

# Fundamentos de Inferência Bayesiana

Victor Fossaluza e Luís Gustavo Esteves

2021-05-18



# Contents

<b>1</b>	<b>Prefácio</b>	<b>5</b>
<b>2</b>	<b>Probabilidade Subjetiva</b>	<b>7</b>
2.1	Definição Axiomática . . . . .	7
2.2	Interpretações de Probabilidade . . . . .	7
2.3	Relação de Crença $\preceq$ . . . . .	9
2.4	Medida de Probabilidade que “representa” $\preceq$ . . . . .	13
2.5	Medida de Probabilidade Condicional . . . . .	16
<b>3</b>	<b>Introdução à Inferência Bayesiana</b>	<b>17</b>
3.1	Conceitos Básicos . . . . .	17
3.2	Teorema de De Finetti . . . . .	24
3.3	Suficiência . . . . .	27
3.4	Distribuição a Priori . . . . .	28
3.5	Alguns Princípios de Inferência . . . . .	45
<b>4</b>	<b>Introdução à Teoria da Decisão</b>	<b>51</b>
4.1	Conceitos Básicos . . . . .	51
4.2	Aleatorização e Decisões Mistas . . . . .	55
4.3	Problemas com Dados . . . . .	55
<b>5</b>	<b>Estimação</b>	<b>59</b>
5.1	Estimação Pontual . . . . .	59
5.2	Estimação por Regiões . . . . .	63
5.3	Custo das Observações . . . . .	70

<b>6</b>	<b>Testes de Hipóteses</b>	<b>73</b>
6.1	Conceitos Básicos . . . . .	73
6.2	Revisão: Abordagem Frequentista . . . . .	74
6.3	Abordagem Bayesiana (via Teoria da Decisão) . . . . .	76
6.4	Probabilidade Posterior de $H_0$ . . . . .	78
6.5	Fator de Bayes . . . . .	80
6.6	Teste de Jeffreys . . . . .	83
6.7	Hipóteses Precisas . . . . .	87
6.8	FBST - <i>Full Bayesian Significance Test</i> . . . . .	87
6.9	P-value - Nivel de Significância Adaptativo . . . . .	91
<b>7</b>	<b>Métodos Computacionais</b>	<b>97</b>
7.1	Método de Monte Carlo . . . . .	98
7.2	Monte Carlo com Amostragem de Importância . . . . .	106
7.3	Método de Rejeição . . . . .	106
7.4	ABC (Approximated Bayesian Computation) . . . . .	109
7.5	MCMC - Monte Carlo via Cadeias de Markov . . . . .	110
7.6	Bibliotecas de R para Inferência Bayesiana . . . . .	117
<b>A</b>	<b>Breve Resumo de Medida e Probabilidade</b>	<b>163</b>
A.1	Conceitos Básicos . . . . .	163
A.2	Valor Esperado de $X$ (OU uma ideia da tal Integral de Lebesgue) . . . . .	165
A.3	Funções de Variáveis Aleatórias . . . . .	172
A.4	Função de Distribuição . . . . .	176
A.5	Probabilidade Condicional . . . . .	181

# Chapter 1

## Prefácio

Esse documento foi criado com base nos cursos de *Inferência Bayesiana* ministrados por nós no Instituto de Matemática e Estatística da Universidade de São Paulo (IME-USP). Essas notas devem ser usadas como um roteiro de estudos e não irão necessariamente apresentar todo o conteúdo dessas disciplinas. Além disso, esta é uma versão preliminar que está bem longe da versão final, de modo que podem haver muitos erros e, assim, correções ou sugestões serão sempre muito bem vindas!



## Chapter 2

# Probabilidade Subjetiva

A construção de probabilidade subjetiva apresentada aqui pode ser encontrada no livro *Optimal Statistical Decisions* (DeGroot, 1970).

- $\Omega$ : *espaço amostral*, conjunto não vazio.
- $\mathcal{A}$ :  $\sigma$ -álgebra de subconjuntos de  $\Omega$ , isto é,
  1.  $\Omega \in \mathcal{A}$ ;
  2.  $A \in \mathcal{A} \implies A^c \in \mathcal{A}$ ;
  3.  $A_1, A_2, \dots \in \mathcal{A} \implies \bigcup_{i \geq 1} A_i \in \mathcal{A}$ .
- Os elementos de  $\mathcal{A}$  são chamados de *eventos* e serão denotados por  $A, B, C, \dots, A_1, A_2, \dots$

### 2.1 Definição Axiomática

- $P : \mathcal{A} \longrightarrow [0, 1]$  é uma *medida de probabilidade* se
  1.  $P(\Omega) = 1$ ;
  2.  $A_1, A_2, \dots \in \mathcal{A}$  com  $A_i \cap A_j = \emptyset$ ,  $P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i)$ .

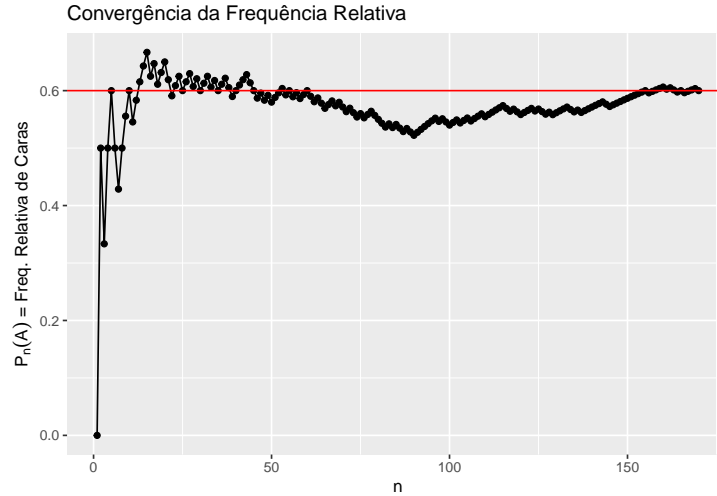
### 2.2 Interpretações de Probabilidade

- **Interpretação Clássica** (De Moivre, Laplace)
  - baseia-se na equiprobabilidade dos resultados;

- $P(A) = \frac{|A|}{|\Omega|}$ .
- **Exemplo:** um lançamento de moeda,  $A = \text{“cara”}$ ,  $P(A) = \frac{1}{2}$ .

- **Interpretação Frequentista** (Venn, von Mises, Reichenbach, etc.)

- quase unânime na primeira metade do século XX e ainda é a mais aceita;
- baseia-se na regularidade das frequências relativas (lei dos grandes números);
- $P(A) = \lim \frac{A_n}{n}$ , onde  $A_n$  é o número de ocorrências de  $A$  em  $n$  realizações *idênticas e independentes* do experimento;
- Supõe que é possível repetir indefinidamente o experimento nas mesmas circunstâncias.
- **Exemplo:** um lançamento de moeda,  $A = \text{“cara”}$ .



- **Interpretação Lógica** (Keynes, Jeffreys, Carnap, etc.)

- medida de “vínculo parcial” entre uma evidência e uma hipótese;
- baseia-se em relações objetivas entre proposições.
- **Exemplo:** considere duas proposições: “até agora todos os lançamentos resultaram em cara” e “será realizado um novo lançamento”. Pode-se afirmar que “provavelmente o resultado do novo lançamento será cara”.



- **Interpretação Subjetivista** (Ramsey, de Finetti, Savage, etc)
  - probabilidade como medida subjetiva de crença;
  - baseada na experiência de cada indivíduo, portanto única.
  - **Exemplo:** suponha que Bruno lançou uma moeda 3 vezes e todos os resultados foram cara. Esse indivíduo, em posse dessa informação, pode acreditar que o resultado cara é mais provável que coroa. Contudo, quando pergunta sobre a probabilidade de cara ao seu colega Olavo, ignorante com relação a moeda, ele responde que é  $1/2$ .

## 2.3 Relação de Crença $\precsim$

$\precsim$  : relação de “crença” em  $\mathcal{A} \times \mathcal{A}$

- $A \prec B$  : acredito mais em  $B$  que em  $A$  ( $B \succ A$ )
- $A \sim B$  : acredito igualmente em  $B$  e  $A$
- $A \precsim B$  : acredito em  $B$  pelo menos tanto quanto em  $A$

**Objetivo:** sob certas condições em  $\precsim$ , obter uma medida de probabilidade  $P$  que representa (concorda) com  $\precsim$ .

$$A \precsim B \iff P(A) \leq P(B)$$

### Suposições sobre $\precsim$

**SP1:** Para  $A, B \in \mathcal{A}$ , exatamente uma das afirmações a seguir deve valer:

$$A \prec B, B \prec A \text{ ou } A \sim B.$$

**SP2:**  $A_1, A_2, B_1, B_2 \in \mathcal{A}$  tais que  $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$  e  $A_i \precsim B_i$ ,  $i = 1, 2$ .  
Então

$$A_1 \cup A_2 \precsim B_1 \cup B_2.$$

Além disso, se  $A_i \prec B_i$  para algum  $i$ , então  $A_1 \cup A_2 \prec B_1 \cup B_2$ .

**SP3:** Se  $A$  é um evento, então  $\emptyset \precsim A$ . Além disso,  $\emptyset \prec \Omega$ .

**SP4:** Se  $A_1, A_2, \dots$  uma sequência decrescente de eventos, isto é,  $A_n \supseteq A_{n+1}, \forall n$ , e  $B$  tal que  $B \precsim A_n, \forall n$  então

$$B \precsim \bigcap_{n \geq 1} A_n.$$

**Lema 1:**  $A, B, D \in \mathcal{A}$  tais que  $A \cap D = B \cap D = \emptyset$ . Então

$$A \precsim B \Leftrightarrow A \cup D \precsim B \cup D$$

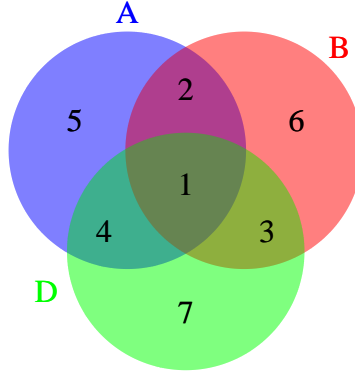
**Demo:**

$$(\Rightarrow) A \precsim B \Rightarrow A \cup D \precsim B \cup D \text{ (SP2)}$$

$$(\Leftarrow) B \prec A \Rightarrow B \cup D \prec A \cup D \text{ (SP2)}$$

**Teorema 1:** Se  $A \precsim B$  e  $B \precsim D$  então  $A \precsim D$ .

**Demo:**



$$(i) (1) \cup (2) \cup (4) \cup (5) \precsim (1) \cup (2) \cup (3) \cup (6) \Rightarrow (4) \cup (5) \precsim (3) \cup (6).$$

$$(ii) \text{ Analogamente, } (2) \cup (6) \precsim (4) \cup (7)$$

$$\text{De (i) e (ii) e pelo Lema 1, } (4) \cup (5) \cup (2) \cup (6) \precsim (3) \cup (6) \cup (4) \cup (7)$$

$$\Rightarrow (2) \cup (5) \precsim (3) \cup (7) \Rightarrow (2) \cup (5) \cup (1) \cup (4) \precsim (3) \cup (7) \cup (1) \cup (4).$$

**Teorema 2 (generalização do SP2):** Se  $A_1, \dots, A_n$  são eventos disjuntos e  $B_1, \dots, B_n$  são também eventos disjuntos tais que  $A_i \lesssim B_i$ , para  $i = 1, \dots, n$ , então

$$\bigcup_{i=1}^n A_i \lesssim \bigcup_{i=1}^n B_i.$$

Se  $A_i \prec B_i$  para algum  $i$ , então  $\bigcup_{i=1}^n A_i \prec \bigcup_{i=1}^n B_i$ .

**Demo:** Basta aplicar SP2  $n - 1$  vezes.

**Teorema 3:** Se  $A \lesssim B$  então  $A^c \gtrsim B^c$ .

**Demo:** Do Lema 1,  $A \cup (A^c \cap B^c) \lesssim B \cup (A^c \cap B^c) \Rightarrow B^c \cup (A \cap B) \lesssim A^c \cup (A \cap B) \Rightarrow B^c \lesssim A^c$ .

**Resultado:** Para todo evento  $A$ ,  $A \lesssim \Omega$ .

**Demo:** Por SP3,  $\emptyset \lesssim A^c$ . Tomando  $D = A$  no Lema 1,  $\emptyset \cup A \lesssim A^c \cup A \Rightarrow A \lesssim \Omega$ .

**Teorema 4:** Se  $A \subseteq B$  então  $A \lesssim B$ .

**Demo:** Suponha,  $B \prec A$ . Tomando  $D = B^c$  no Lema 1,  $B \cup B^c \prec A \cup B^c \Rightarrow \Omega \prec A \cup B^c$ . Absurdo!

**Exemplo 1:**  $\omega_0 \in \Omega$ .  $A \lesssim B \Leftrightarrow \{\omega_0 \in B \text{ ou } \omega_0 \notin (A \cup B)\}$ . Mostre que  $\lesssim$  obedece às SP1 a SP4.

(SP1)

$$A \lesssim B \Leftrightarrow \omega_0 \in B \cup (A \cup B)^c \Rightarrow B \prec A \Leftrightarrow \omega_0 \in B^c \cap (A \cup B) \Leftrightarrow \omega_0 \in A \cap B^c.$$

Analogamente,  $A \prec B \Leftrightarrow \omega_0 \in B \cap A^c$ .

$$A \sim B \Leftrightarrow A \lesssim B \text{ e } B \lesssim A \Leftrightarrow \omega_0 \in [B \cup (A \cup B)^c] \cap [A \cup (A \cup B)^c] \Leftrightarrow \omega_0 \in (A \cap B) \cup (A \cup B)^c.$$

(SP2)

$$A_i \lesssim B_i, i = 1, 2 \Leftrightarrow \omega_0 \in [B_1 \cup (A_1 \cup B_1)^c] \cap [B_2 \cup (A_2 \cup B_2)^c] \Leftrightarrow \omega_0 \in [(B_1 \cup B_2) \cap D^c] \cup (A_1 \cup B_1 \cup A_2 \cup B_2)^c,$$

com  $D = (A_1 \cap B_2) \cup (A_2 \cap B_1)$ .

$A_1 \cup A_2 \precsim B_1 \cup B_2 \Leftrightarrow \omega_0 \in (B_1 \cup B_2) \cup (A_1 \cup A_2 \cup B_1 \cup B_2)^c$

Como  $(B_1 \cup B_2) \cap D^c \subseteq (B_1 \cup B_2)$ , vale o SP2.

(SP3)

$\emptyset \precsim A \Leftrightarrow \omega_0 \in A \cup (\emptyset \cup A)^c \Leftrightarrow \omega_0 \in A \cup A^c = \Omega$ .

Como  $\Omega$  é não-vazio,  $\exists \omega_0 \in \Omega$  e, portanto,  $\emptyset \prec \Omega$ .

(SP4) Exercício!

**Exemplo 2:**  $\Omega = \mathbb{N}$ ,  $\mathcal{A} = \mathcal{P}(\mathbb{N})$ .  $A \precsim B \Leftrightarrow \{B \text{ é infinito ou } A \text{ e } B \text{ são finitos com } |A| \leq |B|\}$ . Verifique se  $\precsim$  satisfaz SP1 a SP4.

**Teorema 5:** Se  $A_1 \subseteq A_2 \subseteq \dots$  é uma sequência crescente de eventos e  $B$  é tal que  $A_n \precsim B, \forall n$  então

$$\bigcup_{n \geq 1} A_n \precsim B.$$

**Demo:**  $A_n^c \supseteq A_{n+1}^c$  e, pelo Teo 3,  $A_n^c \succsim B^c, \forall n$ .

Por SP4,  $\bigcap_{n \geq 1} A_n^c \succsim B^c \Rightarrow \bigcup_{n \geq 1} A_n \precsim B$ .

**Teorema 6:**  $(A_n)_{n \geq 1}$  e  $(B_n)_{n \geq 1}$  sequências tais que  $A_i \cap A_j = B_k \cap B_l = \emptyset, \forall i \neq j, \forall k \neq l$ .

$$A_i \precsim B_i, \forall i \Rightarrow \bigcup_{n \geq 1} A_n \precsim \bigcup_{n \geq 1} B_n.$$

Se existe ao menos um  $j$  tal que  $A_j \prec B_j$  então  $\bigcup_{n \geq 1} A_n \prec \bigcup_{n \geq 1} B_n$ .

**Demo:** Da extensão de SP2, temos que  $\bigcup_{i=1}^n A_i \precsim \bigcup_{i=1}^n B_i, \forall n \geq 1$

$$\Rightarrow \bigcup_{i=1}^n A_i \precsim \bigcup_{i=1}^{\infty} B_i, \forall n \geq 1 \Rightarrow \bigcup_{i=1}^{\infty} A_i \precsim \bigcup_{i=1}^{\infty} B_i \text{ (Teo 5)}$$

$\exists n_0$  tal que  $A_{n_0} \prec B_{n_0}$ . De SP2, temos que, para  $n \geq n_0$ ,

$$\bigcup_{i=1}^{n_0} A_i = \bigcup_{i=1}^{n_0-1} A_i \cup A_{n_0} \prec \bigcup_{i=1}^{n_0-1} B_i \cup B_{n_0} = \bigcup_{i=1}^{n_0} B_i \Rightarrow \bigcup_{i=1}^{n_0} A_i \prec \bigcup_{i=1}^{n_0} B_i.$$

Da primeira parte, temos que  $\bigcup_{i=n_0+1}^{\infty} A_i \precsim \bigcup_{i=n_0+1}^{\infty} B_i$  e, por SP2,

$$\bigcup_{i=1}^{n_0} A_i \cup \bigcup_{i=n_0+1}^{\infty} A_i \prec \bigcup_{i=1}^{n_0} B_i \cup \bigcup_{i=n_0+1}^{\infty} B_i$$

provando o resultado.

## 2.4 Medida de Probabilidade que “representa”

$\preceq$

**SP5:** Existe uma variável aleatória  $X : \Omega \rightarrow \mathbb{R}$ ,  $\mathcal{A}$ -mensurável, tal que  $X(\omega) \in [0, 1]$ ,  $\forall \omega \in \Omega$  e, se  $I_1$  e  $I_2$  são intervalos contidos em  $[0, 1]$ ,  $\{X \in I_1\} \preceq \{X \in I_2\} \Leftrightarrow \lambda(I_1) \leq \lambda(I_2)$ .

- Se  $I = [a, b] \subseteq [0, 1]$ ,  $\lambda(I) = b - a$  é o comprimento do intervalo  $I$  (medida de Lebesgue).
- “Experimento auxiliar” ;  $X \sim \text{Uniforme}[0, 1]$ .
- $\{X \in [a, b]\} \sim \{X \in (a, b]\} \sim \{X \in [a, b)\} \sim \{X \in (a, b)\}$ .

**Teorema 7:** Seja  $A \in \mathcal{A}$ . Então  $\exists! a^* \in [0, 1]$  tal que  $A \sim \{X \in [0, a^*]\}$ .

**Demo:** Seja  $U(A) = \{a \in [0, 1] : A \preceq \{X \in [0, a]\}\}$ .  
 $1 \in U(A)$  pois  $\Omega = \{X \in [0, 1]\} \preceq A \Rightarrow U(A) \neq \emptyset$ .  
Tome  $a^* = \inf U(A)$ .

(i) Considere  $(a_n)_{n \geq 1}$ ,  $a_n \in [0, 1]$ ,  $\forall n \geq 1$ , tal que  $a_n \geq a_{n+1} \geq a^*$  e  $a_n \downarrow a^*$ . Então,  $\forall n \geq 1$ ,  $\{X \in [0, a_n]\} \preceq A$ .

Por SP4,  $\bigcap_{n=1}^{\infty} \{X \in [0, a_n]\} \preceq A \Rightarrow \{X \in [0, a^*]\} \preceq A$

(ii) Se  $a^* = 0$ ,  $\{X \in [0, 0]\} \sim \emptyset \preceq A$  (por SP3).

Se  $a^* > 0$ , considere  $(a_n)_{n \geq 1}$  com  $a_n \leq a_{n+1} < a^*$  e  $a_n \uparrow a^*$ .

$\{X \in [0, a_n]\} \preceq A$ ,  $\forall n \geq 1$  e, pelo Teo 5,  $\bigcup_{n=1}^{\infty} \{X \in [0, a_n]\} \preceq A$   
 $\Rightarrow \{X \in [0, a^*]\} \sim \{X \in [0, a^*]\} \preceq A$ .

De (i) e (ii), temos que  $A \sim \{X \in [0, a^*]\}$ .

$a^*$  é único pois se  $a_1 < a^* < a_2$  são outros valores quaisquer, segue que  $\{X \in [0, a_1]\} \prec \{X \in [0, a^*]\} \prec \{X \in [0, a_2]\}$  e só um desses eventos pode ser equivalente à  $A$ .

**Teorema 8:** A probabilidade do evento  $A$ ,  $P(A)$ , é definida como  $a^* \in [0, 1]$  tal que  $A \sim \{X \in [0, a^*]\}$ . Assim,  $A \sim \{X \in [0, P(A)]\}$ . A função de probabilidade assim definida satisfaz:

$$A \precsim B \Leftrightarrow P(A) \leq P(B).$$

**Demo:** Do Teo 7,  $A \sim \{X \in [0, P(A)]\}$  e  $B \sim \{X \in [0, P(B)]\}$ .  
 $A \precsim B \Leftrightarrow \{X \in [0, P(A)]\} \precsim \{X \in [0, P(B)]\} \Leftrightarrow \lambda([0, P(A)]) \leq \lambda([0, P(B)]) \Leftrightarrow P(A) \leq P(B).$

**Teorema 9:** A função  $P : \mathcal{A} \rightarrow [0, 1]$  que, para cada  $A \in \mathcal{A}$ , associa  $P(A)$  tal que  $A \sim \{X \in [0, P(A)]\}$  é uma medida de probabilidade (no sentido  $\sigma$ -aditiva).

**Demo:** (i)  $P(A) \geq 0$ .  
 $\Omega \sim \{X \in [0, 1]\} \Rightarrow P(\Omega) = 1$ .  
 $\emptyset \sim \{X \in [0, 0]\} \Rightarrow P(\emptyset) = 0$   
 $\emptyset \precsim A \Rightarrow 0 \leq P(A).$

(ii) Seja  $A$  e  $B$  tal que  $A \cap B = \emptyset$ . Vamos mostrar que  $P(A \cup B) = P(A) + P(B)$ .

Pelo Teo 8,  $A \sim \{X \in [0, P(A)]\}$ ,  $B \sim \{X \in [0, P(B)]\}$ ,  $A \cup B \sim \{X \in [0, P(A \cup B)]\}$ .

Como  $A \subseteq A \cup B$  e, por SP3,  $A \precsim A \cup B$ , vale que  $P(A) \leq P(A \cup B)$ .

Vamos verificar que  $B \sim \{X \in (P(A), P(A \cup B)]\}$ .

Suponha, por absurdo,  $B \prec \{X \in (P(A), P(A \cup B)]\}$ .

$A \precsim \{X \in [0, P(A)]\} \xRightarrow{SP2} A \cup B \prec \{X \in [0, P(A)]\} \cup \{X \in (P(A), P(A \cup B)]\} \Rightarrow A \cup B \prec \{X \in [0, P(A)] \cup (P(A), P(A \cup B)]\} \Rightarrow A \cup B \prec \{X \in [0, P(A \cup B)]\}$  (Absurdo!)

Analogamente,  $B \succ \{X \in (P(A), P(A \cup B)]\}$  é absurdo! Logo,

$B \sim \{X \in (P(A), P(A \cup B)]\} \sim \{X \in [0, P(A \cup B) - P(A)]\}$ .

Como  $B \sim \{X \in [0, P(B)]\}$ , temos que  $P(A \cup B) = P(A) + P(B)$ .

**Corolário 1:** Se  $A_1, \dots, A_n$  são eventos disjuntos, então  $P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$ .

**Demo:** Basta repetir o argumento da segunda parte da demonstração anterior  $n - 1$  vezes.

**Teorema 10:** Seja  $A_1 \supseteq A_2 \supseteq \dots$  uma seq. decrescente de eventos tais que  $\bigcap_{i=1}^{\infty} A_i = \emptyset$ . Então  $\lim_{n \uparrow \infty} P(A_n) = 0$ .

**Demo:**  $A_1 \supseteq A_2 \supseteq \dots \Rightarrow P(A_1) \geq P(A_2) \geq \dots$

Além disso,  $\lim_{n \uparrow \infty} P(A_n) = b$ . Como  $P(A_n) \geq b, \forall n$ , segue que  $A_n \succeq$

$\{X \in [0, b]\}, \forall n$ .

Por SP4,  $\emptyset = \bigcap_{i=n}^{\infty} A_i \succeq \{X \in [0, b]\}$ .

Se  $b > 0$ , então  $\{X \in [0, b]\} \succ \{X \in [0, b/2]\} \succeq \emptyset$ . Como essa relação contradiz a anterior, temos que  $b$  deve ser igual a 0.

**Teorema 9: (conclusão)** Usando o Corolário 1 e o Teorema 10 é possível concluir a demonstração do Teorema 9, mostrando que  $P$  é  $\sigma$ -aditiva, isto é,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \quad A_i \cap A_j = \emptyset, \forall i \neq j.$$

**Demo:** Seja  $(A_n)_{n \geq 1}$  sequência de eventos disjuntos. Segue do Corolário 1 que

$$(i) \quad P\left(\bigcup_{i=1}^{\infty} A_n\right) = \sum_{i=1}^n P(A_i) + P\left(\bigcup_{j=n+1}^{\infty} A_j\right), \quad n = 1, 2, \dots$$

Considere  $B_n = \bigcup_{j=n+1}^{\infty} A_j, n \geq 1$ , uma sequência decrescente de eventos tais que  $\bigcap_{n=1}^{\infty} B_n = \emptyset$ . Pelo Teorema 10, segue que  $\lim_{n \uparrow \infty} P(B_n) = 0$ .

Assim, tomando o limite do lado direito de (i), segue que

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \uparrow \infty} \sum_{i=1}^n P(A_i) + \lim_{n \uparrow \infty} P(B_n) = \sum_{i=1}^{\infty} P(A_i).$$

**Teorema 11:** Se a relação de crença  $\precsim$  obedece SP1 a SP5 então  $\exists! P : \mathcal{A} \rightarrow [0, 1]$ , medida de probabilidade, tal que  $P$  representa  $\precsim$ .

**Demo:** Já foi mostrado que  $P$  é uma medida de probabilidade  $\sigma$ -aditiva, de modo que apenas resta mostrar a unicidade de  $P$ .

Considere que existe uma outra medida  $P'$  que concorde com a relação  $\precsim$ . Como  $X \sim \text{Unif}(0, 1)$ ,  $P'(\{X \in [0, a]\}) = a$ . Se  $A$  é um evento, existe um único  $a^*$  tal que  $A \sim \{X \in [0, a^*]\}$  e, como  $P'$  concorda com a relação  $\precsim$ ,

$$P'(A) = P'(\{X \in [0, a^*]\}) = a^* = P(A).$$

## 2.5 Medida de Probabilidade Condicional

**Nova Relação:**  $(A|D) \precsim (B|D)$  (Sabendo que  $D$  ocorreu,  $B$  é preferível a  $A$ ).

- Para  $D = \Omega$ , temos o caso anterior:  $A \precsim B \Leftrightarrow (A|\Omega) \precsim (B|\Omega)$ .
- Suponha que vale as suposições SP1 a SP5 e, adicionalmente,

**SP6:**  $(A|D) \precsim (B|D) \Leftrightarrow (A \cap D) \precsim (B \cap D) \quad \left( (A \cap D|\Omega) \precsim (B \cap D|\Omega) \right)$

**Propriedades decorrentes de SP1 a SP6:**

1.  $\forall A, B, D, (A|D) \precsim (B|D)$  ou  $(B|D) \precsim (A|D)$ .
2. Se  $(A|D) \precsim (B|D)$  e  $(B|D) \precsim (E|D)$  então  $(A|D) \precsim (E|D)$ .
3.  $A, B, D, E$  com  $A \cap D \cap E \sim B \cap D \cap E \sim \emptyset$ .  
 $(A|D) \precsim (B|D) \Leftrightarrow (A \cup E|D) \precsim (B \cup E|D)$ .
4.  $(A|D) \precsim (B|D) \Leftrightarrow (A^c|D) \succsim (B^c|D)$ .
5. Seja  $B, D$  e  $(A_n)_{n \geq 1}$  tal que  $A_n \supseteq A_{n+1}$ .  
 $(B|D) \precsim (A_n|D), \forall n$ , então  $(B|D) \precsim \left( \bigcap_{n=1}^{\infty} A_n | D \right)$ .
6.  $(A_n)_{n \geq 1}$  e  $(B_n)_{n \geq 1}$  tal que  $A_i \cap A_j \sim A_k \cap A_l \sim \emptyset, i \neq j, k \neq l$ , e  
 $(A_n|D) \precsim (B_n|D), \forall n$ . Então  $\left( \bigcup_{n=1}^{\infty} A_n | D \right) \precsim \left( \bigcup_{n=1}^{\infty} B_n | D \right)$

**Teorema 12:**  $\forall A, B, D \in \mathcal{A}$ , considere  $\precsim$  satisfazendo SP1 a SP6. Então  $P : \mathcal{A} \rightarrow [0, 1]$  de modo que para cada  $A \in \mathcal{A}$  é associada  $P(A) \in [0, 1]$  tal que  $A \sim \{X \in [0, P(A)]\}$  é uma medida de probabilidade que representa  $\precsim$ , isto é,

$$(A|\Omega) \precsim (B|\Omega) \Leftrightarrow P(A) \leq P(B).$$

Além disso, se  $D \in \mathcal{A}$  é tal que  $P(D) > 0$ , então

$$(A|D) \precsim (B|D) \Leftrightarrow P(A|D) \leq P(B|D),$$

onde  $P(\cdot|D) : \mathcal{A} \rightarrow [0, 1]$  é uma medida de probabilidade tal que

$$P(A|D) = \frac{P(A \cap D)}{P(D)}.$$



## Chapter 3

# Introdução à Inferência Bayesiana

### 3.1 Conceitos Básicos

- **Inferência Estatística:** fazer afirmações sobre quantidades não observáveis em um determinado contexto.
- $\theta$  : **parâmetro** - quantidade desconhecida de interesse (não-observável em determinado contexto).
- $\Theta$  : **espaço paramétrico** - conjunto onde  $\theta$  toma valores (supostamente conhecido).
- $E = (X, \theta, \{f(x|\theta)\})$ : **experimento** - “*tornar visível algo que antes era invisível*” ou, mais especificamente no nosso contexto, observar uma realização  $x \in \mathfrak{X}$  de um vetor aleatório  $X$  com alguma distribuição  $f(x|\theta)$ . Essa distribuição pertence, na maioria dos casos, à uma família de distribuições fixada mas que depende do parâmetro desconhecido de interesse  $\theta$ . Note que na grande maioria dos problemas do dia a dia de um estatístico ele se utiliza de resultados experimentais para fazer afirmações sobre  $\theta$  e este, por sua vez, é não-observável em geral.
- $\mathfrak{X}$  : **espaço amostral** - conjunto onde  $X$  toma valores (supostamente conhecido).
- $\mathcal{F}$  :  $\sigma$ -álgebra de (sub)conjuntos de  $\mathfrak{X}$ .
- Neste espaço amostral, defini-se uma família  $\mathcal{P} = \{P(\cdot|\theta) : \theta \in \Theta\}$ , isto é, um conjunto de distribuições (condicionais) para  $X$  indexadas por  $\theta$ .
- $(\mathfrak{X}, \mathcal{F}, \mathcal{P})$  : modelo estatístico (clássico).

- $V_x(\theta) = f(x|\theta)$  : função de verossimilhança.

### 3.1.1 Inferência Frequentista (ou Clássica)

- $\theta$  é considerado fixo (apesar de desconhecido) e, portanto, não recebe uma distribuição de probabilidade.
- Baseia-se no " princípio" da amostragem repetida (interpretação frequentista de probabilidade), isto é, supõe que é possível realizar infinitas vezes o experimento. Assim, o  $x$  é apenas um dos possíveis resultados (hipóteses) do experimento.
- Probabilidade somente é definida em (uma  $\sigma$ -álgebra de)  $\mathfrak{X}$ .

### 3.1.2 Inferência Bayesiana

- Baseia-se na interpretação subjetivista de probabilidade, de modo que a *SUA* incerteza sobre algo desconhecido deve ser quantificada (traduzida) em termos de probabilidade.
- Assim, *SUA* incerteza sobre o parâmetro (desconhecido) é representada por uma distribuição de probabilidade,  $\theta$  é tratado como uma variável aleatória (v.a.) e *SUA* distribuição para  $\theta$  antes da realização do experimento,  $f(\theta)$ , é chamada de **distribuição a priori**. Note que a atribuição de uma distribuição a prior para  $\theta$  independe da natureza do parâmetro, ele pode ser a proporção de indivíduos que avalia positivamente o governo atual (quantidade essa que muda a todo instante) ou ainda a milésima casa do  $\pi$  (algum número de 0 a 9, fixo porém desconhecido no momento dessa leitura).
- A atualização de *SUA* incerteza sobre  $\theta$ , incorporando uma nova informação trazida pelos dados  $x$  (representada por  $f(x|\theta)$ ) é feita pelo *Teorema de Bayes*:
- **Teorema de Bayes:**

$$\underbrace{f(\theta|x)}_{\text{dist.posteriori}} = \frac{f(\theta)f(x|\theta)}{\int_{\Theta} f(x|\theta)dP_{\theta}} \propto \underbrace{f(\theta)}_{\text{priori}} \overbrace{f(x|\theta)}^{\text{verossimilhana}}.$$

- Toda a inferência sobre  $\theta$  será baseada exclusivamente em  $f(\theta|x)$ , não sendo necessário considerar pontos amostrais que poderiam mas não foram observados (como é feito na inferência frequentista).

- **Observação:** será utilizada a notação geral para integral (de Lebesgue):

$$\int_{\Theta} f(x|\theta) dP_{\theta} = \begin{cases} \int_{\Theta} f(x|\theta) f(\theta) d\theta & (\text{caso abs. contínuo}) \\ \sum_{\Theta} f(x|\theta) f(\theta) & (\text{caso discreto}) \end{cases}$$

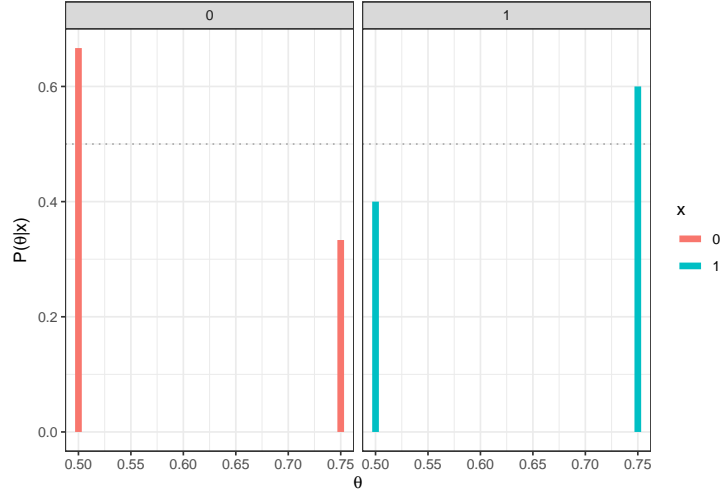
**Exemplo 1a.** Suponha que existem duas moedas, uma delas tem  $\theta = 1/2$  (honestas) e a outra  $\theta = 3/4$  (viesada). Uma moeda é escolhida e é feito um lançamento da moeda selecionada. Nesse experimento, tem-se  $X|\theta \sim \text{Ber}(\theta)$ , com  $\Theta = \{1/2, 3/4\}$  e  $\mathcal{X} = \{0, 1\}$ . Como “chutar” o valor de  $\theta$ ?

Considere que não existe razão para você acreditar que há algum tipo de preferência na escolha de uma ou outra moeda, isto é, considere que a priori  $f(\theta = 1/2) = f(\theta = 3/4) = 1/2$ . Suponha que o lançamento resultou em cara ( $x = 1$ ). Então

$$\begin{aligned} f(\theta = 3/4|X = 1) &= \frac{f(X = 1|\theta = 3/4)f(\theta = 3/4)}{\sum_{\theta} f(X = 1|\theta)f(\theta)} = \frac{\frac{3}{4} \frac{1}{2}}{\frac{3}{4} \frac{1}{2} + \frac{1}{2} \frac{1}{2}} = \frac{3/4}{5/4} = \frac{3}{5} \\ &= 1 - \underbrace{f(\theta = 1/2|X = 1)}_{2/5}. \end{aligned}$$

Se, no entanto, o resultado do lançamento da moeda fosse coroa ( $x = 0$ ), teríamos

$$P(\theta = 3/4|X = 0) = \frac{\frac{1}{4} \frac{1}{2}}{\frac{1}{4} \frac{1}{2} + \frac{1}{2} \frac{1}{2}} = \frac{1/2}{1/2 + 2/2} = \frac{1}{3}.$$



Assim, se sua decisão for escolher o valor mais provável de  $\theta$  após observar  $x$ , a conclusão seria que a moeda é viesada ( $\theta = 3/4$ ) se for observado cara ( $x = 1$ ) e que a moeda é honesta ( $\theta = 1/2$ ) se o resultado for coroa ( $x = 0$ ).

**Exemplo 1b.** Considere agora que serão realizados  $n$  lançamentos da moeda, de modo que agora tem-se  $X|\theta \sim \text{Bin}(n, \theta)$ ,  $\theta \in \{1/2, 3/4\}$ ,  $x \in \{0, 1, \dots, n\}$ . Suponha que observa-se  $X = x$ .

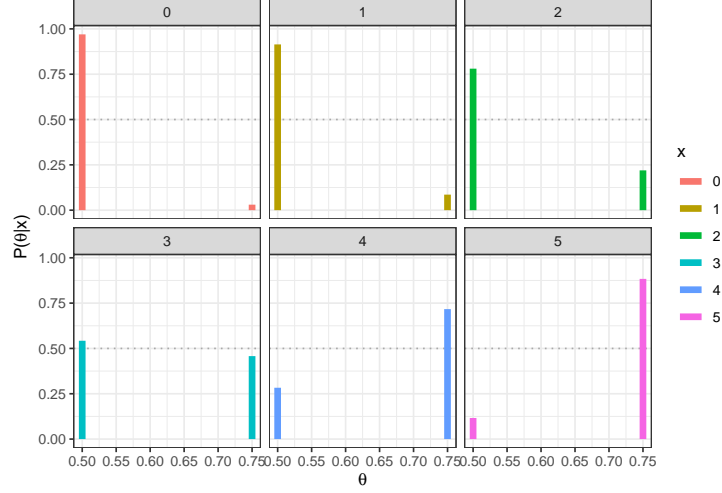
$$\begin{aligned}
 f(\theta = 3/4|X = x) &= \frac{f(x|\theta = 3/4)f(\theta = 3/4)}{\sum_{\theta \in \{1/2, 3/4\}} f(x|\theta)f(\theta)} = \frac{\binom{n}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{n-x} \frac{1}{2}}{\binom{n}{x} \left(\frac{3}{4}\right)^x \left(\frac{1}{4}\right)^{n-x} \frac{1}{2} + \binom{n}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x} \frac{1}{2}} \\
 &= \frac{1}{1 + \left(\frac{2^n}{3^x}\right)} = \frac{3^x}{3^x + 2^n}.
 \end{aligned}$$

```

theta = c(0.5,0.75)
prior=0.5 # priori P(theta[1]) = 1-P(theta[2])
n=5;
post = function(x){
  (prior*dbinom(x,n,theta)) / sum(prior * dbinom(x,n,theta)) }
tibble(x=as.factor(rep(seq(0,n),each=length(theta))),
  x1=rep(theta,(n+1)),x2=rep(theta,(n+1)),y1=0,
  y2=as.vector(apply(matrix(seq(0,n)),1,post))) %>%
ggplot() + geom_hline(yintercept=0.5, col="darkgrey", lty=3) +
geom_segment(aes(x=x1,xend=x2,y=y1,yend=y2,colour=x), lwd=2) +
xlab(expression(theta)) + ylab(expression(paste("P(",theta,"|x)")) +

```

```
theme_bw()+
facet_wrap(~x)
```



Note que o Exemplo 1.a é um caso particular desse exemplo com  $n = 1$ . Se novamente sua decisão é baseada no valor mais provável de  $\theta$ , deve-se escolher  $\theta = 3/4$  se

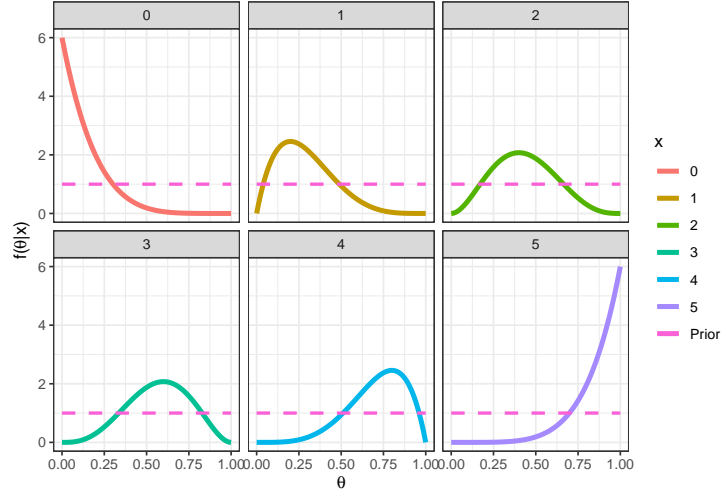
$$f(\theta = 3/4|X = x) > f(\theta = 1/2|X = x) \iff f(\theta = 3/4|X = x) > \frac{1}{2} \iff \frac{3^x}{3^x + 2^n} > \frac{1}{2} \iff 3^x > 2^n \iff \frac{x}{n} = \bar{x} > \log_3 2 \approx 0,63.$$

**Exemplo 1c.** Considere que uma moeda será lançada  $n$  vezes mas que  $\theta$  é desconhecido, de modo que  $\Theta = [0, 1]$ . Para simplificar, vamos assumir  $f(\theta) = \mathbb{I}_{[0,1]}(\theta)$ , isto é,  $\theta \sim Unif(0, 1) \sim Beta(1, 1)$ . Essa priori corresponde ao caso em que você acredita que todos os valores possíveis para  $\theta$  são igualmente “prováveis”, assim como nos exemplos anteriores. Novamente,  $X|\theta \sim Bin(n, \theta)$

$$\begin{aligned} f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{\int_0^1 f(x|\theta)f(\theta)d\theta} = \frac{\binom{n}{x} \theta^x (1-\theta)^{n-x} \mathbb{I}_{[0,1]}(\theta)}{\int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta} = \frac{\frac{\Gamma(1+x)\Gamma(1+n-x)}{\Gamma(1+x)\Gamma(1+n-x)} \theta^x (1-\theta)^{n-x} \mathbb{I}_{[0,1]}(\theta)}{\underbrace{\int_0^1 \frac{\Gamma(1+x)\Gamma(1+n-x)}{\Gamma(1+x)\Gamma(1+n-x)} \theta^x (1-\theta)^{n-x} d\theta}_1} \\ &= \frac{\Gamma(1+x)\Gamma(1+n-x)}{\Gamma(1+x)\Gamma(1+n-x)} \theta^x (1-\theta)^{n-x} \mathbb{I}_{[0,1]}(\theta). \end{aligned}$$

Logo  $\theta|x \sim Beta(1+x, 1+n-x)$ . Nesse exemplo, o valor “mais provável” (com maior densidade a posteriori) para  $\theta$  é a moda da distribuição,  $Moda(\theta|x)$

$= \frac{(1+x)-1}{(1+x)+(1+n-x)-2} = \frac{x}{n} = \bar{x}$ . Suponha que foi observado  $n = 5$  e  $x = 2$ , a posteriori é  $\theta|x = 2 \sim \text{Beta}(3, 4)$  e a moda é  $\text{Moda}(\theta|x) = \frac{1+x-1}{1+1+n-2} = \frac{2}{5} = 0,4$ ;



Algumas medidas resumo da distribuição posterior para esse exemplo são

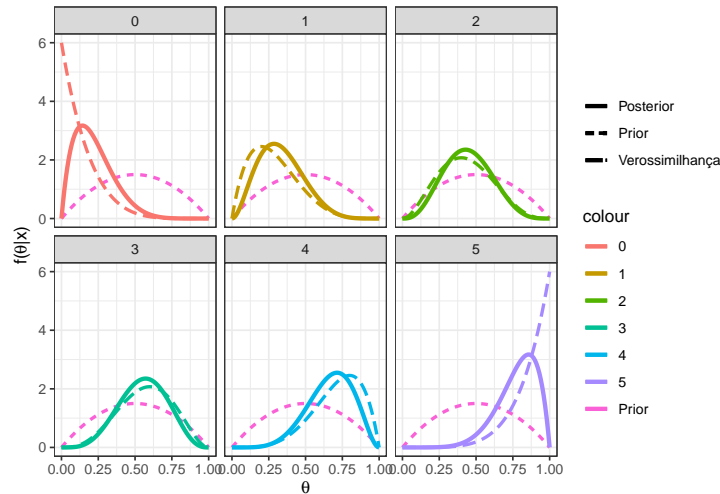
- $\text{Moda}(\theta|x) = \frac{1+x-1}{1+1+n-2} = \frac{2}{5} = 0,4$ ;
- $E[\theta|x] = \frac{1+x}{1+1+n} = \frac{3}{7} = 0,43$ ;
- $\text{Med}(\theta|x) \approx \frac{1+x-1/3}{1+1+n-2/3} = \frac{8/3}{19/3} \approx 0,42$ ;
- $\text{Var}(\theta|x) = \frac{(1+x)(1+n-x)}{(1+1+n)^2(1+1+n+1)} = \frac{12}{392} \approx 0,031$ .

**Exemplo 1d.** Por fim, suponha que no exemplo anterior, sua opinião a priori é representada por uma distribuição beta qualquer com parâmetros  $a$  e  $b$ ,  $a, b > 0$ . Desta forma,  $X|\theta \sim \text{Bin}(n, \theta)$  e  $\theta \sim \text{Beta}(a, b)$ . Calculando a distribuição a posteriori de forma similar ao exemplo anterior, temos que  $\theta|X = x \sim \text{Beta}(a+x, b+n-x)$ . Note que o exemplo anterior é o caso particular em que  $a = b = 1$ .

```

theta = seq(0,1,0.01)
a=2; b=2;
n=5
vero1 = as.vector(apply(matrix(seq(0,n)),1,
                           function(x){dbeta(theta,1+x,1+n-x)}))
post1 = as.vector(apply(matrix(seq(0,n)),1,
                           function(x){dbeta(theta,a+x,b+n-x)}))
tibble(x=as.factor(rep(seq(0,n),each=length(theta))),
        theta=rep(theta,(n+1)),post=post1,vero=vero1) %>%
  ggplot() +
  geom_line(aes(x=theta,y=dbeta(theta,a,b),linetype="Prior",colour="Prior"),lwd=1) +
  geom_line(aes(x=theta,y=post,linetype="Posterior",colour=x),lwd=1.3) +
  geom_line(aes(x=theta,y=vero,linetype="Verossimilhança",colour=x),lwd=1) +
  xlab(expression(theta)) + ylab(expression(paste("f(",theta,"|x)"))) +
  theme_bw()+labs(linetype="")+
  facet_wrap(~x)

```



Suponha agora que  $a = b = 2$ ,  $n = 5$  e  $x = 2$ , de modo que  $\theta|x = 2 \sim \text{Beta}(4, 5)$ . Algumas medidas resumo da distribuição posterior para esse exemplo são

- $\text{Moda}(\theta|x) = \frac{a+x-1}{a+b+n-2} = \frac{3}{7} \approx 0,428;$
- $E[\theta|x] = \frac{a+x}{a+b+n} = \frac{4}{9} \approx 0,444;$
- $\text{Med}(\theta|x) \approx \frac{a+x-1/3}{a+b+n-2/3} = \frac{11/3}{25/3} \approx 0,440;$

$$\bullet \text{ } Var(\theta|x) = \frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)} = \frac{20}{810} \approx 0,0247.$$

## 3.2 Teorema de De Finetti

**Definição.** Uma coleção finita  $X_1, X_2, \dots, X_n$  de quantidades aleatórias é dita *permutável* se a distribuição de  $(X_{\pi_1}, \dots, X_{\pi_n})$  é a mesma para toda permutação  $\pi = (\pi_1, \dots, \pi_n)$  dos índices  $(1, \dots, n)$ . Uma coleção infinita de quantidades aleatórias é *permutável* se toda subcoleção é permutável.

- Segue da definição que cada uma das variáveis  $X_1, \dots, X_n$  tem a mesma distribuição marginal. Além disso,  $(X_i, X_j)$  têm mesma distribuição que  $(X_k, X_l)$ ,  $\forall i \neq j$  e  $k \neq l$ , e assim por diante.

**Proposição.** Uma coleção  $C$  de variáveis aleatórias é permitável se, e somente se, para todo  $n$  finito menor ou igual ao tamanho da coleção  $C$ , toda  $n$ -upla (sequência ordenada de  $n$  elementos) de elementos distintos de  $C$  têm a mesma distribuição que toda outra  $n$ -upla.

**Exemplo 1.** Considere uma coleção  $X_1, X_2, \dots$  uma sequência (finita ou infinita) de variáveis aleatórias independentes e identicamente distribuídas (v.a. i.i.d). Note que  $f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$ ,

$\forall n$ , de modo que  $(X_{i_1}, \dots, X_{i_n})$  têm a mesma distribuição de  $(X_{j_1}, \dots, X_{j_n})$ , para  $i_1 \neq \dots \neq i_n$  e  $j_1 \neq \dots \neq j_n$ . Então, toda coleção de v.a. i.i.d é permutável.

**Exemplo 2:** Foi visto no exemplo anterior que a suposição que uma sequência de v.a. é i.i.d. implica que tal sequência é também permutável. Sabe-se também que independência implica em correlação nula,  $\rho = 0$ . Será então que v.a. identicamente distribuídas e não correlacionadas são também permutáveis?



$X_1 / X_2$	-1	0	+1	$f(x_1)$
-1	0.10	0.05	0.15	0.3
0	0.15	0.20	0.05	0.4
+1	0.05	0.15	0.10	0.3
$f(x_2)$	0.3	0.4	0.3	1.0

$$\text{cor}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}} = \frac{\text{E}[(X_1 - \text{E}[X_1])(X_2 - \text{E}[X_2])]}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}} = \frac{\text{E}[X_1 X_2] - \text{E}[X_1]\text{E}[X_2]}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

$$\text{E}(X_1) = \text{E}(X_2) = 0$$

$$\text{E}(X_1 X_2) = -1 \cdot 0,2 + 0 + 1 \cdot 0,2 = 0 \Rightarrow \text{cor}(X_1, X_2) = 0$$

$(X_1, X_2)$  são identicamente distribuídas e não correlacionadas mas não são permutáveis pois, por exemplo,  $P((X_1, X_2) = (1, -1)) \neq P((X_2, X_1) = (1, -1))$ .

**Exemplo 3:** Suponha que  $X_1, X_2, \dots$  são condicionalmente i.i.d. dado  $Y = y$  com densidade  $f(x_i|y)$ ,  $i = 1, 2, \dots$  e  $Y$  tem densidade  $h(y)$ . Então  $X_1, X_2, \dots$  são permutáveis.

$$f_{X_{i_1}, \dots, X_{i_n}}(x_1, \dots, x_n) = \int \prod_{j=1}^n f(x_j|y) h(y) dy, \text{ para qualquer } n\text{-upla}$$

$X_{i_1}, \dots, X_{i_n}$ . Note que o lado direito não depende dos rótulos  $i_1, \dots, i_n$ .

**Teorema de Representação de De Finetti.** (para v.a. Bernoulli)

Uma sequência infinita  $(X_n)_{n \geq 1}$  de v.a. Bernoulli é permutável se, e somente se, existe uma v.a.  $\theta$  em  $[0, 1]$  tal que, condicional a  $\theta$ ,  $(X_i)_{i \geq 1}$  são i.i.d.  $\text{Ber}(\theta)$ . Além disso, se a sequência é permutável, então a distribuição de  $\theta$  é única e

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \uparrow \infty]{q.c.} \theta.$$

$$P(X_1 = x_1, \dots, X_n = x_n) = \int_0^1 \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} dF(\theta) = \int_0^1 \underbrace{\prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i}}_{f(x_i|\theta)} f(\theta) d\theta,$$

$$\text{onde } F(\theta) = \lim_{n \uparrow \infty} P\left(\frac{\sum_i X_i}{n} \leq \theta\right).$$

**Exemplo 4:** (1.19/1.20 - Schervish)

Seja  $(X_n)_{n \geq 1}$  v.a. Bernoulli.

Considere que o *Estatístico 1* acredita que  $P_1(X_1 = x_1, \dots, X_n = x_n) = \frac{12}{x+2} \frac{1}{\binom{n+4}{x+2}}$ , de modo que  $P_1(X_1 = 1) = \frac{12}{3} \frac{2!}{5!} = \frac{4}{10} = 0,4$ . Por outro lado, o *Estatístico 2* acredita que  $P_2(X_1 = x_1, \dots, X_n = x_n) = \frac{1}{(n+1)\binom{n}{x}}$  e, então,  $P_2(X_1 = 1) = \frac{1}{2} = 0,5$ .

Contudo, pelo Teorema de Finetti, ambos acreditam que o limite  $\theta = \lim_{n \uparrow \infty} \frac{1}{n} \sum_{i=1}^n X_i$  existe com probabilidade 1 e que  $P(X_1 = 1|\theta) = \theta$ , mas não tem opiniões diferentes sobre  $\theta$ .

Suponha agora que foi observado  $x = (x_1, \dots, x_{20})$  com  $\sum_{i=1}^{20} x_i = 14$ .

Então,

$$P_i(X_{21} = 1|X_1 = x_1, \dots, X_{20} = x_{20}) = \frac{P_i(X_1 = x_1, \dots, X_{20} = x_{20}, X_{21} = 1)}{P_i(X_1 = x_1, \dots, X_{20} = x_{20})}$$

de modo que,

$$P_1(X_{21} = 1|\mathbf{X} = \mathbf{x}) = \frac{\frac{12}{17} \frac{1}{\binom{25}{17}}}{\frac{16}{16} \frac{1}{\binom{24}{16}}} = \frac{16}{17} \frac{24!}{16!8!} = \frac{16}{17} \frac{17}{25} = \frac{16}{25} = 0,64$$

$$P_2(X_{21} = 1|\mathbf{X} = \mathbf{x}) = \frac{\frac{1}{22} \frac{20!}{\binom{21}{15}}}{\frac{1}{21} \frac{20!}{\binom{20}{14}}} = \frac{21}{22} \frac{14!6!}{21 \cdot 20!} = \frac{21}{22} \frac{15}{21} = \frac{15}{22} = 0,68$$

**Definição.** Seja  $X_1, \dots, X_n$  uma sequência de variáveis aleatórias permutáveis. A *função de distribuição empírica* é definida como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(x_i \leq x).$$

### Teorema de Representação de De Finetti

Uma sequência de v.a.s  $\{X_n\}_{n \geq 1}$  assumindo valores em (um subconjunto de)  $\mathbb{R}$  é permutável se, e somente se, existe uma medida de probabilidade sobre (uma  $\sigma$ -álgebra do) conjunto de funções de distribuições que “sorteia” uma  $F$  e, dada esta  $F$ , os elementos da sequência  $\{X_n\}_{n \geq 1}$  são i.i.d. com distribuição  $F$ . Isto é,

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \int \prod_{i=1}^n F(x_i) d\mu(F), \forall n.$$

Além disso,  $F_n \xrightarrow{n \uparrow \infty} F$  e a distribuição de  $F = \lim_{n \uparrow \infty} F_n$  é única e é  $\mu$ .

### 3.3 Suficiência

Muitas vezes, a quantidade de dados é muito grande e desejamos “resumir” a informação trazida pelos dados. Uma forma de fazê-lo sem perder informação sobre o parâmetro de interesse é usar uma *estatística suficiente*.

**Definição.** Dizemos que uma função da amostra  $T : \mathfrak{X} \rightarrow \mathbb{R}^p$  é uma *estatística suficiente* (do ponto de vista *frequentista*) se  $f(x|T(x), \theta) = f(x|T(x))$ .

Em palavras, conhecendo o valor da estatística suficiente, a distribuição da amostra (do v.a.  $X$ ) não depende mais do parâmetro  $\theta$ . Isso quer dizer que a informação disponível na amostra  $X$  sobre  $\theta$  está contida em  $T(X)$ . Obter uma estatística suficiente nem sempre é uma tarefa fácil mas o resultado a seguir, conhecido como *critério da fatoração* permite identificar estatísticas suficientes.

**Teorema.** A estatística  $T : \mathfrak{X} \rightarrow \mathbb{R}^p$  é suficiente para a família de distribuições  $\{f(\cdot|\theta) : \theta \in \Theta\}$  se, e somente se, para todo  $x \in \mathfrak{X}$  e para todo  $\theta \in \Theta$ , podemos escrever  $f(x|\theta) = u(x)v(T(x), \theta)$ , onde  $u$  é uma função positiva que não depende de  $\theta$  e  $v$  é uma função não-negativa e depende de  $x$  somente através de  $T(x)$ .

