

---

**CE-227**

**Inferência Estatística**

**Introdução à Inferência Bayesiana**

---

*Stuart Coles*

*Paulo Justiniano Ribeiro Jr*

# **Inferência Estatística (Inferência Bayesiana)**

Stuart Coles

Paulo Justiniano Ribeiro Jr

Laboratório de Estatística e Geoinformação (LEG)

<http://www.leg.ufpr.br>

Departamento de Estatística

Universidade Federal do Paraná (UFPR)

Complementos online: <http://www.leg.ufpr.br/ce227>

Contato: paulojus @ ufpr.br

CE-227 Inferência Bayesiana

Disciplina do Bacharelado em Estatística

Universidade Federal do Paraná

1ª oferta: 1º semestre de 2014

Última atualização: 30 de setembro de 2022

# Prefácio

Este material foi preparado por Stuart Coles e está sendo traduzido, adaptado e expandido para utilização no curso. O material poderá ser atualizado durante o curso – verifique sempre a data da última atualização mostrada abaixo. Em um primeiro momento foi feita uma tradução do texto. Na sequência estão sendo acrescentadas figuras, códigos, texto e exercícios.

P.J.R.Jr

Curitiba, 30 de setembro de 2022

# Sumário

<b>Prefácio</b>	<b>iii</b>
<b>1 Introdução</b>	<b>1</b>
1.1 O que é inferência estatística?	1
1.2 O que é inferência bayesiana?	2
1.3 Distribuição a priori	3
1.4 Características da abordagem bayesiana	4
1.5 Objeções a inferência bayesiana	5
1.6 Revisão do Teorema de Bayes	6
1.7 Exercícios	9
<b>2 Atualização Bayesiana</b>	<b>11</b>
2.1 Introdução	11
2.2 Teorema de Bayes	11
2.3 Tópicos	12
2.3.1 Escolha do modelo de verossimilhança	12
2.3.2 Escolha da priori	12
2.3.3 Computação	13
2.3.4 Inferência	13
2.3.5 Tópicos avançados	13
2.4 Exemplos	13
2.5 Tópicos gerais	20
2.5.1 Atualização sequencial	20
2.5.2 Suficiência	21
2.5.3 O princípio da verossimilhança	22
2.5.4 Distribuição marginal	23
2.6 Códigos computacionais	25
2.7 Exercícios	27
<b>3 Especificação de prioris</b>	<b>29</b>
3.1 Introdução	29
3.2 Prioris conjugadas	29
3.2.1 Uso de prioris conjugadas	30

3.2.2	Obtenção de prioris conjugadas . . . . .	30
3.2.3	Análises conjugadas usuais. . . . .	32
3.3	Prioris impróprias . . . . .	32
3.4	Representações de ignorância . . . . .	33
3.4.1	Exemplos . . . . .	34
3.5	Mistura de prioris . . . . .	35
3.6	Elicitação de prioris . . . . .	39
3.7	Códigos computacionais . . . . .	40
3.8	Exercícios . . . . .	40
<b>4</b>	<b>Problemas com múltiplos parâmetros</b>	<b>43</b>
4.1	Exemplos . . . . .	45
4.2	Exercícios . . . . .	48
<b>5</b>	<b>Resumindo informação a posteriori</b>	<b>51</b>
5.1	Teoria da decisão . . . . .	51
5.2	Estimação pontual . . . . .	54
5.2.1	Perda quadrática . . . . .	55
5.2.2	Perda absoluta . . . . .	55
5.2.3	Perda linear . . . . .	56
5.2.4	Perda 0-1 . . . . .	57
5.2.5	Resumo . . . . .	57
5.3	Intervalos de credibilidade . . . . .	57
5.4	Teste de hipóteses . . . . .	60
5.4.1	Testando hipóteses via regra de decisão . . . . .	60
5.4.2	Teste de Hipóteses via Fator de Bayes . . . . .	61
5.5	Códigos computacionais . . . . .	62
5.6	Exercícios . . . . .	64
<b>6</b>	<b>Predição</b>	<b>67</b>
6.1	Distribuição preditiva . . . . .	67
6.2	Exemplos . . . . .	69
6.3	Exercícios . . . . .	70
<b>7</b>	<b>Propriedades Assintóticas</b>	<b>73</b>
7.1	Introdução . . . . .	73
7.2	Consistência . . . . .	73
7.3	Normalidade assintótica . . . . .	74
7.4	Exemplos . . . . .	75
7.5	Exercícios . . . . .	77
<b>8</b>	<b>Outros tópicos</b>	<b>79</b>
8.1	Bayes Empírico . . . . .	79
8.2	Estimação Linear Bayesiana . . . . .	80

8.3	Robustez . . . . .	80
8.4	Computação . . . . .	80
8.4.1	Amostrador de Gibbs . . . . .	81
8.5	Exemplos . . . . .	82
8.6	Exercícios . . . . .	85

<b>Referências Bibliográficas</b>	<b>86</b>
-----------------------------------	-----------

# Capítulo 1

## Introdução

### 1.1 O que é inferência estatística?

Antes de definir *inferência bayesiana* devemos considerar uma questão mais abrangente, “o que é inferência estatística?”. Muitas definições são possíveis mas a maioria vai de encontro ao princípio de que inferência estatística é a ciência de tirar conclusões sobre uma “população” a partir de uma “amostra” de itens retirados desta população. Isto nos remete a diversas questões sobre o que se quer dizer com “uma população”, qual a relação entre a amostra e a população, como conduzir a amostragem se todas as opções são disponíveis e assim por diante. Mas vamos deixar estes tópicos de lado, e focar a discussão em um exemplo simples.

Suponha que uma Unidade Florestal deseja estimar a proporção de árvores acometidas por uma determinada doença em uma grande floresta. Não é prático ou mesmo factível examinar cada árvore, e então opta-se por selecionar uma amostra de apenas  $n$  árvores. Reiteramos que não vamos discutir aqui como a amostra pode ser escolhida e tomada, mas supomos que a amostra é aleatória, no sentido que se  $\theta$  é a proporção de árvores na floresta que tomadas pela doença, então cada árvore na amostra poderá ter a doença, independentemente de todas demais árvores na floresta, com probabilidade  $\theta$ . Denotando por  $Y$  a variável aleatória que corresponde ao número de árvores doentes na amostra, a Unidade Florestal vai usar o valor observado  $Y = y$ , para inferir sobre o parâmetro populacional  $\theta$ . Esta inferência pode ser na forma de uma *estimativa pontual* ( $\hat{\theta} = 0,10$ ); um *intervalo de confiança* (95% de confiança que o intervalo  $[0,08; 0,12]$  contém o valor de  $\theta$ ); um *teste de hipótese* (rejeita a hipótese de que  $\theta < 0,07$  ao nível de significância de 5%); uma *predição* (prevê-se que 15% das árvores estarão afetadas no ano seguinte); ou *decisão* (decide-se identificar e remover todas árvores infectadas).

Em cada caso, o conhecimento do valor amostral observado  $Y = y$  está sendo usado para fazer inferências sobre a característica  $\theta$  da população. Além disto, essas inferências são feitas especificando-se um modelo de probabilidades,  $f(y|\theta)$ , que

determina como, para um dado valor de  $\theta$ , as probabilidades de diferentes valores de  $X$  são distribuídas.

No exemplo considerado aqui, sob as suposições feitas sobre a amostragem aleatória, nosso modelo seria

$$Y|\theta \sim \text{Bin}(n, \theta).$$

A *inferência estatística* resulta então na uma inferência sobre o parâmetro populacional  $\theta$  baseada na observação  $Y = y$  e, basicamente, infere-se que valores de  $\theta$  que conferem uma alta probabilidade para o valor observado  $y$  são “mais prováveis” do que os valores que conferem uma baixa probabilidade – o princípio da máxima verossimilhança. Note que em um contexto mais amplo, inferência estatística também engloba as questões de escolha de modelos, verificação de modelos, etc, mas restringe-se aqui a atenção na inferência sobre os parâmetros de uma família de modelos paramétricos.

Antes de abordar inferência bayesiana em particular há alguns pontos a serem considerados sobre a abordagem clássica (frequentista) de inferência. O ponto mais fundamental é o de que o parâmetro  $\theta$ , ainda que desconhecido, é tratado como uma *constante* ao invés de *aleatório*. Esta é a pedra fundamental da teoria clássica, mas que leva a problemas de interpretação. Seria desejável que um intervalo de confiança de 95%  $[0,08;0,12]$  significasse que há uma probabilidade de 0,95 que  $\theta$  estivesse entre 0,08 e 0,12. Entretanto tal interpretação *não é possível* uma vez que  $\theta$  não é aleatório. Sendo considerado constante, o valor de  $\theta$  *está* ou *não está* no intervalo, não há idéia da probabilidade associada a isto. O único elemento aleatório neste modelo de probabilidades é o dado. Portanto a interpretação correta do intervalo é a de que, se o procedimento for aplicado “muitas vezes”, então, “em uma longa sequência”, os intervalos que serão construídos vão conter o verdadeiro valor de  $\theta$  em 95% das vezes. Todas inferências baseadas na teoria clássica são forçadas a terem este tipo de interpretação de “frequência em longas sequências” (portanto *frequentista*), muito embora, como no exemplo, tenhamos apenas o intervalo  $[0,08;0,12]$  para interpretar.

## 1.2 O que é inferência bayesiana?

O contexto geral no qual a inferência bayesiana funciona é idêntico ao anterior: há um parâmetro populacional  $\theta$  a respeito do qual deseja-se fazer inferências, e um mecanismo probabilístico  $f(y|\theta)$  que determina a probabilidade de observar cada valor diferente de  $y$ , sob diferentes valores de  $\theta$ . A diferença essencial entretanto é a de que  $\theta$  é tratado como um quantidade *aleatória*. Isto pode parecer bastante inócua, mas de fato leva a abordagens substancialmente diferentes à modelagem e inferência.

Essencialmente, a inferência vai ser baseada em  $f(\theta|y)$  ao invés de  $f(y|\theta)$ ; isto é, a probabilidade do parâmetro condicional aos dados obtidos, ao invés da dos dados, condicional o valor do parâmetro. Em muitas formas tal abordagem leva a inferências muito mais naturais, mas para atingir isto será necessário especificar,



adicionando ao modelo, uma *distribuição a priori* de probabilidades,  $f(\theta)$ , que representa crenças sobre a distribuição de  $\theta$  antes de se considerar qualquer informação proveniente dos dados.

Esta noção de distribuição *a priori* para o parâmetro  $\theta$  está no cerne do pensamento bayesiano. Dependendo se discute-se com um defensor ou oponente da metodologia, tal fato é ou sua principal vantagem sobre a teoria clássica, ou sua maior armadilha<sup>1</sup>.

## 1.3 Distribuição a priori

Em quase todas situações, quando se tenta estimar um parâmetro  $\theta$ , tem-se algum conhecimento ou crença sobre o valor de  $\theta$  antes de considerar os dados. Um exemplo de O'Hagan (1994) deixa claro isto em termos quantitativos.

Voce olha por sua janela e vê algo grande de madeira com galhos cobertos por pequenas coisas verdes. Voce postula duas possíveis hipóteses: uma de que é uma árvore, outra de que é o carteiro. É claro que voce rejeita a hipótese de que é o carteiro porque carteiros em geral não se parecem com isto, enquanto que árvores sim. Desta forma, em linguagem formal, denotando  $A$  o evento de que voce vê uma coisa de madeira coberta de coisinhas verdes,  $B_1$  o evento que é uma árvore e,  $B_2$  o evento de que é um carteiro, voce rejeita  $B_2$  a favor de  $B_1$  porque  $f(A|B_1) > f(A|B_2)$ . Aqui voce está usando o princípio de maximizar a verossimilhança.

Mas, voce pode considerar ainda uma terceira possibilidade,  $B_3$ , de que a “coisa” é uma réplica de uma árvore. Neste caso pode bem ser que  $f(A|B_1) = f(A|B_3)$  e ainda assim voce rejeita esta hipótese a favor de  $B_1$ . Isto é, mesmo que a probabilidade de ver o que voce observou seja a mesma para caso seja uma árvore ou uma réplica, sua crença *a priori* é a de que é mais provável que seja uma árvore do que uma réplica e então voce incluiu esta informação para tomar sua decisão.

Considere um outro exemplo, onde em cada caso o modelo para os seus dados é  $[Y|\theta] \sim \text{Bin}(10, \theta)$  e observamos  $y = 10$  tal que (verificar!!) a hipótese  $H_0 : \theta \leq 0,5$  é rejeitada a favor de  $H_1 : \theta > 0,5$  em uma das seguintes situações.

1. Uma mulher que toma chá afirma que pode detectar em uma xícara de chá com leite se o leite foi colocado antes ou depois do chá. Ela o faz corretamente em 10 xícaras que experimenta.
2. Uma especialista em música afirma que pode distinguir entre uma partitura de Hayden e uma de Mozart. Ela reconhece corretamente 10 peças mostradas a ela.
3. Uma amiga que está embriagada afirma que pode prever o resultado do lançamento de uma moeda honesta e acerta todas as 10 tentativas efetuadas.

---

<sup>1</sup> Estes pontos ficarão mais claros através de exemplos específicos, mas é importante pontuar desde já, para estabelecer os vários argumentos a favor e contra a abordagem bayesiana. Deve-se ler e refletir sobre tais discussões e controvérsias novamente, à medida que avança-se mais na teoria.

Agora, considerando apenas os dados (10 resultados corretos em todos os casos) somos forçados a extrair as mesmas inferências em todos os casos. Mas nossa convicção anterior sugere que é provável que sejamos mais céticos em relação a afirmação da amiga embriagada, um pouco impressionado pela mulher que toma chá, e de forma alguma surpresos sobre a especialista em música.

O ponto essencial é o seguinte: experimentos não são dispositivos abstratos. Invariavelmente, temos algum conhecimento sobre o processo sendo investigado antes de obter os dados. É sensato (muitos diriam essencial) que inferências deveriam ser baseadas na informação combinada que o conhecimento *a priori* e os dados representam.

Apenas para colocar um ponto de vista alternativo, é à esta mesma dependência em crenças prévias que os oponentes da visão bayesiana têm objeção. Crenças prévias diferentes vão levar a diferentes inferências na visão bayesiana, e é exatamente se isto é visto como algo bom ou ruim que determina a aceitação ou não do ponto de vista bayesiano.

## 1.4 Características da abordagem bayesiana

De acordo com O'Hagan (1994) pode-se identificar quatro aspectos fundamentais que caracterizam a abordagem bayesiana para inferência estatística.

- **Informação a priori (anterior)** Todos problemas são únicos e possuem contexto próprio. Este contexto provém de informação anterior, e é a formulação e exploração deste conhecimento anterior que distingue a inferência bayesiana da estatística clássica.
- **Probabilidade Subjetiva** A estatística clássica depende de uma definição objetiva de “frequência em longas sequências” de probabilidades. Mesmo sendo desejável, o que é questionável, isto leva a inferências embaraçosas. Em contraste, a estatística bayesiana formaliza explicitamente a noção de que todas as probabilidades são subjetivas, dependendo de convicções e conhecimentos individuais disponíveis. Deste modo, a análise bayesiana é individualizada, ou seja, é única para as especificações das convicções prévias de cada indivíduo. A inferência é baseada na distribuição *a posteriori*  $f(\theta|y)$ , cuja forma será vista dependente (através do Teorema de Bayes) das particularidades da especificação da priori  $f(\theta)$ .
- **Auto consistência** Tratando o parâmetro  $\theta$  como aleatório, emerge que todo desenvolvimento da inferência bayesiana decorre naturalmente apenas da teoria de probabilidades. Isto tem muitas vantagens e significa que todas as questões inferenciais podem ser expressas por declarações probabilísticas sobre  $\theta$ , que por sua vez são derivadas diretamente da distribuição *a posteriori*  $f(\theta|y)$ .

- **Sem procedimentos “ad-hoc”** Como em inferência clássica não se pode fazer declarações probabilísticas a respeito de  $\theta$ , diversos critérios são desenvolvidos para julgar se um particular estimador é “bom” em algum sentido. Isto gerou uma proliferação de procedimentos, em geral conflitantes uns com os outros. A inferência bayesiana esquivava-se desta tendência de inventar critérios *ad hoc* para julgamento e comparação de estimadores ao contar com a distribuição a posteriori para expressar, em termos probabilísticos, e portanto sem ambiguidade, a inferência completa sobre o  $\theta$  desconhecido.

## 1.5 Objeções a inferência bayesiana

As principais objeções a inferência bayesiana, como descrito acima, são que as conclusões vão depender de escolhas específicas de priori. Como enfatizado anteriormente, argumenta-se por outro lado que esta é a beleza da abordagem bayesiana. Este é (infelizmente) um debate que não tem fim. Mas antes de deixar este tópico de lado, deve-se destacar que mesmo em inferência clássica, e de fato em investigações científicas de modo geral, é implícito que convicções a priori são utilizadas. Algum conhecimento anterior é usado para formular um modelo de verossimilhança adequado. Em testes de hipótese, crenças a priori sobre a plausibilidade de uma hipótese são geralmente ajustadas implicitamente (ou em geral de forma “secreta” - ou implícita) alterando-se o nível de significância de um teste. Acreditando-se que os dados devem conduzir a rejeição de uma hipótese, pode-se assegurar a rejeição escolhendo-se um nível de significância suficientemente baixo para o teste! Então, de certa forma, pode-se argumentar que a incorporação de informação anterior é formalizada explicitamente na inferência bayesiana, o que é, em geral, feito “por trás dos panos” na análise clássica.

Há ainda uma outra forma de pensar sobre inferência bayesiana que parece evitar qualquer dificuldade conceitual ou filosófica com conhecimento a priori. Em inferência clássica, estimadores de máxima verossimilhança são obtidos escolhendo o ponto no espaço paramétrico que maximiza a superfície de verossimilhança. Uma forma de pensar sobre inferência bayesiana é que ela equivale a fazer uma média (ponderada ou integrada) sobre a superfície de verossimilhança, ao invés de maximizar. Isto parece bastante sensato e sem controvérsias. A controvérsia surge devido ao fato de que a média é ponderada de acordo com a distribuição a priori. Mas, mesmo em inferência clássica, é bastante comum atribuir pesos diferentes a diferentes pedaços de informação como, por exemplo, em regressão ponderada. Portanto, novamente parece que os fundamentos da inferência bayesiana aos quais alguns se opõem, são simplesmente procedimentos que, implicitamente, são de fato efetuados em estatística clássica.

## 1.6 Revisão do Teorema de Bayes

Em sua forma básica, o Teorema de Bayes é simplesmente um resultado de probabilidade condicional:

**Teorema 1.1** (Teorema de Bayes em sua forma básica). *Se  $A$  e  $B$  são dois eventos com  $P[A] > 0$  então:*

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]}$$

A prova é trivial. O uso do Teorema de Bayes, em aplicações de probabilidades, é o de reverter o condicionamento dos eventos. Isto é, o teorema mostra como a probabilidade de  $[B|A]$  está relacionada com a de  $[A|B]$ <sup>2</sup>.

Uma pequena extensão do teorema é obtida considerando eventos  $C_1, C_2, \dots, C_k$  que formam uma partição do espaço amostral  $\Omega$ , tal que  $C_i \cap C_j = \emptyset$  se  $i \neq j$  e  $C_1 \cup C_2 \cup \dots \cup C_k = \Omega$ . Então,

$$P[C_i|A] = \frac{P[A|C_i]P[C_i]}{\sum_{j=1}^k P[A|C_j]P[C_j]} \quad \text{para } i = 1, \dots, k$$

**Exemplo 1.1** *Um procedimento de testes de diagnóstico para HIV é aplicado a uma população de alto risco; acredita-se que 10% desta população é positiva para o HIV. O teste de diagnóstico é positivo para 90% das pessoas que de fato são HIV-positivas, e negativo para 85% das pessoas que não são HIV-positivas. Qual a probabilidade de resultados falso-positivo e falso-negativo?*

Notação:

$A$  : a pessoa é HIV-positiva,

$B$  : o resultado do teste é positivo.

Temos dados que  $P[A] = 0,1$ ,  $P[B|A] = 0,9$  e  $P[\bar{B}|\bar{A}] = 0,85$ . Então:

$$\begin{aligned} P[\text{falso positivo}] &= P[\bar{A}|B] \\ &= \frac{P[B|\bar{A}]P[\bar{A}]}{P[B]} \\ &= \frac{0,15 \times 0,9}{(0,15 \times 0,9) + (0,9 \times 0,1)} = 0,6 \end{aligned}$$

<sup>2</sup>Já adiantando, podemos vislumbrar como isto será utilizado como a base para o procedimento de inferência: a verossimilhança nos informa sobre  $[y|\theta]$ , mas deseja-se fazer inferências baseadas em  $[\theta|y]$ . O Teorema de Bayes é a chave para tal inversão.

De forma similar,

$$\begin{aligned}
 P[\text{falso negativo}] &= P[A|\bar{B}] \\
 &= \frac{P[\bar{B}|A]P[A]}{P[\bar{B}]} \\
 &= \frac{0,1 \times 0,1}{(0,1 \times 0,1) + (0,85 \times 0,9)} = 0,0129
 \end{aligned}$$

Não há controvérsia aqui. Mas vamos ver um segundo exemplo que antecipa um pouco mais os tipos de questões que serão encontradas mais adiante.

**Exemplo 1.2** *Em uma sacola há seis bolas de cores desconhecidas. Três bolas são retiradas sem reposição e verifica-se que são pretas. Encontre a probabilidade de que não hajam bolas pretas restantes na urna.*

Seguindo a notação de eventos utilizada na sessão 1.6 denotaríamos por  $C_i$  o número de bolas na sacola e por  $A$  o evento que “foram retiradas três bolas pretas” e a probabilidade de interesse seria  $P[C_3|A]$ . Entretanto, vamos conciliar esta notação de eventos com a notação genérica denotando por  $Y$  a variável aleatória observada e por  $\theta$  o parâmetro que será adotada no próximo Capítulo. Sejam então:

$Y$  : número de bolas pretas entre três retiradas,

$Y = 3$  valor observado de  $Y$ ,

$\theta$  : número de bolas pretas na sacola,

$\theta \in \{0, 1, 2, \dots, 6\}$ ,

e denotamos por  $\theta_i$  quando  $\theta = i$ ,  $i = 0, 1, \dots, 6$ . Então, pelo Teorema de Bayes:

$$P[\theta_3|Y] = \frac{P[Y|\theta_3]P[\theta_3]}{\sum_{j=0}^6 P[Y|\theta_j]P[\theta_j]}$$

Mas há uma questão essencial aqui: quais valores atribuímos a  $P[\theta_0], \dots, P[\theta_6]$ ? Estas são as probabilidades dos diferentes números de bolas pretas na sacola, *antes* (a priori) de ter visto os dados (retirada de três bolas pretas). Não havendo nenhuma outra informação, pode-se muito bem assumir que todos os números de bolas pretas são igualmente prováveis, tomando portanto que  $P[\theta_0] = P[\theta_1] = \dots = P[\theta_6] = 1/7$ . De fato, tal especificação de priori será utilizada no exemplo daqui em diante. Mas, esta especificação é a mais razoável? Poderia-se tomar o ponto de vista de que é bastante provável que todas as bolas na sacola sejam da mesma cor, e consequentemente dar maior probabilidade a  $P[\theta_0]$  e  $P[\theta_7]$ . Ou então, poderia-se obter a informação do fabricante que são produzidas bolas de 10 cores diferentes. A partir desta informação poderia-se adotar o ponto de vista a priori de que cada bola é preta com probabilidade  $1/10$  e usar tal fato como base para calcular as probabilidades à priori. O ponto de que se quer fazer aqui é o de que é necessário *pensar*

bem sobre como expressar opiniões à priori e que a resposta que será obtida ao final poderá depender do que acreditamos ao início.

Esta discussão vai ser retomada mais adiante. Por agora, usando a especificação mencionada de priori mencionada inicialmente, simplesmente aplicamos o Teorema de Bayes para obter:

$$\begin{aligned}
 P[\theta_3|Y] &= \frac{P[Y|\theta_3]P[\theta_3]}{\sum_{i=0}^6 P[Y|\theta_i]P[\theta_i]} \\
 &= \frac{(\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4}) \times \frac{1}{7}}{\frac{1}{7} [0 + 0 + 0 + (\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4}) + (\frac{4}{6} \times \frac{3}{5} \times \frac{2}{4}) + (\frac{5}{6} \times \frac{4}{5} \times \frac{3}{4}) + (\frac{6}{6} \times \frac{5}{5} \times \frac{4}{4})]} \\
 &= \frac{1}{35}
 \end{aligned}$$

Desta forma, os dados atualizaram a opinião prévia que  $P[\theta_3] = 1/7$  para a probabilidade a posteriori  $P[\theta_3] = 1/35$ . Isto é, o evento torna-se muito menos provável após os dados serem obtidos.

Extendendo agora o exemplo pode-se calcular de forma análoga a probabilidade de haver uma, duas ou três bolas pretas restantes na sacola, o que corresponde a que houvessem quatro, cinco ou seis bolas pretas inicialmente, antes da retirada. Em outras palavras, podemos calcular  $P[\theta|Y] \equiv P[\theta_i|Y] \forall i$ , a posteriori completa para todos os possíveis valores de  $\theta$ . Os resultados são mostrados na Tabela 1.1 e as distribuições priori e posteriori  $P[\theta]$  e  $P[\theta|y]$  podem ser visualizadas na Figura 1.1.

