings, descrito a seguir:

Algoritmo 9.26 (Metropolis-Hastings)**.**

1. Defina um valor arbitrário para T_1.
2. Para i de 2 até B:
   a. Obtenha T_i^* da distribuição $h(t_i^*|T_{i-1})$, h é uma distribuição condicional arbitrária.

b. Obtenha $R_i = \frac{f(\theta=T_i^*|x)h(T_{i-1}|T_i^*)}{f(\theta=T_{i-1}^*|x)h(T_i^*|T_{i-1})}$.

c. Gere $U_i \sim \text{Uniforme}(0,1)$.

d. Defina:

$$T_i = \begin{cases} T_i^* & \text{, se } U_i \leq R_i \\ T_{i-1} & \text{, caso contrário.} \end{cases}$$

Em primeiro lugar, note que, para obter R_i, utiliza-se $f(\theta|x)$. Contudo, você está estudando métodos computacionais justamente por não ser possível calcular a posteriori analiticamente. Assim, poderá parecer que o algoritmo acima não é operacional. Contudo, observe que:

$$\begin{aligned} R_i &= \frac{f(\theta = T_i^*|x)h(T_{i-1}|T_i^*)}{f(\theta = T_{i-1}^*|x)h(T_i^*|T_{i-1})} \\ &= \frac{\frac{f(\theta=T_i^*)f(x|\theta=T_i^*)}{f(x)}h(T_{i-1}|T_i^*)}{\frac{f(\theta=T_{i-1}^*)f(x|\theta=T_{i-1}^*)}{f(x)}h(T_i^*|T_{i-1})} \\ &= \frac{f(\theta = T_i^*)f(x|\theta = T_i^*)h(T_{i-1}|T_i^*)}{f(\theta = T_{i-1}^*)f(x|\theta = T_{i-1}^*)h(T_i^*|T_{i-1})} \end{aligned} \tag{27}$$

Assim, uma vez que R_i envolve uma razão de posterioris, é possível calculá-lo utilizando $f(\theta)f(x|\theta)$ ao invés de $f(\theta|x)$. Em geral, enquanto que é possível obter a primeira quantidade analiticamente, não é possível obter a segunda. Assim, calcular R_i a partir da eq. (27) torna o algoritmo operacional.

Também note que, a cada iteração, T_i é gerado utilizando-se apenas as variáveis aleatórias T_{i-1} e U_i. Assim, $T_1, \ldots, T_B$ é uma cadeia de Markov. Além disso, você também pode demonstrar que $f(\theta|x)$ é a distribuição invariante desta cadeia. Para tal, mostrará primeiro que $f(\theta|x)$ é a distribuição reversível de $T_1, \ldots, T_B$.

Lema 9.27. *$f(\theta|x)$ é a distribuição reversível para a cadeia de Markov gerada pelo Algoritmo 9.26.*

Demonstração. Para quaisquer estados, a e b,

$$\begin{aligned} f(\theta = a|x)f(T_i = b|T_{i-1} = a) &= f(\theta = a|x)f(T_i^* = b|T_{i-1} = a)f(T_i = b|T_i^* = b, T_{i-1} = a) \\ &= f(\theta = a|x)h(b|a)f\left(U_i \leq \frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)}\right) \\ &= f(\theta = a|x)h(b|a)\min\left(1, \frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)}\right) \end{aligned} \tag{28}$$

Note que, se $\frac{f(\theta=b|x)h(a|b)}{f(\theta=a|x)h(b|a)} = 1$, então:

$$\begin{aligned} f(\theta = a|x)f(T_i = b|T_{i-1} = a) &= f(\theta = a|x)h(b|a) && \text{eq. (28)}, \frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)} = 1 \\ &= f(\theta = b|x)h(a|b) && \frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)} = 1 \\ &= f(\theta = b|x)f(T_i = a|T_{i-1} = b) && \text{eq. (28)}, \frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)} = 1 \end{aligned}$$

Agora, sem perda de generalidade, suponha que $\frac{f(\theta=b|x)h(a|b)}{f(\theta=a|x)h(b|a)} < 1$. Neste caso,

$$
\begin{aligned}
f(\theta = a|x)f(T_i = b|T_{i-1} = a) &= f(\theta = a|x)h(b|a)\frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)} && \text{eq. (28)}, \frac{f(\theta = b|x)h(a|b)}{f(\theta = a|x)h(b|a)} < 1 \\
&= f(\theta = b|x)h(a|b) \\
&= f(\theta = b|x)h(a|b)\min\left(1, \frac{f(\theta = a|x)h(b|a)}{f(\theta = b|x)h(a|b)}\right) && \frac{f(\theta = a|x)h(b|a)}{f(\theta = b|x)h(a|b)} > 1 \\
&= f(\theta = b|x)f(T_i = a|T_{i-1} = b) && \text{eq. (28)}
\end{aligned}
$$