**Exemplo.** Seja  $X_1, \dots, X_n$  v.a. tais que, condicional ao conhecimento de  $\theta$ , são c.i.i.d. com  $X_1|\theta \sim \text{Exp}(\theta)$ . Então,

$$f(x|\theta) = \prod f(x_i|\theta) = \prod \theta e^{-\theta x_i} \mathbb{I}_{\mathbb{R}_+}(x_i) = \theta^n e^{-\theta \sum x_i} \prod \mathbb{I}_{\mathbb{R}_+}(x_i) \\ = v(\sum x_i, \theta) u(x).$$

Portanto,  $T(x) = \sum x_i$  é estatística suficiente para  $\theta$ . De fato, como  $T(X) = \sum X_i|\theta \sim \text{Gama}(n, \theta)$  e  $\{X_1 = x_1, \dots, X_n = x_n\} \subseteq \{T(X) = \sum X_i = \sum x_i = t\}$ ,

$$f(x|T(x), \theta) = \frac{f(x, T(x)|\theta)}{f(T(x)|\theta)} = \frac{f(x|\theta)}{f(t|\theta)} = \frac{\theta^n e^{-\theta \sum x_i} \prod \mathbb{I}_{\mathbb{R}_+}(x_i)}{\frac{\theta^n}{\Gamma(n)} t^{n-1} e^{-\theta t} \prod \mathbb{I}_{\mathbb{R}_+}(x_i)}$$

$$= \frac{\Gamma(n)}{t^{n-1}} \mathbb{I}_{\mathbb{R}_+}(t),$$

que não depende de  $\theta$ .

Sob o enfoque bayesiano, a definição de suficiência é um pouco mais intuitiva que a frequentista.

**Definição:** Dizemos que uma função da amostra  $T : \mathfrak{X} \rightarrow \mathbb{R}^p$  é uma *estatística suficiente* (no sentido *bayesiano*) se  $f(\theta|T(x)) = f(\theta|x)$ , para todo  $x \in \mathfrak{X}$ .

**Voltando ao exemplo**, suponha agora que, a priori,  $\theta \sim \text{Gama}(a, b)$ . Então,  
 $f(\theta|x) \propto f(x|\theta)f(\theta) \propto \theta^n e^{-\theta \sum x_i} \theta^{a-1} e^{-b\theta} \propto \theta^{a+n-1} e^{-(b+\sum x_i)\theta}$   
 Seja  $T = T(X) = \sum X_i$ , temos que  $T|\theta \sim \text{Gamma}(n, \theta)$ , de modo que  
 $f(\theta|T(x) = t) \propto f(t|\theta)f(\theta) \propto \theta^n t^{n-1} e^{-\theta t} \theta^{a-1} e^{-b\theta} \propto \theta^{a+n-1} e^{-(b+t)\theta}$   
 , com  $t = \sum x_i$ .  
 Assim,  $\theta|x \sim \theta|T(x) \sim \text{Gamma}(a+n, b+\sum x_i)$  e, portanto,  
 $T(X) = \sum X_i$  é estatística suficiente para  $\theta$ .

Pelo teorema da fatoração, temos que  $f(x|\theta) = u(x)v(T(x), \theta)$  e, portanto  $f(\theta|x) \propto f(\theta)f(x|\theta) \propto f(\theta)v(T(x), \theta)$ , que só depende de  $x$  por meio de  $T(x)$ . Para os casos mais comuns, as definições são equivalentes (Schervish, 2012).

Um dos princípios de inferência estatística é o *princípio da suficiência*. Segundo este, se  $T$  é uma estatística suficiente para  $\theta$  e se dois pontos amostrais  $x, y \in \mathfrak{X}$  são tais que  $T(x) = T(y)$  então as inferências baseadas nesses pontos devem ser as mesmas. Adiante, retomaremos esse princípio de forma mais formal.

### 3.4 Distribuição a Priori

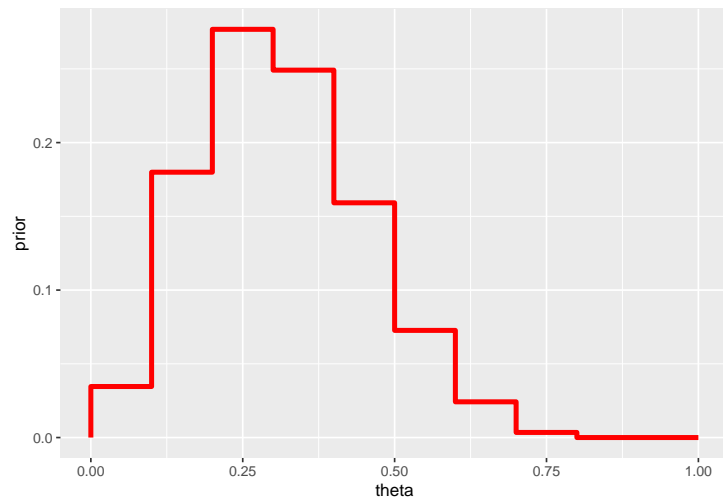
- A priori é sempre subjetiva (assim como a escolha do modelo estatístico)!
  - Por exemplo, dizer que os dados seguem uma distribuição normal, é uma escolha subjetiva, muitas vezes baseadas nas facilidades matemáticas que essa distribuição proporciona.
  - Do mesmo modo, suponha que dois indivíduos que consideram que a distribuição do parâmetro é simétrica, com mesmas suposições sobre média e variância. O primeiro pode optar por representar sua distribuição usando uma distribuição Normal, enquanto o segundo pode utilizar uma distribuição T ou Cauchy.

- Não existe “opinião errada”, existem opiniões diferentes, dado o nível de conhecimento e as experiências prévias do indivíduo. Contudo, algumas “boas práticas” devem ser consideradas como, por exemplo, tomar cuidado para não atribuir probabilidade nula a pontos “possíveis” do espaço paramétrico.
- A priori deve ser sua opinião apenas sobre o parâmetro  $\theta$  e não deve depender de fatores como o desenho do experimento ou o objetivo do estudo.

### 3.4.1 Método do Histograma

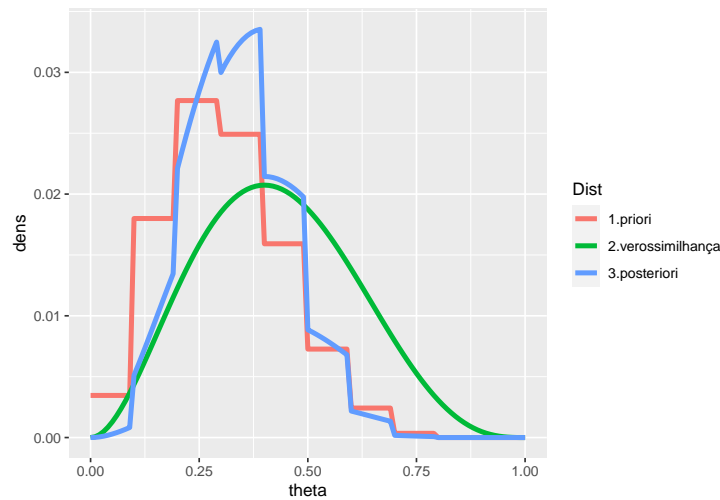
- Muitas vezes, para “extrair” o conhecimento de um especialista, podemos dividir o espaço paramétrico em regiões e pedir para o especialista “ordenar” esses conjuntos, utilizando “pesos” que refletem a crença que o parâmetro esteja em cada uma daquelas regiões.
- **Exemplo 1.** (Albert (2009), pág 27)
  - Seja  $\theta$  uma proporção desconhecida ( $\Theta = [0, 1]$ );
  - Considere a partição  $T = \{[0, 0.1), [0.1, 0.2), \dots, [0.9, 1]\}$ ;
  - Suponha que um especialistas atribui pesos  $p = (1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)$  a esse intervalos;
  - A priori, nesse caso, é o histograma apresentado a seguir.

```
p=c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)
prior = c(0,p/(sum(p)))
tibble(theta=seq(0,1,0.1), prior) %>%
  ggplot(data=.) +
  geom_step(aes(x=theta,y=prior),direction="vh",color="red",lwd=1.5)
```



- Voltando ao exemplo da moeda, suponha novamente que foram observados  $x = 2$  sucessos em  $n = 5$  lançamentos. A posteriori nesse caso pode ser obtida multiplicando a distribuição a priori pela verossimilhança e “padronizando” a função obtida. Assim:

```
n=5
x=2
p = c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)
p = p/(sum(p))
theta = seq(0,1,0.01)
prior = c(rep(p,each=10),0)/sum(c(rep(p,each=10),0))
vero = dbinom(x,n,theta)/sum(dbinom(x,n,theta))
post = (prior * vero)/sum(prior * vero)
pH = tibble(theta=rep(theta,3),dens=c(prior,vero,post),Dist=rep(c('1.priori','2.verossimilhança','3.posteriori'),3))
ggplot(data=pH) +
  geom_line(aes(x=theta,y=dens,colour=Dist),lwd=1.5)
```



### 3.4.2 Elicitação de Hiperparâmetros

- Nessa abordagem, a priori é obtida da seguinte maneira:
  - Escolha uma família de distribuições conveniente. O conceito de “conveniência” aqui pode levar em conta, por exemplo, o suporte da distribuição, se é flexível o suficiente para acomodar diversos tipos de opinião, se permite a obtenção analítica da posteriori e assim por diante;
  - Obtenha um conjunto de medidas resumo (como média, variância, quantis, etc.);
  - Utilize as medidas resumo para calcular hiperparâmetros da distribuição escolhida.
- Exemplo:** Na seção anterior, a priori dada pelo histograma tem média  $m = 0.31$  e variância aproximadamente  $v = 0.02$ . Podemos utilizar como priori, por exemplo, uma distribuição beta com essa média e variância, já que a beta tem um suporte conveniente e facilita as contas, como também já vimos. Assim, vamos considerar uma distribuição  $Beta(a, b)$  e escolher  $a$  e  $b$  satisfazendo:

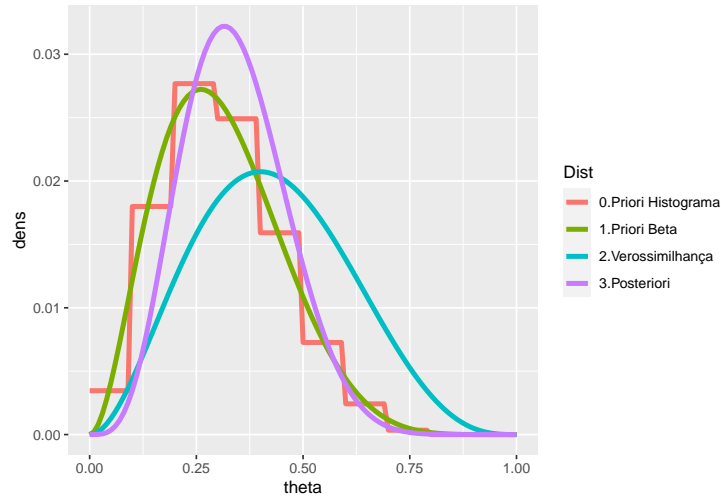
$$(i) \quad E[\theta] = \frac{a}{a+b} = m \Leftrightarrow b = \left( \frac{1-m}{m} \right) a$$

$$(ii) \text{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)} = 0.02 \Leftrightarrow a = \frac{m(m-m^2-v)}{v}$$

Resolvendo o sistema temos, de forma geral, que  $a = \frac{m(m-m^2-v)}{v}$  e  $b = \frac{(1-m)(m-m^2-v)}{v}$ .

Assim, no nosso exemplo, teríamos uma  $Beta(3, 6.7)$ . Além disso, já vimos que, nesse caso, a distribuição a posteriori é  $Beta(3+x, 6.7+n-x)$ . Considerando novamente  $n = 5$  e  $x = 2$ , temos:

```
n=5; x=2
m=0.31; v=0.02
a=m*(m-m^2-v)/v; b=(1-m)*(m-m^2-v)/v
p = c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)
p = p/(sum(p))
theta = seq(0,1,0.01)
prior = dbeta(theta,a,b)/sum(dbeta(theta,a,b))
vero = dbinom(x,n,theta)/sum(dbinom(x,n,theta))
post = dbeta(theta,a+x,b+n-x)/sum(dbeta(theta,a+x,b+n-x))
priorH = c(rep(p,each=10),0)/sum(c(rep(p,each=10),0))
tibble(theta=rep(theta,4),dens=c(prior,vero,post,priorH),
        Dist=rep(c('1.Priori Beta','2.Verossimilhança','3.Posteriori','0.Priori Histograma'),4),
        ggplot(data=.) +
        geom_line(aes(x=theta,y=dens,colour=Dist),lwd=1.5))
```





### 3.4.3 Prioris Conjugadas

Como visto no exemplo da moeda, em que a distribuição a priori era  $Beta(a, b)$ , a posteriori era facilmente obtida e também estava na classe das distribuições  $Beta$ . Em particular, quando observa-se  $x$  sucessos em  $n$  realizações de ensaios de Bernoulli, a distribuição a posteriori é  $Beta(a + x, b + n - x)$ . Isso ocorre pois essa distribuição pertence à uma classe bastante específica de distribuições a priori, chamadas distribuições conjugadas.

**Definição** Seja  $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$  uma família de distribuições (condicionais) para  $X$  e considere  $\mathcal{C} = \{h(\theta|a) : a \in A\}$  uma família de distribuições para  $\theta$ . Dizemos que (a família)  $\mathcal{C}$  é **conjugada** para  $\mathcal{P}$  se,  $\forall h(\theta) \in \mathcal{C}$ ,  $h(\theta|x) \propto f(x|\theta)h(\theta) \in \mathcal{C}, \forall x \in \mathcal{X}$ .

**Resultado 1.** Seja  $X$  v.a. tal que, condicional ao conhecimento de  $\theta$ ,  $X|\theta \sim Bin(n, \theta)$ . Considere que, a priori,  $\theta \sim Beta(a, b)$ . Então,  $\theta|X = x \sim Beta(a + x, b + n - x)$ . Portanto, a família  $\mathcal{C} = \{Beta(a_1, a_2) : (a_1, a_2) \in \mathbb{R}_+^2\}$  é conjugada para  $\mathcal{P} = \{Bin(n, \theta) : \theta \in [0, 1]\}$ .

- Esse resultado também vale se
  1.  $X_1, \dots, X_n$  são v.a.s *condicionalmente independentes e identicamente distribuídas* (c.i.i.d.) com  $X_i|\theta \sim Ber(\theta)$
  2.  $X_i|\theta \sim Geo(\theta), i = 1, \dots, n$  c.i.i.d.
  3.  $X_i|\theta \sim BinNeg(k, \theta)$   
 $\theta \sim Beta(a, b) \Rightarrow \theta|X = x \sim Beta(a + s, b + f)$  em que  $s$  é o número de sucessos e  $f$  é o número de fracassos.

**Resultado 2.** (*generalização do resultado anterior para o caso em que o número de categorias é maior que 2*)

Seja  $X|\theta \sim Multinomial(n, \theta)$ , isto é, sua função de probabilidade é dada por

$$f(x|\theta) = \binom{n}{x_1, x_2, \dots, x_k} \prod_{i=1}^{k-1} \theta^i \underbrace{\left(1 - \sum_{i=1}^{k-1} \theta_i\right)^{n - \sum_{i=1}^{k-1} x_i}}_{\theta_k^{x_k}}$$

em que  $\theta_i \in [0, 1]$  com  $\sum_{i=1}^K \theta_i = 1$ ,  $x_i \in \{0, 1, \dots, n\}$  com  $\sum_{i=1}^n x_i = n$  e

$$\binom{n}{x_1, x_2, \dots, x_k} = \frac{n!}{x_1! x_2! \dots x_k!}.$$

Considere que, a priori,  $\theta \sim \text{Dirichlet}(a_1, \dots, a_k)$ ,  $a_i > 0, i = 1, \dots, k$ , isto é, a f.d.p. a priori para  $\theta$  é dada por

$$f(\theta) = \frac{\Gamma(\sum_{i=1}^K a_i)}{\Gamma(a_1)\Gamma(a_2) \dots \Gamma(a_k)} \prod_{i=1}^{k-1} \theta_i^{a_i-1} \underbrace{\left(1 - \sum_{i=1}^{k-1} \theta_i\right)}_{\theta_k}^{a_k-1}.$$

Então, a distribuição a posteriori para  $\theta$  é  $\theta|X = x \sim \text{Dirichlet}(a_1 + x_1, \dots, a_k + x_k)$ .

**Demo:** Para verificar o resultado, basta ver que

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_{\Theta} f(x|\theta)f(\theta)d\theta} \propto f(x|\theta)f(\theta) \propto \prod_{i=1}^{k-1} \theta_i^{(a_i+x_i)-1} \left(1 - \sum_{i=1}^{k-1} \theta_i\right)^{(a_k+x_k)-1}$$

**Resultado 3.** Seja  $X_1, \dots, X_n$  v.a. c.i.i.d tais que  $X_i|\theta \sim \text{Unif}(0, \theta)$  e considere que, a priori,  $\theta \sim \text{Pareto}(a, b)$ . Então  $\theta|X = x \sim \text{Pareto}(a + n, \max\{b, x_{(n)}\})$ .

**Demo:**

$$\begin{aligned} f(x|\theta) &\stackrel{ci}{=} \prod_{i=1}^n f(x_i|\theta) \stackrel{id}{=} \prod_{i=1}^n \frac{1}{\theta} \mathbb{I}_{[0, \theta]}(x_i) = \frac{1}{\theta^n} \mathbb{I}_{[0, \theta]}(x_{(n)}) \\ &= \frac{1}{\theta^n} \mathbb{I}_{[x_{(n)}, +\infty)}(\theta) \\ \text{em que } x_{(n)} &= \max\{x_1, \dots, x_n\}. \end{aligned}$$

$$f(\theta) = \frac{ab^a}{\theta^{a+1}} \mathbb{I}_{[b, +\infty)}(\theta).$$

Então

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)f(\theta) = \frac{1}{\theta^{a+n+1}} \mathbb{I}_{[x_{(n)}, +\infty)}(\theta) \mathbb{I}_{[b, +\infty)}(\theta) = \frac{1}{\theta^{a+n+1}} \mathbb{I}_{[\max\{b, x_{(n)}\}, +\infty)}(\theta) \\ &\Rightarrow \theta|X = x \sim \text{Pareto}(a + n, \max\{b, x_{(n)}\}). \end{aligned}$$

**Resultado 4.** Seja  $X_1, \dots, X_n, Y_1, \dots, Y_m$  v.a. condicionalmente independentes tais que  $X_i|\theta \sim \text{Exp}(\theta), i = 1, \dots, n$  e  $Y_j|\theta \sim \text{Poisson}(\theta), j = 1, \dots, m$ . Considere que, a priori,  $\theta \sim \text{Gama}(a, b)$ . Então  $\theta|x, y \sim \text{Gama}(a + n + \sum_j y_j, b + m + \sum_i x_i)$ .

**Demo:**

$$\begin{aligned}
 f(x, y|\theta) &\stackrel{ci}{=} f(x|\theta)f(y|\theta) \stackrel{ci}{=} \prod_{i=1}^n f(x_i|\theta) \prod_{j=1}^m f(y_j|\theta) = \prod_{i=1}^n \theta e^{-\theta x_i} \prod_{j=1}^m \frac{\theta^{y_j} e^{-\theta}}{y_j!} = \\
 &\frac{1}{\prod_{j=1}^m y_j!} \theta^{n+\sum_j y_j} e^{-(m+\sum_i x_i)\theta} \\
 f(\theta) &= \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \\
 f(\theta|x, y) &\propto f(x, y|\theta)f(\theta) \propto \theta^{[a+n+\sum_j y_j]-1} e^{-[b+m+\sum_i x_i]\theta} \\
 &\Rightarrow \theta|x, y \sim \text{Gama}(a+n+\sum_j y_j, b+m+\sum_i x_i)
 \end{aligned}$$

**Resultado 5.** Seja  $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$  e  $\mathcal{C} = \{h(\theta|a) : a \in A\}$  uma família conjugada para  $\mathcal{P}$ . Considere  $\mathcal{M} = \{h(\theta) = \sum_{i=1}^m w_i h_i(\theta) : h_i \in \mathcal{C} \text{ e } w_i > 0, \sum_{i=1}^m w_i = 1\}$ . Então  $\mathcal{M}$  é família conjugada para  $\mathcal{P}$ .

**Demo:** Como  $\mathcal{C}$  é conjugada para  $\mathcal{P}$ , para toda função  $h_i \in \mathcal{C}$ , temos que  $f_i(\theta|x) \propto h_i(\theta)f(x|\theta) \in \mathcal{C}$ . Então

$$\begin{aligned}
 h \in \mathcal{M} &\Rightarrow f(\theta|x) \propto h(\theta)f(x|\theta) \propto \sum_{i=1}^m w_i \underbrace{h_i(\theta)f(x|\theta)}_{\in \mathcal{C}} \\
 &\propto \sum_{i=1}^m w_i^* f_i(\theta|x) \in \mathcal{M}.
 \end{aligned}$$

**Exemplo.** Seja  $X|\theta \sim \text{Bin}(n, \theta)$  e  $f(\theta) = wf_1(\theta) + (1-w)f_2(\theta)$ , com  $f_1 \sim \text{Beta}(a_1, b_1)$  e  $f_2 \sim \text{Beta}(a_2, b_2)$ .

$$\begin{aligned}
 f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{\int_0^1 f(x|\theta)f(\theta)} = \frac{f(x|\theta)[wf_1(\theta) + (1-w)f_2(\theta)]}{w \int_0^1 f_1(\theta)f(x|\theta)d\theta + (1-w) \int_0^1 f_2(\theta)f(x|\theta)d\theta} \\
 &\propto \frac{w \binom{n}{x} \frac{\Gamma(a_1+b_1)}{\Gamma(a_1)\Gamma(b_1)} \theta^{a_1+x-1} (1-\theta)^{b_1+n-x-1} + (1-w) \binom{n}{x} \frac{\Gamma(a_2+b_2)}{\Gamma(a_2)\Gamma(b_2)} \theta^{a_2+x-1} (1-\theta)^{b_2+n-x-1}}{\underbrace{w \binom{n}{x} \frac{\Gamma(a_1+b_1)}{\Gamma(a_1)\Gamma(b_1)} \frac{\Gamma(a_1+x)\Gamma(b_1+n-x)}{\Gamma(a_1+b_1+n)}}_A + \underbrace{(1-w) \binom{n}{x} \frac{\Gamma(a_2+b_2)}{\Gamma(a_2)\Gamma(b_2)} \frac{\Gamma(a_2+x)\Gamma(b_2+n-x)}{\Gamma(a_2+b_2+n)}}_B} \\
 &\propto \underbrace{\frac{A}{A+B}}_{w^*} \text{Beta}(a_1+x, b_1+n-x) + \underbrace{\frac{B}{A+B}}_{1-w^*} \text{Beta}(a_2+x, b_2+n-x).
 \end{aligned}$$

Primeiramente, suponha que  $n = 5$ , e temos uma mistura das distribuições  $\text{Beta}(5, 12)$  e  $\text{Beta}(10, 3)$ , com  $w = 0.5$ . O gráfico a seguir apresenta as distribuições a priori, a verossimilhança e a posteriori para cada possível valor de  $x$  em  $\{0, 1, \dots, 5\}$ .

```

a1=5; b1=12
a2=10; b2=3
n=5
w=0.5

theta = seq(0,1,0.01)

A = as.vector(apply(matrix(seq(0,n)),1,
  function(x){w*choose(n,x)*gamma(a1+b1)/(gamma(a1)*gamma(b1))*
    (gamma(a1+x)*gamma(b1+n-x))/gamma(a1+b1+n)}}))

B = as.vector(apply(matrix(seq(0,n)),1,
  function(x){(1-w)*choose(n,x)*gamma(a2+b2)/(gamma(a2)*gamma(b2))*
    (gamma(a2+x)*gamma(b2+n-x))/gamma(a2+b2+n)}}))

w2 = A/(A+B)

prior2 = as.vector(apply(matrix(seq(0,n)),1,
  function(x){w*dbeta(theta,a1,b1)+
    (1-w)*dbeta(theta,a2,b2)}}))

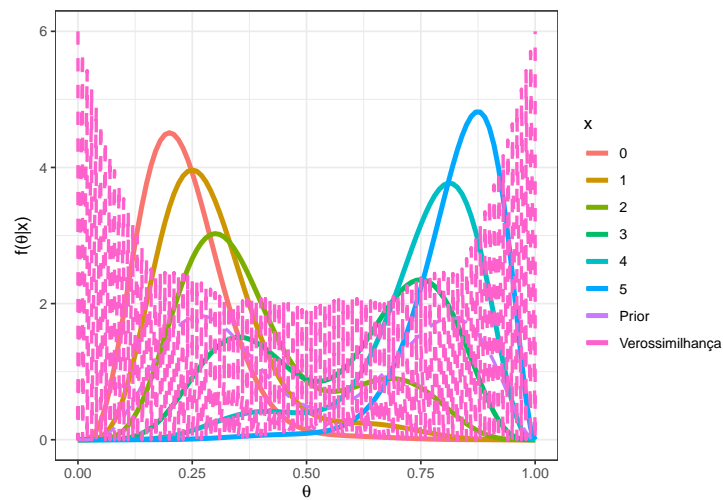
post2 = as.vector(as.matrix(mapply(function(x,w2){
  w2*dbeta(theta,a1+x,b1+n-x)+
  (1-w2)*dbeta(theta,a2+x,b2+n-x)},seq(0,n),w2)))

#vero = as.vector(apply(matrix(seq(0,n)),1,
# function(x){dbinom(x,prob=theta,size=n)}}))

# Verossimilhança proporcional visualmente melhor
vero = as.vector(apply(matrix(seq(0,n)),1,
  function(x){dbeta(theta,x+1,n-x+1)}}))

tibble(x=as.factor(rep(seq(0,n),each=length(theta))),
  w2=rep(w2,each=length(theta)),
  theta=rep(theta,(n+1)),vero=vero,prior=prior2,post=post2) %>%
  ggplot() +
  geom_line(aes(x=theta,y=post, colour=x),lwd=1.5) +
  geom_line(aes(x=theta,y=prior,colour="Prior"),lwd=1,lty=2) +
  geom_line(aes(x=theta,y=vero,colour="Verossimilhança"),lwd=1,lty=2)+
  xlab(expression(theta)) +
  ylab(expression(paste("f(",theta,"|x)")))+
  theme_bw()

```



Agora, suponha que  $n = 5$  e foi observado  $x = 2$ . Novamente, considere a mistura das distribuições  $Beta(5, 12)$  e  $Beta(10, 3)$  mas agora com pesos  $w$  variando no conjunto  $\{0, 0.1, \dots, 0.9, 1\}$ .

```
n=5; x=2
w = seq(0,1,0.1)

A = as.vector(apply(matrix(w),1,
  function(w){w*choose(n,x)*gamma(a1+b1)/(gamma(a1)*
    gamma(b1))*(gamma(a1+x)*gamma(b1+n-x))/gamma(a1+b1+n)}))

B = as.vector(apply(matrix(w),1,
  function(w){(1-w)*choose(n,x)*gamma(a2+b2)/(gamma(a2)*
    gamma(b2))*(gamma(a2+x)*gamma(b2+n-x))/gamma(a2+b2+n)}))

w2 = A/(A+B)

prior2 = as.vector(apply(matrix(w),1,function(w){
  w*dbeta(theta,a1,b1)+(1-w)*dbeta(theta,a2,b2)}))

post2 = as.vector(as.matrix(mapply(function(w,w2){
  w2*dbeta(theta,a1+x,b1+n-x)+
  (1-w2)*dbeta(theta,a2+x,b2+n-x)},w,w2)))

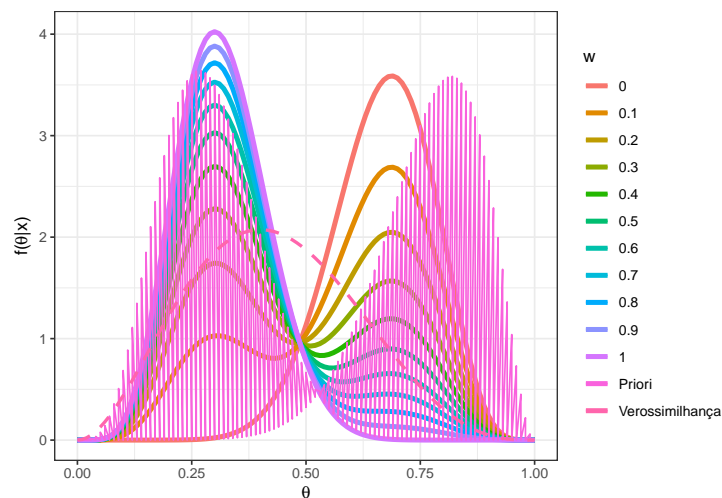
vero = as.vector(apply(matrix(rep(x,2*n+1)),1,
  function(x){dbeta(theta,x+1,n-x+1)}))
```

```

z<-length(w)

tibble(w=as.factor(rep(w,each=length(theta))),
      w2=rep(w2,each=length(theta)),
      theta=rep(theta,z), prior = prior2,
      post = post2, vero = vero) %>%
  ggplot(colour = w) +
  geom_line(aes(x=theta,y=post, colour=w),lwd=1.5) +
  geom_line(aes(x=theta,y=prior,colour="Priori")) +
  geom_line(aes(x=theta,y=vero,colour="Verossimilhança"),lwd=1,lty=2)+
  xlab(expression(theta)) + ylab(expression(paste("f(",theta,"|x)")))+
  theme_bw()

```



### 3.4.4 Prioris “Não-Informativas”

Priors não-informativas são tentativas de representar formalmente um estado de ignorância. Contudo, não existe uma forma única de representar ignorância, tampouco uma priori “objetiva”. Além disso, é bastante raro um cenário onde não há nenhuma informação a priori. De qualquer modo, serão apresentadas aqui algumas formas de representar falta de informação mas a escolha da priori será sempre subjetiva.

### 3.4.4.1 Priori de Bayes-Laplace

**Princípio da Razão Insuficiente.** Quando não existe razão suficiente para acreditar mais em algum subconjunto do espaço paramétrico  $\Theta$ , deve-se adotar equiprobabilidade.

**Exemplo 1.** Se  $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$  então a priori de Bayes-Laplace é  $f(\theta) = 1/k$ ,  $\theta \in \Theta$ .

**Exemplo 2.** Se  $\Theta = [a, b]$  então a priori de Bayes-Laplace é  $f(\theta) = 1/(b-a)$ ,  $\theta \in \Theta$ .

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int_{\Theta} f(\theta)f(x|\theta) d\theta} = \frac{c f(x|\theta)}{c \int_{\Theta} f(x|\theta) d\theta} = \frac{f(x|\theta)}{\int_{\Theta} f(x|\theta) d\theta} \propto f(x|\theta).$$

As principais críticas da priori de Bayes-Laplace são

1. A distribuição é *imprópria* quando o espaço paramétrico  $\Theta$  não é finito ou limitado. Por exemplo,  $\Theta = \mathbb{N}$ ,  $\Theta = \mathbb{Z}$  ou  $\Theta = \mathbb{R}$ . Nesses casos, a priori de Bayes-Laplace é  $f(\theta) \propto \mathbb{I}_{\Theta}(\theta)$ , que não é uma distribuição de probabilidade.
2. Não é *invariante* a reparametrizações. Considere, por exemplo,  $f(\theta)$  uma f.d.p. a priori para  $\theta$  e  $g$  uma transformação um-a-um (injetora) de  $\theta$  tal que  $\psi = g(\theta)$ . A distribuição de  $\psi$  pode ser calculada por  $f_{\psi}(\psi) = f(g^{-1}(\psi)) \left| \frac{dg^{-1}(\psi)}{d\psi} \right|$ . Assim, se  $g$  é uma transformação não linear e a distribuição a priori para  $\theta$  é uniforme, a distribuição para  $\psi$  não é uniforme, em geral.

### 3.4.4.2 Priori de Jeffreys

Seja  $g$  uma transformação um-a-um do parâmetro  $\theta$  e defina  $\psi = g(\theta)$ . Considere uma função  $h : \mathfrak{X} \times \Theta \rightarrow \mathbb{R}$ . Uma classe de distribuições a priori invariantes pode ser definida por

$$f(\theta) \propto \left( \text{Var}_{X|\theta} \left[ \frac{\partial h(X|\theta)}{\partial \theta} \mid \theta \right] \right)^{1/2}.$$

**Demo.** Para mostrar a invariância do método, considere o caso contínuo em que

$$f_\psi(\psi) = f(g^{-1}(\psi)) \left| \frac{\partial g^{-1}(\psi)}{\partial \psi} \right|.$$

Seja  $h^*(x, \psi) = h(x, g^{-1}(\psi))$ . Então

$$\frac{\partial h^*(x, \psi)}{\partial \psi} = \frac{\partial h(x, g^{-1}(\psi))}{\partial \psi} = \frac{\partial h(x, \theta)}{\partial \theta} \bigg|_{\theta=g^{-1}(\psi)} \cdot \frac{\partial g^{-1}(\psi)}{\partial \psi},$$

e, portanto,

$$\text{Var} \left[ \frac{\partial h^*(X, \psi)}{\partial \psi} \mid \theta = g^{-1}(\psi) \right] = \text{Var} \left[ \frac{\partial h(X, \theta)}{\partial \theta} \mid \theta = g^{-1}(\psi) \right].$$

$$\left[ \frac{\partial g^{-1}(\psi)}{\partial \psi} \right]^2 = \left[ f(g^{-1}(\psi)) \left( \frac{\partial g^{-1}(\psi)}{\partial \psi} \right) \right]^2,$$

de modo que

$$f_\psi(\psi) = f(g^{-1}(\psi)) \left| \frac{\partial g^{-1}(\psi)}{\partial \psi} \right| = \text{Var} \left[ \frac{\partial h^*(X, \psi)}{\partial \psi} \mid \theta = g^{-1}(\psi) \right]^{1/2}.$$

A escolha mais usual para  $h$  é  $h(x, \theta) = \log f(x|\theta)$ . Assim, como

$$E \left[ \frac{\partial \log f(X|\theta)}{\partial \theta} \mid \theta \right] = 0, \text{ temos}$$

$$f(\theta) \propto \text{Var} \left[ \frac{\partial \log f(X|\theta)}{\partial \theta} \mid \theta \right]^{1/2} = E \left[ \left( \frac{\partial \log f(X|\theta)}{\partial \theta} \right)^2 \mid \theta \right]^{1/2} = [\mathcal{I}(\theta)]^{1/2},$$

onde  $\mathcal{I}(\theta)$  é a *Informação de Fisher* de  $\theta$ . Neste caso,  $f(\theta) \propto |\mathcal{I}(\theta)|^{1/2}$  é chamada **priori de Jeffreys**.

Uma motivação para o método de Jeffreys é que a informação de Fisher  $\mathcal{I}(\theta)$  é um indicador da quantidade de informação trazida pelo modelo (observações) sobre o parâmetro  $\theta$ . Favorecer os valores de  $\theta$  para o qual  $\mathcal{I}(\theta)$  é grande supostamente minimiza a influência da priori.

**Exemplo 1.** Considere novamente o experimento de lançar uma moeda  $n$  vezes e contar o número de caras, isto é,  $X|\theta \sim \text{Bin}(n, \theta)$ . Então,

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \implies \log f(x|\theta) = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta)$$

$$\frac{\partial \log f(x|\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x-n\theta}{\theta(1-\theta)}.$$

$$\text{Como } E[X|\theta] = n\theta \text{ e } \text{Var}(X|\theta) = E[(X - E[X|\theta])^2 \mid \theta]$$



$$\begin{aligned}
&= E \left[ (X - n\theta)^2 \mid \theta \right] = n\theta(1 - \theta), \text{ a informação de Fisher neste caso} \\
&\text{é} \\
\mathcal{J}_x(\theta) &= E \left[ \left( \frac{\partial \log f(x|\theta)}{\partial \theta} \right)^2 \mid \theta \right] = E \left[ \left( \frac{X - n\theta}{\theta(1 - \theta)} \right)^2 \mid \theta \right] \\
&= \frac{1}{\theta^2(1 - \theta)^2} E \left[ (X - n\theta)^2 \mid \theta \right] = \frac{1}{\theta^2(1 - \theta)^2} \text{Var}(X \mid \theta) \\
&= \frac{n \theta(1 - \theta)}{\theta^2(1 - \theta)^2} = \frac{n}{\theta(1 - \theta)} = n\theta^{-1}(1 - \theta)^{-1},
\end{aligned}$$

de modo que a priori de Jeffreys é

$$f(\theta) \propto [\mathcal{J}_x(\theta)]^{1/2} \propto \theta^{-1/2}(1 - \theta)^{-1/2} \implies \theta \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right).$$

**Exemplo 2.** Considere agora que a mesma moeda é lançada e anota-se o número de caras  $Y$  até que sejam observadas  $r$  coroas, isto é,  $Y|\theta \sim \text{BinNeg}(r, \theta)$ . Então,  $f(y|\theta) = \binom{y+r-1}{y} \theta^y (1 - \theta)^r$   
 $\implies \log f(y|\theta) = \log \binom{y+r-1}{y} + y \log \theta + r \log(1 - \theta)$

$$\frac{\partial \log f(y|\theta)}{\partial \theta} = \frac{y}{\theta} - \frac{r}{1 - \theta} = \frac{1}{\theta} \left[ y - \frac{r \theta}{1 - \theta} \right].$$

Como  $E[X|\theta] = \frac{r \theta}{1 - \theta}$  e  $\text{Var}(X|\theta) = \frac{r \theta}{(1 - \theta)^2}$ , a informação de Fisher neste caso é

$$\begin{aligned}
\mathcal{J}_y(\theta) &= E \left[ \frac{1}{\theta^2} \left( y - \frac{r \theta}{1 - \theta} \right)^2 \mid \theta \right] = \frac{1}{\theta^2} \text{Var}(Y \mid \theta) = \frac{r}{\theta(1 - \theta)^2} = \\
&r\theta^{-1}(1 - \theta)^{-2},
\end{aligned}$$

de modo que a priori de Jeffreys é

$$f(\theta) \propto [\mathcal{J}_y(\theta)]^{1/2} \propto \theta^{-1/2}(1 - \theta)^{-1}.$$

Note que nos exemplos apresentados, a priori depende da *regra de parada*, isto é, a forma como decidimos quando parar de lançar a moeda e que determina se o modelo estatístico é binomial ou binomial negativo. Em outras palavras, a opinião a priori definida dessa forma depende do modelo adotado, mesmo que o parâmetro seja o mesmo nos dois casos. Além disso, a priori de Jeffreys pode ser *imprópria*, como ocorre no exemplo anterior.

### 3.4.4.3 Priori de Máxima Entropia

**Entropia** é um conceito físico que quantifica a desordem ou imprevisibilidade de um sistema, ou da falta de informação sobre ele. O conceito de entropia desempenha um importante papel na teoria da informação. O *princípio da máxima entropia* afirma que a distribuição de probabilidade que melhor representa a falta de informação é aquela com a maior entropia.

**Caso Discreto.** Considere um espaço paramétrico enumerável  $\Theta = \{\theta_1, \theta_2, \dots\}$ . A *entropia* da distribuição  $h$  (Shannon, 1948) é dada por

$$\mathcal{E}(h) = E[-\log h(\theta)] = - \sum_{\theta \in \Theta} \log [h(\theta)] h(\theta) .$$

**Definição.** Considere um espaço paramétrico  $\Theta$  e  $h$  uma f.d.p. para  $\theta$ . A *distribuição da máxima entropia* para  $\theta$  é a função  $h$  que maximiza  $\mathcal{E}(h)$  (Jaynes, 2003)

**Exemplo 1.** Considere o espaço paramétrico  $\Theta = \{\theta_1, \dots, \theta_k\}$  e  $h(\theta_i) = p_i$  uma distribuição discreta para  $\theta$ . A *distribuição da máxima entropia* para  $\theta$  é a função  $h$  que maximiza

$$\mathcal{E}(h) = - \sum_{i=1}^k p_i \log(p_i) \text{ com a restrição } \sum_{i=1}^k h(\theta_i) = \sum_{i=1}^k p_i = 1 .$$

Utilizando o método de multiplicadores de Lagrange, deve-se maximizar a função lagrangiana

$$\mathcal{E}^*(h) = - \sum_{i=1}^k p_i \log(p_i) + \lambda \left( \sum_{i=1}^k p_i - 1 \right)$$

$$\frac{\partial \mathcal{E}^*(h)}{\partial p_i} = - \left[ p_i \frac{1}{p_i} + \log(p_i) \right] + \lambda = 0 \iff p_i = e^{\lambda-1} , \quad i = 1, \dots, k .$$

Assim, como  $p_i$  deve ser constante e  $\sum p_i = 1$ , conclui-se que  $p_i = 1/k$ , para  $i = 1, \dots, k$ .

**Exemplo 2.** Considere agora  $\Theta = \{\theta_1, \theta_2, \dots\}$  e suponha que há  $m$  informações parciais a respeito do parâmetro  $\theta$  que podem ser escritas como  $E[g_j(\theta)] = \mu_j$ ,  $j = 1, \dots, m$ .

Usando novamente o método de Lagrange, deve-se maximizar

$$\mathcal{E}^*(h) = \sum_{i=1}^{\infty} p_i \log(p_i) + \lambda \left( \sum_{i=1}^{\infty} p_i - 1 \right) + \sum_{j=1}^m \lambda_j \left( \sum_{i=1}^{\infty} p_i g_j(\theta_i) - \mu_j \right)$$

$$\frac{\partial \mathcal{E}^*(h)}{\partial p_i} = -\log(p_i) - 1 + \lambda + \sum_{j=1}^m \lambda_j g_j(\theta_i) = 0 \iff p_i \propto$$

$$e^{\lambda - 1 + \sum_{j=1}^m \lambda_j g_j(\theta_i)} \propto e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}, \quad i = 1, \dots, k.$$

Como  $\sum p_i = 1$ ,  $p_i = \frac{e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}}{\sum_{i=1}^k e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}}$  e  $\lambda_j$  é obtido por meio das restrições.

**Exemplo 2a.** Seja  $\Theta = \{0, 1, 2, \dots\}$  e suponha que  $E[\theta] = \mu$ .

Usando o resultado do exemplo anterior com  $g(\theta) = \theta$  e  $\theta_i = i$ ,  $i = 0, 1, 2, \dots$ ,

$$p_i = \frac{e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}}{\sum_{i=0}^{\infty} e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}} = \frac{e^{\lambda i}}{\sum_{i=0}^{\infty} e^{\lambda i}} \stackrel{|e^\lambda| < 1}{=} \frac{e^{\lambda i}}{1/(1 - e^\lambda)}$$

$$= (e^\lambda)^i (1 - e^\lambda) \Rightarrow \theta \sim \text{Geo}(1 - e^\lambda).$$

Como  $E[\theta] = \frac{e^\lambda}{(1 - e^\lambda)} = \mu$ , tem-se que  $\lambda = \log \frac{\mu}{1 + \mu}$ .

**Exemplo 2b.** Considere que  $\Theta = \{1, 2, \dots, k\}$  e suponha que  $\text{Med}(\theta) = m$ .

Nesse caso,  $g(\theta) = \mathbb{I}(\theta \leq m)$  e  $\theta_i = i$ ,  $i = 1, 2, \dots, k$ , de modo que  $E[g(\theta)] = E[\mathbb{I}(\theta \leq m)] = P(\theta \leq m) = 1/2$  e, portanto,  $\sum_{i \leq m} p_i =$

$$\sum_{j > m} p_j = 1/2 \cdot p_i = \frac{e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}}{\sum_{i=1}^k e^{\sum_{j=1}^m \lambda_j g_j(\theta_i)}} = \begin{cases} \frac{e^\lambda}{\sum_{i \leq m} e^\lambda}, & i \leq m \\ \frac{1}{\sum_{i \leq m} 1}, & i > m \end{cases}$$

$$= \begin{cases} \frac{1}{2m}, & i \leq m \\ \frac{1}{2(k - m)}, & i > m \end{cases}$$

(A distribuição de  $\theta$  é uniforme por blocos.)

**Divergência de Kullbach-Leibler.** Considere duas distribuições discretas  $p = (p_1, \dots, p_k)$  e  $q = (q_1, \dots, q_k)$ , tal que  $p_i, q_i > 0$ ,  $i = 1, \dots, k$ , e  $\sum p_i = \sum q_i = 1$ . A *divergência de Kullbach-Leibler* entre  $p$  e  $q$  (Kullback and Leibler, 1951) é dada por

$$D(p \parallel q) = \sum p_i \log \left( \frac{p_i}{q_i} \right).$$

Suponha que  $g = (1/k, \dots, 1/k)$

$$D(p \parallel q) = \sum_{i=1}^k p_i \log \left( \frac{p_i}{1/k} \right) = \sum_{i=1}^k p_i [\ln(p_i) - \ln(1/k)] = \sum_{i=1}^k p_i \ln(p_i) + \ln(k) \sum_{i=1}^k p_i = \ln(k) - \mathcal{E}(p)$$

Assim, exceto por uma constante,  $\mathcal{E}(p)$  está associado com quanto a distribuição  $p$  “diverge” da distribuição uniforme (priori de referência na ausência total de informação).

**Observação:** No caso geral, se  $H$  e  $H_0$  são duas medidas definidas em  $\Theta$  tais que  $H$  é absolutamente contínua com relação à  $H_0$  ( $H \ll H_0$ ), a divergência de Kullbach-Leibler é definida como

$$D(H \parallel H_0) = \int_{\Theta} \log \left( \frac{dH}{dH_0} \right) dH ,$$

em que  $\frac{dH}{dH_0}$  é derivada de Radon-Nikodym. Se  $H$  e  $H_0$  são medidas de probabilidade absolutamente contínuas com relação a medida de Lebesgue  $\lambda$  com f.d.p.  $\frac{dH}{d\lambda} = h$  e  $\frac{dH_0}{d\lambda} = h_0$ , temos que,

$$D(H \parallel H_0) = \int_{\Theta} \log \left( \frac{dH/d\lambda}{dH_0/d\lambda} \right) \frac{dH}{d\lambda} d\lambda = \int_{\Theta} \log \left( \frac{h(\theta)}{h_0(\theta)} \right) h(\theta) d\theta$$

Como a definição anterior de entropia vale apenas para o caso discreto, Jaynes (2003) sugere que no caso contínuo seja utilizada a **entropia relativa**, dada por

$$\mathcal{E}(h) = - \int_{\Theta} h(\theta) \log \left( \frac{h(\theta)}{h_0(\theta)} \right) d\theta = -D(h \parallel h_0) ,$$

onde  $h_0$  é uma priori de referência na ausência total de informação, preferivelmente invariante.

Assim como no caso discreto, se temos  $m$  restrições  $E[g_i(\theta)] = \mu_i$ , a densidade de máxima entropia é

$h(\theta) \propto h_0(\theta) \exp \left\{ \sum_{j=1}^m \lambda_j g_j(\theta) \right\}$  e os  $\lambda_j$ ,  $j = 1, \dots, m$ , são obtidos das restrições.

Por exemplo, se  $g(\theta) = \theta$  com  $E[\theta] = \mu$ , basta fazer

$$\mu = \int_{\Theta} \theta c h_0(\theta) \exp\{\lambda\theta\} d\theta \text{ com } c^{-1} = \int_{\Theta} h_0(\theta) \exp\{\lambda\theta\} d\theta.$$

**Exemplo 1:**  $\Theta = \mathbb{R}_+$  e  $E[\theta] = \mu$ .

Tomando  $h_0(\theta) \propto \mathbb{I}_{\mathbb{R}_+}(\theta)$  (f.d.p. imprópria), tem-se  $h(\theta) \propto e^{\lambda\theta} \mathbb{I}_{\mathbb{R}_+}(\theta) \propto -\lambda e^{\lambda\theta} \mathbb{I}_{\mathbb{R}_+}(\theta) \mathbb{I}_{\mathbb{R}_-}(\lambda)$ .

Como  $E[\theta] = -1/\lambda = \mu$ , tem-se que  $\lambda = -1/\mu$ , isto é,  $\theta \sim \text{Exp}(1/\mu)$ , de modo que  $h(\theta) = \frac{1}{\mu} e^{-\frac{\theta}{\mu}}$ ,  $\mu > 0$ .

**Exemplo 2**  $\Theta = \mathbb{R}$  e  $E[\theta] = \mu$  e  $\text{Var}(\theta) = E[(\theta - \mu)^2] = \sigma^2$ .

Tomando  $g_1(\theta) = \theta$  e  $g_2(\theta) = (\theta - \mu)^2$ , tem-se pelo resultado anterior que

$$\begin{aligned} h(\theta) &\propto \exp\{\lambda_1\theta + \lambda_2(\theta - \mu)^2\} \propto \exp\{\lambda_1\theta + \lambda_2(\theta^2 - 2\theta\mu + \mu^2)\} \\ &\propto \exp\left\{\lambda_2\left[\theta^2 - \left(2\mu - \frac{\lambda_1}{\lambda_2}\right)\theta\right]\right\} \propto \exp\left\{\lambda_2\left[\theta - \left(\mu - \frac{\lambda_1}{2\lambda_2}\right)\right]^2\right\}. \end{aligned}$$

$$\begin{aligned} \text{Considere que } \theta \sim N(\mu, \sigma^2), \text{ isto é, } f(\theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}. \end{aligned}$$

Assim, para concluir que a distribuição de máxima entropia nesse caso é a Normal anterior, basta tomar  $\mu - \frac{\lambda_1}{2\lambda_2} = \mu$  para ver que

$$\lambda_1 = 0 \text{ e } \lambda_2 = -\frac{1}{2\sigma^2}.$$

### 3.5 Alguns Princípios de Inferência

Considere um experimento  $E = (X, \theta, \{f(x|\theta)\})$  que consiste em observar um particular valor  $x \in \mathcal{X}$  do v.a.  $X$  que, para cada possível valor do parâmetro (desconhecido)  $\theta \in \Theta$ , tem f.d.p.  $f(x|\theta)$ . De forma geral, uma *inferência* sobre  $\theta$  baseada no resultado  $x$  do experimento  $E$  será denotada por  $\text{Inf}(E, x)$ .

**Princípio de Suficiência.** Considere um experimento  $E = (X, \theta, \{f(x|\theta)\})$  e suponha que  $T(X)$  é uma estatística suficiente para  $\theta$ . Se  $x_1$  e  $x_2$  são dois pontos amostrais tais que  $T(x_1) = T(x_2)$  então  $\text{Inf}(E, x_1) = \text{Inf}(E, x_2)$ .

**Exemplo 1a.** Seja  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_1 \sim \text{Ber}(\theta)$ .

Considere  $n = 10$  e os pontos amostrais  $x_1 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$  e  $x_2 = (1, 0, 1, 0, 1, 0, 1, 0, 1, 1)$  tais que  $T(x_1) = \sum x_{1i} = 6$  e  $T(x_2) = \sum x_{2i} = 6$ .

Um possível estimador para  $\theta$  nesse exemplo é a média amostral, de modo que  $\bar{x}_1 = \bar{x}_2 = \frac{\sum x_i}{n} = 0,6$ .

**Exemplo 1b.** Ainda no contexto do exemplo anterior, considere que a priori  $\theta \sim \text{Beta}(a, b)$ . Então, se  $T(x_1) = T(x_2) = t$ ,

$$\theta|x_1 \sim \theta|x_2 \sim \theta|T(x_1) = t \sim \text{Beta}(a + t, b + n - t).$$

**Princípio da Condicionalidade.** Suponha que  $E_1 = (X_1, \theta, \{f(x_1|\theta)\})$  e  $E_2 = (X_2, \theta, \{f(x_2|\theta)\})$  são dois experimentos onde somente o parâmetro  $\theta$  precisa ser comum. Considere um experimento misto em que é observada uma v.a.  $J$ , com  $P(J = 1) = P(J = 2) = 1/2$ , independente de  $X_1$ ,  $X_2$  e  $\theta$ , e então o experimento  $E_J$  é realizado. Formalmente, o experimento realizado nesse caso é  $E^* = (X^*, \theta, \{f^*(x^*|\theta)\})$ , onde  $X^* = (J, X_J)$  e  $f^*(x|\theta) = \frac{1}{2} f_j(x_j|\theta)$ . Então,  $\text{Inf}(E^*, (j, x_j)) = \text{Inf}(E_j, x_j)$ .

**Princípio da Verossimilhança.** Suponha dois experimentos  $E_1 = (X_1, \theta, \{f_1(x_1|\theta)\})$  e  $E_2 = (X_2, \theta, \{f_2(x_2|\theta)\})$ , ambos com o mesmo parâmetro  $\theta$ . Suponha que  $x_1$  e  $x_2$  são pontos amostrais de  $E_1$  e  $E_2$ , respectivamente, tais que  $f_1(x_1|\theta) \propto c(x_1, x_2)f_2(x_2|\theta)$ ,  $\forall \theta \in \Theta$ , então,  $\text{Inf}(E_1, x_1) = \text{Inf}(E_2, x_2)$ .

**Teorema de Birnbaum.**  $(P. \text{ Suficiência} \wedge P. \text{ Condicionalidade}) \iff P. \text{ Verossimilhança}.$

**Demo:**

$(\implies)$

Seja  $x_1^*$ ,  $x_2^*$ ,  $E_1$ ,  $E_2$  como no *P. Verossimilhança* e  $E^*$  como no *P.*

*Condicionalidade*. Então,

$$f_1(x_1|\theta) \propto c(x_1, x_2)f_2(x_2|\theta).$$

No espaço do experimento  $E^*$ , defina  $T(j, x_j) = \begin{cases} (1, x_1^*), & \text{se } j = 1, x_1 = x_1^* \\ (j, x_j), & \text{c. c.} \end{cases}$ .

Como  $f^*(x^*|\theta) = f^*((j, x_j)|\theta) = 1/2 f_j(x_j|\theta)$ , pelo o Teorema da Fatoração é possível concluir que  $T(j, x_j)$  é suficiente para  $\theta$  no experimento  $E^*$ .

Então, pelo *P. Suficiência*,  $\text{Inf}(E^*, (1, x_1)) = \text{Inf}(E^*, (2, x_2))$  e, pelo *P. Condicionalidade*,

$\text{Inf}(E^*, (1, x_1^*)) = \text{Inf}(E_1, x_1^*) = \text{Inf}(E^*, (2, x_2)) = \text{Inf}(E_2, x_2^*)$ , de modo que  $\text{Inf}(E_1, x_1^*) = \text{Inf}(E_2, x_2^*)$  e, portanto, vale o *P. Verossimilhança*.

( $\Leftarrow$ )

Como vale o *P. Verossimilhança*,  $f_1(x_1^*|\theta) \propto f_2(x_2^*|\theta)$  e  $\text{Inf}(E_1, x_1^*) = \text{Inf}(E_2, x_2^*)$ .

Além disso, se  $x^* = (1, x_1^*)$ ,

$f^*(x^*|\theta) = f^*((1, x_1^*)|\theta) = 1/2 f_1(x_1^*|\theta) \propto f_1(x_1^*|\theta) \propto 1/2 f_2(x_2^*|\theta) = f^*((2, x_2^*)|\theta)$ ,

e, como vale *P. Verossimilhança*, então  $\text{Inf}(E^*, (1, x_1^*)) = \text{Inf}(E_1, x_1^*)$ .

Usando o mesmo argumento, se  $x^* = (2, x_2^*)$ , conclui-se que  $\text{Inf}(E^*, (2, x_2^*)) = \text{Inf}(E_2, x_2^*)$ .

Portando, vale o *P. Condicionalidade*.

Pelo Teorema de Fatoração,  $f(x|\theta) = g(T(x), \theta) h(x) \propto g(T(x), \theta)$ . Se  $x_1$  e  $x_2$  são pontos amostrais tais que  $T(x_1) = T(x_2)$ ,  $f_1(x_1|\theta) \propto g(T(x_1), \theta) \propto g(T(x_2), \theta) \propto f_2(x_2|\theta)$ , tem-se, pelo *P. Verossimilhança*, que  $\text{Inf}(E_1, x_1) = \text{Inf}(E_2, x_2)$  e, portanto vale o *P. Suficiência*.

**Exemplo.** Seja  $X_1|\theta \sim \text{Bin}(n, \theta)$  e  $X_2|\theta \sim \text{BinNeg}(r, \theta)$ , onde  $n$  é número total de lançamentos (fixado) e  $r$  é número de fracassos (fixado). Então,  $E_1 = (X_1, \theta, \left\{ \binom{n}{x_1} \theta^{x_1} (1-\theta)^{n-x_1} : \theta \in [0, 1] \right\})$  e  $E_2 = (X_2, \theta, \left\{ \binom{r+x_2-1}{x_2} \theta^{x_2} (1-\theta)^r : \theta \in [0, 1] \right\})$ . Note que em ambos os experimentos, o parâmetro  $\theta$  é o mesmo!

(I) Estimação pontual usando Estimador Não-Viesado (ENV) para  $\theta$ , isto é,  $\hat{\theta}_i(X_i)$  tal que  $E[\hat{\theta}_i(X_i)|\theta] = \theta$ . Nesse caso,  $\text{Inf}(E_i, x_i) = \hat{\theta}_i(x_i)$  para  $i = 1, 2$ .

Então,  $\hat{\theta}_1(X_1) = \frac{X_1}{n}$  e  $\hat{\theta}_2(X_2) = \frac{X_2 - 1}{X_2 + r - 1}$  são ENV para  $\theta$  em  $E_1$  e  $E_2$ , respectivamente.

Suponha que  $n = 12, r = 3$  e  $x_1 = x_2 = 9$ . Então, as funções de

verossimilhança são  $f_1(x_1|\theta) = \binom{12}{9}\theta^9(1-\theta)^3 \propto \binom{11}{9}\theta^9(1-\theta)^3 = f_2(x_2|\theta)$ . Contudo,  $\hat{\theta}_1(x_1) = \frac{9}{12} = 0,75 \neq \hat{\theta}_2(x_2) = \frac{8}{11} \approx 0,727\bar{2}$ , e portanto, o ENV **viola** o  $P$ . *Verossimilhança*.