Tabela 1.1: Posteriori de  $\theta$ , o número de bolas pretas na sacola, no Exemplo 1.2

$\theta$	0	1	2	3	4	5	6
$P[\theta Y=3]$	0	0	0	0,0285714	0,1142857	0,2857143	0,5714286

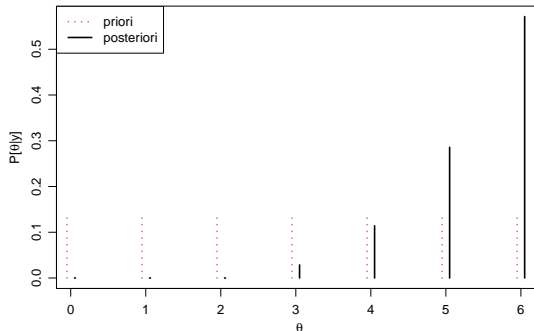


Figura 1.1: Distribuições priori e posteriori para o número de bolas pretas que havia inicialmente na urna no Exemplo 1.2.

## 1.7 Exercícios

**Exercício 1.1** Extratos de rocha A e B são difíceis de distinguir no campo. Através de estudos de laboratório detalhados foi determinado que a única característica que pode ser útil para ajudar discriminar é a presença de fóssil de um determinado animal marinho (brachipodos). Em exposições de rochas de tamanho usualmente encontrados, as probabilidades de presença do fóssil são mostradas na tabela 1.2. Sabe-se também que a rocha do tipo A ocorre com frequência quatro vezes maior do que do tipo B na área de estudo. Se uma amostra é tomada, e o fóssil é encontrado, calcula-se

Extrato	Fóssil Presente	Fóssil ausente
A	0,9	0,1
B	0,2	0,8

Tabela 1.2: Probabilidades de presença e ausência de fóssil em cada extrato.

a distribuição posteriori dos tipos de rocha.

Se um geólogo sempre classifica como A quando o fóssil é encontrado, e classifica como B quando ausente, qual a probabilidade de que vá acertar a próxima classificação?

**Exercício 1.2** Repita o Exemplo 1.2 usando uma escolha diferente de distribuição priori. De que forma esta mudança de priori afeta a probabilidade a posteriori de não haver bolas pretas restando na bolsa?

**Exercício 1.3** Ainda sobre o Exemplo 1.2,

1. Escreva um código para calcular a distribuição de probabilidades da Tabela 1.1 e o gráfico da Figura 1.1 da priori e posteriori.
2. Generalize o código de forma a aceitar diferentes distribuições à priori. Compare posterioris para diferentes opções de prioris.
3. Generalize ainda mais o código de forma a aceitar diferentes números de bolas pretas entre as três retiradas. Compare as posterioris para os diferentes números de bolas pretas retiradas.





## Capítulo 2

# Atualização Bayesiana

### 2.1 Introdução

Como delineado no Capítulo 1, a essência da abordagem bayesiana é tratar o parâmetro desconhecido  $\theta$  como uma variável aleatória, especificar uma distribuição a priori para  $\theta$  que represente as convicções sobre  $\theta$  antes de ver os dados, usar o Teorema de Bayes para atualizar as convicções a priori na forma de probabilidades a posteriori e fazer inferências apropriadas. Portanto há quatro passos característicos da abordagem bayesiana:

1. especificação da verossimilhança do modelo  $L(\theta|y) \equiv f(y|\theta)$ ;
2. determinação da priori  $f(\theta)$ ;
3. cálculo da distribuição posteriori  $f(\theta|y)$ , obtida pelo Teorema de Bayes;
4. extrair inferências da distribuição posteriori.

Neste Capítulo, o Teorema de Bayes será reexpresso em uma forma adequada para variáveis aleatórias ao invés de eventos e serão considerados questões que emergem quando se tenta usar tal resultado no contexto da inferência sobre um parâmetro  $\theta$ . As questões serão abordadas em capítulos subsequentes. Serão também examinados diversos exemplos em que particulares combinações de prioris e verossimilhanças produzem formas matematicamente convenientes para a distribuição a posteriori.

### 2.2 Teorema de Bayes

O Teorema da Bayes expresso em termos de variáveis aleatórias, com densidades denotadas genericamente por  $f(\cdot)$ , tem a forma:

$$f(\theta|y) = \frac{f(\theta)f(y|\theta)}{f(y)} = \frac{f(\theta)f(y|\theta)}{\int f(\theta)f(y|\theta)d\theta}. \quad (2.1)$$

Esta notação será usada em ambos os casos, com  $Y$  discreta ou contínua. No caso contínuo  $f$  é a função de densidade de probabilidade como usual, mas no caso discreto,  $f$  é a função de massa de probabilidade de  $Y$  ( $P[Y = y]$ ). De forma similar,  $\theta$  pode ser discreto ou contínuo mas, caso seja discreto,  $\int f(\theta)f(y|\theta)d\theta$  deve ser interpretado como  $\sum_j f(\theta_j)f(y|\theta_j)$ .

Pode-se então dizer que:

$$posteriori = \frac{priori \times verossimilhança}{marginal}.$$

Note que o denominador no Teorema de Bayes é uma função apenas de  $y$ , resultante de uma integração em  $\theta$  (ou seja,  $\theta$  foi "integrado fora"). Desta forma, uma outra maneira de escrever o Teorema de Bayes é:

$$f(\theta|y) \propto f(\theta)f(y|\theta)$$

ou, em palavras, “a posteriori é proporcional à priori vezes a verossimilhança”.

## 2.3 Tópicos

### 2.3.1 Escolha do modelo de verossimilhança

Isto depende do que se assume ser razoável sobre mecanismo gerador de dados do problema em questão, e é o mesmo problema encarado ao se usar inferência clássica – *qual o modelo adequado para os dados?* Em geral, o conhecimento da estrutura pela qual os dados foram gerados pode sugerir um modelo apropriado (por exemplo, amostragem binomial ou contagem Poisson), mas em geral o modelo será definido hipoteticamente ( $Y$  é relacionado com  $X$  com erros normais independentes, por exemplo) e sua plausibilidade é avaliada posteriormente no contexto dos dados.

### 2.3.2 Escolha da priori

Esta questão é fundamental para a abordagem bayesiana e será discutida em mais detalhes no Capítulo 3. Entretanto destacam-se desde já alguns pontos.

1. Devido ao fato de que a priori representa sua opinião sobre  $\theta$  antes de se observar os dados, segue-se que a análise subsequente é única à voce, ou seja à sua escolha de priori. Uma priori diferente de outra pessoa leva à uma análise à posteriori diferente. Nesse sentido a análise é subjetiva.
2. Será visto adiante que, desde que a priori não seja “completamente não razoável”, o efeito da priori tem menor influência uma vez que mais e mais dados se tornam disponíveis. Desta forma, em certo sentido, uma má especificação da priori deixa de ser importante desde que haja uma quantidade de dados disponíveis suficientes para isto.

3. Em geral, pode-se ter uma “ideia vaga” de como deve ser a priori (talvez a média e variância possam ser fornecidas), mas não se pode ser mais preciso do que isto. Em tais situações pode-se usar uma forma “conveniente” para a priori que seja consistente com a opinião prévia, mas que também torne o tratamento matemático simples e direto. Alguns exemplos disto serão vistos logo adiante.
4. Por vezes, pode-se ter a sensação de que não há conhecimento prévio sobre um parâmetro. Em tais situações pode-se desejar usar uma priori que reflita a ignorância sobre o parâmetro. Muitas vezes isto é possível, mas há algumas dificuldades envolvidas. Isto será discutido em mais detalhes no Capítulo 3.

### 2.3.3 Computação

Embora extremamente simples a princípio, na prática, a implementação do Teorema de Bayes pode ser difícil computacionalmente, principalmente devido a constante normalizadora do denominador. Será visto que para certas combinações priori-verossimilhança, a integração envolvida no cálculo do termo do denominador pode ser evitada mas, em geral, técnicas especializadas são necessárias para simplificar tal cálculo, o que será tratado no Capítulo 8.

### 2.3.4 Inferência

A análise bayesiana fornece a inferência mais completa possível, no sentido de que todo o conhecimento a respeito de  $\theta$ , disponível a partir da priori e dos dados, é representado pela distribuição posteriori. Isto é,  $f(\theta|y)$  é a inferência. Ainda assim, muitas vezes deseja-se resumir a inferência na forma de uma estimativa pontual ou intervalar. Isto será discutido no Capítulo 5.

### 2.3.5 Tópicos avançados

Em estatística clássica discute-se em considerável detalhamento o papel de estatísticas suficientes, ancilares, etc. Tais conceitos possuem papel análogo em inferência bayesiana mas, em geral, são mais atraentemente intuitivos. Por exemplo, em inferência bayesiana, a *suficiência* pode ser caracterizada dizendo que se particionamos os dados em  $y = (y_1, y_2)$ , então  $y_1$  é suficiente para  $\theta$  se  $f(\theta|y)$  é independente de  $y_2$ .

## 2.4 Exemplos

**Exemplo 2.1** *Quando uma máquina em particular se torna defeituosa, a causa pode ser atribuída a uma falha no motor ou a uma falha na transmissão. A localização da falha só pode ser determinada desmontando-se a máquina. Entretanto, a*

falha gera três tipos de sintomas observáveis: apenas aquecimento (AQ), tração irregular (TI), ou ambas. Registros anteriores foram usados para estabelecer as probabilidades na tabela 2.1. Além disto, sabe-se que 60% das falhas em máquinas deste tipo são devidas a transmissão, portanto  $f(\theta_2) = 0,60$ . Obtenha a distribuição a posteriori  $f(\theta|y)$  e interprete adequadamente os resultados.

Tabela 2.1: Tabela de probabilidades das causas de defeitos

Localização da falha	AQ ( $y_1$ )	TI ( $y_2$ )	ambas ( $y_3$ )
Motor	0,10	0,40	0,50
Transmissão	0,50	0,30	0,20

Neste exemplo os valores do parâmetro e da variável são categóricos e as análises podem ser tabuladas como mostrado na Tabela 2.2 sabendo-se que  $f(y, \theta) = f(y|\theta)f(\theta)$  e  $f(y) = \sum_{i=1}^2 f(y, \theta_i)$ .

Tabela 2.2: Tabela de probabilidades das causas de defeitos

$f(\theta)$	$f(y \theta)$	$y_1$	$y_2$	$y_3$
0,4	$\theta_1$	0,10	0,40	0,50
0,6	$\theta_2$	0,50	0,30	0,20
$f(y, \theta)$				
	$\theta_1$	0,04	0,16	0,20
	$\theta_2$	0,30	0,18	0,12
	$f(y)$	0,34	0,34	0,32
$f(\theta y)$	$\theta_1$	4/34	16/34	20/32
	$\theta_2$	30/34	18/34	12/32

Agora, uma forma de interpretar esta análise é verificando as “chances” (*odds*) para cada um dos dois tipos de falhas. Antes da informação sobre  $Y$ , as chances eram de 3:2 a favor de  $\theta_2$  (uma vez que 60% das falhas eram na transmissão). Mas tendo observado  $y$ , as chances mudaram para 15:2 a favor de  $\theta_2$  quando observa-se que  $y = y_1$ , 9:8 a favor de  $\theta_2$  se  $y = y_2$ , e 5:3 a favor de  $\theta_1$  se  $y = y_3$ . Consequentemente, se o critério de decisão é selecionar o diagnóstico de causa de falha mais plausível, observar  $y_1$  ou  $y_2$  levaria à decisão de que a falha é na transmissão, mas observando  $y_3$  levaria à decisão de que a falha seria no motor.<sup>1</sup>

**Exemplo 2.2** (*Amostragem binomial.*) Suponha o modelo  $Y \sim \text{Bin}(n, \theta)$ , e que deseja-se fazer inferências sobre  $\theta$ .

Portanto, a distribuição de probabilidades dos dados é

$$f(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}, \quad y = 0, \dots, n,$$

<sup>1</sup>Note-se que neste ponto está se adotando um critério muito simples de decisão. Na prática, diversas outras considerações relevantes podem ser levadas em conta. Por exemplo, desmontar a transmissão pode ser muito mais trabalhoso do que o motor, e portanto a decisão poderia ser desmontar o motor primeiro, mesmo que a maiores chances a posteriori apontassem para falha na transmissão. Considerações deste tipo serão discutidas em *análise bayesiana de decisão* no Capítulo 5.

e a verossimilhança

$$L(\theta|y) \propto \theta^y (1 - \theta)^{n-y}, \quad 0 < \theta < 1. \quad (2.2)$$

É claro que, em geral, a escolha da especificação da priori para  $\theta$  vai variar de problema para problema e, por definição, vai depender da extensão do conhecimento prévio sobre a situação. Entretanto, procede-se aqui considerando uma possível família de distribuições a priori que, como será visto, dará origem a computações simplificadas. Um ponto relevante a este respeito é que está se assumindo o seguinte: uma vez que a família de distribuições de probabilidade seja bastante vasta e inclua uma diversidade suficiente de formas possíveis, pode-se usar uma priori desta família que se aproxime das verdadeiras opiniões prévias. Se isto ocorre, consegue-se respostas, ou seja, expressões da posteriori, calculadas de forma simples. Se, entretanto, não há dentro desta família uma priori que pareça com o que realmente se acredita, então tal abordagem deve ser evitada.

Desta forma, neste caso, suponha que pode-se representar as opiniões a priori sobre  $\theta$  por uma distribuição Beta:

$$\theta \sim \text{Beta}(p, q)$$

tal que

$$\begin{aligned} f(\theta) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \theta^{p-1} (1-\theta)^{q-1} \quad (0 < \theta < 1) \\ &\propto \theta^{p-1} (1-\theta)^{q-1}. \end{aligned}$$

Então, pelo Teorema de Bayes

$$\begin{aligned} f(\theta|y) &\propto f(\theta)f(y|\theta) \\ &\propto \theta^y (1-\theta)^{n-y} \times \theta^{p-1} (1-\theta)^{q-1} \\ &= \theta^{p+y-1} (1-\theta)^{q+n-y-1}. \end{aligned}$$

Agora, como sabemos que  $f(\theta|y)$  é uma densidade de probabilidades própria, a expressão anterior permite identificar que

$$\theta|y \sim \text{Beta}(p+y; q+n-y). \quad (2.3)$$

É importante notar que na expressão da verossimilhança (2.2) pode-se identificar o núcleo de uma  $\text{Beta}(y+1, n-y+1)$ . Desta forma, através de uma cuidadosa escolha, obtém-se uma distribuição a posteriori que pertence à mesma família que a distribuição a priori, e fazendo isto, evita-se o cálculo de qualquer integral. O efeito dos dados é simplesmente modificar os parâmetros da distribuição priori  $\text{Beta}(p, q)$  para  $\text{Beta}(p+y, q+n-y)$ .

Como um exemplo numérico, considere o conjunto de dados “CANCER” tomado do programa *First Bayes*.<sup>2</sup> O problema consiste em estimar a proporção de

<sup>2</sup>*First Bayes* é um programa computacional, escrito por Tony O'Hagan, que ilustra uma variedade de análise bayesianas. Está disponível em <http://tonyohagan.co.uk/1b/>.

sobreviventes, além de um período especificado, entre pacientes que recebem tratamento segundo um novo protocolo para uma particular forma de câncer. Denote por  $\theta$  a probabilidade de sobrevivência dos pacientes, que é o parâmetro de interesse para inferência. Consultas com médicos especialistas, que são familiarizados com ensaios clínicos similares os levam a expressar o conhecimento a priori que  $E[\theta] = 0,40$  e  $\text{Var}[\theta] = 0,02$ . Agora, se a distribuição beta é considerada razoável para representar os conhecimentos prévios, então deve-se escolher uma distribuição a priori  $\theta \sim \text{Beta}(p, q)$  tal que  $E[\theta] = 0,40$  e  $\text{Var}[\theta] = 0,02$ . Tem-se então que:

$$\frac{p}{p+q} = 0,40 \quad \text{e} \quad \frac{pq}{(p+q)^2(p+q+1)} = 0,02.$$

É possível verificar (*faça isto!*) que, fazendo  $m = E[\theta]$  e  $v = \text{Var}[\theta]$  as equações são resolvidas por:

$$p = \frac{(1-m)m^2}{v} - m \quad \text{e} \quad q = \frac{(1-m)^2 m}{v} - (1-m),$$

o que neste caso leva a  $p = 4,4$  e  $q = 6,6$ . Isto especifica a distribuição da priori para  $\theta$ , sendo um exemplo simples de elicitação de priori. Na prática seria necessário assegurar que toda a distribuição (e não apenas a média e variância dados) seja consistente com as opiniões anteriores.

Passando agora aos dados, tem-se que de 70 pacientes que receberam o tratamento, 34 sobreviveram além do período definido previamente. Presumindo-se que a priori assim definida é adequada e a verossimilhança é binomial, obtém-se a partir de 2.3 a distribuição posteriori como sendo:

$$\theta \sim \text{Beta}(p+y=4,4+34=38,4; q+n-y=6,6+70-34=42,6).$$

Esta distribuição posteriori sumariza toda a informação a respeito de  $\theta$  e representa a completa inferência sobre  $\theta$ . Será discutido posteriormente como, se necessário, pode-se resumir esta inferência, mas por agora pode-se ver como os dados modificam a opinião anterior comparando as esperanças da priori e posteriori:

$$E[\theta] = \frac{p}{p+q} \quad \text{e} \quad E[\theta|y] = \frac{p+y}{p+q+n}.$$

No caso considerado:

$$E[\theta] = 0,4 \quad \text{e} \quad E[\theta|y] = 0,474$$

e portanto, o efeito dos dados observados foi aumentar a estimativa à priori de  $\theta$  de 0,40 para 0,474. Por outro lado, um estimador natural para  $\theta$ , baseado apenas nos dados, seria  $y/n = 34/70 = 0,486$ , que é o estimador de máxima verossimilhança. A estimativa à posteriori é um compromisso entre a opinião prévia e a informação fornecida pelos dados. A Figura 2.1 ilustra os resultados. A função de verossimilhança é padronizada por sua integral de forma a ter integral unitária para poder ser representada na mesma escalas das distribuições priori e posteriori. Neste caso

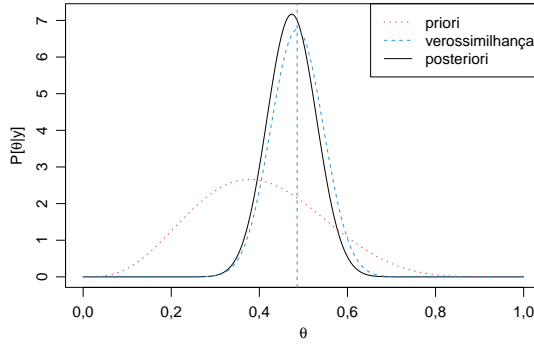


Figura 2.1: Priori, verossimilhança e posteriori para o Exemplo 2.2.

a padronização é trivial uma vez que o núcleo da verossimilhança corresponde a uma Beta( $y + 1, n - y + 1$ ).

De forma mais geral, se  $y$  e  $n$  são grandes em relação a  $p$  e  $q$ , então a esperança na posteriori é aproximadamente  $y/n$ , a estimativa de máxima verossimilhança. Por outro lado, se  $p$  e  $q$  são moderadamente elevados, então terão uma influência razoável na média a posteriori. Também pode ser verificado que a medida que  $y$  e  $n$  aumentam – ou mesmo se valores elevados são atribuídos a  $p$  e  $q$  – então a variância a posteriori diminui.

**Exemplo 2.3** (*Amostragem Poisson.*) Suponha que  $Y_1, \dots, Y_n$  são um conjunto de variáveis aleatórias independentes com distribuição Poisson  $Y|\theta \sim P(\theta)$ .

Tem-se que a verossimilhança é obtida pela expressão da distribuição conjunta que neste caso é:

$$\begin{aligned} L(\theta|y) &\equiv \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n \frac{e^{-\theta} \theta^{y_i}}{y_i!} \\ &\propto e^{-n\theta} \theta^{\sum y_i}. \end{aligned}$$

Aqui a expressão da verossimilhança corresponde ao núcleo de uma distribuição  $\text{Ga}(1 + \sum_{i=1}^n y_i, n)$ .

Como no caso binomial, convicções anteriores sobre  $\theta$  irão variar de problema para problema, mas procura-se aqui por uma forma que acomode diferentes possibilidades, mas sendo também matematicamente tratável. Neste caso supõe-se que as convicções à priori podem ser representadas por uma distribuição Gama,  $\theta \sim \text{Ga}(p, q)$  e então

$$f(\theta) = \frac{q^p}{\Gamma(p)} \theta^{p-1} \exp\{-q\theta\} \quad (\theta > 0).$$

Assim, pelo Teorema de Bayes,

$$\begin{aligned} f(\theta|y) &\propto \frac{q^p}{\Gamma(p)} \theta^{p-1} \exp\{-q\theta\} \times \theta^{\sum y_i} \exp\{-n\theta\} \\ &\propto \theta^{p+\sum y_i-1} \exp\{-(q+n)\theta\} \end{aligned}$$

e conseqüentemente,

$$\theta|y \sim \text{Ga}\left(p + \sum_{i=1}^n y_i, q + n\right), \quad (2.4)$$

uma outra distribuição gama cujos parâmetros, em comparação com a priori, são modificados pelos dados através de  $\sum_{i=1}^n y_i$  e  $n$ . Note-se que os valores individuais de  $y_i$  não são necessários, apenas a sua soma. Dizemos então que  $\sum_{i=1}^n y_i$  é suficiente para  $\theta$ .

Como um exemplo numérico, novamente tomado do programa *First Bayes*, seja  $\theta$  o número médio de gansos de um bando dentro de uma determinada região. Supõe-se que a média e variância a priori para  $\theta$  são 100 e 20, respectivamente. Desta forma, usando o fato de que se  $\theta \sim \text{Ga}(p, q)$  então  $E[\theta] = p/q$  e  $\text{Var}[\theta] = p/q^2$ , resolvendo para  $p$  e  $q$  obtém-se  $p = 500$  e  $q = 5$ , ou seja fica definida a priori  $\theta \sim \text{Ga}(500, 5)$ . Os dados provêm de fotografias aéreas detalhadas de 45 bandos que fornecem  $\sum_{i=1}^{45} y_i = 4019$ . Portanto, a distribuição a posteriori obtida através de 2.4 é  $\theta|y \sim \text{Ga}(4519, 50)$ . A Figura 2.2 permite visualizar a priori, verossimilhança e posteriori do Exemplo.

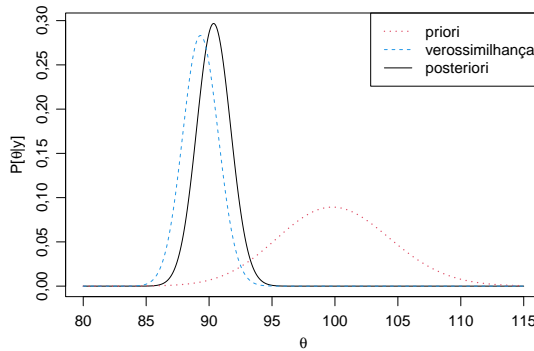


Figura 2.2: Priori, verossimilhança e posteriori para o Exemplo 2.3.

**Exemplo 2.4** (*Média da normal.*) Suponha que  $Y_1, \dots, Y_n$  são um conjunto de variáveis aleatórias independentes com distribuição  $N(\theta, \sigma^2)$ , em que  $\sigma^2$  é conhecido.

Então,

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



e a verossimilhança fica

$$\begin{aligned} L(\theta|y) &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2} \right\} \\ &\propto \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2} \right\} \end{aligned}$$

Após algum algebrismo é possível mostrar que esta expressão pode ser escrita na forma

$$L(\theta|y) \propto \exp \left\{ -\frac{(\theta - \bar{y})^2}{2(\sigma^2/n)} \right\},$$

que corresponde ao núcleo de uma distribuição  $N(\bar{y}, \sigma^2/n)$ .

Agora, suponha que as convicções a priori sobre  $\theta$  podem elas mesmas serem representadas por uma distribuição normal  $\theta \sim N(b, d^2)$ . Novamente, esta escolha visa obter uma análise matemática simples, mas deve apenas ser usada se tal escolha é de fato uma boa aproximação à crença a priori sobre  $\theta$ . Então, pelo Teorema de Bayes,

$$\begin{aligned} f(\theta|y) &\propto \exp \left\{ -\frac{(\theta - b)^2}{2d^2} \right\} \exp \left\{ -\frac{(\theta - \bar{y})^2}{2(\sigma^2/n)} \right\} \\ &= \exp \left\{ -\frac{1}{2} \left[ \frac{(\theta - b)^2}{d^2} + \frac{(\theta - \bar{y})^2}{\sigma^2/n} \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left[ \left( \frac{1}{d^2} + \frac{1}{\sigma^2/n} \right) \theta^2 - 2\theta \left( \frac{b}{d^2} + \frac{\bar{y}}{\sigma^2/n} \right) \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \left( \frac{1}{d^2} + \frac{1}{\sigma^2/n} \right) \left[ \left( \theta - \frac{\frac{b}{d^2} + \frac{\bar{y}}{\sigma^2/n}}{\frac{1}{d^2} + \frac{1}{\sigma^2/n}} \right)^2 \right] \right\} \end{aligned}$$

Portanto,

$$\theta|y \sim N \left( \frac{\frac{b}{d^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}} \right) \quad (2.5)$$

Este resultado é expresso de forma mais concisa definindo-se “precisão” como sendo o recíproco da variância, i.e., seja  $\tau = 1/\sigma^2$  e  $c = 1/d^2$ . Então,

$$\theta|y \sim N \left( \frac{cb + n\tau\bar{y}}{c + n\tau}, \frac{1}{c + n\tau} \right). \quad (2.6)$$

Antes de ver um exemplo numérico, algumas observações podem ser feitas.