Assim, usando a Definição 9.20, conclua dos dois casos anteriormente analisados que $f(\theta|x)$ é a distribuição reversível para a cadeia de Markov gerada pelo Algoritmo 9.26. □

Teorema 9.28. *$f(\theta|x)$ é a distribuição invariante para a cadeia de Markov gerada pelo Algoritmo 9.26.*

Demonstração. Decorre dos Lemas 9.22 e 9.27. □

Como resultado do Teorema 9.28, você obteve que o Algoritmo 9.26 gera uma cadeia de Markov, $T_1, \ldots, T_B$, de medida invariante $f(\theta|x)$. Portanto, utilizando o Teorema 9.25, você pode utilizar $T_1, \ldots, T_B$ para aproximar $\mathbb{E}[g(\theta)|x]$, para qualquer g contínua. Assim, você pode resolver problemas de Inferência, ainda que não consiga calcular a posteriori analiticamente. O Exemplo 9.29, a seguir, ilustra uma implementação genérica do Algoritmo 9.26 no R. Vários exemplos a seguir utilizam esta implementação para ilustrar o funcionamento do Algoritmo 9.26.

Exemplo 9.29 (Implementação genérica do Algoritmo 9.26 no R)**.**

```
#############################################################################
## retorna: um vetor que forma uma cadeia de Markov.                       ##
## B: o tamanho da cadeia retornada.                                       ##
## inicio: a posicao inicial da cadeia retornada.                          ##
## rprop: uma funcao que recebe como argumento elementos da cadeia e retorna uma  ##
## proposta gerada a partir deste elemento                                 ##
## ldprop: o log da densidade condicional da proposta utilizada.           ##
## ldpost: o log de uma funcão proporcional à                              ##
## densidade correspondente à distribuição invariante da cadeia gerada.    ##
#############################################################################
metropolis <- function(B, start, rprop, ldprop, ldtgt)
{
  chain <- as.list(rep(NA, B))
  chain[[1]] <- start
  for(ii in 2:B)
  {
    prop <- rprop(chain[[ii-1]])
    lratio <- ldtgt(prop)-ldtgt(chain[[ii-1]])+
              ldprop(prop,chain[[ii-1]])-
              ldprop(chain[[ii-1]],prop)
    if(log(runif(1)) <= lratio) chain[[ii]] <- prop
    else chain[[ii]] <- chain[[ii-1]]
```

```
  }
  return(chain)
}
```

Exemplo 9.30. Digamos que $\mu \sim \mathrm{T}_1$, isto é, uma distribuição T com 1 grau de liberdade (também conhecida como distribuição Cauchy). Também, dado μ, $X_1, \ldots, X_n$ são i.i.d. e $X_1 \sim N(\mu, 1)$. Você deseja estimar μ minimizando a perda quadrática. Note que:

$$\begin{aligned} f(\mu|\mathbf{x}) &= f(\mu)f(\mathbf{x}|\mu) = f(\mu)\prod_{i=1}^{n} f(x_i|\mu) \\ &\propto f(\mu)\prod_{i=1}^{n} \exp\left(-\frac{(x_i-\mu)^2}{2}\right) \\ &= f(\mu)\exp\left(-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2}\right) \\ &= f(\mu)\exp\left(-\sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{2} - \frac{n(\bar{x}-\mu)^2}{2}\right) \\ &\propto f(\mu)\exp\left(-\frac{n(\bar{x}-\mu)^2}{2}\right) \end{aligned}$$

Portanto, existe uma função proporcional a $f(\mu|x)$, $\tilde{f}(\mu|x)$, tal que

$$\log(\tilde{f}(\mu|\mathbf{x})) = \log(f(\mu)) - \frac{n(\bar{x}-\mu)^2}{2}$$

Digamos que você observou uma amostra de tamanho 100 e $\bar{x} = 3.14$. Assim, pode implementar uma das entradas do Exemplo 9.29 como:

```
ldpost <- function(mu) log(dt(mu,df=1))-(100*(3.14-mu)^2)/2
```

Existe uma variedade enorme de propostas que você pode utilizar. Uma possibilidade é sugerir como proposta a observação anterior somada a ruído branco. Neste caso, obtenha,

```
rprop <- function(ant) ant+rnorm(1)
ldprop <- function(ant,prop) log(dnorm(prop,mean=ant))
```