(II) Estimador de Máxima Verossimilhança (EMV)

$\delta_{MV}$  é um estimador tal que  $\delta_{MV}(x) = \arg \sup_{\theta \in \Theta} f(x|\theta)$ .

$$\delta_{MV}^1(x_1) = \frac{x_1}{n} = \delta_{MV}^2(x_2) = \frac{x_2}{x_2 + r} = \frac{9}{12} = 0,75.$$

Portanto, o EMV **não viola** o  $P$ . *Verossimilhança*.

(III) Suponha que deseja-se testar  $H_0 : \theta \leq 1/2$  ( $\Theta_0$ ) contra  $H_1 : \theta > 1/2$  ( $\Theta_1$ ), com  $\Theta = \Theta_0 \cup \Theta_1$ .

$$\phi(x) = \begin{cases} 1, & T(x) \leq c(\alpha) \\ 0, & T(x) > c(\alpha) \end{cases}$$

em que  $T$  é uma estatística de teste (isto é, valores “grandes” de  $T(x)$  indicam que  $x$  é “favorável” a  $H_0$ ) e  $c(\alpha)$  é tal que  $\alpha = \sup_{\theta_0 \in \Theta_0} P(\text{Rejeitar } H_0 \mid \theta_0) = \sup_{\theta_0 \in \Theta_0} P(\{x \in \mathfrak{X} : T(x) \leq c(\alpha)\} \mid \theta_0)$ .

Considere  $T(x) = RV(x) = \frac{\sup_{\Theta_0} f(x|\theta)}{\sup_{\Theta} f(x|\theta)}$ , de modo que um  $p$ -value

pode ser calculado por  $p(x) = \sup_{\Theta_0} P(T(X) \geq T(x) \mid \theta)$ . Assim, um

teste que conduz a uma decisão equivalente ao descrito anteriormente é *rejeitar*  $H_0$  se, e somente se,  $p(x) \leq \alpha$ . Considere a escolha usual  $\alpha = 0.05$ . Então,

$$p_1(x_1) = P(X_1 \geq 9 \mid \theta = 1/2) = 0.073 > 0.05 \Rightarrow \text{Não rejeita } H_0.$$

$$p_2(x_2) = P(X_2 \geq 9 \mid \theta = 1/2) = 0.0327 < 0.05 \Rightarrow \text{Rejeita } H_0.$$

Portanto, o Teste da Razão de Verossimilhanças viola o  $P$ . *Verossimilhança*.

(IV) Aboragem Bayesiana  $\Rightarrow Inf(E_i, x_i) = f_i(\theta|x_i)$

a) *Bayesiano Subjetivista*

Como o parâmetro  $\theta$  é o mesmo nos dois experimentos, a priori deve ser a mesma.

$f(\theta)$  não depende de  $\{f_i(x|\theta) : \theta \in \Theta\}$

$$f(\theta|x) \propto f(\theta)f(x_1|\theta) \propto f(\theta)f(x_2|\theta)$$

e, portanto, satisfaz o  $P$ . *Verossimilhança*.

b) *Bayesiano Objetivista* (p.e., usando priori de Jeffreys)

Para  $E_1$ ,  $f_1(\theta) \propto |I_F(\theta)|^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \sim \text{Beta}(1/2, 1/2)$



Para  $E_2$ ,  $f_2(\theta) \propto \theta^{-1}(1-\theta)^{-1/2} \sim \text{Beta}(0, 1/2)$  (distribuição imprópria).

Se o número de sucessos é  $x = x_1 = x_2$  e número de fracassos é  $y = n - x_1 = r$ , temos que

$\theta|X_1 = x_1 \sim \text{Beta}(x+1/2, y+1/2)$  e  $\theta|X_2 = x_2 \sim \text{Beta}(x, y+1/2)$ .

Como  $f_1(x_1|\theta) \propto f_2(x_2|\theta)$  mas  $f_1(\theta) \neq f_2(\theta)$ , tem-se que  $f_1(\theta|x_1) \neq f_2(\theta|x_2)$  e, portanto, esse procedimento viola o *P. Verossimilhança*.



## Chapter 4

# Introdução à Teoria da Decisão

A teoria da decisão é uma das possíveis formas de embasar a inferência bayesiana. Sob essa abordagem, considera-se uma *função de perda* (ou *função de utilidade*) que quantifica numericamente as consequências de sua decisão para um dado valor do parâmetro. Essa quantificação de “preferência” é novamente subjetiva e é possível fazer uma construção de função de perda similar ao que fizemos com probabilidade. Ou seja, dado um conjunto de suposições, existe uma função de perda que representa numericamente suas preferências para cada decisão e cada possível valor do parâmetro. Essa construção não será feita aqui mas pode ser encontrada no livro *Optimal Statistical Decisions* (DeGroot, 1970).

### 4.1 Conceitos Básicos

- $d \in \mathcal{D}$  : **decisão** - uma particular afirmação, por exemplo, sobre  $\theta$ . No contexto inferencial, uma decisão pode ser uma estimativa (pontual ou intervalar) para  $\theta$  ou a escolha de uma hipótese específica em um teste de hipóteses.
- $\mathcal{D}$  : **espaço de decisões** - conjunto de todas as possíveis decisões (afirmações).
- $\theta$  : **estado da natureza** - quantidade desconhecida ou parâmetro, no contexto de inferência estatística.
- $\Theta$  : **espaço dos estados da natureza** - espaço paramétrico.

**Exemplo 1.** Suponha que você está saindo de casa pela manhã e precisa tomar uma importante decisão: levar ou não seu guarda-chuva.

- $\mathcal{D} = \{G, G^c\}$ , onde  $G$  : levar guarda-chuva.
- $\Theta = \{C, C^c\}$ , onde  $C$  : chuva.

Suponha que carregar o guarda-chuva é algo que não lhe agrada mas, por outro lado, você odeia ficar molhado e acredita que a pior situação seria não levá-lo e tomar chuva. Você ficará incomodado se levar o guarda-chuva e chover pois, além de tê-lo carregado, voltou para casa com os sapatos molhados. Note que, nessas circunstâncias, o cenário preferido por você seria não levar o guarda-chuva e não chover.

Para quantificar suas preferências, considere uma função de perda  $L : \mathcal{D} \times \Theta \rightarrow \mathbb{R}$ , de modo que, quanto mais algum cenário lhe gera incômodo, maior sua perda. Um exemplo é apresentado a seguir.

	Estados da Natureza	
<b>Decisão</b>	$C$	$C^c$
$G$	2 (ruim)	1 (bom)
$G^c$	3 (pior)	0 (melhor)
$P(\theta)$	p	1-p

Uma possível maneira de tomar uma decisão é escolher a decisão “menos prejudicial”. Se levar o Guarda chuva, no pior caso, sua perda é  $\max_{\theta} L(G, \theta) = 2$  e, se não levá-lo, a maior perda possível é  $\max_{\theta} L(G^c, \theta) = 3$ . Assim, a decisão que tem a menor dentre as maiores perdas é levar o guarda-chuva. Esse procedimento para tomada de decisões é chamado *min-max* e consiste em escolher a decisão  $d'$  tal que  $d' = \operatorname{argmin}_d \max_{\theta} L(d, \theta)$ .

Sendo um pouco mais otimista, você pode escolher a decisão que tenha a maior dentre as menores perdas. Esse procedimento é chamado *max-min* e consiste em escolher a decisão  $d' = \operatorname{argmax}_d \min_{\theta} L(d, \theta)$ . No nosso exemplo, esse procedimento também sugere que você sempre carregue o guarda-chuvas.

Note que a decisão escolhida pelos dois procedimentos descritos anteriormente sugere que você sempre deve carregar o guarda-chuvas. Contudo, isso pode não ser razoável. Imagine que você estava lendo notícias antes de sair de casa e viu que a probabilidade de chuva era 0.01. Nesse caso, não parece fazer sentido você levar o guarda-chuva, já que isso vai te trazer um desconforto e a chance de chover é muito baixa. Assim, a probabilidade de chover deveria ser levada

em consideração em sua tomada de decisão.

Uma maneira de fazer isso é utilizar a **perda esperada**. Note que  $\theta$  é uma quantidade desconhecida e, pelo que já foi discutido anteriormente, você deve descrever sua incerteza em relação a essa quantidade em termos de probabilidade. Suponha que no exemplo  $P(C) = p$ ,  $0 \leq p \leq 1$ .

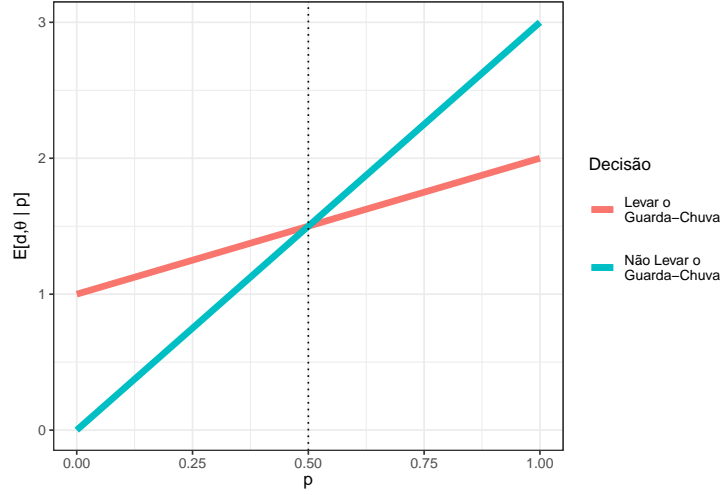
Para cada decisão  $d \in \mathcal{D}$ , é possível calcular o valor esperado da função de perda (**perda esperada** ou **risco** da decisão  $d$  contra a priori  $P$ )

$$\rho(d, P) = E[L(d, \theta) \mid P] = \int_{\Theta} L(\theta) dP(\theta).$$

No exemplo, temos

- $E[L(G, \theta)] = L(G, C)P(C) + L(G, C^c)P(C^c) = 2p + 1(1 - p) = p + 1$ ;
- $E[L(G^c, \theta)] = L(G^c, C)P(C) + L(G^c, C^c)P(C^c) = 3p + 0(1 - p) = 3p$ .

Deste modo, as perdas esperadas associadas a cada decisão dependem da probabilidade de chuva  $p$ . Assim, para cada possível valor de  $p$ , deve-se tomar a decisão que tem menor perda esperada. Por exemplo, se a probabilidade de chuva é  $p = 0.1$ , temos que as perdas esperadas para as decisões de levar ou não o guarda-chuva são, respectivamente,  $E[L(G, \theta)] = 1.1$  e  $E[L(G^c, \theta)] = 0.3$ . Assim, sob essa abordagem, sua decisão seria de não levar o guarda-chuva nesse caso. Por outro lado, se a probabilidade de chuva for  $p = 0.9$ , suas perdas esperadas seriam respectivamente  $E[L(G, \theta)] = 1.9$  e  $E[L(G^c, \theta)] = 2.7$ , de modo que a decisão ótima seria levar o guarda-chuva. O gráfico a seguir apresenta as perdas para cada decisão  $d$  e para cada valor de  $p$ . É possível notar que a decisão ótima é levar o guarda-chuva quando  $p > 0.5$  e não levá-lo caso contrário.



Vamos denotar por  $\rho^*$  o **risco de bayes**, isto é, a perda esperada da **decisão de Bayes** (ou *decisão ótima*)  $d^* \in \mathcal{D}$  tal que  $\rho^*(P) = \rho(d^*, P) = \min_{d \in \mathcal{D}} \rho(d, P)$ .

Para uma argumentação mais formal sobre a escolha pela decisão que minimiza a perda esperada, ver *Optimal Statistical Decisions* (DeGroot, M.H.).

- Vamos denotar um problema de decisão por  $(\Theta, \mathcal{D}, L, P)$ , onde  $\Theta$  é o espaço paramétrico,  $\mathcal{D}$  é o espaço de decisões,  $L : \mathcal{D} \times \Theta \rightarrow \mathbb{R}$  é uma função de perda e  $P$  é a distribuição de probabilidade que representa sua crença sobre a quantidade desconhecida  $\theta$ . Equivalentemente, a função de perda  $L$  pode ser substituída por uma *função de utilidade*  $U$  (por exemplo, tome  $U = -L$ ).

- A solução para um problema de decisão  $(\Theta, \mathcal{D}, L, P)$  é a *decisão de Bayes*,  $d^* \in \mathcal{D}$ , tal que  $\rho^*(P) = \rho(d^*, P) = \inf_{d \in \mathcal{D}} \rho(d, \theta)$ , com

$$\rho(d, P) = \int_{\Theta} L(d, \theta) dP(\theta).$$

## 4.2 Aleatorização e Decisões Mistas

Seja  $D = \{d_1, d_2, \dots\}$  um espaço de decisões e considere  $\mathcal{M}$  o conjunto de todas as *decisões mistas* (ou *aleatorizadas*), isto é, para toda distribuição de probabilidades  $Q = \{q_1, q_2, \dots\}$ , uma decisão  $d \in \mathcal{M}$  se  $d$  consiste em escolher a decisão  $d_i$  com probabilidade  $q_i$ .

Assim, a perda associada à uma decisão  $d \in \mathcal{M}$  é  $L(d, \theta) = \sum q_i L(d_i, \theta)$  e o risco dessa decisão é

$$\rho(d, P) = \int_{\Theta} L(d, \theta) dP(\theta) = \int_{\Theta} \sum q_i L(d_i, \theta) dP(\theta) = \sum q_i \int_{\Theta} L(d_i, \theta) dP(\theta) = \sum q_i \rho(d_i, \theta).$$

Considere a decisão  $d^* \in \mathcal{D}$  tal que  $\rho(d^*, P) = \inf_{d \in \mathcal{D}} \rho(d, \theta)$ .

Então,  $\forall d \in \mathcal{M}$ ,

$$\rho(d, P) = \sum q_i \rho(d_i, \theta) \geq \sum q_i \rho(d^*, \theta).$$

- Em palavras, para toda decisão aleatorizada  $d \in \mathcal{M}$ , existe uma decisão não aleatorizada  $d^* \in \mathcal{D} \subset \mathcal{M}$ , tal que  $\rho(d^*, P) \leq \rho(d, P)$ .

## 4.3 Problemas com Dados

Suponha que antes de escolher uma decisão  $d \in \mathcal{D}$ , é possível observar um v.a.  $X$  que (supostamente) está relacionado com  $\theta$  (isto é,  $X$  traz alguma informação sobre  $\theta$ ).

Desde modo, considere a família  $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$  de funções de distribuição condicionais para  $X$ , isto é, para cada  $\theta \in \Theta$  é possível determinar a distribuição condicional de  $X|\theta$ . Essa distribuição, juntamente com a distribuição a priori  $f(\theta)$ , determina totalmente uma distribuição conjunta  $f(x, \theta) = f(x|\theta)f(\theta)$ .

Pode-se definir uma **função de decisão**  $\delta : \mathfrak{X} \rightarrow \mathcal{D}$  que associa a cada resultado experimental  $x \in \mathfrak{X}$  uma decisão  $d \in \mathcal{D}$ . Denote o conjunto de todas as possíveis funções de decisão por  $\Delta$ .

O risco  $r(\delta, P)$  da função de decisão  $\delta \in \Delta$  é dado por  $r(\delta, P) = E[L(\delta, \theta)] = \int_{\Theta} \int_{\mathfrak{X}} L(\delta(x), \theta) dP(x, \theta)$ .

A *função de decisão de Bayes*,  $\delta^* \in \Delta$ , é tal que  $\rho^*(P) = \rho(\delta^*, P) = \inf_{\delta \in \Delta} \rho(\delta, P)$ .

**Exemplo 1.** Seja  $\Theta = \{\theta_1, \theta_2\}$ ,  $\mathcal{D} = \{d_1, d_2\}$ ,  $X|\theta_1 \sim \text{Ber}(3/4)$ ,  $X|\theta_2 \sim \text{Ber}(1/3)$ ,  $\mathfrak{X} = \{0, 1\}$  e, a priori,  $P(\theta = 3/4) = P(\theta = 1/3) = 1/2$ . Considere a função de perda a seguir.

L	$\theta_1$	$\theta_2$
$d_1$	0	5
$d_2$	10	0

Temos que  $|\Delta| = 2^2 = 4$ , de modo que as possíveis funções de decisão são

$$\delta_1(x) = \begin{cases} d_1, & x = 1 \\ d_2, & x = 0 \end{cases}$$

$$\delta_2(x) = \begin{cases} d_1, & x = 0 \\ d_2, & x = 1 \end{cases}$$

$$\delta_3(x) = d_1 \text{ e } \delta_4(x) = d_2.$$

Para a função  $\delta_1$ , temos

x	$\theta$	$L(\delta_1(x), \theta)$	$P(x \theta)$	$P(\theta)$	$P(x, \theta)$
0	$\theta_1$	10	1/4	1/2	1/8
0	$\theta_2$	0	2/3	1/2	2/6
1	$\theta_1$	0	3/4	1/2	3/8
1	$\theta_2$	5	1/3	1/2	1/6

$$\rho(\delta_1) = \sum_{x=0}^1 \sum_{i=1}^2 L(\delta_1(x), \theta_i) \underbrace{P(X=x|\theta_i)P(\theta_i)}_{P(x, \theta)} = 10 \frac{1}{8} + 5 \frac{1}{6} = \frac{50}{24}$$

De forma análoga,  $\rho(\delta_2, P) = 130/24$ ,  $\rho(\delta_3, P) = 60/24$ ,  $\rho(\delta_4, P) = 120/24$ , e, assim.

$$\delta^*(x) = \delta_1^*(x) = \begin{cases} d_1, & x = 1 \\ d_2, & x = 0 \end{cases}$$

**Risco de Bayes:**  $\rho^*(P) = \rho(\delta^*, P) = 50/24$ .

Em problemas mais complicados, pode ser muito trabalhoso (ou impossível) obter a função de decisão dessa forma, chamada *forma normal*. Sob essa abordagem, é necessário encontrar a função de decisão de bayes  $\delta^*$  dentre todas as possíveis funções de decisão. Nesses casos, pode ser mais fácil resolver o problema usando a *forma extensiva* em que, para cada  $x \in \mathfrak{X}$ , obtem-se a decisão de Bayes  $d_x^*$  que minimiza o *risco posterior*, definido por  $r_x(d) = \int_{\Theta} L(d, \theta) dP(\theta|x)$ .



Assim, é possível obter uma decisão de Bayes  $d_x^*$  para um específico ponto  $x$  observado ou, ainda, construir uma função de decisão de Bayes, fazendo  $\delta^*(x) = d_x^*$  para cada  $x \in \mathfrak{X}$ . A seguir, é mostrado que essas duas formas produzem resultados que minimizam o risco. Note que

$$\begin{aligned} r(\delta, P) &= E[L(\delta, \theta)] = \int_{\Theta} \int_{\mathfrak{X}} L(\delta(x), \theta) dP(x, \theta) = \int_{\Theta} \int_{\mathfrak{X}} L(\delta(x), \theta) dP(x|\theta) dP(\theta) \\ &= \int_{\mathfrak{X}} \left[ \underbrace{\int_{\Theta} L(\delta(x), \theta) dP(\theta|x)}_{r_x(\delta(x))} \right] dP(x). \end{aligned}$$

Note que a integral interna (em  $\theta$ ) pode ser resolvida para cada  $x$  fixado. Para cada  $x$ , considere a decisão  $d_x^*$  tal que  $r_x(d_x^*) = \inf_{d \in \mathcal{D}} r_x(d)$ . Assim

$$\begin{aligned} r(\delta, P) &= \int_{\mathfrak{X}} \left[ \underbrace{\int_{\Theta} L(\delta(x), \theta) dP(\theta|x)}_{r_x(\delta(x))} \right] dP(x) = \int_{\mathfrak{X}} [r_x(\delta(x))] dP(x) \geq \\ &\int_{\mathfrak{X}} [r_x(d_x^*)] dP(x) = \int_{\mathfrak{X}} [r_x(d_x^*)] dP(x). \end{aligned}$$

Assim, a função  $\delta^*(x) = d_x^*$  é uma *função de decisão de Bayes*.

**No Exemplo 1**  $X|\theta_1 \sim \text{Ber}(3/4)$ ,  $X|\theta_2 \sim \text{Ber}(1/3)$  e  $P(\theta_1) = P(\theta_2) = 1/2$ .

$$P(\theta_1|x=0) = \frac{P(X=0|\theta_1)P(\theta_1)}{P(X=0|\theta_1)P(\theta_1) + P(X=0|\theta_2)P(\theta_2)} = \frac{\frac{1}{4} \frac{1}{2}}{\frac{1}{4} \frac{1}{2} + \frac{2}{3} \frac{1}{2}} = \frac{3}{11}$$

$$P(\theta_2|x=0) = \frac{8}{11}$$

$$r_x(d_1, P) = \sum_{i=1}^2 L(d_1, \theta_i) P(\theta_i|x=0) = 0 P(\theta_1|x=0) + 10 P(\theta_2|x=0) = \frac{80}{11}$$

$$r_x(d_2, P) = 5 P(\theta_1|x=0) + 0 P(\theta_2|x=0) = \frac{15}{11}$$

Logo, para  $x=0$ ,  $d_0^* = d_2$ . De forma análoga, para  $x=1$ ,  $d_1^* = d_2$  e, assim,

$$\delta^*(x) = \begin{cases} d_2, & x=0 \\ d_1, & x=1. \end{cases}$$



## Chapter 5

# Estimação

### 5.1 Estimação Pontual

Todos os problemas de inferência estatística podem ser vistos como um caso particular de Teoria da Decisão. Um problema de estimação pontual consiste em encontrar um “chute” para o valor do parâmetro  $\theta$ , de modo que o espaço de decisões é  $\mathcal{D} = \Theta$ . Além disso, nesse tipo de problema é usual considerar funções de perda da forma  $L(d, \theta) = s(\theta)\Delta(d, \theta)$ , onde  $\Delta$  é alguma distância (ou uma medida de discrepância) relacionada ao erro por tomar a decisão  $d$  quando o valor do parâmetro é  $\theta$  e  $s$  é uma função não-negativa relacionada à gravidade do erro para cada  $\theta$  (pode ser constante).

**Exemplo** Considere um problema de estimação pontual, isto é,  $\mathcal{D} = \Theta$ , onde a função de perda é dada por  $L(d, \theta) = (d - \theta)^2$ , conhecida como *perda quadrática*.

$$\begin{aligned} r_x(d) &= \int_{\Theta} (d - \theta)^2 dP(\theta|x) = \int_{\Theta} (d^2 - 2d\theta + \theta^2) dP(\theta|x) = d^2 \int_{\Theta} dP(\theta|x) - \\ &2d \int_{\Theta} \theta dP(\theta|x) + \int_{\Theta} \theta^2 dP(\theta|x) = d^2 - 2d E[\theta|x] + E[\theta^2|x] = g(d). \end{aligned}$$

$$\frac{\partial g(d)}{\partial d} = 2d - 2E[\theta|x] = 0 \Rightarrow d_x^* = E[\theta|x].$$

Logo, um estimador para  $\theta$  contra a perda quadrática é  $\delta^*(X) = E[\theta|X]$ .

- Estimador de Bayes para  $\theta$  contra diferentes funções de perda:

$$1. \text{ Perda Quadrática: } L_2(d, \theta) = (d - \theta)^2 \implies \delta_2^*(X) = E[\theta|X];$$

2. Perda Absoluta:  $L_1(d, \theta) = |d - \theta| \implies \delta_1^*(X) = \text{Med}(\theta|X)$ ;
3. Perda 0-1:  $L_0(d, \theta) = c \mathbb{I}(d \neq \theta) \implies \delta_0^*(X) = \text{Moda}(\theta|X)$ .

**Exemplo 1.** Voltando à *Perda Quadrática*:  $L(d, \theta) = a(d - \theta)^2$ ,  $a > 0$ .

Já vimos que  $\delta^*(X) = E[\theta|X]$ . É importante notar que esse estimador só faz sentido se  $E[\theta|X] \in \mathcal{D}$ . Nos casos em que isso não ocorre, tomamos um valor  $d_x^* \in \mathcal{D}$  próximo a  $E[\theta|X]$  tal que  $r_x(d_x^*)$  é mínimo.

O risco a posteriori para esse estimador é

$$r_x(\delta^*(x)) = r_x(E[\theta|x]) = \int_{\Theta} L(\delta^*(x), \theta) dP(\theta|x) = \int_{\Theta} (\theta - E[\theta|x])^2 dP(\theta|x) = \text{Var}(\theta|x),$$

de modo que o risco de Bayes é dado por

$$\rho^*(P) = \rho(\delta^*(X), P) = \int_{\mathfrak{X}} \underbrace{\int_{\Theta} (\theta - E[\theta|x])^2 dP(\theta|x)}_{\text{Var}[\theta|x]} dP(x) = E[\text{Var}(\theta|X)].$$

A variância da posteriori  $\text{Var}(\theta|x)$  pode ser vista como uma medida de informação, no sentido que quanto menor essa variância, mais concentrada é a distribuição e há “*menos incerteza*” sobre  $\theta$ . Nesse sentido, espera-se que ao observar  $X = x$ , a variância  $\text{Var}(\theta|x)$  diminua em relação a variância da priori  $\text{Var}(\theta)$ .

$$\underbrace{\text{Var}(\theta)}_{\text{constante}} = \underbrace{E[\text{Var}(\theta|X)]}_{\downarrow} + \underbrace{\text{Var}[E(\theta|X)]}_{\uparrow}$$

Aparentemente, quando espera-se que a variância da posteriori diminua, a variância do estimador deveria aumentar. Muitas vezes isso é colocado como se o objetivo fosse encontrar o estimador de maior variância, em contradição com a abordagem frequentista. Contudo, há outra interpretação desse resultado: deseja-se obter um estimador que varie bastante de acordo com o valor observado de  $X$ , isto é, a informação trazida pela amostra muda sua opinião sobre  $\theta$ .

**Exemplo 2.** Considere  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim \text{Poisson}(\theta)$  e, a priori,  $\theta \sim \text{Gama}(a, b)$ .

A função de verossimilhança é  $f(x|\theta) \propto \prod \theta^{x_i} e^{-\theta}$  e a priori é  $f(\theta) \propto \theta^{a-1} e^{-b\theta}$ .

Assim,  $f(\theta|x) \propto \theta^{\sum x_i} e^{-n\theta} \cdot \theta^{a-1} e^{-b\theta} \propto \theta^{a+\sum x_i-1} e^{-(b+n)\theta}$ , de modo que  $\theta|X = x \sim \text{Gama}(a + \sum x_i, b + n)$ .

Como visto anteriormente, o estimador de Bayes contra a perda quadrática é

$$\delta^*(X) = E[\theta|X] = \frac{a + \sum_i X_i}{b + n}.$$

Para calcular o risco de Bayes, note que  $E[X_i|\theta] = \theta$ , de modo que  $E[X_i] = E[E(X_i|\theta)] = E[\theta] = a/b$ . Além disso,  $\text{Var}(\theta) = \frac{a}{b^2}$  e  $\text{Var}(\theta|X) = \frac{a + \sum_i X_i}{(b + n)^2}$ . Então,

$$\rho^*(P) = E[\text{Var}(\theta|X)] = E\left[\frac{a + \sum_i X_i}{(b + n)^2}\right] = \frac{(a + \sum E[X_i])}{(b + n)^2} = \frac{\frac{a}{b}(b + n)}{(b + n)^2} = \frac{a}{b(b + n)}.$$

Por fim, note que a decisão de Bayes pode ser escrita como uma combinação linear convexa da média da distribuição a priori e do estimador de máxima verossimilhança

$$E[\theta|X] = \frac{a + \sum_i X_i}{b + n} = \frac{b}{b + n} \left(\frac{a}{b}\right) + \frac{n}{b + n} \bar{X}.$$

**Resultado:** Seja  $L(d, \theta) = \mathbb{I}(|\theta - d| > \varepsilon) = 1 - \mathbb{I}(d - \varepsilon \leq \theta \leq d + \varepsilon)$ ,  $\varepsilon > 0$ . Então,  $\delta^*(X)$  é centro do intervalo modal, isto é, o intervalo de tamanho  $2\varepsilon$  de maior densidade a posteriori. Em particular, quando  $\varepsilon \downarrow 0$ , temos que  $\delta^*(X) = \text{Moda}(\theta|X)$ .

**Demo:** O risco posterior de uma decisão  $d$  é

$$r_x(d) = E[L(d, \theta)|x] = E[\mathbb{I}(|\theta - d| > \varepsilon)] = E[1 - \mathbb{I}(d - \varepsilon \leq \theta \leq d + \varepsilon)|x] = 1 - P(d - \varepsilon \leq \theta \leq d + \varepsilon|x).$$

O risco  $r_x(d)$  é mínimo quando a probabilidade  $P(d - \varepsilon \leq \theta \leq d + \varepsilon|x)$  é máxima. Assim, basta tomar o intervalo  $[d_x^* - \varepsilon; d_x^* + \varepsilon]$  com maior probabilidade a posteriori e o estimador de Bayes nesse caso será o valor central desse intervalo,  $d_x^*$ .

**Exemplo 3.** Considere o exemplo anterior onde  $\theta|x \sim \text{Gama}(a + \sum x_i, b + n)$  e a função de perda do resultado anterior,  $L(d, \theta) = \mathbb{I}(|\theta - d| > \varepsilon)$ . Temos que

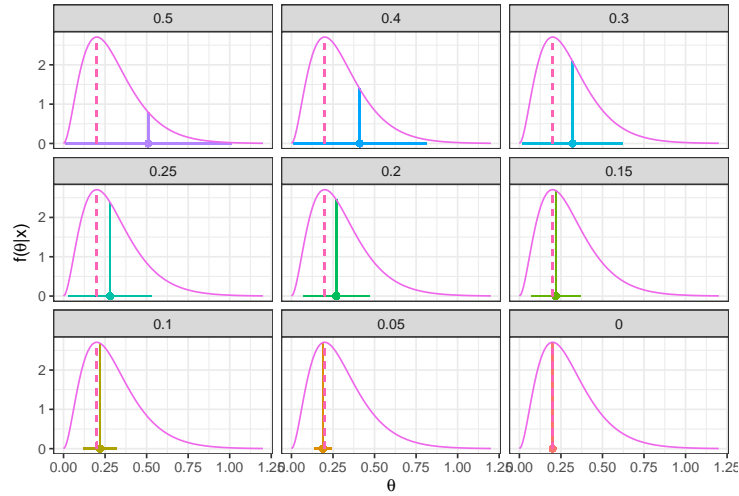
$$f(\theta|x) \propto \theta^{\overbrace{a+\sum x_i-1}^{A_x}} e^{-\overbrace{(b+n)\theta}^{B_x}} = \theta^{A_x} e^{-B_x \theta}$$

```
theta = seq(0,1.2,0.01)
#Parâmetros da dist a posteriori gama
a1 = 3
```

```

b1 = 10
posterior = dgamma(theta, a1, b1 )
# Escolhendo valores de epsilon e calculando a perda mínima associada
e=c(0.5,0.4,0.3,0.25,0.2,0.15,0.1,0.05,0)
loss <- NULL
for(i in 1:length(e)){
  loss[i] <- theta[which.min(as.vector(apply(matrix(theta), 1,
                                              function(d) sum(posterior*(abs(theta-d)>
# Criando o gráfico
n <- length(e)
tibble(x=rep(loss,each=length(theta)),
       e=rep(round(e,2),each=length(theta)),
       theta=rep(theta,(n)), post = rep(posterior,n)) %>%
ggplot() +
geom_segment(aes(x=x, xend=x, y=0,yend=dgamma(x, a1, b1 ),colour=as.factor(e))) +
geom_point(aes(x=x, y=0,colour=as.factor(e))) +
geom_segment(aes(x=x-e, xend=x+e, y=0,yend=0,colour=as.factor(e))) +
geom_line(aes(x=theta,y=post,colour="Dist. a posteriori")) +
geom_segment(aes(x=(a1-1)/b1,xend=(a1-1)/b1,y=0,yend=dgamma((a1-1)/b1,a1,b1),colour=
xlab(expression(theta)) + ylab(expression(paste("f(",theta,"|x)"))) +
theme_bw() + labs(colour = "epsilon") + theme(legend.position="none")+
facet_wrap(~factor(e,levels=as.character(sort(unique(e),decreasing=TRUE))))

```



Como comentado no resultado anterior, quando  $\varepsilon \downarrow 0$ , temos que  $\delta^*(x) = \text{Moda}(\theta|x) = \sup_{\theta} f(\theta|x)$ .

$$\frac{\partial f(\theta|x)}{\partial \theta} = (A_x - B_x \theta) \theta^{A_x-1} e^{-B_x \theta} = 0 \Leftrightarrow \theta = \frac{A_x}{B_x}$$

$$\delta^*(X) = \text{Moda}(\theta|X) = \frac{a + \sum X_i - 1}{b + n} = \frac{b}{b + n} \left( \frac{a - 1}{b} \right) + \frac{n}{b + n} \bar{X}.$$

**Resultado.** Seja  $L(d, \theta) = c_1(d - \theta) \mathbb{I}(d \geq \theta) + c_2(\theta - d) \mathbb{I}(d < \theta)$  com  $c_1 > 0$ ,  $c_2 > 0$ . Então,  $\delta^*(x)$  é tal que  $P(\theta \leq \delta^*(x)|x) = \frac{c_1}{c_1 + c_2}$ . Em particular, se  $c_1 = c_2 = c$ , temos a *perda absoluta*  $L(d, \theta) = c |d - \theta|$  e  $\delta^*(X) = \text{Med}(\theta|X)$ .

**Demo:** exercício.

## 5.2 Estimação por Regiões

Em um problema de estimação por regiões (ou estimação intervalar, no caso univariado), o objetivo é obter um conjunto de valores razoáveis para  $\theta$ . Mais formalmente, temos um problema aonde a decisão consiste em escolher um subconjunto do espaço paramétrico, de modo que  $\mathcal{D} = \mathcal{A}$ , onde  $\mathcal{A}$  é  $\sigma$ -álgebra de subconjuntos de  $\Theta$ . Uma estimativa por região é comumente chamada na literatura Bayesiana de **região de credibilidade** (**intervalo de credibilidade**, no caso univariado) ou **região de probabilidade**  $\gamma = 1 - \alpha$ .

**Exemplo 1.** Suponha que a distribuição a posteriori é  $f(\theta|x) = 4\theta \mathbb{I}_{[0,1/2)}(\theta) + 4(1 - \theta) \mathbb{I}_{[1/2,1]}(\theta)$ . Uma possível estimativa intervalar é um intervalo central ou um intervalo simétrico em torno da média (ou da moda) com uma probabilidade  $\gamma = 1 - \alpha$ . Nesse caso, vamos considerar um intervalo central no sentido que deixa de fora conjuntos caudais de probabilidade  $\alpha/2$ . Note que é possível obter o intervalo de forma analítica nesse exemplo. A seguir, são apresentadas as funções de distribuição  $F$  e quantílica  $Q$  a posteriori e o intervalo de credibilidade  $\alpha$ .

$$F(\theta|x) = \int_0^\theta f(t|x)dt = \int_0^\theta [4t \mathbb{I}_{[0,1/2]}(t) + (4 - 4t) \mathbb{I}_{(1/2,1]}(t)] dt = \begin{cases} 2\theta^2 & , \quad \theta \leq 1/2 \\ -2\theta^2 + 4\theta - 1 & , \quad \theta > 1/2 \end{cases}$$

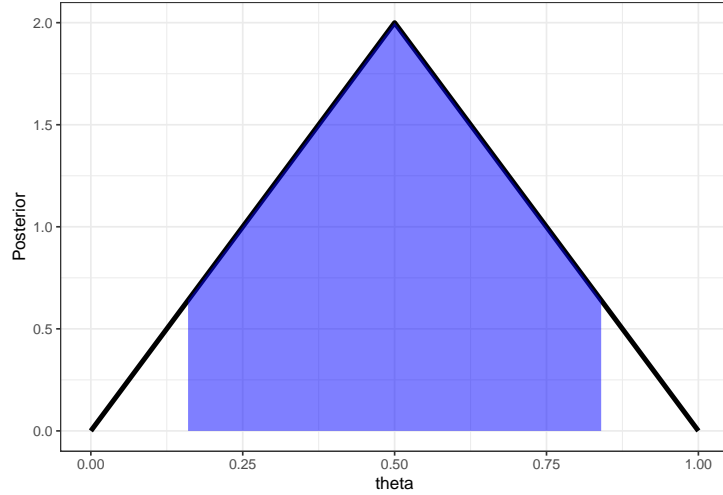
$$Q(p) = \begin{cases} \sqrt{\frac{p}{2}} & , \quad p \leq 1/2 \\ 1 - \sqrt{\frac{1-p}{2}} & , \quad p > 1/2 \end{cases}$$

$$I.C.(1 - \alpha) = \left[ \sqrt{\frac{\alpha/2}{2}}; 1 - \sqrt{\frac{\alpha/2}{2}} \right].$$

```

alpha=0.1
theta = seq(0,1,0.01)
# Densidade a posteriori
dpost = Vectorize(function(t){ 4*t*I(t>=0)*I(t<=0.5)+
  4*(1-t)*I(t>0.5)*I(t<=1) })
# F. Distribuição a posteriori
ppost = Vectorize(function(t){ 2*(t^2)*I(t>=0)*I(t<=0.5)+
  4*(1-t)*I(t>0.5)*I(t<=1)+I(t>1) })
# F. Quantílica a posteriori
qpost = Vectorize(function(t){ ifelse(t<=0.5, sqrt(t/2)*I(t>0),
  (1-sqrt((1-t)/2))*I(t<1)+I(t>=1)) })
post = dpost(theta)
l = c(qpost((alpha/2)),qpost((1-alpha/2)))
X = tibble(theta=theta,Posterior=post)
ggplot(data=X,mapping = aes(x=theta,y=Posterior)) +
  geom_line(lwd=1.5) +
  geom_area(data=subset(X, theta >= l[1] & theta <= l[2]),fill = "blue", alpha=0.5)+
  theme_bw()

```



**Exemplo 2.** Por fim, suponha que a distribuição a posteriori é  $f(\theta|x) = (2 - 4\theta) \mathbb{I}_{[0,1/2)}(\theta) + (4\theta - 2) \mathbb{I}_{[1/2,1]}(\theta)$ . Vamos construir um intervalo como no Exemplo anterior.

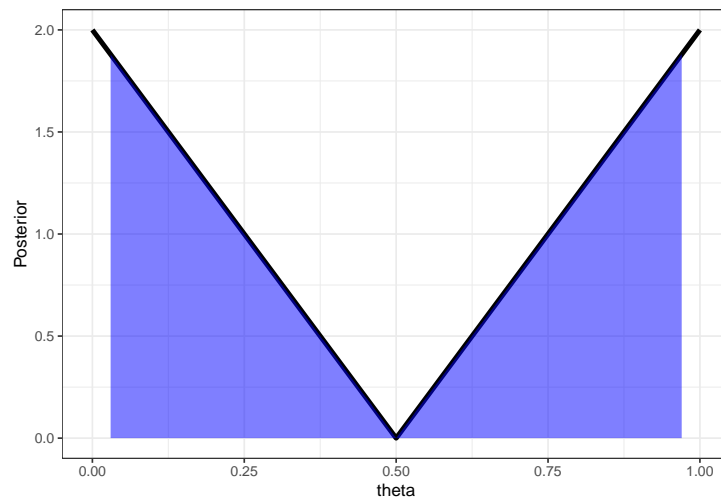
$$F(\theta|x) = \begin{cases} 2\theta(1-\theta) & , \quad \theta \leq 1/2 \\ 2\theta(\theta-1) + 1 & , \quad \theta > 1/2 \end{cases}$$



$$Q(p) = \begin{cases} \frac{1}{2} - \frac{\sqrt{1-2p}}{2}, & p \leq 1/2 \\ \frac{1}{2} + \frac{\sqrt{2p-1}}{2}, & p > 1/2 \end{cases}$$

$$I.C.(1-\alpha) = \left[ \frac{1}{2} - \frac{\sqrt{1-\alpha}}{2}; \frac{1}{2} + \frac{\sqrt{1-\alpha}}{2} \right].$$

```
alpha=0.1
theta = seq(0,1,0.01)
# Densidade a posteriori
dpost = Vectorize(function(t){ (2-4*t)*I(t>=0)*I(t<=0.5)+
  (4*t-2)*I(t>0.5)*I(t<=1) })
# F. Distribuição a posteriori
ppost = Vectorize(function(t){ 2*t*(1-t)*I(t>=0)*I(t<=0.5)+
  (2*(t^2)-2*t+1)*I(t>0.5)*I(t<=1)+I(t>1) })
# F. Quantílica a posteriori
qpost = Vectorize(function(t){ ifelse(t<=0.5, (0.5-sqrt(1-2*t))/2)*I(t>0),
  (0.5+sqrt(2*t-1)/2)*I(t<=1)+I(t>=1) })
post = dpost(theta)
l = c(qpost((alpha/2)),qpost((1-alpha/2)))
X = tibble(theta=theta,Posterior=post)
ggplot(data=X,mapping = aes(x=theta,y=Posterior)) +
  geom_line(lwd=1.5) +
  geom_area(data=subset(X, theta >= l[1] & theta <= l[2]),fill = "blue", alpha=0.5)+
  theme_bw()
```



Note que, nesse exemplo, as regiões que tem mais densidade a posteriori foram excluídas do intervalo. Isso não faz muito sentido pois essas regiões têm maior

chance de conter o  $\theta$  que qualquer outra região de mesmo tamanho.

Uma *função de perda* razoável para um problema de estimação por região deve levar em consideração dois fatores:

- O tamanho da região  $d \in \mathcal{A}$  (deseja-se uma região que seja menor que o espaço paramétrico);
- Pertinência de  $\theta$  na região  $d$ .

Assim, considere uma função de perda da forma

$$L(d, \theta) = \lambda(d) - k \mathbb{I}_d(\theta),$$

onde  $\lambda(d)$  é o “tamanho” da região  $d$ . Por exemplo, a medida de Lebesgue, no caso contínuo, ou a medida de contagem, no caso discreto (no caso geral, considere uma medida que domina a distribuição a posteriori,  $P(\theta|x) \ll \lambda$ ). No caso absolutamente contínuo, o risco a posteriori de uma decisão  $d \in \mathcal{A}$  é

$$\begin{aligned} r_x(d) &= \int_{\Theta} [\lambda(d) - k \mathbb{I}_d(\theta)] dP(\theta|x) = \int_{\Theta} \mathbb{I}_d(\theta) d\theta - \int_{\Theta} k \mathbb{I}_d(\theta) f(\theta|x) d\theta \\ &= \int_d (1 - kf(\theta|x)) d\theta. \end{aligned}$$

Esse risco é mínimo quando  $d = \{\theta \in \Theta : 1 - kf(\theta|x) \leq 0\} \Leftrightarrow d = \{\theta \in \Theta : f(\theta|x) \geq 1/k\}$ .

Assim, a decisão de Bayes contra essa função de perda consiste em escolher uma região  $d \in \mathcal{A}$  que contém os pontos do espaço paramétrico com maior densidade a posteriori.

**Definição:** A região  $R \subseteq \Theta$  é dita ser uma região HPD (Highest Posterior Density) de probabilidade  $\gamma$  se

- $P(\theta \in R|x) = \gamma$ ;
- $\forall \theta \in R$  e  $\forall \theta' \notin R$ ,  $f(\theta|x) \geq f(\theta'|x)$ .

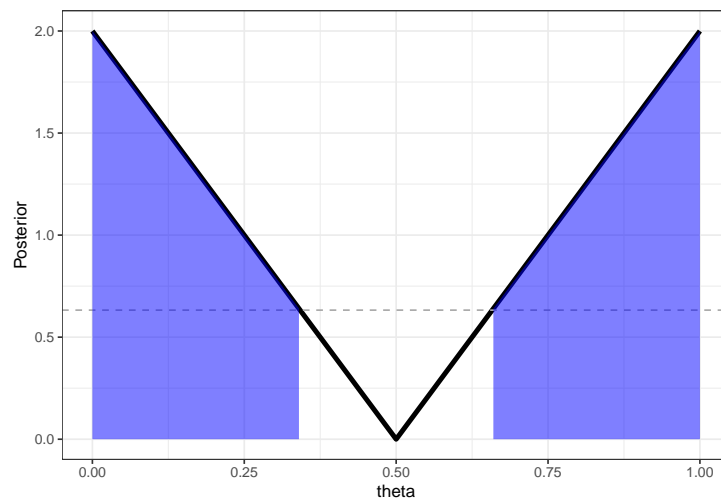
**Voltando ao Exemplo 2.** As regiões centrais nesse exemplo tem menor densidade a posteriori. Assim, uma região HPD de probabilidade  $\gamma = 1 - \alpha$  é dada por

$$I.C.(1 - \alpha) = \left[0 ; \frac{1}{2} - \frac{\sqrt{\alpha}}{2}\right] \cup \left[\frac{1}{2} + \frac{\sqrt{\alpha}}{2} ; 1\right]$$

```

alpha=0.1
theta = seq(0,1,0.01)
# Densidade a posteriori
dpost = Vectorize(function(t){ (2-4*t)*I(t>=0)*I(t<=0.5)+
  (4*t-2)*I(t>0.5)*I(t<=1) })
# F. Distribuição a posteriori
ppost = Vectorize(function(t){ 2*t*(1-t)*I(t>=0)*I(t<=0.5)+
  (2*(t^2)-2*t+1)*I(t>0.5)*I(t<=1)+I(t>1) })
# F. Quantílica a posteriori
qpost = Vectorize(function(t){ ifelse(t<=0.5, (0.5-sqrt(1-2*t))/2)*I(t>0),
  (0.5+sqrt(2*t-1)/2)*I(t<1)+I(t>=1)) })
post = dpost(theta)
l = c(qpost((1-alpha)/2),qpost((alpha+1)/2))
X = tibble(theta=theta,Posterior=post)
ggplot(data=X,mapping = aes(x=theta,y=Posterior)) +
  geom_line(lwd=1.5) +
  geom_hline(yintercept=dpost(l[1]), lty=2, col="darkgray") +
  geom_area(data=subset(X, theta <= l[1]),fill = "blue", alpha=0.5) +
  geom_area(data=subset(X, theta >= l[2]),fill = "blue", alpha=0.5)+
  theme_bw()

```



Note que na solução anterior, o comprimento do intervalo era  $\sqrt{1-\alpha}$ , enquanto que o comprimento do HPD é  $1-\sqrt{\alpha}$ . Tomando, por exemplo,  $\alpha = 0.1$  temos que  $\sqrt{1-\alpha} \approx 0.95$  enquanto  $1-\sqrt{\alpha} \approx 0.68$ . Por conter apenas os pontos com maior densidade, o HPD sempre terá comprimento menor ou igual a qualquer intervalo com mesma probabilidade.

**Exemplo 3.** Considere  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim N(\theta, 1/\tau)$ ,  $\tau = 1/\sigma^2$ , com  $\tau$  conhecido (fixado). Vimos que se, a priori,  $\theta \sim N(m, 1/v)$  então  $\theta|x \sim$

$$N\left(\underbrace{\frac{vm + n\tau\bar{x}}{v + n\tau}}_{M_x}, \underbrace{\frac{1}{v + n\tau}}_{V_x}\right) \text{ e, assim, } Z = \frac{\theta - M_x}{\sqrt{V_x}} \Bigg| x \sim N(0, 1).$$

O intervalo HPD de probabilidade  $\gamma = 1 - \alpha = 0.95$  é

$$I.C.(1 - \alpha) = \left[ M - z_{\alpha/2}\sqrt{V}; M + z_{\alpha/2}\sqrt{V} \right] = \left[ \frac{vm + n\tau\bar{x}}{v + n\tau} \pm 1.96\sqrt{\frac{1}{v + n\tau}} \right].$$

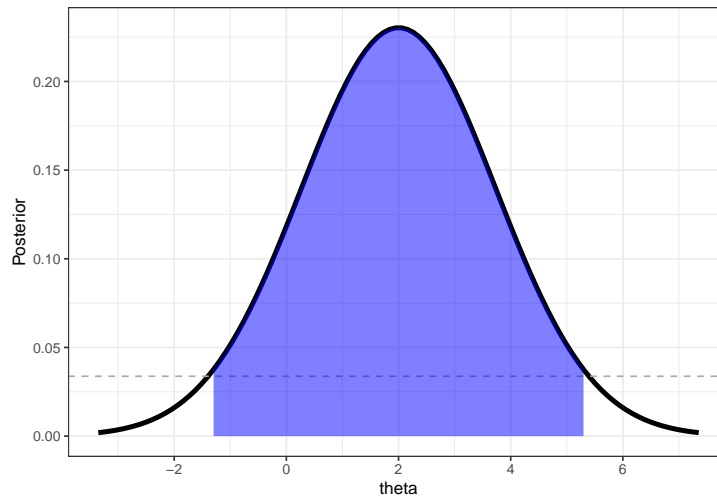
Uma possível forma de representar falta de informação a priori é tomar o limite  $v \downarrow 0$  ( $1/v \uparrow \infty$ ). Dessa forma, tem-se

$$\theta|x \sim N(\bar{x}, 1/(n\tau)) \sim N(\bar{x}, \sigma^2/n),$$

e o intervalo HPD coincide com o I.C. frequentista

$$I.C.(1 - \alpha) = \left[ \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \right].$$

```
mx=2; vx=sqrt(3)
alpha=0.05
theta = seq(qnorm(0.001,mx,vx),qnorm(0.999,mx,vx),length.out=100)
post = dnorm(theta,mx,vx)
l = c(qnorm(alpha/2,mx,vx),qnorm(1-alpha/2,mx,vx))
X = tibble(theta=theta,Posterior=post)
ggplot(data=X,mapping = aes(x=theta,y=Posterior)) +
  geom_line(lwd=1.5) +
  geom_hline(yintercept=dnorm(l[1],mx,vx), lty=2, col="darkgray") +
  geom_area(data=subset(X, theta >= l[1] & theta <= l[2]),fill = "blue", alpha=0.5)+
  theme_bw()
```



**Exemplo 4:** Considere  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim Unif(0, \theta)$ . Vimos que se  $\theta \sim Pareto(a, b)$ , então  $\theta|x \sim Pareto(a + n, \max\{x_{(n)}, b\})$

$$f(\theta|x) = \frac{(a+n)[\max\{x_{(n)}, b\}]^{a+n}}{\theta^{a+n+1}} \mathbb{I}_{[\max\{x_{(n)}, b\}, \infty)}$$

Note que a função de densidade a posteriori é estritamente decrescente de modo que o extremo inferior da região HPD é  $\max\{x_{(n)}, b\}$ . A função de distribuição a posteriori é

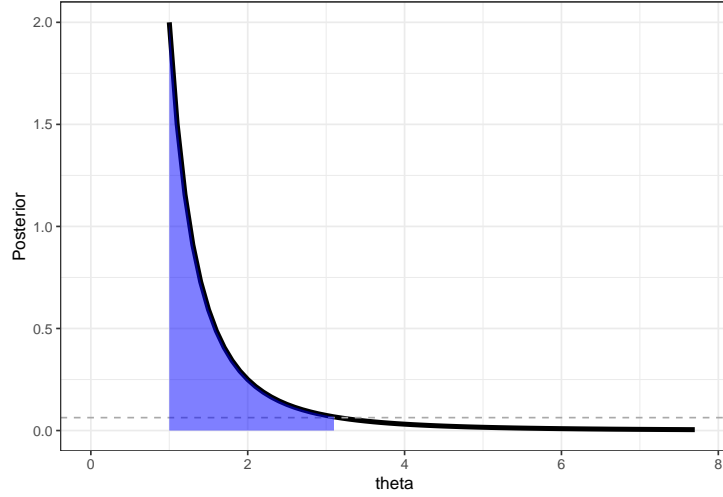
$$F(\theta|x) = 1 - \left( \frac{\max\{x_{(n)}, b\}}{\theta} \right)^{a+n},$$

de modo que o extremo superior do intervalo pode ser obtido fazendo

$$1 - \left( \frac{\max\{x_{(n)}, b\}}{\theta} \right)^{a+n} = \gamma \Leftrightarrow \frac{\max\{x_{(n)}, b\}}{\theta^*} = (1 - \gamma)^{1/(a+n)} \Leftrightarrow \theta^* = \frac{\max\{x_{(n)}, b\}}{(1 - \gamma)^{1/(a+n)}}.$$

$$I.C.(1 - \alpha) = \left[ \max\{x_{(n)}, b\}, \frac{\max\{x_{(n)}, b\}}{\alpha^{1/(a+n)}} \right]$$

```
ax=2; bx=1
maxt=bx/((alpha/3)^(1/ax))
alpha=0.1
limsup=bx/(alpha^(1/ax))
theta = seq(bx,maxt,0.1)
dpareto=Vectorize(function(t){
  ax*(bx^ax)*I(t>=bx) / (t^(ax+1)) })
post = dpareto(theta)
X = tibble(theta=theta,Posterior=post)
ggplot(data=X,mapping = aes(x=theta,y=Posterior)) +
  geom_line(lwd=1.5) +
  xlim(c(0,maxt))+
  geom_hline(yintercept=dpareto(limsup), lty=2, col="darkgray") +
  geom_area(data=subset(X, theta <= limsup),fill = "blue", alpha=0.5)+
  theme_bw()
```



### 5.3 Custo das Observações

Suponha agora que o custo para observar uma amostra de tamanho  $n$  é dado por uma função custo  $c(n)$  e, antes de observar  $X_1, \dots, X_n$ , você precisa decidir qual o tamanho amostral ótimo,  $n^*$ . Desta forma, considere a função de perda  $L(d, \theta, n) = L(d, \theta) + c(n)$  com risco

$$\rho_n(\delta, P) = E[L(d, \theta, n)] = E[L(d, \theta) + c(n)] = \rho_n(\delta, P) + c(n).$$

Note que (supostamente, por simplicidade) o custo  $c(n)$  não depende de  $\theta$ . Se  $\delta^*$  é função de decisão de Bayes contra  $L(d, \theta)$  e a priori  $P$ , o tamanho amostral ótimo  $n^*$  é o valor que minimiza

$$\rho_n(P) = \rho(\delta, P) + c(n) = \rho^*(P) + c(n).$$

**Exemplo.** Considere o exemplo visto anteriormente em que  $\mathcal{D} = \{d_1, d_2\}$ ,  $\Theta = \{\theta_1, \theta_2\} = \{3/4, 1/3\}$ ,  $P(\theta_1) = 1/2$  e a função de perda é  $L(d, \theta) = 10 \mathbb{I}(d_1, \theta_2) + 5 \mathbb{I}(d_2, \theta_1)$ . Se  $X|\theta \sim \text{Ber}(\theta)$ , a função de decisão de Bayes é  $\delta^*(x) = d_1 \mathbb{I}(x = 1) + d_2 \mathbb{I}(x = 0)$ .

Suponha agora que é possível observar  $X_1, \dots, X_n$  v.a. c.i.i.d. tais que  $X_i|\theta \sim \text{Ber}(\theta)$ . Note que  $T(X) = \sum X_i$  é suficiente para  $\theta$  com  $T(X)|\theta \sim \text{Bin}(n, \theta)$  e

$$\begin{aligned}
f(\theta_1|T(X)=t) &= \frac{f(t|\theta_1)P(\theta_1)}{\sum_{i \in \{1,2\}} f(t|\theta_i)P(\theta_i)} = \frac{f(t|\theta_1)}{\sum_{i \in \{1,2\}} f(t|\theta_i)} = \frac{\binom{n}{t} \left(\frac{3}{4}\right)^t \left(\frac{1}{4}\right)^{n-t}}{\binom{n}{t} \left(\frac{3}{4}\right)^t \left(\frac{1}{4}\right)^{n-t} + \binom{n}{t} \left(\frac{1}{3}\right)^t \left(\frac{2}{3}\right)^{n-t}} \\
&= \frac{1}{1 + \left(\frac{1}{6}\right)^t \left(\frac{8}{3}\right)^n} = p_x.
\end{aligned}$$

O risco posterior das decisões  $d_1$  e  $d_2$  são, respectivamente,  $r_x(d_1) = 10(1 - p_x)$  e  $r_x(d_2) = 5p_x$ , de modo que decide-se por  $d_1$  se

$$\begin{aligned}
r_x(d_1) \leq r_x(d_2) &\iff 10(1 - p_x) \leq 5p_x \iff p_x \geq \frac{10}{15} = \frac{2}{3} \iff \frac{1}{1 + \left(\frac{1}{6}\right)^t \left(\frac{8}{3}\right)^n} \geq \\
\frac{2}{3} &\iff \left(\frac{1}{6}\right)^t \left(\frac{8}{3}\right)^n \leq \frac{1}{2} \iff \left(\frac{1}{6}\right)^t \left(\frac{8}{3}\right)^n \leq \frac{1}{2} \iff -t \log(6) \leq n \log\left(\frac{3}{8}\right) - \log(2) \\
&\iff t \geq -n \log_6\left(\frac{3}{8}\right) + \log_6(2) = k_n,
\end{aligned}$$

e a função de decisão de Bayes é

$$\delta^*(X) = \begin{cases} d_1 & , \quad \sum X_i \geq k_n \approx 0.55n + 0.39 \\ d_2 & , \quad \text{caso contrário} \end{cases}$$

O risco de Bayes neste caso é

$$\begin{aligned}
\rho^*(P) &= E[L(\delta^*(x), \theta)] = 10 P(\delta^*(x) = d_1, \theta = \theta_2) + 5 P(\delta^*(x) = d_2, \theta = \theta_1) \\
&= 10 \frac{1}{2} P\left(\sum X_i \geq k_n \mid \theta = \frac{1}{3}\right) + 5 \frac{1}{2} P\left(\sum X_i < k_n \mid \theta = \frac{3}{4}\right) = \\
&= 5 P\left(\text{Bin}\left(n, \frac{1}{3}\right) \geq k_n\right) + 2.5 P\left(\text{Bin}\left(n, \frac{3}{4}\right) < k_n\right).
\end{aligned}$$

Suponha agora que há um custo  $c(n)$  por essas  $n$  observações e que a função de perda é dada por  $L(d, \theta, n) = L(d, \theta) + c(n)$ . Essa função custo  $c : \mathbb{N} \rightarrow \mathbb{R}$  pode depender de questões além das financeiras, como, por exemplo, o tempo de coleta da amostra ou algum risco aos envolvidos no experimento. Considere, por simplicidade, uma função de custo linear  $c(n) = 0.02n$ , de modo que o risco é

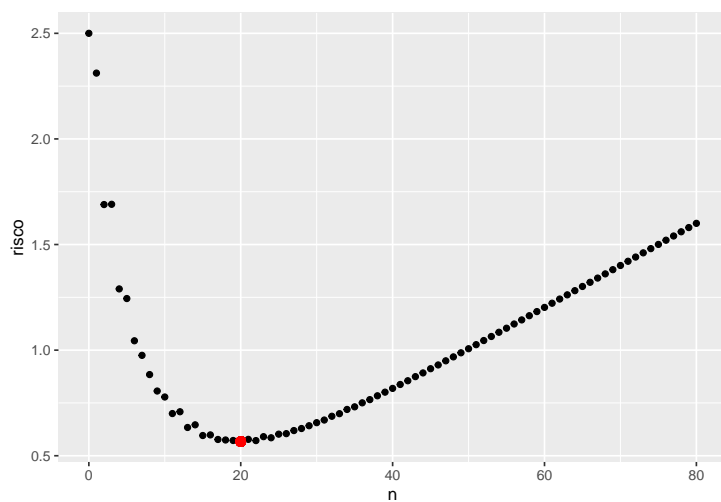
$$\rho_n(P) = \rho^*(P) + c(n) = 5 P\left(\text{Bin}\left(n, \frac{1}{3}\right) \geq k_n\right) + 2.5 P\left(\text{Bin}\left(n, \frac{3}{4}\right) < k_n\right) + 0.02n.$$

A seguir é apresentado um gráfico desse risco para alguns valores de  $n$  e é possível notar que o tamanho amostral ótimo é  $n^* = 20$ .

```

tibble(n=seq(0,80),
       kn=-n*log(3/8,6)+log(2,6),
       risco=5*(1-pbinom(kn,n,1/3))+2.5*pbinom(kn,n,3/4)+0.02*n) %>%
  ggplot() +
  geom_point(aes(x=n,y=risco)) +
  geom_point(aes(x=n[which.min(risco)],y=min(risco)),col="red",cex=2.5)

```



No exemplo anterior, foi apresentado uma maneira de considerar custos das observações e obter um tamanho amostral ótimo para determinado problema de decisão. Quando o custo está relacionado somente a quantidades monetárias, funções de custo lineares não são as mais adequadas. Para uma discussão bastante didática sobre esse problema, veja o artigo *O Paradoxo de São Petersburgo* (Peixoto, C. M. e Wechsler, S.) no Boletim da ISBrA, 6(2).