1. Observe-se que

$$E[\theta|y] = \gamma_n b + (1 - \gamma_n) \bar{y}$$

em que

$$\gamma_n = \frac{c}{c + n\tau},$$

ou seja, a média da posteriori é simplesmente uma média ponderada entre a média da priori e  $\bar{y}$ . Além do mais, o parâmetro ponderador  $\gamma_n$  é determinado pela força relativa da informação na priori em comparação com a dos dados. Isto é, se  $n\tau$  é grande relativamente a  $c$ , então  $\gamma_n \approx 0$  e a média da posteriori é próxima de  $\bar{y}$ .

2. Observe que “precisão à posteriori” = “precisão à priori” +  $n \times$  “precisão de cada dado”.
3. Quando  $n \rightarrow \infty$ , então (vagamente)

$$\theta|y \sim N(\bar{y}, \frac{\sigma^2}{n})$$

de tal forma que no limite a priori não tem efeito.

4. Quando  $d \rightarrow \infty$ , ou equivalentemente,  $c \rightarrow 0$ , novamente se obtém que

$$\theta|y \sim N(\bar{y}, \frac{\sigma^2}{n}).$$

5. Note-se que a distribuição posteriori depende dos dados apenas através de  $\sum_{i=1}^n y_i$  e não através dos valores individuais dos  $y_i$ . Novamente, como ocorria no modelo Poisson-Gama, dizemos que  $\sum_{i=1}^n y_i$  é suficiente para  $\theta$ .

Os pontos 3 e 4 são sutis e serão discutidos posteriormente com maiores detalhes.

Como um exemplo numérico, também tomado do programa *First Bayes*, considera-se um conjunto histórico de dados de densidade do solo em uma região registrados por Henry Cavendish no século XVIII. Agora, supõe-se que, de experimentos e medições prévios, a priori para  $\theta$ , a densidade média do solo, é considerada ser  $N(5,4; 0,1^2)$ . O pesquisador registrou 23 medidas da densidade do solo. Para estes dados,  $\bar{y} = 5,48$  e supõe-se aqui que a variância de seu erro de medida é conhecida e igual a 0,04. Então, temos que a posteriori é simplesmente obtida através de 2.5 como sendo  $\theta|y \sim N(5,468; 0,038^2)$ .

## 2.5 Tópicos gerais

Os princípios e detalhes encontrados nos exemplos anteriores levantam uma série de questionamentos que serão discutidos agora.

### 2.5.1 Atualização sequencial

Foi visto que o Teorema de Bayes fornece um mecanismo pelo qual a informação anterior é atualizada pelos dados fornecendo a informação a posteriori. Então, essa última serve como a “nova” informação a priori antes de mais dados serem disponibilizados. Isto dá origem a uma questão em particular: se uma sequência de

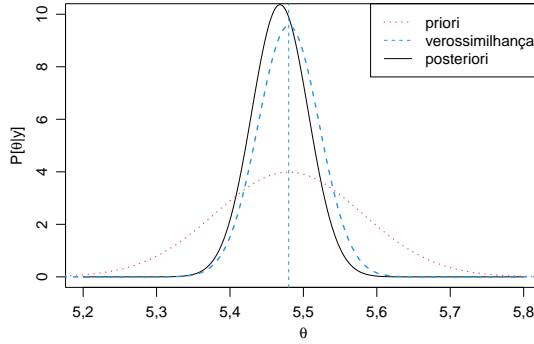


Figura 2.3: Priori, verossimilhança e posteriori para o Exemplo 2.4.

dados é obtida, e a opinião é atualizada na chegada de cada novo dado individualmente, o resultado obtido ao final seria diferente do que o obtido caso se esperasse até que todos os dados fossem disponibilizados e só então atualizar a priori, de uma só vez?

Para responder esta pergunta, considere o caso simples em que se observa  $y_1$ , e atualizando a priori pelo Teorema de Bayes obtém-se:

$$f(\theta|y_1) = c_1 \cdot f(\theta) \cdot f(y_1|\theta) \propto f(\theta) \cdot f(y_1|\theta)$$

em que  $c_1$  é a constante normalizadora. Esta  $f(\theta|y_1)$  se torna a nova distribuição a priori antes de  $y_2$  ser observado. Então,

$$\begin{aligned} f(\theta|y_1, y_2) &= c_2 \cdot f(\theta) \cdot f(y_1|\theta) \cdot f(y_2|\theta) \\ &\propto f(\theta) \cdot f(y_1|\theta) \cdot f(y_2|\theta) \\ &= f(\theta) \cdot f(y_1, y_2|\theta) \end{aligned}$$

que é o mesmo resultado que seria obtido atualizando-se diretamente com base em toda a informação  $(y_1, y_2)$ . Por indução, o argumento estende-se para sequências de qualquer número de observações.

## 2.5.2 Suficiência

Em estatística clássica, a *suficiência* tem um papel central em ambos, desenvolvimentos teóricos e aplicações práticas. O mesmo ocorre em análise bayesiana. Já foram vistos aqui exemplos em que a distribuição a posteriori depende dos dados apenas através de uma estatística suficiente.

Há uma caracterização adicional natural de suficiência sob o enfoque bayesiano, que, embora atrativa intuitivamente, não tem lugar em estatística clássica. Suponha que os dados podem ser particionados em  $y = (y_1, y_2)$ , então  $y_1$  é suficiente para  $\theta$  se  $f(\theta|y)$  é independente de  $y_2$ , (neste caso  $y_2$  é dito ser *ancilar*). Pode-se provar que isto é equivalente a suficiência no sentido de inferência clássica.

### 2.5.3 O princípio da verossimilhança

O princípio da verossimilhança estabelece que se dois experimento produzem a mesma verossimilhança (proporcionalmente), então as inferências sobre  $\theta$  devem ser a mesmas em ambos os casos. Em outras palavras, todos os aspectos de inferência devem ser baseados apenas na função de verossimilhança. A principal virtude da abordagem bayesiana é a de que técnicas bayesianas são inerentemente consistentes com o princípio da verossimilhança, enquanto que diversos procedimentos básicos de estatística clássica (frequentista) violam tal princípio.

**Exemplo 2.5** *Suponha uma urna com bolas pretas e vermelhas e deseja-se fazer inferência sobre a proporção ( $\theta$ ) de bolas pretas na urna. Vamos adotar a priori  $\theta \sim \text{Beta}(1,2; 1,2)$ . Considere ainda que foram conduzidos dois experimentos distintos e separadamente:*

- Foram retiradas oito bolas (com reposição ou considera-se a urna com um número muito grande de bolas) e constatou-se que duas eram pretas.*
- Definiu-se que seriam retiradas bolas da urna até obter a segunda bola preta. Foram retiradas seis vermelhas.*

No primeiro caso temos que a variável aleatória observada  $Y_a$  é o número de bolas pretas, possui distribuição binomial  $Y_a \sim B(n = 8, \theta)$  e o dado observado é  $y_a = 2$ . Temos então que:

$$\begin{aligned} f(y_a|\theta) &= \binom{8}{2} \theta^2 (1-\theta)^6 \\ f(\theta|y_a) &\propto \theta^{2+1,2-1} (1-\theta)^{6+1,2-1} = \theta^{3,2-1} (1-\theta)^{7,2-1} \\ \theta|y_a &\sim \text{Beta}(3,2; 7,2) \end{aligned}$$

Já no segundo experimento, a variável observada  $Y_b$  se refere ao número de vermelhas até obter a segunda preta e portanto distribuição binomial negativa  $Y_b \sim \text{BN}(r = 2, \theta)$  e o dado observado é  $y = 6$ . e temos que

$$\begin{aligned} f(y_b|\theta) &= \binom{7}{1} \theta^2 (1-\theta)^6 \\ f(\theta|y_b) &\propto \theta^{2+1,2-1} (1-\theta)^{6+1,2-1} = \theta^{3,2-1} (1-\theta)^{7,2-1} \\ \theta|y_b &\sim \text{Beta}(3,2; 7,2) \end{aligned}$$

Portanto, como os núcleos das verossimilhanças são iguais, ou seja, as verossimilhanças só diferem por termos constantes, as posteriores são idênticas e portanto as inferências sobre  $\theta$  serão exatamente as mesmas nos dois experimentos.

Um procedimento de teste de hipótese frequentista, com inferência baseada no  $p$ -valor, fornece resultados diferentes nos dois casos. Seja por exemplo testar  $H_0 : p = 0,5$  versus  $H_1 : p < 0,5$ . Teríamos que em cada um dos experimentos os resultados seriam:

a)  $p\text{-valor} = P_{\text{Bin}}[Y_a \leq 2 | \theta = 0,5; n = 8] = 0,145$

b)  $p\text{-valor} = P_{\text{BN}}[Y_b \geq 6 | \theta = 0,5; r = 2] = 0,0352$

ou seja, inferências claramente diferentes, violando o princípio da verossimilhança.

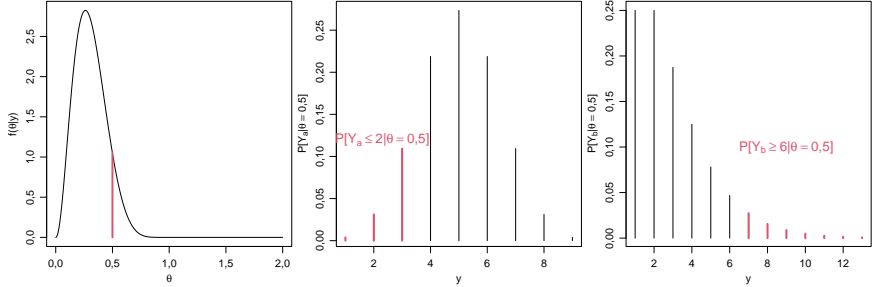


Figura 2.4: Distribuições posteriori e amostrais para inferência sobre parâmetro do Exemplo 2.5.

É ainda interessante notar que, o que difere os dois experimentos é o *critério de parada* levando a diferentes distribuições de probabilidades desconhecidas. Se tivéssemos um terceiro experimento que forneceu seis bolas vermelhas e duas pretas, porém com critério de parada desconhecido (ou não usual) não teríamos como proceder o teste não bayesiano, enquanto que para a inferência bayesiana a verossimilhança ainda seria a mesma (exceto por alguma constante) levando à mesma posteriori  $\theta|y \sim \text{Beta}(3,2; 7,2)$ .

## 2.5.4 Distribuição marginal

Sendo uma priori  $f(\theta)$  e um modelo verossimilhança definido por  $f(y|\theta)$  a distribuição marginal é obtida por

$$f(y) = \int f(y, \theta) d\theta = \int_{\theta} f(y|\theta) f(\theta) d\theta. \quad (2.7)$$

Esta distribuição pode ser chamada de *distribuição preditiva a priori* uma vez que é a distribuição de uma quantidade observável mas que não depende observações prévia do processo.

Tal expressão pode ser reescrita ressaltando que a priori pode depender de hiperparâmetros que denotamos aqui por “a”. Temos então que:

$$f(y|a) = \int f(y, \theta|a) d\theta = \int_{\theta} f(y|\theta, a) f(\theta|a) d\theta.$$

Tal notação deixa claro que especificar a priori é equivalente a especificar a distribuição  $f(y|a)$  para observações. Isto é equivalente à especificação de modelos com efeitos aleatórios.

Em uma expressão alternativa, o Teorema de Bayes em 2.1 pode ser reescrito da forma:

$$f(y) = \frac{f(y|\theta)f(\theta)}{f(\theta|y)}. \quad (2.8)$$

Em poucos casos a expressão de  $f(y)$  pode ser obtida analiticamente e levando à alguma forma conhecida de distribuição de probabilidades. Além da interessante intuição, por vezes é possível extrair resultados de interesse prático.

Seja o contexto do **Exemplo 2.3** da distribuição de Poisson. Neste caso  $a = (p, q)$  e utilizando (2.7) ou (2.8) obtém-se que:

$$f(y) = \binom{p+y-1}{y} \left( \frac{q}{1+q} \right)^p \left( \frac{1}{1+q} \right)^y$$

que é a expressão de uma distribuição binomial negativa com número de sucessos  $p$  e probabilidade de sucesso  $q/(p+q)$ . Por propriedades da distribuição binomial negativa temos que:

$$\begin{aligned} E[Y|p, q] &= p \frac{1/(1+q)}{q/(1+q)} = \frac{p}{q}, \\ \text{Var}[Y|p, q] &= p \frac{1/(1+q)}{(q/(1+q))^2} = \frac{p}{q^2} (1+q). \end{aligned}$$

Entretanto, na maioria dos casos a distribuição marginal de  $Y$  não corresponde a uma distribuição conhecida ou não pode ser obtida analiticamente. Ainda assim podemos calcular as médias e variâncias da marginal como se segue.

$$\begin{aligned} E[Y_i|p, q] &= E_\theta[E_Y[Y_i|\theta]|p, q] \\ &= E_\theta[\theta|p, q] \\ &= \frac{p}{q} \end{aligned}$$

$$\begin{aligned} \text{Var}[Y_i|p, q] &= E_\theta[\text{Var}_Y[Y_i|\theta]|p, q] + \text{Var}_\theta[E_Y[Y_i|\theta]|p, q] \\ &= E_\theta[\theta|p, q] + \text{Var}_\theta[\theta|p, q] \\ &= \frac{p}{q} + \frac{p}{q^2} \\ &= \frac{p}{q^2} (1+q) \\ &= E[Y_i|p, q](1+q) \end{aligned}$$

Como  $(1+q)$  é sempre positivo a variância da marginal é maior que a média, diferentemente da distribuição de Poisson que tem variância igual à média. Desta forma, ao atribuir a priori especificada, o modelo passa a incluir a possibilidade de acomodar dados *superdispersos*. Quanto menor o valor de  $q$  na priori, maior será a superdispersão e, quanto maior, mais a marginal se aproxima da distribuição de Poisson.

## 2.6 Códigos computacionais

**Exemplo 2.1:** As operações para cálculo da posteriori podem ser escritas na forma de operações com matrizes

```
fth <- c(0.4, 0.6)
fy.th <- matrix(c(0.1, 0.5, 0.4, 0.3, 0.5, 0.2), nc = 3, dimnames = list(c("th1",
"th2"), c("y1", "y2", "y3")))
(fyth <- fth * fy.th)

##      y1  y2  y3
## th1 0,04 0,16 0,20
## th2 0,30 0,18 0,12

(fy <- colSums(fyth))

##      y1  y2  y3
## 0,34 0,34 0,32

(fth.y <- MASS::fractions(t(t(fyth)/drop(fy))))

##      y1  y2  y3
## th1 2/17 8/17 5/8
## th2 15/17 9/17 3/8
```

**Exemplo 2.2:** Os comandos a seguir permitem reproduzir a Figura 2.1. Note que a expressão da verossimilhança é o núcleo de uma distribuição Beta( $y + 1, n - y + 1$ ).

```
curve(dbeta(x, 38.4, 42.6), from = 0, to = 1, n = 1001, xlab = expression(theta),
      ylab = expression(paste("P[" , theta, "|y]")))
curve(dbeta(x, 4.4, 6.6), from = 0, to = 1, add = TRUE, col = 2, lty = 3, lwd = 1.5,
      n = 1001)
abline(v = 34/70, lty = 2, col = 4)
curve(dbeta(x, 34 + 1, 70 - 34 + 1), from = 0, to = 1, add = TRUE, col = 4,
      lty = 2, n = 1001)
legend("topright", c("priori", "verossimilhança", "posteriori"), lty = c(3,
2, 1), col = c(2, 4, 1), lwd = c(1.5, 1, 1))
```

**Exemplo 2.3:** Os comandos a seguir permitem reproduzir a Figura 2.2. Note que a expressão da verossimilhança é o núcleo de uma distribuição  $G(\sum_i y_i + 1, n)$ .

```
curve(dgamma(x, 4519, 50), from = 80, to = 115, n = 1001, xlab = expression(theta),
      ylab = expression(paste("P[" , theta, "|y]")))
curve(dgamma(x, 500, 5), from = 80, to = 115, n = 1001, add = TRUE, lty = 3,
      lwd = 1.5, col = 2)
curve(dgamma(x, 4019 + 1, 45), from = 80, to = 115, n = 1001, add = TRUE, lty = 2,
      col = 4)
legend("topright", c("priori", "verossimilhança", "posteriori"), lty = c(3,
2, 1), col = c(2, 4, 1), lwd = c(1.5, 1, 1))
```

**Exemplo 2.4:** Os comandos a seguir permitem reproduzir a Figura 2.3. Note que a expressão da verossimilhança é o núcleo de uma distribuição  $N(\bar{y}, \sigma^2/n)$ .

```
curve(dnorm(x, mpost, sqrt(vpost)), from = 5.2, to = 5.8, n = 401, ylim = c(0,
  10), xlab = expression(theta), ylab = expression(paste("P[", theta, "|y]")))
curve(dnorm(x, mean = yb, sd = d), from = 5, to = 6, n = 401, add = TRUE, lty = 3,
  lwd = 1.5, col = 2)
abline(v = 5.48, col = 4, lty = 2)
curve(dnorm(x, 5.48, sqrt(sig2/n)), from = 5, to = 6, n = 401, add = TRUE, lty = 2,
  lwd = 1.5, col = 4)
legend("topright", c("priori", "verossimilhança", "posteriori"), lty = c(3,
  2, 1), col = c(2, 4, 1), lwd = c(1.5, 1, 1))
```

**Exemplo 2.5:** Os comandos a seguir permitem reproduzir a Figura 2.4.

```
par(mfrow = c(1, 3), mar = c(3, 3, 0.5, 0.5), mgp = c(2, 0.8, 0))
curve(dbeta(x, 3.2, 7.2), n = 501, from = 0, to = 2, xlab = expression(theta),
  ylab = expression(paste("f(", theta, "|y)")))
segments(0.5, 0, 0.5, dbeta(0.5, 3.2, 7.2), col = 2, lwd = 2)
plot(dbinom(0:8, 8, 0.5), type = "h", xlab = "y", ylab = expression(paste("P[",
  Y[a], "|", theta == 0.5, "]")))
lines(dbinom(0:2, 8, 0.5), type = "h", col = 2, lwd = 2)
text(2.3, 0.12, expression(paste("P[", Y[a] <= 2, "|", theta == 0.5, "]")),
  col = 2, cex = 1.2)
plot(dnbinom(0:12, 2, 0.5), type = "h", xlab = "y", ylab = expression(paste("P[",
  Y[b], "|", theta == 0.5, "]")))
lines(7:13, dnbinom(6:12, 2, 0.5), type = "h", col = 2, lwd = 2)
text(9, 0.1, expression(paste("P[", Y[b] >= 6, "|", theta == 0.5, "]")), col = 2,
  cex = 1.2)
```



## 2.7 Exercícios

**Exercício 2.1** Em cada um dos casos a seguir, obtenha a distribuição posteriori:

1.  $y_1, \dots, y_n$  é uma amostra aleatória de uma distribuição com função de probabilidades:

$$f(y|\theta) = \theta^{y-1}(1-\theta); \quad y = 1, 2, \dots$$

com distribuição a priori Beta( $p, q$ ) com densidade:

$$f(\theta) = \frac{\theta^{p-1}(1-\theta)^{q-1}}{B(p, q)}, \quad 0 < \theta < 1.$$

2.  $y_1, \dots, y_n$  é uma amostra aleatória de uma distribuição com função probabilidades:

$$f(y|\theta) = \frac{e^{-\theta}\theta^y}{y!}; \quad y = 0, 1, \dots$$

com distribuição a priori

$$f(\theta) = e^{-\theta}; \quad 0 < \theta.$$

**Exercício 2.2** A proporção  $\theta$  de itens defeituosos em um grande carregamento é desconhecida, mas uma avaliação especializada atribui a  $\theta$  a distribuição a priori Beta(2, 200). Se 100 itens são selecionados ao acaso do carregamento, e são encontrados três defeituosos, qual a distribuição a posteriori de  $\theta$ ?

Qual seria a distribuição a priori adotada por um outro estatístico que, tendo observado três defeituosos, calcula a distribuição a posteriori como sendo uma beta de média 4/102 e variância 0,0003658?

**Exercício 2.3** O diâmetro de um componente em uma longa sequência de produção varia segundo uma distribuição  $N(\theta, 1)$ . Um engenheiro especifica a distribuição a priori de  $\theta$  como sendo  $N(10; 0,25)$ . Em uma sequência de produção 12 componentes são amostrados e encontra-se que a média amostral é de 31/3. Use esta informação para calcular a probabilidade de que o diâmetro médio do componente é de pelo menos 10 unidades.

**Exercício 2.4** O número de defeitos em um rolo de 1200 metros de uma fita magnética possui distribuição Poisson( $\theta$ ). A distribuição priori para  $\theta$  é Gama(3; 1). Quando cinco rolos deste tipo são selecionados ao acaso, o número de defeitos encontrados em cada um deles é: 2, 2, 6, 0 e 3, respectivamente. Determine a distribuição a posteriori de  $\theta$ .

**Exercício 2.5** Suponha que o tempo em minutos necessário para atender um cliente em um banco possui uma distribuição exponencial com parâmetro  $\theta$ . A priori estabelecida para  $\theta$  é a distribuição Gama com média 0,2 e desvio padrão 1. Se o tempo médio observado para atender 20 clientes é de 3,8 minutos, determine a distribuição posteriori para  $\theta$ .

**Exercício 2.6** Uma amostra aleatória  $y_1, \dots, y_n$  é tomada de uma distribuição Poisson de média  $\theta$ . A distribuição a priori para  $\theta$  é Gama com média  $\mu_0$ . Se a média amostral é  $\bar{y}_n$ , mostre que a média da distribuição a posteriori de  $\theta$  vai ser uma média ponderada da forma

$$\gamma_n \bar{y}_n + (1 - \gamma_n) \mu_0,$$

e mostre que  $\gamma \rightarrow 1$  quando  $n \rightarrow \infty$ .

**Exercício 2.7** Considere o **Exemplo 2.3**. Suponha que a amostra tenha sido coletada em três visitas a campo. Na primeira teriam sido avistados 11 bandos e um total de 1150 animais. Na segunda seriam 21 bandos e 1880 animais e finalmente na terceira 13 bandos com o total de 989 animais. Obtenha a distribuição à posteriori que pode ser obtida após coletar os dados de cada visita. Obtenha ainda a média e moda das posteriores. Verifique a característica de atualização sequencial da abordagem bayesiana: repare que os números totais de bandos e animais somados são iguais ao fornecidos no **Exemplo 2.3** bem como a posteriori obtida ao final. Produza um gráfico com a priori inicial e as posteriores.

## Capítulo 3

# Especificação de prioris

### 3.1 Introdução

Foi visto que a diferença fundamental entre estatística clássica e bayesiana é que em estatística bayesiana parâmetros desconhecidos são tratados como variáveis aleatórias, e que o uso do Teorema de Bayes requer a especificação de prioris para tais parâmetros. Ainda que isto facilite a inclusão de convicções anteriores genuínas sobre os parâmetros, por outro lado, a escolha da distribuição a priori não pode ser feita cegamente; deve haver um cuidado considerável e há alguns tópicos bastante relevantes envolvidos. Neste capítulos são vistos alguns destes tópicos.

### 3.2 Prioris conjugadas

As dificuldades computacionais na utilização do Teorema de Bayes surgem quando é necessário avaliar (calcular) a constante normalizadora do denominador,

$$f(y) = \int f(y|\theta) f(\theta) d\theta.$$

Por exemplo, suponha que  $Y_1, \dots, Y_n$  são variáveis aleatórias independentes com distribuição Poisson  $Y|\theta \sim P(\theta)$  e a opinião prévia sobre  $\theta$  é a de que seus valores *definitivamente* estão no intervalo  $[0,1]$ , mas que todos os valores neste intervalo são igualmente prováveis; então,  $f(\theta) = 1; 0 \leq \theta \leq 1$ . Neste caso a constante normalizadora é

$$\int_0^1 e^{-n\theta} \theta^{\sum y_i} d\theta,$$

e esta integral, que corresponde a uma função gama incompleta, não pode ser calculada analiticamente e, portanto, só pode ser avaliada numericamente.