Dado que você está usando a perda quadrática, o estimador ótimo é $\mathbb{E}[\theta|x]$. Decorre do Teorema 9.25 que você pode aproximar esta quantidade por $\frac{\sum_{i=1}^{n} T_i}{B}$, que é obtida por

```
mean(unlist(metropolis(10^5,0,rprop,ldprop,ldpost)))

## [1] 3.134518
```

9.2.3 Monte Carlo para cadeias de Markov na prática

Contudo, para que ambas essas estratégias sejam possíveis, é necessário cumprir dois requisitos:

1. Construir uma Cadeia de Markov de distribuição estacionária $\int g(\theta)\pi(\theta|x)d\theta$.
2. Estimar $\mathbb{V}\left[\frac{\sum_{i=1}^{B} g(T_i)}{B}\right]$.

Nesta seção, discutiremos bibliotecas na linguagem R que realizam ambas estas tarefas para uma variedade grande casos.

Em primeiro lugar, estudaremos como construir uma Cadeia de Markov que tenha a posteriori como a distribuição estacionária. Para tal, usaremos o pacote "rstan" no R. Mais informações sobre esse pacote podem ser encontradas em Gelman et al. (2014) e Stan Development Team (2015). O primeiro passo consiste em criar um arquivo que especifique o modelo de probabilidade do qual se deseja simular. Por exemplo, considere o modelo em que $X_i|\theta \sim N(\theta, 1)$ e $\theta \sim N(0, 1)$. Este modelo poderia ser descrito num arquivo "normal-normal.stan" da seguinte forma:

```
data {
 int<lower=0> J;        //numero de observacoes, minimo=0.
 real x[J];             //vetor de observacoes reais.
}

parameters {
 real theta;            //parametro eh um numero real.
}

model {
 theta ~ normal(0, 1);  //priori.
 x ~ normal(theta, 1);  //verossimilhanca.
}
```

Para obter a amostra de uma Cadeia de Markov que tem como distribuição estacionária $\theta|X$, podemos usar o seguinte código em R

```
library("rstan")
J <- 100            //100 dados.
x <- rnorm(J,10,1)  //gerar dados a partir de uma N(10,1).
amostra <- stan(file="normal-normal.stan",
                data=c("J", "x"), iter=10^4, chains=1) //B=10^4.
```

Utilizando o código acima, a variável amostra conterá diversas informações da Cadeia de Markov que foi gerado pelo Stan. Por exemplo, rodando

```
 summary(amostra)

 stats
 parameter  mean sd     2.5%  25%   50%   75%   97.5%
 theta      9.86 0.098  9.67  9.80  9.86  9.93  10.06
```

obtemos diversas medidas resumo da posteriori que são estimadas a partir do Método de Monte Carlo. Neste caso, observamos a média e variância da posteriori, bem como vários de seus percentis. Além da função summary, também podemos rodar, por exemplo,

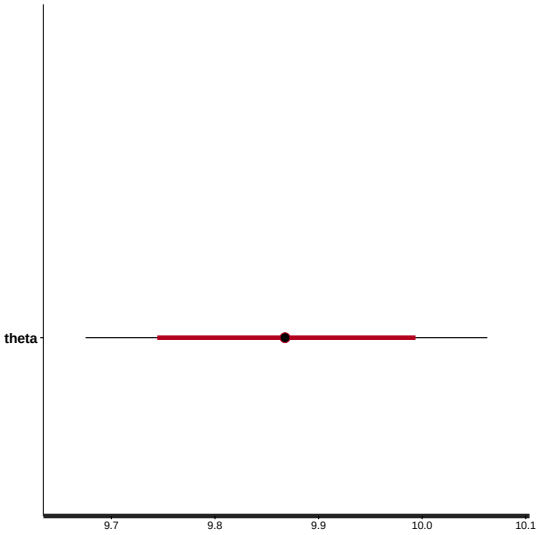

Figura 7: Exemplo de intervalo de credibilidade estimado usando o pacote Stan.

```
 plot(amostra, outer_level=0.95) //credibilidade=0.95.
```

que exibirá estimativas para os intervalos de credibilidade ($\alpha = 5\%$) do modelo descrito (figura 7). O próximo exemplo considera um modelo um pouco mais interessante.