## Chapter 6

# Testes de Hipóteses

### 6.1 Conceitos Básicos

Uma **hipótese estatística** é uma afirmação sobre o parâmetro  $\theta$  (ou a família  $\mathcal{P}$ ). No caso usual, tem-se duas hipóteses:  $H_0 : \theta \in \Theta_0$ , chamada de **hipótese nula**, e  $H_1 : \theta \in \Theta_1 = \Theta_0^c$ , chamada **hipótese alternativa**.

Um **teste de hipótese** é uma regra de decisão  $\varphi : \mathfrak{X} \rightarrow \{0, 1\}$ , onde  $\varphi(x) = 1$  significa rejeitar  $H_0$  (aceitar  $H_1$ ) e  $\varphi(x) = 0$ , não rejeitar (aceitar)  $H_0$ .

Se rejeita-se  $H_0$  (aceita-se  $H_1$ ) quando  $H_0$  é verdadeira, comete-se um **erro do tipo I**. Por outro lado, se não rejeita-se  $H_0$  (aceita  $H_0$ ) quando  $H_0$  é falso, ocorre um **erro do tipo II**.

O conjunto  $\varphi^{-1}(1) = \{x \in \mathfrak{X} : \varphi(x) = 1\}$  recebe o nome de **região de rejeição** (ou **região crítica**). A **função de poder** do teste  $\varphi$  é  $\pi_\varphi(\theta) = P(\varphi^{-1}(1) | \theta) = P(\text{'Rejeitar } H_0' | \theta)$ .

Dizemos que um teste  $\varphi$  tem **nível de significância**  $\alpha$  se  $\sup_{\theta \in \Theta_0} \pi_\varphi(\theta) \leq \alpha$ . Se  $\alpha = \sup_{\theta \in \Theta_0} \pi_\varphi(\theta)$  dizemos que o teste é de **tamanho**  $\alpha$ .

Uma hipótese é dita **simples** se contém apenas um ponto,  $H : \theta = \theta_0$ . Caso contrário é chamada de **hipótese composta**. No caso em que  $H : \theta \in \Theta_0$  é tal que  $\dim(\Theta_0) < \dim(\Theta)$ , diz-se que  $H$  é uma **hipótese precisa** (“*sharp*”).

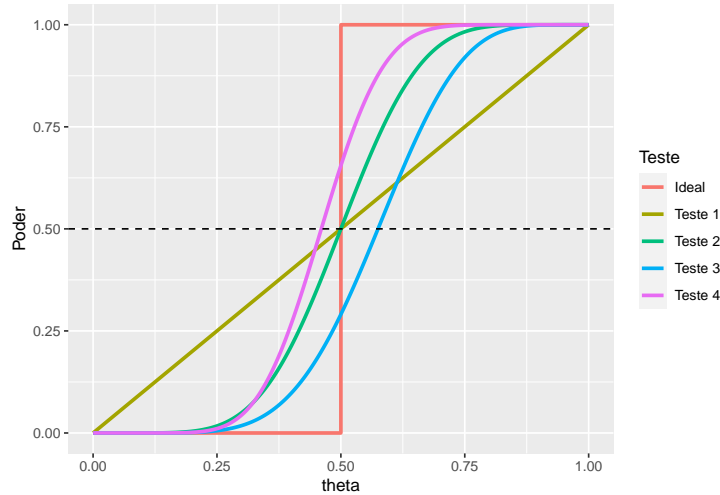
## 6.2 Revisão: Abordagem Frequentista

Um teste de hipótese “ideal” seria aquele que as probabilidades de erros tipo I e tipo II são iguais a zero, isto é,  $\pi_\varphi(\theta) = 0, \forall \theta \in \Theta_0$ , e  $\pi_\varphi(\theta) = 1, \forall \theta \in \Theta_1$ . Contudo, não é possível obter tais testes em geral.

A solução usual é fixar um nível de significância  $\alpha$  e considerar apenas a classe de teste de nível  $\alpha$ , isto é, testes tais que  $\sup_{\theta \in \Theta_0} \pi_\varphi(\theta) \leq \alpha$ . Os testes ótimos

sob o ponto de vista frequentista são aqueles na classe de testes de nível  $\alpha$  que tenha maior função poder  $\pi_\varphi(\theta)$  para  $\theta \in \Theta_1$ . Um teste que satisfaz isso é chamado de **Teste Uniformemente Mais Poderoso (UMP)** mas testes com essa propriedade também só podem ser obtidos em casos específicos.

**Exemplo.** Considere que  $\Theta = [0, 1]$  e deseja-se testar  $H_0 : \theta \leq 0.5$  contra  $H_1 : \theta > 0.5$ . O gráfico a seguir ilustra as funções poder dos quatro testes disponíveis para esse problema. Supondo (apenas para fins didáticos) que  $\alpha = 0.5$ , temos que o teste UMP de nível  $\alpha$  é o teste 2. O teste 4, apesar de ser mais poderoso, não é um teste de nível  $\alpha$ .



Uma das situações onde é possível obter o teste mais poderoso é o caso que as hipóteses nula e alternativa são simples, isto é,  $H_0 : \theta = \theta_0$  contra  $H_1 : \theta = \theta_1$ .

Nesses casos, pode-se considerar o **Lema de Neyman-Pearson**, que afirma que o **teste mais poderoso** é dado por

$$\varphi^*(x) = \begin{cases} 1, & \frac{f(x|\theta_0)}{f(x|\theta_1)} \leq k \\ 0, & \text{c.c.} \end{cases}.$$

Além disso, o teste  $\varphi^*$  minimiza a combinação linear das probabilidades de erro  $a\alpha + b\beta$  com  $b/a$ ,  $\alpha = \pi_{\varphi^*}(\theta_0) = P(\text{'erro tipo I'})$  e  $\beta = 1 - \pi_{\varphi^*}(\theta_1) = P(\text{'erro tipo II'})$ .

Como dito anteriormente, usualmente é fixado um nível  $\alpha$  e isso permite encontrar o valor de  $k$  de modo que o teste construído a partir do lema é o teste mais poderoso de nível  $\alpha$ . Assim,

$$\alpha = \pi_{\varphi^*}(\theta_0) = P(X \in \varphi^{-1}(\{1\})|\theta_0) = P\left(\left\{x : \frac{f(x|\theta_0)}{f(x|\theta_1)} \leq k\right\} \middle| \theta_0\right).$$

Suponha que foi observado  $X = x_o$ , é possível calcular o *nível descritivo* (ou *p-value*) da seguinte forma:

$$p(x_o) = P\left(\left\{x : \frac{f(x|\theta_0)}{f(x|\theta_1)} \leq \frac{f(x_o|\theta_0)}{f(x_o|\theta_1)}\right\} \middle| \theta_0\right)$$

É possível obter testes UMP em alguns casos particulares. Em especial, nos casos em que a família de distribuições para  $X$  condicional a  $\theta$  possui a propriedade de *razão de verossimilhanças monótona*, é possível construir testes UMP para hipóteses do tipo  $H_1 : \theta \leq \theta_0$  contra  $H_1 : \theta > \theta_0$ . Para problemas onde as hipóteses são da forma  $H_1 : \theta = \theta_0$  contra  $H_1 : \theta \neq \theta_0$ , bastante comuns no dia a dia de um estatístico, não existe teste UMP, em geral.

Nos casos em que não existe um teste UMP, o teste mais utilizado sob a abordagem frequentista certamente é o **teste da razão de verossimilhança generalizada** (RVG). Primeiramente, considere a *razão de verossimilhanças generalizada*, dada por

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} f(x|\theta)}{\sup_{\theta \in \Theta} f(x|\theta)}.$$

Note que  $0 \leq \lambda(x) \leq 1$ ,  $\forall x \in \mathfrak{X}$  e  $\forall \Theta_0 \subseteq \Theta$ . Um *teste RVG* é qualquer teste da forma

$$\varphi_{RV}(x) = \begin{cases} 1, & \lambda(x) \leq k \\ 0, & \text{c.c.} \end{cases}.$$

Novamente,  $k$  pode ser escolhido de modo que o teste resultante seja de nível  $\alpha$ , isto é,  $\sup_{\theta \in \Theta_0} P(\lambda(x) \leq k | \theta) \leq \alpha$ . Do mesmo modo, se foi observado  $X = x_o$ , um  $p$ -value é  $p(x_o) = \sup_{\theta \in \Theta_0} P(\{x : \lambda(x) \leq \lambda(x_o)\} | \theta)$ . Por fim, em casos onde é difícil fazer os cálculos de forma exata e o tamanho amostral  $n$  é razoavelmente grande, é possível usar a distribuição assintótica da RVG  $-2 \log \lambda(x) \xrightarrow{\mathcal{D}} \chi_d^2$ , onde  $d = \dim(\Theta) - \dim(\Theta_0)$ .

### 6.3 Abordagem Bayesiana (via Teoria da Decisão)

Sob a abordagem de teoria da decisão, podemos ver um teste de hipóteses como um problema de decisão onde temos duas possíveis decisões  $\mathcal{D} = \{d_0, d_1\}$ , onde  $d_0$  é decidir por  $H_0 : \theta \in \Theta_0$  e  $d_1$  é decidir por  $H_1 : \theta \in \Theta_1$ , com  $\Theta = \Theta_0 \cup \Theta_1$ . Um teste de hipóteses nesse contexto é uma função de decisão  $\varphi : \mathfrak{X} \rightarrow \{0, 1\}$ , de modo que quando  $\varphi(x) = i$ , decide-se por  $d_i$ ,  $i \in \{0, 1\}$ .

Primeiramente, considere o contexto apresentado no Lema de Neyman-Pearson, onde  $\Theta = \{\theta_0, \theta_1\}$  e deseja-se testar  $H_0 : \theta = \theta_0$  contra  $H_1 : \theta = \theta_1$ . Considere que, a priori,  $f(\theta_0) = \pi$ , a função de verossimilhança é  $f(x|\theta)$  e a função de perda apresentada na tabela a seguir.

$L(d, \theta)$

$\theta_0$

$\theta_1$

$d_0$

0

$b$

$d_1$

$a$

0

Então, o risco de uma função de decisão  $\varphi$  é

$$\begin{aligned} \rho(\varphi, P) &= E[L(\varphi(X), \theta)] = \pi E[L(\varphi(X), \theta) | \theta_0] + (1 - \pi) E[L(\varphi(X), \theta) | \theta_1] \\ &= a \pi P(\varphi(x) = 1 | \theta_0) + b (1 - \pi) P(\varphi(x) = 0 | \theta_1) = a \pi \alpha_\varphi + b (1 - \pi) \beta_\varphi \end{aligned}$$

Como o risco acima é uma combinação linear das probabilidades dos erro tipo I e tipo II, podemos aplicar o Lema de Neyman-Pearson e obter a função de decisão  $\varphi^*$  que minimiza o risco

$$\varphi^*(x) = \begin{cases} 1, & \frac{f(x|\theta_0)}{f(x|\theta_1)} \leq \frac{b(1-\pi)}{a\pi} \\ 0, & \text{c.c.} \end{cases}.$$

Esse resultado é apresentado por DeGroot (1986) e é uma espécie de *Lema de Neyman-Pearson Generalizado*.

A solução para esse mesmo problema pode também ser obtida usando a *forma extensiva*. O risco posterior para as suas decisões é  $r_x(d_0) = bf(\theta_1|x)$  e  $r_x(d_1) = af(\theta_0|x)$ , de modo que rejeitamos  $H_0$  (decidimos por  $d_1$  ou  $\varphi(x) = 1$ ) se

$$r_x(d_1) \leq r_x(d_0) \iff af(\theta_0|x) \leq bf(\theta_1|x) \iff af(\theta_0|x) \leq b[1 - f(\theta_0|x)] \iff f(\theta_0|x) \leq \frac{b}{a+b}$$

De modo que o teste de Bayes também pode ser apresentado como

$$\varphi^*(x) = \begin{cases} 1, & f(\theta_0|x) \leq \frac{b}{a+b} \\ 0, & \text{c.c.} \end{cases}.$$

A interpretação nesse caso é mais direta, a hipótese é rejeitada se sua probabilidade posterior é “pequena”. Como vimos, essas soluções são equivalentes. De fato,

$$\begin{aligned} \bullet \quad f(\theta_0|x) &= \frac{f(\theta_0)f(x|\theta_0)}{f(x)} = \pi \frac{f(x|\theta_0)}{f(x)}; \\ \bullet \quad f(\theta_1|x) &= \frac{f(\theta_1)f(x|\theta_1)}{f(x)} = (1-\pi) \frac{f(x|\theta_1)}{f(x)}. \end{aligned}$$

Assim,

$$r_x(d_1) \leq r_x(d_0) \iff a\pi \frac{f(x|\theta_0)}{f(x)} \leq b(1-\pi) \frac{f(x|\theta_1)}{f(x)} \iff \frac{f(x|\theta_0)}{f(x|\theta_1)} \leq \frac{b(1-\pi)}{a\pi}.$$

Considere agora um caso mais geral, onde  $\Theta = \Theta_0 \dot{\cup} \Theta_1$  e deseja-se testar  $H_0 : \theta \in \Theta_0$  contra  $H_1 : \theta \in \Theta_1$ . Considere também a função de perda mais geral apresentada a seguir, com  $a_0 \leq a_1$  e  $b_0 \leq b_1$ .

$$L(d, \theta)$$

$$\Theta_0$$

$$\Theta_1$$

$d_0$  $a_0$  $b_1$  $d_1$  $a_1$  $b_0$ 

O risco posterior de cada uma das decisões é

$$\begin{aligned} r_x(d_0, \theta) &= a_0 P(\theta \in \Theta_0 | x) + b_1 P(\Theta_1 | x) = a_0 P(\theta \in \Theta_0 | x) + b_1 [1 - P(\Theta_0 | x)] \\ &= a_0 P(\theta \in \Theta_0 | x) + b_1 - b_1 P(\Theta_0 | x), \end{aligned}$$

$$r_x(d_1, \theta) = a_1 P(\theta \in \Theta_0 | x) + b_0 - b_0 P(\Theta_0 | x),$$

De modo que rejeita-se  $H_0$ ,  $\varphi(x) = 1$ , se

$$\begin{aligned} r_x(d_1, P) \leq r_x(d_0, P) &\Leftrightarrow (a_1 - b_0)P(\Theta_0 | x) + b_0 \leq (a_0 - b_1)P(\Theta_0 | x) + b_1 \Leftrightarrow \\ P(\Theta_0 | x) &\leq \frac{(b_1 - b_0)}{(a_1 - a_0) + (b_1 - b_0)} \end{aligned}$$

Assim, o teste de bayes nesse caso é

$$\varphi(x) = \begin{cases} 1, & P(\Theta_0 | x) \leq \frac{(b_1 - b_0)}{(a_1 - a_0) + (b_1 - b_0)} \\ 0, & c.c. \end{cases}.$$

## 6.4 Probabilidade Posterior de $H_0$

**Resultado.** Seja  $\Theta = \Theta_0 \dot{\cup} \Theta_1$  e suponha que deseja-se testar  $H_0 : \theta \in \Theta_0$  contra  $H_1 : \theta \in \Theta_1$  considerando a função de perda a seguir, com  $a_0 \leq a_1$  e  $b_0 \leq b_1$ .

 $L(d, \theta)$  $\Theta_0$  $\Theta_1$  $d_0$  $a_0$  $b_1$  $d_1$  $a_1$

$b_0$

Então, o teste de bayes é

$$\varphi(x) = \begin{cases} 1, & P(\Theta_0|x) \leq \frac{(b_1 - b_0)}{(a_1 - a_0) + (b_1 - b_0)} \\ 0, & c.c. \end{cases}.$$

**Exemplo 1.** Considere  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim Ber(\theta)$  com  $\Theta = \{1/2, 3/4\}$ . Suponha que, a priori,  $f(\theta = 1/2) = f(\theta = 3/4) = 1/2$  e deseja-se testar  $H_0 : \theta = 1/2$  contra  $H_1 : \theta = 3/4$ . Tem-se que  $T(X) = \sum X_i \mid \theta \sim Bin(n, \theta)$  é uma estatística suficiente para  $\theta$ . Então,

$$\begin{aligned} P(\theta = 1/2 | T(X) = t) &= \frac{f(t|\theta = 1/2)f(\theta = 1/2)}{\sum_{\theta \in \{1/2, 3/4\}} f(t|\theta)f(\theta)} = \frac{\binom{n}{t} \left(\frac{1}{2}\right)^n}{\binom{n}{t} \left(\frac{1}{2}\right)^n + \binom{n}{t} \left(\frac{3}{4}\right)^t \left(\frac{1}{4}\right)^{n-t}} \\ &= \frac{1}{1 + \frac{3^t}{2^n}}. \end{aligned}$$

Considere a função de perda  $L(d, \theta) = a_0 \mathbb{I}(d_0, \Theta_0) + b_1 \mathbb{I}(d_0, \Theta_1) + a_1 \mathbb{I}(d_1, \Theta_0) + b_0 \mathbb{I}(d_1, \Theta_1)$  como no resultado anterior. Então, rejeita-se  $H_0$  se  $P(\theta \in \Theta_0|x) < K$ , com  $K = \frac{b_1 - b_0}{(a_1 - a_0) + (b_1 - b_0)}$ . Assim,

$$\begin{aligned} P(\theta = 1/2 | T = t) \leq K &\iff \frac{1}{1 + \frac{3^t}{2^n}} \leq K \iff 1 + \frac{3^t}{2^n} \geq \frac{1}{K} \iff 3^t \geq \\ 2^n \left( \frac{1}{K} - 1 \right) &\iff t \geq n \log_3(2) + \log_3 \left( \frac{1 - K}{K} \right) \iff t \geq n \log_3(2) + \\ \log_3 \left( \frac{a_1 - a_0}{b_1 - b_0} \right). \end{aligned}$$

Tomando  $a_1 = b_1 = 1$  e  $a_0 = b_0 = 0$ , rejeita  $H$  se

$$\sum X_i \geq n \log_3(2) + \log_3(1) \iff \sum X_i \geq n \log_3(2) \implies \bar{X} \geq \log_3(2) \approx 0,631.$$

O teste de Bayes é

$$\begin{aligned} \varphi(x) &= \begin{cases} 1, & f(\theta = 1/2|x) \leq \frac{(b_1 - b_0)}{(a_1 - a_0) + (b_1 - b_0)} = \frac{1}{2} \\ 0, & c.c. \end{cases} \implies \varphi(x) = \\ \begin{cases} 1, & \bar{X} \geq \log_3(2) \\ 0, & c.c. \end{cases}. \end{aligned}$$

**Exemplo 2:**  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim N(\theta, \sigma_0^2)$  com  $\sigma_0^2$  conhecido. Suponha que, a priori,  $\theta \sim N(m, v^2)$  e  $\bar{X}$  é estatística suficiente para  $\theta$  com  $\bar{X}|\theta \sim (\theta, \sigma_0^2/n)$ , de modo que  $\theta|X \sim N\left(\frac{\sigma_0^2 m + nv^2 \bar{x}}{\sigma_0^2 + nv^2}, \frac{\sigma_0^2 v^2}{\sigma_0^2 + nv^2}\right)$ . Suponha ainda que o objetivo é testar  $H_0 : \theta \leq \theta_0$  contra  $H_1 : \theta > \theta_0$ .

Utilizando novamente o resultado anterior, temos

$$\begin{aligned} P(\theta \in \Theta_0|x) &= P(\theta \leq \theta_0|x) = P\left(Z \leq \frac{\theta_0 - \frac{\sigma_0^2 m + nv^2 \bar{x}}{\sigma_0^2 + nv^2}}{\sqrt{\frac{\sigma_0^2 v^2}{\sigma_0^2 + nv^2}}} \mid \bar{x}\right) = \Phi\left(\frac{\theta_0 - \frac{\sigma_0^2 m + nv^2 \bar{x}}{\sigma_0^2 + nv^2}}{\sqrt{\frac{\sigma_0^2 v^2}{\sigma_0^2 + nv^2}}}\right) \\ &= \Phi\left(\frac{(\sigma_0^2 + nv^2)\theta_0 - \sigma_0^2 m - nv^2 \bar{x}}{\sigma_0 v \sqrt{\sigma_0^2 + nv^2}}\right), \end{aligned}$$

e deve-se rejeitar  $H_0$  se

$$\begin{aligned} P(\theta \in \Theta_0|x) &\leq \frac{(b_1 - b_0)}{(a_1 - a_0) + (b_1 - b_0)} = K \iff \Phi\left(\frac{(\sigma_0^2 + nv^2)\theta_0 - \sigma_0^2 m - nv^2 \bar{x}}{\sigma_0 v \sqrt{\sigma_0^2 + nv^2}}\right) \leq \\ K &\iff \bar{x} \geq \frac{(\sigma_0^2 + nv^2)\theta_0 - \sigma_0^2 m}{nv^2} - \Phi^{-1}(K) \frac{\sigma_0 \sqrt{\sigma_0^2 + nv^2}}{nv} \iff \bar{x} \geq \frac{\sigma_0^2(\theta_0 - m) + nv^2\theta_0}{nv^2} - \\ &\quad \Phi^{-1}(K) \frac{\sigma_0 \sqrt{\sigma_0^2 + nv^2}}{nv}. \end{aligned}$$

Se  $a_0 = b_0 = 0$  e  $a_1 = b_1 = 1$ , então  $\Phi^{-1}(K = 1/2) = 0$  e rejeita-se  $H_0$  se

$$\bar{x} \geq \frac{\sigma_0^2(\theta_0 - m) + nv^2\theta_0}{nv^2} \xrightarrow{n \uparrow \infty} \theta_0.$$

## 6.5 Fator de Bayes

Voltando ao resultado, tem-se que rejeita-se  $H_0$  se

$$\begin{aligned} r_x(d_0) \geq r_x(d_1) &\iff a_0 P(\Theta_0|x) + b_1 P(\Theta_1|x) \geq a_1 P(\Theta_0|x) + b_0 P(\Theta_1|x) \\ &\iff \frac{P(\Theta_0|x)}{P(\Theta_1|x)} \leq \frac{b_1 - b_0}{a_1 - a_0} \\ &\iff BF(x) = \frac{\frac{P(\Theta_0|x)}{P(\Theta_1|x)}}{\frac{P(\Theta_0)}{P(\Theta_1)}} = \frac{\frac{P(\Theta_0|x)}{P(\Theta_0)}}{\frac{P(\Theta_1|x)}{P(\Theta_1)}} = \frac{f(x|\Theta_0)}{f(x|\Theta_1)} \leq \frac{(b_1 - b_0) P(\Theta_1)}{(a_1 - a_0) P(\Theta_0)}, \end{aligned}$$

onde  $BF$  é o **Fator de Bayes**, frequentemente utilizado na literatura bayesiana para testar hipóteses. Ele pode ser visto como uma razão de chances que representa o aumento na chance da hipótese nula ser mais plausível que a hipótese



alternativa após observar os dados em relação a sua opinião a priori. O  $BF$  também pode ser reescrito como

$$\begin{aligned} BF(x) &= \frac{f_0(x)}{f_1(x)} = \frac{f(x|\Theta_0)}{f(x|\Theta_1)} = \frac{\int_{\Theta} f(x|\theta)f(\theta|\Theta_0)d\theta}{\int_{\Theta} f(x|\theta)f(\theta|\Theta_1)d\theta} = \frac{\int_{\Theta_0} f(x|\theta)dP_0(\theta)}{\int_{\Theta_1} f(x|\theta)dP_1(\theta)} \\ &= \frac{E[f(x|\theta)|\theta \in \Theta_0]}{E[f(x|\theta)|\theta \in \Theta_1]}. \end{aligned}$$

**No exemplo 1.** Lembrando que, a priori,  $P(\theta = 1/2) = P(\theta = 3/4) = 1/2$  e considerando novamente  $a_0 = b_0 = 0$  e  $a_1 = b_1 = 1$ , temos que devemos rejeitar  $H_0$  se  $BF(x) < \frac{(b_1-b_0)}{(a_1-a_0)} \frac{P(\theta=1/2)}{P(\theta=3/4)} = 1$ . Então,

$$BF(x) = \frac{P(\theta = 1/2|x)}{P(\theta = 3/4|x)} \frac{1/2}{1/2} = \frac{\frac{1}{1+3^{t/2^n}}}{\frac{3^{t/2^n}}{1+3^{t/2^n}}} = \frac{2^n}{3^t} \leq 1 \iff \bar{x} \geq \log_3(2),$$

de modo que a decisão baseada no fator de Bayes concorda com o resultado baseado na probabilidade a posteriori da hipótese.

**No exemplo 2.**  $\theta|X \sim N\left(\frac{\sigma_0^2}{\sigma_0^2 + nv^2} \bar{x}, \frac{\sigma_0^2 v^2}{\sigma_0^2 + nv^2}\right)$  e o objetivo é testar  $H_0 : \theta \leq \theta_0$  contra  $H_1 : \theta > \theta_0$ . A probabilidade a posteriori da hipótese  $H_0$  é

$$P(\theta \in \Theta_0|x) = \Phi\left(\frac{(\sigma_0^2 + nv^2)\theta_0 - \sigma_0^2 \bar{x}}{\sigma_0 v \sqrt{\sigma_0^2 + nv^2}}\right),$$

e, a priori,  $P(\theta \in \Theta_0) = P(\theta \leq \theta_0) = \Phi\left(\frac{\theta_0 - m}{v}\right)$ , de modo que o fator de Bayes nesse caso é

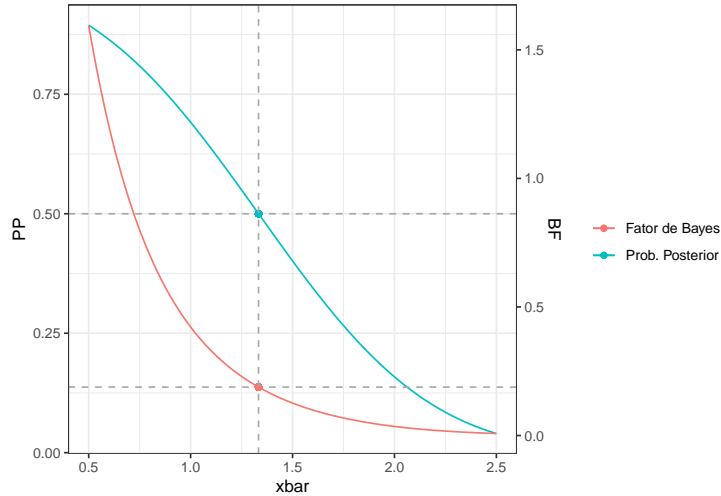
$$BF(x) = \frac{P(\Theta_0|x) P(\Theta_1)}{P(\Theta_1|x) P(\Theta_0)} = \frac{\Phi\left(\frac{(\sigma_0^2 + nv^2)\theta_0 - \sigma_0^2 \bar{x}}{\sigma_0 v \sqrt{\sigma_0^2 + nv^2}}\right)}{\left[1 - \Phi\left(\frac{(\sigma_0^2 + nv^2)\theta_0 - \sigma_0^2 \bar{x}}{\sigma_0 v \sqrt{\sigma_0^2 + nv^2}}\right)\right]} \frac{\left[1 - \Phi\left(\frac{\theta_0 - m}{v}\right)\right]}{\Phi\left(\frac{\theta_0 - m}{v}\right)}.$$

```
m=0; v2=1 # Média e Variância da Priori
sigma02=1 # Variância Populacional (conhecido)
n=3 # tamanho amostral
theta0=1 # H0: theta <= theta0
a0=0; b0=0; a1=1; b1=1 # Função de Perda
K1=(b1-b0)/(a1-a0+b1-b0) # corte Prob. Posterior
K2=((b1-b0)*(1-pnorm((theta0-m)/sqrt(v2)))) / ((a1-a0)*pnorm((theta0-m)/sqrt(v2))) #corte Fator de Bayes
K3=((sigma02+n*v2)*theta0-sigma02*m)/(n*v2) - qnorm(K1)*sqrt(sigma02*(sigma02+n*v2))/(n*sqrt(v2))
```

```

# Probabilidade a Posteriori de H0 (como função de Xbar)
postH = function(xbar){
  pnorm(((sigma02 + n*v2)*theta0 - sigma02*m - n*v2*xbar)/ sqrt(sigma02*v2*(sigma02+n*v2)))
# Fator de Bayes de H0 (como função de Xbar)
bf = function(xbar){
  (postH(xbar)*(1-pnorm((theta0-m)/sqrt(v2))))/ ((1-postH(xbar))*pnorm((theta0-m)/sqrt(v2)))
xbar=seq(0.5,2.5,0.001)
PP=postH(xbar)
BF=bf(xbar)
FS=(max(PP)-min(PP))/(max(BF)-min(BF)) # var. aux. para transformação dos eixos
tibble(xbar,PP,BF) %>%
  ggplot() +
    geom_line(aes(x=xbar,y=PP,colour="Prob. Posterior")) +
    geom_line(aes(x=xbar,y=((BF-min(BF))*FS+min(PP)),colour="Fator de Bayes"))+
    scale_y_continuous(sec.axis = sec_axis(~./FS-min(PP)/FS+min(BF), name = "BF"))+
    geom_hline(aes(yintercept=K1),lty=2, col="darkgrey") +
    geom_point(aes(x=K3,y=K1,colour="Prob. Posterior")) +
    geom_hline(aes(yintercept=((K2-min(BF))*FS+min(PP))),lty=2, col="darkgrey") +
    geom_point(aes(x=K3,y=((K2-min(BF))*FS+min(PP)),colour="Fator de Bayes")) +
    geom_vline(aes(xintercept=K3),lty=2, col="darkgrey") +
    theme_bw() + labs(colour = "")

```



Nesse exemplo, é possível ver que tanto o Fator de Bayes quanto a probabilidade posterior da hipótese nula “ordenam” o espaço amostral, representado aqui pela estatística suficiente,  $\bar{X}$ . Deste modo, quanto menores os valores dessas estatísticas de teste, mais desfavorável é o ponto amostral para a hipótese nula. Como visto anteriormente, as regras de decisão baseadas nessas estatísticas são equivalentes e, portanto, a ordenação do espaço amostral é a mesma.

**Problema:** Suponha agora que, nesse mesmo exemplo, deseja-se testar  $H_0 : \theta = 0$  contra  $H_1 : \theta \neq 0$ . Como a posteriori é Normal, temos que  $P(\theta \in \Theta_0|x) = P(\theta = 0|x) = 0$ ,  $\forall x \in \mathfrak{X}$  e, desta forma, a hipótese nula  $H_0$  sempre é rejeitada.

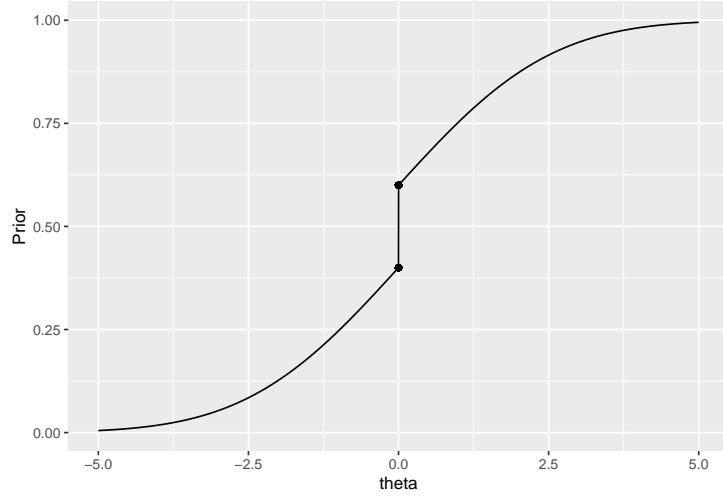
De fato, para qualquer cenário em que a hipótese  $H_0$  tem medida nula a priori,  $P(\theta \in \Theta_0) = 0$ , tem-se que, a posteriori,  $P(\theta \in \Theta_0|x) = 0$ . Isso ocorre frequentemente nos casos em que  $H_0$  é uma *hipótese precisa*, isto é,  $\dim(\Theta_0) < \dim(\Theta)$ . Neste cenário, é necessário definir procedimentos alternativos para testar hipóteses.

## 6.6 Teste de Jeffreys

O teste de Jeffreys (1961?) consiste em atribuir uma probabilidade positiva para o conjunto que define a hipótese nula,  $p_0 = P(\theta \in \Theta_0) > 0$ .

*Exemplo 2.* Suponha que deseja-se testar  $H_0 : \theta = 0$  contra  $H_1 : \theta \neq 0$ . Suponha que sua opinião a priori é  $\theta \sim \text{Normal}(0, 2)$ . Contudo, já foi visto que  $P(\theta = 0|x) = 0$ ,  $\forall x \in \mathfrak{X}$ . Deste modo, você opta por atribuir uma probabilidade positiva  $p_0 = 0.2$  para o ponto  $\theta = 0$ , ou seja, você vai considerar uma distribuição mista  $f(\theta) = p_0 \mathbb{I}(\theta = 0) + (1 - p_0) f_N(\theta)$ , onde  $f_N$  é a densidade da  $\text{Normal}(0, 1)$ . Sua função de distribuição a priori,  $F(\theta) = p_0 \mathbb{I}(\theta \geq 0) + (1 - p_0) \Phi(\theta/\sqrt{2})$ , está representada no gráfico a seguir.

```
theta=c(seq(-5,-0.001,0.001),seq(0.001,5,0.001))
p=0.2
pprior = function(t){
  p*I(t>=0)+(1-p)*pnorm(t,0,2)*I(t!=0)
}
tibble(theta,Prior=pprior(theta)) %>%
  ggplot()+geom_line(aes(x=theta,y=Prior))+
  geom_point(aes(x=0,y=(1-p)*pnorm(0,0,2)))+
  geom_point(aes(x=0,y=(1-p)*pnorm(0,0,2)+p))
```



**Exercício.** Calcule  $P(\theta = 0|X = x)$ .

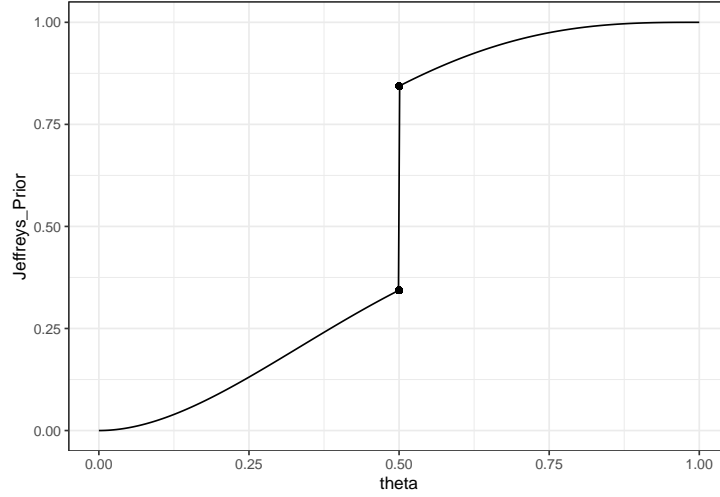
**Exemplo 3.** Seja  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim \text{Ber}(\theta)$  e considere que, a priori,  $\theta \sim \text{Beta}(a, b)$ . Como  $X = \sum X_i$  é estatística suficiente com  $X|\theta \sim \text{Bin}(n, \theta)$ , tem-se que  $\theta|x = \sum x_i \sim \text{Beta}(a + \sum x_i, b + n - \sum x_i)$ .

A distribuição marginal de  $X$  é chamada **distribuição preditiva a priori** e pode ser calculada por

$$\begin{aligned} f(x) &= \int_0^1 f(x, \theta) d\theta = \int_0^1 f(x|\theta) f(\theta) d\theta = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a+x-1} (1-\theta)^{b+n-x-1} d\theta \\ &= \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+x)\Gamma(b+n-x)}{\Gamma(a+b+n)} = \binom{n}{x} \frac{\beta(a+x, b+n-x)}{\beta(a, b)} \mathbb{I}_{\{0, \dots, n\}}(x) \end{aligned}$$

$$\Rightarrow X \sim \text{Beta} - \text{Binomial}(n, a, b).$$

Suponha agora que deseja-se testar  $H_0 : \theta = \theta_0$  contra  $H_1 : \theta \neq \theta_0$ , com  $\theta_0 = 1/2$ , utilizando o teste de Jeffreys. Desta forma, considere  $p_0 = P(\theta = 1/2) = 1/2$  e sua *priori de Jeffreys* é  $f_J(\theta) = p_0 \mathbb{I}(\theta = \theta_0) + (1-p_0) f_\beta(\theta) \mathbb{I}(\theta \neq \theta_0)$ , onde  $f_\beta$  é a densidade da  $\text{Beta}(a, b)$ .



A distribuição preditiva com relação a priori  $f_J$  é

$$f_J(x) = p_0 f(x|\theta_0) \mathbb{I}(\theta = \theta_0) + (1 - p_0) \overbrace{\int_0^1 f(x|\theta) f_\beta(\theta) \mathbb{I}(\theta \neq \theta_0) d\theta}^{f(x)}$$

$$= p_0 \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x} \mathbb{I}(\theta = \theta_0) + (1 - p_0) \binom{n}{x} \frac{\beta(a+x, b+n-x)}{\beta(a, b)} \mathbb{I}(\theta \neq \theta_0),$$

de modo que a distribuição a posteriori é

$$f_J(\theta|x) = \frac{f(x|\theta) f_J(\theta)}{f_J(x)} = \frac{p_0 \binom{n}{x} (1/2)^n}{f_J(x)} \mathbb{I}(\theta = 1/2) + \frac{(1 - p_0) \binom{n}{x} \theta^{a+x-1} (1 - \theta)^{b+n-x-1}}{\beta(a, b) f_J(x)} \mathbb{I}(\theta \neq 1/2).$$

A probabilidade posterior da hipótese  $H_0 : \theta = 1/2$  é

$$p_x = P(\theta = 1/2|x) = \frac{p_0 \binom{n}{x} (1/2)^n}{p_0 \binom{n}{x} (1/2)^n + (1 - p_0) \binom{n}{x} \frac{\beta(a+x, b+n-x)}{\beta(a, b)}} = \frac{1}{1 + \frac{(1 - p_0)}{p_0} \frac{\beta(a+x, b+n-x)}{(1/2)^n \beta(a, b)}}.$$

E, assim, o Fator de Bayes é dado por

$$B_j(x) = \frac{\frac{p_x}{1 - p_0}}{\frac{p_0}{1 - p_0}} = \frac{\frac{1}{\frac{(1 - p_0)}{p_0} \frac{\beta(a+x, b+n-x)}{(1/2)^n \beta(a, b)}}}{\frac{1}{\frac{(1 - p_0)}{p_0} \frac{\beta(a+x, b+n-x)}{(1/2)^n \beta(a, b)}}} = \frac{(1/2)^n \beta(a, b)}{\beta(a+x, b+n-x)}.$$

Note que, nesse caso,  $BF(x)$  não depende da probabilidade a priori  $p_0$  da hipótese  $H_0$ .

```

theta0=1/2
n=6; p=1/2
a=1;b=1
x=seq(0,n)
# Fator de Bayes para cada x
BF=(theta0^x)*((1-theta0)^(n-x))*beta(a,b)/beta(a+x,b+n-x)
# Probabilidade a posteriori para cada x
PP=(1 + (((1-p)*beta(a+x,b+n-x))/(p*(theta0^x)*((1-theta0)^(n-x))*beta(a,b))))^(-1)
tab=t(tibble(BF=round(BF,4),PP=round(PP,4)))
colnames(tab)=x
kable(tab, booktabs=TRUE, escape=FALSE)

```

```

0
1
2
3
4
5
6
BF
0.1094
0.6562
1.6406
2.1875
1.6406
0.6562
0.1094
PP
0.0986
0.3962
0.6213
0.6863
0.6213
0.3962

```

0.0986

Na tabela acima, são calculados  $P(\theta = 1/2|x)$  e  $BF(x)$  para cada  $x$  com  $n = 6$ ,  $p_0 = 1/2$  e os parâmetros da Beta sendo  $a = b = 1$ . Considerando  $a_0 = b_0 = 0$  e  $a_1 = b_1 = 1$ , os valores de corte para a probabilidade a posteriori e o  $BF$  são, respectivamente,  $1/2$  e  $1$ . Desta forma, rejeita-se a hipótese nula para os valores “extremos”  $\{0, 1, 5, 6\}$ .

## 6.7 Hipóteses Precisas

- Probabilidade a posteriori da hipótese  $H_0$ ,  $P(\Theta_0|x)$ .
- Fator de Bayes  $BF(x)$ .
- No caso absolutamente contínuo, quando  $H_0$  é *hipótese precisa*,  $P(\Theta_0|x) = 0$ . Isso faz com que os testes anteriores sempre levem à rejeição de  $H_0$ .
- Primeira alternativa: *teste de Jeffreys*. Problema: a priori deve dar probabilidade positiva à hipótese nula, conduzindo assim a uma priori “artificial” (mista).
- Serão apresentados dois procedimentos alternativos de teste: *FBST* e *P-value*. O primeiro deles foi pensado especificamente para hipóteses precisas ( $\dim(\Theta_0) < \dim(\Theta)$ ) mas ambos podem ser aplicados para hipóteses gerais.

## 6.8 FBST - *Full Bayesian Significance Test*

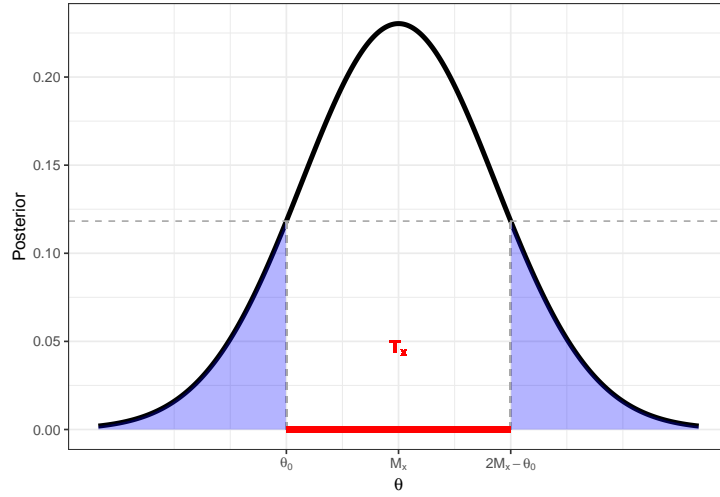
Essa solução foi apresentada por Pereira e Stern em 1999. Suponha que o objetivo é testar  $H_0 : \theta \in \Theta_0$  contra  $H_1 : \theta \in \Theta_1 = \Theta_0^c$ . Seja

$$T_x = \left\{ \theta \in \Theta : f(\theta|x) \geq \sup_{\theta \in \Theta_0} f(\theta|x) \right\} \text{ a região tangente à hipóteses } H_0,$$

formada pelos pontos densidade posterior maior ou igual que qualquer ponto da hipótese nula. Se esse conjunto é “grande” (muito provável), a hipótese nula está em uma região de pouca densidade posterior e deve ser rejeitada. Assim, a *medida de evidência* (de Pereira-Stern) ou *e-value* é definido por  $Ev(\Theta_0, x) = 1 - P(\theta \in T_x|x)$ , e deve-se rejeitar  $H_0$  se o *e-value* for “pequeno”.

**Exemplo.**  $X_1, \dots, X_n$  c.i.i.d.  $N(\theta, \sigma_0^2)$ , com  $\sigma_0^2$  conhecido. Novamente, considere  $\theta \sim N(m, v^2)$ , de modo que  $\theta|x \sim N\left(\frac{\sigma_0^2 m + nv^2 \bar{x}}{\sigma_0^2 + nv^2}, \frac{\sigma_0^2 v^2}{\sigma_0^2 + nv^2}\right)$  e denote a média e a variância da posteriori por  $M_x$  e  $V_x$ , respectivamente. Suponha que o interesse é testar  $H_0 : \theta = \theta_0$  contra  $H_1 : \theta \neq \theta_0$ .

Sem perda de generalidade, suponha que  $M_x \geq \theta_0$ . Então, como a normal é simétrica em torno de  $M_x$ , a região tangente é da forma  $T_x = [\theta_0, 2M_x - \theta_0]$ .



Note que quanto mais próximo  $M_x$  está de  $\theta_0$ , menor a região  $T_x$  e, portanto, maior o valor da evidência em favor de  $H_0$ . O valor da evidência pode ser calculado por

$$\begin{aligned} Ev(\Theta_0, x) &= 1 - P(\theta_0 \leq \theta \leq 2M - \theta_0 | x) = 1 - P\left(\frac{\theta_0 - M}{\sqrt{V}} \leq Z \leq \frac{2M - \theta_0 - M}{\sqrt{V}} | x\right) = \\ &= 2\Phi\left(-\frac{|\theta_0 - M|}{\sqrt{V}}\right) = 2\Phi\left(-\frac{\left|\frac{\sigma_0^2 m + nv^2 \bar{x}}{\sigma_0^2 + nv^2} - \theta_0\right|}{\frac{\sigma_0 v}{\sqrt{\sigma_0^2 + nv^2}}}\right) = 2\Phi\left(-\frac{\sqrt{\sigma_0^2 + nv^2}}{\sigma_0 v} \frac{|\sigma_0^2(m - \theta_0) + nv^2(\bar{x} - \theta_0)|}{\sqrt{\sigma_0^2 + nv^2}}\right) = \\ &= 2\Phi\left(-\frac{1}{\sqrt{\sigma_0^2 + nv^2}} \left|\frac{\sigma_0}{v}(m - \theta_0) + \frac{\sqrt{nv}}{\sigma_0}(\bar{x} - \theta_0)\right|\right) = 2\Phi\left(-\frac{1}{\sqrt{\sigma_0^2 + nv^2}} \left|\frac{(m - \theta_0)}{v/\sigma_0} + \sqrt{nv} \frac{(\bar{x} - \theta_0)}{\sigma_0/\sqrt{n}}\right|\right) \end{aligned}$$

Sob a abordagem frequentista, temos que o  $p$ -value é

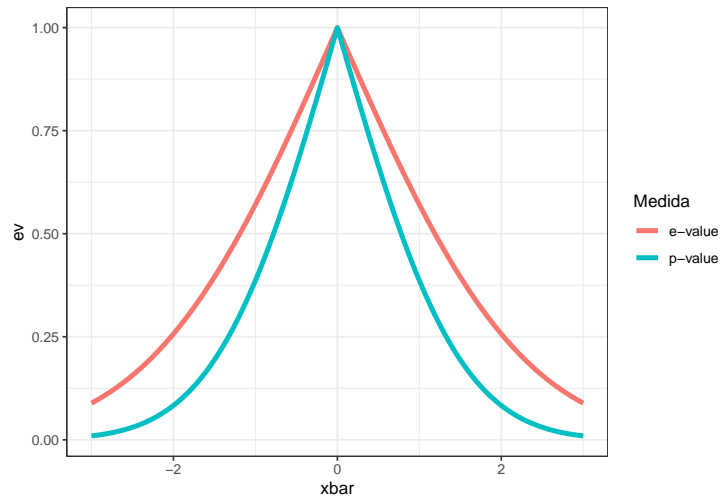
$$\begin{aligned} p(x) &= 1 - P\left(-\frac{|\bar{X} - \theta_0|}{\sigma_0/\sqrt{n}} \leq Z \leq \frac{|\bar{X} - \theta_0|}{\sigma_0/\sqrt{n}}\right) = 2\Phi\left(-\frac{|\bar{X} - \theta_0|}{\sigma_0/\sqrt{n}}\right) \Leftrightarrow \\ &= -\frac{|\bar{X} - \theta_0|}{\sigma_0/\sqrt{n}} = \Phi^{-1}\left(\frac{p\text{-valor}}{2}\right), \end{aligned}$$



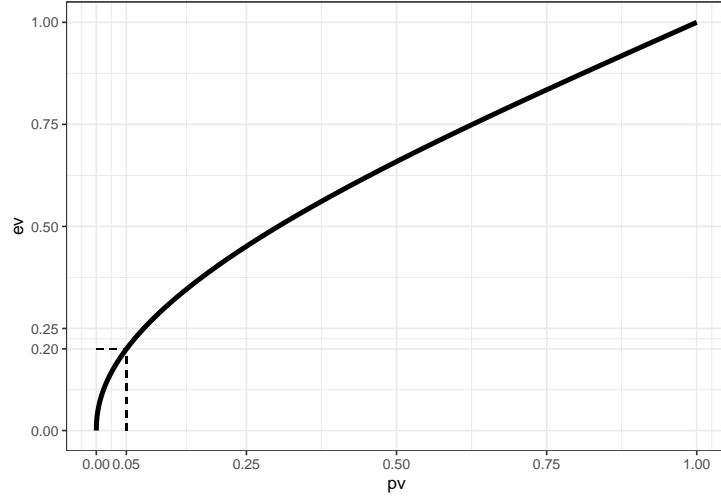
de modo que, nesse exemplo, podemos escrever

$$Ev(\Theta_0, x) = 2\Phi \left( -\frac{1}{\sqrt{\sigma_0^2 + nv^2}} \left| \frac{(m - \theta_0)}{v/\sigma_0} + \sqrt{nv} \Phi \left( \frac{p(x)}{2} \right) \right| \right).$$

A seguir, são apresentados gráficos do *e-value* e do *p-value* como função de  $\bar{x}$  e do *e-value* como função do *p-value* usando da relação anterior.



```
sigma02=4
m=0; v2=1
theta0 = 0
n=3
p=seq(0,1,length.out=5000)
ep = function(p){
  2*pnorm(-abs(sqrt(sigma02/v2)*(m-theta0) + sqrt(n*v2)*qnorm(p/2))/ sqrt(sigma02+n*v2))
}
graf=tibble(pv=p,ev=ep(p)) %>%
  ggplot() +
  geom_line(aes(x=pv,y=ev),lwd=1.5) +
  geom_segment(x=0.05,xend=0.05,y=0,yend=round(ep(0.05),2),lty=2) +
  geom_segment(x=0,xend=0.05,y=round(ep(0.05),2),yend=round(ep(0.05),2),lty=2) +
  scale_y_continuous(breaks=c(0.00,round(ep(0.05),2),0.25,0.50,0.75,1.00)) +
  scale_x_continuous(breaks=c(0.00,0.05,0.25,0.50,0.75,1.00)) +
  theme_bw()
if(knitr::is_latex_output()){
  graf
} else { plotly::ggplotly(graf) }
```



Suponha que um estatístico frequentista decida rejeitar  $H_0$  se o  $p$ -value for menor que  $0.05$ . Para que a decisão baseada no  $e$ -value concorde com o resultado frequentista (para esse particular exemplo), deve-se rejeitar a hipótese se o  $e$ -value for menor que  $0.2$ , aproximadamente. Quando a variância da priori ou o tamanho amostral aumentam, os valores dessas duas medidas se aproximam.

#### Resultados Assintóticos (para esse exemplo)

Suponha que  $H_0 : \theta = \theta_0$  seja falso, isto é, o “verdadeiro” valor do parâmetro é  $\theta^* \neq \theta_0$ . Quando  $n \uparrow \infty$ , pela Lei dos Grandes Números,

$$\bar{X} \xrightarrow{q.c.} \theta^* \implies \frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma_0} \rightarrow +\infty \implies p(X) = 2\Phi\left(-\frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma_0}\right) \rightarrow 0,$$

com probabilidade 1. Por outro lado, sob  $H_0 : \theta = \theta_0$ , pelo Teorema Central do Limite,

$$\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma_0} \xrightarrow{\mathcal{D}} Z \sim N(0, 1) \implies p(X) = 2\Phi\left(-\frac{\sqrt{n}|\bar{X} - \theta_0|}{\sigma_0}\right) \xrightarrow{\mathcal{D}} U = 2\Phi(-|Z|) \sim Unif(0, 1).$$

Esses resultados para o  $p$ -value são bastante conhecidos. No contexto desse exemplo, é possível obter resultados similares para o  $e$ -value. Novamente, considere que  $H_0$  é falso e, sem perda de generalidade,  $\theta = \theta^* > \theta_0$ . Note que

$$\frac{\sigma_0(m - \theta_0)}{v\sqrt{\sigma_0^2 + nv^2}} \rightarrow 0$$

e, pela LGN,

$$\begin{aligned} \bar{X} \xrightarrow{q.c.} \theta^* &\implies (\bar{X} - \theta_0) \rightarrow (\theta^* - \theta_0) > 0 \implies \frac{nv(\bar{X} - \theta_0)}{\sigma_0^2 \sqrt{\sigma_0^2 + nv^2}} \rightarrow +\infty \\ &\implies Ev(\Theta_0, X) = 2\Phi \left( - \left| \frac{\sigma_0(m - \theta_0)}{v\sqrt{\sigma_0^2 + nv^2}} + \frac{nv(\bar{x} - \theta_0)}{\sigma_0 \sqrt{\sigma_0^2 + nv^2}} \right| \right) \rightarrow 2\Phi(-\infty) = 0. \end{aligned}$$

Além disso,  $\frac{v\sqrt{n}}{\sqrt{\sigma_0^2 + nv^2}} \rightarrow 1$  e, sob  $H_0 : \theta = \theta_0$ ,

$$\frac{v\sqrt{n}}{\sqrt{\sigma_0^2 + nv^2}} \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma_0} \xrightarrow{\mathcal{D}} Z \sim N(0, 1),$$

de modo que

$$\begin{aligned} Ev(\Theta_0, X) &= 2\Phi \left( - \left| \frac{\sigma_0(m - \theta_0)}{v\sqrt{\sigma_0^2 + nv^2}} + \frac{nv(\bar{x} - \theta_0)}{\sigma_0 \sqrt{\sigma_0^2 + nv^2}} \right| \right) \rightarrow 2\Phi(-|Z|) \sim \\ &Unif(0, 1). \end{aligned}$$

Esse resultado pode não valer em outros contextos. Por exemplo, quando  $\dim(\Theta_0) \geq 2$ , a distribuição de  $Ev(\Theta_0, X)$  sob  $H_0$  não é  $Unif(0, 1)$ , em geral.