Portanto, nota-se que mesmo escolhas simples de prioris como a priori uniforme desse exemplo, podem os levar a problemas numéricos intratáveis. Entretanto, foram vistos três exemplos no capítulo anterior nos quais a escolha criteriosa

da priori leva a cálculos da posteriori que não exigem nenhuma integração. Em cada um destes casos foi possível identificar a distribuição priori para a qual a distribuição posteriori seja da mesma família de distribuições que a priori; tais prioris são chamadas *prioris conjugadas*. Considera-se agora ainda um outro exemplo em que a priori conjugada pode ser encontrada.

**Exemplo 3.1** (*Amostragem Gama.*)  $Y_1, \dots, Y_n$  são variáveis aleatórias independentes com distribuição  $Y \sim \text{Ga}(k, \theta)$ , em que  $k$  é conhecido. (Note que o caso  $k = 1$  corresponde a distribuição exponencial).

Então a verossimilhança é dada por

$$L(\theta|y) \propto \theta^{nk} \exp\{-\theta \sum y_i\}.$$

Agora, estudando esta forma, tomada como uma função de  $\theta$ , nota-se o núcleo de uma distribuição gamma  $\text{Ga}(nk + 1, \sum y_i)$ , o que sugere que pode-se tomar uma priori da forma

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

isto é,  $\theta \sim \text{Ga}(p, q)$ , então pelo Teorema de Bayes

$$f(\theta|y) \propto \theta^{p+nk-1} \exp\{-(q + \sum y_i)\theta\},$$

a então  $\theta|y \sim \text{Ga}(p + nk, q + \sum y_i)$ .

### 3.2.1 Uso de prioris conjugadas

O uso de prioris conjugadas deve ser visto como o que realmente é: um instrumento matematicamente conveniente. Entretanto, a expressão de convicções e opiniões anteriores na forma de uma distribuição paramétrica sempre é uma aproximação. Em diversas situações, a família conjugada de prioris é ampla o bastante para que uma priori conjugada seja encontrada de forma a ser suficientemente próxima às convicções prévias para que este nível de aproximação seja aceitável. Entretanto, se este não for o caso, tais prioris não devem ser utilizadas apenas para facilitar a matemática.

### 3.2.2 Obtenção de prioris conjugadas

Desde de não estejam em conflito com as convicções prévias, e desde que tal família possa ser encontrada, a simplicidade induzida pelo uso de prioris conjugadas é muito atrativa. Isto leva à questões sobre em quais situações a família conjugada pode ser encontrada. Surge que o único caso em que conjugadas podem ser facilmente encontrados são para modelos na família exponencial. Isto é, a distribuição de  $Y$  é da forma:

$$f(y|\theta) = h(y)g(\theta) \exp\{t(y)c(\theta)\}$$

para funções  $h$ ,  $g$  e  $c$  tais que

$$\int f(y|\theta) dy = g(\theta) \int h(y) \exp\{t(y)c(\theta)\} dy = 1.$$

Isto pode parecer restritivo, mas na realidade inclui distribuições frequentemente utilizadas como a *exponencial*, a *Poisson*, a *Gama com um parâmetro*, a *binomial* e a *normal* (com variância conhecida).

Então, com uma priori  $f(\theta)$  e considerando-se uma amostra aleatória  $y_1, y_2, \dots, y_n$ ,

$$\begin{aligned} f(\theta|y) &\propto f(\theta) f(y|\theta) \\ &= f(\theta) \prod_{i=1}^n \{h(y_i)\} g^n(\theta) \exp\left\{\sum_{i=1}^n t(y_i)c(\theta)\right\} \\ &\propto f(\theta) g^n(\theta) \exp\left\{\sum_{i=1}^n t(y_i)c(\theta)\right\} \end{aligned}$$

Portanto escolhendo

$$f(\theta) \propto g^d(\theta) \exp\{bc(\theta)\}$$

obtém-se

$$\begin{aligned} f(\theta|y) &\propto g^{n+d}(\theta) \exp\{c(\theta)[\sum_{i=1}^n t(y_i) + b]\} \\ &= g^{\bar{d}}(\theta) \exp\{\bar{b}c(\theta)\}, \end{aligned}$$

o que leva a uma posteriori na mesma família da priori, porém com parâmetros modificados.

Pode-se facilmente verificar que todos os exemplos de prioris conjugadas vistos até aqui podem ser obtidos desta forma geral. Por exemplo, no caso binomial:

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

Na notação da família exponencial tem-se que  $h(y) = \binom{n}{y}$ ,  $g(\theta) = (1-\theta)^n$ ,  $t(y) = y$ , e  $c(\theta) = \log(\frac{\theta}{1-\theta})$ . A priori conjugada assume a forma:

$$\begin{aligned} f(\theta) &\propto [(1-\theta)^n]^d \exp\left\{b \log\left(\frac{\theta}{1-\theta}\right)\right\} \\ &= (1-\theta)^{nd-b} \theta^b \end{aligned}$$

que é simplesmente um membro da família beta de distribuições.

### 3.2.3 Análises conjugadas usuais.

Tabela 3.1: Análises conjugadas usuais. Para a geométrica e binomial negativa  $Y$  é definida como o número de “tentativas” até o  $r$ -ésimo sucesso.

Verossimilhança	Priori	Posteriori
$Y \sim B(n, \theta)$	$\theta \sim \text{Beta}(p, q)$	$\theta y \sim \text{Beta}(p + y, q + n - y)$
$Y \sim P(\theta)$	$\theta \sim \text{Ga}(p, q)$	$\theta y \sim \text{Ga}(p + \sum_{i=1}^n y_i, q + n)$
$Y \sim N(\theta, \tau^{-1}), (\tau \text{ conhecido})$	$\theta \sim N(b, c^{-1})$	$\theta y \sim N(\frac{cb + n\tau\bar{y}}{c + n\tau}, \frac{1}{c + n\tau})$
$Y \sim \text{Ga}(k, \theta), (k \text{ conhecido})$	$\theta \sim \text{Ga}(p, q)$	$\theta y \sim \text{Ga}(p + nk, q + \sum_{i=1}^n y_i)$
$Y \sim \text{Geo}(\theta)$	$\theta \sim \text{Beta}(p, q)$	$\theta y \sim \text{Beta}(p + n, q + \sum_{i=1}^n y_i - n)$
$Y \sim \text{BN}(r, \theta)$	$\theta \sim \text{Beta}(p, q)$	$\theta y \sim \text{Beta}(p + r, q + y - r)$

## 3.3 Prioris impróprias

Considere novamente a análise da posteriori obtida ao estimar a média da normal com variância conhecida e usando a priori normal. Neste caso  $Y_1, \dots, Y_n \sim N(\theta, \tau^{-1}), \theta \sim N(b, c^{-1})$ , levando a  $\theta|y \sim N(\frac{cb + n\tau\bar{y}}{c + n\tau}, \frac{1}{c + n\tau})$ . A força das opiniões prévias sobre  $\theta$  são determinadas pela variância ou, equivalentemente, pela precisão  $c$  da priori normal. Um valor grande de  $c$  corresponde a convicções prévias muito fortes, e por outro lado, pequenos valores de  $c$  refletem uma opinião prévia fraca de  $\theta$ . Suponha agora que o conhecimento prévio sobre  $\theta$  é tão pequeno que fazemos  $c \rightarrow 0$ . Neste caso bastante simples, a posteriori passa a ser  $\theta|y \sim N(\bar{y}, \frac{1}{n\tau})$ , ou, em uma notação mais familiar,  $\theta|y \sim N(\bar{y}, \frac{\sigma^2}{n})$ . Portanto, aparentemente se obtém uma distribuição posteriori perfeitamente válida através deste procedimento limite.

Mas há uma armadilha. Considere o que ocorre com a priori quando  $c \rightarrow 0$ . De fato, obtém-se uma priori  $\theta \sim N(0, \infty)$ , que não é genuinamente uma distribuição de probabilidades. Na verdade, quando  $c \rightarrow 0$ , a distribuição  $\theta \sim N(b, c^{-1})$  se torna cada vez mais plana (*flat*) tal que no limite

$$f(\theta) \propto 1; \theta \in \mathbb{R}.$$

Entretanto, esta não pode ser uma função de densidade de probabilidades válida uma vez que

$$\int_{\mathbb{R}} f(\theta) d\theta = \infty.$$

Desta forma, a posteriori  $\theta|y \sim N(\bar{y}, \frac{\sigma^2}{n})$ , obtida fazendo  $c \rightarrow 0$  na análise conjugada usual, não pode surgir através do uso de qualquer distribuição priori própria. Mas, de fato, surge se utiliza uma especificação de priori  $f(\theta) \propto 1$ , a qual é um exemplo do que é chamado uma distribuição priori *imprópria*.

Portanto, a questão que se coloca é: seria válido usar uma distribuição posteriori obtida pela especificação de uma priori imprópria para refletir conhecimento vago? Embora hajam algumas dificuldades adicionais envolvidas (ver a seguir), em

geral o uso de prioris impróprias é considerado aceitável. O ponto é que, se escolhermos  $c$  para ser qualquer outro valor positivo não nulo, uma priori perfeitamente própria seria obtida e não haveriam receios sobre as análises subsequentes. Desta forma, pode-se escolher  $c$  arbitrariamente próximo de zero (ou equivalentemente, uma variância à priori arbitrariamente grande) e obter uma posteriori arbitrariamente próxima a obtida utilizando  $f(\theta) \propto 1$ .

### 3.4 Representações de ignorância

Na sessão anterior foi visto que uma tentativa de representar “ignorância” na análise conjugada padrão da média de uma distribuição normal levou ao conceito de prioris impróprias. Mas há ainda problemas mais fundamentais. Se, digamos, for especificada uma priori da forma  $f(\theta) \propto 1$ , e se considerar uma reparametrização definida por  $\phi = \theta^2$ , então, pelo teorema de transformação de variáveis,

$$f(\phi) = f(\theta^2) \cdot \left| \frac{d\theta}{d\phi} \right| \propto \frac{1}{\sqrt{\phi}},$$

uma função não uniforme e decrescente em  $\phi$ . Por outro lado, sendo “ignorante” a respeito de  $\theta$ , certamente se é também “ignorante” à respeito de  $\phi$ , e desta forma deveria-se fazer igualmente a especificação  $f(\phi) \propto 1$ . Portanto, a ignorância a priori representada por uniformidade nas crenças a priori, não se propaga entre diferentes escalas.

Um particular ponto de vista é o de que a especificação de ignorância a priori *deve necessariamente* ser consistente entre transformações 1-1 dos parâmetros. Isto leva ao conceito de “priori de Jeffreys”, que se baseia no conceito da informação de Fisher:

$$I(\theta) = -E \left\{ \frac{d^2 \log f(y|\theta)}{d\theta^2} \right\} = E \left\{ \left( \frac{d \log f(y|\theta)}{d\theta} \right)^2 \right\}.$$

E a priori de Jeffreys é definida como sendo:

$$f_0(\theta) \propto |I(\theta)|^{1/2}$$

A consistência entre transformações é verificada no seguinte sentido. Suponha que  $\phi = g(\theta)$  seja uma transformação 1-1 de  $\theta$ . Pela regra de mudança de variável tem-se que:

$$\begin{aligned} f(\phi) &\propto f_0(\theta) \cdot \left| \frac{d\theta}{d\phi} \right| \\ &= I^{1/2}(\theta) \cdot \left| \frac{d\theta}{d\phi} \right| \end{aligned}$$

mas, por definição,  $I(\phi) = I(\theta) \cdot (d\theta/d\phi)^2$ , e então

$$f(\phi) \propto |I(\phi)|^{1/2}.$$

Desta forma, vê-se que a priori de Jeffreys para  $\theta$  se transforma naturalmente na priori de Jeffreys para  $\phi$ . Em outras palavras, é invariante sob reparametrização.

### 3.4.1 Exemplos

**Exemplo 3.2** (*Média da normal.*) Supondo que  $Y_1, \dots, Y_n$  são variáveis aleatórias independentes com distribuição  $N(\theta, \sigma^2)$ , com  $\sigma^2$  conhecido.

Então,

$$f(y|\theta) \propto \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2} \right\}$$

logo,

$$\log(f(y|\theta)) \propto -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2}$$

e,

$$\begin{aligned} I(\theta) &= -E \left\{ \frac{d^2 \log f(y|\theta)}{d\theta^2} \right\} \\ &= -E \left\{ -\frac{n}{\sigma^2} \right\} \\ &= \frac{n}{\sigma^2} \end{aligned}$$

Por isso,  $f_0(\theta) \propto 1$ . Note que trabalhou-se aqui com a verossimilhança completa, para todas as observações. Alternativamente, poderia-se ter trabalhado com a verossimilhança de uma única observação  $y_1$  obtendo  $I_1(\theta) = 1/\sigma^2$  e usar a propriedade de que, sob independência,  $I_n(\theta) = nI_1(\theta) = n/\sigma^2$ , onde  $I_1$  e  $I_n$  são as informações de 1 e  $n$  valores independentes de  $y$ , respectivamente. Assim, obtém-se (como esperado) a mesma priori de Jeffreys independentemente de quantas observações fazemos subsequentemente<sup>1</sup>.

**Exemplo 3.3** (*Amostra binomial.*) Supõe-se que  $Y|\theta \sim \text{Bin}(n, \theta)$ .

Então,

$$\log(f(y|\theta)) = y \log(\theta) + (n - y) \log(1 - \theta),$$

logo,

$$\frac{d^2 \log f(y|\theta)}{d\theta^2} = -\frac{y}{\theta^2} - \frac{(n - y)}{(1 - \theta)^2},$$

e como  $E(y) = n\theta$ ,

$$\begin{aligned} I(\theta) &= \frac{n\theta}{\theta^2} + \frac{(n - n\theta)}{(1 - \theta)^2} \\ &= n\theta^{-1}(1 - \theta)^{-1} \end{aligned}$$

o que leva a,

$$f_o(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2} = \theta^{1/2-1}(1 - \theta)^{1/2-1}$$

<sup>1</sup>Entretanto, diferentes experimentos podem levar a diferentes verossimilhanças e, consequentemente, a diferentes informações. Em tais situações, o uso da priori de Jeffreys viola o princípio de verossimilhança.



que é uma distribuição  $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ .

No gráfico da esquerda da Figura 3.1 são mostradas três priors para  $\theta$  discutidos até aqui: a priori  $\text{Beta}(4,4; 6,6)$  utilizada no Exemplo 2.2, a priori  $\text{Beta}(1; 1)$  que corresponde a uma distribuição uniforme no intervalo  $[0,1]$  e a priori de Jeffreys. À direita podem ser visualizadas as respectivas posteriores.

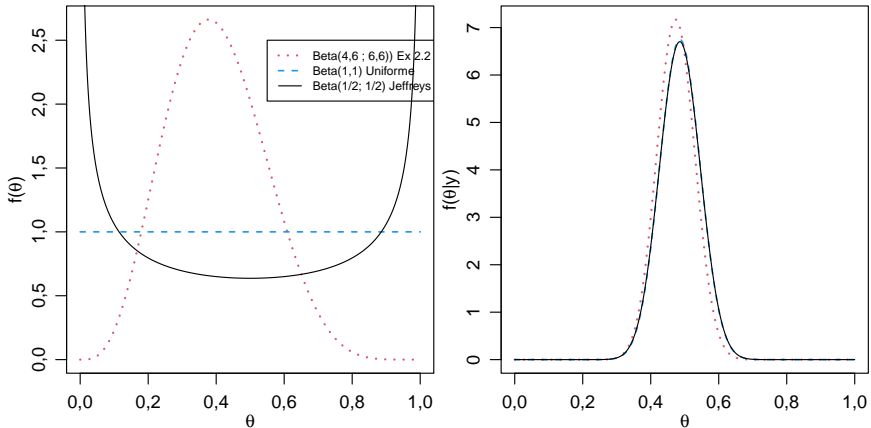


Figura 3.1: Diferentes priors Beta para a verossimilhança binomial (esquerda) e respectivas posteriores (direita).

### 3.5 Mistura de priors

Argumentos para o uso de famílias conjugadas de distribuições à priori foram delineados na sessão 3.2. Entretanto, enfatiza-se novamente, que tais famílias só devem ser utilizadas se um membro adequado da família pode ser encontrado estando de acordo com as opiniões prévias. Em algumas situações a família conjugada natural como, por exemplo, as listadas na tabela 3.1 pode ser por demais restritiva para que isto seja possível. Uma possibilidade para obter formas mais flexíveis para especificação de priors é o uso de uma priori dada por uma *mistura* das distribuições. Vamos considerar aqui apenas o caso de uma mistura discreta e com número finito de termos.

Considere-se o seguinte exemplo, frequentemente citado na literatura. Quando uma moeda é lançada, quase invariavelmente, há uma chance de 0,5 de sair “cara”. Entretanto, se ao invés de lançada, a moeda é girada sobre uma mesa, em geral tem-se que é mais provável que pequenas imperfeições na borda da moeda façam com que ela tenha uma tendência de “preferir” ou cara ou coroa. Levando isto em consideração, pode-se desejar atribuir à probabilidade que a moeda mostre a face “cara” uma distribuição que favoreça valores que sejam, por exemplo, 0,3 ou 0,7. Isto é, nossas opiniões prévias podem ser razoavelmente representadas por uma distribuição bimodal (ou mesmo trimodal se desejarmos atribuir peso extra a al-

gum outro ponto como, por exemplo, a possibilidade não tendenciosa de  $\theta = 0,5$ . O modelo de verossimilhança para o número de “caras” em  $n$  giros da moeda será binomial:  $Y|\theta \sim \text{Bin}(n, \theta)$  e, portanto, a priori conjugada está na família beta. Entretanto, não há nenhum membro desta família para o qual a distribuição seja multimodal. Uma possível solução é usar uma mistura de distribuições conjugadas. Esta família estendida será também uma distribuição conjugada pela seguinte razão. Suponha  $f_1(\theta), \dots, f_n(\theta)$  são todas distribuições conjugadas para  $\theta$ , levando a posterioris  $f_1(\theta|y), \dots, f_n(\theta|y)$ . Considere a família de mistura de distribuições:

$$f(\theta) = \sum_{i=1}^k p_i f_i(\theta).$$

Então,

$$\begin{aligned} f(\theta|y) &\propto f(\theta) f(y|\theta) \\ &= \sum_{i=1}^k p_i f_i(\theta) f(y|\theta) \\ &\propto \sum_{i=1}^k p_i^* f_i^*(\theta|y) \end{aligned}$$

e portanto a posteriori pertence à mesma família de misturas que a priori. Note, entretanto, que as proporções  $p_i^*$  de mistura na posteriori serão, em geral, diferentes da priori.

A obtenção dos pesos  $p_i^*$  é obtida a partir dos pesos das prioris e irá depender dos termos constantes da priori e posteriori de cada termo da mistura, como se justifica a seguir.

Sejam as distribuições da priori e posteriori denotadas por  $f_i(\theta) = C_i K_i(\theta)$  e  $f_i(\theta|y) = C_i^* K_i(\theta|y)$ , separando portanto os termos independentes do parâmetro do restante das expressões das distribuições.

$$\begin{aligned} f(\theta|y) &\propto \sum_{i=1}^k p_i f_i(\theta) f(y|\theta) \\ &\propto \sum_{i=1}^k p_i C_i K_i(\theta) L(\theta|y) \\ &\propto \sum_{i=1}^k p_i C_i K_i(\theta|y) \\ &\propto \sum_{i=1}^k p_i \frac{C_i}{C_i^*} f_i(\theta|y) \\ &\propto \sum_{i=1}^k w_i f_i(\theta|y). \end{aligned}$$

Portanto, os pesos na posteriori são

$$p_i^* = \frac{w_i}{\sum_i w_i} \text{ em que } w_i = \frac{p_i C_i}{C_i^*}. \quad (3.1)$$

**Exemplo 3.4** (*Amostragem binomial.*) Suponha o modelo  $Y \sim \text{Bin}(n, \theta)$  e uma priori dada por uma mistura de distribuições  $\theta \sim 0,2\text{Beta}(6, 14) + 0,8\text{Beta}(15, 5)$ . De-sejam obter a posteriori.

Supomos que obtivemos de um conjunto de dados com  $n = 30$  e  $y = 12$ . Vamos ainda denotar a função beta por  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ . Temos então que:

$$\begin{aligned} f(\theta) &= 0,2 B^{-1}(6, 14) \theta^{6-1} (1-\theta)^{14-1} + 0,8 B^{-1}(15, 5) \theta^{15-1} (1-\theta)^{5-1}, \\ L(\theta) &\propto \theta^{12} (1-\theta)^{30-12}, \\ f(\theta|y) &\propto 0,2 B^{-1}(6, 14) \frac{B(18, 32)}{B(6, 14)} \theta^{6+12-1} (1-\theta)^{14+(30-12)-1} + \\ &\quad 0,8 B^{-1}(15, 5) \frac{B(27, 23)}{B(15, 5)} \theta^{15+12-1} (1-\theta)^{5+(30-12)-1} \\ f(\theta|y) &\propto 0,2 \frac{B(18, 32)}{B(6, 14)} f_1(\theta|y) + 0,8 \frac{B(27, 23)}{B(15, 5)} f_2(\theta|y), \end{aligned}$$

em que  $f_1(\theta|y) \sim \text{Beta}(18, 32)$  e  $f_2(\theta|y) \sim \text{Beta}(27, 23)$ . Portanto,

$$\theta|y \sim p_1^* \text{Beta}(18, 32) + p_2^* \text{Beta}(27, 23)$$

Em que cada um dos pesos na posteriori são calculados como  $p_i^* = w_i / \sum_i w_i$  com

$$w_1 = 0,2 \frac{B(18, 32)}{B(6, 14)} \text{ e } w_2 = 0,8 \frac{B(27, 23)}{B(15, 5)}.$$

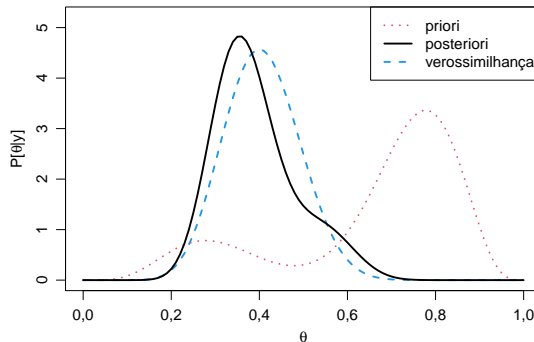


Figura 3.2: Verossimilhança e distribuições priori e posteriori para o Exemplo 3.4.

A Figura 3.2 mostra as funções de verossimilhança, priori e posteriori. Nesse exemplo os pesos calculados na posteriori são  $p_1^* = 0,819$  e  $p_2^* = 0,181$ .

**Exemplo 3.5** (*Amostra binomial com priori de mistura com três componentes.*) Supõe-se que  $Y|\theta \sim \text{Bin}(n, \theta)$  com priori sendo uma mistura de  $\text{Beta}(1,5; 3,5)$ ,  $\text{Beta}(2,5; 2,5)$  e  $\text{Beta}(3,5; 1,5)$  com pesos iguais de cada distribuição. Uma amostra forneceu  $y = 8$  e  $n = 20$ .

Então temos a priori ilustrada na Figura 3.3 dada por

$$f(\theta) = \sum_{i=1}^k p_i f_i(\theta)$$

em que  $k = 3$  e  $p_1 = p_2 = p_3 = 1/3$ , e portanto

$$f(\theta) = \frac{1}{3} \text{Beta}(1,5; 3,5) + \frac{1}{3} \text{Beta}(2,5; 2,5) + \frac{1}{3} \text{Beta}(3,5; 1,5).$$

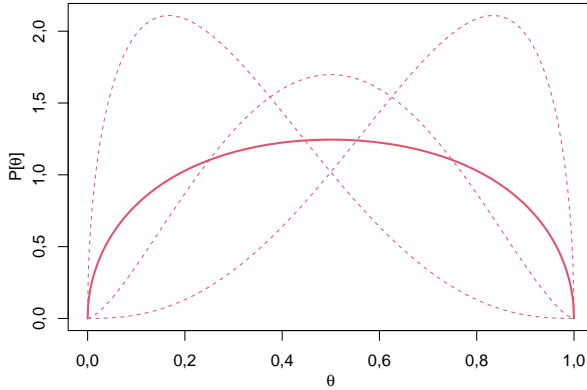


Figura 3.3: Priori como mistura de distribuição beta. Componentes da mistura (linhas tracejadas) e a distribuição resultante de mistura (linha sólida).

