CE-227

Inferência Estatística

Introdução a Inferência Bayesiana

Stuart Coles

Trad. Paulo Justiniano Ribeiro Jr

Inferência Estatística (Inferência Bayesiana)

Stuart Coles
Trad. 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 @ leg.ufpr.br

CE-227 Inferência Bayesiana
Disciplina do Bacharelado em Estatística
Universidade Federal do Paraná
1ªoferta: 1ºsemestre de 2014
Última atualização: 12 de maio de 2016

Prefácio

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

P.J.R.Jr
Curitiba, 12 de maio de 2016

Sumário

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 definição são possíveis mas a maioria vão ao 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 que 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 θ 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 θ. Denotando por X a variável aleatória que corresponde ao número de árvores doentes na amostra, a Unidade Florestal vai usar o valor observado $X = x$, para inferir sobre o parâmetro populacional θ. Esta inferência pode ser na forma de uma *estimativa pontual* ($\hat{\theta} = 0{,}1$); um *intervalo de confiança* (95% de confiança que θ se encontra entre $[0{,}08; 0{,}12]$; 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 $X = x$ está sendo usado para fazer inferências sobre a característica θ da população. Além disto, essas inferências são feitas especificando-se um modelo de probabilidades, $f(x|\theta)$, que

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

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

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

Inferência estatística resulta então a uma inferência sobre o parâmetro populacional θ baseada na observação $X = x$, e basicamente infere-se que valores de θ que conferem uma alta probabilidade para o valor observado x 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 paramétrica de modelos.

Antes de abordar inferência bayesiana em particular, há alguns pontos a serem considerados sobre a abordagem clássica de inferência. O ponto mais fundamental é o de que o parâmetro θ, 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 θ estivesse entre 0,08 e 0,12. Entretanto tal interpretação *não é possível* uma vez que θ não é aleatório: sendo considerado constante, o valor de θ ou *está* no intervalo ou *não está* – probabilidades não está (e não pode estar) 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 θ em 95% das vezes. Todas inferências baseadas na teoria clássica são forçadas a terem este tipo de interpretação de "frequência em longas sequências", 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 θ a respeito do qual deseja-se fazer inferências, e um mecanismo probabilístico $f(x|\theta)$ que determina a probabilidade de observar cada valor diferente de x, sob diferentes valores de θ. A diferença essencial entretanto é a de que θ é tratado como um quantidade *aleatória*. Isto pode parecer inócuo, mas de fato leva a abordagens substancialmente diferentes à modelagem e inferência.

Em essência, a inferência vai ser baseada em $f(\theta|x)$ ao invés de $f(x|\theta)$; 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 θ antes de se considerar qualquer informação contida nos dados.

Esta noção de distribuição *a priori* para o parâmetro θ está no cerne do pensamento bayesiano. Dependendo se fala com um defensor ou oponente da metodologia, é, ou sua principal vantagem sobre a teoria clássica, ou sua maior armadilha[1].

1.3 Distribuição a priori

Em quase todas situações, quado se tenta estimar um parâmetro θ, tem-se algum conhecimento ou crença sobre o valor de θ 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, denotanto 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 é a mesma sendo 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 $X|\theta \sim \text{Bin}(10,\theta)$ e observamos $x = 10$ tal que (verificar) as hipóteses $H_0 : \theta \leq 0{,}5$ é rejeitada a favor de $H_1 : \theta > 0{,}5$ cada vez:

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 predizer o resultado do lançamento de uma moeda honesta e acerta todas as 10 tentativas efetuadas.

[1] 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 tirar 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 tem objeção. Crenças prévias diferentes vão levar a diferentes inferências na visão bayesiana das coisas, e é exatamente se isto é visto como uma coisa boa ou ruim que determina a aceitação 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. Inferência é baseada na distribuição *a posteriori* $f(\theta|x)$, cuja forma será vista dependente (através do Teorema de Bayes) das particularidades da especificação da priori $f(\theta)$.

- **Auto consistência** Tratando o parâmetro θ como aleatório, surge 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 θ, que por sua vez são derivadas diretamente da distribuição a posteriori $f(\theta|x)$.

- **Sem procedimentos "ad-hoc"** Devido ao fato de que em inferência clássica não se pode fazer declarações probabilísticas a respeito de θ, 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. Inferência bayesiana esquiva-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 sem ambiguidade a inferência completa sobre o θ 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. 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) alterando-se o nível de significância de um teste. Se acredita-se que os dados devem conduzir a rejeição de uma hipótese, isto pode ser assegurado escolhendo-se um nível de significância alto para o teste! Então, de certa forma, a incorporação de informação anterior é formalizada 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 (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 $A|B$ [2].

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

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

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-positivos e falso-negativos?*

Notação: A: pessoal é HIV-positiva, e B o resultado do teste é positivo. Temos $P[A] = 0{,}1$, $P[B|A] = 0{,}9$ e $P[\overline{B}|\overline{A}] = 0{,}85$. Então:

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

De forma similar,

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

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

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

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.*

Seja A : 3 bolas pretas são retiradas, e C_i: haviam i bolas pretas na sacola. Então pelo teorema de Bayes:

$$P[C_3|A] = \frac{P[A|C_3]P[C_3]}{\sum_{j=0}^{6} P[A|C_j]P[C_j]}$$

Mas há uma questão essencial aqui: quais valores atribuímos a $P[C_0], \ldots, P[C_6]$? 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). Sem nenhuma outra informação ao contrário, pode-se bem assumir que todos os números de bolas pretas são igualmente prováveis tomando portanto $P[C_0] = P[C_1] = \ldots = P[C_6] = 1/7$. 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 $P[C_0]$ e $P[C_7]$. Ou, 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 é que é necessário *pensar bem* sobre como expressar opiniões à priori, e que a resposta que será obtida vai depender do que acreditamos ao início.

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

$$\begin{aligned} P[C_3|A] &= \frac{P[A|C_3]P[C_3]}{\sum_{j=0}^{6} P[A|C_j]P[C_j]} \\ &= \frac{\frac{1}{7} \times (\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4})}{\frac{1}{7}\left[0+0+0+(\frac{3}{6} \times \frac{2}{5} \times \frac{1}{4})+(\frac{4}{6} \times \frac{3}{5} \times \frac{2}{4})+(\frac{5}{6} \times \frac{4}{5} \times \frac{3}{4})+(\frac{6}{6} \times \frac{5}{5} \times \frac{4}{4})\right]} \\ &= \frac{1}{35} \end{aligned}$$

Desta forma, os dados atualizaram a opinião prévia que $P[C_3] = 1/7$ para a probabilidade a posteriori $P[C_3] = 1/35$. Isto é, o evento torna-se muito menos provável após os dados serem obtidos.

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 en-*

contrados, as probabilidades de presença do fóssil são mostradas na tabela 1.1. 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 1.1: 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 vai 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?*

Capítulo 2

Atualização Bayesiana

2.1 Introdução

Como delineado no Capítulo 1, a essência da abordagem bayesiana é tratar o parâmetro desconhecido θ como uma variável aleatória, especificar uma distribuição a priori para θ que represente as convicções sobre θ 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 $f(x|\theta)$;
2. determinação da priori $f(\theta)$;
3. cálculo da distribuição posteriori $f(\theta|x)$, 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 θ. 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:

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

Esta notação será usada em ambos os casos, com X 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 X ($P[X = x]$). De forma similar, θ pode ser discreto ou contínuo mas, caso seja discreto, $\int f(\theta) f(x|\theta) \mathrm{d}\theta$ deve ser interpretado como $\sum_j f(\theta_j) f(x|\theta_j)$.

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

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

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 mecanismo 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 contagens 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 a opinião sobre θ 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 – veja 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 θ disponível a partir da priori e dos dados é representado pela distribuição posteriori. Isto é, $f(\theta|x)$ 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 suficiente, ancilares, etc. Estes 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 $x = (x_1, x_2)$, então x_1 é suficiente para θ se $f(\theta|x)$ é independente de x_2.

2.4 Exemplos

Exemplo 2.1 *Quando uma máquina em particular se torna defeituosa, a causa pode ser atribuída à uma falha no motor ou à 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* $f(\theta|x)$ *e interprete adequadamente os resultados.*

Tabela 2.1: Tabela de probabilidades das causas de defeitos

Localização da falha	AQ (x_1)	TI (x_2)	ambas (x_3)
Motor	0,10	0,40	0,50
Transmissão	0,50	0,30	0,20

Para este simples exemplo de um caso discreto as análises podem ser tabuladas como mostrado na tabela 2.2 que se usa-se que $f(x,\theta) = f(x|\theta)f(\theta)$ e $f(x) = \sum_{i=1}^{2} f(x,\theta_i)$.

Tabela 2.2: Tabela de probabilidades das causas de defeitos

$f(\theta)$	$f(x\|\theta)$	x_1	x_2	x_3
0,4	θ_1	0,10	0,40	0,50
0,6	θ_2	0,50	0,30	0,20
$f(x,\theta)$	θ_1	0,04	0,16	0,20
	θ_2	0,30	0,18	0,12
	$f(x)$	0,34	0,34	0,32
$f(\theta\|x)$	θ_1	4/34	16/34	20/32
	θ_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 X, as chances eram de 3:2 a favor de θ_2 (uma vez que 60% das falhas eram na transmissão). Mas tendo observado x, estas chances mudaram para 15:2 a favor de θ_2 se observa-se que $x = x_1$, 9:8 a favor de θ_2 se observa-se que $x = x_2$, e 5:3 a favor de θ_1 se observa-se que $x = x_3$. Consequentemente, se o critério de decisão é selecionar o diagnóstico de causa de falha mais plausível, observar x_1 ou x_2 levaria à decisão de que a falha é na transmissão, mas observando x_3 levaria à decisão de que a falha seria no motor.[1]

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

Portanto,

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

É claro que, em geral, a escolha da especificação da priori para θ 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. O ponto a este respeito é que espera-se que, uma vez que a família seja bastante vasta e inclua uma diversidade suficiente de formas possíveis,

[1] 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.

pode-se usar uma priori desta família que se aproxima das verdadeiras opiniões prévias. Se isto ocorre, consegue-se respostas calculadas de forma simples. Se, entretanto, não há uma priori dentro desta família 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 θ por uma distribuição Beta:

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

tal que

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

Então, pelo Teorema de Bayes

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

Agora, como sabemos que $f(\theta|x)$ é uma densidade de probabilidades própria, há que ser que

$$\theta|x \sim \text{Be}(p+x; q+n-x).$$

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 (p,q) da distribuição priori beta para $(p+x, q+n-x)$.

Como um exemplo numérico considera o conjunto da dados "CANCER" tomado do programa *First Bayes*.[2] De 70 pacientes que receberam tratamento segundo um novo protocolo para uma particular forma de câncer, 34 sobreviveram além de um período especificado. Denote por θ a probabilidade de sobrevivência dos pacientes. Consultas com médicos especialistas, que são familiarizados com ensaios clínicos similares os levam a expressar o conhecimento a priori que $\text{E}[\theta] = 0{,}40$ 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 $\text{E}[\theta] = 0{,}40$ 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 fácil verificar (*faça isto!*) que, fazendo $m = \text{E}[\theta]$ e $v = \text{Var}[\theta]$ as equações são resolvidas por:

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

[2] *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/*.

o que neste caso leva a $p = 4{,}4$ e $q = 6{,}6$. Isto especifica a distribuição da priori para θ, 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 especialistas anteriores. Presumindo-se que é, obtém-se agora a distribuição posteriori como sendo:

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

Esta distribuição posteriori sumariza toda a informação a respeito de θ e representa a completa inferência sobre θ. 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:

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

No caso considerado:

$$\text{E}[\theta] = 0{,}40 \quad \text{e } \text{E}[\theta|x] = 0{,}474,$$

e portanto, o efeito dos dados observados foi aumentar a estimativa à priori de θ de 0,40 para 0,474. Por outro lado, um estimador natural para θ, baseado apenas nos dados, seria $x/n = 37/80 = 0{,}486$, que é o estimador de máxima verossimilhança. A estimativa à posteriori é um compromisso entre a opinião prévia e a informação fornecida pelos dados.

De forma mais geral, se x e n são grandes em relação a p e q estão a esperança na posteriori é aproximadamente x/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 x 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 $X_1, \ldots, X_n$ são um conjunto de variáveis aleatórias independentes com distribuição* $\text{N}(\theta)$.

Tem-se que:

$$\begin{aligned} l(\theta|x) &= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \\ &\propto e^{-n\theta}\theta^{\sum x_i}. \end{aligned}$$

Como no caso binomial, convicções anteriores sobre θ 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

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

Assim, pelo Teorema de Bayes,

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

e consequentemente,

$$\theta|x \sim \text{Ga}(p+\sum_{i=1}^{n} x_i, q+n),$$

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

Como um exemplo numérico, novamente tomado do programa *First Bayes*, seja θ o número médio de gansos de um bando dentro de uma determinada região. Fotografias aéreas detalhadas de 45 bandos fornecem $\sum_{i=1}^{45} x_i = 4019$. Supõe-se que a média e variância a priori para θ são 100 e 20, respectivamente. Desta forma, usando o fato de que se $Y \sim \text{Ga}(p,q)$ então $\text{E}[X] = p/q$ e $\text{Var}[X] = p/q^2$, resolvendo para p e q obtém-se $p = 500$ e $q = 5$. Portanto a distribuição a posteriori obtida é $\theta|x \sim \text{Ga}(4519, 50)$.

Exemplo 2.4 *(Média da normal.) Suponha que $X_1, \ldots, X_n$ são um conjunto de variáveis aleatórias independentes com distribuição $\text{N}(\theta, \sigma^2)$, em que σ^2 é conhecido.*

Então,

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

e a verossimilhança fica

$$l(\theta|x) = (2\pi\sigma^2)^{-n/2}\exp\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}\} \propto \exp\{-\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{2\sigma^2}\}.$$

Agora, suponha que as convicções a priori sobre θ podem elas mesmas serem representadas por uma distribuição normal $\theta \sim \text{N}(b, d^2)$. 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 θ. Então, pelo Teorema

de Bayes,

$$\begin{aligned} f(\theta|x) &\propto \exp\left\{-\frac{(\theta-b)^2}{2d^2}\right\} \exp\left\{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{(\theta-b)^2}{2d^2} - \frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{(\theta^2-2b\theta+b^2}{2d^2} - \frac{\sum_{i=1}^n x_i^2 - 2n\overline{x}\theta + n\theta^2}{2\sigma^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\theta^2\left(\frac{1}{d^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{b}{d^2}+\frac{n\overline{x}}{\sigma^2}\right)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{d^2}+\frac{n}{\sigma^2}\right)\left[\theta - \frac{\frac{b}{d^2}+\frac{n\overline{x}}{\sigma^2}}{\frac{1}{d^2}+\frac{n}{\sigma^2}}\right]^2\right\} \end{aligned}$$

Portanto,

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

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,

$$\theta|x \sim \mathrm{N}(\frac{cb_n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau}).$$

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

1. Observe-se que
$$\mathrm{E}[\theta|x] = \gamma_n b + (1-\gamma_n)\overline{x}$$
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 $\overline{x}$. Além do mais, o parâmetro ponderador γ_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 $\overline{x}$.

2. Observe que "precisão à posteriori" = "precisão à priori" + n × "precisão de cada dado".

3. Quando $n \to \infty$, então (vagamente)
$$\theta|x \sim \mathrm{N}(\overline{x}, \frac{\sigma^2}{n})$$
de tal forma que no limite a priori não tem efeito.

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

$$\theta|x \sim \mathrm{N}(\overline{x}, \frac{\sigma^2}{n}).$$

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

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 registrados por Henry Cavendish no século XVIII. Ele tomou 23 medidas da densidade da terra. Para estes dados, $\overline{x} = 5{,}48$ e supõe-se aqui que a variância de seu erro de medida é conhecida e igual a 0,04. Agora, supõe-se que, de experimentos prévios, a priori para θ, a densidade da terra, é considerada ser $\mathrm{N}(5{,}4; 0{,}01)$. Então, temos que a posteriori é simplesmente obtida como sendo $\theta|x \sim \mathrm{N}(5{,}46; 0{,}00303)$.

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 sequencia de 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 x_1, e atualizando a priori pelo Teorema de Bayes obtém-se:

$$f(\theta|x_1) = c_1 \cdot f(\theta) \cdot f(x_1|\theta) \propto f(\theta) \cdot f(x_1|\theta)$$

em que c_1 é a contante normalizador. Esta $f(\theta|x_1)$ se torna a nova distribuição a priori antes de x_2 ser observado. Então,

$$\begin{aligned} f(\theta|x_1, x_2) &= c_2 \cdot f(\theta) \cdot f(x_1|\theta) \cdot f(x_2|\theta) \\ &\propto f(\theta) \cdot f(x_1|\theta) \cdot f(x_2|\theta) \\ &= f(\theta) \cdot f(x_1, x_2|\theta) \end{aligned}$$

que é o mesmo resultado que seria obtido atualizando-se diretamente com base em toda a informação (x_1, x_2). Por indução, o argumento estende-se para sequencias de qualquer número de observações.

2.5.2 Suficiência

Em estatística clássica, a *suficiência* tem um papel central em ambos, desenvolvimentos teóricos a aplicações práticas. O mesmo ocorre em análise bayesiana, e já foram vistos diversos 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 $x = (x_1, x_2)$, então x_1 é suficiente para θ se $f(\theta|x)$ é independente de x_2, (neste caso x_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 θ 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 violam tal princípio.

2.6 Exercícios

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

1. *$x_1, \ldots, x_n$ é uma amostra aleatória de uma distribuição com função de probabilidades:*

$$f(x|\theta) = \theta^{x-1}(1-\theta)\,;\ x = 1, 2, \ldots$$

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

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

2. *$x_1, \ldots, x_n$ é uma amostra aleatória de uma distribuição com função probabilidades:*

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}\,;\ x = 0, 1, \ldots$$

com distribuição a priori

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

Exercício 2.2 *A proporção θ de itens defeituosos em um grande carregamento é desconhecida, mas uma avaliação especializada atribui a θ 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 θ?*

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 sequencia de produção varia segundo uma distribuição* $\mathrm{N}(\theta,1)$. *Um engenheiro especifica a distribuição a priori de θ como sendo* $\mathrm{N}(10;0{,}25)$. *Em uma sequencia 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(θ). A distribuição priori para θ 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 θ.*

Exercício 2.5 *Suponha que o tempo em minutos necessário para atender um cliente em um banco possui uma distribuição exponencial com parâmetro θ. A priori estabelecida para θ 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 θ.*

Exercício 2.6 *Uma amostra aleatória $x_1,\ldots,x_n$ é tomada de uma distribuição Poisson de média θ. A distribuição a priori para θ é Gama com média μ_0. Se a média amostral é $\bar{x}_n$, mostre que a média da distribuição a posteriori de θ vai ser uma média ponderada da forma*

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

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

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,

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

Por exemplo, suponha que $X_1, \ldots, X_n$ são variáveis aleatórias independentes com distribuição Poisson $X|\theta \sim \mathrm{P}(\theta)$ e a opinião prévia sobre θ 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

$$\int_0^1 \exp(-n\theta)\theta^{\sum x_i} d\theta,$$

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 impraticáveis. En-

tretanto, 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.) $X_1,\ldots,X_n$ são variáveis aleatórias independentes com distribuição $X \sim \text{Ga}(k,\theta)$, em que k é conhecido. (Note que o caso $k=1$ corresponde a distribuição exponencial).* Então

$$l(\theta|x) \propto \theta^{nk} \exp\{-\theta \sum x_i\}.$$

Agora, estudando esta forma, tomada como uma função de θ 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

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

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

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 x é da forma:

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

para funções h, g e c tais que

$$\int f(x|\theta)\mathrm{d}x = g(\theta)\int h(x)\exp\{t(x)c(\theta)\}\mathrm{d}x = 1.$$

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 $x_1, x_2, \ldots, x_n$,

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

Portanto escolhendo

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

obtém-se

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

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

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

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

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

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

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 X é definida como o número de "tentativas" até o r-ésimo sucesso.

Verossimilhança	Priori	Posteriori
$X \sim \mathrm{B}(n,\theta)$	$\theta \sim \mathrm{Beta}(p,q)$	$\theta\|x \sim \mathrm{Beta}(p+x, q+n-x)$
$X \sim \mathrm{P}(\theta)$	$\theta \sim \mathrm{Ga}(p,q)$	$\theta\|x \sim \mathrm{Ga}(p+\sum_{i=1}^{n} x_i, q+n)$
$X \sim \mathrm{N}(\theta,\tau^{-1})$, ($\tau$ conhecido)	$\theta \sim \mathrm{N}(b,c^{-1})$	$\theta\|x \sim \mathrm{N}(\frac{cb+n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau})$
$X \sim \mathrm{Ga}(k,\theta)$, ($k$ conhecido)	$\theta \sim \mathrm{Ga}(p,q)$	$\theta\|x \sim \mathrm{Ga}(p+nk, q+\sum_{i=1}^{n} x_i)$
$X \sim \mathrm{Geo}(\theta)$	$\theta \sim \mathrm{Beta}(p,q)$	$\theta\|x \sim \mathrm{Beta}(p+n, q+\sum_{i=1}^{n} x_i - n)$
$X \sim \mathrm{BN}(r,\theta)$	$\theta \sim \mathrm{Beta}(p,q)$	$\theta\|x \sim \mathrm{Beta}(p+r, q+x-r)$

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 $X_1,\ldots,X_n \sim \mathrm{N}(\theta,\tau^{-1}), \theta \sim \mathrm{N}(b,c^{-1})$, levando a $\theta|x \sim \mathrm{N}(\frac{cb+n\tau\overline{x}}{c+n\tau}, \frac{1}{c+n\tau})$ A força das opiniões prévias sobre θ 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 θ. Suponha agora que o conhecimento prévio sobre θ é tão pequeno que fazemos $c \to 0$. Neste caso simples o bastante, a posteriori passa a ser $\theta|x \sim \mathrm{N}(\overline{x}, \frac{1}{n\tau})$, ou em uma notação mais familiar, $\theta|x \sim \mathrm{N}(\overline{x}, \frac{\sigma^2}{n})$. 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 $c \to 0$. De fato, obtém-se uma priori $\theta \sim \mathrm{N}(0,\infty)$, que não é genuinamente uma distribuição de probabilidades. Na verdade, quando $c \to 0$, a distribuição $\theta \sim \mathrm{N}(b,c^{-1})$ se torna cada vez mais plana (*flat*) tal que no limite

$$f(\theta) \propto 1; \theta \in \Re$$

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

$$\int_{\Re} f(\theta)\, d\theta = \infty.$$

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

Portanto, 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 um parâmetro definido por $\phi = \theta^2$, então (pelo teorema de transformação de variáveis),

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

Por outro lado, sendo "ignorante" a respeito de θ certamente se é ignorante à respeito de ϕ, 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:

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

E a priori de Jeffreys é definida como sendo:

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

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

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

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

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

Desta forma, vê-se que a priori de Jeffreys para θ se transforma naturalmente na priori de Jeffreys para ϕ. Em outras palavras, é invariante sob reparametrização.

3.4.1 Exemplos

Exemplo 3.1 *(Média da normal.) Supondo que $X_1, \ldots, X_n$ são variáveis aleatórias independentes com distribuição $N(\theta, \sigma^2)$, com σ^2 conhecido.* Então,

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

logo,

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

e,

$$\begin{aligned} I(\theta) &= -E\left\{\frac{\mathrm{d}^2 \log f(x|\theta)}{\mathrm{d}\theta^2}\right\} \\ &= E\left\{\frac{1}{\sigma^2}\right\} \\ &= \frac{1}{\sigma^2} \end{aligned}$$

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 x_1 e usar a propriedade de que, sob independência, $I_n(\theta) = nI_1(\theta)$, onde I_1 e I_n são as informações de 1 e n valores independentes de x, respectivamente. Assim, obtém-se (como esperado) a mesma priori de Jeffreys independentemente de quantas observações fazemos subsequentemente.[1]

Exemplo 3.2 *(Amostra binomial.) Supõe-se que $X|\theta \sim Bin(n, \theta)$.* Então,

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

logo,

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

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

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

o que leva a,

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

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

[1] 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.

3.5 Mistura de prioris

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. 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 algum 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: $X|\theta \sim \text{Bin}(n,\theta)$ 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. Supunha $f_1(\theta),\ldots,f_n(\theta)$ são todas distribuições conjugadas para θ, levando a posterioris $f_1(\theta|x),\ldots,f_n(\theta|x)$. Considere a família de mistura de distribuições:

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

Então,

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

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.

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 Exercícios

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

Exercício 3.2 *Encontre a priori de Jeffreys para* θ *no modelo geométrico:*

$$f(x|\theta) = (1-\theta)^{x-1}\theta\,;\ x = 1,2,\ldots$$

para uma amostra $x_1,\ldots,x_n$.

Exercício 3.3 *Suponha que* X *tenha uma distribuição de Pareto* $\mathrm{Pa}(a,b)$, *em que* a *é conhecido mas* b *é desconhecido. Desta forma,*

$$f(x|b) = ba^b x^{-b-1}\,;\ (x > a).$$

Encontre a priori de Jeffreys e a correspondente distribuição posteriori para b.

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. Pode acontecer casos em que somente um parâmetro seja de interesse, mas usualmente há também 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 clássica. De fato, não existe absolutamente nenhuma nova teoria além do que já foi visto até o momento. Tem-se agora um vetor $\theta = (\theta_1, \ldots, \theta_d)$ de parâmetros sobre os quais deseja-se fazer inferência. Especifica-se uma distribuição priori (multivariada) $f(\theta)$ para θ que, assim como no caso de um parâmetro, é combinada com a verossimilhança $f(x|\theta)$ pelo Teorema de Bayes obtendo-se:

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

É 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 θ é obtida por cálculos de probabilidade diretamente na distribuição conjunta a posteriori. Por exemplo, a distribuição posteriori marginal de θ_1 é obtida "integrando-se fora" os outros componentes de θ. Assim,

$$f(\theta_1|x) = \int_{\theta_2} \ldots \int_{\theta_d} f(\theta|x)\mathrm{d}\theta_1 \ldots \mathrm{d}\theta_\mathrm{d}.$$

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

Especificação da priori. As prioris 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 avalir 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.

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 prioris conjugadas ainda mais valioso e cria 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 o tamanho do vetor θ. A estrutura da distribuição posteriori pode ser altamente complexa, e isto pode exigir considerável habilidade (e um computador com bons recursos gráficos) 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 inferências decorrem de regras elementares de probabilidades.

4.1 Exemplos

Exemplo 4.1 *Suponha que, uma determinada máquina seja satisfatória ($x = 1$) ou insatisfatória ($x = 2$). A probabilidade da máquina ser satisfatória 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 $x = 1$ são apresentadas na tabela 4.1. Além disto, a distribuição priori conjunta para (θ_1, θ_2) é dada na tabela 4.2.*

Tabela 4.1: Probabilidades condicionais da máquina ser satisfatória.

$\Pr(x = 1 \mid \theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	0,6	0,8
$\theta_2 = 1$	0,7	0,6

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

Tabela 4.2: Priori das condições da sala.

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

Tabela 4.3: Posteriori das condições da sala

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

Assim, somando-se ao longo das margens, obtemos a distribuição posteriori marginal:

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

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 α e β. Agora suponha que a informação priori para α e β pode ser expressa como: $\alpha \sim Be(p,q)$ e $\beta \sim Ga(p+q,1)$ com α e β 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:

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

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

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 α e β 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, $\alpha|y_1,y_2 \sim \text{Be}(y_1+p, y_2+q)$ 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 $X_1,\ldots,X_n$ um conjunto de variáveis independentes com distribuição normal $\text{N}(\theta,\phi)$, em que ambos, a média e a variância, θ e ϕ, respectivamente, são desconhecidos.*

Então,

$$f(x_1|\theta,\phi) = \frac{1}{\sqrt{2\pi\phi}}\exp\left\{-\frac{(x_i-\theta)^2}{2\phi}\right\},$$

e a verossimilhança é da forma:

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

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

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, e considera-se aqui o caso mais simples onde a

priori de referência que corresponde a expressar a ignorância, é usada. Em particular, toma-se:

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

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

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

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

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

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

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

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

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

então,

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

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

Similarmente,

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

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

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:

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

e

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

4.2 Exercícios

4.2.1 Exercícios

Exercício 4.1 *A qualidade de um componente elétrico ou é excelente ($x = 1$), bom ($x = 2$) ou ruim ($x = 3$). 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 $x =$ 3 são dadas pela Tabela 4.4. Além disso, distribuição conjunta de (θ_1, θ_2) é dada na Tabela 4.5. Encontre a distribuição posteriori conjunta de $\theta_1, \theta_2 | x = 3$ e cada uma das distribuições marginais. Tendo observado $x = 3$ qual a combinação fábrica/máquina é a mais provável de ter produzido esse componente?*

Tabela 4.4: Probabilidade condicional $x = 3$

$\Pr(x = 3 \mid \theta_1, \theta_2)$	$\theta_1 = 0$	$\theta_1 = 1$
$\theta_2 = 0$	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 Θ 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 θ.

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.*

	Reação			
Imune	Desprezível	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 ter vacinar uma pessoa imune.

Então, nesse caso particular, temos:

1) O espaço paramétrico $\Theta = \{\theta_1, \theta_2\}$, em que θ_1 e θ_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.

$L(\theta, a)$	θ_1	θ_2
a_1	8	0
a_2	0	20

Tabela 5.2: Função perda

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

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

$$\rho(a, x) = \int L(\theta, a)\, p(\theta|x) \mathrm{d}\theta.$$

Assim, tendo observado x, 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 $d(x)$ é a ação que minimiza a perda esperada posteriori.

		x: Reação			
$p(\theta)$	$p(x\|\theta)$	Desprezível	Leve	Moderada	Forte
0,6	θ_1	0,35	0,30	0,21	0,14
0,4	θ_2	0,09	0,17	0,25	0,49
$p(\theta,x)$	θ_1	0,210	0,180	0,126	0,084
	θ_2	0,036	0,068	0,100	0,196
	$p(x)$	0,246	0,248	0,226	0,280
$p(\theta\|x)$	θ_1	210/246	180/248	126/226	84/280
	θ_2	36/246	68/248	100/226	196/280
$\rho(a,x)$	a_1	1680/246	1440/248	1008/226	672/280
	a_2	720/246	1360/248	2000/226	3920/280
	$d(x)$	a_2	a_2	a_1	a_1

Tabela 5.3: Tabulação dos cálculos da análise de decisão.

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

$$BR(d) = \int \rho(d(x),x)\,p(x)\mathrm{d}x$$

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 observado, a decisão é para vacinar. O risco de Bayes associado a esta estratégia é

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

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 θ. 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 θ por a.

Há uma grande variedade de funções naturais de perda que poderíamos usar, e a escolha particular para qualquer problema especificado vai depender do contexto. Os mais usados são:

1) Perda quadrática: $L(\theta, a) = (\theta - a)^2$;

2) Perda erro absoluto: $L(\theta, a) = |\theta - a|$;

3) Perda $0--1$:

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

4) Perda linear: Para $g, h > 0$

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

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 θ para aquela particular escolha da função de perda.

5.2.1 Perda quadrática

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

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

uma vez que a integral do meio é zero. Isto é, a perda é minimizada quando $a = E(\theta, x)$.

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

5.2.2 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,

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

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

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

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

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

de forma que,

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

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, $\rho(a,x) \leq \rho(q,x) \forall a$, 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. Poranto, para perda em erro absoluto, o estimador de Bayes é a mediana da distribuição posteriori.

5.2.3 Perda 0-1

Claramente, neste caso:

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

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

5.2.4 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 $C_\alpha(x)$ é uma *região de credibilidade* $100(1-\alpha)\%$ para θ se

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

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

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:

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

em que γ é escolhido para garantir que:

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

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 posterioris univariadas padrão, haja valores tabulados para diversos α. De passagem, note que que ocorre o usual "perde/ganha" ao se escolher um α apropriado: um valor pequeno de α resultará em um intervalo largo; enquanto que um valor grande de α resultará em um intervalo para o qual o parâmetro tem uma baixa probabilidade de a ele pertencer.

Exemplo 5.2 *(Média normal.) Sejam $X_1,\ldots,X_n$ variáveis independentes de uma distribuição $N(\theta,\sigma^2)$, (σ^2 conhecido) com uma priori para θ da forma $\theta \sim N(b,d^2)$.*

Com essas informações, obtém-se a posteriori:

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

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

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

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 $d \to \infty$ o intervalo se torna:

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

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

Exemplo 5.3 *Sejam $X_1,\ldots,X_n$ variáveis independentes com distribuição $N(\theta,\phi)$ (ϕ desconhecido) e assumimos que a existe uma priori da forma:*

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

Isso leva a distribuições posterioris marginais:

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

e

$$S^2/\phi \sim \chi^2_{n-1}\,.$$

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

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

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 ϕ é ligeiramente mais problemático. Como $S^2/\phi \sim \chi^2_{n-1}$, segue-se que $\phi/S^2 \sim \chi^{-2}_{n-1}$, a chamada distribuição de qui-quadrado inversa. Pontos críticos de intervalos de maior densidade posteriori para uma variedade de valores de α estão disponíveis em tabelas.

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

Assim, temos a seguinte distribuição posteriori

$$\theta|x \sim Beta(p+x, q+n-x)\,.$$

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

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

e

$$\frac{1}{\text{Be}(p+x,q+n-x)}a^{p+x-1}(1-a)^{q+n-x-1} =$$
$$= \frac{1}{\text{Be}(p+x,q+n-x)}b^{p+x-1}(1-b)^{q+n-x-1} = \gamma$$

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

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 Ω_1 e Ω_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*:

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

Grandes valores de λ indicam que é mais provável que os dados observados x tenham ocorrido se θ_1 é o verdadeiro valor de θ do que teria com θ_0.

Na visão bayesiana, devemos também levar em consideração a informação prévia que temos sobre θ. A abordagem natural é basear as considerações do teste nas probabilidades posterioris relativas dos valores hipotetizados. Assim:

$$\lambda_B = \frac{f(\theta_1|x)}{f(\theta_0|x)} = \frac{f(\theta_1)f(x|\theta_1)}{f(\theta_0)f(x|\theta_0)}. \tag{5.1}$$

Esta quantidade é usualmente chamado da "chances à posteriori" (*posterior odds*). Observe, em particular, que não há nenhuma exigência para calcular os fatores de normalização já que o mesmo fator que aparece no numerador e denominador e cancela-se. Mais uma vez, altos valores de λ_B seriam evidências a favor de H_1.

Existe um conceito relacionado a este conhecido como *fator de Bayes*. Podemos ver a partir da equação 5.1 que as probabilidades posterioris são o produto das prioris pela razão de verossimilhança. Neste contexto, a razão de verossimilhanças é definida como sendo o fator de Bayes. O propósito ao focar no fator de Bayesé o de que esta é uma medida do peso da informação contida nos dados a favor de H_1 sobre H_0. Se o fator de Bayes é suficientemente grande, então ele vai superar qualquer preferência que possa haver à priori para H_0 de forma que a preferência posteriori pode ser passar a ser para H_1.

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 Exercícios

Exercício 5.1 *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 5.2 *Para o problema da média da Normal, descrito no Exemplo 2.4, encontre a estimativa pontual de θ usando cada uma das quatro funções perda descritas nesse capítulo.*

Exercício 5.3 *Repita o exercício anterior para a análise conjugada da distribuição Binomial com paramêtro θ desconhecido.*

Exercício 5.4 *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.

Capítulo 6

Predição

6.1 Distribuição preditiva

Até agora, focou-se na estimativa ou inferência sobre os parâmetros. 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 estimativas de parâmetros na forma de uma distribuição posteriori. Usualmente, 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 teoria clássica correspondente. 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". Ou seja, apenas a segunda fonte de incerteza está incluída na análise, levando à predições que se julgam-se ser mais precisas do que realmente são. 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 sobre a incerteza acerca das estimativas dos parâmetros, informação que é completamente fornecida pela posteriori.

Suponha que tem-se observações anteriores $x = (x_1, \ldots, x_n)$ de uma variável com função de densidade (ou verossimilhança) $p(x|\theta)$ e que deseja-se fazer inferências sobre a distribuição de um valor futuro y a partir deste processo. Com a especificação de uma distribuição a priori $p(\theta)$, o teorema de Bayes leva à uma distribuição posteriori $p(\theta|x)$. Em seguida, a "função de densidade preditiva" de y dado x é obtida por:

$$f(y|x) = \int f(y|\theta) f(\theta|x) \mathrm{d}\theta.$$

A densidade preditiva é portanto a integral da verossimilhança (de uma única observação) multiplicada pela posteriori, ou seja, a verossimilhança ponderada pela posteriori. Novamente, é importante notar que esta definição decorre simplesmente das leis usuais de manipulação de probabilidades e, da 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 θ e basear a inferência/predição na distribuição $f(y|\hat{\theta})$, a distribuição "estimada"[1]. Para enfatizar mais uma vez, isso desconsidera a variabilidade (incerteza) incorridos como resultado de se utilizar um valor estimado de θ, dando assim uma falsa sensação de precisão uma vez que a densidade preditiva $f(y|x)$ é usualmente mais variável por ponderar as predições ao longo da distribuição posteriori de θ.

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* $X|\theta \sim \mathrm{Bin}(n,\theta)$ *e a priori (conjugada) é* $\theta \sim \mathrm{Be}(p,q)$. *Como já visto anteriormente, a posteriori para* θ *é dada por:*

$$[\theta|x] \sim \mathrm{Be}(p+x, q+n-x)$$

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* $z|\theta \sim \mathrm{Bin}(N,\theta)$. Portanto a verossimilhança para a observação futura é

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

[1] 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 θ

Então, para $z = 0, 1, \ldots, N$,

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

Esta é expressão é reconhecida como a densidade de uma distribuição Beta-Binomial[2].

Exemplo 6.2 *(Amostra Gama.) Como no capítulo 2, suponha $X_1, \ldots, X_n$ são variáveis independentes com distribuição $X|\theta \sim \mathrm{Ga}(k, \theta)$, em que k é conhecido. Adota-se a priori conjugada $\theta \sim \mathrm{Ga}(p, q)$*

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

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

A verossimilhança para uma observação futura y é

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

e então a distribuição preditiva tem a forma

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

6.3 Exercícios

Exercício 6.1 *Uma amostra aleatória $x_1, \ldots, x_n$ é observada de uma variável aleatória com distribuição $[X|\theta] \sim \mathrm{Po}(\theta)$. A priori é $\theta \sim \mathrm{Ga}(g, h)$. Mostrar que a distribuição preditiva para uma observação futura, y, desta distribuição $X|\theta \sim \mathrm{Po}(\theta)$ é:*

$$P(y|x) = \binom{y+G-1}{G-1} \left(\frac{1}{1+H}\right)^y \left(1 - \frac{1}{1+H}\right)^G; \qquad y = 0, 1, \ldots$$

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 $\mathrm{N}(\theta; 4)$. O gerente de produção acredita que θ varia de lote para lote, de acordo com uma*

[2] Lembrando que a função beta é definida por $\mathrm{B}(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} \mathrm{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

distribuição $\mathrm{N}(110;0{,}4)$. *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* θ.
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* $m \to \infty$*?*

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* $\mathrm{Po}(l\theta)$. *Pouco se sabe sobre* θ. *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 θ da distribuição normal, com $X_1, \ldots, X_n \sim \mathrm{N}(\theta, \tau^{-1})$, com uma priori $\theta \sim \mathrm{N}(b, c^{-1})$ obteve-se

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

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

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

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 $\bar{x}$, o qual, pela forte lei dos grandes números, converge para o verdadeiro valor de θ. Estes argumentos são formalizados e generalizados como se segue.

7.2 Consistência

Se o valor real de θ é θ_0 e a probabilidade a priori de θ_0 (ou, no caso contínuo, de uma vizinhança arbitrária de θ_0) não é igual a zero, então, com quantidades crescentes de dados x, a probabilidade à posteriori de que $\theta = \theta_0$ (ou em uma vizinhança de θ_0) tende à unidade. Isto é provado como se segue.

Seja $x_1, \ldots, x_n$ observações iid, cada uma com distribuição $g(x|\theta)$. Então a den-

sidade posteriori é

$$\begin{aligned} f(\theta|x_1,\ldots,x_n) &\propto f(\theta)\prod_{i=1}^{n} g(x_i|\theta) \\ &= f(\theta)\exp\left\{\sum_{i=1}^{n}\log g(x_i|\theta)\right\} \\ &= f(\theta)\exp\{nL_n(\theta)\}. \end{aligned}$$

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

$$\bar{L}(\theta) = \int \{\log g(x|\theta)\} g(x|\theta_0)\mathrm{d}x$$

Pode-se mostrar que a expressão é maximizada quando $\theta = \theta_0$. Assim, para $\theta \neq \theta_0$, segue-se que a razão $\exp\{nL_n(\theta_0,x)\}/\exp\{nL_n(\theta,x)\} \Rightarrow \infty$ com probabilidade 1 quando $n \to \infty$. 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 θ, eventualmente a probabilidade à posteriori se concentrará no verdadeiro valor.

7.3 Normalidade assintótica

Quando θ 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, $\exp\{nL_n(\theta_0,x)\}/\exp\{nL_n(\theta,x)\}$ é desprezível exceto em uma vizinhança θ_0 gradativamente menor. Desta forma, $f(\theta)$ pode ser considerado constante nesse intervalo e obtém-se:

$$f(\theta|x_1,\ldots,x_n) \propto \exp\{n\bar{L}(\theta)\}$$

Além disto, expandindo-se $\bar{L}(\theta)$ ao redor de θ_0 por uma série de Taylor,

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

em que,

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

a recíproca da informação para θ_0. Assim, finalmente, obtém-se a aproximação

$$f(\theta|x_1,\ldots,x_n) \propto \exp\{-(\theta-\theta_0)^2\}/(2v),$$

isto é,

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

Então, com $n \to \infty$ a distribuição posteriori é aproximadamente normal, ao redor do verdadeiro valor de θ, 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 θ.

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 da posteriori por uma normal segue os seguintes passos:

(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 $\mathrm{N}(\hat{\theta}\,;\, -H^{-1}(\hat{\theta}))$

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

7.4 Exemplos

Exemplo 7.1 *(Média da Normal.) Seja $X_1, \ldots, X_n$ um conjunto de variáveis independentes com distribuição $\mathrm{N}(\theta, \sigma^2)$, com σ^2 conhecido.*

De forma usual, a verossimilhança é dada por:

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

Assim, pode-se tomar

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

e então,

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

e

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

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

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

O resultado é válido para qualquer distribuição priori com probabilidadades não nulas na vizinhança do verdadeiro valor de θ.

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

Então,

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0,\ldots,n.$$

Assim,

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

Desta forma,

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

e

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

Consequentemente,

$$\hat{\theta} = \frac{x}{n},$$
$$I_n(\hat{\theta}) = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Então, quando $n \to \infty$,

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

7.5 Exercícios

Exercício 7.1 *Encontre a distribuição assintótica posteriori para* θ *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 $x_1, x_2, \ldots x_n$ tomadas de v.a. independentes, cada uma com distribuição* $\mathrm{N}(\theta_i, 1)$ *com* $\theta^T = (\theta_1, \theta_2, \ldots, \theta_n)$. *A priori assumida para θ é tal que os θ_i são considerados independentes com distribuição* $\mathrm{N}(\mu, 1)$, *para um hiperparâmetro desconhecido μ.*

Na análise bayesiana completa, o valor de μ 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 μ. O estimador óbvio é a média amostral $\bar{x}$, com o qual a priori passa a ser $\mathrm{N}(\bar{x}, 1)$, o que então, da forma usual, leva à posterioris

$$\theta_i | x \sim \mathrm{N}\left(\frac{x_i + \bar{x}}{2}, \frac{1}{2}\right),$$

independentes para cada θ_i. Desta forma, a estimativa dada pela moda da posteriori de cada θ_i passa a ser $\hat{\theta}_i = (x_i + \bar{x})/2$. É interessante notar que as estimativas de máxima verossimilhança seriam $\hat{\theta}_i = x_i$ 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 posterioris que são também razoavelmente similares. Isto é, 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 posterioris 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 talvez mais utilizada é o *amostrador de Gibbs* cujo algorítmo é descrito a seguir..

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

1. Simular θ_1 de $[\theta_1|x, \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)}]$.
2. Trocar $\theta_1^{(0)}$ por θ_1.
3. Simular θ_2 de $[\theta_2|x, \theta_1^{(0)}, \theta_3^{(0)}, \ldots, \theta_d^{(0)}]$.
4. Trocar $\theta_2^{(0)}$ por θ_2.
5. Continuar ciclando o algorítimo para $\theta_1, \theta_2, \ldots, \theta_d$, um grande número de vezes.

Sob condições bastante gerais a sequência de θ's obtidos desta maneira forma uma cadeia de Markov (dado o valor atual, o valor subsequente de θ 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 e uma diversidade de técnicas de simulação. Em notação, o algorítmo consiste em simular de cada distribuição condicional $[\theta_i|x, \theta_{-i}^{(0)}]$ em que θ_i é o i-ésimo elemento do vetor de parâmetros que está sendo simulado, x 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 de diretamente de $[\theta_i|x, \theta_{-i}^{(0)}]$, pois a distribuição pode não tem forma analítica, aproximações e/ou outros métodos mais gerais de MCMC são necessários.

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. Os dados referem-se 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. Um gráfico dos dados é mostrado na Figura 8.1.

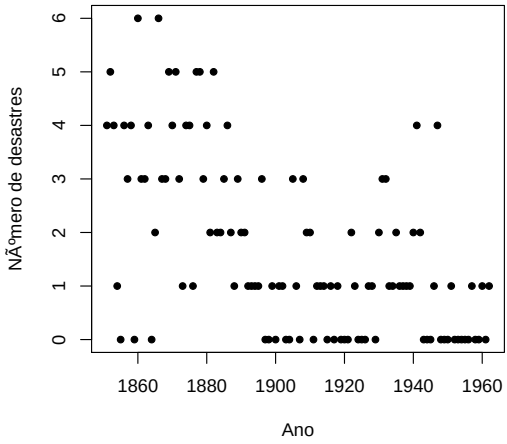

Figura 8.1: Número de desastres por ano em minas de carvão.

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							

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

$$Y_i|\theta \sim \text{Po}(\theta) \ ; \ i = 1, \ldots, k$$
$$Y_i|\lambda \sim \text{Po}(\lambda) \ ; \ i = k+1, \ldots, n,$$

ou seja, o modelo é para um número anual de desastres segundo a distribuição de Poisson, com uma taxa média de θ até o k-ésimo ano (considerado como o ano da possível mudança), e uma taxa média de λ 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

prioris 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 *hiperprioris*, 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 prioris leva às distribuições condicionais a seguir.

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

e

$$p[k | Y, \lambda, \theta, b_1, b_2] = \frac{L(Y; k, \theta, \lambda)}{\sum_{j=1}^{n} L(Y; j, \theta, \lambda)}$$

com

$$L(Y; k, \theta, \lambda) = \exp\{k(\lambda - \theta)\} (\theta / \lambda)^{\sum_{i=1}^{k} Y_i} .$$

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 sequências de valores para θ, λ e k, simulados em 1100 iterações do algorítmo. As prioris e hiperprioris foram especificadas com $a_1 = a_2 = 0{,}5$, $c_1 = c_2 = 0$, $d_1 = d_2 = 1$.

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.

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 θ e λ são fornecidas nos gráficos à esquerda e no centro da Figura 8.4.

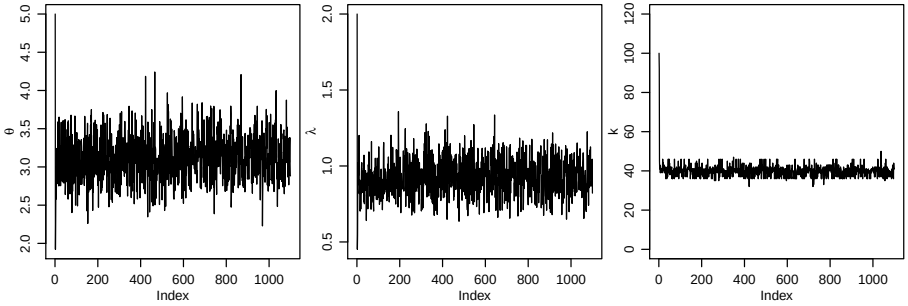

Figura 8.2: Cadeias geradas para os parâmetros θ, λ e k.

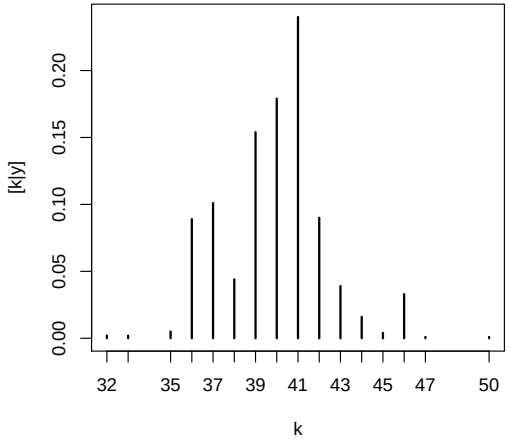

Figura 8.3: Distribuição a posteriori de k.

É claro por estes gráficos que λ é quase certamente menor que θ, 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, $[(\theta - \lambda)|x]$. 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 $[(\theta - \lambda)|x]$. 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 es-

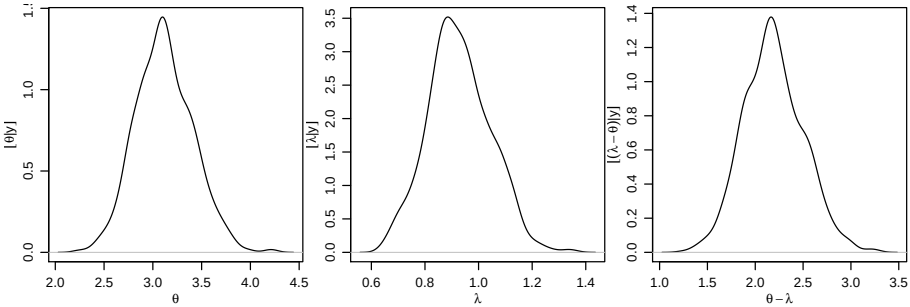

Figura 8.4: Posterioris de θ, λ e de $(\theta - \lambda)$.

timativa suavizada de $[(\theta - \lambda)|x]$ 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).

8.5 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 os distribuições condicionais e implemente computacionalmente o amostrador de Gibbs. Compare os resultados (distribuição posteriori e resumos pontuais e intervalares) por simulação com os obtidos analíticamente.*

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. **Statistical Inference: An Integrated Approach**. Arnold, 1999.

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.