Exemplo 9.31. Considere que, dado θ, $X_i = \alpha + \sum_{j=1}^{k} \beta_j X_{i-j} + \epsilon_i$, onde ϵ_i são i.i.d. e $\epsilon_i \sim N(0,1)$. Este é um modelo AR(k), uma generalização do modelo que vimos no Exercício 8.8. Considere que $X_i = 0$ para $1 \leq i \leq k$ e que, a priori, θ_i são i.i.d. e $\theta_i \sim N(0,1)$. Podemos especificar este modelo no stan da seguinte forma:

```
data {
 int<lower=1> k;  //AR(k)
 int<lower=0> n;  //numero de observacoes, minimo=0.
 real y[n];       //vetor de observacoes reais.
}

parameters {
 real alpha;     //media geral.
 real beta[k];   //parametros autoregressivos.
}

model {
 for(ii in (k+1):n) {
  real mu;
  mu <- alpha;
  for(jj in 1:k) {
   mu <- mu + y[ii-jj]*beta[jj];
  }
  y[ii] ~ normal(mu, 1);
 }
}
```

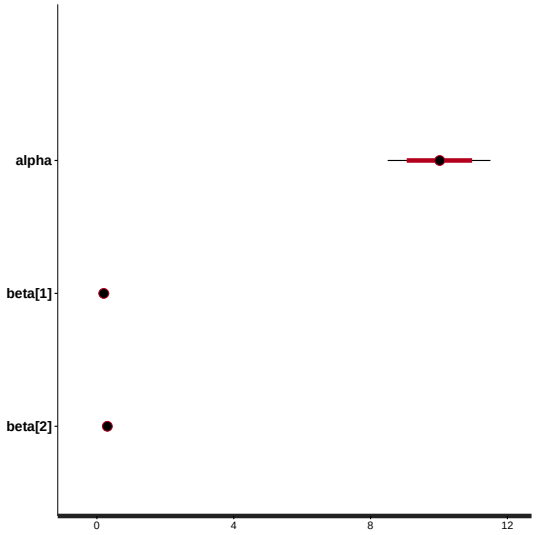

Figura 8: Exemplo de intervalo de credibilidade estimado pelo stan para os parâmetros do Exemplo 9.31.

Este modelo pode ser rodado no R usando os seguintes comandos:

```
k <- 2
n <- 100
y <- rep(0, n)
alpha <- 10
beta <- c(0.2,0.3)
for(ii in 3:n) y[ii] <- rnorm(1,alpha+sum(y[(ii-k):(ii-1)]*beta[k:1]),1)
amostra <- stan(file="ar-k.stan",
                data=c("k", "n", "y"), iter=10^4, chains=1)
plot(amostra)
```

Os intervalos de credibilidade estimados pelo stan podem ser encontrados na figura 8.

A saída do stan também permite recuperar a Cadeia de Markov. Por exemplo, no Exemplo 9.31, a Cadeia de Markov para o parâmetro α pode ser recuperada pelo comando

```
cadeia <- extract(amostra, "alpha")[["alpha"]]
```

Podemos usar esta cadeia para estimar a variância de sua média, assim como no Teorema 9.25. Para tal, utilizaremos o pacote "mcmc" (Geyer and Johnson, 2015). A variância da média da cadeia é estimada por

```
library("mcmc")
olbm(cadeia, 100) //tomamos 100 como 0.01 do tamanho da cadeia.
[1,] 0.0001342983
```

Podemos também estimar a variância do estimador para uma função de α, como no exemplo α^2, por

```
olbm(cadeia^2, 100)
[1,] 0.05281069
```

Exemplo 9.32 (Regressão linear)**.**

Exemplo 9.33 (Regressão linear revisitada)**.**

Exemplo 9.34 (Regressão Poisson)**.**

Exemplo 9.35 (ANOVA)**.**

Vimos que, usando o pacote "rstan" é possível simular Cadeias de Markov que tem como distribuição estacionária a posteriori. Este pacote também estima diversas funções da posteriori, como seus percentis, média e variância. Também, o pacote "mcmc" permite avaliar a precisão das estimativas acima. Unindo estes dois pacotes, é possível aplicar a inferência bayesiana para uma classe ampla de problemas.

9.2.4 Exercícios

Exercício 9.36. Considere o Exercício 7.35.

(a) Gere dados no R considerando que $\mu_1 = 50$, $\mu_2 = 10$, $n = 100$ e $\tau_0^2 = 1$.

(b) Exiba o código para obter uma Cadeia de Markov para μ_1 e μ_2 no stan, considerando que $\nu = 20$, $\tau_1^2 = 25$ e $\tau_2^2 = 10$.

(c) Estime μ_1 e μ_2 usando a utilidade derivada do erro quadrático.

(d) Construa um intervalo de credibilidade de 95% para cada parâmetro. Interprete a razão de os intervalos para μ_1 e μ_2 serem menores do que o intervalo para μ.