## 6.9 P-value - Nível de Significância Adaptativo

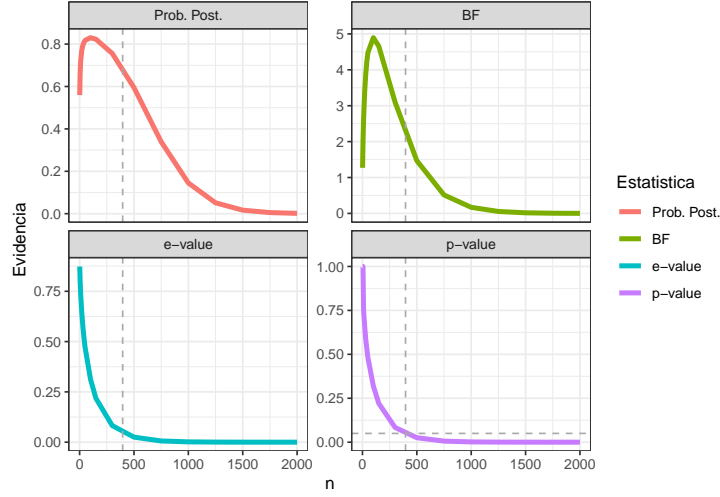
Recentemente, o *p-value* e a utilização do famoso nível  $\alpha = 0.05$  têm sido muito questionados, não apenas na área de testes de hipóteses mas na ciência como um todo. A ideia de fixar um nível de significância é que a probabilidade de *erro tipo I* fica “controlada” e a probabilidade do *erro tipo II* diminui quanto maior o tamanho amostral. Por essa razão, é comum nas áreas de planejamento de experimentos e amostragem, o cálculo do tamanho amostral para um determinado estudo. Infelizmente, na maior parte dos problemas do dia a dia de um estatístico, não há um planejamento cuidadoso ou as amostras disponíveis são “amostras de conveniência”. Simultaneamente, com a “revolução da informação”, a quantidade de dados disponível é cada vez maior. A consequência disso no cenário de testes de hipóteses é que os testes ficam muito poderosos e há uma tendência maior de rejeitar a hipótese nula.

**Exemplo.** Seja  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i|\theta \sim Ber(\theta)$ ,  $\theta \sim Beta(a, b)$  de modo que  $\theta|x = \sum x_i \sim Beta(a + x, b + n - x)$  e suponha que deseja-se testar  $H_0 : \theta = 1/2$ . A seguir, para diferentes tamanhos amostrais  $n$  e supondo que em todos os casos  $\bar{x}_n = 0.55$ , são apresentados os testes para esse caso vistos até aqui: *p-value* do teste RVG, probabilidade posterior e *BF* do teste de Jeffreys e *e-value* do FBST.

```

a=1; b=1
p=0.5
alpha=0.05
theta0=0.5
xbar=0.55
N=c(1,5,10,20,30,40,50,100,150,300,seq(500,2000,250))
p_v=Vectorize(function(n){
  x=n*xbar
  l = c(min(x,n-x),max(x,n-x))
  pbinom(l[1],n,theta0) + 1-pbinom(l[2],n,theta0) })
Nalpha=seq(max(N[(p_v(N)-alpha)>0]),min(N[(p_v(N)-alpha)<0]))
Nalpha=Nalpha[which.min(abs(p_v(Nalpha)-alpha))]
bf=Vectorize(function(n){
  x=n*xbar
  # exp(log(BF))
  exp(x*log(theta0) + (n-x)*log(1-theta0) + lbeta(a,b) - lbeta(a+x,b+n-x)) })
prob_post=Vectorize(function(n){
  x=n*xbar
  l = log(1-p)+lbeta(a+x,b+n-x)-log(p)-x*log(theta0)-(n-x)*log(1-theta0)-lbeta(a,b)
  1/(1+exp(l)) })
e_v=Vectorize(function(n){
  x=n*xbar
  f_Tx=function(t){ dbeta(t,a+x,b+n-x)-dbeta(theta0,a+x,b+n-x) }
  moda=(a+x-1)/(a+b+n-2)
  if(theta0==moda){ return(1) }
  if(theta0<moda){
    Tx=c(theta0,uniroot(f=f_Tx,lower=moda,upper=1)$root)
  }else{
    Tx=c(uniroot(f=f_Tx,lower=0,upper=moda)$root,theta0)
  }
  pbeta(Tx[1],a+x,b+n-x)+1-pbeta(Tx[2],a+x,b+n-x)
})
Dados=tibble(n=rep(N,4), Evidencia=c(p_v(N),prob_post(N),bf(N),e_v(N)),
  Estatistica=factor(rep(c("p-value","Prob. Post.,"BF","e-value"),
    each=length(N)),levels=c("Prob. Post.,"BF","e-value","p-value")),
  corte=c(p_v(Nalpha),rep(NA,4*length(N)-1)))
ggplot(Dados)+
  geom_line(aes(x=n,y=Evidencia, colour=Estatistica),lwd=1.5)+
  geom_vline(xintercept=Nalpha,lty=2,col="darkgrey")+
  facet_wrap(~Estatistica, scales="free_y")+
  geom_hline(data=subset(Dados,Estatistica="p-value"),aes(yintercept=corte), lty=2, col="darkgrey")+
  theme_bw()

```



Note que todas as medidas de suporte da hipótese tendem a zero conforme aumenta o tamanho amostral. Isso indica que se  $|\bar{x}_n - \theta_0| = \varepsilon$  e for fixado um valor de corte para essas medidas que não dependa do tamanho amostral (por exemplo, o  $\alpha = 0.05$  para o  $p$ -value), existe um  $n^*$  tal que  $H_0$  será rejeitado para todo  $n \geq n^*$ . No gráfico, a linha vertical tracejada indica o valor  $n^*$  para o  $p$ -value considerando o corte  $\alpha = 0.05$ .

Como visto anteriormente, o Lema de Neyman-Pearson Generalizado (DeGroot, 1986) garante que os testes Bayesianos (baseados na probabilidade posterior ou no BF) minimizam  $a\alpha + b\beta$ . Baseado nesse resultado, O professor Carlos A. B. Pereira recentemente propôs um novo procedimento de teste para evitar o problema descrito anteriormente.

1. Seja  $BF(x) = \frac{f_0(x)}{f_1(x)} = \frac{\int_{\Theta_0} f(x|\theta) dP_0(\theta)}{\int_{\Theta_1} f(x|\theta) dP_1(\theta)}$ , onde  $P_i$  é a medida de probabilidade a priori para  $\theta$  restrito à hipótese  $H_i$ ,  $i = 0, 1$ ;
2. Defina  $\alpha_n = P\left(\left\{x : BF(x) \leq \frac{b}{a}\right\} \mid \Theta_0\right)$  e  $\beta_n = P\left(\left\{x : BF(x) > \frac{b}{a}\right\} \mid \Theta_1\right)$ ;
3. Suponha que foi observado  $X = x_o$ . O **P-value** é dado por  $P\text{-value}(x_o) = P\left(\{x : BF(x) \leq BF(x_o)\} \mid \Theta_0\right)$ ;
4. O procedimento de teste consiste em rejeitar  $H_0$  se  $P\text{-value}(x) < \alpha_n$ .

**Voltando ao Exemplo.** As distribuições preditivas sob  $H_0 : \theta = \theta_0 = 1/2$  e  $H_1 : \theta \neq 1/2$  são

$$f_0(x) = f(x|\theta_0) = \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x} = \binom{n}{x} (1/2)^n ;$$

$$f_1(x) = \int_{\Theta_{\theta_0}} f(x|\theta) f(\theta) d\theta = \int_0^1 \binom{n}{x} \frac{\theta_0^{a+x-1} (1 - \theta_0)^{b+n-x-1}}{\beta(a, b)} d\theta = \binom{n}{x} \frac{\beta(a+x, b+n-x)}{\beta(a, b)} .$$

Deste modo,

$$BF(x) = \frac{f_0(x)}{f_1(x)} = \frac{\beta(a, b) \theta_0^x (1 - \theta_0)^{n-x}}{\beta(a+x, b+n-x)} = \frac{\beta(a, b) (1/2)^n}{\beta(a+x, b+n-x)},$$

e, assim,

$$\alpha_n = (1/2)^n \sum_{\{x: BF(x) \leq \frac{b}{a}\}} \binom{n}{x},$$

$$\beta_n = \frac{1}{\beta(a, b)} \sum_{\{x: BF(x) > \frac{b}{a}\}} \binom{n}{x} \beta(a+x, b+n-x),$$

$$P\text{-value}(x_0) = (1/2)^n \sum_{\{x: BF(x) \leq BF(x_0)\}} \binom{n}{x}$$

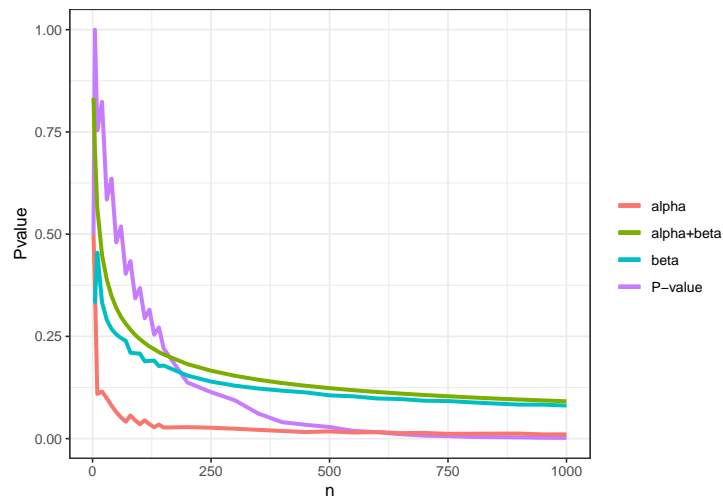
Supondo novamente que foi observado  $\bar{x} = 0.55$ , o gráfico abaixo apresenta esses valores para diversos tamanhos amostrais.

```
a=1; b=1
a1=1; b1=1
p=0.5
alpha=0.05
theta0=0.5
xbar=0.55
N=c(2,5,seq(10,150,10),seq(150,1000,50))
bf=Vectorize(function(x,n){
  exp(x*log(theta0) + (n-x)*log(1-theta0) + lbeta(a,b) - lbeta(a+x,b+n-x))
}, vectorize.args = c("x"))
alphaN = Vectorize(function(n){
  x=n*xbar
  s=seq(0,n)
  s=s[bf(s,n)<=b1/a1]
  (0.5)^n*sum(choose(n,s))
})
betaN = Vectorize(function(n){
```

```

x=n*xbar
s=seq(0,n)
s=s[bf(s,n)>b1/a1]
sum(extraDistr::dbbinom(s,n,a,b))
})
P_v=Vectorize(function(n){
  x=n*xbar
  s=seq(0,n)
  s=s[bf(s,n)<=bf(x,n)]
  (0.5)^n*sum(choose(n,s))
})
Dados=tibble(n=N, alpha=alphaN(N), beta=betaN(N), Pvalue=P_v(N))
ggplot(Dados)+
  geom_line(aes(x=n,y=Pvalue, colour="P-value"),lwd=1.2)+
  geom_line(aes(x=n,y=alpha, colour="alpha"),lwd=1.2)+
  geom_line(aes(x=n,y=beta, colour="beta"),lwd=1.2)+
  geom_line(aes(x=n,y=alpha+beta, colour="alpha+beta"),lwd=1.2)+
  theme_bw() + labs(colour="")

```







## Chapter 7

# Métodos Computacionais

Como visto, a inferência Bayesiana é baseada na aplicação monótona do teorema de Bayes

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int_{\Theta} f(x|\theta)f(\theta)d\theta} = c(x)f(x|\theta)f(\theta) \propto f(x|\theta)f(\theta),$$

e na obtenção de medidas resumo dessa distribuição, como  $E[\theta|x]$ , regiões HPD ou probabilidades a posteriori.

A maior dificuldade na aplicação de Inferência Bayesiana está justamente no cálculo das integrais envolvidas, tanto no cálculo de  $f(x)$  para a obtenção da posteriori, quanto na obtenção das medidas resumos citadas anteriormente. Devido a isso, a inferência bayesiana ganhou muito força com o avanço computacional das últimas décadas. A seguir, serão apresentados um breve resumo de alguns recursos que podem ser utilizados na prática Bayesiana.

Muitos dos métodos descritos baseiam-se na *Lei dos Grande Números* (LGN) e são uma bela aplicação de ideias frequentistas em um cenário controlado onde as suposições de *i.i.d.* são satisfeitas.

(Frac)

**Lei Forte dos Grande Números.** Seja  $X_1, X_2, \dots$  uma sequência de v.a. i.i.d com  $E[X_1] = \mu$  e  $Var[X_1] = \sigma^2 < \infty$ , então

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[q.c.]{P} E[X_1] = \mu.$$

As integrais de interesse aqui serão escritas como o valor esperado de funções de variáveis aleatórias, isto é,

$$\int g(x) dP(x) = E[g(X)].$$

Deste modo, suponha que  $X_1, X_2, \dots$  é uma sequência de v.a. i.i.d e  $g : \mathbb{R} \rightarrow \mathbb{R}$  é uma função (mensurável) tal que  $Var[g(X_1)] < \infty$ . Então, pela LGN,

$$\frac{1}{n} \sum_{i=1}^n g(X_i) \rightarrow E[g(X_1)]$$

## 7.1 Método de Monte Carlo

Suponha que deseja-se calcular  $\int_{\Theta} g(\theta) f(\theta|x) d\theta = E[g(\theta)|x]$  e é possível simular realizações  $\theta_1, \dots, \theta_m$  da distribuição de  $\theta|X = x$ ,  $f(\theta|x)$ .

Então, a integral acima pode ser aproximada por  $\frac{1}{m} \sum_{i=1}^m g(\theta_i)$

- A precisão da aproximação é usualmente estimada pelo erro padrão da estimativa

$$EP \left[ \frac{1}{m} \sum_{i=1}^m g(\theta_i) \right] \approx \sqrt{\frac{1}{m} \left( \frac{1}{m} \sum_{i=1}^m [g(\theta_i)]^2 - \left[ \frac{1}{m} \sum_{j=1}^m g(\theta_j) \right]^2 \right)}$$

**Exemplo 1.** Suponha que deseja-se estimar o número  $\pi$  usando o método de Monte Carlo. Considere então que o v.a.  $(X, Y)$  tem distribuição uniforme em um quadrado centrado na origem,  $\mathfrak{X} = [-1, 1] \times [-1, 1]$ , e um círculo  $A$  de raio 1 inscrito nesse quadrado,  $x^2 + y^2 \leq 1$ . Como a distribuição é uniforme no quadrado, a probabilidade de escolher um ponto no círculo é

$$P(A) = \frac{\text{área da círculo}}{\text{área do quadrado}} = \frac{\pi}{4} = \int_A f(x, y) dx dy = \int_{\mathfrak{X}} \mathbb{I}_A(x, y) \frac{1}{4} dx dy = E[\mathbb{I}_A(X, Y)].$$

Suponha que é possível gerar uma amostra  $\{(x_1, y_1), \dots, (x_m, y_m)\}$  de  $(X, Y)$ , de modo que podemos aproximar o valor de  $\pi$  por

$$\pi = 4 P(A) = E[4 \mathbb{I}_A(X, Y)] \approx \frac{1}{m} \sum_{i=1}^m 4 \mathbb{I}_A(x_i, y_i),$$

e, denotando por  $t = \sum_{i=1}^m \mathbb{I}_A(x_i, y_i)$ , o erro estimado é

$$\sqrt{\frac{1}{m} \left( \frac{1}{m} \sum_{i=1}^m [4 \mathbb{I}_A(x_i, y_i)]^2 - \left[ \frac{1}{m} \sum_{j=1}^m 4 \mathbb{I}_A(x_j, y_j) \right]^2 \right)} = \sqrt{\frac{1}{m} \left( \frac{16}{m} t - \left[ \frac{4}{m} t \right]^2 \right)}$$

$$= \sqrt{\frac{16}{m} \frac{t}{m} \left(1 - \frac{t}{m}\right)} \leq \sqrt{\frac{16}{m} \frac{1}{4}} = \frac{2}{\sqrt{m}}.$$

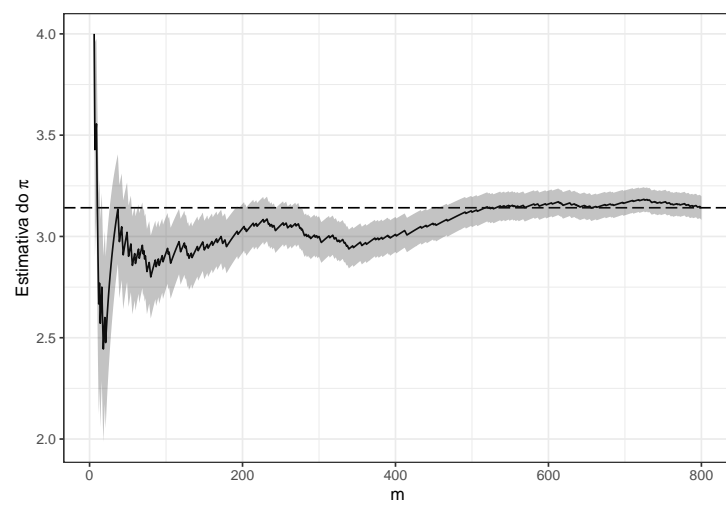
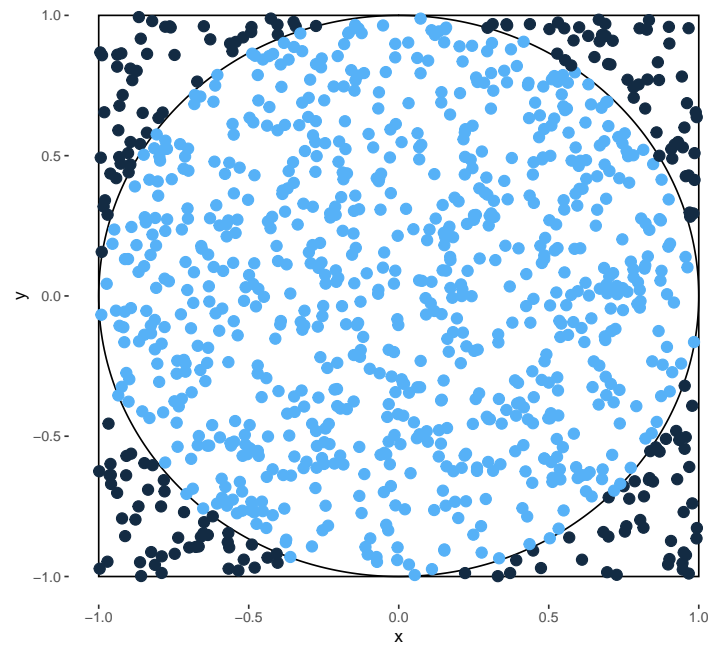
```

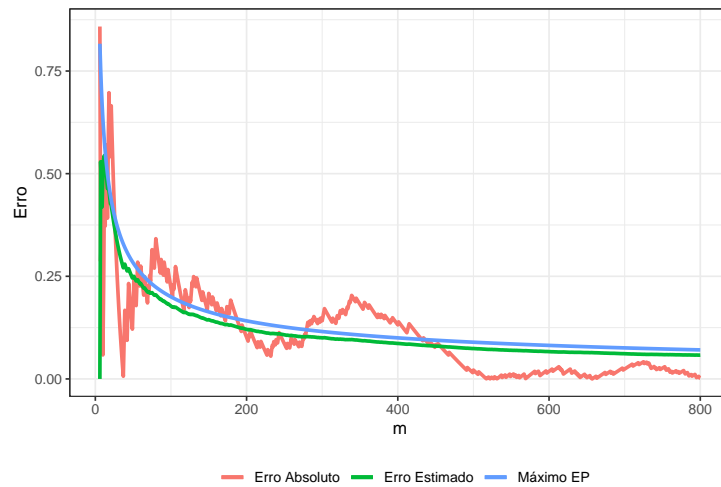
set.seed(666)
M = 1000 # número de iterações
df = tibble(t = 1:M, x = runif(length(t), -1, 1),
            y = runif(length(t), -1, 1)) %>%
  mutate(Circ=ifelse(x^2+y^2<=1,1,0),
         pi_est=round(4*cumsum(Circ)/t,4),
         erro=round(abs(pi-pi_est),4),
         erro_est=round(sqrt((cumsum(16*Circ)/t-pi_est^2)/t),4))
p <- ggplot() + theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank()) +
  ggforce::geom_circle(aes(x0 = 0, y0 = 0, r = 1), color = "black") +
  geom_rect(aes(xmin = -1, ymin = -1, xmax = 1, ymax = 1),
           color = "black", alpha = 0) +
  guides(color = FALSE) +
  geom_point(data = df, aes(x = x, y = y, colour = Circ), size = 3)
p+labs(title = expression(paste("Método de Monte-Carlo para a estimação do ",pi)), subtitle = pas

```

Método de Monte-Carlo para a estimação do  $\pi$

$m = 1000$  ;  $\text{pi\_est} = 4 * (792 / 1000) = 3.168$  ;  $\text{erro} = 0.0264$  ;  $\text{erro\_est} = 0.051$ :





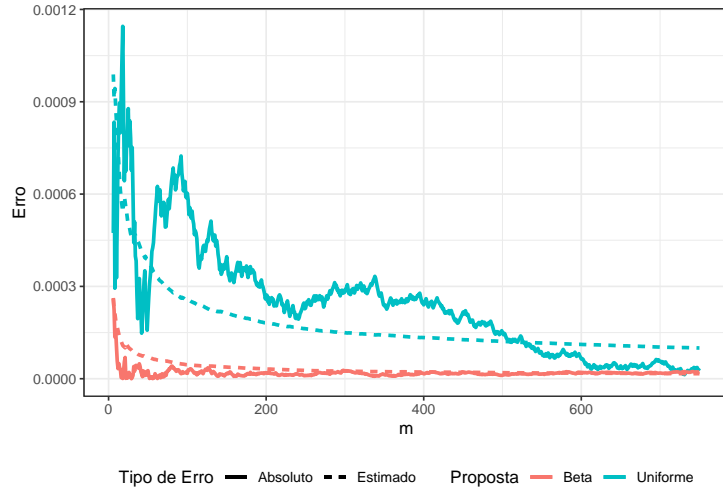
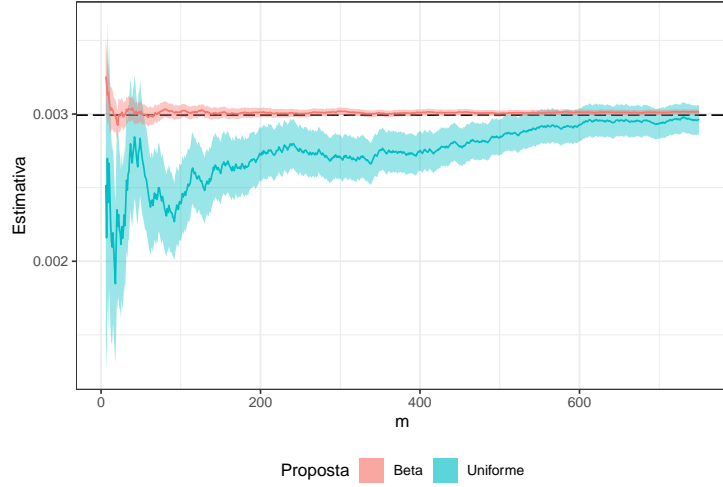
**Exemplo 2.** Suponha que você não sabe que

$$\int_0^1 x^3(1-x)^5 e^x dx = 74046 - 27240e \approx 0.0029928$$

e deseja estimar o resultado usando o método de Monte Carlo. Assim, considere as duas propostas a seguir

1.  $U \sim Unif(0, 1)$  e a integral pode ser escrita como  $E[U^3(1-U)^5 e^U]$ ;
2.  $Y \sim Beta(4, 6)$  de modo que

$$\int_0^1 y^3(1-y)^5 e^y dy = \beta(4, 6) \int_0^1 e^y \frac{y^{4-1}(1-y)^{6-1}}{\beta(4, 6)} dy = \beta(4, 6) E[e^Y].$$



**Exemplo 3. Região HPD** Suponha que  $\theta = (\mu, \sigma^2) \sim \text{Normal-Inv.Gama}(m, v, a, b)$  e deseja-se obter estimativas pontuais e por região para  $\theta$ .

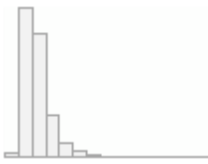
Se não houver um simulador da distribuição Normal-Inv.Gamma diretamente, é possível gerar um ponto  $\theta_i = (\mu_i, \sigma_i^2)$  tomando  $\sigma_i^2 \sim \text{Inv.Gama}(a, b)$  (ou  $\tau_i \sim \text{Gama}(a, b)$  e fazer  $\sigma_i^2 = 1/\tau_i$ ) e  $\mu_i \sim \text{Normal}(m, \sigma_i^2/v)$ . Nesse exemplo é fácil simular uma amostra da distribuição posterior e é possível obter estimativas pontuais

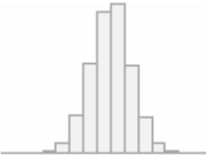
simplesmente obtendo estatística resumo da amostra simulada, como média, moda, mediana, variância e desvio padrão.

Para construir a região HPD, primeiramente note que  $R = \{\theta \in \Theta : f(\theta|x) \geq h\} = \{\theta \in \Theta : f(x|\theta)f(\theta) \geq h^* = c \cdot h\}$ , de modo que não é necessário conhecer a constante  $c = f(x)$  para realizar essa tarefa. Como nesse exemplo a distribuição a posteriori é conhecida e fácil de ser simulada, considere o algoritmo a seguir para estimar a região HPD de probabilidade  $\gamma$ .

1. Simular  $\theta_1, \dots, \theta_m$  de  $f(\theta|x)$ ;
2. Encontrar  $h$  tal que  $\frac{1}{m} \sum_{i=1}^m \mathbb{I}_R(\theta_i) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(\theta_i|x) \geq h) \approx \gamma$ 
  - i. Calcule  $f(\theta_i|x)$ ,  $i = 1, \dots, m$ ;
  - ii. Ordene esses valores e tome  $h$  como o percentil de ordem  $\gamma$ ;
3. Fazer o gráfico da curva de nível  $f(\theta|x) = h$ .

```
set.seed(666)
a=7; b=7; m=0; v=0.5 # parametros da posteriori
M=10000 # No. de simulações
dpost=Vectorize(function(t1,t2){ #densidade posterior
  extraDistr::dinvgamma(t2,a,b)*dnorm(t1,m,sqrt(t2/v))})
# simulações
df = tibble(sigma2=extraDistr::rinvgamma(M,a,b)) %>%
  mutate(mu=rnorm(M,m,sqrt(sigma2/v)))
#summarytools::dfSummary(df, graph.magnif = 0.75, valid.col = FALSE, na.col = FALSE)
summarytools::dfSummary(df, plain.ascii = FALSE, style = "grid",
  graph.magnif = 0.75, valid.col = FALSE, na.col = FALSE,
  varnumbers = FALSE, headings = FALSE, tmp.img.dir = "./tmp")
```

Variable	Stats / Values	Freqs (% of Valid)	Graph
sigma2 [numeric]	Mean (sd) : 1.2 (0.5) min < med < max: 0.3 < 1.1 < 7.3 IQR (CV) : 0.6 (0.4)	10000 distinct values	

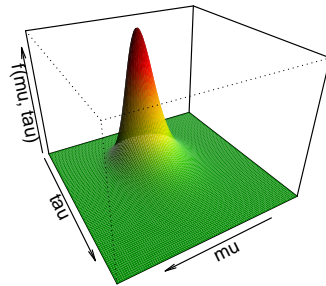
Variable	Stats / Values	Freqs (% of Valid)	Graph
mu [numeric]	Mean (sd) : 0 (1.5) min < med < max: -7.3 < 0 < 6.1 IQR (CV) : 2 (-253.7)	10000 distinct values	

```

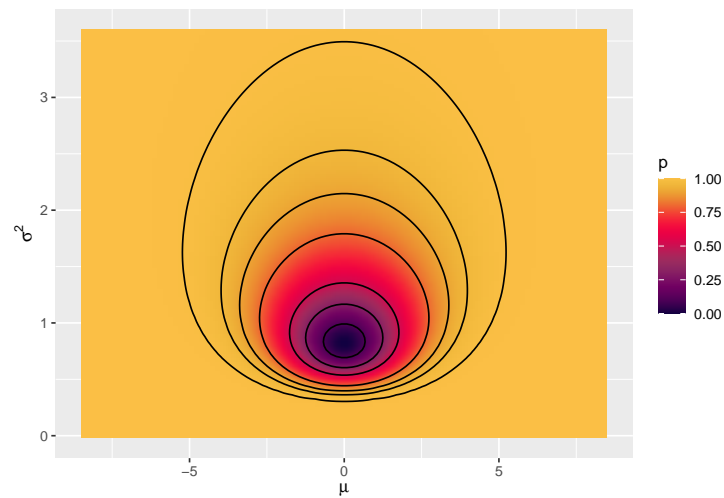
df = df %>% mutate(post=dpost(mu,sigma2))
# variáveis para os gráficos
gama=c(0.99,0.95,0.9,0.8,0.5,0.3,0.1) # prob das regiões
l=quantile(df$post,1-gama)
d=100
x=seq(-4*extraDistr::qinvgamma(0.5,a,b)/v,4*extraDistr::qinvgamma(0.5,a,b)/v,length.out=d)
y=seq(0,extraDistr::qinvgamma(0.996,a,b),length.out = d)
z=matrix(apply(cbind(rep(x,d),rep(y,each=d)),1,function(t){dpost(t[1],t[2])}),ncol=d)
# gráfico da posteriori
# Create a function interpolating colors in the range of specified colors
jet.colors <- colorRampPalette(c('green','yellow','orange','red','darkred'))
# Generate the desired number of colors from this palette
nbcol <- 300
cores <- jet.colors(nbcol)
# Compute the z-value at the facet centres
zfacet <- z[-1, -1] + z[-1, -ncol(z)] +
          z[-nrow(z), -1] + z[-nrow(z), -ncol(z)]
# Recode facet z-values into color indices
facetcol <- cut(zfacet, nbcol)
persp(x, y, z, col = cores[facetcol],theta=150,phi=30,expand=0.75,
      ticktype="simple", xlab=expression(mu), ylab=expression(tau),
      zlab=expression(f(mu,tau)),axes=TRUE,border=NA,shade=0.9)

```





```
# gráfico das regiões HPD de prob. gama=c(0.99,0.95,0.9,0.8,0.5,0.3,0.1)
tibble(x1=rep(x,d),y1=rep(y,each=d),z1=as.vector(z)) %>%
  arrange(z1) %>% mutate(p=1-(cumsum(z1)/sum(z1))) %>%
  ggplot(aes(x1,y1,z=z1,fill = p)) +
  geom_raster(interpolate = TRUE) +
  jcolors::scale_fill_jcolors_contin("pal3") +
  #scale_fill_distiller(palette = "YlOrRd") +
  geom_contour(breaks=1,col="black") +
  xlab(expression(mu)) + ylab(expression(sigma^2))
```



## 7.2 Monte Carlo com Amostragem de Importância

Considere  $f(\theta|x) \propto f(x|\theta)f(\theta)$  e suponha que não se sabe simular observações desta distribuição mas tem-se interesse na quantidade  $E[g(\theta)|x] = \int_{\Theta} g(\theta)f(\theta|x)d\theta$ .

Suponha também que existe uma distribuição  $h(\theta)$  que seja uma aproximação para  $f(\theta|x)$  (preferencialmente com caudas mais pesadas) da qual sabe-se simular. Então,

$$\begin{aligned} E[g(\theta)|x] &= \int_{\Theta} g(\theta)f(\theta|x)d\theta = \frac{\int_{\Theta} g(\theta)f(x|\theta)f(\theta)d\theta}{\int_{\Theta} f(x|\theta)f(\theta)d\theta} = \frac{\int_{\Theta} g(\theta) \left( \frac{f(x|\theta)f(\theta)}{h(\theta)} \right) h(\theta)d\theta}{\int_{\Theta} \left( \frac{f(x|\theta)f(\theta)}{h(\theta)} \right) h(\theta)d\theta} \\ &= \frac{\int_{\Theta} g(\theta)w(\theta)h(\theta)d\theta}{\int_{\Theta} w(\theta)h(\theta)d\theta} \approx \sum_{i=1}^m \frac{w_i}{\sum_{j=1}^m w_j} g(\theta_i), \end{aligned}$$

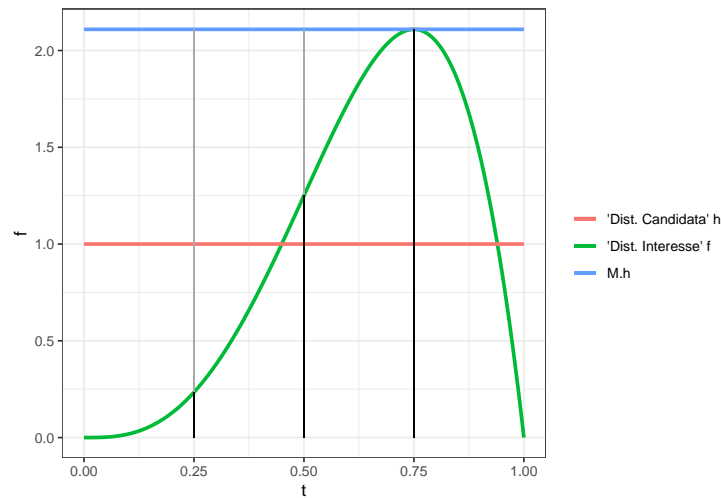
onde  $w_i = w(\theta_i) = \frac{f(x|\theta_i)f(\theta_i)}{h(\theta_i)}$ .

## 7.3 Método de Rejeição

Considere novamente que o objetivo é simular observações de  $f(\theta|x)$  mas não é possível fazer isso diretamente. Por outro lado, sabe-se simular dados de uma “distribuição candidata”,  $h(\theta)$ , tal que  $f(\theta|x) \leq Mh(\theta)$ ,  $\forall \theta \in \Theta$  e para alguma constante  $M$ . A ideia do método é rejeitar pontos gerados em regiões em que  $h$  atribui maior probabilidade que  $f$  com probabilidade  $1 - [f(\theta|x) / Mh(\theta)]$ . Para que a afirmação anterior faça sentido,  $M$  deve ser tal que  $f(\theta|x) / Mh(\theta) \leq 1$ ,  $\forall \theta \in \Theta$ , de modo que a melhor escolha para  $M$  é  $M^* = \sup_{\Theta} \frac{f(\theta|x)}{h(\theta)}$ .

```
r = function(t){dbeta(t,4,2)/dbeta(t,1,1)}
M = optimize(r,c(0,1),maximum = TRUE)
tibble(t = seq(0,1,length.out = 1000)) %>%
  mutate(f=dbeta(t,4,2), h=dbeta(t,1,1),
         Mh=M$objective*dbeta(t,1,1)) %>%
  ggplot() + theme_bw() +
  geom_line(aes(x=t,y=f,colour="Dist. Interesse' f"),lwd=1.1) +
  geom_line(aes(x=t,y=h,colour="Dist. Candidata' h"),lwd=1.1) +
  geom_line(aes(x=t,y=Mh,colour="M.h"),lwd=1.1) +
```

```
geom_segment(x=0.5,xend=0.5,y=dbeta(0.5,4,2),yend=M$objective*dbeta(0.5,1,1),col="darkgrey") +
geom_segment(x=0.5,xend=0.5,y=0,yend=dbeta(0.5,4,2)) +
geom_segment(x=0.25,xend=0.25,y=dbeta(0.25,4,2),yend=M$objective*dbeta(0.25,1,1),col="darkgrey") +
geom_segment(x=0.25,xend=0.25,y=0,yend=dbeta(0.25,4,2)) +
geom_segment(x=0.75,xend=0.75,y=dbeta(0.75,4,2),yend=M$objective*dbeta(0.75,1,1),col="darkgrey") +
geom_segment(x=0.75,xend=0.75,y=0,yend=dbeta(0.75,4,2)) +
labs(colour="")
```



No exemplo apresentado no gráfico acima, suponha que foram gerados os “candidatos” 0.25, 0.5 e 0.75. É possível notar que o ponto 0.75 deve ser aceito, o ponto 0.5 deve ser aceito com probabilidade 0.59 e o ponto 0.25 deve ser aceito com probabilidade 0.11. A seguir é apresentado o pseudo-algoritmo do método da rejeição.

**Para**  $i = 1, \dots, m$

**Repita**

Simule  $u \sim \text{Unif}(0, 1)$

Simule  $\theta'$  da distribuição candidata  $h(\theta)$

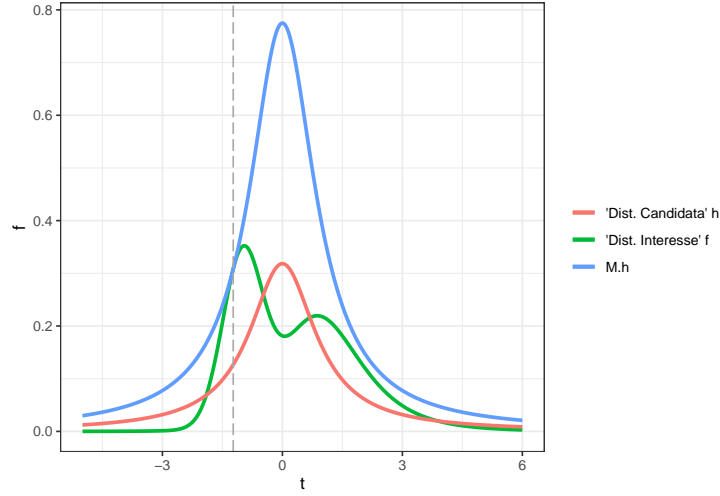
**Até**  $u \leq \frac{f(\theta'|x)}{Mh(\theta')}$

$\theta_i = \theta'$

**Fim\_Para.**

```
r = function(t){(0.4*dnorm(t,-1,1/2)+0.6*dt(t,5,1))/dt(t,1)}
M = optimize(r,c(-8,10),maximum = TRUE)
tibble(t = seq(-5,6,length.out = 1000)) %>%
  mutate(f=(0.4*dnorm(t,-1,1/2)+0.6*dt(t,5,1)), h=dt(t,1),
         Mh=M$objective*dt(t,1)) %>%
```

```
ggplot() + theme_bw() +
  geom_line(aes(x=t,y=f,colour="'Dist. Interesse' f"),lwd=1.1) +
  geom_line(aes(x=t,y=h,colour="'Dist. Candidata' h"),lwd=1.1) +
  geom_line(aes(x=t,y=Mh,colour="M.h"),lwd=1.1) +
  geom_vline(xintercept=M$maximum,linetype="longdash",col="darkgrey") +
  labs(colour="")
```



A linha tracejada representa o ponto na escolha ótima para  $M$ . Nesse exemplo é possível notar que na região central, onde é mais “provável” gerar observações de  $h$ , a razão  $f(\theta|x) / Mh(\theta)$  é menor que 0.25, de modo que há uma grande probabilidade de rejeição. Isso justifica a escolha de distribuições candidatas com caudas pesadas. No caso geral, a *probabilidade de aceitação* do método é

$$\begin{aligned}
 P\left(\left\{(U, \theta) : U \leq \frac{f(\theta|x)}{Mh(\theta)}\right\}\right) &= E_{U, \theta} \left[ \mathbb{I}\left(U \leq \frac{f(\theta|x)}{Mh(\theta)}\right) \right] = E_{\theta} \left[ E_{U|\theta} \left[ \mathbb{I}\left(U \leq \frac{f(\theta|x)}{Mh(\theta)}\right) \right] \right] \\
 &= E_{\theta} \left[ P\left(U \leq \frac{f(\theta|x)}{Mh(\theta)} \mid \theta\right) \right] = E_{\theta} \left[ \frac{f(\theta|x)}{Mh(\theta)} \right] = \int_{\Theta} \frac{f(\theta|x)}{Mh(\theta)} h(\theta) d\theta = \\
 &= \frac{1}{M} \int_{\Theta} f(\theta|x) d\theta = \frac{1}{M}.
 \end{aligned}$$

No exemplo, a probabilidade de aceitação é  $1/2.434 = 0.41$ , ou seja, mais de metade das observações geradas seriam descartadas.

## 7.4 ABC (Aproximated Bayesian Computation)

O método ABC é uma forma bastante simples de gerar pontos da distribuição a posteriori. Para sua utilização é suficiente saber gerar pontos da distribuição dos dados e da priori, de modo que a verossimilhança nem precisa ser analiticamente conhecida, fato esse que faz com que o método seja dito ser “*likelihood-free*”.

Suponha o caso em que  $X$  é **discreto** com função de verossimilhança  $f(x|\theta)$ , a priori é  $f(\theta)$  e foi observado  $X = x_o$ . Abaixo é apresentado o *pseudo-algoritmo* para simular observações da posteriori  $f(\theta|x_o)$  usando o método ABC.

### Algoritmo ABC ( $X$ discreto)

```

Para  $i = 1, \dots, m$ 
  Repita
    Gere  $\theta'$  de  $f(\theta)$  (priori)
    Gere  $y = (y_1, \dots, y_n)$  de  $f(x|\theta')$  (verossimilhança)
  Até  $y = x_o$ 
     $\theta_i = \theta'$ 
Fim_Para

```

Para verificar que o método funciona no caso discreto, basta ver que

$$f(\theta_i) = \sum_{y \in \mathfrak{X}} f(\theta_i) f(y|\theta_i) \mathbb{I}(y = x_o) = f(\theta_i) f(x_o|\theta_i) \propto f(\theta|x_o).$$

No caso em que  $X$  é contínuo, a probabilidade de gerar uma nova amostra  $Y$  exatamente igual ao ponto observado  $x_o$  é zero,  $P(Y = x_o) = 0$ . Nesse caso, o algoritmo é adaptado de modo que são aceitos pontos gerados com  $\Delta(\eta(y), \eta(x_o)) \leq \varepsilon$ , onde  $\Delta$  é uma medida de distância conveniente,  $\eta$  é uma estatística (que pode não ser suficiente para  $\theta$ ) e  $\varepsilon$  é uma constante de tolerância. O *pseudo-algoritmo* é apresentado a seguir.

### Algoritmo ABC ( $X$ qualquer)

```

Para  $i = 1, \dots, m$ 
  Repita
    Gere  $\theta'$  de  $f(\theta)$ 
    Gere  $y$  de  $f(x|\theta')$ 
  Até  $\Delta(\eta(x), \eta(y)) \leq \varepsilon$ 
     $\theta_i = \theta'$ 
Fim_Para

```

## 7.5 MCMC - Monte Carlo via Cadeias de Markov

### 7.5.1 Pequena Introdução às Cadeias de Markov

**Definição** Um *processo estocástico* (em tempo discreto) é uma sequência de v.a.  $X_0, X_1, X_2, \dots$  indexada em  $\mathbb{N}$  (os índices podem indicar, por exemplo, tempo ou espaço ou ?). O conjunto  $E$  onde  $X_i$  toma valores é chamado de *espaço de estados*.

**Definição** Um processo estocásticos é dito uma *Cadeia de Markov* (em tempo discreto) se,  $\forall n \geq 1$  e  $\forall A \subseteq E$ ,

$$P(X_{n+1} \in A | X_n = x_n, \dots, X_1 = x_1, X_0 = x_0) = P(X_{n+1} \in A | X_n = x_n)$$

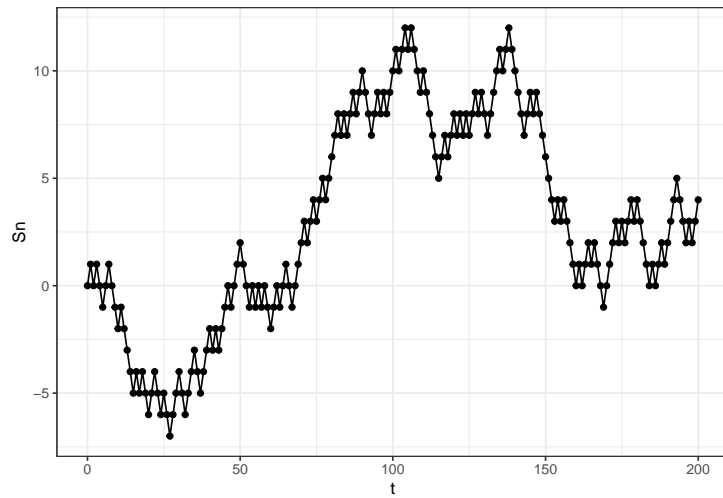
**Exemplo 1.** Suponha uma sequência de v.a.  $(X_n)_{n \geq 1}$  i.i.d. tais

que  $p = P(X_1 = 1) = 1 - P(X_1 = -1)$ . Defina  $S_n = \sum_{i=1}^n X_i$  e

$X_0 = c$ . O processo estocástico  $(S_n)_{n \geq 0}$  é uma Cadeia de Markov.

De fato,

$$\begin{aligned} P(S_n = s_n | S_{n-1} = s_{n-1}, \dots, S_0 = s_0) &= P(X_n + S_{n-1} = s_n | S_{n-1} = s_{n-1}, \dots, S_0 = s_0) \\ &= P(X_n = s_n - s_{n-1} | S_{n-1} = s_{n-1}, \dots, S_0 = s_0) = P(X_n = s_n - s_{n-1} | S_{n-1} = s_{n-1}) \\ &= P(S_n = s_n | S_{n-1} = s_{n-1}) \end{aligned}$$



Uma Cadeia de Markov é caracterizada pela distribuição do *estado inicial*  $X_0$  e pelas *probabilidades de transição*  $Q(x, A) = P(X_n \in A | X_{n-1} = x)$ . Se  $Q(x, A)$  não depende de  $n$ , dizemos que é *homogênea no tempo*.

Para cada  $n$ , a cadeia pode

1. Continuar no estado anterior  $x$ , ou seja,  $X_{n+1} = x$ , com probabilidade  $r(x)$ ,  $0 \leq r < 1$ , ou
2. Saltar para um estado  $y$  segundo uma função de densidade de probabilidade  $q(x, y)$ , onde  $0 < \int_E q(x, y) dy = 1 - r(x) \leq 1$  (sub-probabilidade).  
No caso discreto vamos considerar  $q(x, x) = 0$ .

$$\text{Assim, } Q(x, A) = P(X_{n+1} \in A | X_n = x) = \int_A q(x, y) dy + r(x) \mathbb{I}_A(x).$$

Suponha que para um dado  $n$ ,  $X_n$  tem densidade  $\lambda$ , isto é,  $P(X_n \in A) = \int_A \lambda(x) dx$ . Então, a densidade de  $X_{n+1}$  pode ser obtida por

$$\begin{aligned} P(X_{n+1} \in A) & \stackrel{\text{regra da prob. total}}{=} \int_E \lambda(x) Q(x, A) dx = \int_E \lambda(x) \left[ \int_A q(x, y) dy + r(x) \mathbb{I}_A(x) \right] dx \\ &= \int_A \int_E \lambda(x) q(x, y) dx dy + \int_E \lambda(x) r(x) \mathbb{I}_A(x) dx = \int_A \int_E \lambda(x) q(x, y) dx dy + \\ & \int_A \lambda(y) r(y) dy = \int_A \underbrace{\left[ \int_E \lambda(x) q(x, y) dx + \lambda(y) r(y) \right]}_{\text{f.d.p. de } X_{n+1}} dy. \end{aligned}$$

$$\text{Assim, a f.d.p de } X_{n+1} \text{ é } \lambda Q(y) = \int_E \lambda(x) q(x, y) dx + \lambda(y) r(y)$$

Dizemos que a densidade  $\pi$  é *invariante (estacionária)* se as densidades de  $X_n$  e  $X_{n+1}$  são iguais (q.c), isto é,  $\pi = \pi Q$  ou  $\int_A \pi(x) dx = \int_E \pi(x) Q(x, A) dx$ .

**Resultado 1.** A afirmação anterior é equivalente a  $\int \pi(x) q(x, y) dx = (1 - r(x)) \pi(y)$ .

**Resultado 2.** Se a função  $q(x, y)$  satisfaz a condição de *reversibilidade*, isto é,  $\pi(x)q(x, y) = \pi(y)q(y, x)$ , então  $\pi$  é uma *medida invariante* da cadeia com função de transição  $Q(x, \cdot)$ .

**Demo 1.**

$$\int_E \pi(x)q(x, y)dx = \int_E \pi(y)q(y, x)dx = \pi(y) \underbrace{\int_E q(y, x)dx}_{1-r(y)}$$

**Demo 2.**

$$\begin{aligned} \int_E \pi(x)Q(x, A)dx &= \int_E \pi(x) \left[ \int_A q(x, y)dy \right] dx + \int_E \pi(x)r(x)\mathbb{I}_A(x)dx \\ &= \int_A \left[ \int_E \pi(x)q(x, y)dx \right] dy + \int_A \pi(x)r(x)dx = \int_A \left[ \int_E \pi(y)q(y, x)dx \right] dy + \\ &\quad \int_A \pi(y)r(y)dy = \int_A \pi(y) \left[ \int_E q(y, x)dx \right] dy + \int_A \pi(y)r(y)dy \\ &= \int_A \pi(y)[1 - r(y)]dy + \int_A \pi(y)r(y)dy = \int_A \pi(y)[1 - r(y) + r(y)]dy \\ &= \int_A \pi(y)dy. \end{aligned}$$

### 7.5.2 O algoritmo de Metrópolis-Hastings

Suponha que deseja-se gerar observações de  $\pi(\theta) \propto f(x|\theta)f(\theta) \propto f(\theta|x)$ . Defina uma Cadeia de Markov  $(Y_n)_{n \geq 1}$  tal que, no instante  $n$ ,  $Y_n = y$ . No instante  $n + 1$ , um candidato  $z$  é gerado segundo a densidade  $q(y, z)$  e é aceito com probabilidade  $\alpha(y, z)$ . Isto é, se  $Y_n = y$ ,

$$Y_{n+1} = \begin{cases} z, & \text{com probabilidade } \alpha(y, z) \\ y, & \text{com probabilidade } 1 - \alpha(y, z) \end{cases},$$

em que  $\alpha$  é dado por

$$\alpha(y, z) = \begin{cases} \min \left\{ \frac{\pi(z)q(z, y)}{\pi(y)q(y, z)}, 1 \right\}, & \text{se } \pi(y)q(y, z) > 0 \\ 1, & \text{c.c.} \end{cases}$$

**Resultado:** O algoritmo de M-H gera uma cadeia reversível com respeito a  $\pi$  e, portanto, tem  $\pi$  como distribuição estacionária.



**Demo.** Deve-se mostrar que  $\pi(y) \underbrace{q(y, z)\alpha(y, z)}_{p(y, z)} = \pi(z) \underbrace{q(z, y)\alpha(z, y)}_{p(z, y)}$ .

Suponha  $\pi(z)q(z, y) \geq \pi(y)q(y, z)$  (o caso  $\leq$  é análogo)

i) Se  $\pi(z)q(z, y) = 0 \Rightarrow \pi(y)q(y, z) = 0$  e vale a reversibilidade.

ii)  $\pi(z)q(z, y) > 0 \Rightarrow \alpha(y, z) = 1$  e  $\alpha(z, y) = \frac{\pi(y)q(y, z)}{\pi(z)q(z, y)}$ .

Nesse caso,  $\pi(z)q(z, y)\alpha(z, y) = \pi(z)q(z, y) \frac{\pi(y)q(y, z)}{\pi(z)q(z, y)} = \pi(y)q(y, z)$   
 $= \pi(y)q(y, z)\alpha(y, z)$

### 7.5.3 Amostrador de Gibbs

Suponha que a  $\dim(\Theta) > 1$  e deseja-se amostrar  $f(\theta|x)$  e suponha que é possível obter amostras das distribuições *condicionais completas*, isto é, de  $f(\theta_i|\theta_{-i}, x)$ , onde  $\theta_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$ . Note que  $f(\theta_i|\theta_{-i}, x) \propto f(\theta|x) = f(x|\theta)f(\theta)$ . O método do *Amostrador de Gibbs* é um caso particular do algoritmo de Metrópolis-Hastings em que é gerada uma cadeia  $(\theta^{(n)})_{n \geq 1}$  com  $\alpha(y, z) = 1$  e  $q(y, z) = f(\theta_j = z \mid \theta_{-j} = y, x)$ , gerada segundo o algoritmo a seguir.

#### Algoritmo - Amostrador de Gibbs

Defina uma “chute inicial”  $\theta^{(0)}$  (por exemplo, gerado da priori  $f(\theta)$  ou fixado) **Para**  $i = 1, \dots, m$

Gere  $\theta_1^{(i)}$  de  $f(\theta_1|\theta_{-1}^{(i-1)}, x)$

Gere  $\theta_2^{(i)}$  de  $f(\theta_2|\theta_1^{(i)}, \theta_3^{(i-1)}, \dots, \theta_k^{(i-1)}, x)$

$\vdots$

Gere  $\theta_{k-1}^{(i)}$  de  $f(\theta_{k-1}|\theta_1^{(i)}, \dots, \theta_{k-2}^{(i)}, \theta_k^{(i-1)}, x)$

Gere  $\theta_k^{(i)}$  de  $f(\theta_k|\theta_{-k}^{(i)}, x)$

**Fim\_Para**

Os métodos de Metrópolis-Hastings descritos anteriormente geram observações de cadeias de Markov com distribuição estacionária que coincide com a posteriori. Contudo, deve-se tomar dois cuidados para a utilização de métodos de Monte Carlo usando essas observações. O primeiro é que é necessário verificar se a cadeia já atingiu a estacionariedade. Essa verificação é feita, em geral, observando os gráficos das cadeias geradas e, em geral, as primeiras  $b$  observações são descartadas (*burn-in*). Outra possibilidade para verificar a estacionariedade

da cadeia, bem como a influência do chute inicial, é gerar duas ou mais cadeias iniciando-se de pontos distintos. Outro problema é a dependência entre as observações geradas. Para contornar esse problema, normalmente uma distância  $k$  entre as observações que serão consideradas na amostra final (*thin*) e as observações entre estas são descartadas. Assim, a amostra final é formada pelos pontos  $\theta_b, \theta_{b+k}, \theta_{b+2k}, \dots, \theta_M$ .

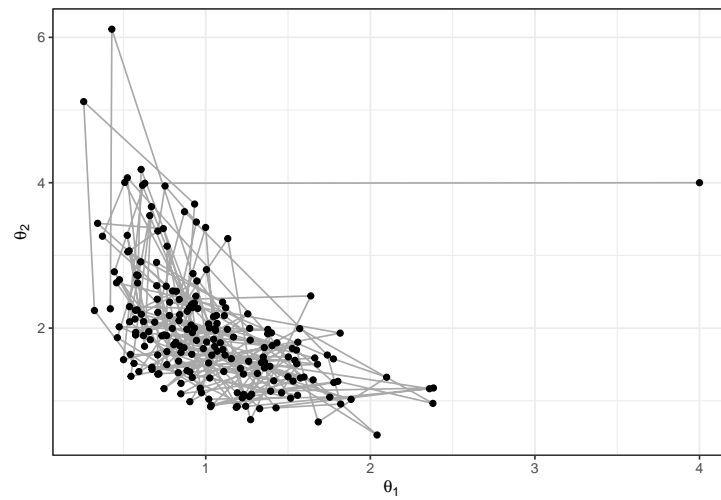
**Exemplo 1.** Seja  $X_1, \dots, X_n$  c.i.i.d. tais que  $X_i | \theta_1, \theta_2 \sim \text{Exp}(\theta_1 \theta_2)$  e considere que a priori  $\theta_i \sim \text{Gama}(a_i, b_i)$ ,  $i = 1, 2$ . Assim,  

$$f(\theta|x) \propto f(x|\theta)f(\theta_1)f(\theta_2) \propto (\theta_1 \theta_2)^n e^{-\theta_1 \theta_2 \sum x_i} \theta_1^{a_1-1} e^{-b_1 \theta_1} \theta_2^{a_2-1} e^{-b_2 \theta_2}$$

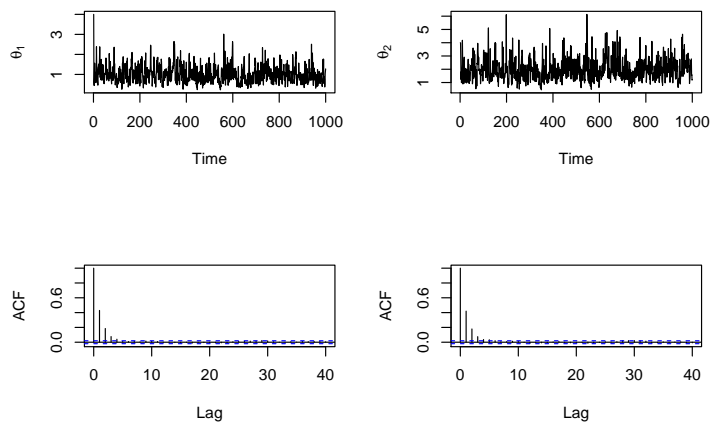
$$\propto \theta_1^{a_1+n-1} \theta_2^{a_2+n-1} e^{-b_1 \theta_1 - b_2 \theta_2 - \theta_1 \theta_2 \sum x_i}.$$
Essa distribuição não é conhecida mas é possível obter as distribuições condicionais completas  

$$f(\theta_i | \theta_j, x) \propto \theta_i^{a_i+n-1} e^{-[b_i \theta_j \sum x_i] \theta_i} \implies \theta_i | \theta_j, x \sim \text{Gama}(a_i + n, b_i + \theta_j \sum x_i),$$
e, portanto, é possível simular observações da posteriori usando o amostrador de Gibbs.

