Introdução a Inferência Bayesiana

Ricardo S. Ehlers

Versão Revisada em junho de 2003

Sumário

Capítulo 1

Introdução

A *informação* que se tem sobre uma quantidade de interesse θ é fundamental na Estatística. O verdadeiro valor de θ é desconhecido e a idéia é tentar reduzir este desconhecimento. Além disso, a intensidade da incerteza a respeito de θ pode assumir diferentes graus. Do ponto de vista Bayesiano, estes diferentes graus de incerteza são representados através de *modelos probabilísticos* para θ. Neste contexto, é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre θ (especificando modelos distintos). Sendo assim, não existe nenhuma distinção entre quantidades observáveis e os parâmetros de um modelo estatístico, todos são considerados quantidades aleatórias.

1.1 Teorema de Bayes

Considere uma quantidade de interesse desconhecida θ (tipicamente não observável). A informação de que dispomos sobre θ, resumida probabilisticamente através de $p(\theta)$, pode ser aumentada observando-se uma quantidade aleatória X relacionada com θ. A distribuição amostral $p(x|\theta)$ define esta relação. A idéia de que após observar $X = x$ a quantidade de informação sobre θ aumenta é bastante intuitiva e o teorema de Bayes é a regra de atualização utilizada para quantificar este aumento de informação,

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d\theta}. \tag{1.1}$$

Note que $1/p(x)$, que não depende de θ, funciona como uma constante normalizadora de $p(\theta|x)$.

Para um valor fixo de x, a função $l(\theta; x) = p(x|\theta)$ fornece a *plausibilidade* ou *verossimilhança* de cada um dos possíveis valores de θ enquanto $p(\theta)$ é chamada distribuição *a priori* de θ. Estas duas fontes de informação, priori e verossimilhança, são combinadas levando à distribuição *a posteriori* de θ, $p(\theta|x)$. Assim, a forma usual do teorema de Bayes é

$$p(\theta|x) \propto l(\theta;x)p(\theta). \tag{1.2}$$

Em palavras temos que

distribuição a posteriori $\propto$ verossimilhança $\times$ distribuição a priori.

Note que, ao omitir o termo $p(x)$, a igualdade em (1.1) foi substituída por uma proporcionalidade. Esta forma simplificada do teorema de Bayes será útil em problemas que envolvam estimação de parâmetros já que o denominador é apenas uma constante normalizadora. Em outras situações, como seleção de modelos, este termo tem um papel crucial.

É intuitivo também que a probabilidade a posteriori de um particular conjunto de valores de θ será pequena se $p(\theta)$ ou $l(\theta;x)$ for pequena para este conjunto. Em particular, se atribuirmos probabilidade a priori igual a zero para um conjunto de valores de θ então a probabilidade a posteriori será zero qualquer que seja a amostra observada.

A constante normalizadora da posteriori pode ser facilmente recuperada pois $p(\theta|x) = kp(x|\theta)p(\theta)$ onde

$$k^{-1} = \int p(x|\theta)p(\theta)d\theta = E_\theta[p(X|\theta)] = p(x)$$

chamada distribuição *preditiva*. Esta é a distribuição esperada para a observação x dado θ. Assim,

- Antes de observar X podemos checar a adequação da priori fazendo predições via $p(x)$.
- Se X observado recebia pouca probabilidade preditiva então o modelo deve ser questionado.

Se, após observar $X = x$, estamos interessados na previsão de uma quantidade Y, também relacionada com θ, e descrita probabilisticamente por $p(y|\theta)$ então

$$\begin{aligned} p(y|x) = \int p(y,\theta|x)d\theta &= \int p(y|\theta,x)p(\theta|x)d\theta \\ &= \int p(y|\theta)p(\theta|x)d\theta \end{aligned}$$

onde a última igualdade se deve a independência entre X e Y condicionado em θ. Esta hipótese de independência condicional está presente em muitos problemas estatísticos. Note que as previsões são sempre verificáveis uma vez que Y é uma quantidade observável. Finalmente, segue da última equação que

$$p(y|x) = E_{\theta|x}[p(Y|\theta)].$$

Fica claro também que os conceitos de *priori* e *posteriori* são relativos àquela observação que está sendo considerada no momento. Assim, $p(\theta|x)$ é a posteriori de θ em relação a X (que já foi observado) mas é a priori de θ em relação a Y (que não foi observado ainda). Após observar $Y = y$ uma nova posteriori (relativa a $X = x$ e $Y = y$) é obtida aplicando-se novamente o teorema de Bayes. Mas será que esta posteriori final depende da ordem em que as observações x e y foram processadas? Observando-se as quantidades $x_1, x_2, \cdots, x_n$, independentes dado θ e relacionadas a θ através de $p_i(x_i|\theta)$ segue que

$$\begin{aligned}
p(\theta|x_1) &\propto l_1(\theta; x_1)p(\theta) \\
p(\theta|x_2, x_1) &\propto l_2(\theta; x_2)p(\theta|x_1) \\
&\propto l_2(\theta; x_2)l_1(\theta; x_1)p(\theta) \\
\vdots \quad &\quad \vdots \\
p(\theta|x_n, x_{n-1}, \cdots, x_1) &\propto \left[\prod_{i=1}^{n} l_i(\theta; x_i)\right] p(\theta) \\
&\propto l_n(\theta; x_n)\, p(\theta|x_{n-1}, \cdots, x_1).
\end{aligned}$$

Ou seja, a ordem em que as observações são processadas pelo teorema de Bayes é irrelevante. Na verdade, elas podem até ser processadas em subgrupos.

Exemplo 1.1 : (Gamerman e Migon, 1993) Um médico, ao examinar uma pessoa, “desconfia” que ela possa ter uma certa doença. Baseado na sua experiência, no seu conhecimento sobre esta doença e nas informações dadas pelo paciente ele assume que a probabilidade do paciente ter a doença é 0,7. Aqui a quantidade de interesse desconhecida é o indicador de doença

$$\theta = \begin{cases} 1, & \text{se o paciente tem a doença} \\ 0, & \text{se o paciente não tem a doença} \end{cases}$$

Para aumentar sua quantidade de informação sobre a doença o médico aplica um teste X relacionado com θ através da distribuição

$$P(X = 1 \mid \theta = 0) = 0,40 \qquad \text{e} \qquad P(X = 1 \mid \theta = 1) = 0,95$$

e o resultado do teste foi positivo ($X = 1$).

É bem intuitivo que a probabilidade de doença deve ter aumentado após este resultado e a questão aqui é quantificar este aumento. Usando o teorema de Bayes segue que

$$P(\theta = 1 \mid X = 1) \propto l(\theta = 1; X = 1)p(\theta = 1) = (0,95)(0,7) = 0,665$$

$$P(\theta = 0 \mid X = 1) \propto l(\theta = 0; X = 1)p(\theta = 0) = (0,40)(0,3) = 0,120.$$

A constante normalizadora é tal que $P(\theta = 0 \mid X = 1) + P(\theta = 1 \mid X = 1) = 1$, i.e., $k(0,665) + k(0,120) = 1$ e $k = 1/0,785$. Portanto, a distribuição a posteriori de θ é

$$P(\theta = 1 \mid X = 1) = 0,665/0,785 = 0,847$$

$$P(\theta = 0 \mid X = 1) = 0,120/0,785 = 0,153.$$

O aumento na probabilidade de doença não foi muito grande porque a verossimilhança $l(\theta = 0; X = 1)$ também era grande (o modelo atribuia uma plausibilidade grande para $\theta = 0$ mesmo quando $X = 1$).

Agora o médico aplica outro teste Y cujo resultado está relacionado a θ através da seguinte distribuição

$$P(Y = 1 \mid \theta = 0) = 0,04 \qquad \text{e} \qquad P(Y = 1 \mid \theta = 1) = 0,99.$$

Mas antes de observar o resultado deste teste é interessante obter sua distribuição preditiva. Como θ é uma quantidade discreta segue que

$$p(y|x) = \sum_{\theta} p(y|\theta)p(\theta|x)$$

e note que $p(\theta|x)$ é a priori em relação a Y. Assim,

$$\begin{aligned}
P(Y = 1 \mid X = 1) &= P(Y = 1 \mid \theta = 0)P(\theta = 0 \mid X = 1) \\
&+ P(Y = 1 \mid \theta = 1)P(\theta = 1 \mid X = 1) \\
&= (0,04)(0,153) + (0,99)(0,847) = 0,845 \\
P(Y = 0 \mid X = 1) &= 1 - P(Y = 1 \mid X = 1) = 0,155.
\end{aligned}$$

O resultado deste teste foi negativo ($Y = 0$). Neste caso, é também intuitivo que a probabilidade de doença deve ter diminuido e esta redução será quantificada por uma nova aplicação do teorema de Bayes,

$$\begin{aligned}
P(\theta = 1 \mid X = 1, Y = 0) &\propto l(\theta = 1; Y = 0)P(\theta = 1 \mid X = 1) \\
&\propto (0,01)(0,847) = 0,0085 \\
P(\theta = 0 \mid X = 1, Y = 0) &\propto l(\theta = 0; Y = 0)P(\theta = 0 \mid X = 1) \\
&\propto (0,96)(0,153) = 0,1469.
\end{aligned}$$

A constante normalizadora é 1/(0,0085+0,1469)=1/0,1554 e assim a distribuição a posteriori de θ é

$$P(\theta = 1 \mid X = 1, Y = 0) = 0,0085/0,1554 = 0,055$$

$$P(\theta = 0 \mid X = 1, Y = 0) = 0,1469/0,1554 = 0,945.$$

Verifique como a probabilidade de doença se alterou ao longo do experimento

$$P(\theta = 1) = \begin{cases} 0,7, & \text{antes dos testes} \\ 0,847, & \text{após o teste } X \\ 0,055, & \text{após } X \text{ e } Y. \end{cases}$$

Note também que o valor observado de Y recebia pouca probabilidade preditiva. Isto pode levar o médico a repensar o modelo, i.e.,

(i) Será que $P(\theta = 1) = 0,7$ é uma priori adequada?

(ii) Será que as distribuições amostrais de X e Y estão corretas ? O teste X é tão inexpressivo e Y é realmente tão poderoso?

Um outro resultado importante ocorre quando se tem uma única observação da distribuição normal com média desconhecida. Se a média tiver priori normal então os parâmetros da posteriori são obtidos de uma forma bastante intuitiva.

Teorema 1.1 *Se $X|\theta \sim N(\theta, \sigma^2)$ com σ^2 conhecido e $\theta \sim N(\mu_0, \tau_0^2)$ então $\theta|x \sim N(\mu_1, \tau_1^2)$ onde*

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + \sigma^{-2}x}{\tau_0^{-2} + \sigma^{-2}} \quad e \quad \tau_1^{-2} = \tau_0^{-2} + \sigma^{-2}.$$

Note que, definindo *precisão* como o inverso da variância, segue do teorema que a precisão a posteriori é a soma das precisões a priori e da verossimilhança e não depende de x. Interpretando precisão como uma medida de informação e definindo $w = \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2}) \in (0,1)$ então w mede a informação relativa contida na priori com respeito à informação total. Podemos escrever então que

$$\mu_1 = w\mu_0 + (1 - w)x$$

ou seja, μ_1 é uma *combinação linear convexa* de μ_0 e x e portanto $\mu_0 \leq \mu_1 \leq x$.

Exemplo 1.2 : (Box & Tiao, 1992) Os físicos A e B desejam determinar uma constante física θ. O físico A tem mais experiência nesta área e especifica sua priori como $\theta \sim N(900, 20^2)$. O físico B tem pouca experiência e especifica uma priori muito mais incerta em relação à posição de θ, $\theta \sim N(800, 80^2)$. Assim, não é difícil verificar que

$$\text{para o físico } A\text{:} \quad P(860 < \theta < 940) \approx 0,95$$

$$\text{para o físico } B\text{:} \quad P(640 < \theta < 960) \approx 0,95.$$

Faz-se então uma medição X de θ em laboratório com um aparelho calibrado com distribuição amostral $X|\theta \sim N(\theta, 40^2)$ e observou-se $X = 850$. Aplicando o teorema 1.1 segue que

$$(\theta|X = 850) \sim N(890, 17, 9^2) \quad \text{para o físico } A$$
$$(\theta|X = 850) \sim N(840, 35, 7^2) \quad \text{para o físico } B.$$

Note também que os aumentos nas precisões a posteriori em relação às precisões a priori foram,

- para o físico A: precisão(θ) passou de $\tau_0^{-2} = 0,0025$ para $\tau_1^{-2} = 0,00312$ (aumento de 25%).
- para o físico B: precisão(θ) passou de $\tau_0^{-2} = 0,000156$ para $\tau_1^{-2} = 0,000781$ (aumento de 400%).

A situação está representada graficamente na Figura 1.1 a seguir. Note como a distribuição a posteriori representa um compromisso entre a distribuição a priori e a verossimilhança. Além disso, como as incertezas iniciais são bem diferentes o mesmo experimento fornece muito pouca informação adicional para o físico A enquanto que a incerteza do físico B foi bastante reduzida.

1.2 Princípio da Verossimilhança

O exemplo a seguir (DeGroot, 1970, páginas 165 e 166) ilustra esta propriedade. Imagine que cada item de uma população de itens manufaturados pode ser classificado como defeituoso ou não defeituoso. A proporção θ de itens defeituosos na população é desconhecida e uma amostra de itens será selecionada de acordo com um dos seguintes métodos:

(i) n itens serão selecionados ao acaso.

(ii) Itens serão selecionados ao acaso até que y defeituosos sejam obtidos.

(iii) Itens serão selecionados ao acaso até que o inspetor seja chamado para resolver um outro problema.

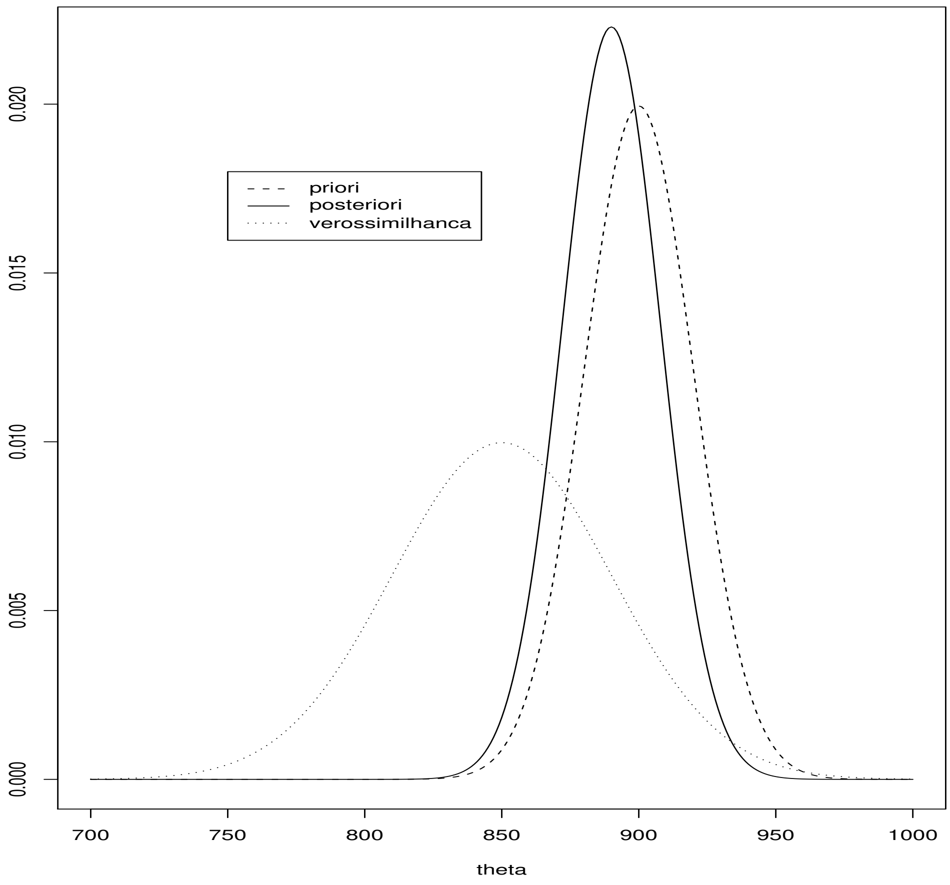
(iv) Itens serão selecionados ao acaso até que o inspetor decida que já acumulou informação suficiente sobre θ.

Qualquer que tenha sido o esquema amostral, se foram inspecionados n itens $x_1, \cdots, x_n$ dos quais y eram defeituosos então

$$l(\theta; x) \propto \theta^y (1-\theta)^{n-y}.$$

O Princípio da Verossimilhança postula que para fazer inferência sobre uma quantidade de interesse θ só importa aquilo que foi realmente observado e não aquilo que "poderia" ter ocorrido mas efetivamente não ocorreu.

Figura 1.1: Densidades a priori e a posteriori e função de verossimilhança para o exemplo 1.2.



1.3 Exercícios

1. No exemplo 1.2, obtenha também a distribuição preditiva de X e compare o valor observado com a média desta preditiva para os 2 físicos. Faça uma previsão para uma $2^{\underline{a}}$ medição Y feita com o mesmo aparelho.

2. Uma máquina produz 5% de itens defeituosos. Cada item produzido passa por um teste de qualidade que o classifica como "bom ", "defeituoso "ou "suspeito ". Este teste classifica 20% dos itens defeituosos como bons e 30% como suspeitos. Ele também classifica 15% dos itens bons como defeituosos e 25% como suspeitos.

    (a) Que proporção dos itens serão classificados como suspeitos ?