Para obtenção da posteriori cada uma das distribuições  $\text{Beta}(a_i, b_i)$  da priori possui a sua posteriori conjugada  $\text{Beta}(a_i^*, b_i^*)$  e a posteriori é da forma:

$$\begin{aligned} f(\theta|y) &= \sum_{i=1}^k p_i f_i(\theta) f(y|\theta) \\ &= p_1^* \theta^{1,5-1} (1-\theta)^{3,5-1} \theta^8 (1-\theta)^{12} + p_2^* \theta^{2,5-1} (1-\theta)^{2,5-1} \theta^8 (1-\theta)^{12} + \\ &\quad + p_3^* \theta^{3,5-1} (1-\theta)^{1,5-1} \theta^8 (1-\theta)^{12}. \end{aligned}$$

Os termos constantes (em relação a  $\theta$ ) são inversas das funções Beta:

$$C_i = B^{-1}(a_i^*, b_i^*) = \frac{\Gamma(a_i^* + b_i^*)}{\Gamma(a_i^*) \Gamma(b_i^*)}$$

e os pesos são:

$$p_1^* = 0,384 \quad p_2^* = 0,42 \quad p_3^* = 0,196.$$

Portanto a posteriori vista na Figura 3.4 fica da forma:

$$f(\theta|y) = \sum_{i=1}^k p_i f_i(\theta) f(y|\theta)$$

$$= 0,384 \text{ Beta}(9,5; 15,5) + 0,42 \text{ Beta}(10,5; 14,5) + 0,196 \text{ Beta}(11,5; 13,5).$$

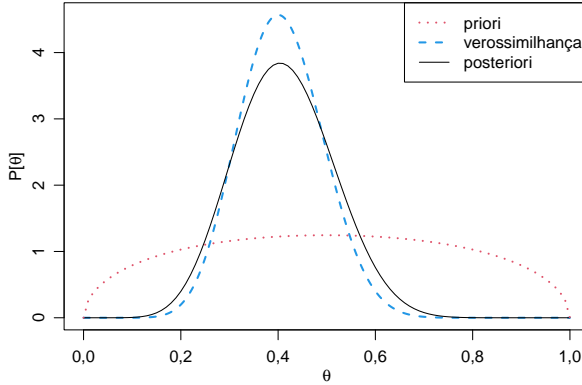


Figura 3.4: Priori, verossimilhança e posteriori para o Exemplo 3.5.

Pode ser demonstrado que uma mistura finita de prioris conjugadas pode ser definida para ser arbitrariamente próxima a *qualquer* distribuição. Entretanto, o número de termos necessários na mistura pode ser grande, e pode ser que seja possível representar as convicções prévias de forma muito mais sucinta usando outras famílias não conjugadas de prioris.

### 3.6 Elicitação de prioris

Não há respostas diretas sobre como melhor elicitar a informação prévia, mas o ponto que deve ser mencionado aqui é que isto é algo importante. Deve-se lembrar que não há como fazer uma quantidade infinita (ou mesmo muitas) de avaliações e atribuições de probabilidades. A especificação da distribuição a priori deve ser vista como uma tentativa de reconciliar as convicções que o analista tem de uma forma unificada. Como visto, uma possível abordagem é tomar uma particular família de distribuições (digamos, a família conjugada), solicitar a informação prévia sobre aspectos que resumam esta distribuição, tais como a média e variância à priori, e então, escolher como uma priori o membro da família conjugada que possua tais valores em particular. No entanto, de forma geral, pode ser extremamente difícil conciliar as opiniões prévias de um especialista (ou mesmo de vários especialistas) na forma de uma distribuição à priori.

## 3.7 Códigos computacionais

A Figura 3.1 é produzida com o código a seguir.

```
par(mfrow = c(1, 2), mar = c(3, 3, 0.1, 0.1), mgp = c(1.7, 0.8, 0))
curve(dbeta(x, 4.4, 6.6), from = 0, to = 1, n = 1001, col = 2, lty = 3, lwd = 2,
      xlab = expression(theta), ylab = expression(paste("f(", theta, ")")))
curve(dbeta(x, 1, 1), from = 0, to = 1, add = TRUE, col = 4, n = 1001, lty = 2,
      lwd = 1.5)
curve(dbeta(x, 1/2, 1/2), from = 0, to = 1, add = TRUE, n = 1001)
legend(0.55, 2.5, c("Beta(4,6 ; 6,6)) Ex 2.2", "Beta(1,1) Uniforme", "Beta(1/2; 1/2) Jeffreys"),
      lwd = c(2, 1.5, 1), lty = c(3, 2, 1), col = c(2, 4, 1), cex = 0.7)

curve(dbeta(x, 34 + 4.4, (70 - 34) + 6.6), from = 0, to = 1, n = 1001, col = 2,
      lty = 3, lwd = 2, xlab = expression(theta), ylab = expression(paste("f(",
      theta, "|y)")))
curve(dbeta(x, 34 + 1, (70 - 34) + 1), from = 0, to = 1, add = TRUE, col = 4,
      n = 1001, lty = 2, lwd = 1.5)
curve(dbeta(x, 34 + 1/2, (70 - 34) + 1/2), from = 0, to = 1, add = TRUE, n = 1001)
```

Os códigos para o cálculo dos pesos e gráficos da mistura de prioris do Exemplo 3.4 são dados a seguir.

```
n <- 30
y <- 12
curve(dbeta(x, y + 1, (n - y) + 1), from = 0, to = 1, col = 4, lty = 2, lwd = 2,
      ylim = c(0, 5.2), xlab = expression(theta), ylab = expression(paste("P(",
      theta, "|y)")))
dbm2.f <- function(x, ws, th) ws[1] * dbeta(x, th[1, 1], th[1, 2]) + ws[2] *
      dbeta(x, th[2, 1], th[2, 2])
ps <- c(p1 = 0.2, p2 = 0.8)
curve(dbm2.f(x, ws = ps, th = rbind(c(6, 14), c(15, 5))), from = 0, to = 1,
      col = 2, lty = 3, add = T, lwd = 2)
Ci <- c(lbeta(6, 14), lbeta(15, 5))
Csi <- c(lbeta(y + 6, (n - y) + 14), lbeta(y + 15, (n - y) + 5))
wi <- ps * exp(Csi - Ci)
ppos <- wi/sum(wi)
curve(dbm2.f(x, ws = ppos, th = rbind(c(y + 6, (n - y) + 14), c(y + 15, (n -
      y) + 5))), from = 0, to = 1, lwd = 2, add = T)
legend("topright", c("priori", "posteriori", "verossimilhança"), lty = c(3,
      1, 2), col = c(2, 1, 4), lwd = 2)
```

## 3.8 Exercícios

**Exercício 3.1** Verifique cada uma das análises conjugadas dadas na tabela 3.1.

**Exercício 3.2** Encontre a priori de Jeffreys para  $\theta$  no modelo geométrico:

$$f(y|\theta) = (1 - \theta)^{y-1}\theta; \quad y = 1, 2, \dots$$

para uma amostra  $y_1, \dots, y_n$ .

**Exercício 3.3** *Suponha que  $Y$  tenha uma distribuição de Pareto  $\text{Pa}(a, \theta)$ , em que  $a$  é conhecido mas  $\theta$  é o parâmetro desconhecido. Desta forma,*

$$f(y|\theta) = \theta a^\theta y^{-\theta-1}; \quad (y > a).$$

*Encontre a priori de Jeffreys e a correspondente distribuição posteriori para  $\theta$ .*





## Capítulo 4

# Problemas com múltiplos parâmetros

Todos os exemplos apresentados até o momento envolvem apenas um parâmetro, tipicamente, a média ou a variância da população. Porém, a maioria dos problemas estatísticos envolvem modelos estatísticos que contém mais do que um parâmetro. Podem ocorrer casos em que apenas há interesse em um parâmetro, mas há ainda outros parâmetros cujos valores são desconhecidos.

O método de analisar problemas multiparamétricos em estatística Bayesiana é muito mais direto, pelo menos a princípio, do que os métodos correspondentes em estatística frequentista. De fato, não existe absolutamente nenhuma nova teoria além do que já foi visto até o momento. Tem-se agora um vetor  $\theta = (\theta_1, \dots, \theta_d)$  de parâmetros sobre os quais deseja-se fazer inferência. Especifica-se uma distribuição priori (multivariada)  $f(\theta)$  para  $\theta$  que, assim como no caso de um parâmetro, é combinada com a verossimilhança  $f(y|\theta)$  pelo Teorema de Bayes obtendo-se assim como antes:

$$f(\theta|y) = \frac{f(\theta)f(y|\theta)}{\int f(\theta)f(y|\theta)d\theta},$$

em que agora  $f(\theta)$  e  $f(\theta|y)$  são distribuições multivariadas com dimensão dada pelo número de parâmetros. É portanto claro que a distribuição posteriori também será uma distribuição multivariada. A simplicidade da abordagem bayesiana decorre do fato de que a inferência sobre qualquer subconjunto de parâmetros dentro de  $\theta$  é obtida por um resultado básico de probabilidades que permite obter uma distribuição marginal diretamente a partir da distribuição posteriori conjunta. Por exemplo, a distribuição posteriori marginal de  $\theta_1$  é obtida “integrando-se fora” os outros componentes de  $\theta$ . Assim,

$$f(\theta_1|y) = \int_{\theta_2} \dots \int_{\theta_d} f(\theta|y) d\theta_2 \dots d\theta_d.$$

Resultados usuais de probabilidades para variáveis multidimensionais são utilizados e, embora nenhuma nova teoria seja necessária, o incremento de dimensão dá origem a alguns problemas práticos:

**Especificação da priori.** As priors são agora distribuições multivariadas. Isso significa que a distribuição priori precisa refletir não somente o comportamento de cada parâmetro individualmente, mas também a dependência entre diferentes combinações dos parâmetros. Por exemplo, se considera-se que um parâmetro possui valor alto, é provável que o outro parâmetro tenha um valor correspondente baixo? Em outras palavras, é necessário avaliar como os parâmetros variam conjuntamente. Escolher adequadamente a família de distribuições para a priori e resumir desta forma as informações a priori de especialistas pode ser bem mais complicado em comparação com o caso univariado. Portanto especificação da priori implica em especificar a distribuição conjunta para o vetor de parâmetros  $\theta$ :

$$f(\theta) = f(\theta_1, \dots, \theta_d).$$

Uma alternativa é especificar priors para os termos da fatoração

$$f(\theta) = f(\theta_1, \dots, \theta_d) = f(\theta_1) \cdot f(\theta_2|\theta_1) \cdot f(\theta_d|\theta_1 \dots \theta_{d-1}).$$

Uma simplificação, muitas vezes adotada na prática, embora não sem consequências e portando deve haver cautela em adotá-la, é supor independência e teríamos

$$f(\theta) = f(\theta_1, \dots, \theta_d) = f(\theta_1) \cdot f(\theta_2) \cdot f(\theta_d).$$

Neste último caso pode-se considerar que cada  $\theta_i$  é um escalar ou, de forma mais geral, que possam ser vetores. No último caso a fatoração é feita entre grupos de parâmetros independentes.

**Computação.** Mesmo em problemas de apenas uma dimensão viu-se que o uso de famílias conjugadas simplifica substancialmente as análises usando o Teorema de Bayes. Em problemas multivariados as integrais envolvidas na obtenção da posteriori são mais difíceis de resolver. Isto torna o uso de família de priors conjugadas ainda mais valioso, quando puder ser adotado. De forma mais geral, surge a necessidade de métodos computacionais para obter inferências quando famílias conjugadas não são disponíveis ou adequadas.

**Interpretação.** Toda a inferência está baseada na distribuição posteriori, que terá tantas dimensões quanto for o tamanho do vetor  $\theta$ . A estrutura da distribuição posteriori pode ser altamente complexa, e isto pode exigir considerável habilidade e um bom uso de recursos gráficos do computador para identificar as relações mais importantes que ela contém.

Apesar destes aspectos práticos é importante reenfatizar que exatamente a mesma teoria que foi utilizada para problemas com um único parâmetro está sendo utilizada para problemas multiparâmetros. A estrutura bayesiana implica que todas as inferências decorrem de regras elementares de probabilidades.

## 4.1 Exemplos

**Exemplo 4.1** Suponha que o funcionamento de uma determinada máquina seja satisfatório ( $y = 1$ ) ou insatisfatório ( $y = 2$ ). A probabilidade do funcionamento da máquina ser satisfatório depende da temperatura do ambiente ( $\theta_1 = 0$ : frio;  $\theta_1 = 1$ : quente) e da umidade ( $\theta_2 = 0$ : seco;  $\theta_2 = 1$ : úmido). As probabilidades de  $y = 1$  são apresentadas na tabela 4.1. Além disto, a distribuição priori conjunta para  $(\theta_1, \theta_2)$  é dada na tabela 4.2.

Tabela 4.1: Probabilidades condicionais do funcionamento da máquina ser satisfatório.

$P[y = 1 \theta_1, \theta_2]$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0,6	0,8
$\theta_2 = 1$	0,7	0,6

Tabela 4.2: Priori das condições da sala.

$P[\theta_1, \theta_2]$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0,3	0,2
$\theta_2 = 1$	0,2	0,3

A distribuição posteriori conjunta pode ser calculada, conforme apresentado na tabela 4.3.

Tabela 4.3: Probabilidades à posteriori das condições da sala

		$\theta_1 = 0$	$\theta_1 = 1$
$\Pr(y = 1 \theta_1, \theta_2) \times \Pr(\theta_1, \theta_2)$	$\theta_2 = 0$	0,18	0,16
	$\theta_2 = 1$	0,14	0,18
$\Pr(\theta_1, \theta_2 y = 1)$	$\theta_2 = 0$	18/66	16/66
	$\theta_2 = 1$	14/66	18/66

Assim, somando-se ao longo das margens, obtemos as distribuições posteriores marginais:

$$\Pr(\theta_1 = 0) = 32/66 \quad , \quad \Pr(\theta_1 = 1) = 34/66$$

e

$$\Pr(\theta_2 = 0) = 34/66 \quad , \quad \Pr(\theta_2 = 1) = 32/66$$

**Exemplo 4.2** Seja  $Y_1 \sim P(\alpha\beta)$  e  $Y_2 \sim P((1 - \alpha)\beta)$  com  $Y_1$  e  $Y_2$  (condicionalmente) independentes, dados  $\alpha$  e  $\beta$ . Agora suponha que a informação priori para  $\alpha$  e  $\beta$  pode ser expressa como:  $\alpha \sim \text{Beta}(p, q)$  e  $\beta \sim \text{Ga}(p + q, 1)$  com  $\alpha$  e  $\beta$  independentes para os hiperparâmetros  $p$  e  $q$  especificados. Note que a especificação da priori independente é uma parte importante da especificação da priori.

Como

$$Y_1 \sim P(\alpha\beta) \Rightarrow f(y_1|\alpha\beta) = \frac{e^{-\alpha\beta}(\alpha\beta)^{y_1}}{y_1!}$$

e

$$Y_2 \sim P((1-\alpha)\beta) \Rightarrow f(y_2|\alpha, \beta) = \frac{e^{-(1-\alpha)\beta}((1-\alpha)\beta)^{y_2}}{y_2!},$$

e sendo  $Y_1$  e  $Y_2$  independentes, a verossimilhança é dada por:

$$f(y_1, y_2|\alpha, \beta) = \frac{e^{-\alpha\beta}(\alpha\beta)^{y_1}}{y_1!} \times \frac{e^{-(1-\alpha)\beta}((1-\alpha)\beta)^{y_2}}{y_2!}.$$

As prioris (marginais) escolhidas são:

$$\alpha \sim \text{Beta}(p, q) \Rightarrow f(\alpha) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \alpha^{p-1} (1-\alpha)^{q-1}$$

$$\beta \sim \text{Ga}(p+q, 1) \Rightarrow f(\beta) = \frac{1}{\Gamma(p+q)\beta^{p+q}} \beta^{p+q-1} e^{-\beta} = \frac{1}{\Gamma(p+q)} \beta^{p+q-1} e^{-\beta}$$

e como supõe-se independência na priori tem-se que,

$$\begin{aligned} f(\alpha, \beta) &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \alpha^{p-1} (1-\alpha)^{q-1} \times \frac{1}{\Gamma(p+q)} \beta^{p+q-1} e^{-\beta} \\ &= \frac{1}{\Gamma(p)\Gamma(q)} \alpha^{p-1} (1-\alpha)^{q-1} \beta^{p+q-1} e^{-\beta} \end{aligned}$$

Deste modo, pelo Teorema de Bayes, a posteriori é da seguinte forma:

$$\begin{aligned} f(\alpha, \beta|y_1, y_2) &\propto e^{-\alpha\beta-(1-\alpha)\beta-\beta} \alpha^{y_1} \beta^{y_1} (1-\alpha)^{y_2} \beta^{y_2} \alpha^{p-1} (1-\alpha)^{q-1} \beta^{p+q-1} \\ &= e^{-2\beta} \beta^{y_1+y_2+p+q-1} \alpha^{y_1+p-1} (1-\alpha)^{y_2+q-1} \end{aligned}$$

Esta é a distribuição posteriori conjunta para  $\alpha$  e  $\beta$  e contém toda a informação vinda da priori e dos dados. Segue-se que as distribuições posteriori marginais são obtidas por:

$$f(\alpha|y_1, y_2) = \int_0^\infty f(\alpha, \beta|y_1, y_2) d\beta \propto \alpha^{y_1+p-1} (1-\alpha)^{y_2+q-1},$$

e

$$f(\beta|y_1, y_2) = \int_0^1 f(\alpha, \beta|y_1, y_2) d\alpha \propto \beta^{y_1+y_2+p+q-1} e^{-2\beta}$$

Isto é,  $\alpha|y_1, y_2 \sim \text{Beta}(y_1+p, y_2+q)$  e  $\beta|y_1, y_2 \sim \text{Ga}(y_1+y_2+p+q, 2)$ .

**Exemplo 4.3** (*Média da distribuição normal com variância desconhecida.*) Seja  $Y_1, \dots, Y_n$  um conjunto de variáveis independentes com distribuição normal  $N(\theta, \phi)$ , em que ambos, a média e a variância,  $\theta$  e  $\phi$ , respectivamente, são desconhecidos.

Então,

$$f(y_1|\theta, \phi) = \frac{1}{\sqrt{2\pi\phi}} \exp \left\{ -\frac{(y_1 - \theta)^2}{2\phi} \right\},$$

e a verossimilhança é da forma:

$$\begin{aligned} l(\theta; y) &\propto \phi^{-n/2} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\phi} \right\} \\ &= \phi^{-n/2} \exp \left\{ -\frac{1}{2\phi} \left( \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \theta)^2 \right) \right\} \\ &= \phi^{-n/2} \exp \left\{ -\frac{1}{2\phi} (S^2 + n(\bar{y} - \theta)^2) \right\} \end{aligned}$$

em que  $S^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ .

É possível fazer uma análise conjugada completa neste modelo usando uma priori conjugada conhecida como a distribuição Normal-Qui-quadrado. Os detalhes algébricos são levemente tediosos. Considera-se aqui o caso mais simples no qual é adotada a priori de referência que corresponde a expressar a ignorância. Em particular, toma-se:

$$f(\theta, \phi) \propto \frac{1}{\phi} \quad -\infty < \theta < \infty; 0 < \phi.$$

Essa é uma priori imprópria, pois sua integral não converge. Com esta priori, pelo Teorema de Bayes, obtém-se:

$$f(\theta, \phi|y) \propto \phi^{-\frac{n}{2}-1} \exp \left\{ -\frac{1}{2\phi} (S^2 + n(\bar{y} - \theta)^2) \right\}.$$

Note que essa expressão não pode ser fatorada, então as distribuições posteriores marginais para  $\theta$  e  $\phi$  não são independentes. Elas podem ser calculadas como mostrado a seguir. Para simplificar a notação, seja  $A = S^2 + n(\bar{y} - \theta)^2$ . Então,

$$p(\theta|y) \propto \int_0^\infty \phi^{-\frac{n}{2}-1} \exp \left\{ -\frac{A}{2\phi} \right\} d\phi.$$

Fazendo a mudança de variável:  $u = A/(2\phi)$ , tem-se que:

$$\begin{aligned} f(\theta|y) &\propto \int_0^\infty \phi^{-\frac{n}{2}-1} e^{-u} \frac{2\phi^2}{A} du \\ &= \left( \frac{2}{A} \right)^{n/2} \int_0^\infty u^{n/2-1} e^{-u} du \\ &= \left( \frac{2}{A} \right)^{n/2} \Gamma(n/2) \end{aligned}$$

Desta,  $f(\theta|y) \propto A^{-n/2} = \{S^2 + n(\bar{y} - \theta)^2\}^{-n/2}$ . Essa não está na forma padrão, mas se considerarmos:

$$t = \frac{\theta - \bar{y}}{s/\sqrt{n}} \quad \text{em que,} \quad s^2 = \frac{S^2}{n-1},$$

então,

$$\begin{aligned} f(t|y) &\propto \{(n-1)s^2 + (st)^2\}^{-n/2} \\ &\propto \left\{1 + \frac{t^2}{n-1}\right\}^{-n/2} \end{aligned}$$

Essa é a densidade da distribuição t-Student com  $n-1$  graus de liberdade, ou seja,  $[t|y] \sim t_{n-1}$ .

Similarmente,

$$\begin{aligned} f(\phi|y) &\propto \int_{-\infty}^{\infty} \phi^{-\frac{n}{2}-1} \exp\left\{-\frac{1}{2\phi}(S^2 + n(\bar{y} - \theta)^2)\right\} d\theta \\ &\propto \phi^{-(n+1)/2} \exp\{-S^2/(2\phi)\} \int_{-\infty}^{\infty} (2\pi\phi/n)^{-1/2} \exp\left\{-\frac{n}{2\phi}(\theta - \bar{y})^2\right\} d\theta \\ &= \phi^{-(n+1)/2} \exp\{-S^2/(2\phi)\} \end{aligned}$$

E portanto,  $S^2/\phi \sim \chi_{n-1}^2$ .

Assim, em resumo, para um modelo Normal com média e variância desconhecidas, tendo a prior de referencia  $p(\theta, \phi) \propto 1/\phi$ , tem-se que:

$$t = \frac{\theta - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}$$

e

$$S^2/\phi \sim \chi_{n-1}^2.$$

## 4.2 Exercícios

**Exercício 4.1** A qualidade de um componente elétrico ou é excelente ( $y = 1$ ), bom ( $y = 2$ ) ou ruim ( $y = 3$ ). A probabilidade de vários níveis de qualidade dependem da fábrica onde é produzido ( $\theta_1 = 0$ : fábrica A,  $\theta_1 = 1$ : fábrica B) e do tipo de máquina ( $\theta_2 = 0$ : máquina I,  $\theta_2 = 1$ : máquina II,  $\theta_2 = 3$ : máquina III). As probabilidades de  $y = 3$  são dadas pela Tabela 4.4. Além disso, distribuição conjunta de  $(\theta_1, \theta_2)$  é dada na Tabela 4.5. Encontre a distribuição posteriori conjunta de  $\theta_1, \theta_2|y = 3$  e cada uma das distribuições marginais. Tendo observado  $y = 3$  qual a combinação fábrica/máquina é a mais provável de ter produzido esse componente?

Tabela 4.4: Probabilidade condicional  $y = 3$ 

$\Pr(y = 3 \theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0,2	0,3
$\theta_2 = 1$	0,4	0,1
$\theta_2 = 2$	0,5	0,2

Tabela 4.5: Probabilidade a priori das combinações fábrica/máquina

$\Pr(\theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0,1	0,2
$\theta_2 = 1$	0,2	0,3
$\theta_2 = 2$	0,1	0,1





## Capítulo 5

# Resumindo informação a posteriori

### 5.1 Teoria da decisão

Essa é uma área extremamente importante. Porém, neste material abordam-se apenas algumas questões principais.

Muitos problemas do mundo real podem ser encarados como problemas de decisão: “Eu deveria votar em um partido político ou em outro?”; “Deveria sair hoje à noite ou concluir um trabalho?”; “Deveria aceitar um novo emprego ou fico na esperança de que apareça uma outra oportunidade de trabalho melhor?”.

A inferência estatística também pode ser pensada como a tomada de decisão: após ter observado um determinado conjunto de dados, qual o valor representará o parâmetro estimado? Existem muitas abordagens para a teoria de decisão, mas, de longe, o mais coerente é uma abordagem baseada na análise bayesiana.

De fato, pode ser mostrado que, se certos axiomas são obedecidos (isto é, um número de regras de decisão de bom senso são adotados), a análise bayesiana é a abordagem lógica para tomada de decisão. Isso é muitas vezes usado como um argumento para justificar a preferência da inferência bayesiana em relação a inferência clássica.

Os elementos necessários para a construção de um problema de decisão são os seguintes:

- 1) Um espaço paramétrico  $\Theta$  que contenha os possíveis *estados na natureza*;
- 2) Um conjunto  $A$  de possíveis *ações* que estão disponíveis para o tomador da decisão;
- 3) Uma *função perda*  $L$ , em que  $L(\theta, a)$  é a perda decorrente de adotar a ação  $a$  quando o verdadeiro estado da natureza é  $\theta$ .

Estes termos são ilustrados a seguir no contexto de um exemplo específico:

**Exemplo 5.1** *Um oficial de saúde pública está buscando uma política de vacinação contra uma doença relativamente branda que faz com que os funcionários se ausentem do trabalho. Pesquisas sugerem que 60% dos indivíduos da população já estão imunes, mas serão realizados testes laboratoriais para detectar vulnerabilidade em alguns indivíduos, já que os processos para triagem em massa são muitos caros. Um teste simples da pele foi desenvolvido, mas não é totalmente confiável. As probabilidades de resposta ao teste estão resumidas na Tabela 5.1.*

Imune	Desprezível	Reação (y)		
		Leve	Moderada	Forte
Sim	0,35	0,30	0,21	0,14
Não	0,09	0,17	0,25	0,49

Tabela 5.1: Probabilidades de reação dado o status imunológico.