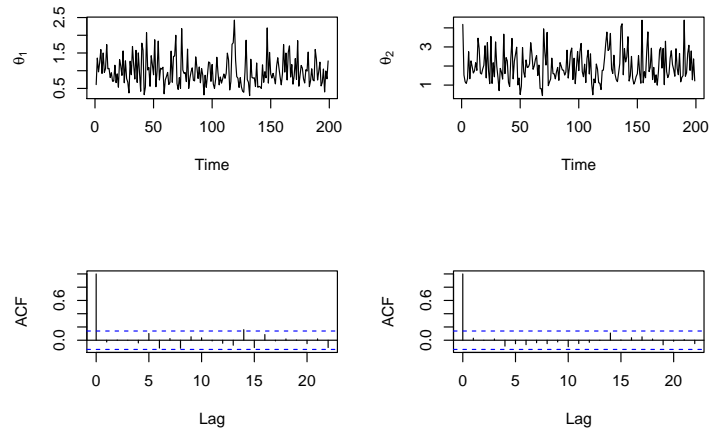
```
set.seed(666)
a1=2; b1=3
a2=3; b2=2
n=8
sumx=4
M=10000
theta1=vector(length = M)
theta2=vector(length = M)
theta1[1] = 4
theta2[1] = 4
for(i in 2:M){
  theta1[i] = rgamma(1,a1+n,b1+theta2[i-1]*sumx)
  theta2[i] = rgamma(1,a2+n,b2+theta1[i]*sumx)
}
m=200; sel=seq(1,m)
tibble(theta1=theta1[sel],theta2=theta2[sel],t=seq(1,length(sel))) %>%
  ggplot() + theme_bw() +
  geom_path(aes(theta1,theta2),col="darkgrey") +
  geom_point(aes(theta1,theta2)) +
  xlab(expression(theta[1])) + ylab(expression(theta[2]))
```



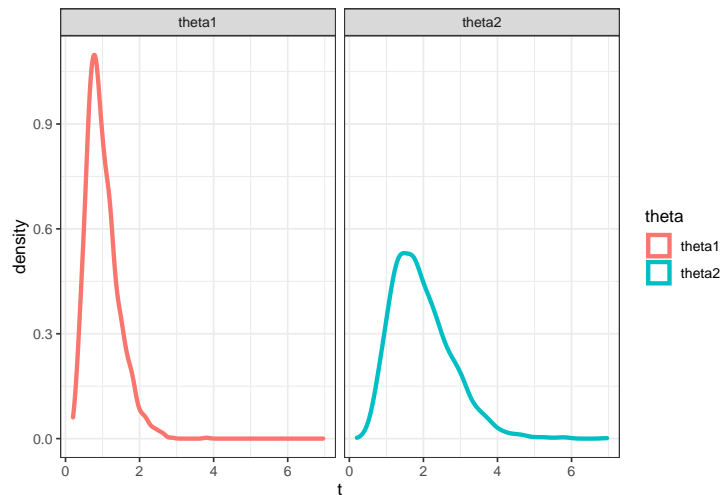
A seguir, são apresentados os gráficos das cadeias geradas e das autocorrelações.



Aparentemente, a cadeia converge rapidamente para a distribuição estacionária mas as autocorrelações entre observações consecutivas é alta. Assim, vamos descartar as 10 primeiras observações e considerar saltos de tamanho 5. Os novo gráficos são apresentados abaixo.



Por fim são apresentadas as estimativas das densidades marginais e as regiões HPD.

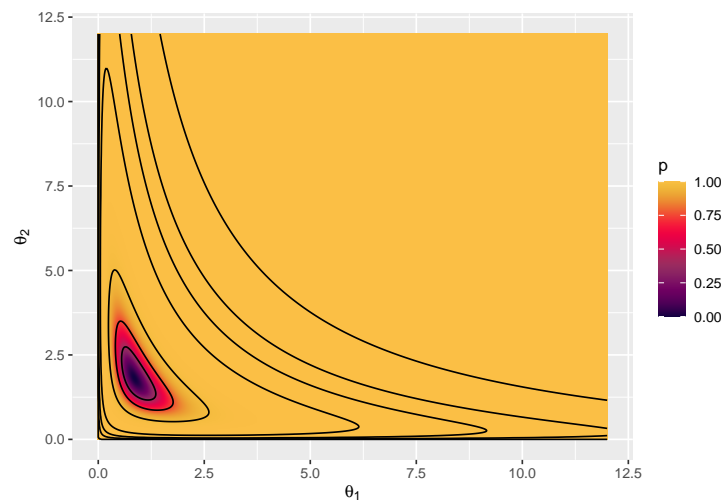


```
sel=seq(10,M,5)
dpost=Vectorize(function(t1,t2){ #densidade posterior
  exp((n-1)*log(t1*t2)+dexp(sumx,t1*t2,log=TRUE)+dgamma(t1,a1,b1,log=TRUE)+dgamma(t2,a2,b2,log=TRUE))
# simulações
df = tibble(theta1=theta1[sel],theta2=theta2[sel]) %>%
  mutate(post=dpost(theta2,theta2))
# variáveis para os gráficos
gama=c(0.99,0.95,0.9,0.8,0.5,0.3,0.1) # prob das regiões
```

```

l=quantile(df$post,1-gama)
d=1000
x=seq(0,12,length.out = d)
y=seq(0,12,length.out = d)
z=apply(cbind(rep(x,d),rep(y,each=d)),1,function(t){dpost(t[1],t[2])})
# gráfico das regiões HPD de prob. gama=c(0.99,0.95,0.9,0.8,0.5,0.3,0.1)
tibble(x1=rep(x,d),y1=rep(y,each=d),z1=z) %>%
  arrange(z1) %>% mutate(p=1-(cumsum(z1)/sum(z1))) %>%
  ggplot(aes(x1,y1,z=z1,fill = p)) +
  geom_raster(interpolate = TRUE) +
  jcolors::scale_fill_jcolors_contin("pal3") +
  geom_contour(breaks=l,col="black") +
  xlab(expression(theta[1])) + ylab(expression(theta[2]))

```



## 7.6 Bibliotecas de R para Inferência Bayesiana

Nessa seção serão apresentadas algumas bibliotecas do R para inferência Bayesiana, em especial, `LaplacesDemon` e `Stan`, que são bibliotecas utilizadas para simular dados da posteriori. Alguns dos gráficos apresentados nessa seção serão construídos com bibliotecas específicas para avaliar amostras da posteriori, `ggmcmc` e `bayesplot`.

### 7.6.1 O Modelo de Regressão Linear

Inicialmente, será apresentado como exemplo o modelo de regressão linear, possivelmente um dos métodos mais usados nas aplicações de inferência estatística.

Considere  $n$  observações de uma variável aleatória de interesse (chamada de *variável dependente* ou *variável resposta*) e de  $p - 1$  características associadas a cada uma dessas observações (chamadas de *variáveis dependentes* ou *explicativas* ou *covariáveis*), supostamente fixadas. Um modelo de regressão linear pode ser escrito como

$$Y = X\beta + \epsilon$$

com  $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$  ;  $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix}$  ;  $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$  ;

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} ; \quad Z = [X, Y] \quad ,$$

em que  $Z$  é a matriz de dados (observada),  $\beta$  é o vetor de parâmetros e  $\epsilon_i$  é o “erro aleatório” associado a  $i$ -ésima observação, supostamente c.i.i.d. com distribuição  $Normal(0, \sigma^2)$ .

De forma equivalente, o modelo pode ser escrito como  $Y|X, \beta, \sigma \sim Normal_n(\mu, \Sigma)$  com  $\mu = X\beta$  e  $\Sigma = \sigma^2 I$ .

Na abordagem *frequentista*, se  $X'X$  é não singular, os estimadores de máxima verossimilhança para os parâmetros  $(\beta, \sigma^2)$  são, respectivamente,  $\hat{\beta} = (X'X)^{-1}X'Y$  e  $s^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - p}$ .

**Exemplo.** Vamos considerar um simples exemplo de regressão linear, com apenas uma covariável. Para isso, considere as variáveis `speed` e `dist` do conjunto de dados `cars`, disponível no R. Um ajuste usando a abordagem frequentista é apresentado a seguir.

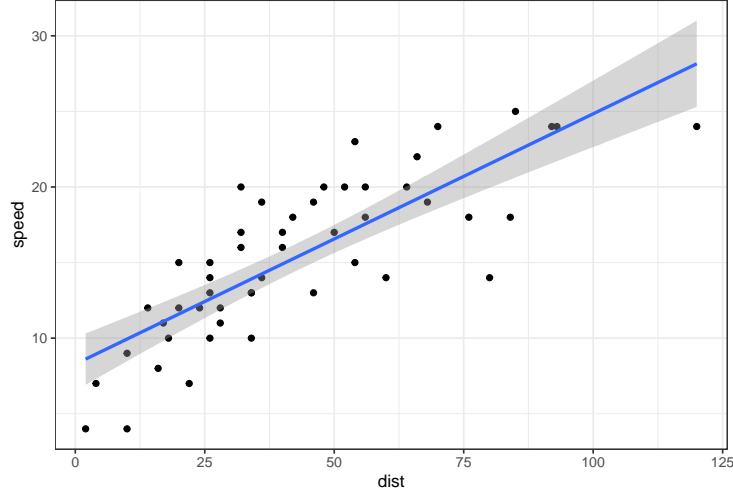
```
# a boring regression
fit = lm(speed ~ 1 + dist, data = cars)
coef(summary(fit)) # estimativa dos betas
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 8.2839056 0.87438449 9.473985 1.440974e-12
## dist        0.1655676 0.01749448 9.463990 1.489836e-12
```

```
(summary(fit)$sigma)**2 # estimativa do sigma^2
```

```
## [1] 9.958776
```

```
ggplot(cars, aes(y=speed, x=dist)) + theme_bw() +  
  geom_point() + geom_smooth(method=lm)
```



Sob a abordagem *bayesiana*, a distribuição Normal-Inversa Gama (*NIG*) é uma priori conjugada para  $\theta = (\beta, \sigma^2)$  neste modelo. Assim,

$$(\beta, \sigma^2) \sim NIG(\beta_0, V_0, a_0, b_0) .$$

Isto é,

$$\beta | \sigma^2 \sim Normal_p(\beta_0, \sigma^2 V_0) \quad ; \quad \sigma^2 \sim InvGamma(a_0, b_0)$$

ou, equivalentemente,

$$\beta \sim T_p\left(2a_0; \beta_0, \frac{b_0 V_0}{a_0}\right) \quad ; \quad \sigma^2 | \beta \sim InvGamma\left(a_0 + \frac{p}{2}, b_0 + \frac{(\beta - \beta_0)^T V_0^{-1} (\beta - \beta_0)}{2}\right) ,$$

com  $\beta_0 \in \mathbb{R}^p$ ,  $V_0$  matriz simétrica positiva definida e  $a_0, b_0 \in \mathbb{R}_+$ .

Então:

$$(\beta, \sigma^2) | Z \sim NIG(\beta_1, V_1, a_1, b_1)$$

$$\begin{aligned} \text{com } \beta_1 &= V_1 \left( V_0^{-1} \beta_0 + X^T X \hat{\beta} \right) \quad ; \quad V_1 = \left( V_0^{-1} + X^T X \right)^{-1} \quad ; \quad a_1 = a_0 + \frac{n}{2} \quad ; \\ b_1 &= b_0 + \frac{\beta_0^T V_0^{-1} \beta_0 + Y^T Y - \beta_1^T V_1^{-1} \beta_1}{2} . \end{aligned}$$

**Observação.** Uma das maneiras de representar falta de informação nesse contexto é utilizar a priori de Jeffreys,  $f(\theta) = \left| \mathcal{J}(\theta) \right|^{1/2} \propto 1/\sigma^2$ . Nesse caso, distribuição a posteriori é

$$(\beta, \sigma^2) | Z \sim \text{NIG} \left( \hat{\beta}, (X^T X)^{-1}, \frac{n-p}{2}, \frac{(n-p)s^2}{2} \right).$$

**Exemplo.** Considere que, a priori,  $(\beta, \sigma^2) \sim \text{NIG}(\beta_0, V_0, a_0, b_0)$ , com  $\beta_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ ;  $V_0 = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$ ;  $a_0 = 3$ ;  $b_0 = 100$ .

A seguir são apresentadas as distribuições marginais dos parâmetros, a distribuição marginal e as regiões HPD bivariadas do parâmetro  $\beta$ .

```
x = cars$dist      # variável resposta
y = cars$speed     # variável explicativa
n = length(x)      # n=50
X = cbind(1,x)     # Matrix de planejamento
p = ncol(X)        # p=2
g = n-p            # gl=48
beta_est = solve(t(X)%*%X)%*(t(X)%*%y) # (-17.6, 3.9)
sigma_est =
  as.double(t(y-X%*%beta_est)%*(y-X%*%beta_est)/(n-p)) #236.5
beta0 = c(0,0)      # média priori betas
V0 = matrix(c(100,0,0,100),ncol=2) # matriz de escala beta
a0 = 3              # priori sigma
b0 = 100            # priori sigma
# parâmetros da posteriori
V1 = solve(solve(V0) + t(X)%*%X)
beta1 = V1%*(solve(V0)%*%beta0 + t(X)%*%X%*%beta_est)
a1 = a0 + n/2
b1 = as.double(b0 + (t(beta0)%*%solve(V0)%*%beta0 + t(y)%*%y - t(beta1)%*%solve(V1)%*%
V = b1*V1/a1 # Matrix de escala da posteriori marginal de beta

beta1lim=c(beta1[1]-qt(0.9999,2*a1)*sqrt(V[1,1]),beta1[1]+qt(0.9999,2*a1)*sqrt(V[1,1]))
beta2lim=c(beta1[2]-qt(0.9999,2*a1)*sqrt(V[2,2]),beta1[2]+qt(0.9999,2*a1)*sqrt(V[2,2]))
sigma2lim=c(extraDistr::qinvgamma(0.0001,a1,b1),extraDistr::qinvgamma(0.9999,a1,b1))

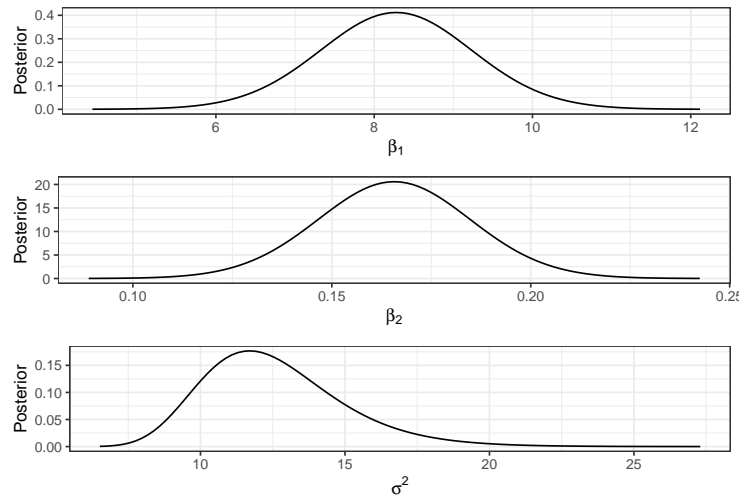
biplot <- ggplot(data.frame(x=beta1lim), aes(x=x), colour = "0.Posterior") +
  stat_function(fun = mnormt::dmt,args = list(mean = beta1[1], S = V[1,1], df=2*a1)) +
  theme_bw() + xlab(expression(beta[1])) + ylab("Posterior")
```



```

b2plot <- ggplot(data.frame(x=beta2lim), aes(x=x), colour = "0.Posterior") +
  stat_function(fun = mnormt::dmt, args = list(mean = beta1[2], S = V[2,2], df=2*a1)) +
  theme_bw() + xlab(expression(beta[2])) + ylab("Posterior")
s2plot <- ggplot(data.frame(x=sigma2lim), aes(x=x), colour = "0.Posterior")+
  stat_function(fun = extraDistr::dinvgamma, args = list(alpha = a1, beta = b1)) +
  theme_bw() + xlab(expression(sigma^2)) + ylab("Posterior")
multiplot(b1plot,b2plot,s2plot)

```



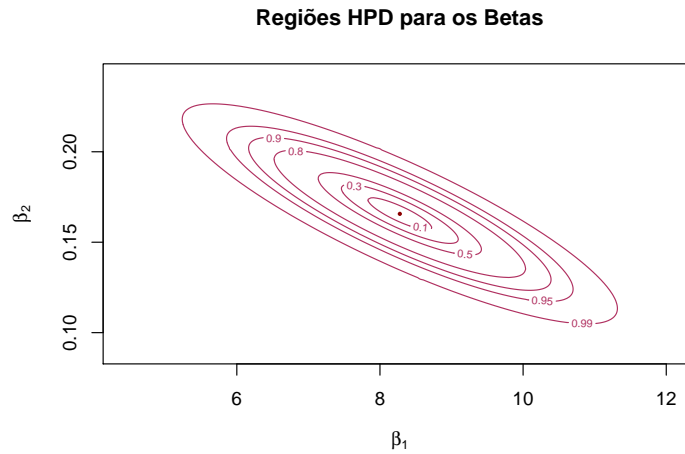
```

# posteriori marginal bivariada dos betas
posterior <- function(theta0,theta1) { apply(cbind(theta0,theta1),1,function(w){ mnormt::dmt(w, m
# Gráfico da posteriori marginal bivariada dos betas
grx <- seq(beta1lim[1], beta1lim[2],length.out=200)
gry <- seq(beta2lim[1], beta2lim[2],length.out=200)
z1 <- outer(grx,gry,posterior)
#persp(grx,gry,z1)
plotly::plot_ly(alpha=0.1) %>%
  plotly::add_surface(x=grx, y=gry, z=t(z1), colorscale = list(c(0,'#BA52ED'), c(1,'#FCB040')), s

```

WebGL is not  
supported by  
your browser  
- visit  
<https://get.webgl.org>  
for more info

```
# Curvas de Probabilidade
l = c(0.1,0.3,0.5,0.8,0.9,0.95,0.99)
z1v = sort(as.vector(z1),decreasing = TRUE)
v1 <- z1v/sum(z1v)
a=0; j=1; l1=NULL
for(i in 1:length(v1)) {
  a <- a+v1[i]
  if(j<=length(l) & a>l[j]) {
    l1 <- c(l1,z1v[i-1])
    j <- j+1
  }
}
contour(grx,gry,z1,col=colors()[455],main="Regiões HPD para os Betas",xlab=expression(
points(beta1[1],beta1[2],col="darkred",pch=16,cex=0.5)
```



### 7.6.2 Laplace's Demon

LaplacesDemon é uma biblioteca do R que oferece diversos algoritmos implementados de MCMC, permitindo fazer Inferência Bayesiana aproximada. Os algoritmos de MCMC disponíveis são

1. Automated Factor Slice Sampler (AFSS)
2. Adaptive Directional Metropolis-within-Gibbs (ADMG)
3. Adaptive Griddy-Gibbs (AGG)
4. Adaptive Hamiltonian Monte Carlo (AHMC)
5. Adaptive Metropolis (AM)
6. Adaptive Metropolis-within-Gibbs (AMWG)
7. Adaptive-Mixture Metropolis (AMM)
8. Affine-Invariant Ensemble Sampler (AIES)
9. Componentwise Hit-And-Run Metropolis (CHARM)
10. Delayed Rejection Adaptive Metropolis (DRAM)
11. Delayed Rejection Metropolis (DRM)
12. Differential Evolution Markov Chain (DEMC)
13. Elliptical Slice Sampler (ESS)
14. Gibbs Sampler (Gibbs)
15. Griddy-Gibbs (GG)
16. Hamiltonian Monte Carlo (HMC)
17. Hamiltonian Monte Carlo with Dual-Averaging (HMCDA)
18. Hit-And-Run Metropolis (HARM)

19. Independence Metropolis (IM)
20. Interchain Adaptation (INCA)
21. Metropolis-Adjusted Langevin Algorithm (MALA)
22. Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC)
23. Metropolis-within-Gibbs (MWG)
24. Multiple-Try Metropolis (MTM)
25. No-U-Turn Sampler (NUTS)
26. Oblique Hyperrectangle Slice Sampler (OHSS)
27. Preconditioned Crank-Nicolson (pCN)
28. Random Dive Metropolis-Hastings (RDMH)
29. Random-Walk Metropolis (RWM)
30. Reflective Slice Sampler (RSS)
31. Refractive Sampler (Refractive)
32. Reversible-Jump (RJ)
33. Robust Adaptive Metropolis (RAM)
34. Sequential Adaptive Metropolis-within-Gibbs (SAMWG)
35. Sequential Metropolis-within-Gibbs (SMWG)
36. Slice Sampler (Slice)
37. Stochastic Gradient Langevin Dynamics (SGLD)
38. Tempered Hamiltonian Monte Carlo (THMC)
39. t-walk (twalk)
40. Univariate Eigenvector Slice Sampler (UESS)
41. Updating Sequential Adaptive Metropolis-within-Gibbs (USAMWG)
42. Updating Sequential Metropolis-within-Gibbs (USMWG)

<https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/BayesianInference.pdf>

<https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/LaplacesDemonTutorial.pdf>

<https://cran.r-project.org/web/packages/LaplacesDemon/vignettes/Examples.pdf>

<https://cran.r-project.org/web/packages/LaplacesDemon/LaplacesDemon.pdf>

- Especificação do modelo

```
require(LaplacesDemon)

parm.names=as.parm.names(list(beta=rep(0,p), sigma2=0))
pos.beta=grep("beta", parm.names)
pos.sigma=grep("sigma2", parm.names)

MyData <- list(J=p, X=X, y=y, mon.names="LP",
```

```

    parm.names=parm.names,pos.beta=pos.beta,pos.sigma=pos.sigma)

Model <- function(parm, Data)
{
  ### Parameters
  beta <- parm[Data$pos.beta]
  sigma2 <- interval(parm[Data$pos.sigma], 1e-100, Inf)
  parm[Data$pos.sigma] <- sigma2
  ### Log-Prior
  sigma.prior <- dinvgamma(sigma2, a0, b0, log=TRUE)
  beta.prior <- dmvn(beta, beta0, sigma2*V0, log=TRUE)
  ### Log-Likelihood
  mu <- tcrossprod(Data$X, t(beta))
  LL <- sum(dnormv(Data$y, mu, sigma2, log=TRUE))
  ### Log-Posterior
  LP <- LL + beta.prior + sigma.prior
  Modelout <- list(LP=LP, Dev=-2*LL, Monitor=LP,
                  yhat=rnorm(length(mu), mu, sigma2), parm=parm)
  return(Modelout)
}

Initial.Values <- c(beta_est,sigma_est)

burnin <- 2000
thin <- 3
N=(2000+burnin)*thin

```

---

• **Exemplo 1:** Metropolis-within-Gibbs (MWG)

```

set.seed(666)
Fit1 <- LaplacesDemon(Model, Data=MyData, Initial.Values,
  Covar=NULL, Iterations=N, Status=N/5, Thinning=thin,
  Algorithm="MWG", Specs=NULL)

##
## Laplace's Demon was called on Tue Nov 03 14:52:16 2020
##
## Performing initial checks...
## Algorithm: Metropolis-within-Gibbs
##
## Laplace's Demon is beginning to update...
## Iteration: 2400, Proposal: Componentwise, LP: -146.2

```

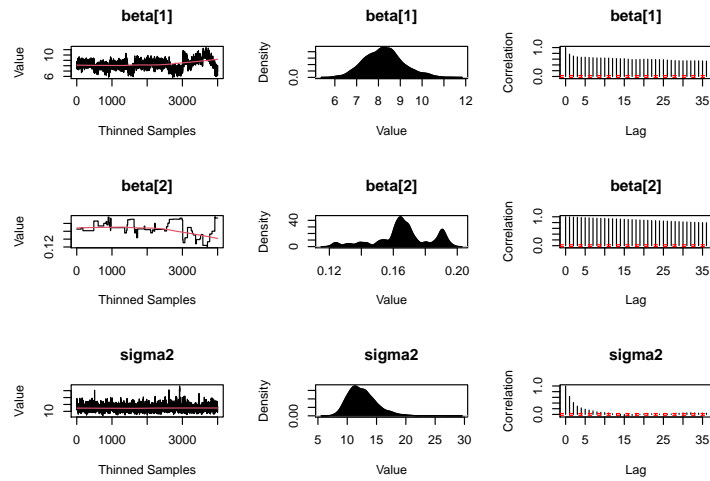
```
## Iteration: 4800, Proposal: Componentwise, LP: -147.2
## Iteration: 7200, Proposal: Componentwise, LP: -142.7
## Iteration: 9600, Proposal: Componentwise, LP: -143.5
## Iteration: 12000, Proposal: Componentwise, LP: -145.1
##
## Assessing Stationarity
## Assessing Thinning and ESS
## Creating Summaries
## Estimating Log of the Marginal Likelihood
## Creating Output
##
## Laplace's Demon has finished.
```

```
#names(Fit1)
print(Fit1)
```

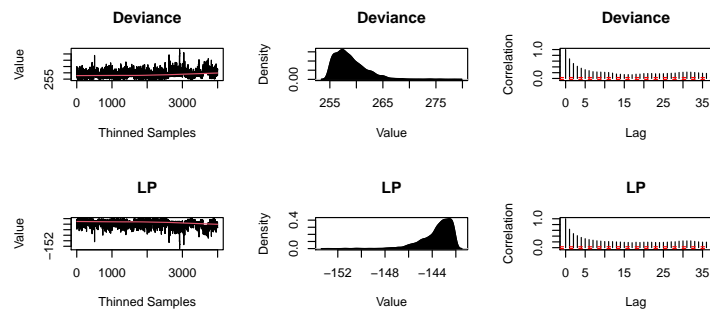
```
## Call:
## LaplacesDemon(Model = Model, Data = MyData, Initial.Values = Initial.Values,
## Covar = NULL, Iterations = N, Status = N/5, Thinning = thin,
## Algorithm = "MWG", Specs = NULL)
##
## Acceptance Rate: 0.35375
## Algorithm: Metropolis-within-Gibbs
## Covariance Matrix: (NOT SHOWN HERE; diagonal shown instead)
## beta[1] beta[2] sigma2
## 1.890044 1.890044 1.890044
##
## Covariance (Diagonal) History: (NOT SHOWN HERE)
## Deviance Information Criterion (DIC):
## All Stationary
## Dbar 258.781 259.440
## pD 4.338 4.237
## DIC 263.119 263.677
## Initial Values:
## [1] 8.2839056 0.1655676 9.9587760
##
## Iterations: 12000
## Log(Marginal Likelihood): -126.8415
## Minutes of run-time: 0.3
## Model: (NOT SHOWN HERE)
## Monitor: (NOT SHOWN HERE)
## Parameters (Number of): 3
## Posterior1: (NOT SHOWN HERE)
## Posterior2: (NOT SHOWN HERE)
## Recommended Burn-In of Thinned Samples: 3200
```

```
## Recommended Burn-In of Un-thinned Samples: 9600
## Recommended Thinning: 36
## Specs: (NOT SHOWN HERE)
## Status is displayed every 2400 iterations
## Summary1: (SHOWN BELOW)
## Summary2: (SHOWN BELOW)
## Thinned Samples: 4000
## Thinning: 3
##
##
## Summary of All Samples
##           Mean           SD           MCSE           ESS           LB
## beta[1]      8.2478022 0.08164035 0.094717679 24.80187 6.6527588
## beta[2]      0.1665176 0.01702535 0.003042654 12.48151 0.1240373
## sigma2     12.5840386 2.48412045 0.125660593 830.27788 8.7682617
## Deviance    258.7813640 2.94533567 0.202218161 136.68013 255.0038715
## LP          -143.5465138 1.23958817 0.087107901 109.45812 -146.5361533
##           Median           UB
## beta[1]      8.2225293 10.1866171
## beta[2]      0.1655676 0.1937371
## sigma2     12.2724404 18.2059419
## Deviance    258.1752506 265.5673438
## LP          -143.2451510 -142.1163971
##
##
## Summary of Stationary Samples
##           Mean           SD           MCSE           ESS           LB
## beta[1]      8.9659139 1.01640981 0.264577460 6.932756 6.5707408
## beta[2]      0.1498047 0.02004864 0.007080864 3.675837 0.1240373
## sigma2     12.4235537 2.25114323 0.223124198 148.522644 8.9164125
## Deviance    259.4399279 2.91101586 0.562252817 35.086016 255.3738557
## LP          -143.8933516 1.24169010 0.191055588 27.276956 -146.8229317
##           Median           UB
## beta[1]      9.0304813 10.6078726
## beta[2]      0.1504279 0.1937371
## sigma2     12.0978375 17.6632608
## Deviance    258.9769661 265.8590137
## LP          -143.6581852 -142.1644366
```

```
Post1 <- data.frame(Fit1$Posterior1, Algorithm="1.MWG")
colnames(Post1) <- c("beta1", "beta2", "sigma2", "Algorithm")
#head(Post1)
plot(Fit1, BurnIn=0, MyData, PDF=FALSE, Params=NULL)
```



```
#plot(Fit1, BurnIn=burnin, MyData, PDF=FALSE, Params=NULL)
```




---

- **Exemplo 2:** Adaptative Metropolis-within-Gibbs (MWG)

```
set.seed(666)
Fit2 <- LaplacesDemon(Model, Data=MyData, Initial.Values,
  Covar=NULL, Iterations=N, Status=N/5, Thinning=thin,
  Algorithm="AMWG", Specs=NULL)
```

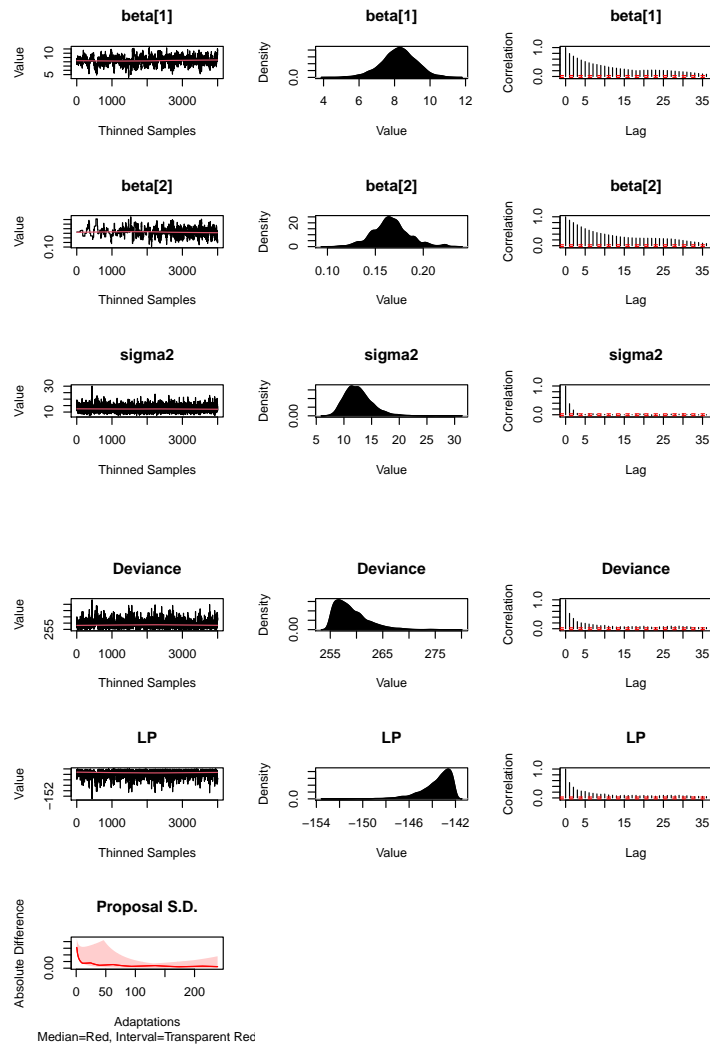


```
##
## Laplace's Demon was called on Tue Nov 03 14:52:38 2020
##
## Performing initial checks...
## Algorithm: Adaptive Metropolis-within-Gibbs
##
## Laplace's Demon is beginning to update...
## Iteration: 2400, Proposal: Componentwise, LP: -145.3
## Iteration: 4800, Proposal: Componentwise, LP: -143
## Iteration: 7200, Proposal: Componentwise, LP: -144.1
## Iteration: 9600, Proposal: Componentwise, LP: -143.1
## Iteration: 12000, Proposal: Componentwise, LP: -146.2
##
## Assessing Stationarity
## Assessing Thinning and ESS
## Creating Summaries
## Creating Output
##
## Laplace's Demon has finished.
```

```
#names(Fit2)
#print(Fit2)

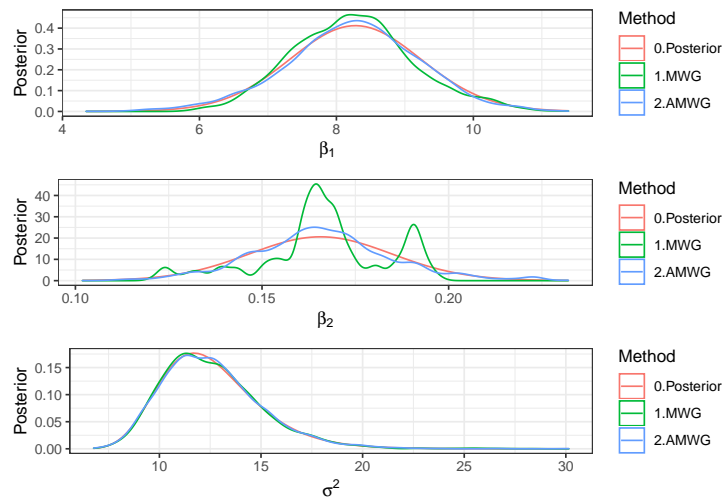
Post2 <- data.frame(Fit2$Posterior1, Algorithm="2.AMWG")
colnames(Post2) <- c("beta1", "beta2", "sigma2", "Algorithm")
#head(Post2)

#plot(Fit2, BurnIn=burnin, MyData, PDF=FALSE, Params=NULL)
plot(Fit2, BurnIn=0, MyData, PDF=FALSE, Params=NULL)
```



```
Post <- rbind(Post1,Post2)
b1plot <- ggplot(Post) +
  stat_function(aes(colour="0.Posterior"), fun = mnormt::dmt,args = list(mean = beta1[
  geom_density(aes(beta1, colour = Algorithm)) + theme_bw() +
  xlab(expression(beta[1])) + ylab("Posterior") + labs(colour = "Method")
b2plot <- ggplot(Post) +
  stat_function(aes(colour="0.Posterior"), fun = mnormt::dmt, args = list(mean = beta1
  geom_density(aes(beta2, colour = Algorithm)) + theme_bw() +
  xlab(expression(beta[2])) + ylab("Posterior") + labs(colour = "Method")
s2plot <- ggplot(Post)+
  stat_function(aes(colour="0.Posterior"), fun = extraDistr::dinvgamma, args = list(al
```

```
geom_density(aes(sigma2, colour = Algorithm)) + theme_bw() +
  xlab(expression(sigma^2)) + ylab("Posterior") + labs(colour = "Method")
multiplot(b1plot,b2plot,s2plot)
```



- **Exemplo 3:** Consort e Automated Factor Slice Sampler (AFSS)

```
Consort(Fit2)
```

```
##
## #####
## # Consort with Laplace's Demon #
## #####
## Call:
## LaplacesDemon(Model = Model, Data = MyData, Initial.Values = Initial.Values,
##   Covar = NULL, Iterations = N, Status = N/5, Thinning = thin,
##   Algorithm = "AMWG", Specs = NULL)
##
## Acceptance Rate: 0.34264
## Algorithm: Adaptive Metropolis-within-Gibbs
## Covariance Matrix: (NOT SHOWN HERE; diagonal shown instead)
##   beta[1]   beta[2]   sigma2
## 1.17679178 0.02838563 9.45831955
##
## Covariance (Diagonal) History: (NOT SHOWN HERE)
## Deviance Information Criterion (DIC):
```

```

##           All Stationary
## Dbar 258.957    258.957
## pD    4.957     4.957
## DIC  263.915    263.915
## Initial Values:
## [1] 8.2839056 0.1655676 9.9587760
##
## Iterations: 12000
## Log(Marginal Likelihood): NA
## Minutes of run-time: 0.3
## Model: (NOT SHOWN HERE)
## Monitor: (NOT SHOWN HERE)
## Parameters (Number of): 3
## Posterior1: (NOT SHOWN HERE)
## Posterior2: (NOT SHOWN HERE)
## Recommended Burn-In of Thinned Samples: 0
## Recommended Burn-In of Un-thinned Samples: 0
## Recommended Thinning: 105
## Specs: (NOT SHOWN HERE)
## Status is displayed every 2400 iterations
## Summary1: (SHOWN BELOW)
## Summary2: (SHOWN BELOW)
## Thinned Samples: 4000
## Thinning: 3
##
##
## Summary of All Samples
##           Mean          SD          MCSE          ESS          LB
## beta[1]    8.2386566 0.97788924 0.085914819 151.0105    6.1651888
## beta[2]    0.1661183 0.01903241 0.002099194 209.2080    0.1294318
## sigma2    12.5980978 2.45984860 0.071949821 1784.2000    8.7806779
## Deviance  258.9573762 3.14870026 0.163068368 595.3985   255.1347857
## LP        -143.6311506 1.33506450 0.072413145 530.5654  -147.0582460
##           Median          UB
## beta[1]    8.2641629    10.1040670
## beta[2]    0.1655676     0.2078131
## sigma2    12.3274919    18.1748415
## Deviance  258.2440979 266.8653177
## LP        -143.2679823 -142.1324005
##
##
## Summary of Stationary Samples
##           Mean          SD          MCSE          ESS          LB
## beta[1]    8.2386566 0.97788924 0.085914819 151.0105    6.1651888
## beta[2]    0.1661183 0.01903241 0.002099194 209.2080    0.1294318
## sigma2    12.5980978 2.45984860 0.071949821 1784.2000    8.7806779

```

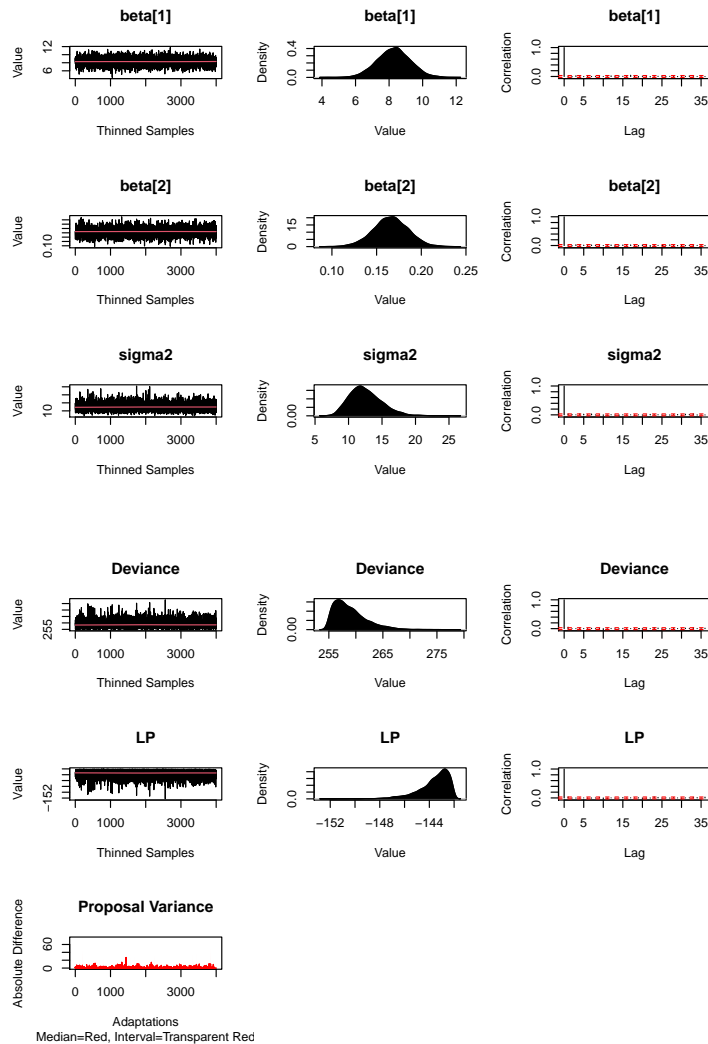
```
## Deviance 258.9573762 3.14870026 0.163068368 595.3985 255.1347857
## LP -143.6311506 1.33506450 0.072413145 530.5654 -147.0582460
##           Median          UB
## beta[1] 8.2641629 10.1040670
## beta[2] 0.1655676 0.2078131
## sigma2 12.3274919 18.1748415
## Deviance 258.2440979 266.8653177
## LP -143.2679823 -142.1324005
##
## Demonic Suggestion
##
## Due to the combination of the following conditions,
##
## 1. Adaptive Metropolis-within-Gibbs
## 2. The acceptance rate (0.3426389) is within the interval [0.15,0.5].
## 3. At least one target MCSE is >= 6.27% of its marginal posterior
##    standard deviation.
## 4. Each target distribution has an effective sample size (ESS)
##    of at least 100.
## 5. Each target distribution became stationary by
##    1 iteration.
##
## Quantiles of Absolute Posterior1 Correlation:
##           0%          25%          50%          75%          100%
## 0.02856704 0.03170699 0.85367854 1.00000000 1.00000000
##
## Possibly excessive posterior correlation for a componentwise algorithm.
##
## WARNING: Diminishing adaptation did not occur.
##           A new algorithm will be suggested.
##
## Laplace's Demon has not been appeased, and suggests
## copy/pasting the following R code into the R console,
## and running it.
##
## Initial.Values <- as.initial.values(Fit2)
## Fit2 <- LaplacesDemon(Model, Data=MyData, Initial.Values,
##           Covar=Fit2$Covar, Iterations=420000, Status=40000, Thinning=105,
##           Algorithm="AFSS", Specs=list(A=Inf, B=NULL, m=100,
##           n=0, w=1))
##
## Laplace's Demon is finished consorting.
```

```
Initial.Values <- as.initial.values(Fit2)
```

```
set.seed(666)
Fit3 <- LaplacesDemon(Model, Data=MyData, Initial.Values,
  Covar=NULL, Iterations=N, Status=N/5, Thinning=thin,
  Algorithm="AFSS", Specs=list(A=Inf, B=NULL, m=100, n=0, w=1))
```

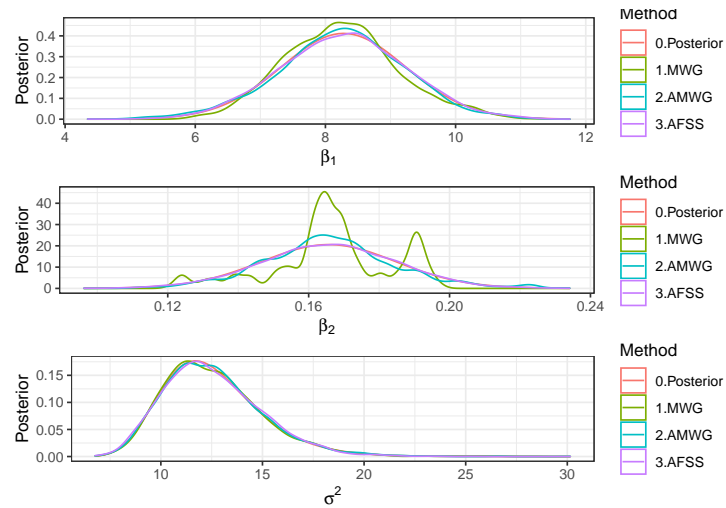
```
##
## Laplace's Demon was called on Tue Nov 03 14:53:24 2020
##
## Performing initial checks...
## Algorithm: Automated Factor Slice Sampler
##
## Laplace's Demon is beginning to update...
##
## Eigendecomposition will occur every 120 iterations.
##
## Iteration: 2400,   Proposal: Multivariate,   LP: -142.1
## Iteration: 4800,   Proposal: Multivariate,   LP: -144.2
## Iteration: 7200,   Proposal: Multivariate,   LP: -145.2
## Iteration: 9600,   Proposal: Multivariate,   LP: -144.1
## Iteration: 12000,  Proposal: Multivariate,   LP: -142.5
##
## Assessing Stationarity
## Assessing Thinning and ESS
## Creating Summaries
## Estimating Log of the Marginal Likelihood
## Creating Output
##
## Laplace's Demon has finished.
```

```
Post3 <- data.frame(Fit3$Posterior1, Algorithm="3.AFSS")
colnames(Post3) <- c("beta1", "beta2", "sigma2", "Algorithm")
plot(Fit3, BurnIn=0, MyData, PDF=FALSE, Params=NULL)
```

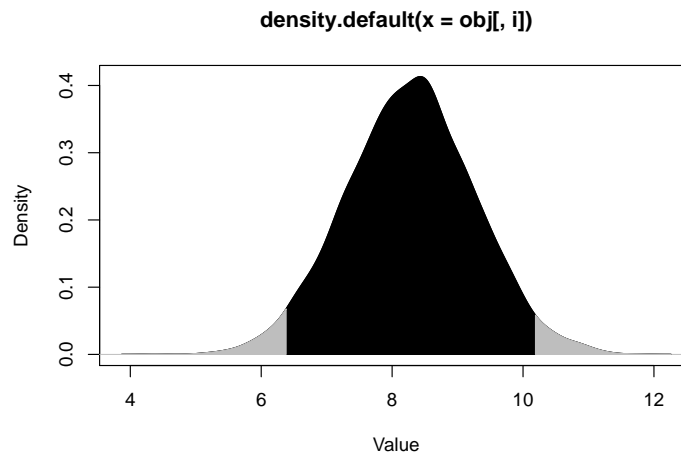


```
Post <- rbind(Post1,Post2,Post3)
b1plot <- ggplot(Post) +
  stat_function(aes(colour="0.Posterior"), fun = mnormt::dmt, args = list(mean = beta1[1], S = V[1,1])) +
  geom_density(aes(beta1, colour = Algorithm)) + theme_bw() +
  xlab(expression(beta[1])) + ylab("Posterior") + labs(colour = "Method")
b2plot <- ggplot(Post) +
  stat_function(aes(colour="0.Posterior"), fun = mnormt::dmt, args = list(mean = beta1[2], S = V[1,2])) +
  geom_density(aes(beta2, colour = Algorithm)) + theme_bw() +
  xlab(expression(beta[2])) + ylab("Posterior") + labs(colour = "Method")
s2plot <- ggplot(Post)+
  stat_function(aes(colour="0.Posterior"), fun = extraDistr::dinvgamma, args = list(alpha = a1, b1 = b1)) +
  geom_density(aes(sigma2, colour = Algorithm)) + theme_bw() +
  xlab(expression(sigma^2)) + ylab("Posterior") + labs(colour = "Method")
```

```
geom_density(aes(sigma2, colour = Algorithm)) + theme_bw() +
  xlab(expression(sigma^2)) + ylab("Posterior") + labs(colour = "Method")
multiplot(b1plot,b2plot,s2plot)
```



```
p.interval(Post3$beta1, HPD=FALSE, MM=FALSE, plot=TRUE)
```



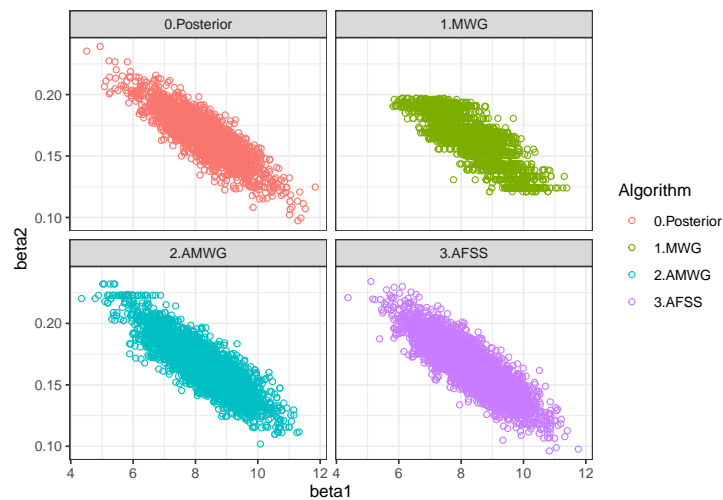
```
##           Lower  Upper
## [1,]  6.383021 10.175
## attr("Probability.Interval")
## [1] 0.95
```



```

set.seed(666)
S0 <- as.matrix(extraDistr::rinvgamma(N/thin-burnin,a1,b1))
M0 <- apply(S0,1,function(s){mnormt::rmnorm(1,mean=beta1,varcov=s*V1)})
Post0 <- data.frame(t(M0),S0,"0.Posterior")
colnames(Post0) <- c("beta1","beta2","sigma2","Algorithm")
Post = rbind(Post0,Post1,Post2,Post3)
ggplot(Post) + theme_bw() +
  geom_point(aes(beta1,beta2,colour=Algorithm), shape=1) +
  facet_wrap(Algorithm ~ .)

```



### 7.6.3 Stan

O Stan é uma plataforma de modelagem estatística de alto desempenho. Em particular, permite fazer inferência bayesiana usando o método de Monte Carlo Hamiltoniano (HMC) e a variação No-U-Turn Sampler (NUTS). Esses recursos convergem para distribuições alvo de altas dimensões muito mais rapidamente que métodos mais simples, como o amostrador de Gibbs ou outras variações do método de Metropolis-Hastings. A linguagem utilizada é independente da plataforma e existem bibliotecas para R (**rstan**) e Python.

<https://mc-stan.org/>

**Voltando ao Exemplo**

```

library(rstan)
# Parametros do método
Initial.Values <- c(beta_est,sigma_est)
burnin <- 2000
thin <- 3
N=(2000+burnin)*thin
# Conjunto de dados
stan_data <- list(N = n, J = p, y = y, x = X)

# Especificação do modelo
rs_code <- '
  data {
    int<lower=1> N;
    int<lower=1> J;
    matrix[N,J] x;
    vector[N] y;
  }
  parameters {
    vector[J] beta;
    real<lower=0> sigma2;
  }
  model {
    sigma2 ~ inv_gamma(3, 100);
    beta ~ normal(0, sqrt(sigma2*100));
    y ~ normal(x * beta, sqrt(sigma2));
  }
'

stan_mod <- stan(model_code = rs_code, data = stan_data, init=Initial.Values,
                 chains = 1, iter = N, warmup = burnin, thin = thin)

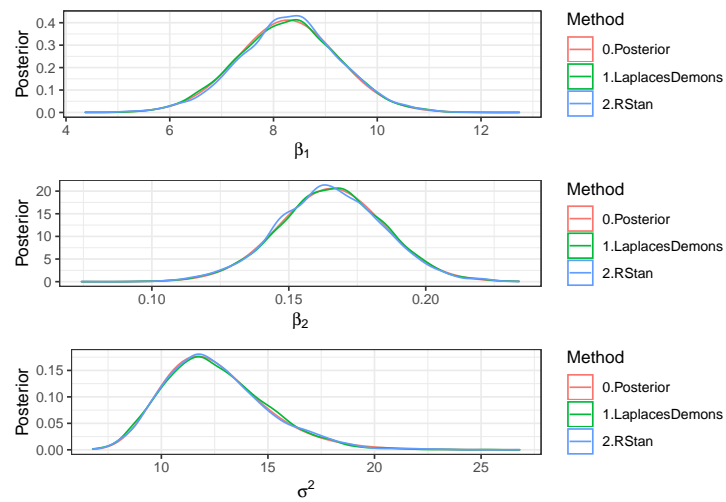
##
## SAMPLING FOR MODEL '6b158d2712aa5479f81bd408884c9453' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:      1 / 12000 [  0%] (Warmup)
## Chain 1: Iteration:   1200 / 12000 [ 10%] (Warmup)
## Chain 1: Iteration:   2001 / 12000 [ 16%] (Sampling)
## Chain 1: Iteration:   3200 / 12000 [ 26%] (Sampling)
## Chain 1: Iteration:   4400 / 12000 [ 36%] (Sampling)
## Chain 1: Iteration:   5600 / 12000 [ 46%] (Sampling)
## Chain 1: Iteration:   6800 / 12000 [ 56%] (Sampling)

```

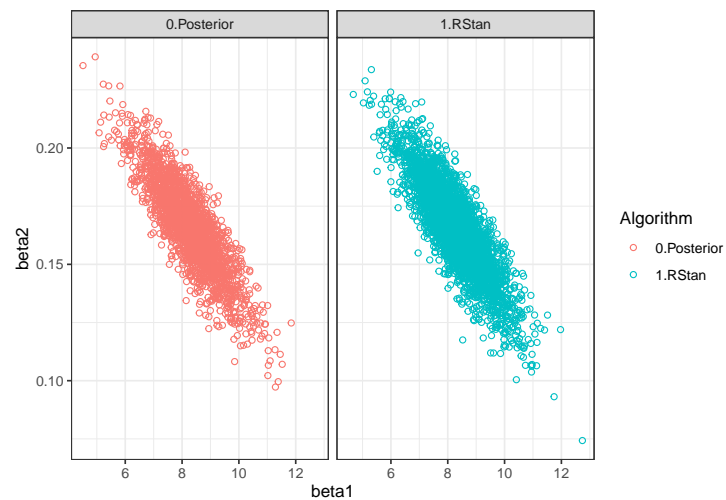
```
## Chain 1: Iteration: 8000 / 12000 [ 66%] (Sampling)
## Chain 1: Iteration: 9200 / 12000 [ 76%] (Sampling)
## Chain 1: Iteration: 10400 / 12000 [ 86%] (Sampling)
## Chain 1: Iteration: 11600 / 12000 [ 96%] (Sampling)
## Chain 1: Iteration: 12000 / 12000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 0.289 seconds (Warm-up)
## Chain 1: 1.451 seconds (Sampling)
## Chain 1: 1.74 seconds (Total)
## Chain 1:
```

```
posterior <- rstan::extract(stan_mod)
Post4 <- data.frame(posterior$beta[,1],posterior$beta[,2],posterior$sigma2,Algorithm="4.RStan")
colnames(Post4) <- c("beta1","beta2","sigma2","Algorithm")

# gráficos posterioris marginais
Post <- rbind(Post3,Post4) %>%
  mutate(Algorithm=ifelse(Algorithm=="3.AFSS","1.LaplaceDemos","2.RStan"))
b1plot <- ggplot(Post) +
  stat_function(aes(colour="0.Posterior"), fun = mnormt::dmt,args = list(mean = beta1[1], S = V[1,1])) +
  geom_density(aes(beta1, colour = Algorithm)) + theme_bw() +
  xlab(expression(beta[1])) + ylab("Posterior") + labs(colour = "Method")
b2plot <- ggplot(Post) +
  stat_function(aes(colour="0.Posterior"), fun = mnormt::dmt, args = list(mean = beta1[2], S = V[1,2])) +
  geom_density(aes(beta2, colour = Algorithm)) + theme_bw() +
  xlab(expression(beta[2])) + ylab("Posterior") + labs(colour = "Method")
s2plot <- ggplot(Post)+
  stat_function(aes(colour="0.Posterior"), fun = extraDistr::dinvgamma, args = list(alpha = a1, b = b1)) +
  geom_density(aes(sigma2, colour = Algorithm)) + theme_bw() +
  xlab(expression(sigma^2)) + ylab("Posterior") + labs(colour = "Method")
multiplot(b1plot,b2plot,s2plot)
```



```
Post = rbind(Post0,Post4) %>%
  mutate(Algorithm=ifelse(Algorithm=="4.RStan","1.RStan","0.Posterior"))
ggplot(Post) + theme_bw() +
  geom_point(aes(beta1,beta2,colour=Algorithm), shape=1) +
  facet_wrap(Algorithm ~ .)
```



```
stan_mod2 <- stan(model_code = rs_code, data = stan_data, init=Initial.Values,
  chains = 2, iter = N, warmup = burnin, thin = thin)
```

```

##
## SAMPLING FOR MODEL '6b158d2712aa5479f81bd408884c9453' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:      1 / 12000 [ 0%] (Warmup)
## Chain 1: Iteration:    1200 / 12000 [ 10%] (Warmup)
## Chain 1: Iteration:    2001 / 12000 [ 16%] (Sampling)
## Chain 1: Iteration:    3200 / 12000 [ 26%] (Sampling)
## Chain 1: Iteration:    4400 / 12000 [ 36%] (Sampling)
## Chain 1: Iteration:    5600 / 12000 [ 46%] (Sampling)
## Chain 1: Iteration:    6800 / 12000 [ 56%] (Sampling)
## Chain 1: Iteration:    8000 / 12000 [ 66%] (Sampling)
## Chain 1: Iteration:    9200 / 12000 [ 76%] (Sampling)
## Chain 1: Iteration:   10400 / 12000 [ 86%] (Sampling)
## Chain 1: Iteration:   11600 / 12000 [ 96%] (Sampling)
## Chain 1: Iteration:   12000 / 12000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 0.278 seconds (Warm-up)
## Chain 1:                0.999 seconds (Sampling)
## Chain 1:                1.277 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL '6b158d2712aa5479f81bd408884c9453' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:      1 / 12000 [ 0%] (Warmup)
## Chain 2: Iteration:    1200 / 12000 [ 10%] (Warmup)
## Chain 2: Iteration:    2001 / 12000 [ 16%] (Sampling)
## Chain 2: Iteration:    3200 / 12000 [ 26%] (Sampling)
## Chain 2: Iteration:    4400 / 12000 [ 36%] (Sampling)
## Chain 2: Iteration:    5600 / 12000 [ 46%] (Sampling)
## Chain 2: Iteration:    6800 / 12000 [ 56%] (Sampling)
## Chain 2: Iteration:    8000 / 12000 [ 66%] (Sampling)
## Chain 2: Iteration:    9200 / 12000 [ 76%] (Sampling)
## Chain 2: Iteration:   10400 / 12000 [ 86%] (Sampling)
## Chain 2: Iteration:   11600 / 12000 [ 96%] (Sampling)
## Chain 2: Iteration:   12000 / 12000 [100%] (Sampling)
## Chain 2:

```

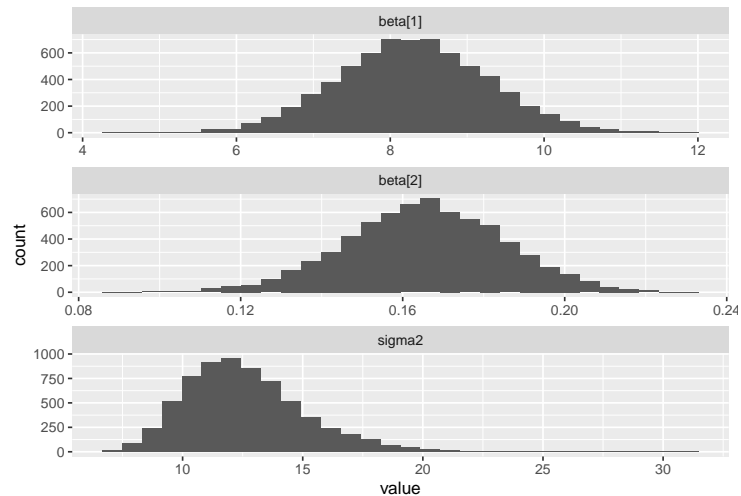
```
## Chain 2: Elapsed Time: 0.235 seconds (Warm-up)
## Chain 2: 1.011 seconds (Sampling)
## Chain 2: 1.246 seconds (Total)
## Chain 2:
```

```
library(ggmcmc)
#ggs(stan_mod2) %>% ggmcmc(., file = "ggmcmc.html")
```

### Gráficos ggmcmc

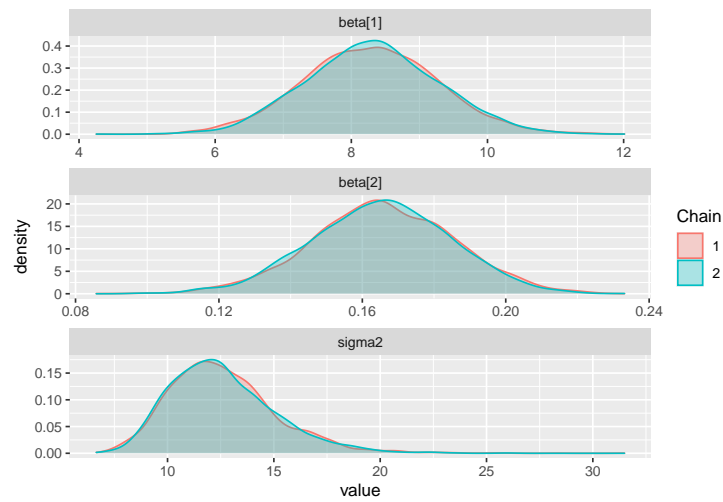
1. *Histograma* com as cadeias geradas combinadas.

```
ggs(stan_mod2) %>% ggs_histogram(.)
```



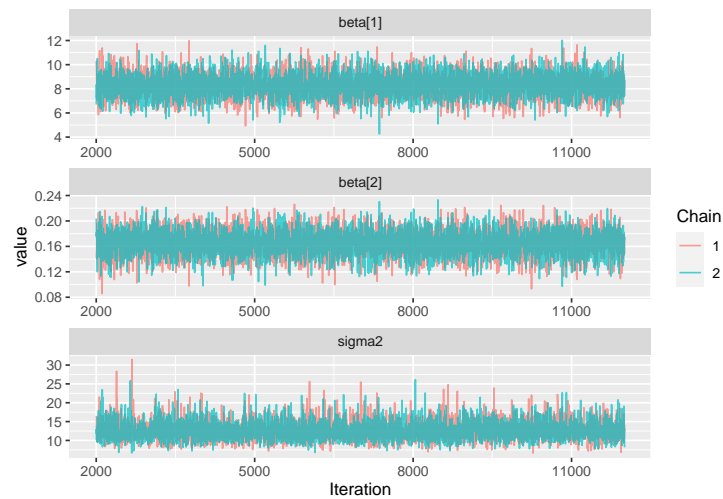
2. *Gráficos das Densidades* sobrepostos com cores diferentes por cadeia, permite comparar se as cadeias convergiram para distribuições semelhantes.

```
ggs(stan_mod2) %>% ggs_density(.)
```



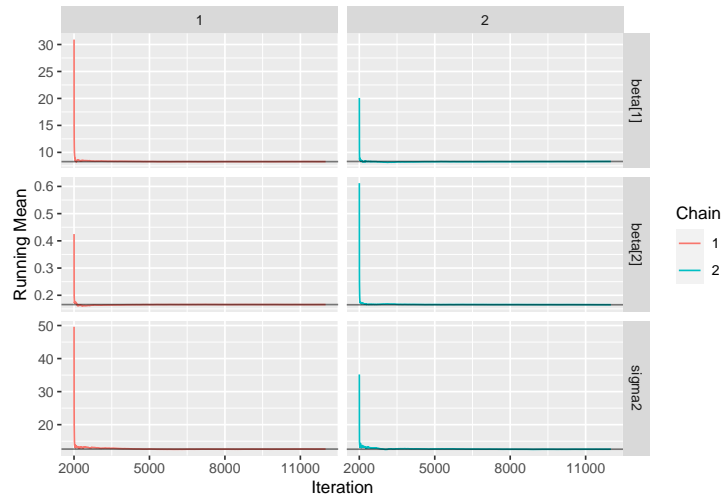
3. *Séries Temporais* das cadeias geradas. É esperado que as cadeias geradas apresentem distribuições semelhantes em torno de uma mesma média, indicando assim que atingiu-se a “estacionariedade”.

```
ggs(stan_mod2) %>% ggs_traceplot(.)
```



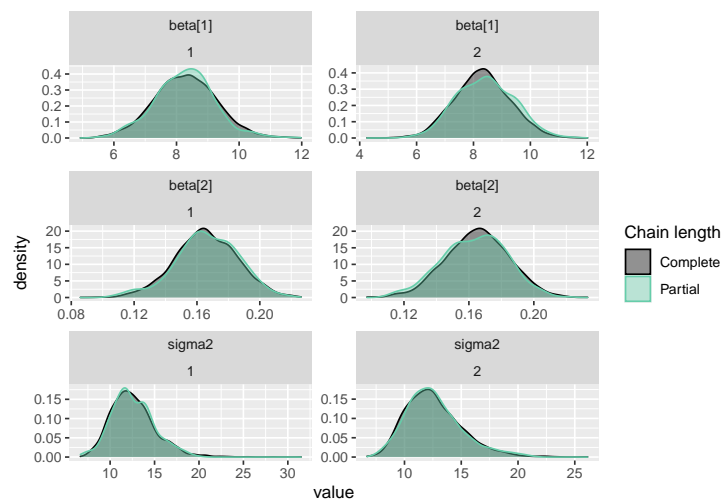
4. *Gráfico de Médias Móveis*. É esperado que as curvas das médias das cadeias geradas se aproximem rapidamente de um mesmo valor.

```
ggs(stan_mod2) %>% ggs_running(.)
```



5. *Densidades parcial e completa sobrepostas.* Compara a última parte da cadeia (por padrão, os últimos 10% dos valores, em verde) com a cadeia inteira (em preto). Idealmente, as partes inicial e final da cadeia devem ser amostradas na mesma distribuição alvo, de modo que as densidades sobrepostas devem ser semelhantes.

