Inferência Bayesiana

Joaquim Neto
joaquim.neto@ufjf.edu.br
www.ufjf.br/joaquim_neto

Departamento de Estatística - ICE
Universidade Federal de Juiz de Fora

Versão 1.0 - 2010

Sumário

4. Revisão: normal multivariada

5. Introdução à inferência bayesiana
   - O paradigma bayesiano
   - Quem foi Thomas Bayes?
   - Teorema de Bayes
   - Densidade conjunta, marginal e condicional
   - Função de verossimilhança e distribuição a priori
   - A distribuição a posteriori e a inferência paramétrica
   - Núcleo e constante de proporcionalidade
   - Constante de proporcionalidade da posteriori
   - Verossimilhança marginal
   - Distribuição preditiva

6. Exemplos
   - Regressão linear simples
   - Regressão não linear
   - Ponto de mudança
   - Modelo hierárquico
   - Modelo linear generalizado
   - Análise de sobrevivência

Informações gerais

Contato

E-mail

joaquim.neto@ufjf.edu.br

Site pessoal

http://www.ufjf.br/joaquim_neto

Site do Departamento de Estatística (UFJF)

http://www.ufjf.br/estatistica

Referências Bibliográficas

Barry, R. James
(1981)
Probabilidade: um curso em cível intermediário.
Rio de Janeiro: Instituto de Matemática Pura e Aplicada (Projeto Euclides).

Degroot, M. H. & Schervish, M. J.
(2001)
Probability and Statistics, 3rd Edition, 3rd edn.
Addison Wesley.

Gamerman, D. & Lopes, H. F.
(2006)
Markov Chain Monte Carlo - Stochastic Simulation for Bayesian Inference, 2nd edn.
Chapman & Hall.

Pena, Sérgio Danilo
(2006)
Thomas.
Revista Ciência Hoje, **38,** no. 228, 22–29.

Turnbull, Bruce W., Bryon Wm. Brown, Jr. & HU, Marie
(1974)
Survivorship analysis of heart transplant data.
Journal of the American Statistical Association, **69,** no. 345, 74–80.

Introdução à probabilidade

Espaço amostral

Definição 1: *Suponhamos um experimento realizado sob certas condições fixas. O* **espaço amostral** Ω *do experimento é um conjunto que contém representações de todos os resultados possíveis, onde por "resultado possível", entende-se resultado elementar e indivisível do experimento.* Ω *deve satisfazer as seguintes condições:*

- *A todo resultado possível corresponde um, e somente um, elemento* $\omega \in \Omega$.
- *Resultados distintos correspondem a elementos distintos em* Ω*, ou seja,* $\omega \in \Omega$ *não pode representar mais de um resultado.*

Eventos

Quando se realiza um experimento, há certos eventos que ocorrem ou não. Por exemplo, ao jogar um dado e observar o resultado, alguns eventos são:

- observar um número par,
- observar o número 2,
- observar um número maior ou igual a 4,
- etc ...

Todo evento associado à um experimento pode ser identificado a um subconjunto do espaço amostral Ω. Reciprocamente, todo subconjunto A de Ω pode ser associado ao evento "resultado do experimento pertence a A". Assim, podemos associar

- o conjunto $\{2, 4, 6\}$ ao evento observar um número par,
- o conjunto $\{2\}$ ao evento observar um número 2
- e o conjunto $\{4, 5, 6\}$ ao evento observar um número maior ou igual a 4.

Definição 2: *Seja Ω o espaço amostral do experimento. Todo subconjunto $A \subset \Omega$ será chamado* **evento**. *Ω é o evento certo e $\emptyset$ é o evento impossível. Se $\omega \in \Omega$, o evento $\{\omega\}$ é dito* **elementar** *(ou* **simples**).

Definições de probabilidade

Definição clássica de probabilidade

Se Ω é finito, a definição clássica da probabilidade $P(A)$ de um evento $A \subset \Omega$ é dada por

$$P(A) = \frac{\#A}{\#\Omega} = \frac{\text{número de elementos de } A}{\text{número de elementos de } \Omega}.$$

Esta definição basea-se no conceito de resultados equiprováveis, ou melhor, no princípio da indiferença. Por exemplo, em um experimento que consiste em lançar um dado e observar o resultado, podemos usar $\Omega = \{1, 2, ..., 6\}$ e, diante da indiferença entre os resultados, temos $P(\{i\}) = \frac{1}{6}$, $\forall i \in \Omega$.

Exemplo 1: *Suponhamos um experimento que consiste em retirar uma carta em um baralho. Usando a definição clássica de probabilidade, qual é a probabilidade de tirar um 7?*

Solução: *Suponhamos que*

$$\Omega = \{A\heartsuit, 2\heartsuit, ..., J\clubsuit, K\clubsuit\}$$

é o nosso espaço amostral e que

$$A = \{7\clubsuit, 7\diamondsuit, 7\heartsuit, 7\spadesuit\}$$

é o nosso evento de interesse.

Assim,

$$P(A) = \frac{\#A}{\#\Omega} = \frac{4}{52}.$$

Definição frequentista de probabilidade

Outro método de definir a probabilidade $P(A)$ de um evento A é usando o limite da frequência relativa da ocorrência de A em n repetições independentes do experimento, com n tendendo ao infinito, ou seja,

$$P(A) = \lim_{n \to \infty} \frac{1}{n} \times \left(\begin{array}{c} \text{número de ocorrências de } A \text{ em } n \text{ realizações} \\ \text{independentes do experimento} \end{array} \right)$$

Esta é a **definição frequentista de probabilidade**.

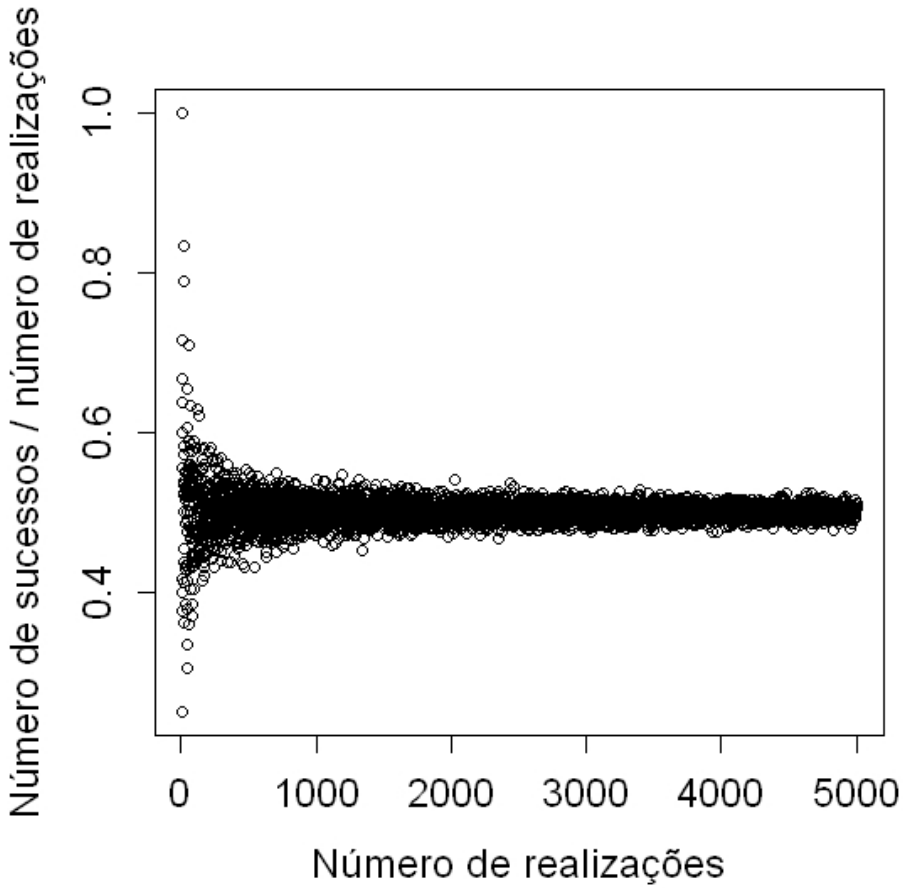


Figura: Número de arremessos de uma moeda honesta versos proporções de coroas obtidas.

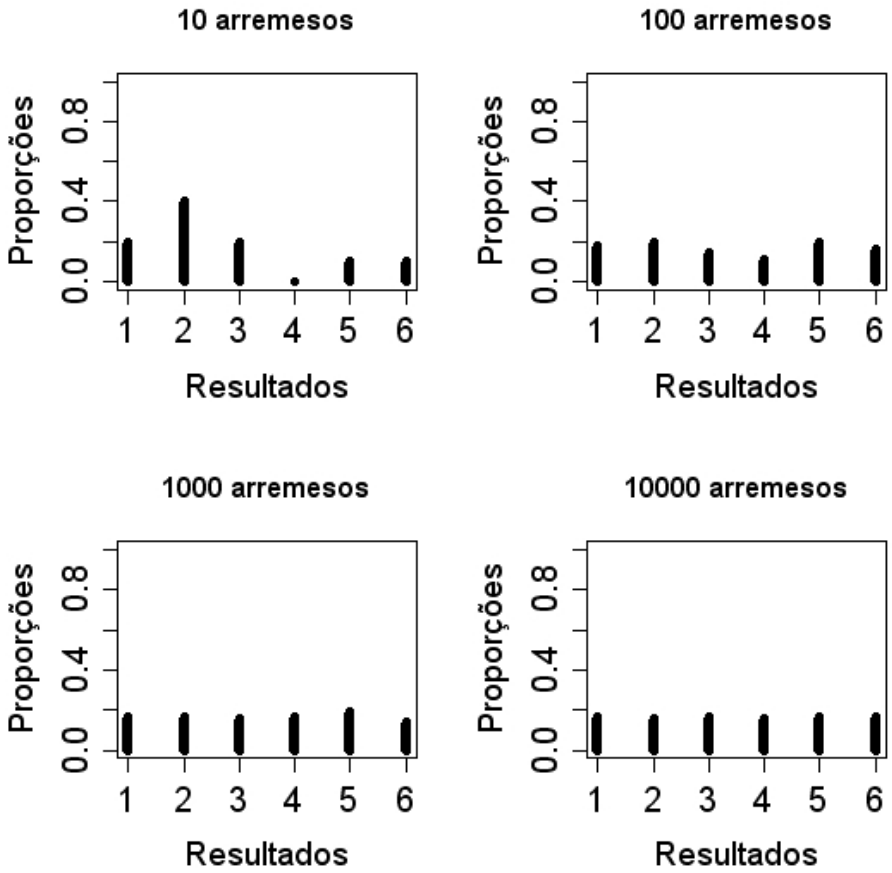


Figura: Proporção de resultados em 10, 100, 1000 e 10000 arremessos de um dado.

Probabilidade geométrica

Consideremos um experimento que consiste em escolher um ponto ao acaso no círculo unitário. Podemos definir a probabilidades $P(A)$ de um evento A como

$$P(A) = \frac{\text{área de } A}{\text{área de } \Omega} = \frac{\text{área de } A}{\pi}.$$

Acontece, que nem todo subconjunto de Ω tem área bem definida, ou seja, nem todo evento teria uma probabilidade (hipótese do contínuo).

Disjuntos 2 a 2 e conjunto das partes

Definição 3: *Os conjuntos* $A_1, A_2, \ldots$ *são* **disjuntos 2 a 2**, *se* $A_i \cap A_j = \emptyset$, $\forall i \neq j$.

Definição 4: *O* **conjunto das partes** $\mathscr{P}(A)$ *de um conjunto* A *é definido por*

$$\mathscr{P}(A) = \{B | B \subset A\}$$

Exemplo 2: *Se* $A = \{3, 5, 7\}$, *então*

$$\mathscr{P}(A) = \{\{3\}, \{5\}, \{7\}, \{3, 5\}, \{3, 7\}, \{5, 7\}, \{3, 5, 7\}, \emptyset\}$$

Definição 5: *Um evento A ao qual atribuímos probabilidade é chamado de evento aleatório.*

Axiomas de probabilidade

Não vamos nos preocupar, doravante, com o problema de como definir probabilidade para cada experimento. Simplesmente, vamos admitir que as probabilidades estão definidas em um certo conjunto $\mathscr{A}$[1] de eventos, chamados de eventos aleatórios. Vamos supor que a todo $A \in \mathscr{A}$ seja associado um número real $P(A)$, chamado de probabilidade de A, de modo que os axiomas a seguir sejam satisfeitos.

- **Axioma 1:** $P(A) \geq 0, \forall A \in \mathscr{A}$
- **Axioma 2:** $P(\Omega) = 1$
- **Axioma 3:** Se $A_1, A_2, ... \in \mathscr{A}$ são disjuntos 2 a 2, então

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P\left(A_n\right)$$

[1] Geralmente usamos $\mathscr{A} = \mathscr{P}(\Omega)$. Para saber mais sobre restrições de $\mathscr{A}$, consulte Barry (1981).

Espaço de probabilidade

Definição 6: *Um* **espaço de probabilidade**[2] *é um trio* $(\Omega, \mathscr{A}, P)$, *onde*

- Ω *é um conjunto não vazio e*
- P *é uma probabilidade em* $\mathscr{A}$.

[2] Para saber mais sobre restrições de $\mathscr{A}$, consulte Barry (1981).

Proposições

Proposição 1: $P(\emptyset) = 0$.

Prova: *Temos que*

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \emptyset \cup \emptyset \cup ...) \Rightarrow \\ P(\Omega) &= P(\Omega) + P(\emptyset) + P(\emptyset) + ... \Rightarrow \\ 0 &= P(\emptyset) + P(\emptyset) + ... \end{aligned}$$

Proposição 2: *Se* $A_1, A_2, ..., A_n \in \mathscr{A}$ *são disjuntos 2 a 2 então*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right)$$

Prova:

Outros resultados de probabilidade são:

- $P(A^c) = 1 - P(A)$
- $A_1 \subset A_2 \Rightarrow P(A_1) \leq P(A_2)$
- $0 \leq P(A) \leq 1$
- $P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P\left(A_i\right)$
- $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P\left(A_i\right)$
- $P(A_1 \cap A_2^c) = P(A_1) - P(A_1 \cap A_2)$

Probabilidade condicional

Definição 7: *Seja* $(\Omega, \mathscr{A}, P)$ *um espaço de probabilidade. Se* $B \in \mathscr{A}$ *e* $P(B) > 0$, *a probabilidade condicional de* $A \in \mathscr{A}$ *dado* B *é definida por*

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

OBS:

- Se $P(B) = 0$, $P(A|B)$ pode ser arbitrariamente definida. Mas, por independência, é conveniente fazer $P(A|B) = P(A)$, como veremos adiante.
- Decorre da definição que $P(A \cap B) = P(B)P(A|B)$, e esta igualdade também é válida quando $P(B) = 0$ (verifique!).

Exemplo 3: *Considere um experimento que consiste em retirar 2 cartas do baralho, ao acaso e sem reposição. Usando a definição clássica de probabilidade, qual a probabilidade de tirar um rei na segunda extração dado que foi obtido um rei na primeira extração?*

Passos para a solução:

- *Defina o espaço amostral.*
- *Defina os eventos de interesse.*
- *Use a fórmula de probabilidade condicional.*

Solução:

Teorema da Multiplicação

Teorema 1: *Seja* $(\Omega, \mathscr{A}, P)$ *um espaço de probabilidade com* $A_1, A_2, ..., A_N \in \mathscr{A}$. *Então*

$$\begin{aligned} P(A_1 \cap A_2 \cap ... \cap A_N) &= P(A_N | A_1 \cap ... \cap A_{N-1}) \\ &\times P(A_{N-1} | A_1 \cap ... \cap A_{N-2}) \\ &\times ... \times \\ &\times P(A_2 | A_1) P(A_1) \end{aligned}$$

Em particular, para $N = 2$, *temos*

$$P(A_1 \cap A_2) = P(A_2 \mid A_1) P(A_1) = P(A_1 \mid A_2) P(A_2).$$

Exemplo 4: *Considere um experimento que consiste em retirar 2 cartas do baralho, ao acaso e sem reposição. Usando a definição clássica de probabilidade, qual a probabilidade de tirar dois reis?*

Passos para a solução:

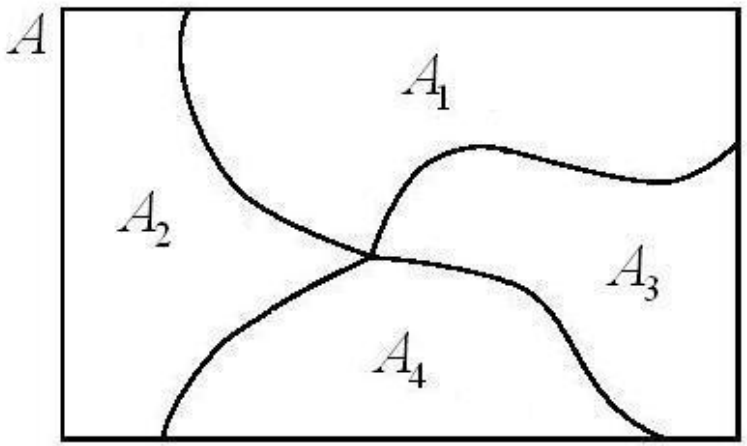
- *Defina o espaço amostral.*
- *Defina os eventos de interesse.*
- *Use o Teorema da Multiplicação.*

Solução:

Partição

Definição 8: *Uma sequência $A_1, A_2, \ldots$ finita ou enumerável de conjuntos é uma partição de um conjunto A quando*

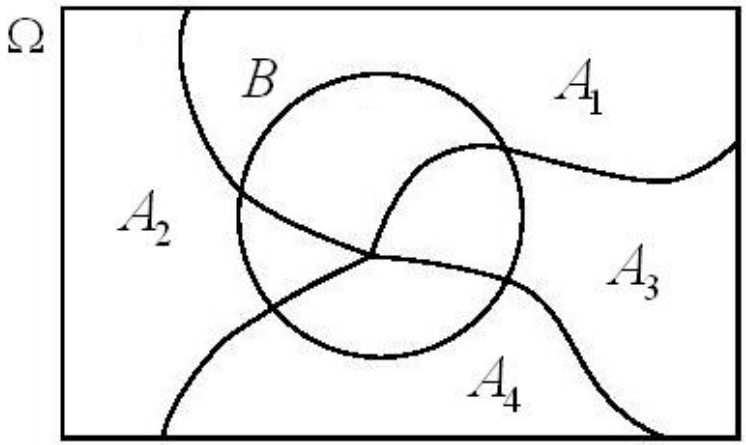
- *for uma sequência de conjuntos disjuntos 2 a 2 e*
- $\bigcup_i A_i = A$.



Teorema da Probabilidade Total

Teorema 2: *Seja* $(\Omega, \mathscr{A}, P)$ *um espaço de probabilidade. Se a sequência (finita ou enumerável)* $A_1, A_2, ..., A_N \in \mathscr{A}$ *formar uma partição de* Ω*, então*

$$P(B) = \sum_i P(B|A_i)\, P(A_i)$$



Exemplo 5: *Considere um experimento que consiste em retirar 2 cartas do baralho, ao acaso e sem reposição. Usando a definição clássica de probabilidade, qual a probabilidade de tirar um rei na segunda extração?*

Passos para a solução:

- *Defina o espaço amostral.*
- *Defina os eventos de interesse.*
- *Use o Teorema da Probabilidade Total.*

Solução:

Revisão: normal multivariada

Façamos uma breve revisão da distribuição normal multivariada.

Definição 9: *Seja Σ uma matriz $p \times p$ positiva definida. Dizemos que um vetor aleatório $Y = (Y_1, ..., Y_p)$ tem distribuição normal multivariada (de dimensão p) com vetor de médias $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ e matriz de covariâncias $\boldsymbol{\Sigma}$, se sua densidade for*

$$p(y \mid \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{p}{2}} (\det(\Sigma))^{-\frac{1}{2}} \exp\left(-0.5 (\mathbf{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\right),$$

onde $\mathbf{y} = (y_1, ..., y_p) \in \mathbb{R}^p$ e $det(\Sigma)$ é o determinante de Σ. Se $p = 1$, dizemos que a distribuição é normal univariada e, se $p = 2$, normal bivariada.

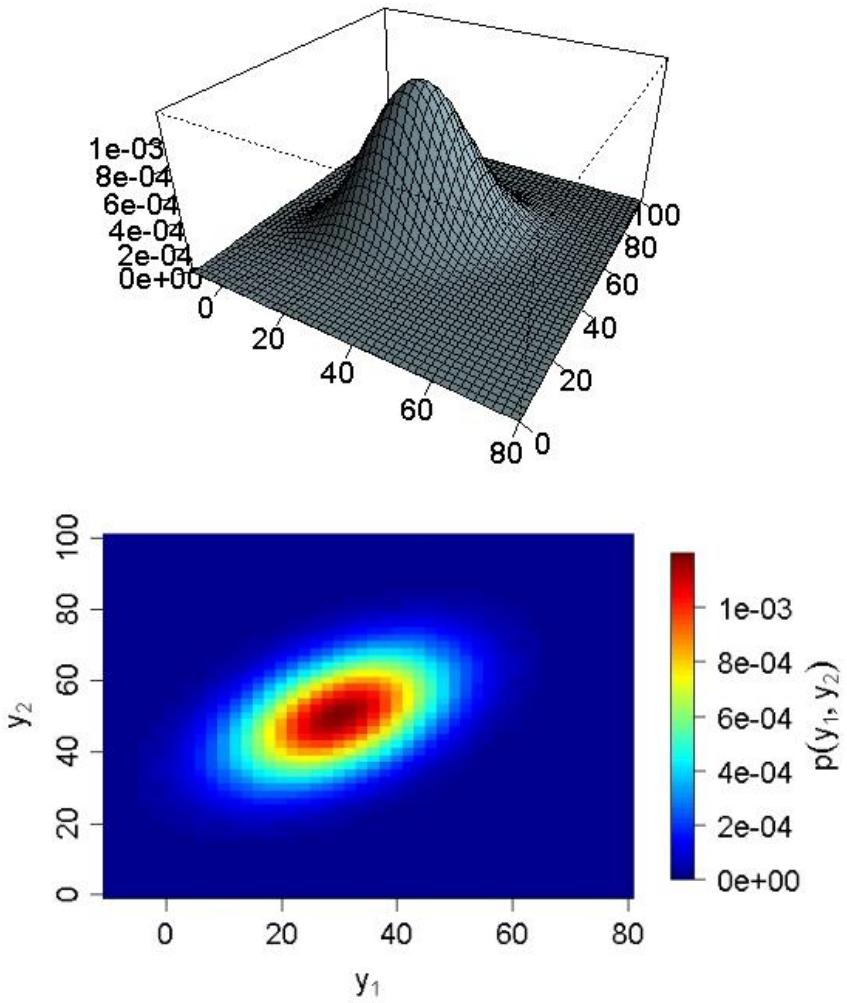


Figura: Densidade da normal bivariada com vetor de médias $\mu = [30, 50]^T$ e matriz de covariâncias $\Sigma = \begin{bmatrix} 150 & 70 \\ 70 & 150 \end{bmatrix}$

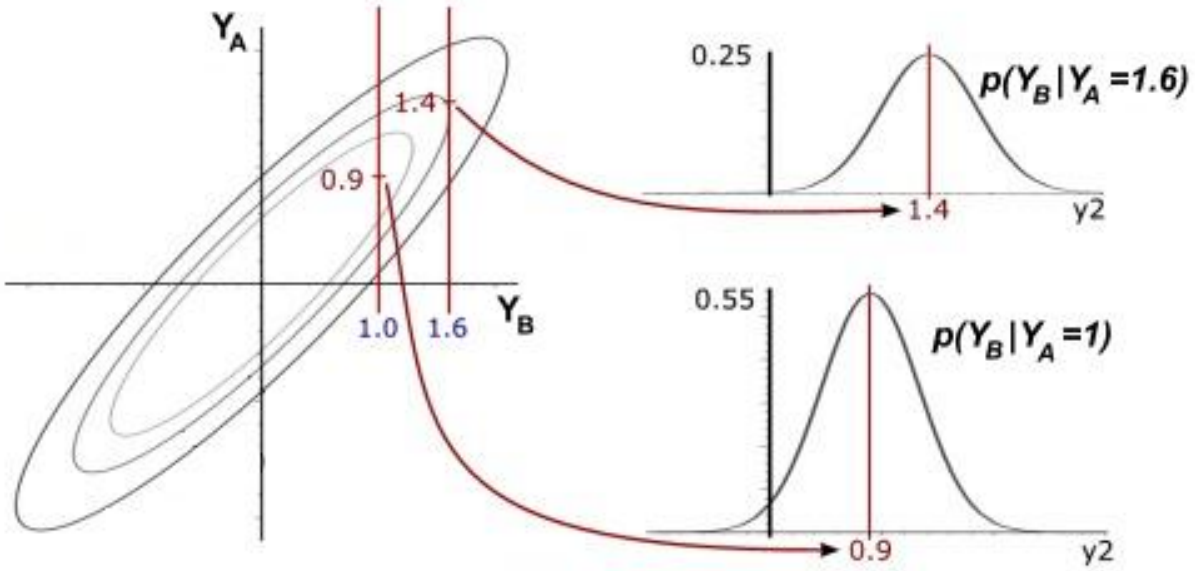


Figura: Distribuição condicional.

Suponhamos as partições

$$Y = \begin{bmatrix} Y_A \\ Y_B \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \quad \text{e} \quad \Sigma = \begin{bmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{bmatrix}.$$

A distribuição condicional de Y_B dado Y_A é normal multivariada com vetor de médias

$$\mu_{B|A} = \mu_B + \Sigma_{B,A}\Sigma_{A,A}^{-1}(Y_A - \mu_A)$$

e matriz de covariancias

$$\Sigma_{B|A} = \Sigma_{B,B} - \Sigma_{B,A}\Sigma_{A,A}^{-1}\Sigma_{A,B}.$$

Exercício 1: *Use o software R para plotar a densidade de uma distribuição normal bivariada qualquer.*

Introdução à inferência bayesiana

O paradigma bayesiano

O pároco de um vilarejo da Inglaterra do século 18, até certo ponto obscuro em seu tempo, é hoje festejado e considerado avançado nos meios científicos atuais, tudo por ter escrito um pequeno ensaio sobre probabilidade. O processo de raciocínio idealizado por Thomas Bayes neste texto, que ele mesmo sequer levou a público, é tido hoje como uma novo paradigma na estatística e como a base de uma verdadeira revolução em diferentes campos do conhecimento, da genética à teologia. Mas o que é o raciocínio bayesiano e por que vem ganhando tanto prestígio?

Ao visitar o campus de uma universidade norte-americana, é provável que encontre estudantes usando camisetas com a inscrição Bayes rules! A tradução para o português seria algo como "Bayes é o 'cara'!" (em inglês, a frase contém um trocadilho). Curioso, você decide checar quem é esse Bayes e o melhor lugar para isso é certamente a internet. Ao digitar o nome "Bayes" em uma página de busca, descobre-se que o nome completo dele é Thomas Bayes, que há um teorema de Bayes e que esse nome é citado (em junho de 2006) em nada menos que 9.3 milhões de páginas de internet! Se usarmos a palavra inglesa bayesian (bayesiano), o total de páginas sobe para 23.2 milhões.

Quem foi Thomas Bayes?

Considerando sua imensa importância na estatística, sabemos pouco sobre Thomas Bayes.

- Ele foi um reverendo presbiteriano que viveu no início do século 18 na Inglaterra.
- Estudou teologia na Universidade de Edimburgo (Escócia), de onde saiu em 1722.
- Em 1731 assumiu a paróquia de Tunbridge Wells, no condado de Kent, a 58 km de Londres. Neste mesmo ano, apareceu na Inglaterra um livro anônimo, hoje creditado a Bayes, chamado Benevolência divina.
- Cinco anos depois, publicou seu primeiro e único livro de matemática, chamado "*The doctrine of fluxions*" (A doutrina dos fluxions[3]).
- Com base nesse livro e em outras possíveis contribuições sobre as quais não temos dados precisos, Bayes foi eleito em 1752 para a Real Sociedade, entidade científica britânica criada em 1645.
- Dois anos após sua morte, um amigo, o filósofo Richard Price (1723-1791), apresentou à Real Sociedade um artigo que aparentemente encontrou entre os papéis do reverendo, com o nome "*An essay towards solving a problem in the doctrine of chances*" (Ensaio buscando resolver um problema na doutrina das probabilidades). Neste artigo estava a demonstração do famoso teorema de Bayes. Após sua publicação, o trabalho caiu no esquecimento, do qual só foi resgatado pelo matemático francês Pierre-Simon de Laplace (1749-1827), que o revelou ao mundo.

[3] O nome *fluxion* foi dado pelo matemático e físico Isaac Newton (1642-1727) para a derivativa de uma função contínua (que Newton chamava de *fluent*).

Figura: O reverendo Thomas Bayes na única representação que existe dele.

Teorema de Bayes

Teorema 3: *Seja* $(\Omega, \mathscr{A}, P)$ *um espaço de probabilidade. Se a sequência (finita ou enumerável)* $A_1, A_2, ..., A_N \in \mathscr{A}$ *formar uma partição de* Ω*, então*

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_j P(B|A_j)\,P(A_j)}$$

Exemplo 6 (probabilidade subjetiva): *Uma pessoa vai ao médico reclamando de dores. Após uma detalhada conversa e um cuidadoso exame físico, o médico acredita que o paciente pode ter uma determinada doença.*

Seja θ uma quantidade desconhecida que indica se o paciente tem a doença ou não. Se ele possui a doença então $\theta = 1$, caso contrário $\theta = 0$. Subjetivamente, o médico assume que $P(\theta = 1|H) = 0.6$, onde H representa toda a informação disponível até a consulta. Para simplificar, iremos omitir H fazendo $P(\theta = 1|H) = P(\theta = 1) = 0.6$.

Um pouco antes do fim da consulta, o médico prescreve um exame laboratorial. Seja X uma variável associada ao resultado deste exame, de modo que $X = 1$ se o exame acusa a doença e $X = 0$ caso contrário. O exame fornece um resultado incerto com as seguintes probabilidades

$$P(X = 1 \mid \theta = 0) = 0.10 \quad e \quad P(X = 1 \mid \theta = 1) = 0.95$$

Suponhamos que o resultado do exame tenha acusado a doença, $X = 1$. Assim, para o médico, a probabilidade do paciente ter a doença passa a ser

$$\begin{aligned} P(\theta = 1 \mid X = 1) &= \frac{P(X = 1 \mid \theta = 1)P(\theta = 1)}{P(X = 1 \mid \theta = 1)P(\theta = 1) + P(X = 1 \mid \theta = 0)P(\theta = 0)} \\ &= \frac{0.95 \times 0.6}{0.95 \times 0.6 + 0.1 \times 0.4} = 0.9344262. \end{aligned}$$

Exemplo 7: *Recomenda-se que, a partir dos 40 anos, as mulheres façam mamografias anuais. Nesta idade, 1% das mulheres são portadoras de um tumor assintomático de mama.*

Seja θ uma quantidade desconhecida que indica se uma paciente (desta faixa etária) tem a doença ou não. Se ela possui a doença, então $\theta = 1$, caso contrário, $\theta = 0$. Assim, podemos assumir que

$$P(\theta = 1) = 0.01 \quad e \quad P(\theta = 0) = 0.99.$$

Sabe-se que a mamografia indica a doença em 80% das mulheres com câncer de mama, mas esse mesmo resultado ocorre também com 9.6% das mulheres sem o câncer. Assim, seja X uma variável aleatória associada ao resultado da mamografia, de modo que se $X = 1$ o exame acusou a doença e $X = 0$ caso contrário. Temos então que

$$\begin{aligned} P(X = 1 \mid \theta = 0) &= 0.096 \\ P(X = 1 \mid \theta = 1) &= 0.80 \end{aligned}$$

Imagine agora que você encontra uma amiga de 40 e poucos anos aos prantos, desesperada, porque fez uma mamografia de rotina e o exame acusou a doença. Qual a probabilidade de ela ter um câncer de mama?

Solução: Temos que

$$\begin{aligned} P(\theta = 1 \mid X = 1) &= \frac{P(X = 1 \mid \theta = 1)P(\theta = 1)}{P(X = 1 \mid \theta = 1)P(\theta = 1) + P(X = 1 \mid \theta = 0)P(\theta = 0)} \\ &= \frac{0.80 \times 0.01}{0.80 \times 0.01 + 0.096 \times 0.99} = 0.07763975 \end{aligned}$$

Logo, a probabilidade dela ter a doença é de aproximadamente 7.8%.

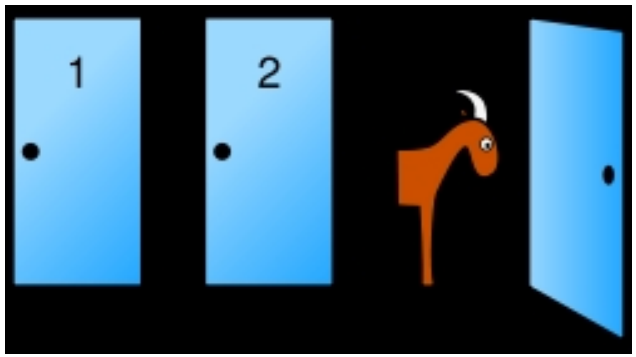
Ao apresentar este problema a várias pessoas, inclusive estudantes de medicina, observa-se uma tendência a superestimar a probabilidade a posteriori da doença. Isso revela que o raciocínio bayesiano não é intuitivo. Parece haver uma tendência geral a ignorar o fato de que a probabilidade a priori de doença é pequena, fenômeno denominado "falácia da probabilidade de base" pelo psicólogo norte-americano (de origem israelense) Daniel Kahneman, premiado com o Nobel de Economia em 2002 por estudos sobre o comportamento de investidores. Num sentido específico: "as pessoas não são racionais".

Exemplo 8: *O* **problema de Monty Hall** *é um problema matemático que surgiu a partir de um concurso televisivo dos Estados Unidos da América chamado Let's Make a Deal, exibido na década de 1970.*

O jogo consiste no seguinte: Monty Hall (o apresentador) apresentava 3 portas aos concorrentes, sabendo que atrás de uma delas está um carro (prêmio bom) e que as outras têm prêmios de pouco valor.

1. *Na 1ª etapa o concorrente escolhe uma porta (que ainda não é aberta).*
2. *Em seguida, Monty abre uma das outras duas portas que o concorrente não escolheu, sabendo que o carro não se encontra nela.*
3. *Agora, com duas portas apenas para escolher e sabendo que o carro está atrás de uma delas, o concorrente tem que se decidir se permanece com a porta que escolheu no início do jogo e abre-a ou se muda para a outra porta que ainda está fechada para então a abrir.*

Neste caso, existe uma estratégia mais lógica? Ficar com a porta escolhida inicialmente ou mudar de porta? Será que com alguma das portas ainda fechadas o concorrente tem mais probabilidades de ganhar? Por que?



O problema de Monty Hall demonstra muito bem como nossa intuição é falha em certos problemas que envolvem chances. Felizmente, pode-se resolver o problema de Monty Hall de forma simples e sem erro usando o teorema de Bayes.

Justificativa pelo Teorema de Bayes: Consideremos os eventos

- A_1 = "Carro está na primeira porta",
- A_2 = "Carro está na segunda porta",
- A_3 = "Carro está na terceira porta" e
- C = "O apresentador abre a terceira porta".

Naturalmente, iremos assumir $P(C \mid A_1) = 0.5$, $P(C \mid A_2) = 1$ e $P(C \mid A_3) = 0$. Assim, pelo teorema da probabilidade total, temos

$$\begin{aligned} P(C) &= P(C|A_1)\,P(A_1) + P(C|A_2)\,P(A_2) + P(C|A_3)\,P(A_3) = \\ &= \frac{1}{2}\cdot\frac{1}{3} + 1\cdot\frac{1}{3} + 0\cdot\frac{1}{3} = \frac{1}{2} = 0.5 \end{aligned}$$

Agora, usando o teorema de Bayes, temos

$$\begin{aligned} P(A_1 \mid C) &= \frac{P(C \mid A_1)P(A_1)}{P(C)} = \frac{\frac{1}{2}\times\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}, \\ P(A_2 \mid C) &= \frac{P(C \mid A_2)P(A_2)}{P(C)} = \frac{1\times\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} \quad \text{e} \\ P(A_3 \mid C) &= \frac{P(C \mid A_3)P(A_3)}{P(C)} = \frac{0\times\frac{1}{3}}{\frac{1}{2}} = 0. \end{aligned}$$

Portanto, escolhendo trocar de porta a chance de ganhar o carro é maior.

Densidade conjunta, marginal e condicional

Sejam $\mathbf{X}$ e $\mathbf{Y}$ vetores aleatórios com densidades $p(x)$ e $p(y)$, respectivamente. Suponhamos ainda que $p(\mathbf{x}, \mathbf{y})$ é a densidade conjunta do vetor $(\mathbf{X}, \mathbf{Y})$ e $p(\mathbf{x} \mid \mathbf{y})$ é a densidade da distribuição de $\mathbf{Y}$ dado $\mathbf{X}$. Uma importante equação relaciona estas densidades:

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{y})} \quad \text{(probabilidade condicional)}.$$

Já a densidade da distribuição marginal de $\mathbf{X}$ pode ser obtida com

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y})\, d\mathbf{y} = \int p(\mathbf{x} \mid \mathbf{y})\, p(\mathbf{y})\, d\mathbf{y} \quad \text{(teorema da probabilidade total)}.$$

Temos ainda a seguinte relação

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{y})\, p(\mathbf{y})}{\int p(\mathbf{x} \mid \mathbf{y})\, p(\mathbf{y})\, d\mathbf{y}} \quad \text{(teorema de Bayes)}.$$

OBS: Para distribuições discretas, as integrais acima assumem a forma de somatórios.

Função de verossimilhança e distribuição a priori

Suponhamos agora que um modelo probabilístico é utilizado para representar matematicamente a aleatoriedade inerente às observações $y_1, ..., y_n$. Suponhamos ainda que este modelo depende de um vetor paramétrico $\boldsymbol{\theta}$. Em outras palavras, consideremos um problema típico de inferência paramétrica.

Seguindo o paradigma bayesiano, informações prévias sobre $\boldsymbol{\theta}$ são representadas matematicamente usando uma distribuição de probabilidade, chamada de **distribuição a priori** (ou simplesmente **priori**), que estabelece (pondera) quais valores de $\boldsymbol{\theta}$ são mais prováveis, segundo informações disponíveis antes de conhecer as observações. Uma distribuição a priori deve então representar a informação do pesquisador sobre $\boldsymbol{\theta}$ antes de conhecer as observações.

OBS: A priori não é uma distribuição para $\boldsymbol{\theta}$, uma vez que este é fixo, mas sim uma distribuição que representa a incerteza do pesquisador diante do valor desconhecido $\boldsymbol{\theta}$. No entanto, num abuso de linguagem e notação, é comum dizermos "priori para $\boldsymbol{\theta}$" e usarmos $p(\boldsymbol{\theta})$ para a densidade de $\boldsymbol{\theta}$, por exemplo.

A distribuição a posteriori e a inferência paramétrica

A distribuição condicional de $\boldsymbol{\theta}$ dado um conjunto de observações $y_1, ..., y_n$ é chamada de **distribuição a posteriori** (ou simplesmente **posteriori**) de $\boldsymbol{\theta}$. A densidade ou função de probabilidade da posteriori será denotada por $p(\boldsymbol{\theta} \mid y_1, ..., y_n)$.

Pelo teorema de Bayes, temos que a posteriori pode ser obtida com

$$\begin{aligned} p\left(\boldsymbol{\theta}|y_1, ..., y_n\right) &= \frac{p\left(y_1, ..., y_n \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{p\left(y_1, ..., y_n\right)} \\ &= \frac{p\left(y_1, ..., y_n|\theta\right) p\left(\boldsymbol{\theta}\right)}{\int p\left(y_1, ..., y_n \mid \boldsymbol{\theta}\right) p\left(\theta\right) d\boldsymbol{\theta}} \end{aligned}$$

OBS: Naturalmente, no caso discreto a integral acima assume a forma de um somatório.

Núcleo e constante de proporcionalidade

Definição 10: *Se uma densidade (ou função de probabilidade)* $p(\mathbf{x} \mid \boldsymbol{\theta})$ *for escrita como* $p(\mathbf{x} \mid \boldsymbol{\theta}) = f(\boldsymbol{\theta})g(\mathbf{x}, \boldsymbol{\theta})$, *onde* f *é uma função de* $\boldsymbol{\theta}$ *(apenas) e* g *é uma função de* $(\mathbf{x}, \boldsymbol{\theta})$, *diremos que* $g(\mathbf{x}, \boldsymbol{\theta})$ *e* $f(\boldsymbol{\theta})$ *são, respectivamente, um* **núcleo** *e uma* **constante de proporcionalidade** *de* $p(\mathbf{x} \mid \boldsymbol{\theta})$.

A partir do núcleo de uma densidade (ou f.p.) $p(\mathbf{x} \mid \boldsymbol{\theta})$, podemos obter a constante de proporcionalidade via integração. Para isto, basta lembrar que toda densidade (ou f.p.) deve integrar 1, ou seja, basta fazer

$$\int p(\mathbf{x} \mid \boldsymbol{\theta})\, d\mathbf{x} = 1 \Rightarrow \int f(\boldsymbol{\theta})\, g(\mathbf{x}, \boldsymbol{\theta})\, d\mathbf{x} = 1 \Rightarrow f(\boldsymbol{\theta}) = \frac{1}{\int g(\mathbf{x}, \boldsymbol{\theta})\, d\mathbf{x}}$$

Assim, podemos dizer que "o núcleo contém toda a informação de uma distribuição". A identificação de núcleos será útil mais adiante para reconhecer prioris conjugadas.

Constante de proporcionalidade da posteriori

Por $1/p(y_1, ..., y_n)$ não depender de $\boldsymbol{\theta}$, dizemos que esta quantidade é uma **constante de proporcionalidade da posteriori** (que denotaremos por k) . Assim, observe que a posteriori é proporcional ao produto da verossimilhança pela priori:

$$p(\boldsymbol{\theta} \mid y_1, ..., y_n) = kp(y_1, ..., y_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) \propto \underbrace{p(y_1, ..., y_n \mid \boldsymbol{\theta})}_{\text{verossimilhança}} \underbrace{p(\boldsymbol{\theta})}_{\text{priori}}$$

A constante de proporcionalidade pode ser recuperada com

$$k = \left(\int p(y_1, ..., y_n|\theta)\, p(\theta) d\theta \right)^{-1}.$$

Verossimilhança marginal

Definição 11: *A densidade (ou função de probabilidade)* $p(y_1, ..., y_n)$ *é chamada de* **verossimilhança marginal** *e pode ser obtida com*

$$p\left(y_1, ..., y_n\right) = \int p\left(y_1, ..., y_n|\theta\right) p\left(\theta\right).$$

Distribuição preditiva

Em muitas aplicações como, por exemplo, em séries temporais e geoestatística, o maior interesse é prever um processo em pontos não observados do tempo ou espaço. Suponha então que, após observar $y_1, ..., y_n$, estamos interessados na previsão de quantidades $Y_1^*, ..., Y_p^*$, também relacionadas à $\boldsymbol{\theta}$ e descritas probabilisticamente por uma distribuição $(Y_1^*, ..., Y_p^* \mid y_1, ..., y_n, \boldsymbol{\theta})$.

Definição 12: *A distribuição de* $(Y_1^*, ..., Y_p^* \mid y_1, ..., y_n)$ *é chamada de* **distribuição preditiva** *e sua densidade (ou f.p.) pode ser obtida por integração com*

$$\begin{aligned} p\left(y_1^*, ..., y_p^* | y_1, ..., y_n\right) &= \int p\left(y_1^*, ..., y_p^*, \theta | y_1, ..., y_n\right) d\theta \\ &= \int p\left(y_1^*, ..., y_p^* | \theta, y_1, ..., y_n\right) p\left(\theta | y_1, ..., y_n\right) d\theta. \end{aligned} \tag{1}$$

Em muitos problemas estatísticos a hipótese de independência condicional[4] entre $(Y_1, ..., Y_n)$ e $(Y_1^*, ..., Y_p^*)$ dado θ está presente e a distribuição preditiva pode ser representada por

$$p\left(y_1^*, ..., y_p^* \mid y_1, ..., y_n\right) = \int p\left(y_1^*, ..., y_p^* | \theta\right) p\left(\theta | y_1, ..., y_n\right) d\theta. \quad (2)$$

OBS: Em muitas aplicações práticas, a integral em (1) (ou (2)) não tem solução analítica e precisa ser obtida por algum método de aproximação.

[4] Esta hipótese de independência condicional não é uma hipótese razoável para dados espacialmente distribuídos, onde admite-se que exista alguma estrutura de correlação no espaço.

Exemplos

Exemplo 9 (regressão linear simples): *Um biólogo investiga o efeito de diferentes quantidades de fertilizante na produção de grama em solo calcário. Dez áreas de 1 m^2 foram escolhidas ao acaso e diferentes quantidades do fertilizante foram aplicadas a cada área. Dois meses depois, as produções de grama foram anotadas. Os dados desta investigação são apresentados na tabala abaixo.*

Massa de fertilizante (g/m^2)	**Produção de grama** (g/m^2)
25	84
50	80
75	90
100	154
125	148
150	169
175	206
200	244
225	212
250	248

Considere o modelo de regressão linear descrito por

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \quad e \\ \mu_i &= \alpha X_i + \beta. \end{aligned}$$

Assuma a priori que

$$\begin{aligned} \alpha &\sim N(0, 10^6), \\ \beta &\sim N(0, 10^6) \text{ e} \\ \phi &= \frac{1}{\sigma^2} \sim Ga(0.1, 0.1). \end{aligned}$$

a) Construa o gráfico de dispersão com as massas de fertilizante no eixo das abscissas e as produções de grama no eixo das ordenadas.
b) Construa uma tabela com as médias a posteriori e intervalos de 95% de credibilidade a posteriori para os parâmetros α, β e σ^2.
c) Acrescente ao gráfico de dispersão obtido no item (a) a função $f(x) = \hat{\alpha}x + \hat{\beta}$, onde $\hat{\alpha}$ e $\hat{\beta}$ são, respectivamente, a média a posteriori de α e β.
d) Compare as médias a posteriori com as estimativas de máxima verossimilhança dos parâmetros α, β e σ^2.
e) Usando 15 g/m^2 de fertilizante, qual é o intervalo de 95% de credibilidade para a produção de grama em 1 m^2.
d) Suponha que a grama foi plantada em 50 m^2 usando 15 g/m^2 de fertilizante. Qual é o intervalo de 95% de credibilidade para a produção de grama em toda a área plantada.

Exemplo 10 (regressão não linear): *Dugongues são animais da mesma ordem dos peixes-bois (ordem Sirenia). Dentre suas principais características, temos: narinas no topo da cabeça, lábio superior voltado para baixo e nadadeira dividida em duas partes (como a das baleias e golfinhos). Os Dugongues são encontrados na costa leste da África, Índia, Indonésia, Malásia e Austrália.*

Figura: Foto de um dugongue

A tabela abaixo exibe tamanhos e idades de 27 Dugongues.

Idades x_i	**Tamanhos** y_i	**Idades** x_i	**Tamanhos** y_i
1	1.80	10	2.50
1.5	1.85	12	2.32
1.5	1.87	12	2.32
1.5	1.77	13	2.43
2.5	2.02	13	2.47
4	2.27	14.5	2.56
5	2.15	15.5	2.65
5	2.26	15.5	2.47
7	2.47	16.5	2.64
8	2.19	17	2.56
8.5	2.26	22.5	2.70
9	2.40	29	2.72
9.5	2.39	31.5	2.57
9.5	2.41		

Carlin & Gelfand (1991) propõem o seguinte modelo para os dados da tabela:

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \alpha - \beta\gamma^{x_i} \end{aligned}$$

a) Construa um gráfico de dispersão com as idades no eixo das abscissas e os tamanhos no eixo das ordenadas.
b) Construa uma tabela as médias a posteriori e intervalos de 95% de credibilidade a posteriori para os parâmetros α, β, γ e σ^2.
c) Construa um gráfico com as idades no eixo das abscissas e médias a posteriori dos parâmetros $\mu_1, ..., \mu_{27}$ no eixo das ordenadas. Neste mesmo gráfico, exiba os intervalos de credibilidade destes parâmetros.

Exemplo 11 (ponto de mudança): *Sejam $Y_1, Y_2, ..., Y_n$ uma sequência de variáveis aleatórias com distribuição de Poisson e y_i uma observação de Y_i, $\forall i \in \{1, ..., n\}$. Suponhamos ainda que existe uma suspeita de "mudança de ponto", ou seja, suspeita-se que, para algum $m \in \{1, ..., m\}$, a sequência $Y_1, ..., Y_m$ têm média λ_1 e a sequência $Y_{m+1}, ..., Y_n$ têm média λ_2. Com as $n = 112$ observações $y_1, ..., y_{112}$ apresentadas na tabela abaixo, estime os parâmetros λ_1, λ_2 e m assumindo as distribuições a priori independentes: $\lambda \sim Ga(0.1, 0.1)$, $\phi \sim Ga(0.1, 0.1)$ e m uniformemente distribuído em $\{1,...,n\}$.*

4	5	4	1	0	4	3	4	0	6	3	3	4	0
2	6	3	3	5	4	5	3	1	4	4	1	5	5
3	4	2	5	2	2	3	4	2	1	3	2	2	1
1	1	1	3	0	0	1	0	1	1	0	0	3	1
0	3	2	2	0	1	1	1	0	1	0	1	0	0
0	2	1	0	0	0	1	1	0	2	3	3	1	1
2	1	1	1	1	2	4	2	0	0	0	1	4	0
0	0	1	0	0	0	0	0	1	0	0	1	0	1

Tabela: Observações $y_1, ..., y_{112}$

Exemplo 12: *(modelo hierárquico) Souza (1999) considera modelos hierárquicos para descrever o ganho de peso de 68 mulheres grávidas que visitaram, de 5 à 7 vezes, o Instituto de Puericultura e Pediatria Martagão Gesteira da Universidade Federal do Rio de Janeiro. Os dados deste exemplo podem ser obtidos na página*

http://www.ufjf.br/joaquim_neto .

Um dos modelos propostos por Souza (1999) é descrito por

$$Y_{i,j} \mid \alpha_i, \beta_i, \sigma^2 \sim N\left(\alpha_i + \beta_i X_{i,j}, \sigma^2\right),$$
$$\alpha_i \sim N\left(\mu_\alpha, \sigma_\alpha^2\right), \quad \beta_i \sim N\left(\mu_\beta, \sigma_\beta^2\right),$$
$$\mu_\alpha \sim N\left(0, 1000\right), \quad \mu_\beta \sim N\left(0, 1000\right),$$
$$\sigma^{-2} \sim Ga\left(0.1, 0.1\right), \quad \sigma_\alpha^{-2} \sim Ga\left(0.1, 0.1\right) \quad e \quad \sigma_\beta^{-2} \sim Ga\left(0.1, 0.1\right).$$

Aqui, $Y_{i,j}$ está associada à j-ésima medida de peso da i-ésima mulher e $x_{i,j}$ é o tempo (em semanas após o início do estudo) em que ocorre esta medida.

a) Usando os diferentes valores de i e j para os quais existe uma observação $y_{i,j}$, faça um gráfico cartesiano com segmentos de reta conectando os pontos $(x_{i,j}, y_{i,j})$ e $(x_{i,j+1}, y_{i,j+1})$.
b) Construa uma tabela com médias a posteriori e limites dos intervalos de 95% de credibilidade a posteriori para os parâmetros α, β, σ^2, σ^2_α, σ^2_β.
c) Construa uma tabela com médias a posteriori e limites dos intervalos de 95% de credibilidade a posteriori para os parâmetros $\alpha_1, ..., \alpha_{68}, \beta_1, ..., \beta_{68}$.
d) Para cada observação $x_{i,j}$ sem observação de ganho de peso, encontre as médias a posteriori e os intervalos de 95% de credibilidade a posteriori para estes ganhos. Exiba estas informações em uma tabela.

Exemplo 13 (modelo linear generalizado): *Suponha que 8 grupos de besouros são expostos à diferentes níveis de concentração de disulfato de carbono gasoso. A tabela abaixo exibe o número de besouros em cada grupo e o número de besouros mortos após 5 horas de exposição.*

Numeração dos Grupos (i)	**Dose** (x_i)	**Número de besouros** (m_i)	**Número de mortos** (y_i)
1	1.6907	59	6
2	1.7242	60	13
3	1.7552	62	18
4	1.7842	56	28
5	1.8113	63	52
6	1.8369	59	53
7	1.8610	62	61
8	1.8839	60	60

a) Construa um gráfico de dispersão com as doses (x_i) no eixo das abscissas e as proporções de mortos (y_i/m_i) no eixo das ordenadas.

b) Assuma que y_i é uma amostra aleatória da v.a. Y_i, de modo que $Y_i \sim Bin(\pi_i, m_i)$, $\forall i \in \{1, ..., 8\}$ (modelo binomial). Supondo uma função de ligação logito, caracterizada pelas equações equivalentes

$$logito(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i$$

e

$$\pi_i = \frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)} \quad ,$$

encontre a média a posteriori e o intervalo de 95% de credibilidade a posteriori para os parâmetros β_1 e β_2.

Exemplo 14 (análise de sobrevivência): *Turnbull* et al. *(1974) descrevem algumas abordagens para analisar dados do programa de transplante de coração de Stanford. Um dos objetivos é decidir se transplantes de coração estendem o tempo de vida do paciente.*

Para cada paciente $i \in \{1, ..., 82\}$ *no programa, seja* $T_{1,i}$ *sua data de entrada no programa e* $T_{i,2}$ *a "última informação" que se tem do paciente, que pode ser: a data de sua morte ou a data de fechamento do estudo para análise dos dados (1 de março de 1973). Se o paciente realizou um transplante, seja* $T_{i,3}$ *a data da operação. Assim, para um paciente transplantado, temos* $T_{i,1} \leq T_{i,3} \leq T_{i,2}$. *Para os pacientes que não realizaram transplante, seja* $X_i = T_{i,2} - T_{i,1}$ *seu tempo de sobrevivência. Já para os pacientes que receberam o transplante, seja* $Y_i = T_{i,3} - T_{i,1}$ *seu tempo de espera até realizar o transplante e* $Z_i = T_{i,2} - T_{i,3}$ *seu tempo de sobrevivência após o transplante.*

Suponhamos ainda que $X_1, ..., X_n, Y_1, ..., Y_m, Z_1, ..., Z_m$ *são variáveis associadas a pacientes que faleceram e que* $X_{n+1}, ..., X_N, Y_{m+1}, ..., Y_M, Z_{m+1}, ..., Z_M$ *estão associadas a pacientes ainda vivos na data de fechamento do estudo.*

Em um dos modelos, Turnbull *et al.* (1974) assumem que os tempos de vida dos pacientes no grupo de não transplantados seguem uma distribuição exponencial[5] com média $1/\theta$. Por outro lado, para os pacientes transplantados, o tempo de sobrevida[6] segue uma distribuição exponencial com média $1/(\theta\tau)$. Assim, para um paciente transplantado, o modelo assume que riso do paciente muda de acordo com um fator τ. Especificamente, se $\tau = 1$, não há mudança no risco devido à realização do transplante.

O banco de dados deste exemplo e o artigo Turnbull *et al.* (1974) podem ser obtidos na página

http://www.ufjf.br/joaquim_neto

contém 4 variáveis: "transplant", que assume o valor 1 para transplantados e 0 caso contrário, "state", que assume o valor 1 para paciente vivo e zero caso contrário, "timetotransplant", que contém os dias de espera até o transplante e, finalmente, "survtime", que contém os dias de sobrevivência do paciente desde sua entrada no estudo.

[5] A distribuição exponencial é muito usada em análise de sobrevivência para modelar o tempo até a primeira ocorrência de um evento. Vale lembrar ainda que ao assumir um processo Poisson, o tempo até a primeira ocorrência segue distribuição exponencial.

[6] tempo de vida após transplante

a) Proponha uma distribuição a priori para o vetor de parâmetros desconhecidos.
b) Escreva a função de verossimilhanças.
c) Estime as médias a posteriori e os intervalos de 95% de credibilidade para os parâmetros do modelo.
d) Você diria que o transplante aumenta, diminui ou não afeta o risco do paciente? Por que?

Conjugação

Para os modelos estatísticos mais populares, existem famílias de distribuições com uma característica muito especial.

Definição 13: *Suponhamos que uma distribuição a priori pertence à uma determinada família de distribuições. Se, para um determinado modelo e parâmero, a posteriori pertencer a mesma família, dizemos que esta é uma* **família conjugada** *de distribuições a priori para o parâmetro.*

Resultado 1 (normal - normal): *Seja $Y_1, ..., Y_n$ uma amostra aleatória da $N(\mu, \sigma^2)$, com σ^2 conhecido. Supondo que $\mu \sim N(m, v^2)$ então $\mu \mid Y_1, ..., Y_n$ tem distribuição normal com*

$$E(\mu|Y_1, ..., Y_n) = \frac{\sigma^2 m + nv^2 \overline{Y}}{\sigma^2 + nv^2}$$

$$Var(\mu|Y_1, ..., Y_n) = \frac{\sigma^2 v^2}{\sigma^2 + nv^2}, \text{ onde } \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

Prova:

Exercício 2: *Seja $Y_1, ..., Y_n$ uma amostra aleatória da $N(\mu, \sigma^2)$, com σ^2 conhecido. Supondo uma priori $\mu \sim N(m, v^2)$:*

a) *verifique o limite de $E(\mu \mid y_1, ..., y_n)$ quando v^2 tende à ∞,*

b) *encontre o EMV para esta amostra aleatória e*

c) *compare os resultados dos itens (a) e (b).*

Resultado 2 (binomial - beta): *Seja $Y_1, ..., Y_n$ uma amostra aleatória da $Bin(m, \theta)$, com m conhecido. Supondo uma priori $\theta \sim Be(a, b)$ temos que*

$$(\theta \mid Y_1, ..., Y_n) \sim Be\left(a + \sum_{i=1}^{n} Y_i, nm + b - \sum_{i=1}^{n} Y_i\right).$$

Prova:

OBS: Lembre-se que a distribuição uniforme é um caso particular da beta (basta fazer $a = 1$ e $b = 1$). Assim, pelo resultado acima, um modelo binomial combinado com uma priori uniforme produz uma posteriori beta.

Resultado 3 (Poisson - gamma): *Seja $Y_1, ..., Y_n$ uma amostra aleatória da $Poi(\theta)$. Supondo uma priori $\theta \sim Ga(a, b)$ temos que*

$$(\theta \mid Y_1, ..., Y_n) \sim Ga\left(a + \sum_{i=1}^{n} Y_i, b + n\right).$$

Prova:

Resultado 4 (normal - gamma): *Seja* $Y_1, ..., Y_n$ *uma amostra aleatória da* $N(\mu, \phi^{-1})$, *com* μ *conhecido. Supondo uma priori* $\phi \sim Ga(a, b)$ *então*

$$(\phi \mid Y_1, ..., Y_n) \sim Ga\left(\frac{n}{2} + a, b + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2}\right).$$

Algoritmos de simulação

Aqui, veremos apenas dois algoritmos para simular amostras de uma distribuição de probabilidade: o amostrador de Gibbs e o Metropolis-Hastings. Para uma revisão mais detalhada destes algoritmos e para conhecer outros, veja Gamerman and Lopes (2006).

Condicionais completas

Definição 14: *As distribuições de*

$$
\begin{aligned}
&(Y_1|Y_2, Y_3, ..., Y_n)\,,\\
&(Y_2|Y_1, Y_3, Y_4, ..., Y_n)\,,\\
&(Y_3|Y_1, Y_2, Y_4, Y_5, ..., Y_n)\,,\\
&\vdots\\
&(Y_n|Y_1, Y_2, ..., Y_{n-1})
\end{aligned}
$$

são chamadas de **condicionais completas** *da distribuição conjunta de* $(Y_1, ..., Y_n)$.

Pela teoria de probabilidades, a densidade (ou f.p.) condicional $p(y_i \mid y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$ pode ser obtida com

$$
p(y_i \mid y_1, ..., y_{i-1}, y_{i+1}, ..., y_n) = \frac{p\,(y_1, ..., y_n)}{\int p\,(y_1, ..., y_n)\,dy_i}.
$$

Amostrador de Gibbs

O **amostrador de Gibbs** (AG) é um método MCMC que têm núcleo de transição formado pelas distribuições condicionais completas.

O amostrador de Gibbs é um dos algoritmos de simulação mais utilizados na inferência bayesiana. Seu objetivo é simular de uma distribuição conjunta (que no contexto bayesiano é a posteriori) e, para isto, valores são simulados sucessivamente das distribuições condicionais completas.

O problema a ser resolvido então envolve a simulação de uma distribuição conjunta quando a simulação das condicionais completas é acessível (ou trivial) e esquemas para simular diretamente da conjunta [7] são muito complicados ou simplesmente não disponíveis.

[7] Ver Gamerman and Lopes (2006)

O amostrador de Gibbs é descrito pelos seguintes passos:

1) Inicialize um contador de iterações $j = 1$ e defina ("chute") um vetor de valores iniciais $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)}, ..., y_n^{(0)})$.

2) Obtenha um novo vetor $\mathbf{y}^{(1)} = (y_1^{(1)}, y_2^{(1)}, ..., y_n^{(1)})$ simulando

$$
\begin{aligned}
y_1^{(j)} \text{ de } p\left(y_1 | y_2^{(j-1)}, y_3^{(j-1)}, ..., y_n^{(j-1)}\right) \\
y_2^{(j)} \text{ de } p\left(y_2 | y_1^{(j)}, y_3^{(j-1)}, y_4^{(j-1)}, ..., y_n^{(j-1)}\right) \\
y_3^{(j)} \text{ de } p\left(y_3 | y_1^{(j)}, y_2^{(j)}, y_4^{(j-1)}, y_5^{(j-1)}, ..., y_n^{(j-1)}\right) \\
\vdots \\
y_n^{(j)} \text{ de } p\left(y_n | y_1^{(j)}, y_2^{(j)}, ..., y_{n-1}^{(j)}\right).
\end{aligned}
$$

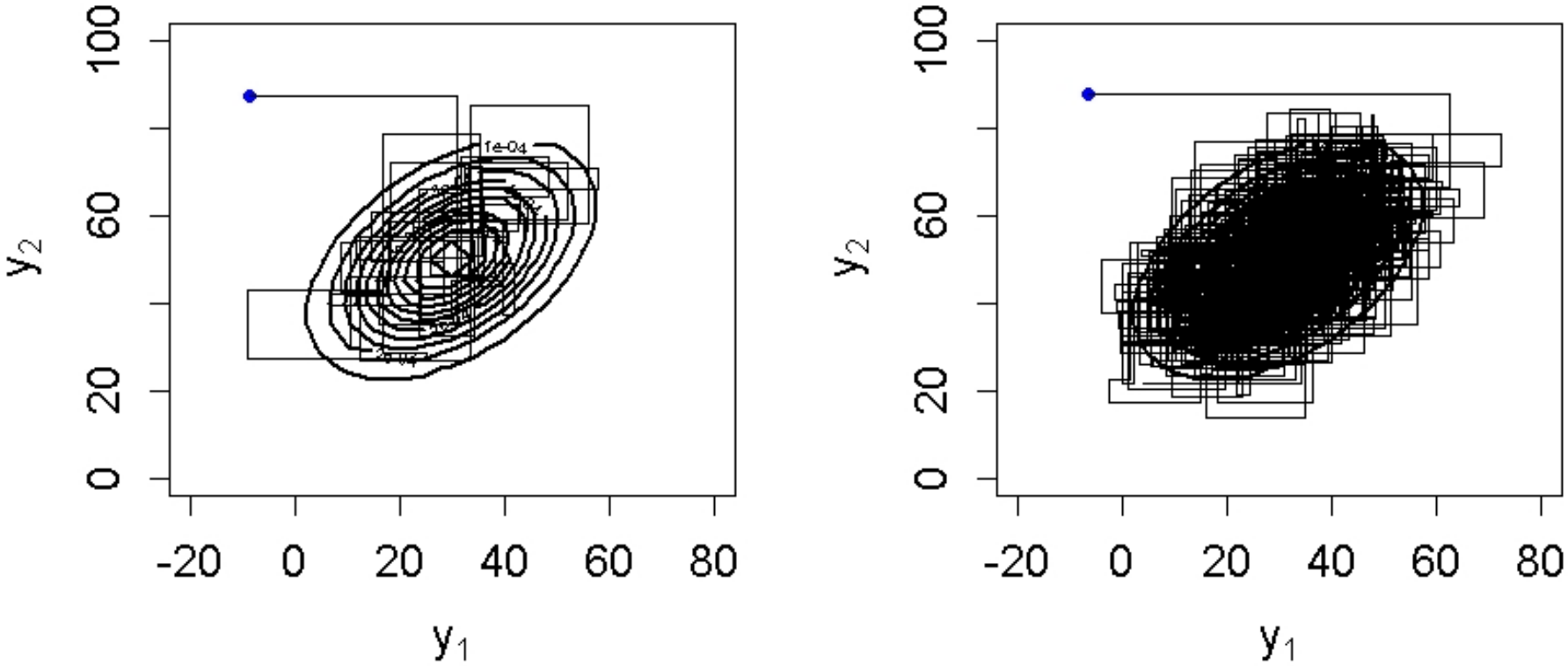
3) Se o contador for menor que k, mude o contador de j para $j + 1$ e retorne ao passo 2.

O algoritmo produz sequências $y_i^{(1)}, y_i^{(2)}, ..., y_i^{(k)}$, para $i \in \{1, ..., n\}$. Na prática[8], o procedimento utilizado para obter amostras da distribuição conjunta a partir destas sequências consiste em descartar um número b de valores iniciais e, em seguida, escolher valores com um espaçamento igual a t. As quantidades b e t são chamadas de aquecimento (*burnin*) e espaçamento (*thinning*), respectivamente.

[8] Para mais informações, consulte Gamerman and Lopes (2006).

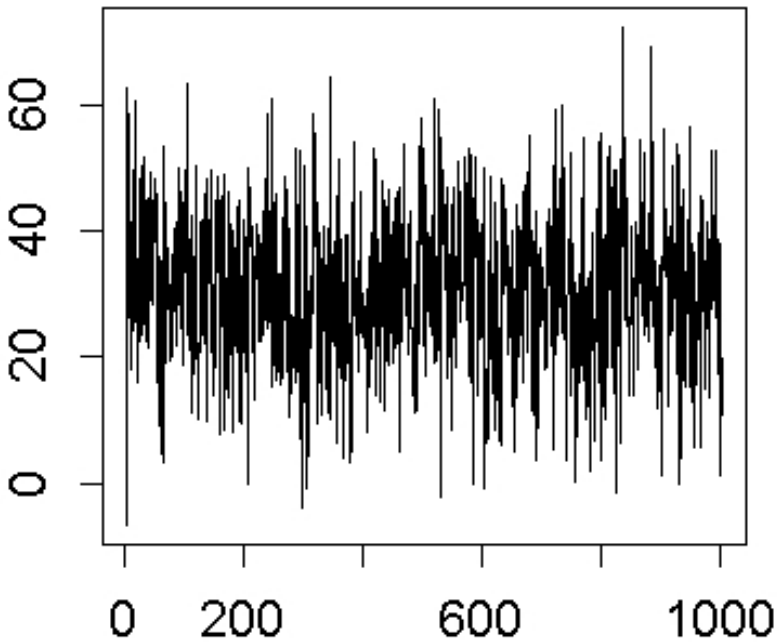
Exemplo 15 (Didático): *Use o amostrador de Gibbs para simular amostras de uma distribuição normal bivariada*[9] *com vetor de médias* $\mu = (30, 50)^T$ *e matriz de covariâncias* $\Sigma = \begin{bmatrix} 150 & 70 \\ 70 & 150 \end{bmatrix}$.

[9] Neste exemplo, a escolha da normal foi didática, uma vez que a maioria dos softwares estatísticos possuem comandos para simular desta distribuição.

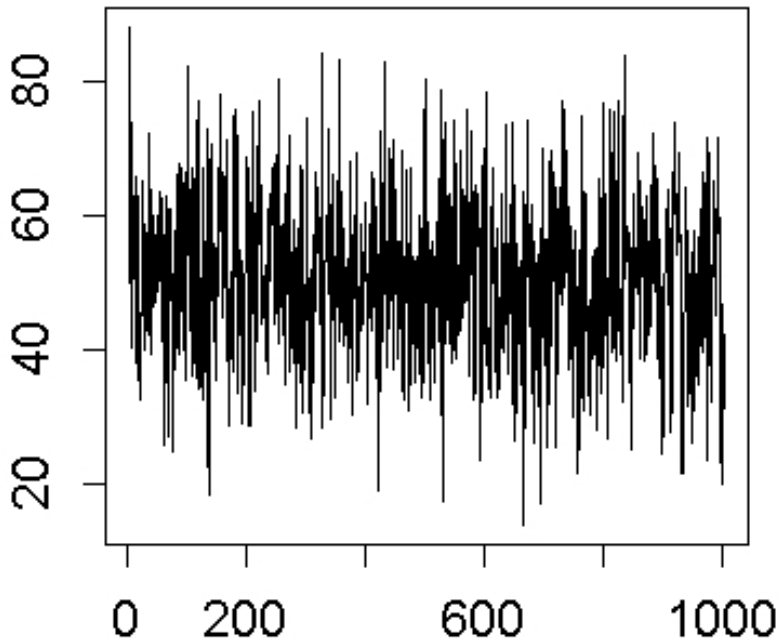


50 iterações. 1000 iterações.

Figura: Cadeias obtidas com o amostrador de Gibbs para o exemplo 15. O ponto azul marca o valor inicial da cadeia.



y_1.



y_2.

Figura: Cadeias obtidas com o amostrador de Gibbs para o exemplo 15.

Metropolis-Hastings

O algoritmo Metropolis-Hastings é descrito pelos seguintes passos:

1) Inicialize um contador de iterações $j = 1$ e defina ("chute") um vetor de valores iniciais $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)}, ..., y_n^{(0)})$.
2) Obtenha um novo vetor $\mathbf{y}^* = (y_1^*, y_2^*, ..., y_n^*)$ simulando de uma densidade (ou f.p.) proposta $q(\mathbf{y}^* \mid \mathbf{y}^{(j-1)})$, que pode ou não depender de $\mathbf{y}^{(j-1)}$.
2) Simule uma amostra u da distribuição $U[0, 1]$.
3) Calcule a razão

$$r(\mathbf{y}^{(j-1)}, \mathbf{y}^*) = \frac{p\left(\mathbf{y}^* \mid \mathbf{y}^{(j-1)}\right)}{p\left(\mathbf{y}^{(j-1)} \mid \mathbf{y}^*\right)} \frac{q\left(\mathbf{y}^{(j-1)} \mid \mathbf{y}^*\right)}{p\left(\mathbf{y}^* \mid \mathbf{y}^{(j-1)}\right)}$$

.

4) Se $u < r(\mathbf{y}^{(j-1)}, \mathbf{y}^*)$ o valor proposto é aceito fazendo $\mathbf{y}^{(j)} = \mathbf{y}^*$, caso contrário, o valor proposto é rejeitado fazendo $\mathbf{y}^{(j)} = \mathbf{y}^{(j-1)}$.
5) Se o contador for menor que k, mude o contador de j para $j + 1$ e retorne ao passo 2.

Assim como o amostrador de Gibbs, o algoritmo produz sequências $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(k)}$ e para obter amostras da distribuição basta retirar observações iniciais (*burnin*) e usar um espaçamento (*thinning*).

Um algoritmo de Metropolis muito utilizado é baseado em um passeio aleatório. Neste caso, se usarmos uma distribuição proposta com variância σ^2, duas situações extremas podem ocorrer:

- se σ^2 for muito pequena os valores gerados estarão próximos do valor atual e serão aceitos com frequência. Neste caso, serão necessárias muitas iterações para percorrer o espaço paramétrico.
- valores grandes de σ^2 levam a uma taxa de rejeição excessivamente alta e a cadeia se movimentará lentamente.

Exemplo 16 (Didático): *Use o algoritmo Metropolis-Hastings para simular valores da distribuição $N(5, 25)$. Use como proposta um passeio aleatório normal com variância τ^2. Compare as taxas de aceitação obtidas com diferentes valores de τ^2.*

Exercício 3: *Considere a densidade*

$$p\left(y_1, y_2|\mu_1, \mu_2, \Sigma_1, \Sigma_2\right) = 0.7p\left(y_1, y_2|\mu_1, \Sigma_1\right) + 0.3p\left(y_1, y_2|\mu_2, \Sigma_2\right)$$

de uma mistura de normais, onde $p\left(y_1, y_2 \mid \mu_i, \Sigma_i\right)$ é a densidade de uma normal bivariada com vetor de médias μ_i e matriz de covariâncias Σ_i, para $i = 1, 2$. Especificamente, fixe

$$\mu_1 = (4, 5)^T, \mu_2 = (0.7, 3.5)^T, \Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \text{ e } \Sigma_2 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 1 \end{bmatrix}.$$

a) *Plote a densidade $p\left(y_1, y_2|\mu_1, \mu_2, \Sigma_1, \Sigma_2\right)$ (superfície tridimensional).*

b) *Use o algoritmo Metropolis-Hastings para simular amostras de $p\left(y_1, y_2|\mu_1, \mu_2, \Sigma_1, \Sigma_2\right)$.*

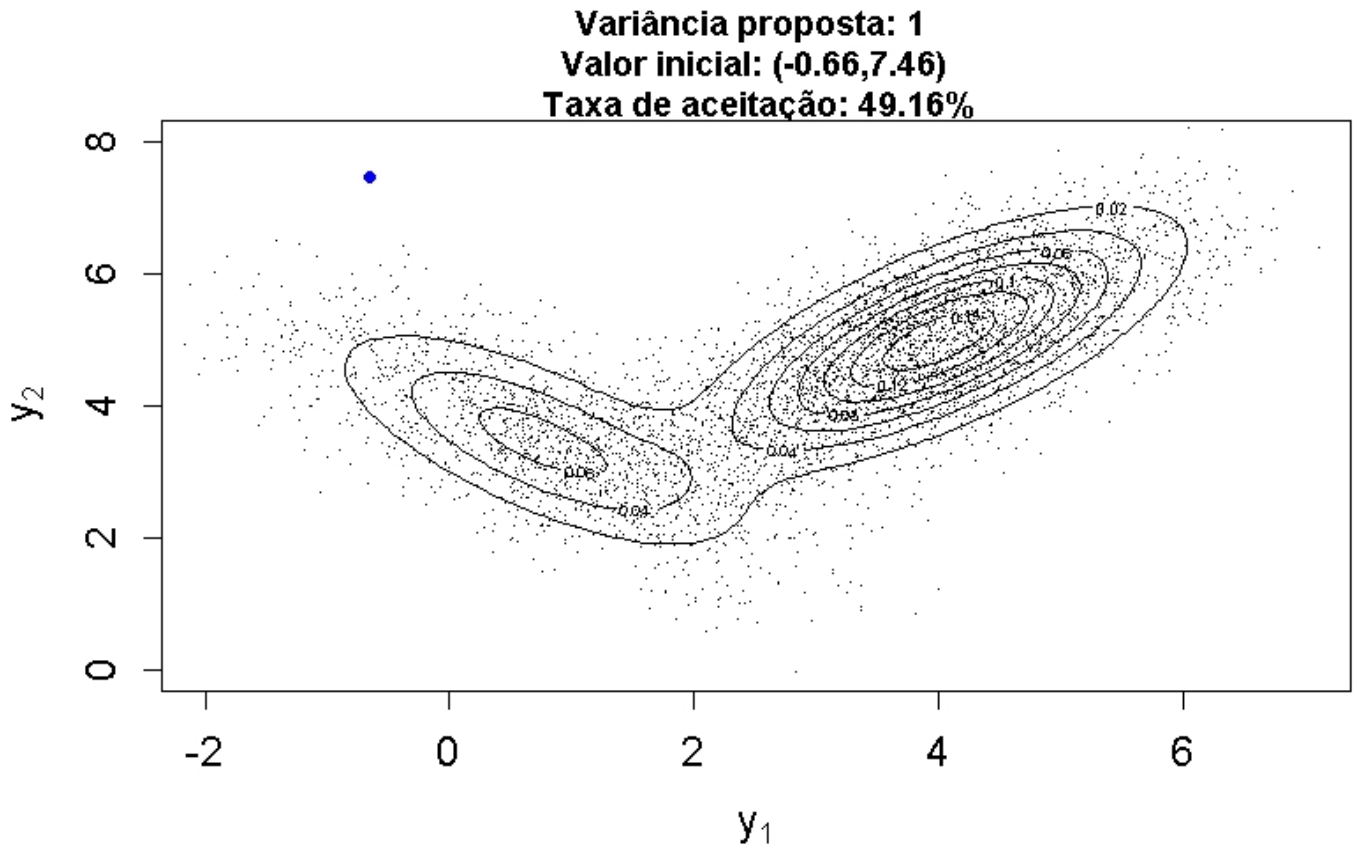


Figura: Cadeia obtida com o algoritmo Metropolis-Hastings. O ponto azul marca o valor inicial da cadeia.

Introdução a teoria da decisão

Relembremos um típico problema de inferência paramétrica, no qual uma variável aleatória **X** (possivelmente multidimensional) é usada para representar matematicamente a incerteza de observações amostrais. Suponhamos um vetor $\boldsymbol{\theta}$, chamado de **vetor paramétrico**, com componentes desconhecidas. Suponhamos ainda que $\boldsymbol{\theta}$ pertence a um conjunto Θ de valores possíveis, chamado **espaço paramétrico**.

Um problema de decisão pode ser completamente especificado usando 3 conjuntos:

- o espaço paramétrico Θ,
- o conjunto de possíveis resultados do experimento Ω e
- um conjunto de possíveis ações $\mathcal{A}$.

Uma **regra de decisão** (ou simplesmente **decisão**) δ é uma função definida em Ω e que assume valores em $\mathcal{A}$, ou seja $\delta : \Omega \rightarrow \mathcal{A}$. Uma **função perda** $L(\delta, \theta)$ associa a cada decisão δ e valor possível $\theta \in \Theta$ um número real positivo. Este número pode ser interpretado como uma punição[10] por tomar a decisão δ sendo que o valor do parâmetro é θ.

[10] Esta punição é denominada também pelos termos: custo, perda e prejuízo.

Definição 15: *O* **risco** *de uma regra de decisão δ, denotado por $R(\delta)$, é definido como o valor esperado da função perda, ou seja,*

$$R(\delta) = E(L(\delta, \boldsymbol{\theta})).$$

Se nenhuma amostra de X foi observada, o valor esperado da equação acima é baseado na priori, ou seja,

$$R(\delta) = E(L(\delta, \boldsymbol{\theta})) = \int L(\delta, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Por outro lado, se uma amostra $\mathbf{x}$ *de* $\mathbf{X}$ *já tiver sido observada, o valor esperado é com relação a distribuição a posteriori e*

$$R(\delta) = E(L(\delta, \boldsymbol{\theta})) = \int L(\delta, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{x})\, d\boldsymbol{\theta}.$$

Uma regra de decisão δ^* é dita **ótima** se tem risco mínimo, ou seja, se para qualquer outra regra de decisão δ e observação $\mathbf{x}$, $R(\delta^*) < R(\delta)$. O risco é uma medida que permite comparar diferentes regras de decisão e, naturalmente, a regra de decisão com o menor risco tem menor perda esperada.

Exemplo 17: *Consideremos novamente a situação exposta no exemplo 6. Porém, agora o médico precisa escolher entre duas decisões: submeter o paciente a uma determinada cirurgia, denotada por d_1, ou não, denotada por d_2. A tabela abaixo apresenta uma possível atribuição de perdas para cada decisão e estado do paciente. Qual é a regra de decisão ótima?*

		Decisão	
		Não faz a cirurgia (d_2)	**Faz a cirurgia (d_1)**
Estado θ	**Sem a doença (0)**	*0*	*500*
	Com a doença (1)	*1000*	*100*

Solução:

Definição 16: *Um estimador é uma regra de decisão ótima com respeito a uma determinada função perda. Seu valor observado é chamado de estimativa.*

Proposição 3: *Seja* $L(\delta, \boldsymbol{\theta}) = L(\delta(\omega), \boldsymbol{\theta})$ *a função perda associada a estimação de um parâmetro* θ*. Esta função é geralmente chamada de perda quadrática. O estimador de* θ *é* $E(\boldsymbol{\theta} \mid \mathbf{x})$*, se uma amostra* x *foi observada, e* $E(\boldsymbol{\theta}$*, caso contrário.*

FIM!