(e) Teste a hipótese de M usando a utilidade na tabela 11 e $c = 1$.

(f) Teste $H_0 : \mu \geq 20$ usando a utilidade na tabela 11 e $c = 1$.

Exercício 9.37. Considere o Exercício 7.34.

(a) Gere dados no R considerando que $\theta_1 = 100$ e $\theta_2 = 150$ e $n = 30$.

(b) Exiba o código para obter uma Cadeia de Markov para θ_1 e θ_2 no stan, usando $a_1 = a_2 = b_1 = b_2 = 1$.

(c) Construa um intervalo de credibilidade de 95% para cada parâmetro.

(d) Teste a hipótese de M usando a utilidade na tabela 11 e $c = 1$.

(e) Como você levaria em conta que o número de acessos pode ser influenciado pelo dia da semana? Sugira um modelo que leve este efeito em consideração e rode-o no stan.

10 Revisão final

Exercício 10.1. Considere que $X|\theta \sim \text{Poisson}(\theta)$. Um estatístico deseja usar $\pi(\theta) \propto \theta^{-1}$.

(a) A posteriori para $\theta|X = x$ é proporcional a qual função?

(b) Note que a posteriori para $\theta|X = 0$ não é integrável e, portanto, não é uma densidade de probabilidade. Qual falha permitiu que isso ocorresse?

Referências

Adams, E. (1962). On rational betting systems. *Arch. Math. Log.*, 6(1-2):7–29.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

Billingsley, P. (1986). *Probability and Measure*. Wiley, New York, 2nd edition.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Bonassi, F. V., Stern, R. B., Peixoto, C. M., and Wechsler, S. (2015). Exchangeability and the law of maturity. *Theory and Decision*, 78(4):603–615.

de Bragança Pereira, C. A. and Stern, J. M. (1999). Evidence and credibility: full bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110.

De Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio.

de Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fundam. Math.*, 17:298–329.

De Finetti, B. (1973). Foresight. its logical laws, its subjective sources. reprinted in he kyburg & he smokler. *Studies in Subjective Probability*, pages 93–158.

DeGroot, M. (1986). *Probability and Statistics*. Addison-Wesley.

DeGroot, M. H. (2005). *Optimal statistical decisions*, volume 82. John Wiley & Sons.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Geinitz, S. and Furrer, R. (2013). Conjugate distributions in hierarchical bayesian anova for computational efficiency and assessments of both practical and statistical significance. *arXiv preprint arXiv:1303.3390*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

Geyer, C. J. and Johnson, L. T. (2015). *mcmc: Markov Chain Monte Carlo*. R package version 0.9-4.

Hacking, I. (1972). The logic of pascal's wager. *American Philosophical Quarterly*, 9(2):186–192.

Heath, D. and Sudderth, W. (1976). De finetti's theorem on exchangeable variables. *The American Statistician*, 30(4):188–189.

Izbicki, R. and Esteves, L. G. (2015). Logical consistency in simultaneous statistical test procedures. *Logic Journal of IGPL*, 23(5):732–758.

Jones, G. L. et al. (2004). On the markov chain central limit theorem. *Probability surveys*, 1(299-320):5–1.

Kadane, J. B. (2011). *Principles of uncertainty*. CRC Press.

Kadane, J. B. et al. (2016). Sums of possibly associated bernoulli variables: The conway–maxwell-binomial distribution. *Bayesian Analysis*, 11(2):403–420.

Kingman, J. F. et al. (1978). Uses of exchangeability. *The Annals of Probability*, 6(2):183–197.

Madruga, M. R., Esteves, L. G., and Wechsler, S. (2001). On the bayesianity of pereira-stern tests. *Test*, 10(2):291–299.

O'Neill, B. and Puza, B. (2005). In defence of the reverse gambler's belief. *Mathematical Scientist*, 30(1):13–16.

Rodrigues, F. and Wechsler, S. (1993). A discrete Bayes explanation of a failure-rate paradox. *IEEE Transactions on Reliability*, 42(1):132–133.

Schervish, M. J. (2012). *Theory of statistics*. Springer Science & Business Media.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Stan Development Team (2015). Stan modeling language user's guide and reference manual.

Stern, R. B. and Kadane, J. B. (2015). Coherence of countably many bets. *Journal of Theoretical Probability*, 28(2):520–538.

Wechsler, S., Pereira, C. A. d. B., et al. (2008). Birnbaum's theorem redux. In *BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: Proceedings of the 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 1073, pages 96–100. AIP Publishing.