*Estima-se que o equivalente em dinheiro por horas-homem perdidas, por deixar de vacinar um indivíduo vulnerável é de vinte (20) unidades, e que o custo desnecessário de vacinar uma pessoa imune é oito (8) unidades, e que não há nenhum custo incorrido em vacinar uma pessoa vulnerável ou não vacinar uma pessoa imune.*

Então, nesse caso particular, temos:

- 1) O espaço paramétrico  $\Theta = \{\theta_1, \theta_2\}$ , em que  $\theta_1$  e  $\theta_2$  correspondem ao indivíduo sendo imune e vulnerável, respectivamente.
- 2) O conjunto de ações  $A = \{a_1, a_2\}$  em que  $a_1$  e  $a_2$  correspondem a vacinar e não vacinar, respectivamente.
- 3) A função de perda  $L(\theta, a)$  definida na Tabela 5.2.

$L(\theta, a)$	$\theta_1$	$\theta_2$
$a_1$	8	0
$a_2$	0	20

Tabela 5.2: Função perda

Agora, voltando para a configuração geral, depois de ter observado um dado  $y$ , e também ter informação prévia  $f(\theta)$ , o objetivo é tomar a decisão que minimiza a perda esperada. Em primeiro lugar, o teorema de Bayes possibilita o cálculo das probabilidades posteriores  $f(y|\theta)$ .

Então, para qualquer ação particular  $a$ , a *perda esperada à posteriori* é a perda média da distribuição posteriori em  $\theta$ :

$$\rho(a, y) = E[L(\theta, a)] = \int L(\theta, a) f(\theta|y) d\theta,$$

que no caso de  $\theta$  discreto fica

$$\rho(a, y) = E[L(\theta, a)] = \sum_i L(\theta_i, a) f(\theta_i|y),$$

$f(\theta)$	$f(y \theta)$	Reação ( $y$ )			
		Desprezível	Leve	Moderada	Forte
0,6	$\theta_1$	0,35	0,30	0,21	0,14
0,4	$\theta_2$	0,09	0,17	0,25	0,49
$f(\theta, y)$	$\theta_1$	0,210	0,180	0,126	0,084
	$\theta_2$	0,036	0,068	0,100	0,196
	$f(y)$	0,246	0,248	0,226	0,280
$f(\theta y)$	$\theta_1$	210/246	180/248	126/226	84/280
	$\theta_2$	36/246	68/248	100/226	196/280
$\rho(a, y)$	$a_1$	1680/246	1440/248	1008/226	672/280
	$a_2$	720/246	1360/248	2000/226	3920/280
	$d(y)$	$a_2$	$a_2$	$a_1$	$a_1$

Tabela 5.3: Tabulação dos cálculos da análise de decisão.

Por exemplo, para  $y = \text{Desprezível}$ :

$$\rho(a_1, y) = E[L(\theta, a_1)] = L(\theta_1, a_1)P(\theta_1|y) + L(\theta_2, a_1)P(\theta_2|y) = 8 \times \frac{210}{246} + 0 \times \frac{36}{246} = \frac{1680}{246}$$

e

$$\rho(a_2, y) = E[L(\theta, a_2)] = L(\theta_1, a_2)P(\theta_1|y) + L(\theta_2, a_2)P(\theta_2|y) = 0 \times \frac{210}{246} + 20 \times \frac{36}{246} = \frac{720}{246}.$$

Assim, tendo observado  $y$ , esta é a perda na qual espera-se incorrer tomando medidas  $a$ . Obviamente que prefere-se minimizar as perdas, de modo que a regra de decisão de Bayes  $d(y)$  é a ação que minimiza a perda esperada posteriori.

Pode-se dar um passo adiante e calcular o risco associado à certa política, por meio da incerteza nas observações  $y$ . Ou seja, define-se o risco de Bayes por:

$$BR(d) = E_y[\rho(d(y), y)]$$

sendo então  $BR(d) = \sum_i \rho(d(y_i), y_i)f(y_i)$  para variável  $Y$  discreta e  $BR(d) = \int \rho(d(y), y)f(y)dy$  para o caso contínuo. Todos os cálculos para o exemplo em questão são indicados na Tabela 5.3.

Então, em resumo, se uma reação desprezível ou leve é observada, a decisão de Bayes é a de não vacinar, enquanto que se uma reação moderada ou forte é observada, a decisão é para vacinar. O risco de Bayes associado a esta estratégia é

$$\begin{aligned} BR(d) &= \sum \rho(d(y), y)f(y) \\ &= \frac{720}{246} \times 0,246 + \frac{1360}{248} \times 0,248 + \frac{1008}{226} \times 0,226 + \frac{672}{280} \times 0,280 = 3,76 \end{aligned}$$

O valor do risco de Bayes pode ser comparado com o custo de estratégias alternativas. Por exemplo, se adotamos uma política de vacinação para todas as pessoas, então a perda esperada por indivíduo seria  $0,6 \times 8 = 4,8$ .

## 5.2 Estimação pontual

Destacou-se ao longo do texto que toda a distribuição posteriori é o resumo completo da inferência sobre um parâmetro  $\theta$ . Em essência, **a distribuição a posteriori é a inferência**. No entanto, para algumas aplicações, é desejável (ou necessário) resumir essa informação de alguma maneira. Em particular, por vezes desejamos a “melhor” estimativa do parâmetro desconhecido. Note-se aqui a distinção com estatística clássica em que as estimativas pontuais dos parâmetros são a consequência natural de uma inferência, e expressar a incerteza da estimativa a parte mais problemática.

Assim, no âmbito Bayesiano, como é que vamos reduzir a informação de uma distribuição posteriori para dar a “melhor” estimativa?

Na verdade, a resposta depende do que queremos dizer com “melhor”, e este por sua vez, é especificado por transformar o problema em um problema de decisão. Ou seja, especificamos uma função perda  $L(\theta, a)$  que mede a “perda” na estimativa de  $\theta$  por  $a$ .

Há uma grande variedade de funções naturais de perda que podemos usar. A escolha particular para qualquer problema específico vai depender do contexto. As mais usadas são:

- 1) Perda quadrática:  $L(\theta, a) = (\theta - a)^2$ ;
- 2) Perda erro absoluto:  $L(\theta, a) = |\theta - a|$ ;
- 3) Perda linear: para  $g, h > 0$

$$L(\theta, a) = \begin{cases} g(a - \theta) & \text{se } a > \theta \\ h(\theta - a) & \text{se } a < \theta \end{cases}$$

- 4) Perda 0 – 1:

$$L(\theta, a) = \begin{cases} 0 & \text{se } |a - \theta| \leq \epsilon \\ 1 & \text{se } |a - \theta| > \epsilon \end{cases}$$

Em cada um destes casos, minimizando a perda esperada à posteriori, obtém-se formas simples para a regra de decisão de Bayes, que é tomada como sendo a estimativa pontual de  $\theta$  para aquela particular escolha da função de perda.

### 5.2.1 Perda quadrática

Nesse caso, a perda esperada à posteriori é dada por:

$$\begin{aligned}
 \rho(a, y) &= \int L(\theta, a) f(\theta|y) d\theta \\
 &= \int (\theta - a)^2 f(\theta|y) d\theta \\
 &= \int (\theta - E(\theta|y) + E(\theta|y) - a)^2 f(\theta|y) d\theta \\
 &= \int (\theta - E(\theta|y))^2 f(\theta|y) d\theta + 2 \int (\theta - E(\theta|y))(E(\theta|y) - a) f(\theta|y) d\theta \\
 &\quad + \int (E(\theta|y) - a)^2 f(\theta|y) d\theta \\
 &= \text{Var}(\theta|y) + [E(\theta|y) - a]^2
 \end{aligned}$$

uma vez que a integral do segundo termo é zero e  $(E(\theta|y) - a)$  é uma constante. Desta forma, a perda é minimizada quando  $a = E(\theta|y)$ .

Assim, a decisão de Bayes é estimar  $\theta$  pela esperança posteriori e, nesse caso, a perda esperada mínima é a variância da posteriori.

### 5.2.2 Perda absoluta

Note que a minimização de  $E[|\theta - a| | Y]$  envolve uma função não diferenciável, em todo ponto, com respeito a  $a$ . Portanto, é necessário manipular a equação resultante de forma a podermos minimizar  $E[|\theta - a| | Y]$  usando técnicas de derivação. Sendo assim, temos:

$$\begin{aligned}
 E(L(\theta, a) | Y) &= E[|\theta - a| | Y] = \int_{-\infty}^{\infty} |\theta - a| f(\theta|Y) d\theta \\
 &= \int_{-\infty}^a (a - \theta) f(\theta|Y) d\theta + \int_a^{\infty} (\theta - a) f(\theta|Y) d\theta \\
 &= a \left[ \int_{-\infty}^a f(\theta|Y) d\theta - \int_a^{\infty} f(\theta|Y) d\theta \right] - \int_{-\infty}^a \theta f(\theta|Y) d\theta \\
 &\quad + \int_a^{\infty} \theta f(\theta|Y) d\theta \\
 &= a \left[ \int_{-\infty}^a f(\theta|Y) d\theta - \left( 1 - \int_{-\infty}^a f(\theta|Y) d\theta \right) \right] - \int_{-\infty}^a \theta f(\theta|Y) d\theta \\
 &\quad + \int_a^{\infty} \theta f(\theta|Y) d\theta + \int_{-\infty}^a \theta f(\theta|Y) d\theta - \int_{-\infty}^a \theta f(\theta|Y) d\theta \\
 &= a \left[ 2 \int_{-\infty}^a f(\theta|Y) d\theta - 1 \right] + E(\theta|Y) - 2 \int_{-\infty}^a \theta f(\theta|Y) d\theta.
 \end{aligned}$$

Como a função

$$a \left[ 2 \int_{-\infty}^a f(\theta|Y) d\theta - 1 \right] + E(\theta|Y) - 2 \int_{-\infty}^a \theta f(\theta|Y) d\theta$$

é diferenciável (em todo ponto) com respeito a  $a$ , segue:

$$\begin{aligned}\frac{\partial}{\partial a} E(L(\theta, a)|Y) &= \left[ a \left[ 2 \int_{-\infty}^a f(\theta|Y) d\theta - 1 \right] + E(\theta|Y) - 2 \int_{-\infty}^a \theta f(\theta|Y) d\theta \right]' \\ &= \left[ 2 \int_{-\infty}^a f(\theta|Y) d\theta - 1 \right] + a 2f(a|Y) - 2a f(a|Y) \\ &= 2 \int_{-\infty}^a f(\theta|Y) d\theta - 1.\end{aligned}$$

Daí, temos que a decisão ótima é  $a^* = \text{Mediana}(\theta|Y)$ .

### 5.2.3 Perda linear

Note que o caso de perda erro absoluto é um caso especial de perda linear com  $g = h = 1$  (i.e. função identidade). Prova-se a seguir o caso mais geral e que portanto também resultar na perda em erro absoluto como caso particular.

Se  $q$  denota o quantil  $\frac{h}{g+h}$  da distribuição posteriori, ou seja,

$$\frac{h}{g+h} = \int_{-\infty}^q f(\theta|y) d\theta;$$

e supõe-se (sem perda de generalidade) que  $a > q$ . Então

$$L(\theta, q) - L(\theta, a) = \begin{cases} \frac{g(q-a)}{(g+h)\theta - hq - ga} & \text{se } \theta \leq q \\ \frac{h(a-q)}{h(a-q)} & \text{se } q < \theta < a \\ & \text{se } a < \theta \end{cases}.$$

Mas, para  $q < \theta < a$ ,

$$(g+h)\theta - hq - ga < h(a-q)$$

de forma que,

$$L(\theta, q) - L(\theta, a) \leq \begin{cases} \frac{g(q-a)}{h(a-q)} & \text{se } \theta \leq q \\ & \text{se } q < \theta \end{cases}$$

Então,

$$E(L(\theta, q) - L(\theta, a)) \leq g(q-a) \frac{h}{g+h} + h(a-q) \left( 1 - \frac{h}{g+h} \right) = 0$$

Isto é,  $\rho(a, y) \leq \rho(q, y) \forall a$ , então a regra de decisão de Bayes neste caso é  $a = q$ , o quantil  $\frac{h}{g+h}$  da distribuição posteriori.

Note que, quando  $g = h = 1$ ,  $q$  é a mediana da distribuição. Portanto, para perda em erro absoluto, o estimador de Bayes é a mediana da distribuição posteriori.

### 5.2.4 Perda 0-1

Claramente, neste caso:

$$\rho(a, y) = P(|\theta - a| > \epsilon | y) = 1 - P(|\theta - a| \leq \epsilon | y).$$

Consequentemente, se definirmos um intervalo modal de comprimento  $2\epsilon$ , como o intervalo  $[y - \epsilon, y + \epsilon]$ , que tem maior probabilidade, então a estimativa de Bayes é até ponto médio do intervalo com maior probabilidade. Ao escolher  $\epsilon$  arbitrariamente pequeno, esse procedimento chegará na estimativa de Bayes posteriori com essa perda em particular.

### 5.2.5 Resumo

Em resumo, no âmbito bayesiano, uma estimativa pontual de um parâmetro é uma estatística de resumo da distribuição posteriori.

A metodologia de teoria da decisão leva a escolhas ótimas de estimativas pontuais ao definir a qualidade de um estimador através de uma função de perda. Em particular, as opções mais naturais de função de perda apresentadas levam à média, mediana e moda posteriori, respectivamente, como estimadores pontuais ideais.

## 5.3 Intervalos de credibilidade

A idéia de um intervalo de credibilidade é fornecer um análogo ao intervalo de confiança obtido em estatística frequentista. O raciocínio é que estimativas pontuais não dão nenhuma medida de precisão, por isso é preferível informar no resultado das análises um intervalo, dentro do qual é “provável” que o parâmetro se encontre.

Isto provoca problemas na estatística frequentista, uma vez os parâmetros não são considerados como aleatórios, de modo que não é possível dar um intervalo com a interpretação de que existe uma certa probabilidade de que o parâmetro pertença ao intervalo. Em vez disso, os intervalos de confiança devem ter a interpretação de que, se a amostragem for repetida inúmeras vezes, há uma probabilidade especificada de que o intervalo por hora obtido deverá conter o parâmetro, ou seja, é o intervalo que é aleatório e não o parâmetro.

Não existe tal dificuldade na abordagem bayesiana, pois os parâmetros são tratados como aleatórios. Assim, uma região  $C_\alpha(y)$  é uma *região de credibilidade*  $100(1 - \alpha)\%$  para  $\theta$  se

$$\int_{C_\alpha(y)} f(\theta | y) d\theta = 1 - \alpha.$$

Ou seja, existe uma probabilidade de  $1 - \alpha$ , com base na distribuição posteriori, que  $\theta$  esteja contido em  $C_\alpha(y)$ .

Alguns bayesianos argumentam que intervalos de credibilidade têm pouco valor, uma vez que é toda a distribuição posteriori que contém as informações para

inferência, e que, os intervalos de credibilidade só foram propostos a fim de fornecer algo comparável aos intervalos de confiança.

Uma dificuldade com o intervalo de credibilidade (que também ocorre com intervalos de confiança), é que eles não são unicamente definidos. Qualquer região com probabilidade  $1 - \alpha$  é um intervalo válido. Uma vez que deseja-se um intervalo que contenha apenas os valores mais plausíveis do parâmetro, é habitual impor uma restrição adicional, que a largura do intervalo seja tão pequena quanto possível. Isso equivale a um intervalo (ou região) da forma:

$$C_\alpha(y) = \{\theta : f(\theta|y) \geq \gamma\},$$

em que  $\gamma$  é escolhido para garantir que:

$$\int_{c_\alpha(y)} f(\theta|y) d\theta = 1 - \alpha.$$

Tais regiões são chamadas de regiões com “mais alta de densidade posteriori” – HPD (sigla em inglês para *highest posterior density*). Tipicamente, estes intervalos são encontrados numericamente, embora, para a maioria das distribuições posteriores univariadas padrão, haja valores tabulados para diversos  $\alpha$ . De passagem, note que que ocorre o usual “perde/ganha” ao se escolher um  $\alpha$  apropriado: um valor pequeno de  $\alpha$  resultará em um intervalo largo; enquanto que um valor grande de  $\alpha$  resultará em um intervalo para o qual o parâmetro tem uma baixa probabilidade de a ele pertencer.

**Exemplo 5.2** (*Média normal.*) Sejam  $Y_1, \dots, Y_n$  variáveis independentes de uma distribuição  $N(\theta, \sigma^2)$ , ( $\sigma^2$  conhecido) com uma priori para  $\theta$  da forma  $\theta \sim N(b, d^2)$ .

Com essas informações, obtém-se a posteriori:

$$\theta|y \sim N\left(\frac{\frac{b}{d^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\right)$$

Agora, como a distribuição Normal é unimodal e simétrica, a região HPD de  $100(1 - \alpha)\%$  para  $\theta$  é:

$$\left(\frac{\frac{b}{d^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\right) \pm z_{\alpha/2} \left(\frac{1}{\frac{1}{d^2} + \frac{n}{\sigma^2}}\right)^{\frac{1}{2}}$$

em que  $z_{\alpha/2}$  é o ponto com porcentagem desejada da distribuição normal padrão  $N(0, 1)$ .

Note, além do mais que à medida que  $d \rightarrow \infty$  o intervalo se torna:

$$\bar{y} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

que é precisamente o intervalo de confiança para  $100(1 - \alpha)\%$  de  $\theta$  obtido em inferência clássica. Neste caso especial, o intervalo de credibilidade e o intervalo de confiança são idênticos, embora as suas interpretações sejam bastante diferentes.



**Exemplo 5.3** Sejam  $Y_1, \dots, Y_n$  variáveis independentes com distribuição  $N(\theta, \phi)$  ( $\phi$  desconhecido) e assumimos que a existe uma priori da forma:

$$f(\theta, \phi) \propto \frac{1}{\phi}; \quad -\infty < \theta < \infty, 0 < \phi.$$

Isso leva a distribuições posteriores marginais:

$$t = \frac{\theta - \bar{y}}{s/\sqrt{n}} \sim t_{n-1}$$

e

$$S^2/\phi \sim \chi_{n-1}^2.$$

Então, devido à simetria da distribuição t-Student, o intervalo de credibilidade  $100(1 - \alpha)\%$  para  $\theta$  é:

$$\bar{y} \pm t_{n-1, \alpha/2} \frac{S}{\sqrt{n}}$$

em que  $t_{n-1, \alpha/2}$  é o ponto com porcentagem desejada na distribuição  $t_{n-1}$ .

Por outro lado, o intervalo de credibilidade para  $\phi$  é ligeiramente mais problemático. Como  $S^2/\phi \sim \chi_{n-1}^2$ , segue-se que  $\phi/S^2 \sim \chi_{n-1}^{-2}$ , a chamada distribuição de qui-quadrado inversa. Pontos críticos de intervalos de maior densidade posteriori para uma variedade de valores de  $\alpha$  estão disponíveis em tabelas, podendo ser obtidos numericamente,

**Exemplo 5.4** Suponha  $Y \sim \text{Bin}(n, \theta)$  com priori  $\theta \sim \text{Beta}(p, q)$ .

Assim, temos a seguinte distribuição posteriori

$$\theta|y \sim \text{Beta}(p + y, q + n - y).$$

O intervalo  $[a, b]$  com  $100(1 - \alpha)\%$  de credibilidade e a maior densidade à posteriori, satisfaz a:

$$\frac{1}{B(p + y, q + n - y)} \int_a^b \theta^{p+y-1} (1 - \theta)^{q+n-y-1} d\theta = 1 - \alpha,$$

e

$$\begin{aligned} & \frac{1}{B(p + y, q + n - y)} a^{p+y-1} (1 - a)^{q+n-y-1} = \\ & = \frac{1}{B(p + y, q + n - y)} b^{p+y-1} (1 - b)^{q+n-y-1} = \gamma \end{aligned}$$

De forma geral a solução é obtida numericamente. No caso especial  $y = 0$  obtém-se solução analítica.

Sejam os valores  $p = 2$ ,  $q = 2$ ,  $n = 12$  e  $y = 4$ . Neste caso  $\theta|y \sim \text{Beta}(6, 10)$  e o intervalo de credibilidade de 95% obtido numericamente é  $[0,154, 0,605]$ , ou seja,  $P_{\theta|y}[0,154 < \theta < 0,605] = 0,95$ .

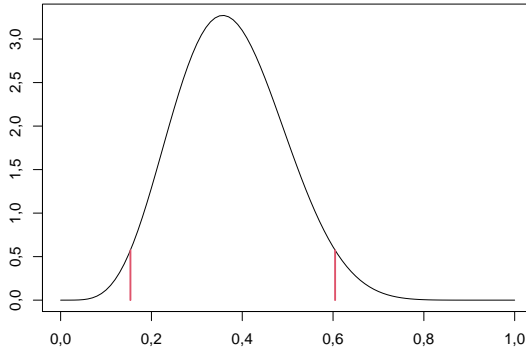


Figura 5.1: Distribuição posteriori e intervalo HPD de 95% de credibilidade para o Exemplo 5.4.

## 5.4 Teste de hipóteses

Testes de hipóteses são decisões na forma de escolha entre duas hipóteses.  $H_0 : \theta \in \Omega_0$  ou  $H_1 : \theta \in \Omega_1$ . Vamos considerar o caso simples, em que  $\Omega_1$  e  $\Omega_2$  são pontos individuais, de modo que o teste é da forma  $H_0 : \theta = \theta_0$  contra  $H_1 : \theta = \theta_1$ . A abordagem clássica para este problema é, geralmente, baseada no *teste de razão da verossimilhança*:

$$\lambda = \frac{f(y|\theta_1)}{f(y|\theta_0)}.$$

Grandes valores de  $\lambda$  indicam que é mais provável que os dados observados  $y$  tenham ocorrido se  $\theta_1$  é o verdadeiro valor de  $\theta$  do que teria com  $\theta_0$ . Na abordagem bayesiana, testes de hipóteses são estruturados em função de condições de preferência que envolvem probabilidades/densidades a posteriori e uma regra de decisão.

### 5.4.1 Testando hipóteses via regra de decisão

Considere um modelo estatístico  $f(y|\theta)$  com  $\theta \in \Theta$ . Seja  $\Theta_0$  e  $\Theta_1$  uma partição de  $\Theta$ , ou seja,  $\Theta_0 \cup \Theta_1 = \Theta$  e  $\Theta_0 \cap \Theta_1 = \emptyset$ .

Agora, considere que estamos interessados em testar

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

Defina a seguinte estrutura de perda

$L(\theta, a)$	Decisão	
	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$\theta \in \Theta_0$	0	$\alpha$
$\theta \in \Theta_1$	$\beta$	0

sendo  $\alpha, \beta > 0$ . Sob essa formulação, qual o estimador de Bayes associado?

Quando nos referimos ao estimador de Bayes, o objetivo é encontrar a decisão  $a$  que minimiza  $E[L(\theta, a)|Y]$ . Logo, temos

$$\begin{aligned} E[L(\theta, a)|Y] &= \int_{\theta \in \Theta} L(\theta, a) f(\theta|Y) d\theta = \int_{\theta \in \Theta_0} L(\theta, a) f(\theta|Y) d\theta + \int_{\theta \in \Theta_1} L(\theta, a) f(\theta|Y) d\theta \\ &= \alpha I_{\{\theta \in \Theta_1\}}(a) \int_{\theta \in \Theta_0} f(\theta|Y) d\theta + \beta I_{\{\theta \in \Theta_0\}}(a) \int_{\theta \in \Theta_1} f(\theta|Y) d\theta. \end{aligned}$$

Portanto

$$E[L(\theta, a)|Y] = \begin{cases} \alpha P(\theta \in \Theta_0|Y), & \text{se decidimos por } H_1, \\ \beta P(\theta \in \Theta_1|Y), & \text{se decidimos por } H_0. \end{cases}$$

Assim,  $H_0$  é preferível a  $H_1$  quando  $\beta P(\theta \in \Theta_1|Y) < \alpha P(\theta \in \Theta_0|Y)$ . Daí, segue que

$$\begin{aligned} \beta P(\theta \in \Theta_1|Y) &< \alpha P(\theta \in \Theta_0|Y) \quad (\Leftrightarrow) \\ \beta(1 - P(\theta \in \Theta_0|Y)) &< \alpha P(\theta \in \Theta_0|Y) \quad (\Leftrightarrow) \\ \beta &< \alpha P(\theta \in \Theta_0|Y) + \beta P(\theta \in \Theta_0|Y) \quad (\Leftrightarrow) \\ P(\theta \in \Theta_0|Y) &> \frac{\beta}{\alpha + \beta}. \end{aligned}$$

Conclui-se assim que  $H_0$  é preferível a  $H_1$  quando  $P(\theta \in \Theta_0|Y) > \frac{\beta}{\alpha + \beta}$ .

**Exemplo 5.5** Suponha  $Y \sim \text{Bin}(n, \theta)$  com priori  $\theta \sim \text{Beta}(2, 2)$ . Deseja-se testar  $H_0 : \theta \leq 0,35$  versus  $H_1 : \theta > 0,35$ . Suponha ainda a função de perda como na tabela ?? com  $\alpha = 2$  e  $\beta = 3$ . Qual seria a decisão para  $n = 12$  e  $y = 4$ ?