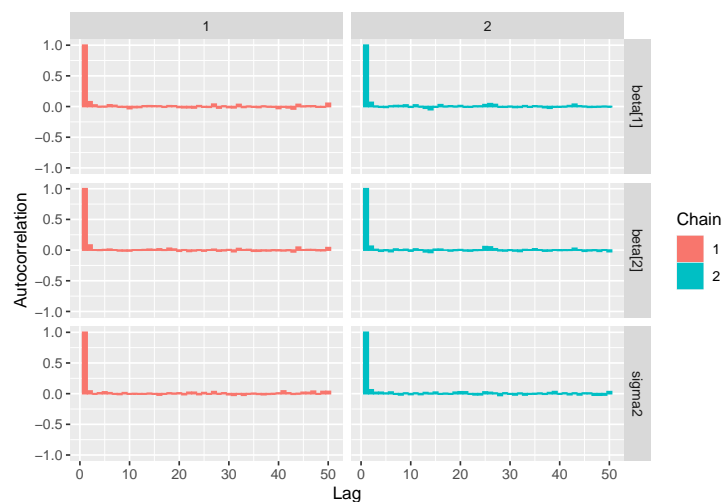
```
ggs(stan_mod2) %>% ggs_compare_partial(.)
```





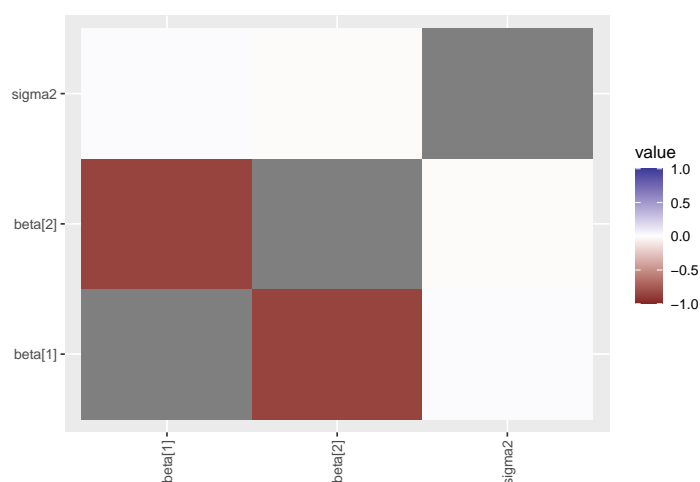
6. *Gráfico de Autocorrelação.* Espera-se alta correlação apenas no primeiro lag. Quando há um comportamento diferente do esperado, deve-se aumentar o tamanho dos saltos (*thin*) entre as observações da cadeia gerada que serão consideradas na amostra final.

```
ggs(stan_mod2) %>% ggs_autocorrelation(.)
```

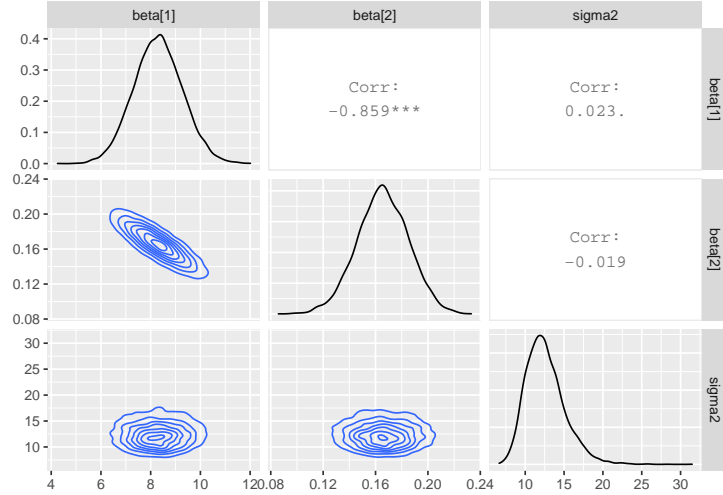


6. *Correlação Cruzada.* Quando há alta correlação entre os parâmetros é possível que a convergência da cadeia seja mais lenta.

```
ggs(stan_mod2) %>% ggs_crosscorrelation(.)
```



```
ggs(stan_mod2) %>% ggs_pairs(., lower = list(continuous = "density"))
```



#### 7.6.4 MLG

O modelos lineares generalizados (MLG) são uma extensão natural dos modelos lineares para casos em que a distribuição da variável resposta não é normal. Como exemplo, vamos considerar o particular caso onde a resposta é binária, conhecido como *regressão logística*.

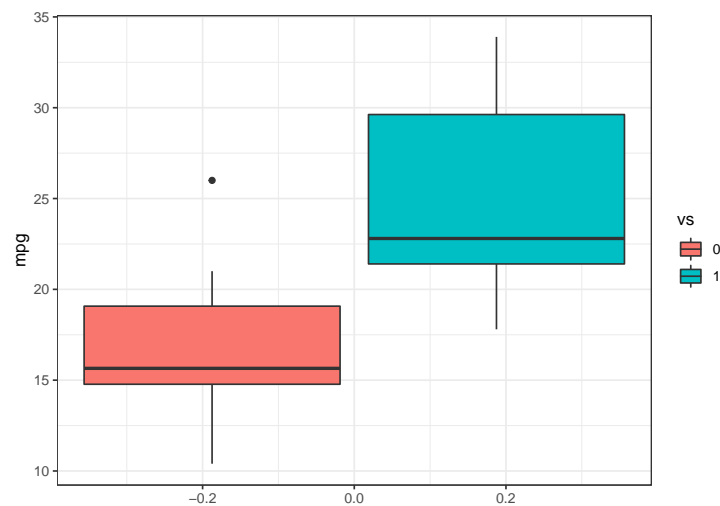
Considere  $Y_1, \dots, Y_n$  condicionalmente independentes tais que  $Y_i | \theta_i \sim \text{Ber}(\theta_i)$ , em que  $\theta_i$  é tal que  $\log\left(\frac{\theta_i}{1 - \theta_i}\right) = x_i' \beta$  ou, em outras palavras,  $\theta_i = \frac{1}{1 + e^{-x_i' \beta}} = \frac{e^{x_i' \beta}}{1 + e^{x_i' \beta}}$  com  $x_i$  as covariáveis da  $i$ -ésima observação e o vetor de parâmetros  $\beta = (\beta_1, \dots, \beta_p)$ .

**Exemplo.** Considere as variáveis *vs* (0 = motor em forma de V, 1 = motor reto) e *mpg* (milhas/galão(EUA)) do conjunto de dados *mtcars* do R. Suponha um modelo de regressão logística para a variável resposta *vs* com a covariável *mpg* em que, a priori,  $\beta_i \sim \text{Laplace}(0, b_i)$ ,  $i = 1, 2$ , independentes. Deste modo, a posteriori é dada por

$$f(\beta|y, x) \propto f(y|\beta, x)f(\beta) \propto \prod_{i=1}^n \left( \frac{1}{1 + e^{x'_i \beta}} \right)^{y_i} \left( \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}} \right)^{1-y_i} \prod_{j=1}^p \frac{1}{2b_j} e^{-\frac{|\beta_j|}{b_j}}.$$

```
library(rstan)

dados <- as_tibble(mtcars)
# mpg: Miles/(US)gallon ; vs: Engine(0=V-shaped,1=straight)
dados %>% ggplot(aes(group=as.factor(vs), y=mpg, fill=as.factor(vs))) +
  geom_boxplot() + scale_fill_discrete(name="vs") + theme_bw()
```



```
y <- dados %>% select(vs) %>% pull()
x <- dados %>% select(mpg) %>% pull()
n <- length(y)
X <- as.matrix(cbind(1,x))
p <- ncol(X)

stan_data <- list(N = n, J = p, y = y, x = X)

rs_code <- '
data {
  int<lower=1> N;
  int<lower=1> J;
  int<lower=0,upper=1> y[N];
  matrix[N,J] x;
}
parameters {
  vector[J] beta;
```

```

    }
    model {
      beta ~ double_exponential(0, 100);
      y ~ bernoulli_logit(x * beta);
    }
  }

N=2000
thin=10
burnin=1000

stan_log <- stan(model_code = rs_code, data = stan_data, init = c(0,0),
  chains = 1, iter = N*thin, warmup = burnin, thin = thin)

```

```

##
## SAMPLING FOR MODEL '6525180a46ff3a51f06bbb110d962a1c' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:      1 / 20000 [  0%] (Warmup)
## Chain 1: Iteration:  1001 / 20000 [  5%] (Sampling)
## Chain 1: Iteration:  3000 / 20000 [ 15%] (Sampling)
## Chain 1: Iteration:  5000 / 20000 [ 25%] (Sampling)
## Chain 1: Iteration:  7000 / 20000 [ 35%] (Sampling)
## Chain 1: Iteration:  9000 / 20000 [ 45%] (Sampling)
## Chain 1: Iteration: 11000 / 20000 [ 55%] (Sampling)
## Chain 1: Iteration: 13000 / 20000 [ 65%] (Sampling)
## Chain 1: Iteration: 15000 / 20000 [ 75%] (Sampling)
## Chain 1: Iteration: 17000 / 20000 [ 85%] (Sampling)
## Chain 1: Iteration: 19000 / 20000 [ 95%] (Sampling)
## Chain 1: Iteration: 20000 / 20000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 0.232 seconds (Warm-up)
## Chain 1:           3.101 seconds (Sampling)
## Chain 1:           3.333 seconds (Total)
## Chain 1:

```

```
print(stan_log)
```

```

## Inference for Stan model: 6525180a46ff3a51f06bbb110d962a1c.
## 1 chains, each with iter=20000; warmup=1000; thin=10;
## post-warmup draws per chain=1900, total post-warmup draws=1900.

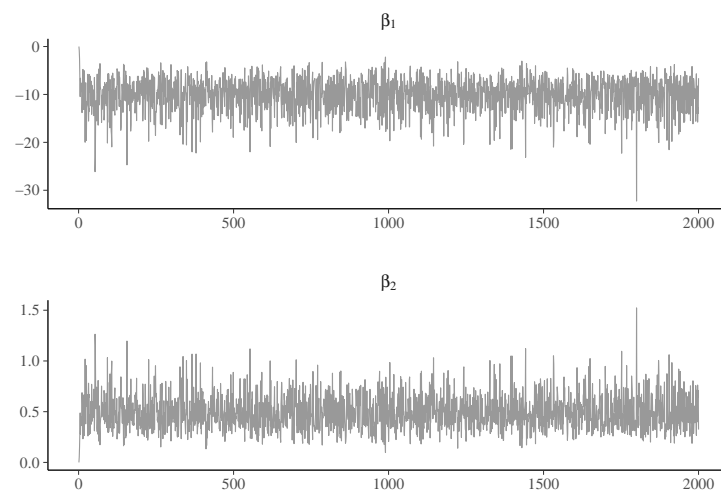
```

```
##
##           mean se_mean  sd  2.5%  25%   50%   75%  97.5% n_eff Rhat
## beta[1] -10.15    0.08 3.51 -17.92 -12.17  -9.65  -7.61  -4.64 1763    1
## beta[2]  0.50    0.00 0.17  0.22  0.37  0.47  0.60  0.89 1735    1
## lp__    -13.89    0.03 1.05 -16.67 -14.31 -13.55 -13.15 -12.88 1733    1
##
## Samples were drawn using NUTS(diag_e) at Tue Nov 10 14:23:37 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

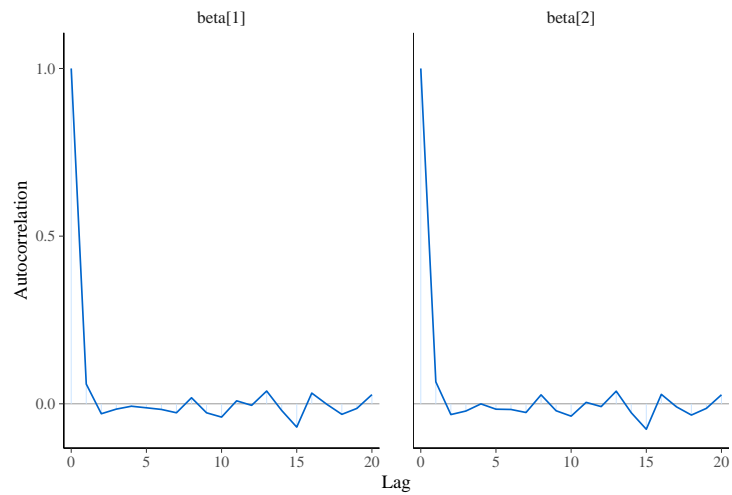
```
library(bayesplot)

post_log <- extract(stan_log, inc_warmup = TRUE, permuted = FALSE)

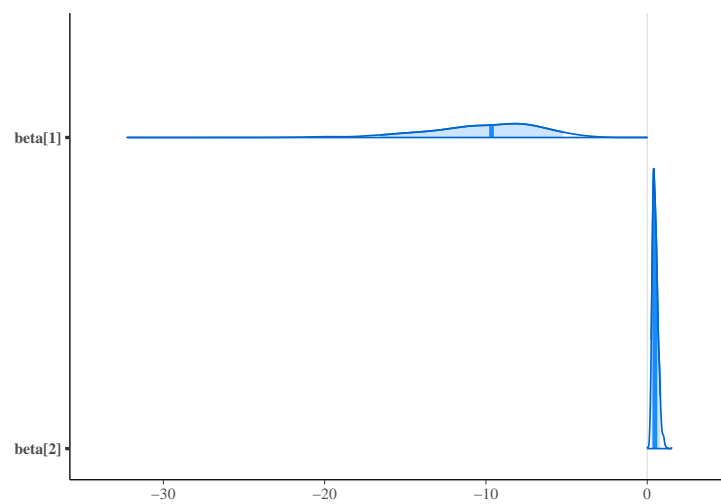
color_scheme_set("mix-brightblue-gray")
mcmc_trace(post_log, pars = c("beta[1]", "beta[2]"), n_warmup = 0,
            facet_args = list(nrow = 2, labeller = label_parsed))
```



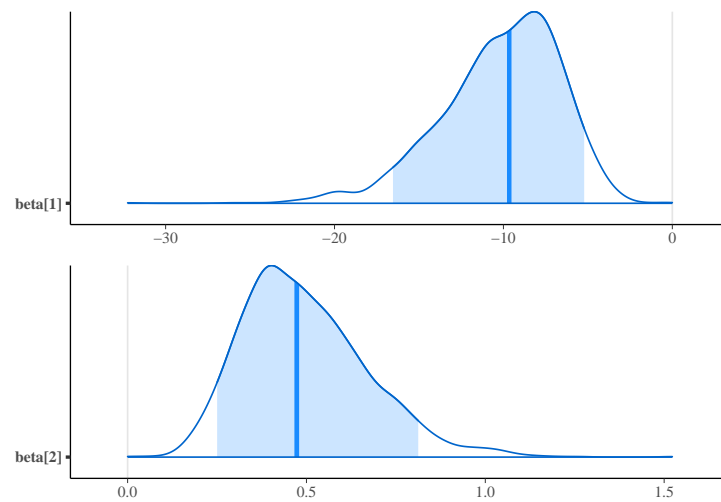
```
mcmc_acf(post_log, pars = c("beta[1]", "beta[2]"))
```



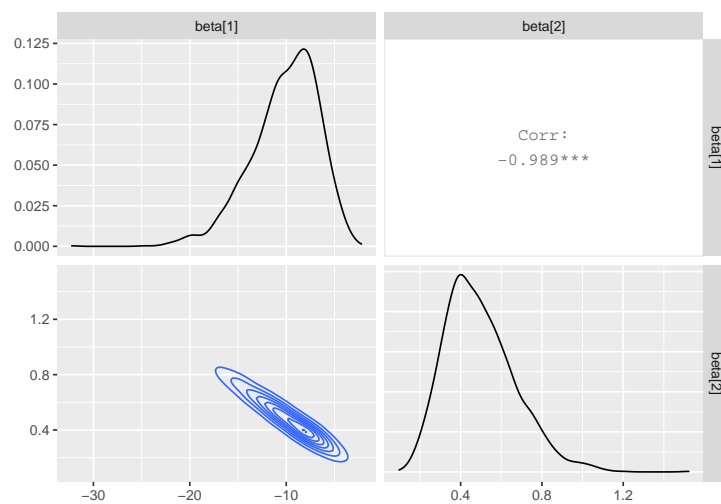
```
mcmc_areas(post_log, pars = c("beta[1]", "beta[2]"), prob=0.9)
```



```
multiplot(mcmc_areas(post_log, pars = c("beta[1]"), prob=0.9),  
          mcmc_areas(post_log, pars = c("beta[2]"), prob=0.9))
```



```
library(ggmcmc)
#ggs(stan_log) %>% ggmcmc(., file = "ggmcmc_log.html")
ggs(stan_log) %>% ggs_pairs(., lower = list(continuous = "density"))
```



- Existe uma biblioteca para modelos lineares bayesianos usando o Stan chamada **rstanarm**. Nesta biblioteca, a função **stan\_glm** pode ser utilizada para o ajuste de MLGs sob o ponto de vista bayesiano.

– <https://cran.r-project.org/web/packages/rstanarm/>

### 7.6.5 Modelos Dinâmicos

- A Biblioteca **walker** para do R que usa o RStan para fazer inferência bayesiana em modelos lineares com coeficientes variando no tempo (modelos dinâmicos).
- Modelo de Regressão Dinâmico Bayesiano

$$y_t = x_t \beta_t + \epsilon_t, \quad \epsilon_t \sim \text{Normal}(0, \sigma_y^2)$$

$$\beta_{t+1} = \beta_t + \eta_t, \quad \eta_t \sim \text{Normal}_k(0, D)$$

onde

- $y_t$ : variável resposta no instante  $t$ ;
- $x_t$ : vetor com  $k$  variáveis preditoras no instante  $t$ ;
- $\epsilon_t$  e  $\eta_t$ : ruídos brancos;
- $\beta_t$ : vetor dos  $k$  coeficientes de regressão no instante  $t$ ;
- $D = \text{diag}(\sigma_{\eta_i})$ ;
- $\sigma = (\sigma_y, \sigma_{\eta_1}, \dots, \sigma_{\eta_k})$ : vetor de parâmetros de variância.

As distribuição a priori são dadas por

$$\beta_1 \sim \text{Normal}(m_\beta, s_\beta^2)$$

$$\sigma_i^2 \sim \text{NormalTrunc}(m_{\sigma_i}, s_{\sigma_i}^2)$$

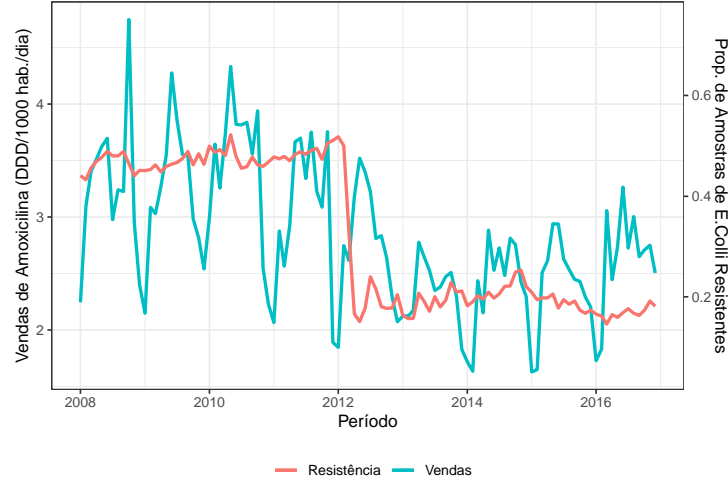
- Sobre a biblioteca **walker**:
  - <https://cran.r-project.org/web/packages/walker/vignettes/walker.html>
  - <https://rdrr.io/cran/walker/man/walker.html>



### 7.6.5.1 Dados de Amoxicilina (fonte: CEA)

- Dados mensais de venda de Amoxicilina e resistência da bactéria *E. coli* no período de 01/2008 à 12/2016.
- A venda de Amoxicilina foi avaliada pela média mensal das doses diárias definidas por 1000 habitantes-dia (DDD/1000hab/dia), que indica a quantidade média de determinado antibiótico que uma dada população consome diariamente (*IMS Health Brazil/Pfizer*).
- Para avaliar a resistência da bactéria *E. coli* à Amoxicilina foram utilizados dados obtidos a partir de amostras da rede de apoio do laboratório DASA, que atende principalmente a rede privada de assistência à saúde mas também inclui hospitais que atendem pacientes pelo Sistema Único de Saúde (SUS). Foram incluídas amostras resultantes de exames de sangue e urina positivas, avaliando a proporção de cepas isoladas resistentes à amoxicilina.
- **Objetivo.** Avaliar o impacto da regulamentação RDC 44 da ANVISA que obriga a retenção da prescrição médica para a venda de antimicrobianos, implementada em 26 de Outubro de 2010.

```
library(lubridate)
dados = read_csv("Amoxicillin.csv") %>%
  select(Período=Período,Vendas=DDD1000hab_dia,Resistência=Resistencia_Amoxicilina.Clavulanato.k)
dados$Vendas[61:62]=mean(c(2.073005,2.173923))
mV=mean(dados$Vendas); sdV=sd(dados$Vendas)
mR=mean(dados$Resistência); sdR=sd(dados$Resistência)
Tr <- sdV/sdR # var. aux. para transformação dos eixos
dados %>% ggplot(aes(x=Período)) + theme_bw() +
  geom_line(aes(y=Vendas,colour="Vendas"),lwd=1) + theme(legend.position="bottom") +
  geom_line(aes(x=Período,y=((Resistência-mR)*Tr+mV),colour="Resistência"),lwd=1)+
  labs(y="Vendas de Amoxicilina (DDD/1000 hab./dia)",colour="") +
  scale_y_continuous(sec.axis = sec_axis(~./Tr+(mR-mV/Tr), name = "Prop. de Amostras de E.Coli F
```



Para verificar se o efeito da venda de antimicrobianos influencia na resistência bacteriana e identificar possíveis mudanças após a implementação da lei, foi considerado um modelo de regressão dinâmico bayesiano, descrito a seguir.

$$(Y_t - \bar{Y}_0) = \beta_t(X_t - \bar{X}_0) + \epsilon_t, \quad \epsilon_t \sim \text{Normal}(0, \sigma_y^2)$$

$$\beta_{t+1} = \beta_t + W_t, \quad W_t \sim \text{Normal}(0, \sigma_r^2)$$

$$\sigma_y^2 \sim \text{Half} - \text{Normal}(100)$$

$$\beta_t \sim \text{Normal}(0, 100)$$

$$\sigma_r^2 \sim \text{Half} - \text{Normal}(100)$$

\*  $Y_t$ : proporção de testes de resistência positiva do microbiano *E. coli* no instante  $t$ ;

\*  $X_t$ : quantidade de doses consumidas de Amoxicilina (DDD/1000hab/dia) no instante  $t$ ;

\*  $\beta_t$ : parâmetro que representa o efeito da venda do antimicrobiano na resistência bacteriana no instante  $t$ ;

\*  $\epsilon_t$  e  $W_t$ : ruídos brancos.

```

library(walker)
set.seed(666)
# resistência média e venda média do período anterior a RDC 44
mAntes = dados %>% filter(year(Período)<2011) %>%
  summarise(mean(Vendas),mean(Resistência)) %>% t()
fit1 <- dados %>%
  mutate(Vendas=Vendas-mAntes[1],Resistência=Resistência-mAntes[2]) %>%
  walker(data=., formula = Resistência ~ -1+rw1(~ -1+Vendas,beta=c(0,100),
    sigma=c(0,100)),sigma_y_prior = c(0,100), chain=2)

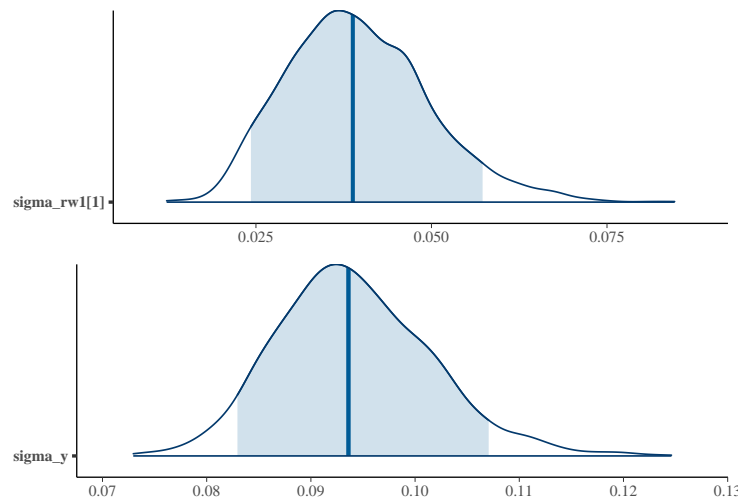
##
## SAMPLING FOR MODEL 'walker_lm' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.001 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 10 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [ 0%] (Warmup)
## Chain 1: Iteration:  200 / 2000 [10%] (Warmup)
## Chain 1: Iteration:  400 / 2000 [20%] (Warmup)
## Chain 1: Iteration:  600 / 2000 [30%] (Warmup)
## Chain 1: Iteration:  800 / 2000 [40%] (Warmup)
## Chain 1: Iteration: 1000 / 2000 [50%] (Warmup)
## Chain 1: Iteration: 1001 / 2000 [50%] (Sampling)
## Chain 1: Iteration: 1200 / 2000 [60%] (Sampling)
## Chain 1: Iteration: 1400 / 2000 [70%] (Sampling)
## Chain 1: Iteration: 1600 / 2000 [80%] (Sampling)
## Chain 1: Iteration: 1800 / 2000 [90%] (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 7.369 seconds (Warm-up)
## Chain 1:                8.175 seconds (Sampling)
## Chain 1:                15.544 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'walker_lm' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0.001 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 10 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 2000 [ 0%] (Warmup)
## Chain 2: Iteration:  200 / 2000 [10%] (Warmup)

```

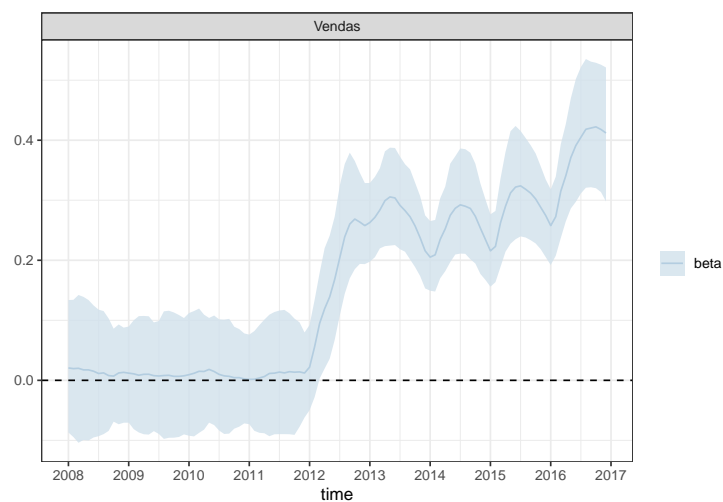
```
## Chain 2: Iteration: 400 / 2000 [ 20%] (Warmup)
## Chain 2: Iteration: 600 / 2000 [ 30%] (Warmup)
## Chain 2: Iteration: 800 / 2000 [ 40%] (Warmup)
## Chain 2: Iteration: 1000 / 2000 [ 50%] (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%] (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%] (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%] (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%] (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 2:
## Chain 2: Elapsed Time: 7.61 seconds (Warm-up)
## Chain 2:           7.921 seconds (Sampling)
## Chain 2:           15.531 seconds (Total)
## Chain 2:
```

```
#Default: chain=4, iter=2000, warmup=1000, thin=1
```

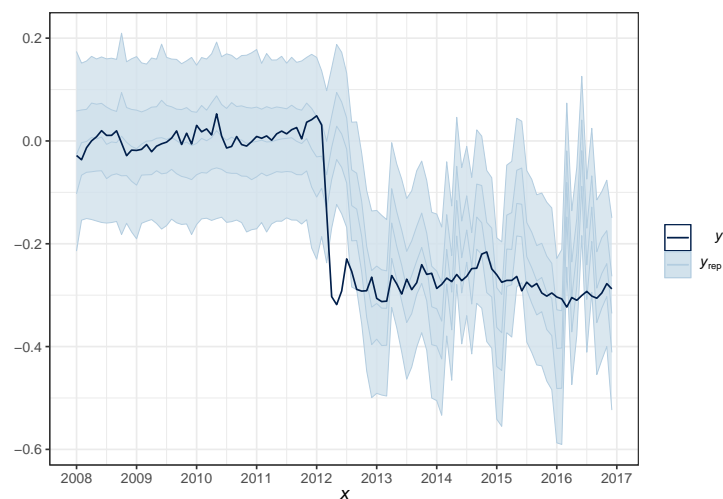
```
multiplot(
  #mcmc_areas(as.matrix(fit1$stanfit), regex_pars = c("beta_fixed"), prob=0.9),
  mcmc_areas(as.matrix(fit1$stanfit), regex_pars = c("sigma_rw1"), prob=0.9),
  mcmc_areas(as.matrix(fit1$stanfit), regex_pars = c("sigma_y"), prob=0.9))
```



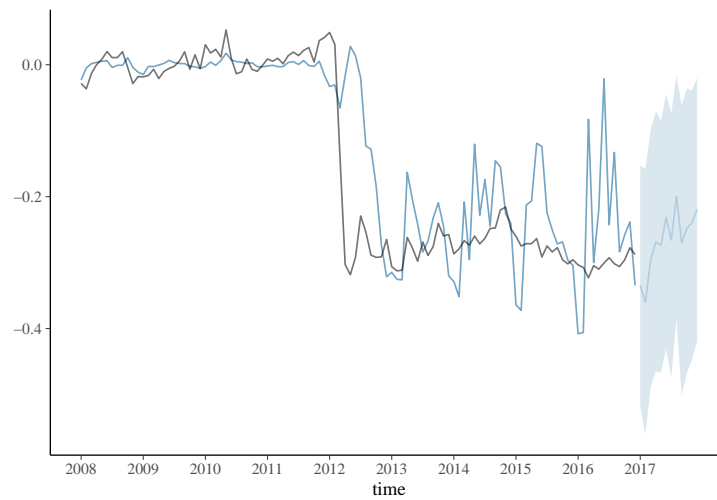
```
plot_coefs(fit1, scales = "free", alpha=0.8) + theme_bw() +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))
```



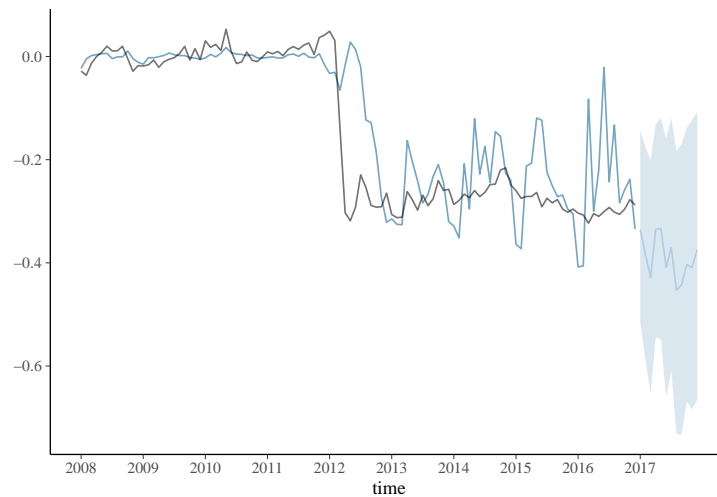
```
pp_check(fit1, alpha=0.8) + theme_bw() +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))
```



```
new_data <- data.frame(Vendas=seq(dados$Vendas[108]-mAntes[1],-0.4,length.out=12)+c(0,rnorm(11,0,
pred1 <- predict(fit1, new_data)
plot_predict(pred1, alpha=0.8) +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))
```



```
new_data <- data.frame(Vendas=seq(dados$Vendas[108]-mAntes[1],-1.1,length.out=12)+c(0,
pred1 <- predict(fit1, new_data)
plot_predict(pred1, alpha=0.8) +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))))
```



#### 7.6.5.2 Extensões: Efeitos mais suaves e modelos não gaussianos

Ao modelar os coeficientes de regressão como uma passeio aleatório simples, as estimativas posteriores desses coeficientes podem ter grandes variações de curto prazo que podem não ser realistas na prática. Uma maneira de impor mais

suavidade às estimativas é alternar dos coeficientes do passeio aleatório para coeficientes de passeio aleatório de segunda ordem integrados:

$$\beta_{t+1} = \beta_t + \nu_t$$

$$\nu_{t+1} = \nu_t + \eta_t, \quad \eta_t \sim \text{Normal}_k(0, D)$$

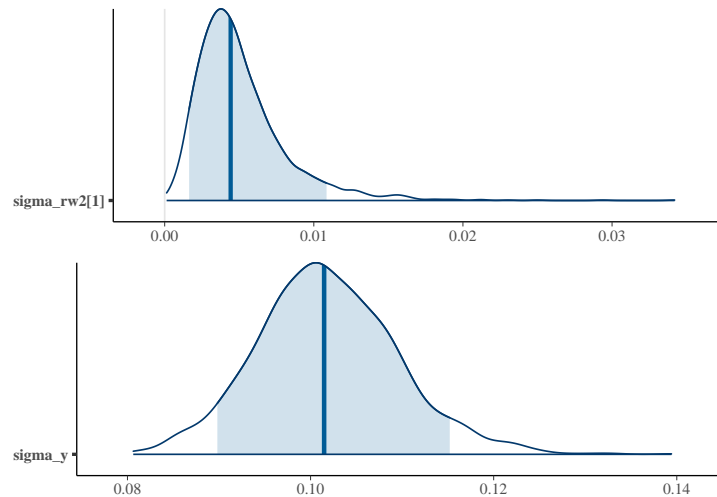
```
library(walker)
set.seed(666)
fit2 <- dados %>%
  mutate(Vendas=Vendas-mAntes[1],Resistência=Resistência-mAntes[2]) %>%
  walker(data=., formula = Resistência ~ -1+rw2(~ -1+Vendas, beta=c(0,100),
    sigma=c(0,100),nu=c(0,100)),sigma_y_prior=c(0,100), chain=2)

##
## SAMPLING FOR MODEL 'walker_lm' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.001 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 10 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [ 0%] (Warmup)
## Chain 1: Iteration:  200 / 2000 [10%] (Warmup)
## Chain 1: Iteration:  400 / 2000 [20%] (Warmup)
## Chain 1: Iteration:  600 / 2000 [30%] (Warmup)
## Chain 1: Iteration:  800 / 2000 [40%] (Warmup)
## Chain 1: Iteration: 1000 / 2000 [50%] (Warmup)
## Chain 1: Iteration: 1001 / 2000 [50%] (Sampling)
## Chain 1: Iteration: 1200 / 2000 [60%] (Sampling)
## Chain 1: Iteration: 1400 / 2000 [70%] (Sampling)
## Chain 1: Iteration: 1600 / 2000 [80%] (Sampling)
## Chain 1: Iteration: 1800 / 2000 [90%] (Sampling)
## Chain 1: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 8.569 seconds (Warm-up)
## Chain 1:                8.853 seconds (Sampling)
## Chain 1:                17.422 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'walker_lm' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0.002 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 20 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
```

```
## Chain 2:
## Chain 2: Iteration:    1 / 2000 [  0%] (Warmup)
## Chain 2: Iteration:   200 / 2000 [ 10%] (Warmup)
## Chain 2: Iteration:   400 / 2000 [ 20%] (Warmup)
## Chain 2: Iteration:   600 / 2000 [ 30%] (Warmup)
## Chain 2: Iteration:   800 / 2000 [ 40%] (Warmup)
## Chain 2: Iteration:  1000 / 2000 [ 50%] (Warmup)
## Chain 2: Iteration: 1001 / 2000 [ 50%] (Sampling)
## Chain 2: Iteration: 1200 / 2000 [ 60%] (Sampling)
## Chain 2: Iteration: 1400 / 2000 [ 70%] (Sampling)
## Chain 2: Iteration: 1600 / 2000 [ 80%] (Sampling)
## Chain 2: Iteration: 1800 / 2000 [ 90%] (Sampling)
## Chain 2: Iteration: 2000 / 2000 [100%] (Sampling)
## Chain 2:
## Chain 2: Elapsed Time: 8.753 seconds (Warm-up)
## Chain 2:                6.997 seconds (Sampling)
## Chain 2:                15.75 seconds (Total)
## Chain 2:
```

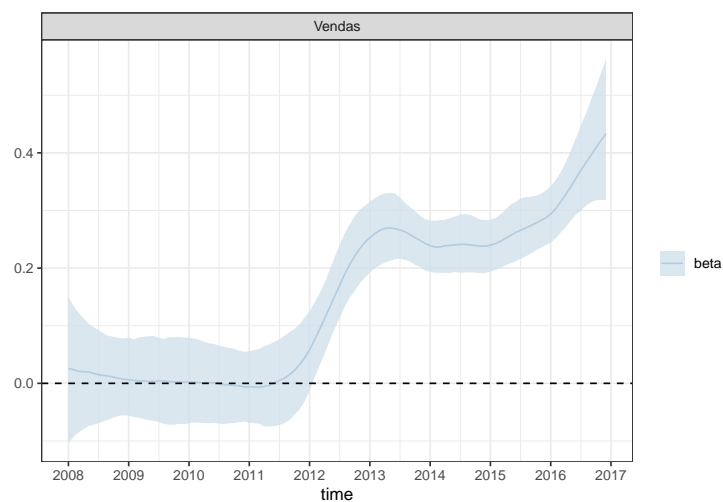
```
#Default: chain=4, iter=2000, warmup=1000, thin=1
```

```
multiplot(
  mcmc_areas(as.matrix(fit2$stanfit), regex_pars = c("sigma_rw2"), prob=0.9),
  mcmc_areas(as.matrix(fit2$stanfit), regex_pars = c("sigma_y"), prob=0.9))
```

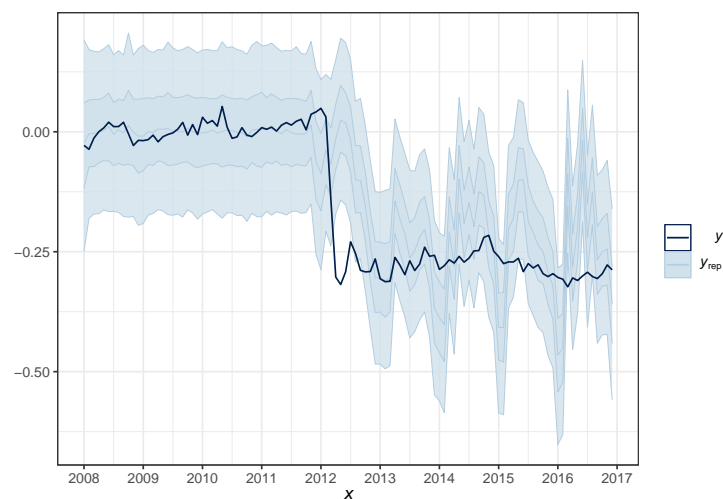


```
plot_coefs(fit2, scales = "free", alpha=0.8) + theme_bw() +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))
```

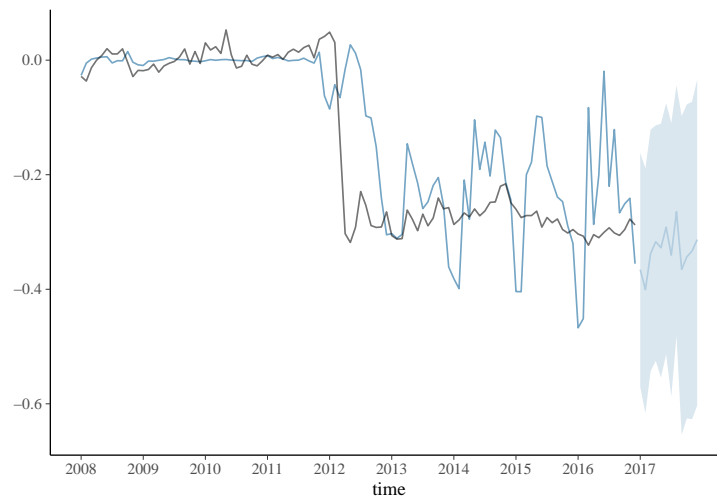




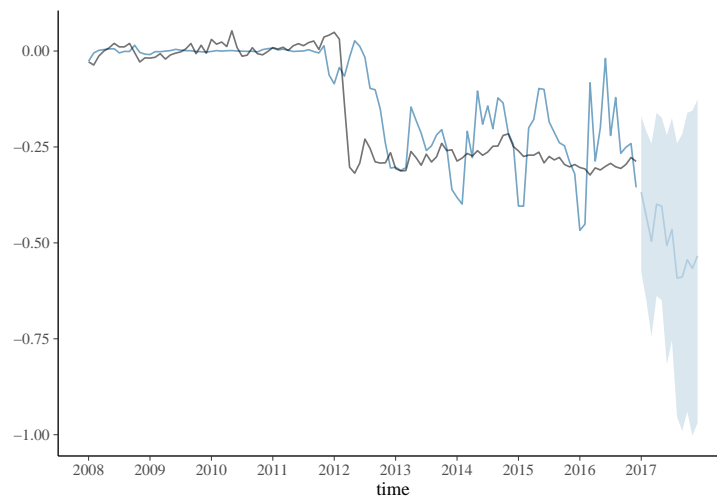
```
pp_check(fit2, alpha=0.8) + theme_bw() +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))
```



```
new_data <- data.frame(Vendas=seq(dados$Vendas[108]-mAntes[1],-0.4,length.out=12)+c(0,rnorm(11,0,
pred2 <- predict(fit2, new_data)
plot_predict(pred2, alpha=0.8) +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))
```