(b) Qual a probabilidade de um item classificado como suspeito ser defeituoso ?

(c) Outro teste, que classifica 95% dos itens defeituosos e 1% dos itens bons como defeituosos, é aplicado somente aos itens suspeitos.

(d) Que proporção de itens terão a suspeita de defeito confirmada ?

(e) Qual a probabilidade de um item reprovado neste 2º teste ser defeituoso ?

Capítulo 2

Distribuições a Priori

A utilização de informação a priori em inferência Bayesiana requer a especificação de uma distribuição a priori para a quantidade de interesse θ. Esta distribuição deve representar (probabilisticamente) o conhecimento que se tem sobre θ antes da realização do experimento. Neste capítulo serão discutidas diferentes formas de especificação da distribuição a priori.

2.1 Prioris Conjugadas

A partir do conhecimento que se tem sobre θ, pode-se definir uma família paramétrica de densidades. Neste caso, a distribuição a priori é representada por uma forma funcional, cujos parâmetros devem ser especificados de acordo com este conhecimento. Estes parâmetros indexadores da família de distribuições a priori são chamados de *hiperparâmetros* para distingui-los dos parâmetros de interesse θ.

Esta abordagem em geral facilita a análise e o caso mais importante é o de prioris conjugadas. A idéia é que as distribuições a priori e a posteriori pertençam a mesma classe de distribuições e assim a atualização do conhecimento que se tem de θ envolve apenas uma mudança nos hiperparâmetros. Neste caso, o aspecto sequencial do método Bayesiano pode ser explorado definindo-se apenas a regra de atualização dos hiperparâmetros já que as distribuições permanecem as mesmas.

Definição 2.1 *Se $F = \{p(x|\theta), \theta \in \Theta\}$ é uma classe de distribuições amostrais então uma classe de distribuições P é conjugada a F se*

$$\forall \; p(x|\theta) \in F \quad e \quad p(\theta) \in P \Rightarrow p(\theta|x) \in P.$$

Gamerman (1996, 1997 Cap. 2) alerta para o cuidado com a utilização indiscriminada de prioris conjugadas. Essencialmente, o problema é que a priori conjugada nem sempre é uma representação adequada da incerteza a priori. Sua utilização está muitas vezes associada à tratabilidade analítica decorrente.

Uma vez entendidas suas vantagens e desvantagens a questão que se coloca agora é "como" obter uma família de distribuições conjugadas.

(i) Identifique a classe P de distribuições para θ tal que $l(\theta;x)$ seja proporcional a um membro desta classe.

(ii) Verifique se P é *fechada por amostragem*, i.e., se $\forall\ p_1, p_2 \in P\ \exists\ k$ tal que $kp_1p_2 \in P$.

Se, além disso, existe uma constante k tal que $k^{-1} = \int l(\theta;x)d\theta < \infty$ e todo $p \in P$ é definido como $p(\theta) = k\ l(\theta;x)$ então P é a *família conjugada natural* ao modelo amostral gerador de $l(\theta;x)$.

Exemplo 2.1 : Sejam $X_1,\dots,X_n \sim \text{Bernoulli}(\theta)$. Então a densidade amostral conjunta é

$$p(\boldsymbol{x}|\theta) = \theta^t(1-\theta)^{n-t}, \quad 0 < \theta < 1 \quad \text{onde} \quad t = \sum_{i=1}^n x_i$$

e pelo teorema de Bayes segue que

$$p(\theta|\boldsymbol{x}) \propto \theta^t(1-\theta)^{n-t}p(\theta).$$

Note que $l(\theta;x)$ é proporcional à densidade de uma distribuição $\text{Beta}(t+1, n-t+1)$. Além disso, se p_1 e p_2 são as densidades das distribuições $\text{Beta}(a_1,b_1)$ e $\text{Beta}(a_2,b_2)$ então

$$p_1p_2 \propto \theta^{a_1+a_2-2}(1-\theta)^{b_1+b_2-2},$$

ou seja p_1p_2 é proporcional a densidade da distribuição $\text{Beta}(a_1+a_2-1, b_1+b_2-1)$. Conclui-se que a família de distribuições Beta com parâmetros inteiros é conjugada natural à família Bernoulli. Na prática esta classe pode ser ampliada para incluir todas as distribuições Beta, i.e. incluindo todos os valores positivos dos parâmetros.

2.2 Conjugação na Família Exponencial

A família exponencial inclui muitas das distribuições de probabilidade mais comumente utilizadas em Estatística, tanto contínuas quanto discretas. Uma característica essencial desta família é que existe uma estatística suficiente com dimensão fixa. Veremos adiante que a classe conjugada de distribuições é muito fácil de caracterizar.

Definição 2.2 *A família de distribuições com função de (densidade) de probabilidade $p(x|\theta)$ pertence à família exponencial a um parâmetro se podemos escrever*

$$p(x|\theta) = a(x)\exp\{u(x)\phi(\theta) + b(\theta)\}.$$

Note que pelo critério de fatoração de Neyman $U(x)$ é uma estatística suficiente para θ.

Neste caso, a classe conjugada é facilmente identificada como,

$$p(\theta) = k(\alpha,\beta)\exp\{\alpha\phi(\theta) + \beta b(\theta)\}.$$

e aplicando o teorema de Bayes segue que

$$p(\theta|x) = k(\alpha + u(x), \beta + 1)\exp\{[\alpha + u(x)]\phi(\theta) + [\beta + 1]b(\theta)\}.$$

Agora, usando a constante k, a distribuição preditiva pode ser facilmente obtida sem necessidade de qualquer integração. A partir da equação $p(x)p(\theta|x) = p(x|\theta)p(\theta)$ e após alguma simplificação segue que

$$p(x) = \frac{p(x|\theta)p(\theta)}{p(\theta|x)} = \frac{a(x)k(\alpha,\beta)}{k(\alpha + u(x), \beta + 1)}.$$

Exemplo 2.2 : Uma extensão direta do exemplo 2.1 é o modelo binomial, i.e. $X|\theta \sim \text{Binomial}(n,\theta)$. Neste caso,

$$p(x|\theta) = \binom{n}{x}\exp\left\{x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}$$

e a família conjugada natural é Beta(r,s). Podemos escrever então

$$\begin{aligned}
p(\theta) &\propto \theta^{r-1}(1-\theta)^{s-1} \\
&\propto \exp\left\{(r-1)\log\left(\frac{\theta}{1-\theta}\right) + \left(\frac{s+r-2}{n}\right)n\log(1-\theta)\right\} \\
&\propto \exp\left\{\alpha\phi(\theta) + \beta b(\theta)\right\}.
\end{aligned}$$

A posteriori também é Beta com parâmetros $\alpha + x$ e $\beta + 1$ ou equivalentemente $r + x$ e $s + n - x$, i.e.

$$\begin{aligned}
p(\theta|x) &\propto \exp\left\{(r+x-1)\phi(\theta) + \left[\frac{s+r-2+n}{n}\right]b(\theta)\right\} \\
&\propto \theta^{r+x-1}(1-\theta)^{s+n-x-1}.
\end{aligned}$$

Então distribuição preditiva é dada por

$$p(x) = \binom{n}{x}\frac{B(r+x, s+n-x)}{B(r,s)}, \quad x = 0,1,\ldots,n, \quad n \geq 1,$$

onde B^{-1} é a constante normalizadora da distribuição Beta. Esta distribuição é denominada Beta-Binomial.

No caso geral em que se tem uma amostra $X_1, \ldots, X_n$ da família exponencial a natureza sequencial do teorema de Bayes permite que a análise seja feita por replicações sucessivas. Assim a cada observação x_i os parâmetros da distribuição a posteriori são atualizados via

$$\begin{aligned} \alpha_i &= \alpha_{i-1} + u(x_i) \\ \beta_i &= \beta_{i-1} + 1 \end{aligned}$$

com $\alpha_0 = \alpha$ e $\beta_0 = \beta$. Após n observações temos que

$$\begin{aligned} \alpha_n &= \alpha + \sum_{i=1}^{n} u(x_i) \\ \beta_n &= \beta + n \end{aligned}$$

e a distribuição preditiva é dada por

$$p(\boldsymbol{x}) = \left[\prod_{i=1}^{n} a(x_i)\right] \frac{k(\alpha, \beta)}{k(\alpha + \sum u(x_i), \beta + n)}.$$

Finalmente, a definição de família exponencial pode ser extendida ao caso multiparamétrico, i.e.

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \left[\prod_{i=1}^{n} a(x_i)\right] \exp\left\{\sum_{j=1}^{r} \left[\sum_{i=1}^{n} u_j(x_i)\right] \phi_j(\theta) + nb(\theta)\right\}$$

onde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$. Neste caso, pelo critério de fatoração, temos que $\sum U_1(x_i), \ldots, \sum U_r(x_i)$ é uma estatística conjuntamente suficiente para o vetor de parâmetros $\boldsymbol{\theta}$.

2.3 Principais Famílias Conjugadas

Já vimos que a família de distribuições Beta é conjugada ao modelo Bernoulli e binomial. Não é difícil mostrar que o mesmo vale para as distribuições amostrais geométrica e binomial-negativa. A seguir veremos resultados para outros membros importantes da família exponencial.

2.3.1 Distribuição normal com variância conhecida

Para uma única observação vimos pelo teorema 1.1 que a família de distribuições normais é conjugada ao modelo normal. Para uma amostra de tamanho n, a função de verssimilhança pode ser escrita como

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

onde os termos que não dependem de θ foram incorporados à constante de proporcionalidade. Portanto, a verossimilhança tem a mesma forma daquela baseada em uma única observação bastando substituir x por $\overline{x}$ e σ^2 por σ^2/n. Logo vale o teorema 1.1 com as devidas substituições, i.e. a distribuição a posteriori de θ dado $\boldsymbol{x}$ é $N(\mu_1, \tau_1^2)$ onde

$$\mu_1 = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\overline{x}}{\tau_0^{-2} + n\sigma^{-2}} \quad \text{e} \quad \tau_1^{-2} = \tau_0^{-2} + n\sigma^{-2}.$$

2.3.2 Distribuição de Poisson

Seja $X_1,\dots,X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ. Sua função de probabilidade conjunta é dada por

$$p(x|\theta) = \frac{e^{-n\theta}\theta^t}{\prod x_i!} \propto e^{-n\theta}\theta^t, \quad \theta > 0, \quad t = \sum_{i=1}^{n} x_i.$$

O núcleo da verossimilhança é da forma $\theta^a e^{-b\theta}$ que caracteriza a família de distribuições Gama que é fechada por amostragem. Assim, a priori conjugada natural de θ é Gama com parâmetros positivos α e β, i.e.

$$p(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}, \quad \alpha,\beta > 0 \quad \theta > 0.$$

A densidade a posteriori fica

$$p(\theta|x) \propto \theta^{\alpha+t-1}\exp\{-(\beta+n)\theta\}$$

que corresponde à densidade Gama$(\alpha+t, \beta+n)$. A distribuição preditiva também é facilmente obtida pois

$$p(x|\theta) = \left[\prod_{i=1}^{n}\frac{1}{x_i!}\right]\exp\{t\theta - n\theta\}$$

e portanto

$$p(x) = \left[\prod_{i=1}^{n}\frac{1}{x_i!}\right]\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+t)}{(\beta+n)^{\alpha+t}}.$$

2.3.3 Distribuição multinomial

Denotando por $\boldsymbol{X} = (X_1, \ldots, X_p)$ o número de ocorrências em cada uma de p categorias em n ensaios independentes, e por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ as probabilidades associadas deseja-se fazer inferência sobre estes p parâmetros. No entanto, note que existem efetivamente $k-1$ parâmetros já que temos a seguinte restrição $\sum_{i=1}^p \theta_i = 1$. Além disso, a restrição $\sum_{i=1}^p X_i = n$ obviamente também se aplica. Dizemos que $\boldsymbol{X}$ tem distribuição multinomial com parâmetros n e $\boldsymbol{\theta}$ e função de probabilidade conjunta das p contagens $\boldsymbol{X}$ é dada por

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^p x_i!} \prod_{i=1}^p \theta_i^{x_i}.$$

Note que esta é uma generalização da distribuição binomial que apenas duas categorias. Não é difícil mostrar que esta distribuição também pertence à família exponencial. A função de verossimilhança para $\boldsymbol{\theta}$ é

$$l(\boldsymbol{\theta}; \boldsymbol{x}) \propto \prod_{i=1}^p \theta_i^{x_i}$$

que tem o mesmo núcleo da função de densidade de uma distribuição de Dirichlet. A família Dirichlet com parâmetros inteiros $a_1, \ldots, a_p$ é a conjugada natural do modelo multinomial, porém na prática a conjugação é extendida para parâmetros não inteiros. A distribuição a posteriori é dada por

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \prod_{i=1}^p \theta_i^{x_i} \prod_{i=1}^p \theta_i^{a_i-1} = \prod_{i=1}^p \theta_i^{x_i+a_i-1}.$$

Note que estamos generalizando a análise conjugada para amostras binomiais com priori beta.

2.3.4 Distribuição normal com média conhecida e variância desconhecida

Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ conhecido e $\phi = \sigma^{-2}$ desconhecido. Neste caso a função de densidade conjunta é dada por

$$p(\boldsymbol{x}|\theta, \phi) \propto \phi^{n/2} \exp\{-\frac{\phi}{2n} \sum_{i=1}^n (x_i - \theta)^2\}.$$

Note que o núcleo desta verossimilhança tem a mesma forma daquele de uma distribuição Gama. Como sabemos que a família Gama é fechada por amostragem podemos considerar uma distribuição a priori Gama com parâmetros $n_0/2$ e $n_0\sigma_0^2/2$, i.e.

$$\phi \sim Gama\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

Equivalentemente, podemos atribuir uma distribuição a priori qui-quadrado com n_0 graus de liberdade para $n_0\sigma_0^2\phi$. A forma funcional dos parâmetros da distribuição a priori é apenas uma conveniência matemática como veremos a seguir.

Definindo $ns_0^2 = \sum_{i=1}^{n}(x_i - \theta)^2$ e aplicando o teorema de Bayes obtemos a distribuição a posteriori de ϕ,

$$\begin{aligned} p(\phi|\boldsymbol{x}) &\propto \phi^{n/2}\exp\left\{-\frac{\phi}{2}ns_0^2\right\}\phi^{n_0/2-1}\exp\left\{-\frac{\phi}{2}n_0\sigma_0^2\right\} \\ &= \phi^{(n_0+n)/2-1}\exp\left\{-\frac{\phi}{2}(n_0\sigma_0^2+ns_0^2)\right\}. \end{aligned}$$

Note que esta expressão corresponde ao núcleo da distribuição Gama, como era esperado devido à conjugação. Portanto,

$$\phi|\boldsymbol{x} \sim Gama\left(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2+ns_0^2}{2}\right).$$

Equivalentemente podemos dizer que $(n_0\sigma_0^2 + ns_0^2)\phi \mid \boldsymbol{x} \sim \chi^2_{n_0+n}$.

2.3.5 Distribuição normal com média e variância desconhecidos

Seja $X_1, \dots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com ambos θ e σ^2 desconhecidos. Neste caso a distribuição a priori conjugada será especificada em dois estágios. No primeiro estágio,

$$\theta|\phi \sim N(\mu_0, (c_0\phi)^{-1}), \quad \phi = \sigma^{-2}$$

e a distribuição a priori marginal de ϕ é a mesma do caso anterior, i.e.

$$\phi \sim \text{Gama}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

A distribuição conjunta de (θ, ϕ) é geralmente chamada de Normal-Gama com parâmetros $(\mu_0, c_0, n_0, \sigma_0^2)$ e sua função de densidade conjunta é dada por,

$$\begin{aligned} p(\theta,\phi) &= p(\theta|\phi)p(\phi) \\ &\propto \phi^{1/2}\exp\left\{-\frac{c_0\phi}{2}(\theta-\mu_0)^2\right\}\phi^{n_0/2-1}\exp\left\{-\frac{n_0\sigma_0^2\phi}{2}\right\} \\ &= \phi^{(n_0+1)/2-1}\exp\left\{-\frac{\phi}{2}(n_0\sigma_0^2+c_0(\theta-\mu_0)^2)\right\}. \end{aligned}$$

A partir desta densidade conjunta podemos obter a distribuição marginal de θ por integração

$$\begin{aligned}
p(\theta) &= \int p(\theta|\phi)p(\phi)d\phi \\
&\propto \int_0^\infty \phi^{1/2}\exp\left\{-\frac{c_0\phi}{2}(\theta-\mu_0)^2\right\}\phi^{n_0/2-1}\exp\left\{-\frac{n_0\sigma_0^2}{2}\phi\right\}d\phi \\
&\propto \int_0^\infty \phi^{(n_0+1)/2-1}\exp\left\{-\frac{\phi}{2}[n_0\sigma_0^2+c_0(\theta-\mu_0)^2]\right\}d\phi \\
&\propto \left[\frac{n_0\sigma_0^2+c_0(\theta-\mu_0)^2}{2}\right]^{-\frac{n_0+1}{2}} \propto \left[1+\frac{(\theta-\mu_0)^2}{n_0(\sigma_0^2/c_0)}\right]^{-\frac{n_0+1}{2}},
\end{aligned}$$

que é o núcleo da distribuição t de Student com n_0 graus de liberdade, parâmetro de locação μ_0 e parâmetro de escala σ_0^2/c_0. Denotamos $\theta \sim t_{n_0}(\mu_0, \sigma_0^2/c_0)$. A distribuição condicional de ϕ dado θ também é facilmente obtida como

$$\begin{aligned}
p(\phi|\theta) &\propto p(\theta|\phi)p(\phi) \\
&\propto \phi^{(n_0+1)/2-1}\exp\left\{-\frac{\phi}{2}[n_0\sigma_0^2+c_0(\theta-\mu_0)^2]\right\},
\end{aligned}$$

e portanto,

$$\phi|\theta \sim \text{Gama}\left(\frac{n_0+1}{2}, \frac{n_0\sigma_0^2+c_0(\theta-\mu_0)^2}{2}\right).$$

A posteriori conjunta de (θ, ϕ) também é obtida em 2 etapas como segue. Primeiro, para ϕ fixo podemos usar o resultado da seção 2.3.1 de modo que a distribuição a posteriori de θ dado ϕ fica

$$\theta|\phi, \boldsymbol{x} \sim N(\mu_1, (c_1\phi)^{-1})$$

onde

$$\mu_1 = \frac{c_0\phi\mu_0 + n\phi\overline{x}}{c_0\phi + n\phi} = \frac{c_0\mu_0 + n\overline{x}}{c_0 + n} \quad \text{e} \quad c_1 = c_0 + n.$$

Na segunda etapa, combinando a verossimilhança com a priori de ϕ obtemos que

$$\phi|\boldsymbol{x} \sim \text{Gama}\left(\frac{n_1}{2}, \frac{n_1\sigma_1^2}{2}\right)$$

onde

$$n_1 = n_0 + n \quad \text{e} \quad n_1\sigma_1^2 = n_0\sigma_0^2 + \sum(x_i - \overline{x})^2 + c_0 n(\mu_0 - \overline{x})^2/(c_0 + n).$$

Equivalentemente, podemos escrever a posteriori de ϕ como $n_1\sigma_1^2\phi \sim \chi^2_{n_1}$. Assim, a posteriori conjunta é $(\theta, \phi|\boldsymbol{x}) \sim$ Normal-Gama$(\mu_1, c_1, n_1, \sigma_1^2)$ e portanto a posteriori marginal de θ fica

$$\theta \mid \boldsymbol{x} \sim t_{n_1}(\mu_1, \sigma_1^2/c_1).$$

Em muitas situações é mais fácil pensar em termos de algumas características da distribuição a priori do que em termos de seus hiperparâmetros. Por exemplo, se $E(\theta) = 2$, $Var(\theta) = 5$, $E(\phi) = 3$ e $Var(\phi) = 3$ então

(i) $\mu_0 = 2$ pois $E(\theta) = \mu_0$.

(ii) $\sigma_0^2 = 1/3$ pois $E(\phi) = 1/\sigma_0^2$.

(iii) $n_0 = 6$ pois $Var(\phi) = 2/(n_0\sigma_0^4) = 18/n_0$.

(iv) $c_0 = 1/10$ pois $Var(\theta) = \left(\dfrac{n_0}{n_0-2}\right)\dfrac{\sigma_0^2}{c_0} = \dfrac{1}{2c_0}$

2.4 Priori não Informativa

Esta seção refere-se a especificação de distribuições a priori quando se espera que a informação dos dados seja dominante, no sentido de que a nossa informação a priori é *vaga*. Os conceitos de "conhecimento vago", "não informação", ou "ignorância a priori" claramente não são únicos e o problema de caracterizar prioris com tais características pode se tornar bastante complexo.

Por outro lado, reconhece-se a necessidade de alguma forma de análise que, em algum sentido, consiga captar esta noção de uma priori que tenha um efeito mínimo, relativamente aos dados, na inferência final. Tal análise pode ser pensada como um ponto de partida quando não se consegue fazer uma elicitação detalhada do "verdadeiro" conhecimento a priori. Neste sentido, serão apresentadas aqui algumas formas de "como" fazer enquanto discussões mais detalhadas são encontradas em Berger (1985), Box e Tiao (1992), Bernardo e Smith (1994) e O'Hagan (1994).

A primeira idéia de "não informação" a priori que se pode ter é pensar em todos os possíveis valores de θ como igualmente prováveis, i.e., com uma distribuição a priori uniforme. Neste caso, fazendo $p(\theta) \propto k$ para θ variando em um subconjunto da reta significa que nenhum valor particular tem preferência (Bayes, 1763). Porém esta escolha de priori pode trazer algumas dificuldades técnicas

(i) Se o intervalo de variação de θ for ilimitado então a distribuição é imprópria, i.e.

$$\int p(\theta)d\theta = \infty.$$

(ii) Se $\phi = g(\theta)$ é uma reparametrização não linear monótona de θ então $p(\phi)$ é não uniforme já que pelo teorema de transformação de variáveis

$$p(\phi) = p(\theta(\phi))\left|\frac{d\theta}{d\phi}\right| \propto \left|\frac{d\theta}{d\phi}\right|.$$

Na prática, como estaremos interessados na distribuição a posteriori não daremos muita importância à impropriedade da distribuição a priori. No entanto devemos sempre nos certificar de que a posterior é própria para antes de fazer qualquer inferência.

A classe de prioris não informativas proposta por Jeffreys (1961) é invariante a transformações 1 a 1, embora em geral seja imprópria e será definida a seguir. Antes porém precisamos da definição da medida de informação de Fisher.

Definição 2.3 *Considere uma única observação X com função de (densidade) de probabilidade $p(x|\theta)$. A medida de informação esperada de Fisher de θ através de X é definida como*

$$I(\theta) = E\left[-\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\right]$$

Se $\boldsymbol{\theta}$ for um vetor paramétrico define-se então a matriz de informação esperada de Fisher de $\boldsymbol{\theta}$ através de X como

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right].$$

Note que o conceito de informação aqui está sendo associado a uma espécie de curvatura média da função de verossimilhança no sentido de que quanto maior a curvatura mais precisa é a informação contida na verossimilhança, ou equivalentemente maior o valor de $I(\theta)$. Em geral espera-se que a curvatura seja negativa e por isso seu valor é tomado com sinal trocado. Note também que a esperança matemática é tomada em relação à distribuição amostral $p(x|\theta)$.

Podemos considerar então $I(\theta)$ uma medida de informação global enquanto que uma medida de informação local é obtida quando não se toma o valor esperado na definição acima. A medida de informação observada de Fisher $J(\theta)$ fica então definida como

$$J(\theta) = -\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}$$

e que será utilizada mais adiante quando falarmos sobre estimação.

Definição 2.4 *Seja uma observação X com função de (densidade) de probabilidade $p(x|\theta)$. A priori não informativa de Jeffreys tem função de densidade dada por*

$$p(\theta) \propto [I(\theta)]^{1/2}.$$

Se $\boldsymbol{\theta}$ for um vetor paramétrico então $p(\theta) \propto |\det I(\theta)|^{1/2}$.

Exemplo 2.3 : Seja $X_1, \dots, X_n \sim \text{Poisson}(\theta)$. Então o logaritmo da função de probabilidade conjunta é dado por

$$\log p(\boldsymbol{x}|\theta) = -n\theta + \sum_{i=1}^{n} x_i \log \theta - \log \prod_{i=1}^{n} x_i!$$

$$\frac{\partial^2 \log p(x|\theta)}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left[-n + \frac{\sum_{i=1}^n x_i}{\theta}\right] = -\frac{\sum_{i=1}^n x_i}{\theta^2}$$

$$I(\theta) = \frac{1}{\theta^2} E\left[\sum_{i=1}^n x_i\right] = n/\theta \propto \theta^{-1}.$$

Portanto, a priori não informativa de Jeffreys para θ no modelo Poisson é $p(\theta) \propto \theta^{-1/2}$. Note que esta priori é obtida tomando-se a conjugada natural Gama(α, β) e fazendo-se $\alpha = 1/2$ e $\beta \to 0$.

Em geral a priori não informativa é obtida fazendo-se o parâmetro de escala da distribuição conjugada tender a zero e fixando-se os demais parâmetros convenientemente. Além disso, a priori de Jeffreys assume formas específicas em alguns modelos que são frequentemente utilizados como veremos a seguir.

Definição 2.5 *X tem um modelo de locação se existem uma função f e uma quantidade θ tais que $p(x|\theta) = f(x - \theta)$. Neste caso θ é chamado de parâmetro de locação.*

A definição vale também quando θ é um vetor de parâmetros. Alguns exemplos importantes são a distribuição normal com variância conhecida, e a distribuição normal multivariada com matriz de variância-covariância conhecida. Pode-se mostrar que para o modelo de locação a priori de Jeffreys é dada por $p(\theta) \propto$ constante.

Definição 2.6 *X tem um modelo de escala se existem uma função f e uma quantidade σ tais que $p(x|\sigma) = (1/\sigma)f(x/\sigma)$. Neste caso σ é chamado de parâmetro de escala.*

Alguns exemplos são a distribuição exponencial com parâmetro θ, com parâmetro de escala $\sigma = 1/\theta$, e a distribuição $N(\theta, \sigma^2)$ com média conhecida e escala σ. Pode-se mostrar que para o modelo de escala a priori de Jeffreys é dada por $p(\sigma) \propto \sigma^{-1}$.

Definição 2.7 *X tem um modelo de locação e escala se existem uma função f e as quantidades θ e σ tais que*

$$p(x|\theta, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \theta}{\sigma}\right).$$

Neste caso θ é chamado de parâmetro de locação e σ de parâmetro de escala.

Alguns exemplos são a distribuição normal (uni e multivariada) e a distribuição de Cauchy. Em modelos de locação e escala, a priori não informativa pode ser obtida assumindo-se independência a priori entre θ e σ de modo que $p(\theta, \sigma) = p(\theta)p(\sigma) \propto \sigma^{-1}$.

Exemplo 2.4 : Seja $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ com μ e σ^2 desconhecidos. Neste caso,

$$p(x|\mu,\sigma^2) \propto \frac{1}{\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

portanto (μ, σ) é parâmetro de locação-escala e $p(\mu, \sigma) \propto \sigma^{-1}$ é a priori não informativa. Então, pela propriedade da invariância, a priori não informativa para (μ, σ^2) no modelo normal é $p(\mu, \sigma^2) \propto \sigma^{-2}$.

Vale notar entretanto que a priori não informativa de Jeffreys viola o princípio da verossimilhança, já que a informação de Fisher depende da distribuição amostral.

2.5 Prioris Hierárquicas

A idéia aqui é dividir a especificação da distribuição a priori em estágios. Além de facilitar a especificação esta abordagem é natural em determinadas situações experimentais.

A distribuição a priori de θ depende dos valores dos hiperparâmetros ϕ e podemos escrever $p(\theta|\phi)$ ao invés de $p(\theta)$. Além disso, ao invés de fixar valores para os hiperparâmetros podemos especificar uma distribuição a priori $p(\phi)$ completando assim o segundo estágio na hierarquia. A distribuição a priori marginal de θ pode ser então obtida por integração como

$$p(\theta) = \int p(\theta, \phi) d\phi = \int p(\theta|\phi) p(\phi) d\phi.$$

Exemplo 2.5 : Sejam $X_1, \dots, X_n$ tais que $X_i \sim N(\theta_i, \sigma^2)$ com σ^2 conhecido e queremos especificar uma distribuição a priori para o vetor de parâmetros $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$. Suponha que no primeiro estágio assumimos que $\theta_i \sim N(\mu, \tau^2)$, $i = 1, \dots, n$. Neste caso, se fixarmos o valor de $\tau^2 = \tau_0^2$ e assumirmos que μ tem distribuição normal então $\boldsymbol{\theta}$ terá distribuição normal multivariada. Por outro lado, fixando um valor para $\mu = \mu_0$ e assumindo que τ^{-2} tem distribuição Gama implicará em uma distribuição t de Student multivariada para $\boldsymbol{\theta}$.

Teoricamente, não há limitação quanto ao número de estágios, mas devido às complexidades resultantes as prioris hierárquicas são especificadas em geral em 2 ou 3 estágios. Além disso, devido à dificuldade de interpretação dos hiperparâmetros em estágios mais altos é prática comum especificar prioris não informativas para este níveis.

Uma aplicação interessante do conceito de hierarquia é quando a informação a priori disponível só pode ser convenientemente resumida através de uma *mistura* de distribuições. Isto implica em considerar uma distribuição discreta para

ϕ de modo que

$$p(\theta) = \sum_{i=1}^{k} p(\theta|\phi_i)p(\phi_i).$$

Não é difícil verificar que a distribuição a posteriori de θ é também uma mistura com veremos a seguir. Aplicando o teorema de Bayes temos que,

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\displaystyle\int p(\theta)p(x|\theta)d\theta} = \frac{\displaystyle\sum_{i=1}^{k} p(x|\theta)p(\theta|\phi_i)p(\phi_i)}{\displaystyle\sum_{i=1}^{k} p(\phi_i)\int p(x|\theta)p(\theta|\phi_i)d\theta}.$$

Mas note que a posteriori condicional de θ dado ϕ_i é

$$p(\theta|x,\phi_i) = \frac{p(x|\theta)p(\theta|\phi_i)}{\displaystyle\int p(x|\theta)p(\theta|\phi_i)d\theta} = \frac{p(x|\theta)p(\theta|\phi_i)}{m(x|\phi_i)}.$$

Assim, podemos escrever a posteriori de θ como

$$p(\theta|x) = \frac{\displaystyle\sum_{i=1}^{k} p(\theta|x,\phi_i)m(x|\phi_i)p(\phi_i)}{\displaystyle\sum_{i=1}^{k} m(x|\phi_i)p(\phi_i)} = \sum_{i=1}^{k} p(\theta|x,\phi_i)p(\phi_i|x)$$

Note também que $p(x) = \sum m(x|\phi_i)p(\phi_i)$, isto é a distribuição preditiva, é uma mistura de preditivas condicionais.

Exemplo 2.6 : Se $\theta \in (0,1)$, a família de distribuições a priori Beta é conveniente. Mas estas são sempre unimodais e assimétricas à esquerda ou à direita. Outras formas interessantes, e mais de acordo com a nossa informação a priori, podem ser obtidas misturando-se 2 ou 3 elementos desta família. Por exemplo,

$$\theta \sim 0,25\text{Beta}(3,8) + 0,75\text{Beta}(8,3)$$

representa a informação a priori de que $\theta \in (0,5;0,95)$ com alta probabilidade (0,71) mas também que $\theta \in (0,1;0,4)$ com probabilidade moderada (0,20). As modas desta distribuição são 0,23 e 0,78. Por outro lado

$$\theta \sim 0,33\text{Beta}(4,10) + 0,33\text{Beta}(15,28) + 0,33\text{Beta}(50,70)$$

representa a informação a priori de que $\theta > 0,6$ com probabilidade desprezível. Estas densidades estão representadas graficamente nas Figuras 2.1 e 2.2 a seguir. Note que a primeira mistura deu origem a uma distribuição a priori bimodal enquanto a segunda originou uma priori assimétrica à esquerda com média igual a 0,35.

Figura 2.1: Mistura de funções de densidade Beta(3,8) e Beta(8,3) com pesos 0,25 e 0,75.

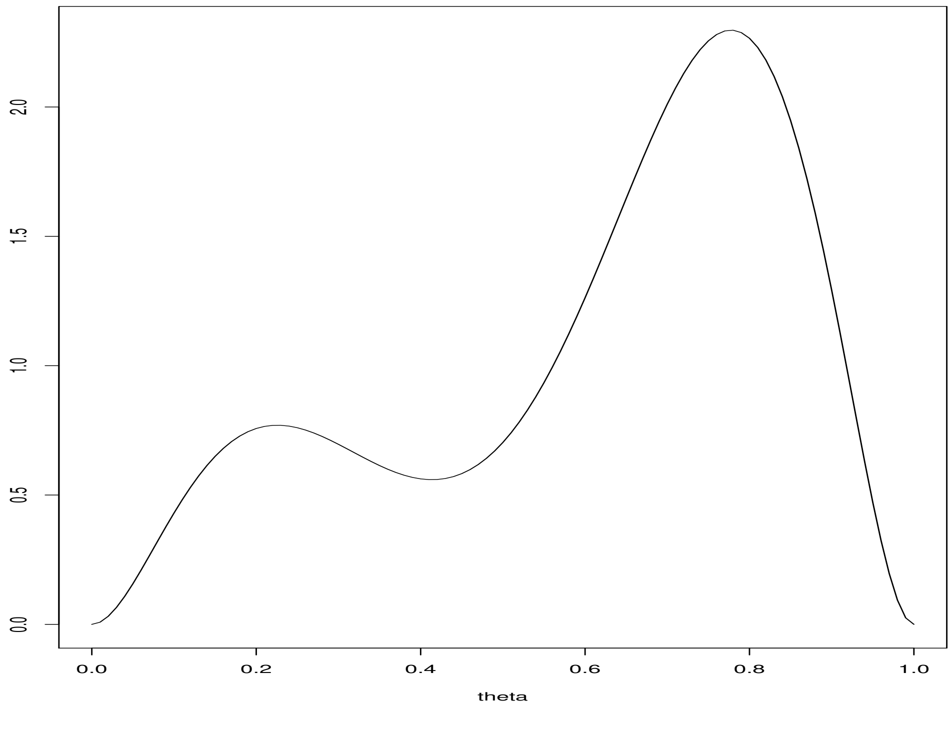
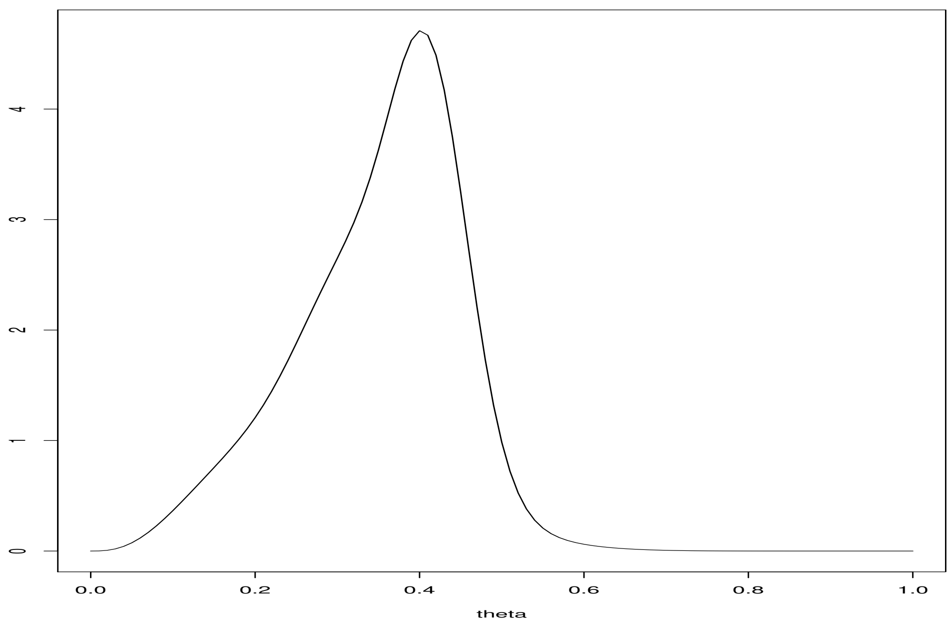


Figura 2.2: Mistura de funções de densidade de Beta(4,10), Beta(15,28) e Beta(50,70) com pesos iguais a 0,33.



2.6 Problemas

1. Mostre que a família de distribuições Beta is conjugada em relação às distribuições amostrais binomial, geométrica e binomial negativa.

2. Para uma amostra aleatória de 100 observações da distribuição normal com média θ e desvio-padrão 2 foi especificada uma priori normal para θ.

   (a) Mostre que o desvio-padrão a posteriori será sempre menor do que 1/5. Interprete este resultado.

   (b) Se o desvio-padrão a priori for igual a 1 qual deve ser o menor número de observações para que o desvio-padrão a posteriori seja 0,1?

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ conhecido. Utilizando uma distribuição a priori Gama para σ^{-2} com coeficiente de variação 0,5, qual deve ser o tamanho amostral para que o coeficiente de variação a posteriori diminua para 0,1?

4. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ e σ^2 desconhecidos, e considere a priori conjugada de (θ, ϕ).

   (a) Determine os parâmetros $(\mu_0, c_0, n_0, \sigma_0^2)$ utilizando as seguintes informações a priori: $E(\theta) = 0$, $P(|\theta| < 1,412) = 0,5$, $E(\phi) = 2$ e $E(\phi^2) = 5$.

   (b) Em uma amostra de tamanho $n = 10$ foi observado $\overline{X} = 1$ e $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 8$. Obtenha a distribuição a posteriori de θ e esboce os gráficos das distribuições a priori, a posteriori e da função de verossimilhança, com ϕ fixo.

   (c) Calcule $P(|Y| > 1|\boldsymbol{x})$ onde Y é uma observação tomada da mesma população.

5. Suponha que o tempo, em minutos, para atendimento a clientes segue uma distribuição exponencial com parâmetro θ desconhecido. Com base na experiência anterior assume-se uma distribuição a priori Gama com média 0,2 e desvio-padrão 1 para θ.

   (a) Se o tempo médio para atender uma amostra aleatória de 20 clientes foi de 3,8 minutos, qual a distribuição a posteriori de θ.

   (b) Qual o menor número de clientes que precisam ser observados para que o coeficiente de variação a posteriori se reduza para 0,1?

6. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ.

(a) Determine os parâmetros da priori conjugada de θ sabendo que $E(\theta) = 4$ e o coeficiente de variação a priori é 0,5.

(b) Quantas observações devem ser tomadas até que a variância a posteriori se reduza para 0,01 ou menos?

(c) Mostre que a média a posteriori é da forma $\gamma_n \overline{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \to 1$ quando $n \to \infty$. Interprete este resultado.

7. O número médio de defeitos por 100 metros de uma fita magnética é desconhecido e denotado por θ. Atribui-se uma distribuição a priori Gama(2,10) para θ. Se um rolo de 1200 metros desta fita foi inspecionado e encontrou-se 4 defeitos qual a distribuição a posteriori de θ?

8. Seja $X_1, \dots, X_n$ uma amostra aleatória da distribuição Bernoulli com parâmetro θ e usamos a priori conjugada $Beta(a, b)$. Mostre que a média a posteriori é da forma $\gamma_n \overline{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \to 1$ quando $n \to \infty$. Interprete este resultado.

9. Para uma amostra aleatória $X_1, \dots, X_n$ tomada da distribuição $U(0, \theta)$, mostre que a família de distribuições de Pareto com parâmetros a e b, cuja função de densidade é $p(\theta) = ab^a/\theta^{a+1}$, é conjugada à uniforme.

10. Para uma variável aleatória $\theta > 0$ a família de distribuições Gama-invertida tem função de densidade de probabilidade dada por

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \quad \alpha, \beta > 0.$$

Mostre que esta família é conjugada ao modelo normal com média μ conhecida e variância θ desconhecida.

11. Suponha que $\boldsymbol{X} = (X_1, X_2, X_3)$ tenha distribuição trinomial com parâmetros n (conhecido) e $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ com $\pi_1 + \pi_2 + \pi_3 = 1$. Mostre que a priori não informativa de Jeffreys para $\boldsymbol{\pi}$ é $p(\boldsymbol{\pi}) \propto [\pi_1 \pi_2 (1 - \pi_1 - \pi_2)]^{-1/2}$.

12. Para cada uma das distribuições abaixo verifique se o modelo é de locação, escala ou locação-escala e obtenha a priori não informativa para os parâmetros desconhecidos.

(a) Cauchy(0,β).

(b) $t_\nu(\mu, \sigma^2)$, ν conhecido.

(c) Pareto(a, b), b conhecido.

(d) Uniforme $(\theta - 1, \theta + 1)$.

(e) Uniforme $(-\theta, \theta)$.

13. Seja uma coleção de variáveis aleatórias independentes X_i com distribuições $p(x_i|\theta_i)$ e seja $p_i(\theta_i)$ a priori não informativa de θ_i, $i = 1, \dots, k$. Mostre que a priori não informativa de Jeffreys para o vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ é dada por $\prod_{i=1}^k p_i(\theta_i)$.

14. Se θ tem priori não informativa $p(\theta) \propto k$, $\theta > 0$ mostre que a priori de $\phi = a\theta + b$, $a \neq 0$ também é $p(\phi) \propto k$.

15. Se θ tem priori não informativa $p(\theta) \propto \theta^{-1}$ mostre que a priori de $\phi = \theta^a$, $a \neq 0$ também é $p(\phi) \propto \phi^{-1}$ e que a priori de $\psi = \log \theta$ é $p(\psi) \propto k$.

Capítulo 3

Estimação

A distribuição a posteriori de um parâmetro θ contém toda a informação probabilística a respeito deste parâmetro e um gráfico da sua função de densidade a posteriori é a melhor descrição do processo de inferência. No entanto, algumas vezes é necessário resumir a informação contida na posteriori através de alguns poucos valores numéricos. O caso mais simples é a estimação pontual de θ onde se resume a distribuição a posteriori através de um único número, $\hat{\theta}$. Como veremos a seguir, será mais fácil entender a escolha de $\hat{\theta}$ no contexto de teoria da decisão.

3.1 Introdução à Teoria da Decisão

Um problema de decisão fica completamente especificado pela descrição dos seguintes espaços:

(i) Espaço do parâmetro ou estados da natureza, Θ.

(ii) Espaço dos resultados possíveis de um experimento, Ω.

(iii) Espaço de possíveis ações, A.

Uma regra de decisão δ é uma função definida em Ω que assume valores em A, i.e. $\delta : \Omega \to A$. A cada decisão δ e a cada possível valor do parâmetro θ podemos associar uma perda $L(\delta, \theta)$ assumindo valores positivos. Definimos assim uma função de perda.

Definição 3.1 *O risco de uma regra de decisão, denotado por $R(\delta)$, é a perda esperada a posteriori, i.e. $R(\delta) = E_{\theta|\boldsymbol{x}}[L(\delta, \theta)]$.*

Definição 3.2 *Uma regra de decisão δ^* é ótima se tem risco mínimo, i.e. $R(\delta^*) < R(\delta)$, $\forall \delta$. Esta regra será denominada regra de Bayes e seu risco, risco de Bayes.*

Exemplo 3.1 : Um laboratório farmaceutico deve decidir pelo lançamento ou não de uma nova droga no mercado. É claro que o laboratório só lançará a droga se achar que ela é eficiente mas isto é exatamente o que é desconhecido. Podemos associar um parâmetro θ aos estados da natureza: droga é eficiente ($\theta = 1$), droga não é eficiente ($\theta = 0$) e as possíveis ações como lança a droga ($\delta = 1$), não lança a droga ($\delta = 0$). Suponha que foi possível construir a seguinte tabela de perdas levando em conta a eficiência da droga,

	eficiente	não eficiente
lança	-500	600
não lança	1500	100

Vale notar que estas perdas traduzem uma avaliação subjetiva em relação à gravidade dos erros cometidos. Suponha agora que a incerteza sobre os estados da natureza é descrita por $P(\theta = 1) = \pi$, $0 < \pi < 1$ avaliada na distribuição atualizada de θ (seja a priori ou a posteriori). Note que, para δ fixo, $L(\delta, \theta)$ é uma variável aleatória discreta assumindo apenas dois valores com probabilidades π e $1 - \pi$. Assim, usando a definição de risco obtemos que

$$
\begin{aligned}
R(\delta = 0) &= E(L(0, \theta)) = \pi 1500 + (1 - \pi)100 = 1400\pi + 100 \\
R(\delta = 1) &= E(L(1, \theta)) = \pi(-500) + (1 - \pi)600 = -1100\pi + 600
\end{aligned}
$$

Uma questão que se coloca aqui é, para que valores de π a regra de Bayes será de lançar a droga. Não é difícil verificar que as duas ações levarão ao mesmo risco, i.e. $R(\delta = 0) = R(\delta = 1)$ se somente se $\pi = 0,20$. Além disso, para $\pi < 0,20$ temos que $R(\delta = 0) < R(\delta = 1)$ e a regra de Bayes consiste em não lançar a droga enquanto que $\pi > 0,20$ implica em $R(\delta = 1) < R(\delta = 0)$ e a regra de Bayes deve ser de lançar a droga.

3.2 Estimadores de Bayes

Seja agora uma amostra aleatória $X_1, \ldots, X_n$ tomada de uma distribuição com função de (densidade) de probabilidade $p(x|\theta)$ aonde o valor do parâmetro θ é desconhecido. Em um problema de inferência como este o valor de θ deve ser estimado a partir dos valores observados na amostra.

Se $\theta \in \Theta$ então é razoável que os possíveis valores de um estimador $\delta(\boldsymbol{X})$ também devam pertencer ao espaço Θ. Além disso, um bom estimador é aquele para o qual, com alta probabilidade, o erro $\delta(\boldsymbol{X}) - \theta$ estará próximo de zero. Para cada possível valor de θ e cada possível estimativa $a \in \Theta$ vamos associar uma perda $L(a, \theta)$ de modo que quanto maior a distância entre a e θ maior o valor da perda. Neste caso, a perda esperada a posteriori é dada por

$$E[L(a,\theta)|\boldsymbol{x}] = \int L(a,\theta)p(\theta|\boldsymbol{x})d\theta$$

e a regra de Bayes consiste em escolher a estimativa que minimiza esta perda esperada.

Aqui vamos discutir apenas funções de perda simétricas, já que estas são mais comumente utilizadas. Dentre estas a mais utilizada em problemas de estimação é certamente a função de perda quadrática, definida como $L(a,\theta) = (a-\theta)^2$. Neste caso, pode-se mostrar que o estimador de Bayes para o parâmetro θ será a média de sua distribuição atualizada.

Exemplo 3.2 : Suponha que queremos estimar a proporção θ de itens defeituosos em um grande lote. Para isto será tomada uma amostra aleatória $X_1,\dots,X_n$ de uma distribuição de Bernoulli com parâmetro θ. Usando uma priori conjugada Beta(α,β) sabemos que após observar a amostra a distribuição a posteriori é Beta$(\alpha+t,\beta+n-t)$ onde $t=\sum_{i=1}^n x_i$. A média desta distribuição Beta é dada por $(\alpha+t)/(\alpha+\beta+n)$ e portanto o estimador de Bayes de θ usando perda quadrática é

$$\delta(\boldsymbol{X}) = \frac{\alpha+\sum_{i=1}^n X_i}{\alpha+\beta+n}.$$

A perda quadrática é as vezes criticada por penalizar demais o erro de estimação. A função de perda absoluta, definida como $L(a,\theta)=|a-\theta|$, introduz punições que crescem linearmente com o erro de estimação e pode-se mostrar que o estimador de Bayes associado é a mediana da distribuição atualizada de θ.

Para reduzir ainda mais o efeito de erros de estimação grandes podemos considerar funções que associam uma perda fixa a um erro cometido, não importando sua magnitude. Uma tal função de perda, denominada perda 0-1, é definida como

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

para todo $\epsilon > 0$. Neste caso pode-se mostrar que o estimador de Bayes é a moda da distribuição atualizada de θ. A moda da posteriori de θ também é chamado de estimador de máxima verossimilhança generalizado (EMVG) e é o mais fácil de ser obtido dentre os estimadores vistos até agora. No caso contínuo devemos obter a solução da equação

$$\frac{\partial p(\theta|\boldsymbol{x})}{\partial\theta} = 0.$$

Exemplo 3.3 : Se $X_1, \dots, X_n$ é uma amostra aleatória da $N(\theta, \sigma^2)$ com σ^2 conhecido e usarmos a priori conjugada, i.e. $\theta \sim N(\mu_0, \tau_0^2)$ então a posteriori também será normal e neste caso média, mediana e moda coincidem. Portanto, o estimador de Bayes de θ é dado por

$$\delta(\boldsymbol{X}) = \frac{\tau_0^{-2}\mu_0 + n\sigma^{-2}\overline{\boldsymbol{X}}}{\tau_0^{-2} + n\sigma^{-2}}.$$

Exemplo 3.4 : No exemplo 3.2 suponha que foram observados 100 itens dos quais 10 eram defeituosos. Usando perda quadrática a estimativa de Bayes de θ é

$$\delta(\boldsymbol{x}) = \frac{\alpha + 10}{\alpha + \beta + 100}$$

Assim, se a priori for Beta(1,1), ou equivalentemente $U(0,1)$, então $\delta(\boldsymbol{x}) = 0,108$. Por outro lado se especificarmos uma priori Beta(1,2), que é bem diferente da anterior, então $\delta(\boldsymbol{x}) = 0,107$. Ou seja, as estimativas de Bayes são bastante próximas, e isto é uma consequência do tamanho amostral ser grande. Note também que ambas as estimativas são próximas da proporção amostral de defeituosos 0,1, que é a estimativa de máxima verossimilhança.

3.3 Estimação por Intervalos

Voltamos a enfatizar que a forma mais adequada de expressar a informação que se tem sobre um parâmetro é através de sua distribuição a posteriori. A principal restrição da estimação pontual é que quando estimamos um parâmetro através de um único valor numérico toda a informação presente na distribuição a posteriori é resumida através deste número. É importante também associar alguma informação sobre o quão precisa é a especificação deste número. Para os estimadores vistos aqui as medidas de incerteza mais usuais são a variância ou o coeficiente de variação para a média a posteriori, a medida de informação observada de Fisher para a moda a posteriori, e a distância entre quartis para a mediana a posteriori.

Nesta seção vamos introduzir um compromisso entre o uso da própria distribuição a posteriori e uma estimativa pontual. Será discutido o conceito de intervalo de credibilidade (ou intervalo de confiança Bayesiano) baseado no distribuição a posteriori.

Definição 3.3 *C é um intervalo de credibilidade de 100(1-α)%, ou nível de credibilidade (ou confiança) $1 - \alpha$, para θ se $P(\theta \in C) \geq 1 - \alpha$.*

Note que a definição expressa de forma probabilística a pertinência ou não de θ ao intervalo. Assim, quanto menor for o tamanho do intervalo mais concentrada é a distribuição do parâmetro, ou seja o tamanho do intervalo informa sobre a dispersão de θ. Além disso, a exigência de que a probabilidade acima possa ser maior do que o nível de confiança é essencialmente técnica pois queremos que o intervalo seja o menor possível, o que em geral implica em usar uma igualdade. No entanto, a desigualdade será útil se θ tiver uma distribuição discreta onde nem sempre é possível satisfazer a igualdade.

Outro fato importante é que os intervalos de credibilidade são invariantes a transformações 1 a 1, $\phi(\theta)$. Ou seja, se $C = [a, b]$ é um intervalo de credibilidade 100(1-α)% para θ então $[\phi(a), \phi(b)]$ é um intervalo de credibilidade 100(1-α)% para $\phi(\theta)$. Note que esta propriedade também vale para intervalos de confiança na inferência clássica.

É possível construir uma infinidade de intervalos usando a definição acima mas estamos interessados apenas naquele com o menor comprimento possível. Pode-se mostrar que intervalos de comprimento mínimo são obtidos tomando-se os valores de θ com maior densidade a posteriori, e esta idéia é expressa matematicamente na definição abaixo.

Definição 3.4 *Um intervalo de credibilidade C de 100(1-α)% para θ é de máxima densidade a posteriori (MDP) se $C = \{\theta \in \Theta : p(\theta|\boldsymbol{x}) \geq k(\alpha)\}$ onde $k(\alpha)$ é a maior constante tal que $P(\theta \in C) \geq 1 - \alpha$.*

Usando esta definição, todos os pontos dentro do intervalo MDP terão densidade maior do que qualquer ponto fora do intervalo. Além disso, no caso de distribuições com duas caudas, e.g. normal, t de Student, o intervalo MDP é obtido de modo que as caudas tenham a mesma probabilidade.

Um problema com os intervalos MDP é que eles não são invariantes a transformações 1 a 1, a não ser para transformações lineares. O mesmo problema ocorre com intervalos de comprimento mínimo na inferência clássica.

3.4 Estimação no Modelo Normal

Os resultados desenvolvidos nos capítulos anteriores serão aplicados ao modelo normal para estimação da média e variância em problemas de uma ou mais amostras e em modelos de regressão linear. A análise será feita com priori conjugada e priori não informativa quando serão apontadas as semelhanças com a análise clássica. Assim como nos capítulos anteriores a abordagem aqui é introdutória. Um tratamento mais completo do enfoque Bayesiano em modelos lineares pode ser encontrado em Broemeling (1985) e Box e Tiao (1992).

Nesta seção considere uma amostra aleatória $X_1, \cdots, X_n$ tomada da distribuição $N(\theta, \sigma^2)$.

3.4.1 Variância Conhecida

Se σ^2 é conhecido e a priori de θ é $N(\mu_0, \tau_0^2)$ então, pelo Teorema 1.1, a posteriori de θ é $N(\mu_1, \tau_1^2)$. Intervalos de confiança Bayesianos para θ podem então ser construídos usando o fato de que

$$\frac{\theta - \mu_1}{\tau_1} | \boldsymbol{x} \sim N(0,1).$$

Assim, usando uma tabela da distribuição normal padronizada podemos obter o valor do percentil $z_{\alpha/2}$ tal que

$$P\left(-z_{\alpha/2} \leq \frac{\theta - \mu_1}{\tau_1} \leq z_{\alpha/2}\right) = 1 - \alpha$$

e após isolar θ, obtemos que

$$P\left(\mu_1 - z_{\alpha/2}\tau_1 \leq \theta \leq \mu_1 + z_{\alpha/2}\tau_1\right) = 1 - \alpha.$$

Portanto $\left(\mu_1 - z_{\alpha/2}\tau_1; \mu_1 + z_{\alpha/2}\tau_1\right)$ é o intervalo de confiança 100(1-α)% MDP para θ, devido à simetria da normal.

A priori não informativa pode ser obtida fazendo-se a variância da priori tender a infinito, i.e. $\tau_0^2 \to \infty$. Neste caso, é fácil verificar que $\tau_1^{-2} \to n\sigma^{-2}$ e $\mu_1 \to \overline{\boldsymbol{x}}$, i.e. a média e a precisão da posteriori convergem para a média e a precisão amostrais. Média, moda e mediana a posteriori coincidem então com a estimativa clássica de máxima verossimilhança, $\overline{\boldsymbol{x}}$. O intervalo de confiança Bayesiano 100(1-α)% é dado por

$$\left(\overline{\boldsymbol{x}} - z_{\alpha/2}\ \sigma/\sqrt{n};\ \overline{\boldsymbol{x}} + z_{\alpha/2}\ \sigma/\sqrt{n}\right)$$

e também coincide numericamente com o intervalo de confiança clássico. Aqui entretanto a interpretação do intervalo é como uma afirmação probabilística sobre θ.

3.4.2 Média e Variância desconhecidas

Neste caso, usando a priori conjugada Normal-Gama vista no Capítulo 2 temos que a distribuição a posteriori marginal de θ é dada por

$$\theta | \boldsymbol{x} \sim t_{n_1}(\mu_1, \sigma_1^2/c_1).$$

Portanto, média, moda e mediana a posteriori coincidem e são dadas por μ_1. Denotando por $t_{\alpha/2,n_1}$ o percentil 100(1-$\alpha/2$)% da distribuição $t_{n_1}(0,1)$ podemos obter este percentil tal que

$$P\left(-t_{\alpha/2,n_1} \leq \sqrt{c_1}\frac{\theta - \mu_1}{\sigma_1} \leq t_{\alpha/2,n_1}\right) = 1 - \alpha$$

e após isolar θ, usando a simetria da distribuição t-Student obtemos que

$$\left(\mu_1 - t_{\alpha/2,n_1}\frac{\sigma_1}{\sqrt{c_1}} \leq \theta \leq \mu_1 + t_{\alpha/2,n_1}\frac{\sigma_1}{\sqrt{c_1}}\right)$$

é o intervalo de confiança Bayesiano 100(1-α)% de MDP para θ.

No caso da variância populacional σ^2 intervalos de confiança podem ser obtidos usando os percentis da distribuição qui-quadrado uma vez que a distribuição a posteriori de ϕ é tal que $n_1\sigma_1^2\phi|\boldsymbol{x} \sim \chi^2_{n_1}$. Denotando por

$$\underline{\chi}^2_{\alpha/2,n_1} \quad \text{e} \quad \overline{\chi}^2_{\alpha/2,n_1}$$

os percentis $\alpha/2$ e $1-\alpha/2$ da distribuição qui-quadrado com n_1 graus de liberdade respectivamente, podemos obter estes percentis tais que

$$P\left(\frac{\underline{\chi}^2_{\alpha/2,n_1}}{n_1\sigma_1^2} \leq \phi \leq \frac{\overline{\chi}^2_{\alpha/2,n_1}}{n_1\sigma_1^2}\right) = 1-\alpha.$$

Note que este intervalo não é de MDP já que a distribuição qui-quadrado não é simétrica. Como $\sigma^2 = 1/\phi$ é uma função 1 a 1 podemos usar a propriedade de invariância e portanto

$$\left(\frac{n_1\sigma_1^2}{\overline{\chi}^2_{\alpha/2,n_1}}; \frac{n_1\sigma_1^2}{\underline{\chi}^2_{\alpha/2,n_1}}\right)$$

é o intervalo de confiança Bayesiano 100(1-α)% para σ^2.

Um caso particular é quanto utilizamos uma priori não informativa. Vimos na Seção 2.4 que a priori não informativa de locação e escala é $p(\theta,\sigma) \propto 1/\sigma$, portanto pela propriedade de invariância segue que a priori não informativa de (θ,ϕ) é obtida fazendo-se $p(\theta,\phi) \propto \phi^{-1}$. Note que este é um caso particular (degenerado) da priori conjugada natural com $c_0 = 0$, $\sigma_0^2 = 0$ e $n_0 = -1$. Neste caso a distribuição a posteriori marginal de θ fica

$$\theta|\boldsymbol{x} \sim t_{n-1}(\overline{\boldsymbol{x}}, s^2/n)$$

onde $s^2 = 1/(n-1)\sum_{i=1}^n (x_i - \overline{x})^2$.

Mais uma vez média, moda e mediana a posteriori de θ coincidem com a média amostral $\overline{x}$ que é a estimativa de máxima verossimilhança. Como $\sqrt{n}(\theta - \overline{x})/s \sim t_{n-1}(0,1)$ segue que o intervalo de confiança 100(1-α)% para θ de MDP é

$$\left(\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}; \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right)$$

que coincide numericamente com o intervalo de confiança clássico.

Para fazer inferências sobre σ^2 temos que

$$\phi|\boldsymbol{x} \sim \text{Gama}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \quad \text{ou} \quad (n-1)s^2\phi|\boldsymbol{x} \sim \chi^2_{n-1}.$$

A estimativa pontual de σ^2 utilizada é $[E(\phi|x)]^{-1} = s^2$ que coincide com a estimativa clássica uma vez que o estimador de máxima verossimilhança $(n-1)S^2/n$ é viciado e normalmente substituido por S^2 (que é não viciado). Os intervalos de confiança 100(1-α)% Bayesiano e clássico também coincidem e são dados por

$$\left(\frac{(n-1)s^2}{\overline{\chi}^2_{\alpha/2,n-1}}; \frac{(n-1)s^2}{\underline{\chi}^2_{\alpha/2,n-1}}\right).$$

Mais uma vez vale enfatizar que esta coincidência com as estimativas clássicas é apenas numérica uma vez que as interpretações dos intervalos diferem radicalmente.

3.4.3 O Caso de duas Amostras

Nesta seção vamos assumir que $X_{11}, \dots, X_{1n_1}$ e $X_{21}, \dots, X_{2n_2}$ são amostras aleatórias das distribuições $N(\theta_1, \sigma_1^2)$ e $N(\theta_2, \sigma_2^2)$ respectivamente e que as amostras são independentes.

Para começar vamos assumir que as variâncias σ_1^2 e σ_2^2 são conhecidas. Neste caso, a função de verossimilhança é dada por

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2|\theta_1, \theta_2) = p(\boldsymbol{x}_1|\theta_1)p(\boldsymbol{x}_2|\theta_2) \propto \exp\left\{-\frac{n_1}{2\sigma_1^2}(\theta_1 - \overline{x}_1)^2\right\} \exp\left\{-\frac{n_2}{2\sigma_2^2}(\theta_2 - \overline{x}_2)^2\right\}$$

isto é, o produto de verossimilhanças relativas a θ_1 e θ_2. Assim, se assumirmos que θ_1 e θ_2 são independentes a priori então eles também serão independentes a posteriori já que

$$p(\theta_1, \theta_2|\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_1|\theta_1)p(\theta_1)}{p(\boldsymbol{x}_1)} \times \frac{p(\boldsymbol{x}_2|\theta_2)p(\theta_2)}{p(\boldsymbol{x}_2)}$$

Se usarmos a classe de prioris conjugadas $\theta_i \sim N(\mu_i, \tau_i^2)$ então as posterioris independentes serão $\theta_i|\boldsymbol{x}_i \sim N(\mu_i^*, \tau_i^{*^2})$ onde

$$\mu_i^* = \frac{\tau_i^{-2}\mu_i + n_i\sigma_i^{-2}\ \overline{\boldsymbol{x}}_i}{\tau_i^{-2} + n_i\sigma_i^{-2}} \quad \text{e} \quad \tau_i^{*^2} = 1/(\tau_i^{-2} + n_i\sigma_i^{-2}), \quad i = 1, 2.$$

Em geral estaremos interessados em comparar as médias populacionais, i.e queremos estimar $\beta = \theta_1 - \theta_2$. Neste caso, a posteriori de β é facilmente obtida, devido à independência, como

$$\beta|\boldsymbol{x}_1, \boldsymbol{x}_2 \sim N(\mu_1^* - \mu_2^*, \tau_1^{*^2} + \tau_2^{*^2})$$

e podemos usar $\mu_1^* - \mu_2^*$ como estimativa pontual para a diferença e também construir um intervalo de credibilidade MDP para esta diferença. Note que se usarmos priori não informativa, i.e. fazendo $\tau_i^2 \to \infty$, $i = 1, 2$ então a posteriori fica

$$\beta|\boldsymbol{x}_1, \boldsymbol{x}_2 \sim N\left(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

e o intervalo obtido coincidirá mais uma vez com o intervalo de confiança clássico.

No caso de variâncias populacionais desconhecidas porém iguais, temos que $\phi = \sigma_1^{-2} = \sigma_2^{-2} = \sigma^2$. A priori conjugada pode ser construída em duas etapas. No primeiro estágio, assumimos que, dado ϕ, θ_1 e θ_2 são a priori condicionalmente independentes, e especificamos

$$\theta_i|\phi \sim N(\mu_i, (c_i\phi)^{-1}), i = 1, 2.$$

e no segundo estágio, especificamos a priori conjugada natural para ϕ, i.e.

$$\phi \sim \text{Gama}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right).$$

Combinando as prioris acima não é difícil verificar que a priori conjunta de $(\theta_1, \theta_2, \phi)$ é

$$\begin{aligned} p(\theta_1, \theta_2, \phi) &= p(\theta_1|\phi)p(\theta_2|\phi)p(\phi) \\ &\propto \phi^{n_0/2} \exp\left\{-\frac{\phi}{2}\left[n_0\sigma_0^2 + c_1(\theta_1 - \mu_1)^2 + c_2(\theta_2 - \mu_2)^2\right]\right\}. \end{aligned}$$

Além disso, também não é difícil obter a priori condicional de $\beta = \theta_1 - \theta_2$, dado ϕ, como

$$\beta|\phi \sim N(\mu_1 - \mu_2, \phi^{-1}(c_1^{-1} + c_2^{-1}))$$

e portanto, usando os resultados da Seção 2.3.5 segue que a distribuição a priori marginal da diferença é

$$\beta \sim t_{n_0}(\mu_1 - \mu_2, \sigma_0^2(c_1^{-1} + c_2^{-1})).$$

Podemos mais uma vez obter a posteriori conjunta em duas etapas já que θ_1 e θ_2 também serão condicionalmente independentes a posteriori, dado ϕ. Assim, no primeiro estágio usando os resultados obtidos anteriormente para uma amostra segue que

$$\theta_i|\phi, \boldsymbol{x} \sim N(\mu_i^*, (c_1^*\phi)^{-1}), \quad i = 1, 2$$

onde

$$\mu_i^* = \frac{c_i\mu_i + n_i\overline{x}_i}{c_i + n_i} \quad \text{e} \quad c_i^* = c_i + n_i.$$

Na segunda etapa temos que combinar a verossimilhança com a priori de $(\theta_1, \theta_2, \phi)$. Definindo a variância amostral combinada

$$s^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

e denotando $\nu = n_1 + n_2 - 2$, a função de verossimilhança pode ser escrita como

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2 | \theta_1, \theta_2, \phi) = \phi^{(n_1+n_2)/2} \exp\left\{ -\frac{\phi}{2}\left[\nu s^2 + n_1(\theta_1 - \overline{\boldsymbol{x}}_1)^2 + n_2(\theta_2 - \overline{\boldsymbol{x}}_2)^2 \right] \right\}$$

e após algum algebrismo obtemos que a posteriori é proporcional a

$$\phi^{(n_0+n_1+n_2)/2} \exp\left\{ -\frac{\phi}{2}\left[n_0\sigma_0^2 + \nu s^2 + \sum_{i=1}^{2} \frac{c_i n_i}{c_i^*}(\mu_i - \overline{\boldsymbol{x}}_i)^2 + c_i^*(\theta_i - \mu_i^*)^2 \right] \right\}.$$

Como esta posteriori tem o mesmo formato da priori segue por analogia que

$$\phi | \boldsymbol{x} \sim \text{Gama}\left(\frac{n_0^*}{2}, \frac{n_0^* {\sigma_0^*}^2}{2} \right)$$

onde $n_0^* = n_0 + n_1 + n_2$ e $n_0^* {\sigma_0^*}^2 = n_0\sigma_0^2 + \nu s^2 + \sum_{i=1}^{2} c_i n_i (\mu_i - \overline{\boldsymbol{x}}_i)^2 / c_i^*$. Ainda por analogia com o caso de uma amostra, a posteriori marginal da diferença é dada por

$$\beta | \boldsymbol{x} \sim t_{n_0^*}(\mu_1^* - \mu_2^*, {\sigma_0^*}^2({c_1^*}^{-1} + {c_2^*}^{-1})).$$

Assim, média, moda e mediana a posteriori de β coincidem e a estimativa pontual é $\mu_1^* - \mu_2^*$. Também intervalos de credibilidade de MDP podem ser obtidos usando os percentis da distribuição t de Student. Para a variância populacional a estimativa pontual usual é ${\sigma_0^*}^2$ e intervalos podem ser construídos usando os percentis da distribuição qui-quadrado já que $n_0^* {\sigma_0^*}^2 \phi \mid \boldsymbol{x} \sim \chi^2_{n_0^*}$

Vejamos agora como fica a análise usando priori não informativa. Neste caso, $p(\theta_1, \theta_2, \phi) \propto \phi^{-1}$ e isto equivale a um caso particular (degenerado) da priori conjugada com $c_i = 0$, $\sigma_0^2 = 0$ e $n_0 = -2$. Assim, temos que $c_i^* = n_i$, $\mu_i^* = \overline{\boldsymbol{x}}_i$, $n_0^* = \nu$ e $n_0^* {\sigma_0^*}^2 = \nu s^2$ e a estimativa pontual concide com a estimativa de máxima verossimilhança $\hat{\beta} = \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2$. O intervalo de $100(1-\alpha)\%$ de MDP para β tem limites

$$\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 \pm t_{\frac{\alpha}{2},\nu}\, s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

que coincide numericamente com o intervalo de confiança clássico.

O intervalo de $100(1-\alpha)\%$ para σ^2 é obtido de maneira análoga ao caso de uma amostra usando a distribuição qui-quadrado, agora com ν graus de liberdade, i.e.

$$\left(\frac{\nu s^2}{\overline{\chi}^2_{\frac{\alpha}{2},\nu}}, \frac{\nu s^2}{\underline{\chi}^2_{\frac{\alpha}{2},\nu}} \right).$$

3.4.4 Variâncias desiguais

Até agora assumimos que as variâncias populacionais desconhecidas eram iguais (ou pelo menos aproximadamente iguais). Na inferência clássica a violação desta suposição leva a problemas teóricos e práticos uma vez que não é trivial encontrar uma quantidade pivotal para β com distribuição conhecida ou tabelada. Na verdade, se existem grandes diferenças de variabilidade entre as duas populações pode ser mais apropriado analisar conjuntamente as consequências das diferenças entre as médias e as variâncias. Assim, caso o pesquisador tenha interesse no parâmetro β deve levar em conta os problemas de ordem teóricas introduzidos por uma diferença substancial entre σ_1^2 e σ_2^2.

Do ponto de vista Bayesiano o que precisamos fazer é combinar informação a priori com a verossimilhança e basear a estimação na distribuição a posteriori. A função de verossimilhança agora pode ser fatorada como

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2 | \theta_1, \theta_2, \sigma_1^2 \sigma_2^2) = p(\boldsymbol{x}_1 | \theta_1, \sigma_1^2) p(\boldsymbol{x}_2 | \theta_2, \sigma_2^2)$$

e vamos adotar prioris conjugadas normal-gama independentes com parâmetros $(\mu_i, c_i, \nu_i, \sigma_{0i}^2)$ para cada uma das amostras. Fazendo as operações usuais para cada amostra, e usando a conjugação da normal-gama, obtemos as seguintes distribuições a posteriori independentes

$$\theta_i | \boldsymbol{x} \sim t_{n_{0i}^*}(\mu_i^*, \sigma_{0i}^{*^2} / c_i^*) \quad \text{e} \quad \phi_i | \boldsymbol{x} \sim \text{Gama}\left(\frac{n_{0i}^*}{2}, \frac{n_{0i}^* \sigma_{0i}^{*^2}}{2}\right), \quad i = 1, 2.$$

Pode-se mostrar que β tem uma distribuição a posteriori chamada Behrens-Fisher, que é semelhante à t de Student e é tabelada. Assim, intervalos de credibilidade podem ser construídos usando-se estes valores tabelados.

Outra situação de interesse é a comparação das duas variâncias populacionais. Neste caso, faz mais sentido utilizar a razão de variâncias ao invés da diferença já que elas medem a escala de uma distribuição e são sempre positivas. Neste caso temos que obter a distribuição a posteriori de $\sigma_2^2 / \sigma_1^2 = \phi_1 / \phi_2$. Usando a independência a posteriori de ϕ_1 e ϕ_2 e após algum algebrismo pode-se mostrar que

$$\frac{\sigma_{01}^{*^2}}{\sigma_{02}^{*^2}} \frac{\phi_1}{\phi_2} \sim F(n_{01}^*, n_{02}^*)$$

Embora sua função de distribuição não possa ser obtida analiticamente os valores estão tabelados em muitos livros de estatística e também podem ser obtidos na maioria dos pacotes computacionais. Os percentis podem então ser utilizados na construção de intervalos de credibilidade para a razão de variâncias.

Uma propriedade bastante útil para calcular probabilidade com a distribuição F vem do fato de que se $X \sim F(\nu_2, \nu_1)$ então $X^{-1} \sim F(\nu_1, \nu_2)$ por simples inversão na razão de distribuições qui-quadrado independentes. Assim, denotando os quantis α e $1 - \alpha$ da distribuição $F(\nu_1, \nu_2)$ por $\underline{F}_\alpha(\nu_1, \nu_2)$ e $\overline{F}_\alpha(\nu_1, \nu_2)$ respectivamente segue que

$$\underline{F}_\alpha(\nu_1, \nu_2) = \frac{1}{\overline{F}_\alpha(\nu_2, \nu_1)}.$$

Note que é usual que os livros forneçam tabelas com os percentis superiores da distribuição F para várias combinações de valores de ν_1 e ν_2 devido à propriedade acima. Por exemplo, se temos os valores tabelados dos quantis 0,95 podemos obter também um quantil 0,05. Basta procurar o quantil 0,95 inverterndo os graus de liberdade.

Finalmente, a análise usando priori não informativa pode ser feita para $p(\theta_1, \theta_2, \sigma_1^2, \sigma_2^2) \propto \sigma_1^{-2} \sigma_2^{-2}$ e será deixada como exercício.

Capítulo 4

Computação Bayesiana

Existem várias formas de resumir a informação descrita na distribuição a posteriori. Esta etapa frequentemente envolve a avaliação de probabilidades ou esperanças.

Neste capítulo serão descritos métodos baseados em simulação, incluindo Monte Carlo simples, Monte Carlo com função de importância, o método do Bootstrap Bayesiano e Monte Carlo via cadeias de Markov (MCMC). O material apresentado é introdutório e mais detalhes sobre os estes métodos podem ser obtidos em Gamerman (1997), Davison e Hinckley (1997) e Robert e Casella (1999). Outros métodos computacionalmente intensivos como técnicas de otimização e integração numérica, bem como aproximações analíticas não serão tratados aqui e uma referência introdutória é Migon e Gamerman (1999).

Todos os algoritmos que serão vistos aqui são não determinísticos, i.e. todos requerem a simulação de números (pseudo) aleatórios de alguma distribuição de probabilidades. Em geral, a única limitação para o número de simulações são o tempo de computação e a capacidade de armazenamento dos valores simulados. Assim, se houver qualquer suspeita de que o número de simulações é insuficiente, a abordagem mais simples consiste em simular mais valores.

4.1 Uma Palavra de Cautela

Apesar da sua grande utilidade, os métodos que serão apresentados aqui devem ser aplicados com cautela. Devido à facilidade com que os recursos computacionais podem ser utilizados hoje em dia, corremos o risco de apresentar uma solução para o problema errado (o erro tipo 3) ou uma solução ruim para o problema certo. Assim, os métodos computacionalmente intensivos não devem ser vistos como substitutos do pensamento crítico sobre o problema por parte do pesquisador.

Além disso, sempre que possível deve-se utilizar soluções exatas, i.e. não aproximadas, se elas existirem. Por exemplo, em muitas situações em que precisamos calcular uma integral múltipla existe solução exata em algumas dimensões, enquanto nas outras dimensões temos que usar métodos de aproximação.

4.2 O Problema Geral da Inferência Bayesiana

A distribuição a posteriori pode ser convenientemente resumida em termos de esperanças de funções particulares do parâmetro θ, i.e.

$$E[g(\theta)|\boldsymbol{x}] = \int g(\theta)p(\theta|\boldsymbol{x})d\theta$$

ou distribuições a posteriori marginais quando $\boldsymbol{\theta}$ for multidimensional, i.e.

$$p(\boldsymbol{\theta}_1|\boldsymbol{x}) = \int p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}_2$$

onde $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Assim, o problema geral da inferência Bayesiana consiste em calcular tais valores esperados segundo a distribuição a posteriori de θ. Alguns exemplos são,

1. Constante normalizadora. $g(\theta) = 1$ e $p(\theta|bfx) = kq(\theta)$, segue que
$$k = \left[\int q(\theta)d\theta\right]^{-1}.$$

2. Se $g(\theta) = \theta$, então têm-se $\mu = E(\theta|\boldsymbol{x})$, média a posteriori.

3. Quando $g(\theta) = (\theta - \mu)^2$, então $\sigma^2 = var(\theta) = E((\theta - \mu)^2|\boldsymbol{x})$, a variância a posteriori.

4. Se $g(\theta) = I_A(\theta)$, onde $I_A(x) = 1$ se $x \in A$ e zero caso contrário, então $P(A \mid \boldsymbol{x}) = \int_A p(\theta|\boldsymbol{x})d\theta$

5. Seja $g(\theta) = p(y|\theta)$, onde $y \perp \boldsymbol{x}|\theta$. Nestas condições obtemos $E[p(y|\boldsymbol{x})]$, a distribuição preditiva de y, uma observação futura.

Portanto, a habilidade de integrar funções, muitas vezes complexas e multidimensionais, é extremamente importante em inferência Bayesiana. Inferência exata somente será possível se estas integrais puderem ser calculadas analiticamente, caso contrário devemos usar aproximações. Nas próximas seções iremos apresentar métodos aproximados baseados em simulação para obtenção dessas integrais.

4.3 Método de Monte Carlo Simples

A idéia do método é justamente escrever a integral que se deseja calcular como um valor esperado. Para introduzir o método considere o problema de calcular a integral de uma função $g(\theta)$ no intervalo (a, b), i.e.

$$I = \int_a^b g(\theta) d\theta.$$

Esta integral pode ser reescrita como

$$I = \int_a^b (b-a) g(\theta) \frac{1}{b-a} d\theta = (b-a) E[g(\theta)]$$

identificando θ como uma variável aleatória com distribuição $U(a, b)$. Assim, transformamos o problema de avaliar a integral no problema estatístico de estimar uma média, $E[g(\theta)]$. Se dispomos de uma amostra aleatória de tamanho n, $\theta_1, \dots, \theta_n$ da distribuição uniforme no intervalo (a, b) teremos também uma amostra de valores $g(\theta_1), \dots, g(\theta_n)$ da função $g(\theta)$ e a integral acima pode ser estimada pela média amostral, i.e.

$$\hat{I} = (b-a) \frac{1}{n} \sum_{i=1}^{n} g(\theta_i).$$

Não é difícil verificar que esta estimativa é não viesada já que

$$E(\hat{I}) = \frac{(b-a)}{n} \sum_{i=1}^{n} E[g(\theta_i)] = (b-a) E[g(\theta)] = \int_a^b g(\theta) d\theta.$$

Podemos então usar o seguinte algoritmo

1. gere $\theta_1, \dots, \theta_n$ da distribuição $U(a, b)$;
2. calcule $g(\theta_1), \dots, g(\theta_n)$;
3. calcule a média amostral $\overline{g} = \sum_{i=1}^{n} g(\theta_i)/n$
4. calcule $\hat{I} = (b-a)\overline{g}$

A generalização é bem simples para o caso em que a integral é a esperança matemática de uma função $g(\theta)$ onde θ tem função de densidade $p(\theta)$, i.e.

$$I = \int_a^b g(\theta) p(\theta) d\theta = E[g(\theta)]. \tag{4.1}$$

Neste caso, podemos usar o mesmo algoritmo descrito acima modificando o passo 1 para gerar $\theta_1, \dots, \theta_n$ da distribuição $p(\theta)$ e calculando $\hat{I} = \overline{g}$.

Uma vez que as gerações são independentes, pela Lei Forte dos Grandes Números segue que $\hat{I}$ converge quase certamente para I. Além disso, a variância do estimador pode também ser estimada como

$$v = \frac{1}{n^2} \sum_{i=1}^{n} (g(\theta_i) - \overline{g})^2,$$

i.e. a aproximação pode ser tão acurada quanto se deseje bastando aumentar o valor de n. É importante notar que n está sob nosso controle aqui, e não se trata do tamanho da amostra de dados.

Para n grande segue que

$$\frac{\overline{g} - E[g(\theta)]}{\sqrt{v}}$$

tem distribuição aproximadamente $N(0,1)$. Podemos usar este resultado para testar convergência e construir intervalos de confiança.

No caso multivariado a extensão também é direta. Seja $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ um vetor aleatório de dimensão k com função de densidade $p(\boldsymbol{\theta})$. Neste caso os valores gerados serão também vetores $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ e o estimador de Monte Carlo fica

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i)$$

4.3.1 Monte Carlo via Função de Importância

Em muitas situações pode ser muito custoso ou mesmo impossível simular valores da distribuição a posteriori. Neste caso, pode-se recorrer à uma função $q(\theta)$ que seja de fácil amostragem, usualmente chamada de *função de importância*. O procedimento é comumente chamado de *amostragem por importância*.

Se $q(\theta)$ for uma função de densidade definida no mesmo espaço variação de θ então a integral (4.1) pode ser reescrita como

$$I = \int \frac{g(\theta)p(\theta)}{q(\theta)} q(\theta) dx = E\left[\frac{g(\theta)p(\theta)}{q(\theta)}\right]$$

onde a esperança agora é com respeito a distribuição q. Assim, se dispomos de uma amostra aleatória $\theta_1, \dots, \theta_n$ tomada da distribuição q o estimador de Monte Carlo da integral acima fica

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(\theta_i)p(\theta_i)}{q(\theta_i)}.$$

e tem as mesmas propriedades do estimador de Monte Carlo simples.

Em princípio não há restrições quanto à escolha da densidade de importância q, porém na prática alguns cuidados devem ser tomados. Pode-se mostrar que a escolha ótima no sentido de minimizar a variância do estimador consiste em tomar $q(\theta) \propto g(\theta)p(\theta)$.

Exemplo 4.1 : Para uma única observação X com distribuição $N(\theta, 1)$, θ desconhecido, e priori Cauchy(0,1) segue que

$$p(x|\theta) \propto \exp[-(x-\theta)^2/2] \quad \text{e} \quad p(\theta) = \frac{1}{\pi(1+\theta^2)}.$$

Portanto, a densidade a posteriori de θ é dada por

$$p(\theta|x) = \frac{\dfrac{1}{1+\theta^2}\exp[-(x-\theta)^2/2]}{\displaystyle\int \frac{1}{1+\theta^2}\exp[-(x-\theta)^2/2]d\theta}.$$

Suponha agora que queremos estimar θ usando função de perda quadrática. Como vimos no Capítulo 3 isto implica em tomar a média a posteriori de θ como estimativa. Mas

$$E[\theta|\boldsymbol{x}] = \int \theta p(\theta|\boldsymbol{x})d\theta = \frac{\displaystyle\int \frac{\theta}{1+\theta^2}\exp[-(x-\theta)^2/2]d\theta}{\displaystyle\int \frac{1}{1+\theta^2}\exp[-(x-\theta)^2/2]d\theta}$$

e as integrais no numerador e denominador não têm solução analítica exata. Uma solução aproximada via simulação de Monte Carlo pode ser obtida usando o seguinte algoritmo,

1. gerar $\theta_1, \dots, \theta_n$ independentes da distribuição $N(x, 1)$;
2. calcular $g_i = \dfrac{\theta_i}{1+\theta_i^2}$ e $g_i^* = \dfrac{1}{1+\theta_i^2}$;
3. calcular $\hat{E}(\theta|\boldsymbol{x}) = \dfrac{\sum_{i=1}^n g_i}{\sum_{i=1}^n g_i^*}$.

Este exemplo ilustrou um problema que geralmente ocorre em aplicações Bayesianas. Como a posteriori só é conhecida a menos de uma constante de proporcionalidade as esperanças a posteriori são na verdade uma razão de integrais. Neste caso, a aproximação é baseada na razão dos dois estimadores de Monte Carlo para o numerador e denominador.

Exercícios

1. Para uma única observação X com distribuição $N(\theta, 1)$, θ desconhecido, queremos fazer inferência sobre θ usando uma priori Cauchy(0,1). Gere um valor de X para $\theta = 2$, i.e. $x \sim N(2, 1)$.

(a) Estime θ através da sua média a posteriori usando o algoritmo do exemplo 1.

(b) Estime a variância da posteriori.

(c) Generalize o algoritmo para k observações $X_1, \dots, X_k$ da distribuição $N(\theta, 1)$.

4.4 Métodos de Reamostragem

Existem distribuições para as quais é muito difícil ou mesmo impossível simular valores. A idéia dos métodos de reamostragem é gerar valores em duas etapas. Na primeira etapa gera-se valores de uma distribuição auxiliar conhecida. Na segunda etapa utiliza-se um mecanismo de correção para que os valores sejam representativos (ao menos aproximadamente) da distribuição a posteriori. Na prática costuma-se tomar a priori como distribuição auxiliar conforme proposto em Smith e Gelfand (1992).

4.4.1 Método de Rejeição

Considere uma densidade auxiliar $q(\theta)$ da qual sabemos gerar valores. A única restrição é que exista uma constante A finita tal que $p(\theta|\boldsymbol{x}) < Aq(\theta)$. O método de rejeição consiste em gerar um valor θ^* da distribuição auxiliar q e aceitar este valor como sendo da distribuição a posteriori com probabilidade $p(\theta|\boldsymbol{x})/Aq(\theta)$. Caso contrário, θ^* não é aceito como uma valor gerado da posteriori e o processo é repetido até que um valor seja aceito. O método também funciona se ao invés da posteriori, que em geral é desconhecida, usarmos a sua versão não normalizada, i.e $p(\boldsymbol{x}|\theta)p(\theta)$.

Tomando a priori $p(\theta)$ como densidade auxiliar a constante A deve ser tal que $p(\boldsymbol{x}|\theta) < A$. Esta desigualdade é satisfeita se tomarmos A como sendo o valor máximo da função de verossimilhança, i.e. $A = p(\boldsymbol{x}|\hat{\theta})$ onde $\hat{\theta}$ é o estimador de máxima verossimilhança de θ. Neste caso, a probabilidade de aceitação se simplifica para $p(\boldsymbol{x}|\theta)/p(\boldsymbol{x}|\hat{\theta})$.

Podemos então usar o seguinte algoritmo para gerar valores da posteriori

1. gerar um valor θ^* da distribuição a priori;

2. gerar $u \sim U(0, 1)$;

3. aceitar θ^* como um valor da posteriori se $u < p(\boldsymbol{x}|\theta^*)/p(\boldsymbol{x}|\hat{\theta})$, caso contrário rejeitar θ^* e retornar ao item 1.

Um problema técnico associado ao método é a necessidade de se maximizar a função de verossimilhança o que pode não ser uma tarefa simples em modelos mais complexos. Se este for o caso então o método de rejeição perde o seu principal atrativo que é a simplicidade. Neste caso, o método da próxima seção passa a ser recomendado.

Outro problema é que a taxa de aceitação pode ser muito baixa, i.e. teremos que gerar muitos valores da distribuição auxiliar até conseguir um número suficiente de valores da posteriori. Isto ocorrerá se as informações da priori e da verossimilhança forem conflitantes já que neste caso os valores gerados terão baixa probabilidade de serem aceitos.

4.4.2 Reamostragem Ponderada

Estes métodos usam a mesma idéia de gerar valores de uma distribuição auxiliar porém sem a necessidade de maximização da verossimilhança. A desvantagem é que os valores obtidos são apenas aproximadamente distribuidos segundo a posteriori.

Suponha que temos uma amostra $\theta_1, \dots, \theta_n$ gerada da distribuição auxiliar q e a partir dela construimos os pesos

$$w_i = \frac{p(\theta_i|\boldsymbol{x})/q(\theta_i)}{\sum_{j=1}^n p(\theta_j|\boldsymbol{x})/q(\theta_j)}, \quad i = 1, \dots, n$$

O método consiste em tomar uma segunda amostra (ou reamostra) de tamanho m da distribuição discreta em $\theta_1, \dots, \theta_n$ com probabilidades $w_1, \dots, w_n$. Aqui também não é necessário que se conheça completamente a posteriori mas apenas o produto priori vezes verossimilhança já que neste caso os pesos não se alteram.

Tomando novamente a priori como densidade auxiliar, i.e. $q(\theta) = p(\theta)$ os pesos se simplificam para

$$w_i = \frac{p(\boldsymbol{x}|\theta_i)}{\sum_{j=1}^n p(\boldsymbol{x}|\theta_j)}, \quad i = 1, \dots, n$$

e o algoritmo para geração de valores (aproximadamente) da posteriori então fica

1. gerar valores $\theta_1, \dots, \theta_n$ da distribuição a priori;
2. calcular os pesos w_i, $i = 1, \dots, n$;
3. reamostrar valores com probabilidades $w_1, \dots, w_n$.

Exercícios

1. Em um modelo de regressão linear simples temos que $y_i \sim N(\beta x_i, 1)$. Os dados observados são $\boldsymbol{y} = (-2, 0, 0, 0, 2)$ e $\boldsymbol{x} = (-2, -1, 0, 1, 2)$, e usamos uma priori vaga $N(0, 4)$ para β. Faça inferência sobre β obtendo uma amostra da posteriori usando reamostragem ponderada. Compare com a estimativa de máxima verossimilhança $\hat{\beta} = 0,8$.

2. Para o mesmo modelo do exercício 1 e os mesmos dados suponha agora que a variância é desconhecida, i.e. $y_i \sim N(\beta x_i, \sigma^2)$. Usamos uma priori hierárquica para (β, σ^2), i.e. $\beta|\sigma^2 \sim N(0, \sigma^2)$ e $\sigma^{-2} \sim G(0,01, 0,01)$.
   (a) Obtenha uma amostra da posteriori de (β, σ^2) usando reamostragem ponderada.
   (b) Baseado nesta amostra, faça um histograma das distribuições marginais de β e σ^2.
   (c) Estime β e σ^2 usando uma aproximação para a média a posteriori. Compare com as estimativas de máxima verossimilhança.

4.5 Monte Carlo via cadeias de Markov

Em todos os métodos de simulação vistos até agora obtém-se uma amostra da distribuição a posteriori em um único passo. Os valores são gerados de forma independente e não há preocupação com a convergência do algoritmo, bastando que o tamanho da amostra seja suficientemente grande. Por isso estes métodos são chamados *não iterativos* (não confundir iteração com interação). No entanto, em muitos problemas pode ser bastante difícil, ou mesmo impossível, encontrar uma densidade de importância que seja simultaneamente uma boa aproximação da posteriori e fácil de ser amostrada.

Os métodos de Monte Carlo via cadeias de Markov (MCMC) são uma alternativa aos métodos não iterativos em problemas complexos. A idéia ainda é obter uma amostra da distribuição a posteriori e calcular estimativas amostrais de características desta distribuição. A diferença é que aqui usaremos técnicas de simulação iterativa, baseadas em cadeias de Markov, e assim os valores gerados não serão mais independentes.

Neste capítulo serão apresentados os métodos MCMC mais utilizados, o amostrador de Gibbs e o algoritmo de Metropolis-Hastings. A idéia básica é simular um passeio aleatório no espaço de θ que converge para uma distribuição estacionária, que é a distribuição de interesse no problema. Uma discussão mais geral sobre o tema pode ser encontrada por exemplo em Gamerman (1997).

4.5.1 Cadeias de Markov

Uma cadeia de Markov é um processo estocástico $\{X_0, X_1, \dots\}$ tal que a distribuição de X_t dados todos os valores anteriores $X_0, \dots, X_{t-1}$ depende apenas de X_{t-1}. Matematicamente,

$$P(X_t \in A|X_0, \dots, X_{t-1}) = P(X_t \in A|X_{t-1})$$

para qualquer subconjunto A. Os métodos MCMC requerem ainda que a cadeia seja,

- homogênea, i.e. as probabilidades de trasição de um estado para outro são invariantes;
- irredutível, i.e. cada estado pode ser atingido a partir de qualquer outro em um número finito de iterações;
- aperiódica, i.e. não haja estados absorventes.

e os algoritmos que serão vistos aqui satisfazem a estas condições.

4.5.2 Algoritmo de Metropolis-Hastings

Os algoritmos de Metropolis-Hastings usam a mesma idéia dos métodos de rejeição vistos no capítulo anterior, i.e. um valor é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade. Este mecanismo de correção garante que a convergência da cadeia para a distribuição de equilibrio, que neste caso é a distribuição a posteriori.

Suponha que a cadeia esteja no estado θ e um valor θ' é gerado de uma *distribuição proposta* $q(\cdot|\theta)$. Note que a distribuição proposta pode depender do estado atual da cadeia, por exemplo $q(\cdot|\theta)$ poderia ser uma distribuição normal centrada em θ. O novo valor θ' é aceito com probabilidade

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left(1, \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right). \tag{4.2}$$

onde π é a distribuição de interesse.

Uma característica importante é que só precisamos conhecer π parcialmente, i.e. a menos de uma constante já que neste caso a probabilidade (4.2) não se altera. Isto é fundamental em aplicações Bayesianas aonde não conhecemos completamente a posteriori.

Em termos práticos, o algoritmo de Metropolis-Hastings pode ser especificado pelos seguintes passos,

1. Inicialize o contador de iterações $t = 0$ e especifique um valor inicial $\theta^{(0)}$.
2. Gere um novo valor θ' da distribuição $q(\cdot|\theta)$.
3. Calcule a probabilidade de aceitação $\alpha(\theta, \theta')$ e gere $u \sim U(0, 1)$.
4. Se $u \leq \alpha$ então aceite o novo valor e faça $\theta^{(t+1)} = \theta'$, caso contrário rejeite e faça $\theta^{(t+1)} = \theta$.
5. Incremente o contador de t para $t + 1$ e volte ao passo 2.

Uma useful feature of the algorithm is that the target distribution needs only be known up to a constant of proportionality since only the target ratio

$\pi(\boldsymbol{\theta}')/\pi(\boldsymbol{\theta})$ is used in the acceptance probability. Note also that the chain may remain in the same state for many iterations and in practice a useful monitoring device is given by the average percentage of iterations for which moves are accepted. Hastings (1970) suggests that this acceptance rate should always be computed in practical applications.

The independence sampler is a Metropolis-Hastings algorithm whose proposal distribution does not depend on the current state of the chain, i.e., $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}')$. In general, $q(\cdot)$ should be a good approximation of $\pi(\cdot)$, but it is safest if $q(\cdot)$ is heavier-tailed than $\pi(\cdot)$.

The Metropolis algorithm considers only symmetric proposals, i.e., $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}', \boldsymbol{\theta})$ for all values of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, and the acceptance probability reduces to

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left(1, \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}\right).$$

A special important case is the random-walk Metropolis for which $q(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(|\boldsymbol{\theta} - \boldsymbol{\theta}'|)$, so that the probability of generating a move from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ depends only on the distance between them. Using a proposal distribution with variance σ^2, very small values of σ^2 will lead to small jumps which are almost all accepted but it will be difficult to traverse the whole parameter space and it will take many iterations to converge. On the other hand, large values of σ^2 will lead to an excessively high rejection rate since the proposed values are likely to fall in the tails of the posterior distribution.

Typically, there will be an optimal value for the proposal scale σ determined on the basis of a few pilot runs which lies in between these two extremes (see for example, Roberts, Gelman and Gilks, 1997). We return to this point later and, in particular, discuss an approach for choosing optimal values for the parameters of the proposal distribution for (RJ)MCMC algorithms in Chapter 6.

4.5.3 Amostrador de Gibbs

while in the Gibbs sampler the chain will always move to a new value. Gibbs sampling is an MCMC scheme where the transition kernel is formed by the full conditional distributions, $\pi(\theta_i|\boldsymbol{\theta}_{-i})$, where $\boldsymbol{\theta}_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_d)'$. In general, each one of the components θ_i can be either uni- or multi-dimensional. So, the full conditional distribution is the distribution of the ith component of $\boldsymbol{\theta}$ conditioning on all the remaining components, and it is derived from the joint distribution as follows,

$$\pi(\theta_i|\boldsymbol{\theta}_{-i}) = \frac{\pi(\boldsymbol{\theta})}{\displaystyle\int \pi(\boldsymbol{\theta}) d\theta_i}.$$

If generation schemes to draw a sample directly from $\pi(\boldsymbol{\theta})$ are costly, complicated or simply unavailable but the full conditional distributions are completely known and can be sampled from, then Gibbs sampling proceeds as follows,

1. Initialize the iteration counter of the chain $t = 1$ and set initial values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \dots, \theta_d^{(0)})'$.
2. Obtain a new value of $\boldsymbol{\theta}^{(t)}$ from $\boldsymbol{\theta}^{(t-1)}$ through successive generation of values

$$
\begin{array}{rcl}
\theta_1^{(t)} & \sim & \pi(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)}) \\
\theta_2^{(t)} & \sim & \pi(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \dots, \theta_d^{(t-1)}) \\
 & \vdots & \\
\theta_d^{(t)} & \sim & \pi(\theta_d|\theta_1^{(t)}, \theta_2^{(t)}, \dots, \theta_{d-1}^{(t)})
\end{array}
$$

3. Increment the counter t to $t + 1$ and return to step 2 until convergence is reached.

So, each iteration is completed after d moves along the coordinates axes of the components of $\boldsymbol{\theta}$. When convergence is reached, the resulting value $\boldsymbol{\theta}$ is a draw from $\pi(\boldsymbol{\theta})$. It is worth noting that, even in a high-dimensional problem, all of the simulations may be univariate, which is usually a computational advantage.

However, the Gibbs sampler does not apply to problems where the number of parameters varies because of the lack of irreducibility of the resulting chain. When the length of $\boldsymbol{\theta}$ is not fixed and its elements need not have a fixed interpretation across all models, to resample some components conditional on the remainder would rarely be meaninful.

Note also that the Gibbs sampler is a special case of the Metropolis-Hastings algorithm, in which individual elements of $\boldsymbol{\theta}$ are updated one at a time (or in blocks), with the full conditional distribution as the candidate generating function and acceptance probabilities uniformly equal to 1.

4.5.4 Updating strategies

In the above scheme, all the components of $\boldsymbol{\theta}$ are updated in the same deterministic order at every iteration. However, other scanning or updating strategies are possible for visiting the components of $\boldsymbol{\theta}$. Geman and Geman (1984) showed in a discrete setting that any updating scheme that guarantees that all components are visited infinitely often when the chain is run indefinitely, converges to the joint distribution of interest, i.e., $\pi(\boldsymbol{\theta})$. For example, Zeger and Karim (1991) describe a Gibbs sampling scheme where some components are visited only every kth iteration, which still guarantees that every component is updated infinitely often for finite, fixed k.

Roberts and Sahu (1997) consider a random permutation scan where at each iteration a permutation of $\{1, \dots, d\}$ is chosen and components are visited in that order. In particular, they showed that when π is multivariate normal, convergence for the deterministic scan is faster than for the random scan if the precision matrix is tridiagonal (θ_i depends only on θ_{i-1} and θ_{i+1}) or if it has non-negative partial correlations.

4.5.5 Blocking

In principle, the way the components of the parameter vector $\boldsymbol{\theta}$ are arranged in blocks of parameters is completely arbitrary and includes blocks formed by scalar components as special cases. However, the structure of the Gibbs sampler imposes moves according to the coordinate axes of the blocks, so that larger blocks allow moves in more general directions. This can be very beneficial, although more computationally demanding, in a context where there is high correlation between individual components since these higher dimensional moves incorporate information about this dependence. Parameter values are then generated from the joint full conditional distribution for the block of parameters considered.

Roberts and Sahu (1997) showed that for a multivariate normal π and random scans, convergence improves as the number of blocks decreases. They also proved that blocking can hasten convergence for non-negative partial correlation distributions and even more as the partial correlation of the components in the block gets larger. However, they also provided an example where blocking worsens convergence.

4.5.6 Completion

Even when every full conditional distribution associated with the target distribution π is not explicit there can be a density π^* for which π is a marginal density, i.e.,

$$\int \pi^*(\boldsymbol{\theta}, \boldsymbol{z}) d\boldsymbol{z} = \pi(\boldsymbol{\theta})$$

and such that all the full conditionals associated with π^* are easy to simulate from. Then the Gibbs sampler can be implemented in π^* instead of π and this is called the completion Gibbs sampler because π^* is a completion of π. The required sample from the target distribution is obtained by marginalizing again, i.e., integrating $\boldsymbol{z}$ out.

This approach was actually one of the first appearances of the Gibbs sampler in Statistics with the introduction of data augmentation by Tanner and Wong (1987). It is also worth noting that, in principle, this Gibbs sampler does not require that the completion of π into π^* and of $\boldsymbol{\theta}$ into $(\boldsymbol{\theta}, \boldsymbol{z})$ should be related to the problem of interest and the vector $\boldsymbol{z}$ might have no meaning from a statistical point of view.

4.5.7 The Slice Sampler

This is a very general version of the Gibbs sampler which applies to most distributions and is based on the simulation of specific uniform random variables. In its simplest version when only one variable is being updated, if π can be written as a product of functions, i.e.,

$$\pi(\theta) = \prod_{i=1}^{k} f_i(\theta),$$

where f_i are positive functions but not necessarily densities then f can be completed (or demarginalised) into

$$\prod_{i=1}^{k} I_{0<z_i<f_i(\theta)}.$$

The slice sampler consists of generating $(z_1, \dots, z_k, \theta)$ from their full conditional distributions, i.e.,

- generate $z_i^{(t+1)}$ from $U[0, f_i(\theta^{(t)})]$, $i = 1, \dots, k$ and
- generate $\theta^{(t+1)}$ from a uniform distribution in $A^{(t+1)} = \{y : f_i(y) > z_i^{(t+1)}, i = 1, \dots, k\}$.

Roberts and Rosenthal (1998) study the slice sampler and show that it usually enjoys good theoretical properties. In practice there may be problems as d increases since the determination of the set $A^{(t+1)}$ may get increasingly complex.

Further details about the Gibbs sampler and related algorithms are given, for example, in Gamerman (1997, Chapter 5) and Robert and Casella (1999, Chapter 7).

4.6 Posterior Model Probabilities

The posterior model probability is obtained as

$$p(k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|k)p(k)}{p(\boldsymbol{y})}$$

where the term $p(\boldsymbol{y}|k)$ is sometimes referred to as the marginal likelihood for model k and is calculated as

$$p(\boldsymbol{y}|k) = \int p(\boldsymbol{y}|\boldsymbol{\theta}^{(k)}, k)p(\boldsymbol{\theta}^{(k)}|k)d\boldsymbol{\theta}^{(k)}.$$

Also, $1/p(\boldsymbol{y}|k)$ is the normalisation constant for $p(\boldsymbol{\theta}^{(k)}|k, \boldsymbol{y})$, the posterior density of $\boldsymbol{\theta}$ within model k.

Hence, the posterior probability of a certain model is proportional to the product of the prior probability and the marginal likelihood for that model. It is also worth noting that, in practice, $p(\boldsymbol{y})$ is unknown so that typically the model probabilities are known only up to a normalisation constant.

The above integral is commonly analytically intractable but may be approximated in a number of ways by observing that it can be regarded as the expected value of the likelihood with respect to the prior distribution $p(\boldsymbol{\theta}^{(k)}|k)$. In terms of simulation techniques, the simplest estimate consists of simulating n values $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from the prior, evaluating the likelihood at those values and computing the Monte Carlo estimate

$$\hat{p}(\boldsymbol{y}|k) = \frac{1}{n}\sum_{i=1}^{n} p(\boldsymbol{y}|\boldsymbol{\theta}_i^{(k)}, k).$$

This estimator has high variance with possibly few terms contributing substantially to the sum in cases of disagreement between prior and likelihood. Various alternative estimators are reviewed in Gamerman (1997, Chapter 7) and analytical approximations supported by asymptotic normal theory might also be used. Other alternatives will be explored in the next section.

Having obtained the posterior model probabilities these may be used for either selecting the model with the highest probability or highest Bayes factor from the list of candidate models (model selection), or estimating some quantity under each model and then averaging the estimates according to how likely each model is, that is, using these probabilities as weights (model averaging). In the next section, we present MCMC methods that take into account different models simultaneously.

Capítulo 5

Exercícios

5.1 Lista de exercícios 1

1. No exemplo dos físicos nas notas de aula, obtenha também a distribuição preditiva de X e compare o valor observado com a média desta preditiva para os 2 físicos. Faça uma previsão para uma $2^{\underline{a}}$ medição Y feita com o mesmo aparelho.

2. Uma máquina produz 5% de itens defeituosos. Cada item produzido passa por um teste de qualidade que o classifica como "bom ", "defeituoso "ou "suspeito ". Este teste classifica 20% dos itens defeituosos como bons e 30% como suspeitos. Ele também classifica 15% dos itens bons como defeituosos e 25% como suspeitos.

   (a) Que proporção dos itens serão classificados como suspeitos ?

   (b) Qual a probabilidade de um item classificado como suspeito ser defeituoso ?

   (c) Outro teste, que classifica 95% dos itens defeituosos e 1% dos itens bons como defeituosos, é aplicado somente aos itens suspeitos.

   (d) Que proporção de itens terão a suspeita de defeito confirmada ?

   (e) Qual a probabilidade de um item reprovado neste $2^{\underline{o}}$ teste ser defeituoso ?

5.2 Lista de exercícios 2

1. Mostre que a família de distribuições Beta é conjugada em relação às distribuições amostrais binomial, geométrica e binomial negativa.

2. Para uma amostra aleatória de 100 observações da distribuição normal com média θ e desvio-padrão 2 foi especificada uma priori normal para θ.

   (a) Mostre que o desvio-padrão a posteriori será sempre menor do que $1/5$. Interprete este resultado.

   (b) Se o desvio-padrão a priori for igual a 1 qual deve ser o menor número de observações para que o desvio-padrão a posteriori seja 0,1?

3. Seja $X_1, \dots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ conhecido. Utilizando uma distribuição a priori Gama para σ^{-2} com coeficiente de variação 0,5, qual deve ser o tamanho amostral para que o coeficiente de variação a posteriori diminua para 0,1?

4. Seja $X_1, \dots, X_n$ uma amostra aleatória da distribuição $N(\theta, \sigma^2)$, com θ e σ^2 desconhecidos, e considere a priori conjugada de (θ, ϕ).

   (a) Determine os parâmetros $(\mu_0, c_0, n_0, \sigma_0^2)$ utilizando as seguintes informações a priori: $E(\theta) = 0$, $P(|\theta| < 1,412) = 0,5$, $E(\phi) = 2$ e $E(\phi^2) = 5$.

   (b) Em uma amostra de tamanho $n = 10$ foi observado $\overline{X} = 1$ e $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 8$. Obtenha a distribuição a posteriori de θ e esboce os gráficos das distribuições a priori, a posteriori e da função de verossimilhança, com ϕ fixo.

   (c) Calcule $P(|Y| > 1|\boldsymbol{x})$ onde Y é uma observação tomada da mesma população.

5. Suponha que o tempo, em minutos, para atendimento a clientes segue uma distribuição exponencial com parâmetro θ desconhecido. Com base na experiência anterior assume-se uma distribuição a priori Gama com média 0,2 e desvio-padrão 1 para θ.

   (a) Se o tempo médio para atender uma amostra aleatória de 20 clientes foi de 3,8 minutos, qual a distribuição a posteriori de θ.

   (b) Qual o menor número de clientes que precisam ser observados para que o coeficiente de variação a posteriori se reduza para 0,1?

5.3 Lista de exercícios 3

1. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição de Poisson com parâmetro θ.

   (a) Determine os parâmetros da priori conjugada de θ sabendo que $E(\theta) = 4$ e o coeficiente de variação a priori é 0,5.

   (b) Quantas observações devem ser tomadas até que a variância a posteriori se reduza para 0,01 ou menos?

   (c) Mostre que a média a posteriori é da forma $\gamma_n \overline{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \to 1$ quando $n \to \infty$. Interprete este resultado.

2. O número médio de defeitos por 100 metros de uma fita magnética é desconhecido e denotado por θ. Atribui-se uma distribuição a priori Gama(2,10) para θ. Se um rolo de 1200 metros desta fita foi inspecionado e encontrou-se 4 defeitos qual a distribuição a posteriori de θ?

3. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição Bernoulli com parâmetro θ e usamos a priori conjugada $Beta(a, b)$. Mostre que a média a posteriori é da forma $\gamma_n \overline{x} + (1 - \gamma_n)\mu_0$, onde $\mu_0 = E(\theta)$ e $\gamma_n \to 1$ quando $n \to \infty$. Interprete este resultado.

4. Para uma amostra aleatória $X_1, \ldots, X_n$ tomada da distribuição $U(0, \theta)$, mostre que a família de distribuições de Pareto com parâmetros a e b, cuja função de densidade é $p(\theta) = ab^a/\theta^{a+1}$, é conjugada à uniforme.

5. Para uma variável aleatória $\theta > 0$ a família de distribuições Gama-invertida tem função de densidade de probabilidade dada por

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \quad \alpha, \beta > 0.$$

   Mostre que esta família é conjugada ao modelo normal com média μ conhecida e variância θ desconhecida.

5.4 Lista de exercícios 4

1. Suponha que $\boldsymbol{X} = (X_1, X_2, X_3)$ tenha distribuição trinomial com parâmetros n (conhecido) e $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ com $\pi_1 + \pi_2 + \pi_3 = 1$. Mostre que a priori não informativa de Jeffreys para $\boldsymbol{\pi}$ é $p(\boldsymbol{\pi}) \propto [\pi_1 \pi_2 (1 - \pi_1 - \pi_2)]^{-1/2}$.

2. Para cada uma das distribuições abaixo verifique se o modelo é de locação, escala ou locação-escala e obtenha a priori não informativa para os parâmetros desconhecidos.

   (a) Cauchy(0,β).

   (b) $t_\nu(\mu, \sigma^2)$, ν conhecido.

   (c) Pareto(a, b), b conhecido.

   (d) Uniforme $(\theta - 1, \theta + 1)$.

   (e) Uniforme $(-\theta, \theta)$.

3. Seja uma coleção de variáveis aleatórias independentes X_i com distribuições $p(x_i|\theta_i)$ e seja $p_i(\theta_i)$ a priori não informativa de θ_i, $i = 1, \ldots, k$. Mostre que a priori não informativa de Jeffreys para o vetor paramétrico $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ é dada por $\prod_{i=1}^k p_i(\theta_i)$.

4. Se θ tem priori não informativa $p(\theta) \propto k$, $\theta > 0$ mostre que a priori de $\phi = a\theta + b$, $a \neq 0$ também é $p(\phi) \propto k$.

5. Se θ tem priori não informativa $p(\theta) \propto \theta^{-1}$ mostre que a priori de $\phi = \theta^a$, $a \neq 0$ também é $p(\phi) \propto \phi^{-1}$ e que a priori de $\psi = \log \theta$ é $p(\psi) \propto k$.

5.5 Lista de exercícios 5

Resolva estes problemas usando o pacote estatístico R. Entregue os resultados juntamente com os comandos que utilizou.

1. Ensaios de Bernoulli.

   (a) Gere uma amostra aleatória de tamanho 10 da distribuição de Bernoulli com probabilidade de sucesso $\theta = 0,8$

   (b) Faça um gráfico com as funções de densidade das prioris conjugadas Beta(6,2), Beta(2,6), Beta(1,1).

   (c) Repita o gráfico anterior acrescentando a função de verossimilhança. Note que a verossimilhança deve ser normalizada.

   (d) Faça um gráfico com as funções de densidade das posterioris usando as prioris acima e mais a priori não informativa de Jeffreys. O que você conclui?

   (e) Repita o item anterior com uma amostra de tamanho 100. O que você conclui?

2. Modelo de Poisson.

   (a) Gere uma amostra aleatória de tamanho 10 da distribuição de Poisson com média $\theta = 2,0$

   (b) Faça um gráfico com as funções de densidade das prioris conjugadas Gama(5,2), Gama(2,5), Gama(1,1).

   (c) Repita o gráfico anterior acrescentando a função de verossimilhança. Note que a verossimilhança deve ser normalizada.

   (d) Faça um gráfico com as funções de densidade das posterioris usando as prioris acima e mais a priori não informativa de Jeffreys. O que você conclui?

   (e) Repita o item anterior com uma amostra de tamanho 100. O que você conclui?

5.6 Lista de exercícios 6

Resolva estes problemas usando o pacote estatístico R. Entregue os resultados juntamente com os comandos que utilizou.

1. Para uma única observação X com distribuição $N(\theta, 1)$, θ desconhecido, queremos fazer inferência sobre θ usando uma priori Cauchy(0,1). Gere um valor de X para $\theta = 2$, i.e. $x \sim N(2, 1)$.

   (a) Estime θ através da sua média a posteriori usando o algoritmo do exemplo 4.1 das notas de aula.

   (b) Estime a variância da posteriori.

   (c) Generalize o algoritmo para k observações $X_1, \dots, X_k$ da distribuição $N(\theta, 1)$.

2. Em um modelo de regressão linear simples temos que $y_i \sim N(\beta x_i, 1)$. Os dados observados são $\boldsymbol{y} = (-2, 0, 0, 0, 2)$ e $\boldsymbol{x} = (-2, -1, 0, 1, 2)$, e usamos uma priori vaga $N(0, 4)$ para β.

   (a) Obtenha uma amostra da posteriori de β usando reamostragem ponderada.

   (b) Baseado nesta amostra, faça um histograma e estime β usando uma aproximação para a média a posteriori. Compare com a estimativa de máxima verossimilhança $\hat{\beta} = 0, 8$.

3. Para o mesmo modelo do exercício 1 e os mesmos dados suponha agora que a variância é desconhecida, i.e. $y_i \sim N(\beta x_i, \sigma^2)$. Usamos uma priori hierárquica para (β, σ^2), i.e. $\beta|\sigma^2 \sim N(0, \sigma^2)$ e $\sigma^{-2} \sim G(0, 01, 0, 01)$.

   (a) Obtenha uma amostra da posteriori de (β, σ^2) usando reamostragem ponderada.

   (b) Baseado nesta amostra, faça um histograma das distribuições marginais de β e σ^2.

   (c) Estime β e σ^2 usando uma aproximação para a média a posteriori. Compare com as estimativas de máxima verossimilhança.

Apêndice A

Lista de Distribuições

Neste apêndice são listadas as distribuições de probabilidade utilizadas no texto para facilidade de referência. São apresentadas suas funções de (densidade) de probabilidade além da média e variância. Uma revisão exaustiva de distribuições de probabilidades pode ser encontrada em Johnson *et al.* (1992, 1994, 1995).

A.1 Distribuição Normal

X tem distribuição normal com parâmetros μ e σ^2, denotando-se $X \sim N(\mu, \sigma^2)$, se sua função de densidade é dada por

$$p(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/2\sigma^2], \quad -\infty < x < \infty,$$

para $-\infty < \mu < \infty$ e $\sigma^2 > 0$. Quando $\mu = 0$ e $\sigma^2 = 1$ a distribuição é chamada normal padrão. A distribuição log-normal é definida como a distribuição de e^X.

No caso vetorial, $\boldsymbol{X} = (X_1, \dots, X_p)$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de variância-covariância Σ, denotando-se $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ se sua função de densidade é dada por

$$p(x|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp[-(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2]$$

para $\boldsymbol{\mu} \in \mathbb{R}^p$ e Σ positiva-definida.

A.2 Distribuição Gama

X tem distribuição Gama com parâmetros α e β, denotando-se $X \sim Ga(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,$$

para $\alpha, \beta > 0$.

$$E(X) = \alpha/\beta \quad \text{e} \quad V(X) = \alpha/\beta^2.$$

Casos particulares da distribuição Gama são a distribuição de Erlang, $Ga(\alpha, 1)$, a distribuição exponencial, $Ga(1, \beta)$, e a distribuição qui-quadrado com ν graus de liberdade, $Ga(\nu/2, 1/2)$.

A.3 Distribuição Gama Inversa

X tem distribuição Gama Inversa com parâmetros α e β, denotando-se $X \sim GI(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x}, \quad x > 0,$$

para $\alpha, \beta > 0$.

$$E(X) = \frac{\beta}{\alpha - 1} \quad \text{e} \quad V(X) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)}.$$

Não é difícil verificar que esta é a distribuição de $1/X$ quando $X \sim Ga(\alpha, \beta)$.

A.4 Distribuição Beta

X tem distribuição Beta com parâmetros α e β, denotando-se $X \sim Be(\alpha, \beta)$, se sua função de densidade é dada por

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \quad 0 < x < 1,$$

para $\alpha, \beta > 0$.

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{e} \quad V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

A.5 Distribuição de Dirichlet

O vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_k)$ tem distribuição de Dirichlet com parâmetros $\alpha_1, \ldots, \alpha_k$, denotada por $D_k(\alpha_1, \ldots, \alpha_k)$ se sua função de densidade conjunta é dada por

$$p(\boldsymbol{x}|\alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \ldots, \Gamma(\alpha_k)} x_1^{\alpha_1-1} \ldots x_k^{\alpha_k-1}, \quad \sum_{i=1}^{k} x_i = 1,$$

para $\alpha_1, \ldots, \alpha_k > 0$ e $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

$$E(X_i) = \frac{\alpha_i}{\alpha_0}, \quad V(X_i) = \frac{(\alpha_0 - \alpha_i)\alpha_i}{\alpha_0^2(\alpha_0 + 1)}, \quad \text{e} \quad Cov(X_i, X_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Note que a distribuição Beta é obtida como caso particular para $k = 2$.

A.6 Distribuição t de Student

X tem distribuição t de Student (ou simplesmente t) com média μ, parâmetro de escala σ e ν graus de liberdade, denotando-se $X \sim t_\nu(\mu, \sigma^2)$, se sua função de densidade é dada por

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma((\nu+1)/2)\nu^{\nu/2}}{\Gamma(\nu/2)\sqrt{\pi}\sigma}\left[\nu + \frac{(x-\mu)^2}{\sigma^2}\right]^{-(\nu+1)/2}, \quad x \in \mathbb{R},$$

para $\nu > 0$, $\mu \in \mathbb{R}$ e $\sigma^2 > 0$.

$$E(X) = \mu, \quad \text{para } \nu > 1 \quad \text{e} \quad V(X) = \frac{\nu}{\nu - 2}, \quad \text{para } \nu > 2.$$

Um caso particular da distribuição t é a distribuição de Cauchy, denotada por $C(\mu, \sigma^2)$, que corresponde a $\nu = 1$.

A.7 Distribuição F de Fisher

X tem distribuição F com ν_1 e ν_2 graus de liberdade, denotando-se $X \sim F(\nu_1, \nu_2)$, se sua função de densidade é dada por

$$p(x|\nu_1, \nu_2) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}x^{\nu_1/2-1}(\nu_2 + \nu_1 x)^{-(\nu_1+\nu_2)/2}$$

$x > 0$, e para $\nu_1, \nu_2 > 0$.

$$E(X) = \frac{\nu_2}{\nu_2 - 2}, \quad \text{para } \nu_2 > 2 \quad \text{e} \quad V(X) = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)(\nu_2 - 2)^2}, \quad \text{para } \nu_2 > 4.$$

A.8 Distribuição Binomial

X tem distribuição binomial com parâmetros n e p, denotando-se $X \sim bin(n, p)$, se sua função de probabilidade é dada por

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

para $n \geq 1$ e $0 < p < 1$.

$$E(X) = np \quad \text{e} \quad V(X) = np(1-p)$$

e um caso particular é a distribuição de Bernoulli com $n = 1$.

A.9 Distribuição Multinomial

O vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_k)$ tem distribuição multinomial com parâmetros n e probabilidades $\theta_1, \ldots, \theta_k$, denotada por $M_k(n, \theta_1, \ldots, \theta_k)$ se sua função de probabilidade conjunta é dada por

$$p(\boldsymbol{x}|\theta_1, \ldots, \theta_k) = \frac{n!}{x_1!, \ldots, x_k!}\theta_1^{x_1}, \ldots, \theta_k^{x_k}, \quad x_i = 0, \ldots, n, \quad \sum_{i=1}^k x_i = n,$$

para $0 < \theta_i < 1$ e $\sum_{i=1}^k \theta_i = 1$. Note que a distribuição binomial é um caso especial da multinomial quando $k = 2$. Além disso, a distribuição marginal de cada X_i é binomial com parâmetros n e θ_i e

$$E(X_i) = n\theta_i, \quad V(X_i) = n\theta_i(1-\theta_), \quad \text{e} \quad Cov(X_i, X_j) = -n\theta_i\theta_j.$$

A.10 Distribuição de Poisson

X tem distribuição de Poisson com parâmetro θ, denotando-se $X \sim Poisson(\theta)$, se sua função de probabilidade é dada por

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

para $\theta > 0$.

$$E(X) = V(X) = \theta.$$

A.11 Distribuição Binomial Negativa

X tem distribuição de binomial negativa com parâmetros r e p, denotando-se $X \sim BN(r, p)$, se sua função de probabilidade é dada por

$$p(x|r, p) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, \ldots$$

para $r \geq 1$ e $0 < p < 1$.

$$E(X) = r(1-p)/p \quad \text{e} \quad V(X) = r(1-p)/p^2.$$

Referências

Bayes, T. (1763). An essay towards solving in the doctrine of chances. *Philosophical Transactions of the Royal Society London.*

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory.* Wiley: New York.

Box, G. E. P. and G. C. Tiao (1992). *Bayesian Inference in Statistical Analysis.* Wiley Classics Library ed. Wiley-Interscience.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill Book Co.

Gamerman, D. (1996). *Simulação Estocástica via Cadeias de Markov.* Associação Brasileira de Estatística. Minicurso do 12º SINAPE.

Gamerman, D. (1997). *Markov chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Texts in Statistical Sciences. Chapman and Hall, London.

Gamerman, D. and H. S. Migon (1993). *Inferência Estatística: Uma Abordagem Integrada.* Textos de Métodos Matemáticos. Instituto de Matemática, UFRJ.

O'Hagan, A. (1994). *Bayesian Inference*, Volume 2B. Edward Arnold, Cambridge.