Neste caso, como  $\theta|Y \sim \text{Beta}(6, 10)$  prefere-se  $H_1$  pois

$$P(\theta \in \Theta_0|Y) = P_{\theta|Y}(\theta \leq 0,35) = 0,44 < \frac{\beta}{\alpha + \beta} = \frac{3}{2 + 3} = 0,60.$$

Um questionamento sobre o procedimento descrito é que  $\alpha$  e  $\beta$  são valores dados pelo pesquisador. Logo, devido ao grau de subjetividade, pode ser interessante trabalhar num contexto onde a tomada de decisão não dependa dessa atribuição.

### 5.4.2 Teste de Hipóteses via Fator de Bayes

Consideremos, novamente, o teste de hipóteses  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_1$ . O fator de Bayes (FB) é definido como a razão entre as probabilidades a posteriori de  $H_0$  sobre  $H_1$  e as probabilidades a priori de  $H_0$  e  $H_1$ , ou seja

$$\lambda_B = \frac{P(H_0|Y)/P(H_1|Y)}{P(H_0)/P(H_1)}. \quad (5.1)$$

No caso especial onde  $H_0$  e  $H_1$  correspondem a hipóteses simples, o teste baseado em  $\lambda_B$  coincide com o teste da razão de máxima verossimilhança (a demonstração fica como exercício).

A definição do Fator de Bayes sugere que, se  $\lambda_B > 1$ , então  $H_0$  é preferível. Em particular, Jeffreys (1961) construiu a seguinte escala de evidência empírica, independente de uma regra de decisão, em favor de  $H_0$ :

- Se  $\log_{10}(\lambda_B) \in [0; 0,5)$ , a evidência em favor de  $H_0$  é fraca;
- Se  $\log_{10}(\lambda_B) \in [0,5; 1)$ , a evidência em favor de  $H_0$  é razoável;
- Se  $\log_{10}(\lambda_B) \in [1; 2)$ , a evidência em favor de  $H_0$  é forte;
- Se  $\log_{10}(\lambda_B) \in [2; +\infty)$ , a evidência em favor de  $H_0$  é conclusiva.

Logo, o procedimento de testar hipóteses se limita a calcular  $\lambda_B$  e classificar a tomada de decisão com base nessa relação empírica.

O cenário onde  $\theta$  é contínuo e tem-se o interesse em testar  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$  sugere manipular o modelo a priori de forma a garantir que  $P(H_0|Y) > 0$ . Não entraremos em detalhes metodológicos sobre essa abordagem, mas consideramos ser importante que todos tenham ciência dessa particularidade envolvendo o fator de Bayes.

Neste material não vamos além disto em testes de hipóteses embora hajam outras propostas e procedimentos descritos na literatura. Entre eles destacamos aqui o procedimento chamado de *FBST – Full Bayesian Significance Test* proposto em Pereira & Stern (1999) e também descrito e discutido em Pereira et al. (2008) que procura oferecer um análogo bayesiano aos procedimentos usuais baseados em p-valores.

## 5.5 Códigos computacionais

Os cálculos do Exemplo 5.1 podem ser expressos por operações matriciais conforme a seguir.

```
(th <- c(th1 = 0.6, th2 = 0.4))

## th1 th2
## 0,6 0,4

y.th1 <- c(y1 = 0.35, y2 = 0.3, y3 = 0.21, y4 = 0.14)
y.th2 <- c(y1 = 0.09, y2 = 0.17, y3 = 0.25, y4 = 0.49)
(y.th <- rbind(y.th1, y.th2))

##          y1  y2  y3  y4
## y.th1 0,35 0,30 0,21 0,14
## y.th2 0,09 0,17 0,25 0,49

(yth <- th * y.th)

##          y1  y2  y3  y4
## y.th1 0,210 0,180 0,126 0,084
## y.th2 0,036 0,068 0,100 0,196
```

```

rownames(yth) <- c("yth1", "yth2")
yth

##          y1      y2      y3      y4
## yth1 0,210 0,180 0,126 0,084
## yth2 0,036 0,068 0,100 0,196

(y <- drop(crossprod(th, y.th)))

##          y1      y2      y3      y4
## 0,246 0,248 0,226 0,280

colSums(th * y.th)

##          y1      y2      y3      y4
## 0,246 0,248 0,226 0,280

(th.y <- t(t(yth)/drop(y)))

##          y1          y2          y3 y4
## yth1 0,8536585 0,7258065 0,5575221 0,3
## yth2 0,1463415 0,2741935 0,4424779 0,7

rownames(th.y) <- c("th1.y", "th2.y")
th.y

##          y1          y2          y3 y4
## th1.y 0,8536585 0,7258065 0,5575221 0,3
## th2.y 0,1463415 0,2741935 0,4424779 0,7

L <- diag(c(8, 20))
rownames(L) <- paste("a", 1:2, sep = "")
L

##      [,1] [,2]
## a1      8    0
## a2      0   20

(L.th.y <- L %*% th.y)

##          y1          y2          y3 y4
## a1 6,829268 5,806452 4,460177 2,4
## a2 2,926829 5,483871 8,849558 14,0

D.f <- function(x) ifelse(x[1] < x[2], "a1:Vacina", "a2:Não Vacina")
apply(L.th.y, 2, D.f)

##          y1          y2          y3          y4
## "a2:Não Vacina" "a2:Não Vacina" "a1:Vacina" "a1:Vacina"

apply(L.th.y, 2, min)

##          y1          y2          y3          y4
## 2,926829 5,483871 4,460177 2,400000

sum(apply(L.th.y, 1, min) * y)

## [1] 2,678166

```

Pode-se ainda comparar a estratégia anterior com outras estratégias. Por exemplo, decidindo adotar a estratégia *vacina todo mundo* a função perda fica agora definida pela matrix

$$L = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix}$$

e o risco de Bayes pode ser calculado.

```
LT <- diag(c(8, 0))
(LT.th.y <- LT %*% th.y)
sum(apply(LT.th.y, 2, sum) * y)
sum(th * c(8, 0))
```

Por outro lado se estratégia fosse *não vacinar ninguém* a função perda fica agora definida pela matrix

$$L = \begin{pmatrix} 0 & 0 \\ 0 & 20 \end{pmatrix}$$

e o risco de Bayes pode ser avaliado.

```
LN <- diag(c(0, 20))
(LN.th.y <- LN %*% th.y)
sum(apply(LN.th.y, 2, sum) * y)
sum(th * c(0, 20))
```

Portanto, dentre as três *estratégias consideradas*, (i) testar e vacinar quem tem reação moderada e forte, (ii) vacinar todos, (iii) não vacinar ninguém; a de menor risco de Bayes é a primeira, para as condições informadas.

## 5.6 Exercícios

**Exercício 5.1** Considere o Exemplo 5.1 e calcule o risco de Bayes para as seguintes estratégias:

- Vacinar somente se a reação for Forte.
- Não vacinar somente se a reação for Desprezível.

**Exercício 5.2** A confiabilidade dos julgamentos de dois especialistas em arte, A e B, foram testadas de forma que cada um deles, separadamente, julgou como “genuína” ou “falsificada” cada uma de um grande número de obras de arte de origem conhecida. Os testes apontaram que A tem probabilidade 0,8 de detectar uma falsificação e probabilidade 0,7 de reconhecer um objeto genuíno; B tem uma probabilidade maior, 0,9, de detectar uma falsificação, mas infelizmente classifica um objeto genuíno como falso com probabilidade de apenas 0,4. Um objeto de arte é ofertado por um valor que é uma verdadeira barganha, de apenas US\$ 100. Entretanto, se não for genuíno, então ele não vale nada. Se for genuíno, acredita-se que pode ser revendido imediatamente por US\$ 300. Acredita-se que há chance de 0,5 de que o objeto seja genuíno. Os especialistas A e B cobram US\$ 30 e US\$ 40, respectivamente, por seus serviços. Há vantagem em pagar qualquer um deles para uma avaliação?

**Exercício 5.3** Para o problema da média da Normal, descrito no Exemplo 2.4, encontre a estimativa pontual de  $\theta$  usando cada uma das quatro funções perda descritas nesse capítulo.

**Exercício 5.4** Repita o exercício anterior para a análise conjugada da distribuição Binomial com parâmetro  $\theta$  desconhecido.

**Exercício 5.5** Uma amostra de 6 alturas de colheita são registrados como:

5,3; 5,6; 5,9; 6,1; 6,2; 6,5.

Assumindo um modelo Normal com uma priori “não informativa” (vaga), calcular a região HDR de credibilidade 90% para a média populacional, com:

- a) a variância populacional igual a 1;
- b) a variância populacional desconhecida.

**Exercício 5.6** Suponha uma função perda com as seguintes características:

- i) igual a distância entre  $\theta$  e  $a$ , se  $a$  é menor do que  $\theta$ , e
  - ii) o triplo da distância entre  $\theta$  e  $a$ , se  $a$  é maior do que  $\theta$ .
- a) Obtenha a expressão matemática da função perda.
  - b) Mostre que o estimador de Bayes de  $\theta$  é o primeiro quartil da distribuição a posteriori de  $\theta$ .
  - c) Generalize o resultado em (b) considerando que a perda em (ii) é  $k$  vezes a distância entre  $\theta$  e  $a$ .

**Exercício 5.7** Considere o modelo  $Y|\theta \sim N(\theta, 1)$ , com  $\pi(\theta) \propto 1$ .

- a) Identifique a distribuição a posteriori de  $\theta$ ;
- b) Indique quais são os estimadores de Bayes sob perda quadrática, absoluta e 0-1 (considere uma única observação da variável aleatória  $Y$ ).
- c) Considere o teste de hipóteses:  $H_0 : \theta \leq 0$  vs  $H_1 : \theta > 0$ .
  - 1. Utilize o método discutido da seção 5.4.2 com  $\alpha = \beta$ , para os cenários onde (i)  $Y = -1$ , (ii)  $Y = 0,5$ , (iii)  $Y = 2$
  - 2. Repita o item anterior usando fator de Bayes, e interprete.

**Exercício 5.8** Considere  $Y|\theta \sim \text{Ber}(\theta)$ , e  $\theta \sim U(0,1)$ .

- a) Repita os itens a) e b) do exercício 5.6;
- b) Considere o teste de hipóteses  $H_0 : \theta \leq 1/2$  vs  $H_1 : \theta > 1/2$ .
  - 1. Considerando que o valor observado foi  $Y = 0$ , analise os cenários  $\alpha = \beta$ ,  $\alpha = \beta/4$  e  $\alpha = 3\beta$ . Repita o exercício considerando que o valor observado foi  $Y = 1$ . (Utilize o método discutido da seção 5.4.2)

2. Repita o item anterior usando fator de Bayes, e interprete.

**Exercício 5.9** Considere um único experimento do modelo  $Y|\theta \sim \text{Bin}(n, \theta)$ , e assuma a priori que  $\theta$  tem distribuição uniforme. Suponha que o valor observado foi  $Y = n$ . Mostre que o intervalo HPD  $100(1 - \alpha)\%$  de credibilidade para  $\theta$  é da forma  $[c, 1]$ , com  $c < 1$ .



# Capítulo 6

## Predição

### 6.1 Distribuição preditiva

Até agora, focou-se na inferência sobre o(s) parâmetro(s). Um modelo de probabilidades foi especificado para descrever o processo aleatório que assume-se gerar um conjunto de dados, e mostrou-se que, no contexto bayesiano, as informações de amostras e informações prévias são combinadas para obter inferência sobre o(s) parâmetro(s) na forma de uma distribuição posteriori. Por vezes, o objetivo ao elaborar um modelo estatístico é fazer previsões sobre os valores futuros do processo. Isso é abordado mais elegantemente na estatística bayesiana do que na correspondente teoria clássica, conforme discutimos a seguir.

O ponto essencial do argumento é que, ao fazer previsões sobre valores futuros com base em um modelo estimado existem duas fontes de incerteza:

- a incerteza sobre os valores dos parâmetros que foram estimados com base nos dados obtidos anteriormente;
- a incerteza devido ao fato de que qualquer valor futuro é, em si, um processo aleatório.

Na estatística clássica é usual ajustar um modelo aos dados coletados e, em seguida, fazer previsões de valores futuros sob pressuposto de que este modelo ajustado é correto, a chamada abordagem de “estimativa”. Nesta ótica, apenas a segunda fonte de incerteza está sendo incluída na análise, levando à previsões que julgam-se ser mais precisas do que podem realmente ser. Não existe na abordagem clássica uma maneira completamente satisfatória de contornar este problema, uma vez que os parâmetros não são considerados aleatórios.

Sob o paradigma bayesiano é simples e direto considerar ambas fontes de incerteza. Simplesmente pondera-se a predição sobre a incerteza acerca dos possíveis valores do(s) parâmetro(s), informação que é completamente fornecida pela posteriori.

Suponha que tem-se observações anteriores  $y = (y_1, \dots, y_n)$  de uma variável com função de densidade (ou verossimilhança)  $f(y|\theta)$  e que deseja-se fazer inferências sobre a distribuição de um valor futuro  $y_p$  a partir deste processo. Com a especificação de uma distribuição a priori  $f(\theta)$ , o teorema de Bayes leva à uma distribuição posteriori  $f(\theta|y)$ . Em seguida, a “função de densidade preditiva” de  $y_p$  dado  $y$  é obtida por:

$$f(y_p|y) = \int f(y_p, \theta|y) d\theta = \int f(y_p|\theta, y) f(\theta|y) d\theta.$$

Sob independência condicional, ou seja, sendo  $Y$  e  $Y_p$  independentes dado  $\theta$ , tem-se que

$$f(y_p|y) = \int f(y_p|\theta) f(\theta|y) d\theta. \quad (6.1)$$

A densidade preditiva é portanto a integral da verossimilhança (de uma única observação) multiplicada pela posteriori, ou seja, a verossimilhança da variável a ser predita, ponderada pela posteriori. Novamente, é importante notar que esta definição decorre simplesmente das leis usuais de manipulação de probabilidades e, a partir própria definição, tem-se uma interpretação simples em termos de probabilidades.

A abordagem correspondente na estatística clássica seria, por exemplo, inicialmente obter a estimativa de máxima verossimilhança  $\hat{\theta}$  de  $\theta$  e basear a inferência/predição na distribuição  $f(y_p|\hat{\theta})$ , a distribuição “estimada”<sup>1</sup>. Para enfatizar mais uma vez, isso desconsidera a variabilidade (incerteza) decorrentes de se utilizar um valor estimado de  $\theta$ , dando assim uma falsa sensação de precisão, uma vez que a densidade preditiva  $f(y_p|y)$  é usualmente mais variável por ponderar as predições ao longo da distribuição posteriori de  $\theta$ .

Na prática distribuição preditiva  $f(y_p|y)$  pode ser obtida de diferentes formas e consideramos três delas aqui.

- analiticamente
- por aproximação gaussiana
- por simulação

A obtenção analítica implica em resolver a integral em 6.1 chegando a alguma distribuição conhecida. Esta é a forma mais precisa para inferências preditiva. Entretanto, nem sempre pode ser obtida. Pouco modelos permite que a obtenção de tal expressão seja possível.

A obtenção por aproximação gaussiana consiste em supor que a distribuição preditiva pode ser aproximada por uma distribuição normal  $[Y_p|y] \sim N(\mu_p, \sigma_p^2)$  com

---

<sup>1</sup>Tal procedimento é por vezes chamado na literatura de predição *plug-in*, indicando-se que o valor de  $\hat{\theta}$  é *plugado* no lugar de  $\theta$

os parâmetros sendo calculados por propriedades de esperança e variância condicionais como se segue.

$$\begin{aligned}\mu_p &= E[Y_p|y] = E_{\theta|y} [E_Y[Y_p|\theta, y]|y] \\ \sigma_p^2 &= \text{Var}[Y_p|y] = E_{\theta|y} [\text{Var}_Y[Y_p|\theta, y]|y] + \text{Var}_{\theta|y} [E_Y[Y_p|\theta]|y]\end{aligned}$$

A aproximação pode funcionar mal em alguns casos e gerar probabilidades fora do domínio de  $Y_p$ . Além disto para  $Y$  discreto é necessário "agrupar" as probabilidades nos pontos suportes de  $Y$ .

A obtenção por simulação é a mais versátil de todas no sentido de contornar casos nos quais a solução analítica é intratável e sem impor a forma gaussiana para a preditiva. A solução consiste aproximar a integral em 6.1 obtendo-se amostras desta distribuição, o que pode ser feito em dois passos:

1. simular um valor  $\theta_i$  da posteriori  $f(\theta|y)$ ,
2. simular um valor  $y_{p(i)}$  da distribuição do modelo  $f(y|\theta_i)$ .

Repete-se os passos acima uma quantidade  $N$  arbitrária de vezes. A partir destas amostras pode-se obter inferências tais como gráficos da distribuição preditiva, estatísticas resumo e intervalos.

## 6.2 Exemplos

Embora simples a princípio, os cálculos da distribuição preditiva podem tornar-se difíceis na prática. No entanto, muitas das famílias conjugadas padrão da forma priori-verossimilhança podem induzir formas tratáveis para a distribuição preditiva.

**Exemplo 6.1** (*Amostra Binomial.*) Suponha que tem-se uma observação de  $Y|\theta \sim \text{Bin}(n, \theta)$  e a priori (conjugada) é  $\theta \sim \text{Beta}(p, q)$ . Como já visto anteriormente, a posteriori para  $\theta$  é dada por:

$$[\theta|y] \sim \text{Beta}(p + y, q + n - y)$$

Supõe-se agora que há a intenção de se fazer mais observações  $N$  no futuro, e seja  $z$  o número de sucessos nestes  $N$  ensaios futuros, de modo que  $z|\theta \sim \text{Bin}(N, \theta)$ . Portanto a verossimilhança para a observação futura é

$$f(z|\theta) = C_z^N \theta^z (1 - \theta)^{N-z}.$$

Então, para  $z = 0, 1, \dots, N$ ,

$$\begin{aligned}f(z|y) &= \int_0^1 C_z^N \theta^z (1 - \theta)^{N-z} \times \frac{\theta^{p+y-1} (1 - \theta)^{q+n-y-1}}{B(p + y, q + n - y)} d\theta \\ &= C_z^N \frac{B(z + p + y - 1, N - z + q + n - y)}{B(p + y, q + n - y)}\end{aligned}$$

Esta expressão é reconhecida como a densidade de uma distribuição Beta-Binomial<sup>2</sup>.

**Exemplo 6.2** (*Amostra Gama.*) Como no capítulo 2, suponha  $Y_1, \dots, Y_n$  são variáveis independentes com distribuição  $Y|\theta \sim \text{Ga}(k, \theta)$ , em que  $k$  é conhecido. Adota-se a priori conjugada  $\theta \sim \text{Ga}(p, q)$

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\},$$

o que pelo teorema de Bayes leva a  $\theta|y \sim \text{Ga}(G = p + nk, H = q + \sum y_i)$ .

A verossimilhança para uma observação futura  $y_p$  é

$$f(y_p|\theta) = \frac{\theta^k y_p^{k-1} \exp\{-\theta y_p\}}{\Gamma(k)}$$

e então a distribuição preditiva tem a forma

$$\begin{aligned} f(y_p|y) &= \int_0^\infty \frac{\theta^k y_p^{k-1} \exp\{-\theta y_p\}}{\Gamma(k)} \times \frac{H^G \theta^{G-1} \exp\{-H\theta\}}{\Gamma(k)} d\theta \\ &= \frac{H^G y_p^{k-1}}{B(k, G)(H + y_p)^{G+k}}; \quad (y > 0). \end{aligned}$$

## 6.3 Exercícios

**Exercício 6.1** Uma amostra aleatória  $y_1, \dots, y_n$  é observada de uma variável aleatória com distribuição  $Y|\theta \sim \text{Po}(\theta)$ . A priori é  $\theta \sim \text{Ga}(g, h)$ . Mostrar que a distribuição preditiva para uma observação futura,  $y_p$ , desta distribuição  $Y|\theta \sim \text{Po}(\theta)$  é:

$$P(y_p|y) = \binom{y+G-1}{G-1} \left( \frac{1}{1+H} \right)^{y_p} \left( 1 - \frac{1}{1+H} \right)^G; \quad y_p = 0, 1, \dots$$

para algum valor de  $G$  e  $H$ . Qual é essa distribuição?

**Exercício 6.2** Os pesos dos itens de um determinado processo de produção são independentes e identicamente distribuídos, cada um com uma distribuição  $N(\theta; 4)$ . O gerente de produção acredita que  $\theta$  varia de lote para lote, de acordo com uma distribuição  $N(110; 0,4)$ . Uma amostra de 5 itens é selecionada aleatoriamente, produzindo as medições:

108,0 ; 109,0 ; 107,4 ; 109,6 ; 112,0.

1. Derivar a distribuição posteriori para  $\theta$ .

<sup>2</sup>Lembrando que a função beta é definida por  $B(a, b) = \int_0^1 y^{a-1} (1-y)^{b-1} dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

2. Encontre também a distribuição preditiva para: (a) o peso de um outro item do lote; (b) a média amostral do peso de  $m$  outros itens do lote.
3. O que acontece em (b) quando  $m \rightarrow \infty$ ?

**Exercício 6.3** A distribuição de falhas ao longo do comprimento de uma fibra artificial segue um processo de Poisson, de modo que o número de falhas de um comprimento  $l$  da fibra é  $\text{Po}(l\theta)$ . Pouco se sabe sobre  $\theta$ . O número de falhas obtidos em 5 de fibras de comprimentos de 10, 15, 25, 30 e 40 metros, respectivamente, foram de 3, 2, 7, 6 e 10. Encontre a distribuição preditiva para o número de falhas em outra fibra com 60 metros de comprimento.



# Capítulo 7

## Propriedades Assintóticas

### 7.1 Introdução

Voltando para a análise conjugada para a média  $\theta$  da distribuição normal, com  $Y_1, \dots, Y_n \sim N(\theta, \tau^{-1})$ , com uma priori  $\theta \sim N(b, c^{-1})$  obteve-se

$$\theta|y \sim N\left(\frac{cb + n\tau\bar{y}}{c + n\tau}, \frac{1}{c + n\tau}\right).$$

Com  $n \rightarrow \infty$ , a expressão anterior resume-se a

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

Nota-se assim que, à medida que  $n$  se torna “suficientemente” grande, o efeito da priori desaparece e a posteriori fica determinada unicamente pelos dados. Além disso, a distribuição posteriori se torna cada vez mais concentrada em torno de  $\bar{y}$ , o qual, pela forte lei dos grandes números, converge para o verdadeiro valor de  $\theta$ . Estes argumentos são formalizados e generalizados como se segue.

### 7.2 Consistência

Se o valor real de  $\theta$  é  $\theta_0$  e a probabilidade a priori de  $\theta_0$  (ou, no caso contínuo, de uma vizinhança arbitrária de  $\theta_0$ ) não é igual a zero, então, com quantidades crescentes de dados  $y$ , a probabilidade à posteriori de que  $\theta = \theta_0$  (ou em uma vizinhança de  $\theta_0$ ) tende à unidade. Isto é provado como se segue.

Sejam  $y_1, \dots, y_n$  observações iid, cada uma com distribuição  $g(y|\theta)$ . Então a

densidade posteriori é

$$\begin{aligned} f(\theta|y_1, \dots, y_n) &\propto f(\theta) \prod_{i=1}^n g(y_i|\theta) \\ &= f(\theta) \exp \left\{ \sum_{i=1}^n \log\{g(y_i|\theta)\} \right\} \\ &= f(\theta) \exp\{nL_n(\theta)\}. \end{aligned}$$

Para um  $\theta$  fixo,  $L_n(\theta)$  é a média das  $n$  variáveis aleatórias iid, e portanto converge em probabilidade, para sua esperança:

$$\bar{L}(\theta) = \int \log\{g(y|\theta)\} g(y|\theta_0) dy$$

Pode-se mostrar que a expressão é maximizada quando  $\theta = \theta_0$ . Assim, para  $\theta \neq \theta_0$ , segue-se que a razão  $\exp\{nL_n(\theta_0, y)\} / \exp\{nL_n(\theta, y)\} \Rightarrow \infty$  com probabilidade 1 quando  $n \rightarrow \infty$ . Isto é suficiente para provar a afirmação, desde que  $f(\theta_0) \neq 0$ .

Assim, desde que a distribuição a priori não dê peso zero para o verdadeiro valor de  $\theta$ , eventualmente a probabilidade à posteriori se concentrará no verdadeiro valor.

### 7.3 Normalidade assintótica

Quando  $\theta$  é contínuo, o argumento anterior pode ser estendido para obter uma forma aproximada da distribuição posteriori, quando  $n$  é grande. Pelo argumento na seção anterior, quando  $n$  aumenta,  $\exp\{nL_n(\theta_0, y)\} / \exp\{nL_n(\theta, y)\}$  é desprezível exceto em uma vizinhança  $\theta_0$  gradativamente menor. Desta forma,  $f(\theta)$  pode ser considerado constante nesse intervalo e obtém-se:

$$f(\theta|y_1, \dots, y_n) \propto \exp\{n\bar{L}(\theta)\}$$

Além disto, expandindo-se  $\bar{L}(\theta)$  ao redor de  $\theta_0$  por uma série de Taylor,

$$\bar{L}(\theta) \approx \bar{L}_{\theta_0} - (\theta - \theta_0)^2 / (2\nu)$$

em que,

$$\nu = -1/\bar{L}''(\theta_0) = [I_n(\theta_0)]^{-1}$$

a recíproca da informação para  $\theta_0$ . Assim, finalmente, obtém-se a aproximação

$$f(\theta|y_1, \dots, y_n) \propto \exp\{-(\theta - \theta_0)^2 / (2\nu)\},$$

isto é,

$$\theta|y \sim N(\theta_0, I_n^{-1}(\theta_0))$$



Então, com  $n \rightarrow \infty$  a distribuição posteriori é aproximadamente normal, ao redor do verdadeiro valor de  $\theta$ , e com variância dada por  $[I_n(\theta_0)]^{-1}$ . Observe-se mais uma vez, que este resultado é verdadeiro independentemente da especificação a priori, desde que a priori não seja nula para verdadeiro valor de  $\theta$ .

Esse resultado tem diversos usos. Em primeiro lugar, ele pode ser usado diretamente para obter probabilidades a posteriori aproximadas em situações em que os cálculos para se obter a distribuição posteriori são difíceis. Em segundo lugar, a aproximação pode fornecer valores iniciais úteis para cálculos numéricos usados em situações para as quais soluções analíticas são intratáveis. Entretanto, o talvez mais importante, é que se mostra formalmente que uma vez que obtenha dados suficientes, a preocupação com a seleção da priori torna-se irrelevante. Dois indivíduos podem especificar formas bastante distintas para as suas crenças anteriores, mas, eventualmente, uma vez que uma quantidade suficiente de dados se torne disponível, as inferências a posteriori de ambos serão as mesmas.

Portanto a aproximação assintótica posteriori por uma normal segue os seguintes passos:

- (i) obter a expressão da log-verossimilhança (sem a necessidade da constante normalizadora),
- (ii) obter o MLE  $\hat{\theta}$ , estimador de máxima verossimilhança,
- (iii) obter a curvatura  $H$  da log-verossimilhança (hessiano) ao redor da moda e portanto a informação observada  $I_n(\hat{\theta}) = -H$
- (iv) aproximar a distribuição por  $N(\hat{\theta}; I_n^{-1}(\hat{\theta}))$

O passos acima devem produzir para grandes amostras resultado similar à aproximação da posteriori que seria obtida por:

- (i) obter a log-densidade da posteriori (sem a necessidade da constante normalizadora),
- (ii) obter a moda ( $\hat{\theta}$ ) desta log-densidade,
- (iii) obter a curvatura  $H$  da log-densidade (hessiano) ao redor da moda,
- (iv) aproximar a distribuição por  $N(\hat{\theta}; -H^{-1}(\hat{\theta}))$

Os passos (ii) e (iii) podem ser obtidos analiticamente em alguns casos, mas de forma geral se utilizam algoritmos numéricos.

## 7.4 Exemplos

**Exemplo 7.1** (*Média da Normal.*) Seja  $Y_1, \dots, Y_n$  um conjunto de variáveis independentes com distribuição  $N(\theta, \sigma^2)$ , com  $\sigma^2$  conhecido.

De forma usual, a verossimilhança é dada por:

$$f(y|\theta) \propto \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \theta)^2}{2\sigma^2} \right\}.$$

Assim, pode-se tomar

$$\log(f(y|\theta)) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2$$

e então,

$$\frac{d \log(f(y|\theta))}{d\theta} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \theta)$$

e

$$\frac{d^2 \log(f(y|\theta))}{d^2\theta} = -\frac{n}{\sigma^2}$$

Consequentemente, o estimador de máxima verossimilhança  $\hat{\theta} = \bar{y}$  e  $I_n(\theta) = n/\sigma^2$ . Assim, assintoticamente, com  $n \rightarrow \infty$ ,

$$\theta|y \sim N(\bar{y}, \sigma^2/n).$$

O resultado é válido para qualquer distribuição priori com probabilidades não nulas na vizinhança do verdadeiro valor de  $\theta$ .

**Exemplo 7.2** (*Amostra Binomial.*) Considere novamente o modelo de verossimilhança  $Y \sim \text{Bin}(n, \theta)$ .

Então,

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, \dots, n.$$

Assim,

$$\log(f(y|\theta)) \propto y \log \theta + (n-y) \log(1-\theta).$$

Desta forma,

$$\frac{d \log(f(y|\theta))}{d\theta} = \frac{y}{\theta} - \frac{(n-y)}{(1-\theta)}$$

e

$$\frac{d^2 \log l(\theta)}{d^2\theta} = -\frac{y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}.$$

Consequentemente,

$$\begin{aligned} \hat{\theta} &= \frac{y}{n}, \\ I_n(\hat{\theta}) &= \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}. \end{aligned}$$

Então, quando  $n \rightarrow \infty$ ,

$$\theta|y \sim N\left(\frac{y}{n}, \frac{\frac{y}{n}(1-\frac{y}{n})}{n}\right)$$

## 7.5 Exercícios

**Exercício 7.1** *Encontre a distribuição assintótica posteriori para  $\theta$  para os dois modelos no exercício 2.1.*

**Exercício 7.2** *Encontre a distribuição assintótica posteriori para  $b$  no modelo de Pareto do exercício 3.3.*



# Capítulo 8

## Outros tópicos

Ao longo deste texto foram mencionadas as principais ideias de inferência bayesiana, indicando a simplicidade por trás da metodologia, a facilidade de interpretação e a importância de ser capaz de explorar todas as fontes de informação. Técnicas bayesianas são cada vez mais utilizadas em uma variedade de situações que envolvem modelagens estatísticas complexas, e isto requer, de alguma maneira, técnicas mais avançadas do que as vistas até aqui. Neste capítulo algumas destas ideias são brevemente resumidas.

### 8.1 Bayes Empírico

Esta é uma forma de usar a informação da amostra para auxiliar na especificação da distribuição priori. Desta forma, tal procedimento não é estritamente bayesiano, uma vez que, em um procedimento genuinamente bayesiano, a priori deve ser formulada independentemente dos dados. Entretanto, métodos *bayesianos empíricos* tem sido largamente utilizados na prática de análises bayesianas.

Como ilustração, considere o exemplo a seguir.

**Exemplo 8.1** *Sejam observações  $y_1, y_2, \dots, y_n$  tomadas de v.a. independentes, cada uma com distribuição  $N(\theta_i, 1)$  com  $\theta^T = (\theta_1, \theta_2, \dots, \theta_n)$ . A priori assumida para  $\theta$  é tal que os  $\theta_i$  são considerados independentes com distribuição  $N(\mu, 1)$ , para um hiperparâmetro desconhecido  $\mu$ .*

Na análise bayesiana completa, o valor de  $\mu$  deve ser especificado usando outras fontes de informação que não os dados, por exemplo, a opinião de algum especialista. Na abordagem bayesiana empírica, a própria amostra é usada para estimar  $\mu$ . O estimador óbvio é a média amostral  $\bar{y}$ , com o qual a priori passa a ser  $N(\bar{y}, 1)$ , o que então, da forma usual, leva à posteriores

$$\theta_i | y \sim N\left(\frac{y_i + \bar{y}}{2}, \frac{1}{2}\right),$$

independentes para cada  $\theta_i$ . Desta forma, a estimativa dada pela moda da posteriori de cada  $\theta_i$  passa a ser  $\hat{\theta}_i = (y_i + \bar{y})/2$ . É interessante notar que as estimativas de máxima verossimilhança seriam  $\hat{\theta}_i = y_i$  e portanto as estimativas bayesianas podem ser vistas como uma *suavização* destas últimas.

## 8.2 Estimação Linear Bayesiana

É claramente difícil formular alguma informação a priori de forma muito acurada, e pode ocorrer que seja possível especificar com alguma convicção apenas médias, variâncias e covariâncias. Entretanto, a posteriori vai depender da completa especificação da distribuição priori. O *estimador linear bayesiano* de um parâmetro é um estimador cujo valor depende apenas das médias e covariâncias, sem requerer uma completa especificação da priori.

## 8.3 Robustez

Há uma vasta literatura no assunto, que, de certa forma, é um problema mais sério em inferência estatística bayesiana do que em inferência estatística frequentista. Em estatística frequentista é necessário assegurar que as inferências feitas não são indevidamente sensíveis ao modelo pressuposto, que é imposto de forma possivelmente arbitrária. Isto é, as inferências devem ser robustas a aspectos da escolha de modelos, sobre os quais temos pouca convicção. Isto também ocorre no contexto bayesiano, mas há um tópico adicional sobre robustez também à escolha da priori. A investigação deste tópico específico é por vezes chamada na literatura de *análise de sensibilidade*. A distribuição posteriori será sempre o que de fato é relevante para inferências, para qualquer especificação da distribuição priori. Entretanto, espera-se que prioris que não sejam muito dissimilares levam a posteriors que são também razoavelmente similares. Portanto, é indesejável que a análise dependa demais de uma específica escolha de priori, uma vez que a priori é, na melhor das hipóteses, apenas um resumo grosseiro da reais convicções anteriores que se possa ter.

## 8.4 Computação

Um dos maiores obstáculos à implementação de técnicas bayesianas no passado era a dificuldade nas computações envolvidas no procedimento. Como visto, o uso de prioris conjugadas pode simplificar substancialmente as computações e análises, mas estas podem ser inadequadas ou indisponíveis em problemas com alguma complexidade.

Uma abordagem para contornar tais dificuldades tem sido o desenvolvimento de técnicas altamente especializadas para integração numérica e que podem levar em consideração especificidades da estrutura computacional necessária para a análise à posteriori. Mas, mesmo tais técnicas, possuem limitações para modelos

multidimensionais de alta complexidade. Desenvolvimentos nas últimas décadas tem reconhecido que técnicas de simulação podem ser usadas para gerar amostras cuja distribuição é a posteriori. Isto não é simples, uma vez que as distribuições posteriores podem ser de alta dimensão, com forma analítica indisponível e complexas estruturas de dependências. Devido a tais fatos, uma diversidade de técnicas têm sido desenvolvidas utilizando técnicas de *Cadeias de Markov via Monte Carlo* (MCMC - sigla em inglês para *Monte Carlo Markov Chain*). A técnica MCMC talvez mais simples e mais utilizada é o *amostrador de Gibbs* cujo algoritmo em sua forma básica, é descrito a seguir.

### 8.4.1 Amostrador de Gibbs

Suponha que o vetor de parâmetros de um modelo seja  $\theta = (\theta_1, \theta_2, \dots, \theta_d)^T$  e deseja-se simular da distribuição conjunta  $f(\theta|y)$ . O “truque” é escolher um valor inicial  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})^T$  para o vetor de parâmetros e então simular da seguinte maneira.

1. Simular  $\theta_1$  de  $[\theta_1|y, \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_d^{(0)}]$ .
2. Trocar  $\theta_1^{(0)}$  pelo valor simulado  $\theta_1$ .
3. Simular  $\theta_2$  de  $[\theta_2|y, \theta_1^{(0)}, \theta_3^{(0)}, \dots, \theta_d^{(0)}]$ .
4. Trocar  $\theta_2^{(0)}$  por  $\theta_2$ .
5. Continuar ciclando o algoritmo para  $\theta_1, \theta_2, \dots, \theta_d$ , um grande número de vezes.

Sob condições bastante gerais a sequência de  $\theta$ 's obtidos desta maneira forma uma cadeia de Markov (dado o valor atual, o valor subsequente de  $\theta$  é independente de todos os valores anteriores), e a distribuição estacionária da cadeia é a verdadeira distribuição a posteriori.

Desta forma, em uma aplicação típica, a cadeia é inicializada, rodada por um longo período até que a amostra resultante pareça ter atingido a distribuição estacionária, e, a partir deste ponto, conhecido como aquecimento (*burn-in*) da cadeia, a sequência subsequente de amostras é analisada como sendo uma amostra da distribuição a posteriori.

O algoritmo de Gibbs requer que seja possível simular das distribuições condicionais de cada parâmetro dadas as observações e os valores atuais dos demais parâmetros na cadeia, o que dá origem a uma diversidade de técnicas de simulação. Em notação, o algoritmo consiste em simular de cada distribuição condicional  $[\theta_i|y, \theta_{-i}^{(0)}]$  em que  $\theta_i$  é o  $i$ -ésimo elemento do vetor de parâmetros que está sendo simulado,  $y$  são os dados e  $\theta_{-i}^{(0)}$  é o estado atual da cadeia para os demais elementos do vetor de parâmetros. Em situações mais gerais, em que não é possível simular diretamente de  $[\theta_i|y, \theta_{-i}^{(0)}]$ , pois a distribuição pode não ter forma analítica, aproximações e/ou outros métodos mais gerais de MCMC são necessários.

## 8.5 Exemplos

Para demonstrar o uso do amostrador de Gibbs, considere-se um problema um pouco intrincado, que consiste em um modelo para o ponto de mudança em um processo de Poisson.

**Exemplo 8.1** *(Ponto de mudança em processo de Poisson.) Considere um processo de Poisson em tempos discretos com taxa  $\lambda_1$  em  $t_1, \dots, t_k$  e taxa  $\lambda_2$  em  $t_{k+1}, \dots, t_n$ . Assume-se prioris Gamma independentes, com parâmetros diferentes para as taxas Poisson e hiperprioris para as taxas das prioris. O objetivo é montar o amostrador de Gibbs.*

Para contextualizar o problema considera-se dados referentes a uma série com os registros dos números anuais de desastres em minas de carvão britânicas no período de 1851 a 1962. Os dados são mostrados na Tabela 8.1 e representados graficamente na Figura 8.1.

Tabela 8.1: Número de desastres por ano

	0	1	2	3	4	5	6	7	8	9
185_		4	5	4	1	0	4	3	4	0
186_	6	3	3	4	0	2	6	3	3	5
187_	4	5	3	1	4	4	1	5	5	3
188_	4	2	5	2	2	3	4	2	1	3
189_	2	2	1	1	1	1	3	0	0	1
190_	0	1	1	0	0	3	1	0	3	2
191_	2	0	1	1	1	0	1	0	1	0
192_	0	0	2	1	0	0	0	1	1	0
193_	2	3	3	1	1	2	1	1	1	1
194_	2	4	2	0	0	0	1	4	0	0
195_	0	1	0	0	0	0	0	1	0	0
196_	1	0	1							

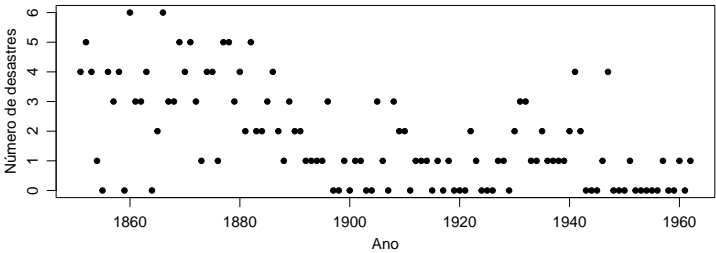


Figura 8.1: Número anual de desastres em minas de carvão.

O gráfico sugere que pode ter havido uma redução na taxa de desastres no período. Carlin et al. (1992) analisam os dados adotando um modelo que tem a



seguinte forma:

$$\begin{aligned} Y_i | \theta &\sim \text{Po}(\theta) ; i = 1, \dots, k \\ Y_i | \lambda &\sim \text{Po}(\lambda) ; i = k+1, \dots, n, \end{aligned}$$

ou seja, o modelo é para um número anual de desastres segundo a distribuição de Poisson, com uma taxa média de  $\theta$  até o  $k$ -ésimo ano (considerado como o ano da possível mudança), e uma taxa média de  $\lambda$  daí em diante. Desta forma, se  $k \geq n$ , em que  $n$  corresponderia ao último ano da série, a conclusão seria a de que não haveria ocorrido mudança na taxa de desastres.

A especificação bayesiana do modelo é completada de forma hierárquica com priors independentes para os parâmetros do modelo,

$$\theta \sim \text{Ga}(a_1; b_1) ; \lambda \sim \text{Ga}(a_2; b_2) \text{ e } k \sim \text{U}_d(1; 112) .$$

Seguimos aqui Carlin et al. (1992) que assume ainda as seguintes *hiperpriors*, também independentes,

$$b_1 \sim \text{InvGa}(c_1; d_1) \text{ e } b_2 \sim \text{InvGa}(c_2; d_2) .$$

Este modelo é dito ser *hierárquico* com três níveis, o primeiro especificando as distribuições para as quantidades observadas, o segundo especificando distribuições para os parâmetros das distribuições do primeiro nível, e o terceiro para os parâmetros no segundo nível.

Não é muito difícil verificar que tal especificação de priors leva às distribuições condicionais a seguir.

$$\begin{aligned} \theta | Y, \lambda, b_1, b_2, k &\sim \text{Ga} \left( a_1 + \sum_{i=1}^k Y_i ; k + 1 / b_1 \right) \\ \lambda | Y, \theta, b_1, b_2, k &\sim \text{Ga} \left( a_2 + \sum_{i=k+1}^n Y_i ; n - k + 1 / b_2 \right) \\ b_1 | Y, \lambda, \theta, b_2, k &\sim \text{Ga} (a_1 + c_1 ; \theta + 1 / d_1) \\ b_2 | Y, \lambda, \theta, b_1, k &\sim \text{Ga} (a_2 + c_2 ; \lambda + 1 / d_2) \\ &\text{e} \\ p[k | Y, \lambda, \theta, b_1, b_2] &= \frac{L(Y; k, \theta, \lambda)}{\sum_{j=1}^n L(Y; j, \theta, \lambda)} \\ &\text{com} \\ L(Y; k, \theta, \lambda) &= \exp\{k(\lambda - \theta)\} (\theta / \lambda)^{\sum_{i=1}^k Y_i} . \end{aligned}$$

Todas as condicionais são conhecidas, e pode-se obter valores simulados diretamente, segundo os passos do amostrador de Gibbs. Na Figura 8.2 são mostradas as seqüências de valores para  $\theta$ ,  $\lambda$  e  $k$ , simulados em 1100 iterações do algoritmo. As priors e hiperpriors foram especificadas com  $a_1 = a_2 = 0,5$ ,  $c_1 = c_2 = 0$ ,  $d_1 = d_2 = 1$ .

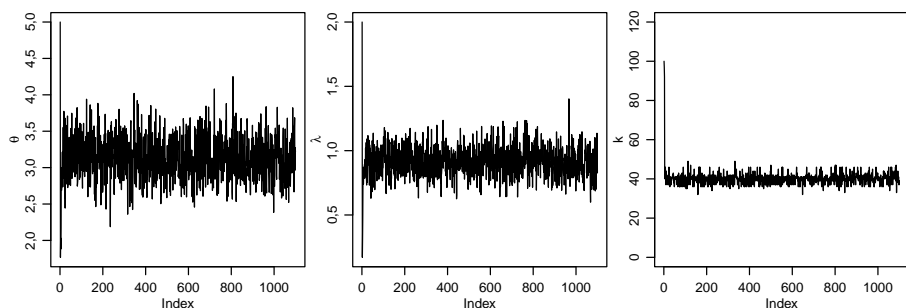


Figura 8.2: Cadeias geradas para os parâmetros  $\theta$ ,  $\lambda$  e  $k$ .

A convergência para a distribuição estacionária parece rápida, após a compensação para a escolha “ruim” para o valor inicial dos parâmetros. Desta forma, opta-se por descartar os 100 valores iniciais, e basear as análises subsequentes nos 1000 pontos restantes. O gráfico de frequências da distribuição posteriori de  $k$  é dada na Figura 8.3.

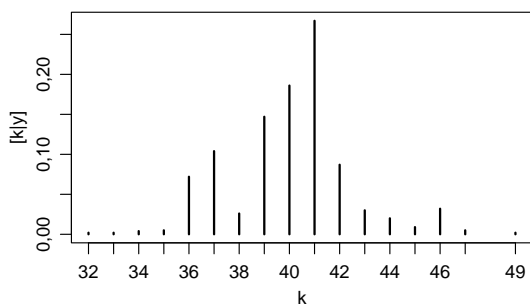


Figura 8.3: Distribuição a posteriori de  $k$ .

Baseando-se nesta estimativa, é praticamente certo que o ponto de mudança de fato ocorreu, com a estimativa de moda à posteriori sendo  $k = 41$ , o que corresponde a um ponto de mudança no ano de 1891.

Estimativas suavizadas das distribuições posteriores de  $\theta$  e  $\lambda$  são fornecidas nos gráficos à esquerda e no centro da Figura 8.4.

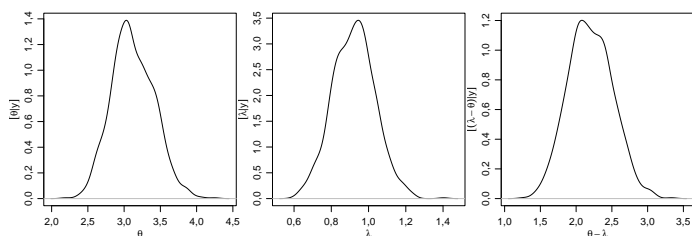


Figura 8.4: Posterioris de  $\theta$ ,  $\lambda$  e de  $(\theta - \lambda)$ .

É claro por estes gráficos que  $\lambda$  é quase certamente menor que  $\theta$ , uma vez que assumem valores em intervalos claramente distintos. Mas isto pode ser investigado de forma ainda mais objetiva obtendo-se a distribuição à posteriori (marginal) para diferença dos dois parâmetros,  $[(\theta - \lambda)|y]$ . Um aspecto importante do amostrador de Gibbs (e, de forma geral, dos métodos que produzem simulações da posteriori) é que, praticamente sem esforço adicional algum, contrastes de interesse envolvendo os parâmetros pode ser examinados de direta e simples. Isto é, observando-se a sequência de valores  $(\theta - \lambda)_i$ , obtida simplesmente aplicando esta operação nos valores simulados, obtém-se a distribuição posteriori marginal desejada  $[(\theta - \lambda)|y]$ . A distribuição assim obtida é portanto marginalizada (numericamente integrada) em relação aos demais parâmetros do modelo, no caso  $(k, b_1, b_2)$ . Nenhuma teoria complicada de transformação de variáveis é portanto necessária para obtenção de tal posteriori, uma vez que se está trabalhando diretamente na amostra. Uma estimativa suavizada de  $[(\theta - \lambda)|y]$  produzida desta forma é mostrada no gráfico da direita da Figura 8.4. O valor zero, que indicaria que não houve mudança na taxa, está claramente fora do intervalo de valores de  $(\theta - \lambda)$ , o que permite concluir que de fato houve redução na taxa.

Desta forma, o amostrador de Gibbs possibilitou conduzir a inferência bayesiana em um problema um pouco complexo, o qual não seria tão facilmente tratado usando qualquer procedimento de inferência clássica (frequentista).

## 8.6 Exercícios

**Exercício 8.1** *Revisite a inferência sobre os parâmetros da distribuição normal no caso que ambos são considerados desconhecidos. Assuma alguma distribuição a priori conveniente para a qual a posteriori seja disponível em forma analítica. Obtenha as distribuições condicionais e implemente computacionalmente o amostrador de Gibbs. Compare os resultados (distribuição posteriori e resumos pontuais e intervalares) por simulação com os obtidos analiticamente.*

**Exercício 8.2** *Implemente computacionalmente o amostrador de Gibbs para o modelo dos desastres em minas de carvão discutido neste capítulo.*



# Referências Bibliográficas

- ALBERT, J. **Bayesian Computation with R**. New York, NY: Springer-Verlag New York, 2009.
- CARLIN, B. P.; GELFAND, A. E.; SMITH, A. F. M. Hierarchical Bayesian analysis of changepoint problems. **Applied Statistics**, v.41, p.389–405, 1992.
- CARLIN, B. P.; LOUIS, T. A. **Bayesian methods for data analysis**. Boca Raton: CRC Press, 2009.
- GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. Boca Raton [etc.]: CRC Press, 2014.
- GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. Cambridge; New York: Cambridge University Press, 2007.
- GILL, J. **Bayesian methods: a social and behavioral sciences approach**. Boca Raton: Chapman & Hall/CRC, 2008.
- KRUSCHKE, J. K. **Doing bayesian data analysis a tutorial with R and BUGS**. Burlington, MA: Academic Press, 2011.
- LEE, P. **Bayesian Statistics – An Introduction**. Wiley, 2012.
- MARIN, J.-M.; ROBERT, C. P. **Bayesian essentials with R**. Springer, 2014.
- MIGON, H.; GAMERMAN, D.; LOUZADA, F. **Statistical Inference: An Integrated Approach**. 2. ed. CRC Press, 2015.
- O'HAGAN, A. **Bayesian Inference**. Edward Arnold, 1994. Kendall's advanced theory of statistics.
- PEREIRA, C. A. B.; STERN, J. M. Evidence and credibility: full Bayesian test for precise hypotheses. **Entropy**, v.1, p.99–110, 1999.
- PEREIRA, C. A. D. B.; STERN, J. M. S.; WECHSLER, S. Can a significance test be genuinely Bayesian. **Bayesian Analysis**, v.3, n.1, p.79–100, 2008.
- STONE, J. V. **Bayes' rule: a tutorial introduction to Bayesian analysis**. Sebtel Press, 2013. <http://jim-stone.staff.shef.ac.uk/BookBayes2012/BayesRuleBookMain.html>.