```
new_data <- data.frame(Vendas=seq(dados$Vendas[108]-mAntes[1],-1.1,length.out=12)+c(0,
pred2 <- predict(fit2, new_data)
plot_predict(pred2, alpha=0.8) +
  scale_x_continuous(breaks=seq(1,length(dados$Período)+1,12),labels=c(unique(year(dados$Período))))
```



- A função `walker_glm` estende o pacote para lidar com observações de com distribuição de *Poisson* e *Binomial*, usando a metodologia similar à mencionada acima.

## Appendix A

# Breve Resumo de Medida e Probabilidade

Essa seção tem o objetivo de apresentar as ideias de probabilidade como uma medida e da integral de Lebesgue. Para maiores detalhes, ver Ash and Doleans-Dade (2000), Billingsley (1986), Shiryaev (1996) ou, para uma versão mais resumida, os Apêndices de Schervish (2012).

### A.1 Conceitos Básicos

- $\Omega$ : espaço amostral (um conjunto não vazio).
- $\mathcal{A}$ :  $\sigma$ -álgebra de subconjuntos de  $\Omega$ , isto é,
  1.  $\Omega \in \mathcal{A}$ ;
  2.  $A \in \mathcal{A} \implies A^c \in \mathcal{A}$ ;
  3.  $A_1, A_2, \dots \in \mathcal{A} \implies \bigcup_{i \geq 1} A_i \in \mathcal{A}$ .
- Os elementos de  $\mathcal{A}$  são chamados de *eventos* e serão denotados por  $A, B, C, \dots, A_1, A_2, \dots$
- Uma coleção de eventos  $A_1, A_2, \dots$  forma uma *partição* de  $\Omega$  se  $A_i \cap A_j = \emptyset$ ,  $\forall i \neq j$ , e  $\bigcup_{i=1}^{\infty} A_i = \Omega$ .
- $(\Omega, \mathcal{A})$ : *espaço mensurável*.
- Usualmente, denota-se a  $\sigma$ -álgebra gerada por um conjunto  $\mathcal{C}$  como  $\sigma(\mathcal{C})$ . Por exemplo:

- $\sigma(\Omega) = \{\emptyset, \Omega\}$  ( $\sigma$ -álgebra trivial);
- Para  $A \subset \Omega$ ,  $\sigma(A) = \{\emptyset, A, A^c, \Omega\}$ ;
- $\sigma(\mathbb{N}) = \mathcal{P}(\mathbb{N})$  (partes de  $\mathbb{N}$ , todos o subconjuntos de  $\mathbb{N}$ );
- $\sigma(\{(-\infty, x) : x \in \mathbb{R}\}) = \mathcal{B}(\mathbb{R})$  (borelianos de  $\mathbb{R}$ )

**Definição:** A função  $\mu : \mathcal{A} \rightarrow \bar{\mathbb{R}}_+$  é uma *medida* se

1.  $\mu(\emptyset) = 0$ ;
  2.  $A_1, A_2, \dots \in \mathcal{A}$  com  $A_i \cap A_j = \emptyset$ ,  $\forall i \neq j$ ,  $\mu\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} \mu(A_i)$ .
- $(\Omega, \mathcal{A}, \mu)$  é chamado de *espaço de medida*.

**Exemplo 1 (medida de contagem):** Seja  $\Omega$  um conjunto não vazio e  $A \subseteq \Omega$ . Defina  $\nu(A) = |A|$  como o número de elementos (cardinalidade) de  $A$ . Assim,  $\nu(\Omega) > 0$ ,  $\nu(\emptyset) = 0$  e, se  $(A_n)_{n \geq 1}$  é uma sequência de eventos disjuntos, então  $\nu(\bigcup A_n) = \sum \nu(A_n)$ . Note que  $\nu(A) = \infty$  é possível se  $\Omega$  tem infinitos elementos.

**Exemplo 2 (medida de Lebesgue):** Seja  $\Omega = \mathbb{R}$  e  $A \subseteq \Omega$  um intervalo. Se  $A$  é limitado, defina  $\lambda(A)$  como o comprimento do intervalo  $A$ . Se  $A$  não é limitado,  $\lambda(A) = \infty$ . Note que  $\lambda(\mathbb{R}) = \infty$ ,  $\lambda(\emptyset) = 0$  e, se  $A_1 \cap A_2 = \emptyset$  e  $A_1 \cup A_2$  é um intervalo (ou uma união de intervalos disjuntos), então  $\lambda(A_1 \cup A_2) = \lambda(A_1) + \lambda(A_2)$ .

**Exemplo 3:** Seja  $f : \mathbb{R} \rightarrow \mathbb{R}_+$  uma função contínua e não nula. Para cada intervalo  $A$ , defina  $\mu(A) = \int_A f(x)dx = \int_{\mathbb{R}} \mathbb{I}_A(x)f(x)dx$ . Então,  $\mu(\mathbb{R}) > 0$ ,  $\mu(\emptyset) = 0$  e, se  $A_1 \cap A_2 = \emptyset$  e  $A_1 \cup A_2$  é um intervalo (ou uma união de intervalos disjuntos), então  $\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$ .

- Se  $\mu(\Omega) < \infty$  dizemos que  $\mu$  é uma *medida finita*. Se existe uma partição enumerável de  $\Omega$ ,  $A_1, A_2, \dots$ , tal que cada elemento da partição tem medida finita,  $\mu(A_i) < \infty$ ,  $\forall i$ , dizemos que  $\mu$  é uma *medida  $\sigma$ -finita*.

**Definição:**  $P : \mathcal{A} \rightarrow [0, 1]$  é uma **medida de probabilidade** se

1.  $P(\Omega) = 1$ ;
2.  $A_1, A_2, \dots \in \mathcal{A}$  com  $A_i \cap A_j = \emptyset$ ,  $P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i)$ .

- $(\Omega, \mathcal{A}, P)$ : espaço de probabilidade

**Definição:** Seja  $(\Omega, \mathcal{A})$  e  $(\mathfrak{X}, \mathcal{F})$  dois espaços mensuráveis. Uma função  $X : \Omega \rightarrow \mathfrak{X}$  é chamado de *quantidade aleatória* se é uma *função mensurável*, isto é, se  $\forall B \in \mathcal{F}$ , o evento  $A = X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$  pertence à  $\sigma$ -álgebra original  $\mathcal{A}$ .

Se  $\mathfrak{X} = \mathbb{R}$  e  $\mathcal{F} = \mathcal{B}(\mathbb{R})$  ( $\sigma$ -álgebra de Borel),  $X$  é chamada **variável aleatória** (v.a.).

- Considere  $(\Omega, \mathcal{A}, P)$ . A medida de probabilidade  $P_X$  induzida por  $X$  recebe o nome de *distribuição de  $X$* . Se  $B \in \mathcal{F}$  e  $A = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$ , a medida induzida por  $X$  é

$$P_X(B) = P_X(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}) = P(A) .$$

- A distribuição de  $X$  é dita ser *discreta* se existe um conjunto enumerável  $A \subseteq \mathfrak{X}$  tal que  $P_X(A) = 1$ . A distribuição de  $X$  é *contínua* se  $P_X(\{x\}) = 0$  para todo  $x \in \mathfrak{X}$ .

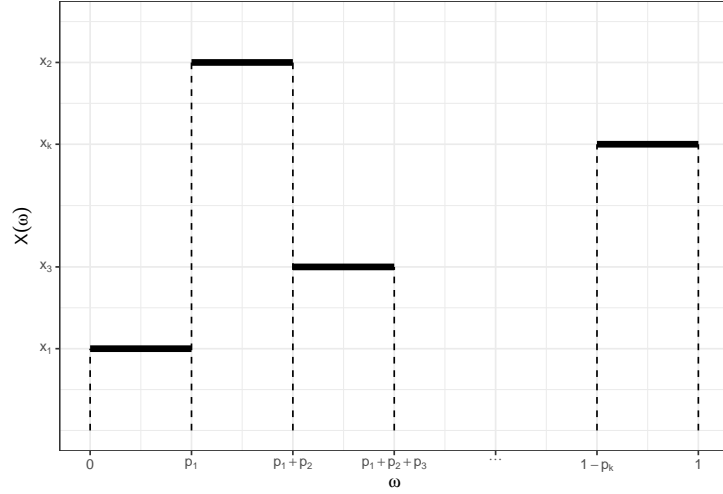
## A.2 Valor Esperado de $X$ (OU uma ideia da tal Integral de Lebesgue)

Por simplicidade, considere o espaço  $(\Omega = [0, 1], \mathcal{A} = \mathcal{B}([0, 1]), P = \lambda)$ .

Seja  $X : \Omega \rightarrow \mathbb{R}_+$  uma variável aleatória discreta, assumindo valores não negativos  $\mathfrak{X} = \{x_1, x_2, \dots, x_k\}$  com probabilidades  $\{p_1, p_2, \dots, p_k\}$ . Nos cursos básicos de probabilidade é visto que o *valor esperado* (ou *esperança*) de  $X$  é  $E[X] = \sum x_i P(X = x_i) = \sum x_i p_i$ .

Podemos definir essa v.a. como

$$X(\omega) = \begin{cases} x_1, & \omega \in [0, p_1] = A_1 \\ x_2, & \omega \in [p_1, p_1 + p_2] = A_2 \\ \vdots & \\ x_j, & \omega \in \left[ \sum_{i=1}^{j-1} p_i, \sum_{i=1}^j p_i \right] = A_j \\ \vdots & \\ x_k, & \omega \in [1 - p_k, 1] = A_k \end{cases}$$



Note que a medida  $\lambda$  define uma distribuição uniforme no espaço  $(\Omega, \mathcal{A})$ . Assim, temos que

- $P_X(X = x_1) = P(X^{-1}(x_1)) = P(\{\omega \in \Omega : X(\omega) = x_1\}) = P(A_1) = \lambda([0, p_1]) = p_1,$
- $P_X(X = x_j) = P(\{\omega \in \Omega : X(\omega) = x_j\}) = \lambda\left(\left[\sum_{i=1}^{j-1} p_i, \sum_{i=1}^j p_i\right]\right) = p_j,$   
 $j \in \{2, \dots, k\}.$

**Definição:** Uma função mensurável  $X : \Omega \rightarrow \mathbb{R}_+$  é dita *simples* se assumir um número finito de valores.

**Definição:** Considere um espaço de probabilidade  $(\Omega, \mathcal{A}, P)$ ,  $X : \Omega \rightarrow \mathbb{R}_+$  v.a. assumindo valores  $\{x_1, x_2, \dots, x_k\}$  e  $A_1, A_2, \dots, A_k$  eventos disjuntos em  $\mathcal{A}$ . Seja

$$X(\omega) = \sum_{i=1}^k x_i \mathbb{1}_{A_i}(\omega), \text{ uma função simples com } A_i = X^{-1}(x_i), i = 1, \dots, k. \text{ A}$$

integral de Lebesgue de  $X$  em relação à medida  $P$  é

$$E[X] = \int_{\Omega} X dP = \sum_{i=1}^k x_i P(A_i).$$

**Propriedades:** se  $X, Y : \Omega \longrightarrow \mathbb{R}_+$  são funções simples, então

1.  $\int_{\Omega} X dP \geq 0$ ;
2.  $\int_{\Omega} cX dP = c \int_{\Omega} X dP$ ;
3.  $\int_{\Omega} (X + Y) dP = \int_{\Omega} X dP + \int_{\Omega} Y dP$ .

**Demo 1.** Segue de  $x_i \geq 0$  e  $P(A_i) \geq 0$ .

**Demo 2.**

Para  $X$  v.a. temos

$$X = \sum_{i=1}^k x_i \mathbb{I}_{A_i} \text{ e } cX = \sum_{i=1}^k c x_i \mathbb{I}_{A_i}. \text{ Logo,}$$

$$\int_{\Omega} cX dP = \sum_{i=1}^k c x_i P(A_i) = c \sum_{i=1}^k x_i P(A_i) = c \int_{\Omega} X dP.$$

**Demo 3.**

$$X = \sum_{i=1}^k x_i \mathbb{I}_{A_i} \text{ e } Y = \sum_{j=1}^l y_j \mathbb{I}_{B_j}.$$

$$X + Y = \sum_{i=1}^k x_i \mathbb{I}_{A_i} + \sum_{j=1}^l y_j \mathbb{I}_{B_j} = \sum_{i=1}^k \sum_{j=1}^l x_i \mathbb{I}_{A_i \cap B_j} + \sum_{i=1}^k \sum_{j=1}^l y_j \mathbb{I}_{A_i \cap B_j}$$

$$= \sum_{i=1}^k \sum_{j=1}^l (x_i + y_j) \mathbb{I}_{A_i \cap B_j}.$$

$$\int_{\Omega} (X + Y) dP = \sum_{i=1}^k \sum_{j=1}^l (x_i + y_j) P(A_i \cap B_j) = \sum_{i=1}^k \sum_{j=1}^l x_i P(A_i \cap B_j) + \sum_{i=1}^k \sum_{j=1}^l y_j P(A_i \cap B_j)$$

$$= \sum_{i=1}^k x_i P(A_i) + \sum_{j=1}^l y_j P(B_j) = \int_{\Omega} X dP + \int_{\Omega} Y dP.$$

A generalização da integral de Lebesgue é feita usando resultados como o *Lema de Fatou* e os teoremas da *convergência monótona* e da *convergência dominada*. Aqui será apresentado apenas uma ideia dessa extensão. Para maiores detalhes, veja as referências citadas anteriormente (Ash and Doleans-Dade, 2000; Schervish, 2012; Billingsley, 1986; Shiryaev, 1996).

**Definição:** Seja  $X : \Omega \rightarrow \mathbb{R}_+$  uma função mensurável não negativa e considere o conjunto de funções  $\mathcal{C}_X = \{f : \Omega \rightarrow \mathbb{R}_+, f \text{ simples}, f \leq X\}$ . O *valor esperado de X* é

$$E[X] = \int_{\Omega} X dP = \sup \left\{ \int_{\Omega} f dP : f \in \mathcal{C}_X \right\}.$$

**Resultado:** Para toda função  $X : \Omega \rightarrow \mathbb{R}_+$ , existe uma sequência  $(X_n)_{n \geq 1}$  de funções simples não-negativas tais que  $X_n(\omega) \leq X_{n+1}(\omega)$ ,  $\forall \omega \in \Omega$ ,  $\forall n \in \mathbb{N}$  com  $X_n(\omega) \uparrow X(\omega)$ ,  $\forall \omega \in \Omega$ .

**Exemplo** de sequência  $(X_n)_{n \geq 1}$  atendendo as condições anteriores

Para cada  $n$ , considere  $1 + n2^n$  conjuntos em  $\mathcal{A}$  :

- $E_j^n = \left\{ \omega \in \Omega : \frac{j}{2^n} \leq X(\omega) \leq \frac{j+1}{2^n} \right\}$ ,  $j = 0, 1, \dots, n2^n - 1$ .
- $E_{n2^n}^n = \left\{ \omega \in \Omega : X(\omega) \geq n \right\}$

e defina  $X_n(\omega) = \sum_{j=0}^{n2^n} \frac{j}{2^n} \mathbb{1}_{E_j^n}(\omega)$ . Pode-se provar que  $(X_n)_{n \geq 1}$  é tal que

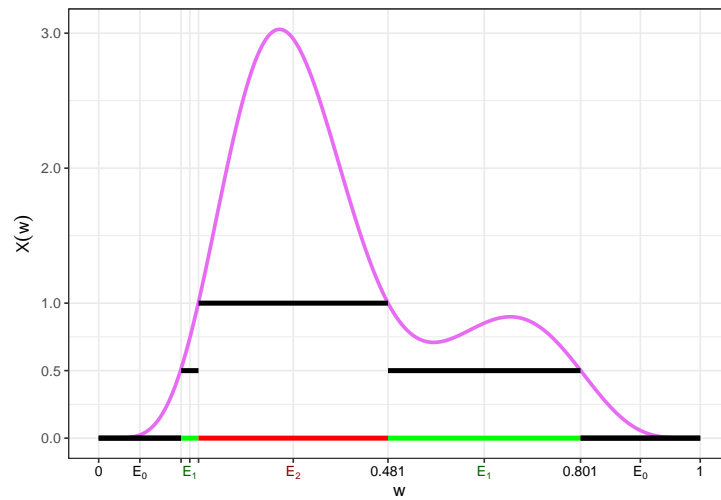
- $X_n$  é simples,  $\forall n \geq 1$
- $X_n \leq X_{n+1}$
- $X_n(\omega) \uparrow X(\omega)$

A primeira função dessa sequência é

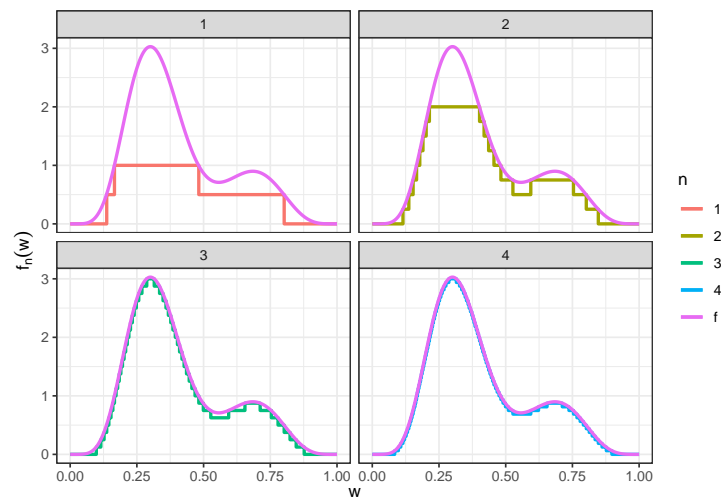
$$X_1(\omega) = \sum_{i=0}^2 \frac{i}{2} \mathbb{1}_{E_i^1}(\omega) = \begin{cases} 0, & \omega \in E_0^1 \\ 0.5, & \omega \in E_1^1 \\ 1, & \omega \in E_2^1 \end{cases}.$$



A.2. VALOR ESPERADO DE  $X$  (OU UMA IDEIA DA TAL INTEGRAL DE LEBESGUE) 169



O gráfico a seguir mostra os quatro primeiras funções da sequência  $(X_n)_{n \geq 1}$  e é possível ter uma ideia da convergência para  $X$ .



**Resultado:**  $X, Y : \Omega \rightarrow \mathbb{R}_+$ , com  $X \leq Y$ . Então  $E[X] \leq E[Y]$ .

**Demo:** Como  $X \leq Y$  (isto é,  $X(\omega) \leq Y(\omega) \forall \omega \in \Omega$ ),  $\mathcal{C}_X \subseteq \mathcal{C}_Y$   
 $\Rightarrow \sup \left\{ \int_{\Omega} f \, dP : f \in \mathcal{C}_X \right\} \leq \sup \left\{ \int_{\Omega} g \, dP : g \in \mathcal{C}_Y \right\} \Rightarrow$

$$\int_{\Omega} X dP \leq \int_{\Omega} Y dP.$$

**Definição:** Seja  $X : \Omega \rightarrow \mathbb{R}_+$  e  $E \in \mathcal{A}$  definimos  $E(X \mathbb{I}_E) = \int_E X dP = \int_{\Omega} X \mathbb{I}_E dP$ .

Se  $E, F \in \mathcal{A}$  com  $E \subseteq F$ ,  $\int_E X dP \leq \int_F X dP$ .

**Propriedades:** se  $X, Y : \Omega \rightarrow \mathbb{R}_+$  são funções mensuráveis positivas, então

1.  $\int_{\Omega} cX dP = c \int_{\Omega} X dP, c \geq 0$ ;
2.  $\int_{\Omega} (X + Y) dP = \int_{\Omega} X dP + \int_{\Omega} Y dP$ .

**Demo 1.** Seja  $X_n \uparrow X$ ,  $X_n \geq 0$  simples. Então  $cX_n \uparrow cX$ ,  $cX_n \geq 0$ , simples.

$$\begin{aligned} \int_{\Omega} cX dP &= \lim_{n \rightarrow \infty} \int_{\Omega} cX_n dP = \lim_{n \rightarrow \infty} c \int_{\Omega} X_n dP = c \lim_{n \rightarrow \infty} \int_{\Omega} X_n dP = \\ &= c \int_{\Omega} X dP. \end{aligned}$$

**Demo 2.** Exercício.

**Exemplo:** Suponha que  $X$  assume valores em  $\mathbb{N}$ . Pode-se escrever

$$X = \sum_{i=1}^{\infty} i \mathbb{I}_{A_i}, \text{ com } A_i = X^{-1}(\{i\}).$$

Defina  $X_n = \sum_{i=1}^{n-1} i \mathbb{I}_{A_i} + n \mathbb{I}_{\cup_{j=n}^{\infty} A_j}$ . Então  $X_n$  é simples,  $X_n \geq 0$ ,  $X_n \leq$

$X_{n+1}$  e  $X_n \uparrow X$ , de modo que  $E(X) = \int_{\Omega} X dP = \lim_{n \rightarrow \infty} \int_{\Omega} X_n dP$ .

Além disso,

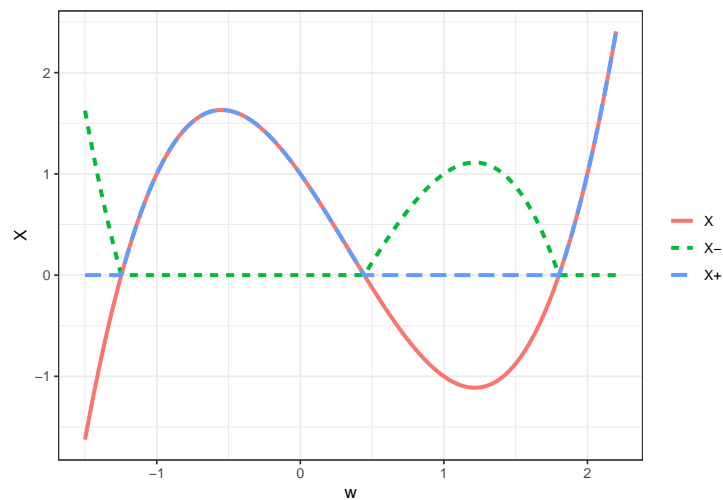
$$\begin{aligned} \int_{\Omega} X_n dP &= \sum_{i=1}^{n-1} i P(A_i) + n P\left(\bigcup_{j=n}^{\infty} A_j\right) = \sum_{i=1}^{n-1} i P(X=i) + n P(X \geq n) \\ &= \sum_{i=1}^{n-1} \sum_{j=1}^i P(X=i) + n P(X \geq n) = \sum_{j=1}^{n-1} \sum_{i=j}^{n-1} P(X=i) + n P(X \geq n) \end{aligned}$$

$$n) = \sum_{j=1}^{n-1} P(j \leq X \leq n-1) + n P(X \geq n) = \sum_{j=1}^n P(X \geq j),$$

então,  $E(X) = \lim_{n \rightarrow \infty} \sum_{j=1}^n P(X \geq j) = \sum_{j=1}^{\infty} P(X \geq j).$

Seja  $X : \Omega \rightarrow \mathbb{R}$  e  $X^-, X^+ : \Omega \rightarrow \mathbb{R}$  dados por

- $X^- = \max\{-X, 0\}$  (parte negativa de  $X$ ) e
- $X^+ = \max\{X, 0\}$  (parte positiva de  $X$ )



Note que  $X = X^+ - X^-$

Se  $\int_{\Omega} X^+ dP < \infty$  ou  $\int_{\Omega} X^- dP < \infty$ , definimos

$$E[X] = \int_{\Omega} X dP = \int_{\Omega} X^+ dP - \int_{\Omega} X^- dP = E[X^+] - E[X^-].$$

Além disso, seja  $|X| = X^+ + X^-$ . Então,  $E[|X|] < \infty$  se  $E(X^+) < \infty$  e  $E(X^-) < \infty$ , e, nesse caso, dizemos que  $X$  é *integrável*.

**Propriedades:** se  $X, Y : \Omega \longrightarrow \mathbb{R}$  são funções mensuráveis, então

1.  $X \leq Y \Rightarrow E(X) \leq E(Y)$ ;
2.  $c \in \mathbb{R}, E(cX) = cE(X)$ ;
3.  $X, Y$  integráveis.  $E(X + Y) = E(X) + E(Y)$ .

**Demo 1.**

$$X \leq Y \Rightarrow \begin{cases} X^+ \leq Y^+ \\ X^- \geq Y^- \end{cases}$$

$$E(X) = E(X^+) - E(X^-) \leq E(Y^+) - E(Y^-) = E(Y).$$

**Demo 2.**

$$(cX)^+ = \begin{cases} cX^+ & , \quad c \geq 0 \\ -cX^- & , \quad c < 0 \end{cases}$$

$$(cX)^- = \begin{cases} cX^- & , \quad c \geq 0 \\ -cX^+ & , \quad c < 0 \end{cases}$$

Para  $c < 0$ ,

$$E[cX] = E[(cX)^+] - E[(cX)^-] = E[-cX^-] - E[-cX^+] = -cE[X^-] + cE[X^+] = cE[X].$$

**Demo 3.**

$$\begin{aligned} & \int_{\Omega} (X^+ + Y^+) dP < \infty \text{ ou } \int_{\Omega} (X^- + Y^-) dP < \infty \\ & X + Y = (X + Y)^+ - (X + Y)^- = X^+ - X^- + Y^+ - Y^- \\ & \Rightarrow (X + Y)^+ + X^- + Y^- = X^+ + Y^+ + (X + Y)^- \\ & \Rightarrow \int_{\Omega} (X + Y)^+ dP + \int_{\Omega} X^- dP + \int_{\Omega} Y^- dP \\ & = \int_{\Omega} X^+ dP + \int_{\Omega} Y^+ dP + \int_{\Omega} (X + Y)^- dP. \\ & |X + Y| = |X^+ - X^- + Y^+ - Y^-| \leq X^+ + X^- + Y^+ + Y^- \\ & \Rightarrow \int_{\Omega} (X + Y)^+ dP - \int_{\Omega} (X + Y)^- dP = \int_{\Omega} X^+ dP - \int_{\Omega} X^- dP + \\ & \int_{\Omega} Y^+ dP - \int_{\Omega} Y^- dP. \\ & \Rightarrow \int_{\Omega} (X + Y) dP = \int_{\Omega} X dP + \int_{\Omega} Y dP \end{aligned}$$

### A.3 Funções de Variáveis Aleatórias

Considere agora uma v.a.  $X : \Omega \longrightarrow \mathbb{R}$  e uma função real  $g : \mathbb{R} \longrightarrow \mathbb{R}$ . Defina  $Y = g(X)$ . Então

$$(\Omega, \mathcal{A}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X) \xrightarrow{g} (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_Y)$$

$$(\Omega, \mathcal{A}, P) \xrightarrow{Y=g(X)} (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_Y)$$

Logo, se  $g$  é uma função mensurável,  $Y = g(X)$  também é v.a. e as medidas induzidas por  $X$  e  $Y$  são

$$P_X(A) = P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\});$$

$$P_Y(B) = P_X(g^{-1}(B)) = P_X(\{x \in \mathbb{R} : g(x) \in B\}) = P(\{\omega \in \Omega : g(X(\omega)) \in B\}).$$

Assim, uma pergunta natural é como obter o valor esperado de  $Y$ .

$$E(Y) = \int_{\Omega} Y dP = \int_{\Omega} g(X) dP \stackrel{?}{=} \int_{\mathbb{R}} g dP_X.$$

**Caso 1.** Seja  $g : \mathbb{R} \rightarrow \mathbb{R}_+$  uma função simples tal que  $g = \sum_{i=1}^k g_i \mathbb{I}_{B_i}$ ,  $g_1, \dots, g_k \in \mathbb{R}$  e  $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ . Então,

$$\begin{aligned} \int_{\Omega} Y dP &= \int_{\Omega} g(X) dP = \int_{\Omega} \left( \sum_{i=1}^k g_i \mathbb{I}_{B_i}(X) \right) dP = \int_{\Omega} \left( \sum_{i=1}^k g_i \mathbb{I}_{X^{-1}(B_i)} \right) dP \\ &\stackrel{def}{=} \sum_{i=1}^k g_i P(X^{-1}(B_i)) = \sum_{i=1}^k g_i P_X(B_i) = \int_{\mathbb{R}} \left( \sum_{i=1}^k g_i \mathbb{I}_{B_i} \right) dP_X \\ &= \int_{\mathbb{R}} g dP_X. \end{aligned}$$

**Caso 2.** Seja  $g : \mathbb{R} \rightarrow \mathbb{R}_+$  uma função não negativa e  $(g_n)_{n \geq 1}$ ,  $g_n \geq 0$ , uma sequência crescente de funções simples tal que  $g_n \uparrow g$ . Como  $g_n$  é simples,

$$\int_{\Omega} g_n(X) dP = \int_{\mathbb{R}} g_n dP_X \xrightarrow{n \uparrow \infty} \int_{\Omega} g(X) dP = \int_{\mathbb{R}} g dP_X.$$

**Caso 3.** Agora para  $g : \mathbb{R} \rightarrow \mathbb{R}$ , temos

$$\int_{\Omega} g^+(X) dP = \int_{\mathbb{R}} g^+ dP_X \text{ e } \int_{\Omega} g^-(X) dP = \int_{\mathbb{R}} g^- dP_X.$$

$$\text{Logo, } \int_{\Omega} g(X) dP = \int_{\mathbb{R}} g dP_X.$$

Suponha agora  $X$  v.a. discreta assumindo valores em  $\{x_1, x_2, \dots\}$  com probabilidade 1. Nesse caso, para  $A \subseteq \mathcal{B}(\mathbb{R})$ ,

$$P_X(A) = P_X(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}) = \sum_{i: x_i \in A} P_X(X = x_i).$$

Vamos “verificar” que  $E[g(X)] = \sum_{i=1}^{\infty} g(x_i)P_X(X = x_i)$ .

**Caso 1.**  $g$  simples com  $g = \sum_{i=1}^k g_i \mathbb{I}_{B_i}$ ,  $g_1, \dots, g_k \in \mathbb{R}_+$ ,  $B_1, \dots, B_k \in \mathcal{B}(\mathbb{R})$ . Então,

$$\begin{aligned} E[g(X)] &= \int_{\Omega} g(X) dP = \sum_{i=1}^k g_i P(X^{-1}(B_i)) = \sum_{i=1}^k g_i P_X(B_i) \\ &= \sum_{i=1}^k g_i \sum_{j: x_j \in B_i} P_X(X = x_j) = \sum_{i=1}^k g_i \sum_{j=1}^{\infty} \mathbb{I}_{B_i}(x_j) P_X(X = x_j) \\ &= \sum_{j=1}^{\infty} \underbrace{\left( \sum_{i=1}^k g_i \mathbb{I}_{B_i}(x_j) \right)}_{g(x_j)} P_X(X = x_j). \end{aligned}$$

**Caso 2.**  $g \geq 0$ ,  $g_n \geq 0$ ,  $g_n$  simples tal que  $g_n \uparrow g$

$$\begin{aligned} \int_{\Omega} g(X) dP &= \lim_{n \rightarrow \infty} \int_{\Omega} g_n(X) dP = \lim_{n \rightarrow \infty} \left\{ \sum_{j=1}^{\infty} g_n(x_j) P_X(X = x_j) \right\} = \\ &= \sum_{j=1}^{\infty} g(x_j) P_X(X = x_j) \end{aligned}$$

**Caso 3.** Agora para  $g: \mathbb{R} \rightarrow \mathbb{R}$ , temos

$$\begin{aligned} \int_{\Omega} g^+(X) dP &= \sum_{j=1}^{\infty} g^+(x_j) P_X(X = x_j) \quad \text{e} \quad \int_{\Omega} g^-(X) dP = \\ &= \sum_{j=1}^{\infty} g^-(x_j) P_X(X = x_j). \end{aligned}$$

Logo,  $\int_{\Omega} g(X) dP = \sum_{j=1}^{\infty} g(x_j) P_X(X = x_j)$ .

Suponha agora  $X$  v.a. absolutamente contínua com função de densidade de probabilidade  $f_X$ , ou seja, pode-se escrever  $P_X(X \in A) = \int_A f_X(t)dt = \int_{\mathbb{R}} \mathbb{I}_A(t)f_X(t)dt$ . Vamos “verificar” que  $E[g(X)] = \int_{\mathbb{R}} g(x)f_X(x)dx$ .

$$\begin{aligned}
 \textbf{Caso 1. } \quad & g \text{ simples com } g = \sum_{i=1}^k g_i \mathbb{I}_{B_i}, \quad g_1, \dots, g_k \in \mathbb{R}_+ \\
 & B_1, \dots, B_k \in \mathcal{B}(\mathbb{R}). \text{ Então,} \\
 E[g(X)] &= \int_{\Omega} g(X) dP = \int_{\Omega} \left( \sum_{i=1}^k g_i \mathbb{I}_{B_i}(X) \right) dP = \int_{\Omega} \left( \sum_{i=1}^k g_i \mathbb{I}_{X^{-1}(B_i)} \right) dP \\
 &= \sum_{i=1}^k g_i P(X^{-1}(B_i)) = \sum_{i=1}^k g_i P_X(B_i) = \sum_{i=1}^k g_i \int_{\mathbb{R}} \mathbb{I}_{B_i}(x)f_X(x)dx \\
 &= \int_{\mathbb{R}} \sum_{i=1}^k g_i \mathbb{I}_{B_i}(x)f_X(x)dx = \int_{\mathbb{R}} g(x)f_X(x)dx.
 \end{aligned}$$

A extensão para funções positivas e para funções reais é análogo ao que foi feito nos exemplos anteriores.

Assim, em geral, vale que:

- $X$  discreto:  $E[g(X)] = \sum_{j=1}^{\infty} g(x_j)P_X(X = x_j)$ ;
- $X$  (absolutamente) contínuo:  $E[g(X)] = \int_{\mathbb{R}} g(x)f_X(x)dx$ .

Esses resultados valem também se  $X : \Omega \rightarrow \mathbb{R}^k$  e  $g : \mathbb{R}^k \rightarrow \mathbb{R}$ .

**Exemplo 1.** Seja  $X$  uma v.a. definida em  $\mathbb{N}$  com função de probabilidade  $P_X(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \mathbb{I}_{\mathbb{N}}(x)$ , para  $\lambda > 0$  fixado.

Dizemos nesse caso que  $X \sim \text{Poisson}(\lambda)$ . Então, o valor esperado de  $X$  é

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} x P_X(X = x) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-1)!} \\ &= \lambda \sum_{x=1}^{\infty} \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!} \stackrel{t=x-1}{=} \lambda \sum_{t=0}^{\infty} \frac{e^{-\lambda} \lambda^t}{t!} \Rightarrow E[X] = \lambda. \end{aligned}$$

Ainda neste exemplo, considere  $g : \mathbb{R} \rightarrow \mathbb{R}$  com  $g(t) = e^t$ . Então,

$$\begin{aligned} E[g(X)] &= \sum_{x=0}^{\infty} g(x) P_X(X = x) = \sum_{x=0}^{\infty} e^x \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e)^x}{x!} \\ &= e^{-\lambda} e^{\lambda e} \underbrace{\sum_{x=0}^{\infty} \frac{e^{-\lambda e} (\lambda e)^x}{x!}}_1 = e^{\lambda e - \lambda} = e^{\lambda(e-1)}. \end{aligned}$$

**Exemplo 2.** Seja  $X$  uma v.a. definida em  $[0, 1]$  com função densidade de probabilidade  $f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \mathbb{I}_{[0,1]}(x)$ , para  $a, b > 0$  fixados. Dizemos nesse caso que  $X \sim \text{Beta}(a, b)$ . Então, o valor esperado de  $X$  é

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} dx \\ &= \frac{\Gamma(a+1)}{\Gamma(a+1+b)} \frac{\Gamma(a+b)}{\Gamma(a)} \int_0^1 \frac{\Gamma(a+1+b)}{\Gamma(a+1)\Gamma(b)} x^{(a+1)-1}(1-x)^{b-1} dx \\ &= \frac{\Gamma(a+1)}{\Gamma(a+1+b)} \frac{\Gamma(a+b)}{\Gamma(a)}. \end{aligned}$$

Considere agora  $g : \mathbb{R} \rightarrow \mathbb{R}$  com  $g(t) = t^c(1-t)^d$ , com  $c, d > 0$  fixados. Então,

$$\begin{aligned} E[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a+c-1}(1-x)^{b+d-1} dx \\ &= \frac{\Gamma(a+c)\Gamma(b+d)}{\Gamma(a+c+b+d)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \frac{\Gamma(a+c+b+d)}{\Gamma(a+b)\Gamma(b+d)} x^{(a+c)-1}(1-x)^{(b+d)-1} dx \\ &= \frac{\Gamma(a+c)\Gamma(b+d)}{\Gamma(a+c+b+d)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{\beta(a+c, b+d)}{\beta(a, b)}. \end{aligned}$$

## A.4 Função de Distribuição

**Definição:** Uma função  $F : \mathbb{R} \rightarrow [0, 1]$  é uma *função de distribuição* (f.d.) se



- (i)  $F$  é não-decrescente e contínua à direita;
- (ii)  $\lim_{x \downarrow -\infty} F(x) = 0$  e  $\lim_{x \uparrow +\infty} F(x) = 1$ .

**Proposição:** Se  $X$  é uma v.a., então  $F_X(x) = P_X(X \leq x)$  é uma f.d. Recíprocamente, se  $F_X$  é uma f.d, então existe uma v.a.  $X$  com f.d.  $F_X$ .

- É possível usar uma f.d.  $F$  para criar uma medida em  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ . Para tal, defina  $P((a, b]) = F(b) - F(a)$  e essa medida pode ser estendida para a  $\sigma$ -álgebra usando o Teorema de Extensão de Caratheodory (veja, por exemplo, Schervish (2012), pág. 578).
- Reciprocamente, se  $P$  é uma medida de probabilidade em  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  então  $F(x) = P((-\infty, x])$  é uma f.d.
- Neste caso, se  $g : \mathbb{R} \rightarrow \mathbb{R}$  é uma função mensurável, não será feita distinção entre  $\int g(x) dF(x) = \int g(x) dP_X(x)$ .
- Se  $P$  é uma medida de probabilidade em  $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$  então  $F(x_1, \dots, x_k) = P((-\infty, x_1] \times \dots \times (-\infty, x_k])$  é a *função de distribuição conjunta* do *vector aleatório*  $X = (X_1, \dots, X_K)$ .

**Definição:** Uma função de distribuição é dita

- (i) *Discreta* se existe um conjunto enumerável  $B = \{x_1, x_2, \dots\} \subset \mathbb{R}$  tal que  $P_X(B) = 1$  e  $F_d(x) = \sum_{x_i \leq x} P_X(X = x_i)$ . Nesse caso,  $f(x_i) = P_X(X = x_i)$  é chamada *função de probabilidade* de  $X$ ;
- (ii) *Absolutamente Contínua* é contínua se existe  $f : \mathbb{R} \rightarrow \mathbb{R}$  tal que  $P_X((a, b]) = F_c(b) - F_c(a) = \int_a^b f(t) dt$ . A função  $f$  é a *função de densidade de probabilidade* de  $X$ ;
- (iii) *Singular* se  $F_s$  é contínua com  $F'_s = 0$   $[\lambda]$  q.c. ( $F_s$  é singular com relação à medida de Lebesgue  $\lambda$ ).

**Resultado:** Toda f.d.  $F$  pode ser escrita como  $F = \alpha_1 F_d + \alpha_2 F_c + (1 - \alpha_1 + \alpha_2) F_s$ , com  $\alpha_1, \alpha_2 \geq 0$  tal que  $\alpha_1 + \alpha_2 \leq 1$ .

**Definição:** Seja  $(\Omega, \mathcal{A})$  um espaço mensurável e  $\mu_1$  e  $\mu_2$  medidas nesse espaço. Dizemos que  $\mu_2$  é *absolutamente contínua* com relação à  $\mu_1$  se,  $\forall A \in \mathcal{A}$ ,  $\mu_1(A) = 0 \Rightarrow \mu_2(A) = 0$ .

- Nesse caso, dizemos que  $\mu_2$  é dominada por  $\mu_1$  ou que  $\mu_1$  é uma medida dominante para  $\mu_2$  e denotamos  $\mu_2 \ll \mu_1$ .

**Teorema (de Radon-Nikodin):** Seja  $\mu_2 \ll \mu_1$  com  $\mu_1$   $\sigma$ -finita. Então,  $\exists f : \Omega \rightarrow [0, +\infty]$  tal que,  $\forall A \in \mathcal{A}$ ,

$$\mu_2(A) = \int_A f(x) d\mu_1(x).$$

Além disso, se  $g : \Omega \rightarrow \mathbb{R}$  é  $\mu_2$ -integrável, então

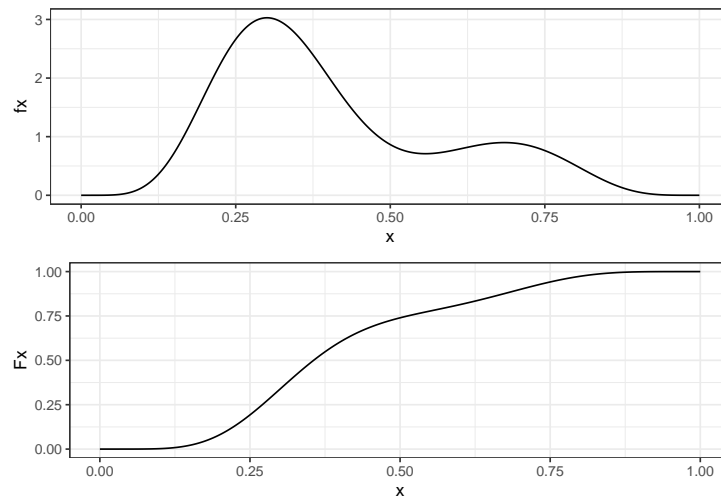
$$\int g(x) d\mu_2(x) = \int g(x) f(x) d\mu_1(x).$$

A função  $f = \frac{d\mu_2}{d\mu_1}$  é chamada de derivada de Radon-Nikodin da medida  $\mu_2$  com relação à medida  $\mu_1$  e é única  $[\mu_1]$  q.c. (ou seja, é única em todo conjunto  $\Omega$  com eventual excessão de um conjunto  $C$  tal que  $\mu_1(C) = 0$ ).

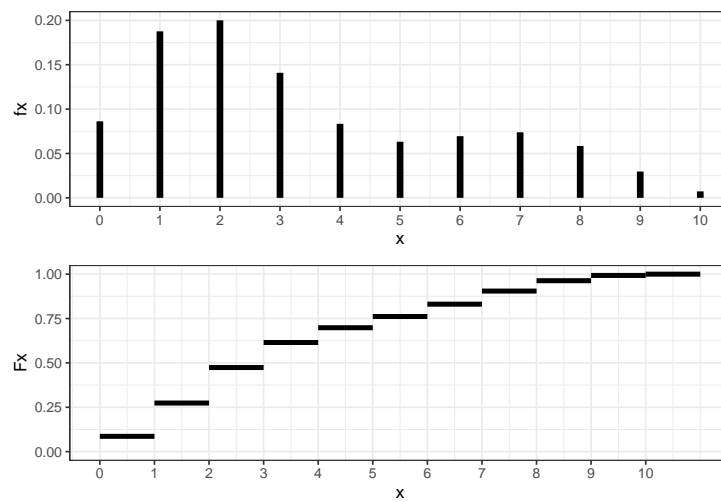
**Definição:**  $(\Omega, \mathcal{A}, P)$  espaço de probabilidade e  $(\mathfrak{X}, \mathcal{F}, \mu)$  espaço mensurável. Considere  $X : \Omega \rightarrow \mathfrak{X}$  uma v.a. e  $P_X$  a medida induzida por  $X$  de  $P$ , i.e.  $P_X(B) = P(X^{-1}(B))$ . Suponha que  $P_X \ll \mu$ . Então, a derivada de Radon-Nikodin  $f_X = \frac{dP_X}{d\mu}$  é chamada *densidade de  $X$  com respeito à  $\mu$* .

**Proposição:** Se  $h : \mathfrak{X} \rightarrow \mathbb{R}$  é mensurável e  $f_X = \frac{dP_X}{d\mu}$  é a densidade de  $X$  com respeito à  $\mu$ , então  $\int h(x) dP_X(x) = \int h(x) f_X(x) d\mu$ .

**Exemplo 1:** Seja  $\Omega = \mathfrak{X} = \mathbb{R}$  com a  $\sigma$ -álgebra de Borel e  $f$  uma função não negativa tal que  $\int f(x) dx = 1$ . Defina  $P(A) = \int_A f(x) dx$  e  $X(\omega) = \omega$ . Então,  $X$  é uma variável aleatória absolutamente contínua com função de densidade de probabilidade (f.d.p.)  $f$  e  $P_X = P$ . Além disso,  $P_X$  é absolutamente contínua com relação à medida de Lebesgue ( $P_X \ll \lambda$ ) e  $\frac{dP_X}{d\lambda} = f$ .



**Exemplo 2:** Seja  $\Omega = \mathbb{R}$  com a  $\sigma$ -álgebra de Borel,  $\mathfrak{X} = \{x_1, x_2, \dots\}$  um conjunto enumerável. Seja  $f$  uma função não negativa definida em  $\mathfrak{X}$  tal que  $\sum_{i=1}^{\infty} f(x_i) = 1$ . Defina  $P_X(A) = \sum_{\{i: x_i \in A\}} f(x_i)$ . Então  $X$  é uma variável aleatória discreta com função de probabilidade (f.d.p.)  $f$ . Além disso,  $P_X$  é absolutamente contínua com relação à medida de contagem ( $P_X \ll \nu$ ) e  $\frac{dP_X}{d\nu} = f$ .



**Resultado** Sejam  $(\Omega, \mathcal{A})$  espaço mensurável,  $P_1, P_2 : \mathcal{A} \rightarrow [0, 1]$  medidas de probabilidade,  $X : \Omega \rightarrow \mathbb{R}$  v.a. e  $P = \alpha P_1 + (1 - \alpha)P_2$  com  $0 \leq \alpha \leq 1$ . Então,

$$\int_{\Omega} X dP = \alpha \int_{\Omega} X dP_1 + (1 - \alpha) \int_{\Omega} X dP_2.$$

**Caso 1.**  $X$  simples,  $X = \sum_{i=1}^k X_i \mathbb{I}_{A_i}$ .

$$\begin{aligned} \int_{\Omega} X dP &= \sum_{i=1}^k x_i P(A_i) = \sum_{i=1}^k x_i [\alpha P_1(A_i) + (1 - \alpha)P_2(A_i)] = \\ &= \alpha \sum_{i=1}^k x_i P_1(A_i) + (1 - \alpha) \sum_{i=1}^k x_i P_2(A_i) = \alpha \int_{\Omega} X dP_1 + (1 - \alpha) \int_{\Omega} X dP_2. \end{aligned}$$

**Caso 2.**  $X \geq 0$ .

Considere a sequência  $(X_n)_{n \geq 1}$  tal que  $X_n \uparrow X$ ,  $X_n \geq 0$  simples. Então,

$$\begin{aligned} \int_{\Omega} X dP &= \lim_{n \rightarrow \infty} \int_{\Omega} X_n dP = \lim_{n \rightarrow \infty} \left\{ \alpha \int_{\Omega} X_n dP_1 + (1 - \alpha) \int_{\Omega} X_n dP_2 \right\} \\ &= \alpha \lim_{n \rightarrow \infty} \int_{\Omega} X_n dP_1 + (1 - \alpha) \lim_{n \rightarrow \infty} \int_{\Omega} X_n dP_2 = \alpha \int_{\Omega} X dP_1 + (1 - \\ &\alpha) \int_{\Omega} X dP_2. \end{aligned}$$

**Caso 3.**  $X$  qualquer.

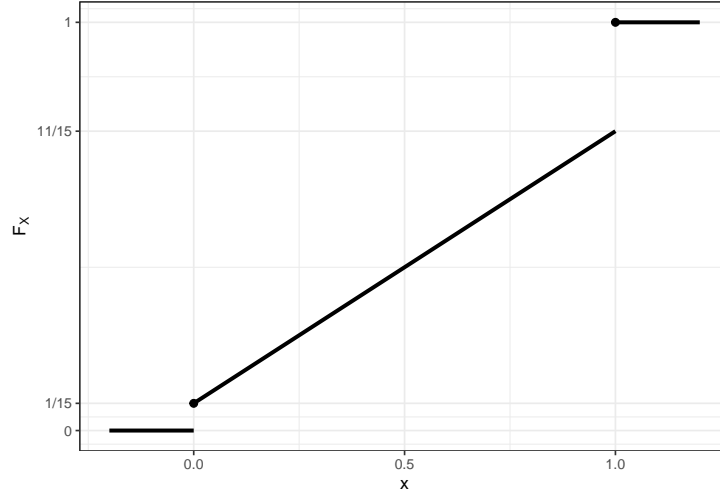
Basta escrever  $X = X^+ - X^-$  e repetir o procedimento anterior.

Seja  $P_1$  uma distribuição discreta com  $P_1(\{x_1, x_2, \dots\}) = 1$ ,  $P_2$  uma distribuição absolutamente contínua com função densidade de probabilidade  $f_x$  e  $X : \Omega \rightarrow \mathbb{R}$  tal que  $P_X(X \in A) = \alpha P_1(X^{-1}(A)) + (1 - \alpha)P_2(X^{-1}(A))$ . Então,

$$\begin{aligned} E(X) &= \int_{\Omega} X dP = \alpha \int_{\Omega} X dP_1 + (1 - \alpha) \int_{\Omega} X dP_2 = \alpha \sum_{i=1}^{\infty} x_i P_1(X = x_i) + (1 - \\ &\alpha) \int_{-\infty}^{\infty} x f_X(x) dx. \end{aligned}$$

**Exemplo.** Considere uma v.a.  $X$  com f.d. dada por

$$F_X(t) = \begin{cases} 0, & t < 0 \\ \frac{1}{15} + \frac{2}{3}t, & 0 \leq t < 1 \\ 1, & t \geq 1 \end{cases}$$



Temos que  $P(X = 0) = 1/15$ ,  $P(X = 1) = 4/15$  e, assim,  $P(0 < X < 1) = 10/15 = 2/3 = 1 - \alpha$ , de modo que

$$\frac{1}{15} = P(X = 0) = \alpha P_1(X = 0) = 1/3 P_1(X = 0) \Rightarrow P_1(X = 0) = \frac{1}{5} = 1 - P_1(X = 1).$$

$$E(X) = \alpha \int_{\Omega} X dP_1 + (1 - \alpha) \int_{\Omega} X dP_2 = \frac{1}{3} \left\{ 0 \cdot \frac{1}{5} + 1 \cdot \frac{4}{5} \right\} + \frac{2}{3} \int_0^1 x f_X(x) dx = \frac{1}{3} \cdot \frac{4}{5} + \frac{2}{3} \int_0^1 x dx = \frac{4}{15} + \frac{1}{3} = \frac{4}{15} + \frac{5}{15} = \frac{9}{15}.$$

## A.5 Probabilidade Condicional

**Motivação:**  $P(B|A) = \frac{P(A \cap B)}{P(A)}$  é bem definido se  $P(A) > 0$ .

Seja  $X, Y : \Omega \rightarrow \mathbb{R}$  v.a. tais que  $P_X([0, 1]) = 1$  e  $P_Y(\{0, 1\}) = 1$ . Considere um experimento em dois estagios onde seleciona-se  $X$  segundo uma distribuição absolutamente contínua  $F_X$  e, dado  $X = x$ ,  $0 \leq x \leq 1$ , uma moeda com probabilidade  $x$  é lançada  $n$  vezes. Nesse caso, é natural definir  $Y | X = x \sim \text{Bin}(n, x)$ , mesmo que  $P(X = x) = 0$ ,  $\forall x \in [0, 1]$ .

### Teorema da Medida Produto (para medidas de probabilidade)

Seja  $(\Omega_1, \mathcal{A}_1, P_1)$  um espaço de probabilidade e  $(\Omega_2, \mathcal{A}_2)$  um espaço mensurável.

Para cada  $\omega_1 \in \Omega_1$ , defina uma medida de probabilidade  $\mu(\omega_1, \cdot)$  em  $\mathcal{A}_2$ . Assuma também que, para cada  $B \in \mathcal{A}_2$ ,  $\mu(\cdot, B)$  é  $\mathcal{A}_1$ -mensurável. Então, existe uma única medida de probabilidade  $P$  em  $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$  tal que

$$P(A \times B) = \int_A \mu(\omega_1, B) dP_1(\omega_1), \quad \forall A \in \mathcal{A}_1, \forall B \in \mathcal{A}_2.$$

Se  $D(\omega_1)$  denota uma secção de  $D$  em  $\omega_1$ , isto é,  $D(\omega_1) = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in D\}$ ,  $D \in \mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ , então  $P(D) = \int_{\Omega_1} \mu(\omega_1, D(\omega_1)) dP_1(\omega_1)$ .

Voltando à probabilidade condicional, interprete (informalmente, por enquanto) a medida  $\mu(x, B)$  do teorema anterior como  $P(Y \in B | X = x)$ . Ainda informalmente, considere o evento  $\{X = x\}$ . Intuitivamente, a probabilidade que  $X \in (x, x + dx]$  é  $dF_X(x)$ . Então, sabendo que  $X = x$  ocorreu, o evento  $\{(X, Y) \in C\}$  ocorre se, e somente,  $Y \in C(x) = \{y : (x, y) \in C\}$  e a probabilidade desse evento é  $\mu(x, C(x))$ . Pela regra da probabilidade total,

$$P(C) = P(\{(X, Y) \in C\}) = \int_{\mathbb{R}} \mu(x, C(x)) dF(x).$$

Em particular, quando  $C = \{(x, y) : x \in A, y \in B\} = A \times B$ ,  $C(x) = B$  se  $x \in A$  e  $C(x) = \emptyset$  se  $x \notin A$ , então

$$P(C) = P(A \times B) = \int_A \mu(x, B) dF(x)$$

Se  $\mu(x, B)$  é mensurável em  $x$  para cada  $B \in \mathcal{B}(\mathbb{R})$ , então, pelo Teorema anterior,  $P$  é única.

**Exemplo 1.** Seja  $X \sim \text{Beta}(a, b)$  e  $Y | X = x \sim \text{Bin}(n, x)$

Considere  $(\Omega_1 = [0, 1], \mathcal{A}_1 = \mathcal{B}([0, 1]), P_X)$ , de modo que, para  $A \in \mathcal{A}_1$ ,

$$P_X(A) = \int_A \mathbb{1}_A dF_X(x) = \int_A f_X(x) dx = \int_A \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} dx.$$

Além disso, considere  $(\Omega_2 = \{0, 1, \dots, n\}, \mathcal{A}_2 = \mathcal{P}(\Omega_2))$  e, para cada  $x \in [0, 1]$ , defina  $\mu(x, B) = P(Y \in B | X = x)$ . Então, para  $k = 0, 1, \dots, n$ ,

$$\mu(x, \{k\}) = P(Y = k | X = x) = \binom{n}{k} x^k (1-x)^{n-k} \quad (\text{que é mensurável em } x).$$

Tomando  $\Omega = \Omega_1 \times \Omega_2$ ,  $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ ,  $P$  é a única medida de probabilidade determinada por  $P_X$  (ou  $F_X$ ) e  $\mu(x, \cdot)$ . Assim, para

$C \in \mathcal{A}$ ,

$$P(C) = \int_{\Omega_1} \mu(x, C(x)) dP_X = \int_0^1 \mu(x, C(x)) dF_X(x) = \int_0^1 \mu(x, C(x)) f_X(x) dx.$$

Por exemplo, se  $C = \Omega_1 \times \{k\}$ , temos

$$\begin{aligned} P(\Omega_1 \times \{k\}) &= P(\{X \in [0, 1], Y = k\}) = P_Y(Y = k) = \\ &= \int_0^1 P(Y = k | X = x) dF_X(x) = \int_0^1 P(Y = k | X = x) f_X(x) dx = \\ &= \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} dx = \binom{n}{k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)} \int_0^1 \frac{\Gamma(a+b+n)}{\Gamma(a+k)\Gamma(b+n-k)} x^{(a+k)-1} (1-x)^{(b+n-k)-1} dx \\ &= \binom{n}{k} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+k)\Gamma(b+n-k)}{\Gamma(a+b+n)} = \binom{n}{k} \frac{\beta(a+k, b+n-k)}{\beta(a, b)}. \end{aligned}$$

Nesse caso, diz-se que  $Y \sim \text{Beta-Bin}(n, a, b)$ .

**Teorema** Considere  $(\Omega, \mathcal{A}, P)$  e  $X : \Omega \rightarrow \mathfrak{X}$ ,  $\mathcal{F}$  uma  $\sigma$ -álgebra de  $\mathfrak{X}$  e  $B \in \mathcal{A}$ . Então existe  $g : \mathfrak{X} \rightarrow \mathbb{R}$  tal que, para cada  $A \in \mathcal{F}$ ,  $P(\{X \in A\} \cap B) = \int_A g(x) dP_X(x)$ .

Além disso,  $g$  é única  $[P_X]$  q.c., isto é,  $g(x) = P(B|X = x)$  é única  $[P_X]$  q.c. para um dado  $B \in \mathcal{A}$ .

**Demo:** segue diretamente do Teorema de Radon-Nikodin: se  $\mu(A) = P(\{X \in A\} \cap B)$  então  $\mu$  é medida finita em  $\mathcal{F}$  com  $\mu \ll P_X$ .

**Exemplo 2.** Seja  $\mathfrak{X} = \{x_1, x_2, \dots\}$  com  $p_i = P(\{X = x_i\}) > 0$ . Para  $i = 1, 2, \dots$ , considere a função  $g$ , uma “proposta” para  $P(B|\{X = x_i\})$ , definida por  $g(x_i) = \frac{P(B \cap \{X = x_i\})}{P(\{X = x_i\})}$ .

Seja  $A \in \mathcal{F} = \mathcal{P}(\mathfrak{X})$ , então

$$\begin{aligned} \int_A g(x) dP_X(x) &= \int_{\mathfrak{X}} g(x) \mathbb{1}_A(x) dP_X(x) = \sum_{i=1}^{\infty} g(x_i) \mathbb{1}_A(x_i) P_X(X = x_i) \\ &= \sum_{x_i \in A} g(x_i) P(\{X = x_i\}) = \sum_{x_i \in A} \frac{P(B \cap \{X = x_i\})}{P(\{X = x_i\})} P(\{X = x_i\}) \\ &= \sum_{x_i \in A} P(B \cap \{X = x_i\}) = P(\{X \in A\} \cap B). \end{aligned}$$

**Exemplo 3.** Considere agora  $\Omega = \mathbb{R}^2$ ,  $\mathcal{A} = \mathcal{B}(\mathbb{R}^2)$ ,  $X(x, y) = x$ ,  $Y(x, y) = y$  e  $(X, Y)$  vetor aleatório (absolutamente) contínuo com densidade conjunta  $f$ , isto é,  $P(A) = \int \int_A f(x, y) dx dy$ ,  $A \in \mathcal{A}$ . Nesse caso  $P(\{X = x\}) = 0$ ,  $\forall x$ .

Seja  $f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy$  a densidade marginal de  $X$  e defina  $f(y|x) = \frac{f(x, y)}{f_1(x)}$  como a densidade condicional de  $Y$  dado  $X = x$ .

Note que  $f(y|x)$  só está definido quando  $f_1(x) \neq 0$ . Contudo, se  $S = \{(x, y) : f_1(x) = 0\}$  então

$$\begin{aligned} P(\{(X, Y) \in S\}) &= \int \int_S f(x, y) dx dy = \int_{\{x: f_1(x)=0\}} \left[ \int_{-\infty}^{\infty} f(x, y) dy \right] dx \\ &= \int_{\{x: f_1(x)=0\}} f_1(x) dx = 0, \text{ de modo que } P(\{(X, Y) \in S\}) = 0 \text{ e} \\ &\text{podemos "ignorar" o conjunto onde } f(y|x) \text{ não está definida.} \end{aligned}$$

Se  $X = x$ ,  $\forall B \in \mathcal{A}$ ,  $B$  ocorre se, e somente se,  $Y \in B(x) = \{y : (x, y) \in B\}$ . Assim, considere a "proposta"

$$g(x) = P(\{Y \in B(x) | X = x\}) = \int_{B(x)} f(y|x) dy = \int_{-\infty}^{\infty} \mathbb{I}_B(x, y) f(y|x) dy.$$

Então, se  $A \in \mathcal{B}(\mathbb{R})$ ,

$$\begin{aligned} P(\{X \in A\} \cap B) &= \int \int_{\{x \in A; (x, y) \in B\}} f(x, y) dx dy = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} \mathbb{I}_B(x, y) f(y|x) dy \right] \mathbb{I}_A(x) f_1(x) dx \\ &= \int_A f_1(x) dx \underbrace{\int_{B(x)} f(y|x) dy}_{g(x)} = \int_A g(x) f_1(x) dx = \int_A g(x) dP_X(x). \end{aligned}$$

Portanto,  $g(x) = P(B|X = x)$ .

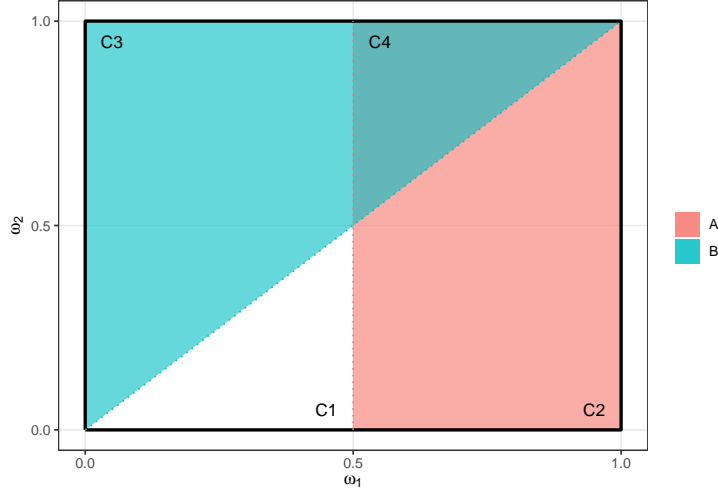
No exemplo anterior, as relações entre as densidades  $f(x, y) = f_1(x)f(y|x)$  ou, equivalentemente,  $f(x, y) = f_2(y)f(x|y)$ , podem ser usadas para obter a probabilidade condicional  $P(Y \in C | X = y) = \int_C f(y|x) dy$ ,  $C \in \mathcal{B}(\mathbb{R})$ . Além disso, para  $A, B \in \mathcal{B}(\mathbb{R})$ , existe uma única medida  $P$  satisfazendo

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A P(B|X = x) f_1(x) dx = \int_A \int_B f(y|x) f_1(x) dy dx = \\ &= \int_A \int_B f(x, y) dx dy = \int_B \int_A f(x|y) f_2(y) dx dy. \end{aligned}$$



**Exemplo 4. Esperança Condicional**

Seja  $(\Omega = [0, 1]^2, \mathcal{A} = \mathcal{B}([0, 1]^2), P = \lambda)$  e considere as partições apresentados na figura a seguir.



Defina as v.a.  $X$  e  $Y$  como

$$X(\omega) = \begin{cases} x_2, & \omega_1 \geq 1/2 \quad (A) \\ x_1, & \omega_1 < 1/2 \quad (A^c) \end{cases}$$

$$Y(\omega) = \begin{cases} y_2, & \omega_1 \leq \omega_2 \quad (B) \\ y_1, & \omega_1 > \omega_2 \quad (B^c) \end{cases}$$

$$P_X(x_2) = P(X^{-1}(\{x_2\})) = P(\omega \in A) = \lambda(A) = 1/2$$

$$P_Y(y_2) = P(Y^{-1}(\{y_2\})) = P(\omega \in B) = \lambda(B) = 1/2$$

$$\sigma_X = \{\emptyset, A, A^c, \Omega\} \subseteq \mathcal{B}([0, 1]^2) \text{ (é sub-}\sigma\text{-álgebra de } \mathcal{A})$$

$$\sigma_Y = \{\emptyset, B, B^c, \Omega\} \subseteq \mathcal{B}([0, 1]^2)$$

Seja  $Z(\omega) = (X(\omega), Y(\omega)) = (X, Y)(\omega)$ . Então,  $Z : \Omega \rightarrow \mathbb{R}^2$ , de

modo que  $Z(\omega) = \sum_{i=1}^4 z_i \mathbb{1}_{C_i}(\omega)$  é uma função simples com

$$Z(\omega) = \begin{cases} z_1 = (x_1, y_1), & \omega \in A^c \cap B^c = C_1 \\ z_2 = (x_2, y_1), & \omega \in A \cap B^c = C_2 \\ z_3 = (x_1, y_2), & \omega \in A^c \cap B = C_3 \\ z_4 = (x_2, y_2), & \omega \in A \cap B = C_4 \end{cases},$$

onde  $C_i = Z^{-1}(\{z_i\}) = \{\omega \in \Omega : (X(\omega), Y(\omega)) = z_i\}$ . Então,

$$P_Z(z_1) = P_Z(z_4) = P_Z((x_1, y_1)) = P_Z((x_2, y_2)) = \frac{1}{8} = \lambda(A^c \cap B^c) \\ = \lambda(A \cap B),$$

$$P_Z(z_2) = P_Z(z_3) = P_Z((x_2, y_1)) = P_Z((x_1, y_2)) = \frac{3}{8} = \lambda(A \cap B^c)$$

$$= \lambda(A^c \cap B) .$$

Pela que foi visto anteriormente, podemos definir

$$P_{Y|X=x_i}(Y = y_j | X = x_i) = \frac{P(\{Y = y_j, X = x_i\})}{P(\{X = x_i\})} = \begin{cases} \frac{1/8}{1/2} = \frac{1}{4}, & i = j \\ \frac{3/8}{1/2} = \frac{3}{4}, & i \neq j \end{cases},$$

e, assim,

$$E[Y | X = x_i] = \int y \, dP_{Y|x_i}(y) = \sum_{j=1}^2 y_j P(Y = y_j | X = x_i) .$$

Considere, por exemplo,  $x_1 = y_1 = 1$  e  $x_2 = y_2 = 2$ . Então,

$$E[Y|X = 1] = 1 \cdot \frac{1}{4} + 2 \cdot \frac{3}{4} = \frac{7}{4},$$

$$E[Y|X = 2] = 1 \cdot \frac{3}{4} + 2 \cdot \frac{1}{4} = \frac{5}{4} .$$

Deste modo, podemos definir uma nova v.a.

$$E[Y|X](\omega) = \begin{cases} 5/4, & \{\omega : X(\omega) = x_2\} = \{\omega \in A\} \\ 7/4, & \{\omega : X(\omega) = x_1\} = \{\omega \in A^c\} \end{cases} .$$

Note que a  $\sigma$ -álgebra gerada pela v.a.  $E[Y|X]$  coincide com a gerada por  $X$ ,  $\sigma_X$ . Dessa forma, podemos definir, de forma equivalente para esse caso, o *valor esperado de  $Y$  condicional à  $\sigma_X$*  por  $E[Y|X] = E[Y|\sigma_X]$ .

# Bibliography

- Albert, J. (2009). *Bayesian computation with R*. Springer.
- Ash, R. B. and Doleans-Dade, C. (2000). *Probability and Measure Theory*. Academic Press, California.
- Billingsley, P. (1986). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, second edition.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. MacGraw-Hill, New York.
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge Univ. Press.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Schervish, M. J. (2012). *Theory of Statistics*. Springer Science & Business Media.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- Shiryaev, A. N. (1996). *Probability*, volume 95 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition.