

Distribuições a Priori

(Versão em preparação)

GUSTAVO L. GILARDONI

2 de Abril de 2022

Conteúdo: prioris conjugadas, “prioris” impróprias, prioris informativas e não informativas.

Referências: Capítulo 3 do texto *The Bayesian Choice* de C. Robert, entre outros.

0 Introdução

Comentários gerais

- Os termos “a priori” e “a posteriori” não são conceitos absolutos. Eles dependem do contexto (o momento no tempo, ou mais precisamente o que a gente sabe que ocorreu)
- Suponha uma observação (x, y) com verossimilhança $p(x, y | \theta) = p(x | \theta) p(y | x, \theta)$.
 - $p(\theta | x) \propto p(\theta) p(x | \theta)$
 - $p(\theta | x, y) \propto p(\theta) p(x, y | \theta) \propto \{p(\theta) p(x | \theta)\} p(y | x, \theta)$
- A nota anterior sugere que pode-se (recomenda-se!) que as distribuições a priori sejam baseadas em informação proporcionada por experimentos anteriores, caso seja possível.
- Da mesma forma, a conveniência ou não de usar uma certa priori depende fortemente do problema sob estudo (veja a questão 12 da Lista 2!)
- Visto que $p(\theta | x) \propto p(\theta) p(x | \theta)$, devemos concentrar o esforço para especificar *elucidar* / *elicitate* a priori na região onde (esperamos que) a verossimilhança vai ser significativamente maior do que zero.

1 Famílias Conjugadas

- Exponential family: $p(x | \theta) \propto c(\theta) \exp\{\sum_{j=1}^p T_j(x) \eta_j(\theta)\}$;
- Uma priori da forma $p(\theta) \propto b(\theta) \exp\{\sum_{j=1}^p \alpha_j \eta_j(\theta)\}$ vai ser conjugada;
- Usualmente chamada a família *conjugada natural*.
- Beta-Binomial, Multinomial-Dirichlet, Poisson-Gama, Exponencial-Gama (mais geral: Gama com α fixo-Gama); Normal-Normal (variância conhecida), Normal-Normal-Gama (precisão desconhecida) etc.
- Priori conjugada fora de famílias exponenciais: Uniforme(0, θ)-Pareto

1.1 Inferência para a média de uma distribuição Normal: Variância conhecida

$$p(x_1, \dots, x_n | \mu) \propto \exp\left\{-\sum_{i=1}^n (x_i - \mu)^2 / 2\sigma^2\right\}$$

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

Considere $X_1, \dots, X_n | \mu \stackrel{iid}{\sim} N(\mu, \sigma^2)$ com σ^2 conhecida e suponha que a priori $\mu \sim N(\mu_0, \tau^2)$. Para calcular a distribuição a posteriori o seguinte Lema será muito útil.

Lema 1. *Seja Y uma variável aleatória que toma valores em toda a reta real tal que a sua densidade é $f(y) = k e^{-Q(y)}$, onde k é uma constante que não depende de y e $Q(y) = ay^2 + by + c$ ($a > 0$) é uma forma quadrática em y . Logo, Y segue uma distribuição Normal com média e variância*

$$\mathbb{E}(Y) = -\frac{b}{2a} = -\frac{Q'(0)}{Q''(0)} \quad e \quad \text{Var}(Y) = \frac{1}{2a} = \frac{1}{Q''(0)}.$$

Demonstração. Primeiro completamos quadrados,

$$ay^2 + by + c = a \left(y + \frac{b}{2a} \right)^2 + c - \frac{b^2}{4a},$$

de forma que é possível escrever

$$f(y) = k' \exp \left\{ -a \left(y + \frac{b}{2a} \right)^2 \right\},$$

onde $k' = k \exp\{c - b^2/(4a)\}$ continua sendo uma constante que não depende de y . Dessa forma, a densidade de y é proporcional a uma densidade Normal com média $\mu = -\frac{b}{2a}$ e variância $\sigma^2 = \frac{1}{2a}$ e, como é uma densidade, a constante de proporcionalidade tem que ser igual a um. \square

Retornando ao problema do início desta Seção, para achar a distribuição a posteriori note que (i) $p(\mu) \propto \exp\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\}$ e (ii) como $\bar{x} = \sum_{i=1}^n x_i$ é um estatístico suficiente e $\bar{x} | \mu \sim N(\mu, \sigma^2/n)$, $p(x_1, \dots, x_n | \mu) \propto p(\bar{x} | \mu) \propto \exp\{-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\}$ (aqui e em todo o exemplo “ \propto ” significa “proporcional a menos de uma constante que não depende de μ ”). Logo,

$$\begin{aligned} p(\mu | x_1, \dots, x_n) &= \frac{p(\mu) p(x_1, \dots, x_n | \mu)}{\int_0^1 p(\mu) p(x_1, \dots, x_n | \mu) d\mu} \propto p(\mu) p(x_1, \dots, x_n | \mu) \propto p(\mu) p(\bar{x} | \mu) \\ &\propto \exp \left\{ -\frac{1}{2\tau^2} (\mu - \mu_0)^2 \right\} \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \right\} \\ &\propto \exp \left\{ -\left[\frac{1}{2\tau^2} (\mu - \mu_0)^2 + \frac{n}{2\sigma^2} (\bar{x} - \mu)^2 \right] \right\}. \quad (1) \end{aligned}$$

Definindo $Q(\mu) = \frac{1}{2\tau^2}(\mu - \mu_0)^2 + \frac{n}{2\sigma^2}(\bar{x} - \mu)^2$ e usando o Lema 1, segue que a distribuição a posteriori é Normal com média e variâncias dadas por

$$\mu^* = \mathbb{E}(\mu | x_1, \dots, x_n) = -\frac{Q'(0)}{Q''(0)} = \frac{\frac{1}{\tau^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \quad (2)$$

e

$$\tau^{*2} = \text{Var}(\mu | x_1, \dots, x_n) = \frac{1}{Q''(0)} = \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}. \quad (3)$$

Algumas vezes na literatura Bayesiana a variância da distribuição a priori é parametrizada pelo recíproco $\phi = \tau^{-2}$, denominada de *precisão*. De forma semelhante, a *precisão* da distribuição amostral de \bar{x} é $[\text{Var}(\bar{x} | \mu)]^{-1} = n/\sigma^2$. Com essa nova terminologia, a equação (3) diz simplesmente que “a *precisão a posteriori* $[\text{Var}(\mu | x_1, \dots, x_n)]^{-1}$ é igual a *precisão a priori* τ^{-2} mais a *precisão amostral* n/σ^2 ”. De forma semelhante, a equação (2) diz que “a *média a posteriori* é uma *média ponderada da média a priori* μ_0 e da *média amostral* \bar{x} com pesos proporcionais as respectivas *precisões*”. Por exemplo, quanto maior for o tamanho amostral n , se τ^2 e σ^2 , mais perto de \bar{x} vai estar a média a posteriori.

Exemplo 1. Um objeto foi pesado $n = 10$ vezes numa balança retornando média $\bar{x} = 3.53\text{g}$. Segundo o fabricante da balança, para esse tipo de objetos a balança é muito precisa e o desvio padrão das pesadas deve ser da ordem de $\sigma = 0.2\text{g}$.

Uma outra balança, menos precisa, também estava disponível e nela o objeto foi pesado $m = 50$ vezes retornando média e desvio padrão iguais a $\bar{y} = 3.25$ e $s_y = 0.5\text{g}$.

Construa um modelo para analisar esses dados e com base nele obtenha um intervalo que contem μ com probabilidade a posteriori igual a 95%.

Podemos usar a informação da balança menos precisa para justificar uma distribuição a priori $\mu \sim \text{Normal}(3.25, 0.5^2)$ (formalmente, a distribuição a posteriori resultante de usar uma priori não informativa com os dados da balança menos precisa seria t de Student com 49 graus de liberdade—veja as Seções 1.2.1 e 1.2.2 abaixo, mas como n é grande, a distribuição Normal é uma boa aproximação). Logo, podemos tomar $\mu_0 = \bar{y} = 3.25$ e $\tau^2 = s_y^2 = 0.5^2$. Usando o valor de $\bar{x} = 3.53$ usamos as equações (2) e (3) para calcular

$$\mu^* = \frac{\frac{1}{0.5^2}(3.25) + \frac{10}{0.2^2}3.53}{\frac{1}{0.5^2} + \frac{10}{0.2^2}} \doteq 3.5256$$

e

$$\tau^{*2} = \frac{1}{\frac{1}{0.5^2} + \frac{10}{0.2^2}} \doteq (0.0627)^2.$$

Assim a posteriori, $\mu | x_1, \dots, x_n \sim \text{Normal}(3.5256, 0.0627^2)$. O intervalo desejado é então

$$\mu^* \pm z_{\alpha/2} \tau^* \doteq 3.5256 \pm (1.96)(0.0627) \doteq (3.4026; 3.6486).$$

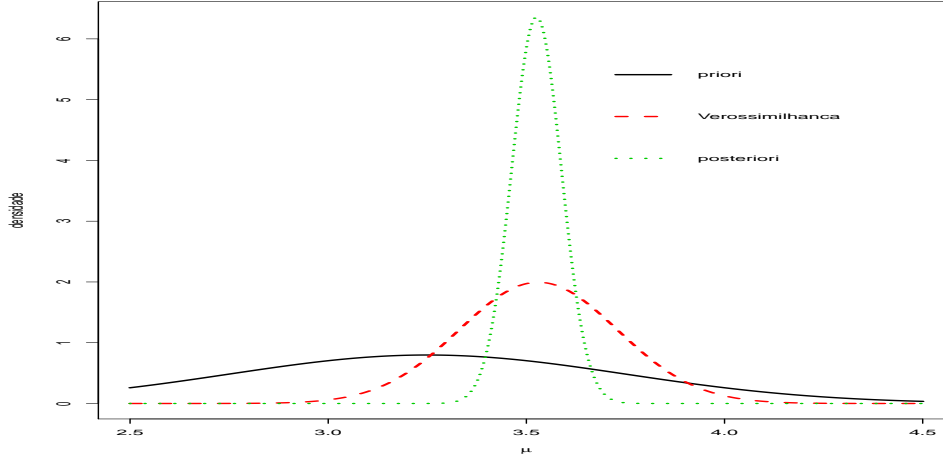


Figura 1: Versossimilhança e densidades a priori e a posteriori para o Exemplo 1.

A Figura 1 mostra a densidade a priori, a verossimilhança e a densidade a posteriori. Note que a distribuição a priori contém muita menos informação que a priori, por isso a moda da densidade a posteriori fica perto da moda da verossimilhança (em outras palavras, $n/\sigma^2 = 250$ é grande cpm respeito a $1/\tau^2 = 4$, veja a equação (2)). Ainda assim, a distribuição a posteriori é mais precisa tanto que a verossimilhança quanto que a priori (esse é uma consequência do modelo Normal-Normal, veja a equação (3); não ocorre necessariamente para outros modelos).

1.1.1 Distribuição a priori não informativa

Na situação descrita nesta Seção tem duas formas de pensar numa distribuição a priori *pouco informativa*. Uma, é fazer a variância τ^2 da distribuição a priori muito grande. Em particular, quando $\tau^2 \rightarrow \infty$, tomando limites nas equações (2) e (3), vemos que a distribuição a posteriori tende a uma Normal com média \bar{x} e variância σ^2/n , que depende somente da verossimilhança (veja que o valor de μ_0 também some uma vez que fazemos $\tau^2 \rightarrow \infty$). Outra forma de proceder é pensar numa “distribuição uniforme” para μ , algo assim como especificar em (1) $p(\mu)$ igual a uma constante $c > 0$ (como a constante c no numerador cancela com a do denominador, usualmente escreve-se $p(\mu) \propto 1$). Porém, a distribuição uniforme não está definida num intervalo infinito como a reta real ou, em outras palavras, não é possível normalizar a densidade para que integre um pois $\int_{-\infty}^{\infty} c d\mu = \infty$ (lembre da discussão sobre $w(\theta)$ e $w^*(\theta) = w(\theta) / \int w(\theta) d\theta$ na Seção ??), de forma que essa possibilidade está, ao menos formalmente, fora do paradigma Bayesiano. Ainda, é possível pôr $p(\mu) \propto 1$ na equação (1) e, formalmente, fazer as mesmas contas que faríamos com uma distribuição a priori “*verdadeira*”. Se o fizermos,

chegaríamos a que

$$p(\mu | x_1, \dots, x_n) \propto p(x_1, \dots, x_n | \theta) \propto p(\bar{x} | \theta) \propto \exp\left\{-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right\},$$

o que implica que $\mu | x_1, \dots, x_n \sim \text{Normal}(\bar{x}, \sigma^2/n)$, que é o mesmo resultado ao qual chegamos fazendo $\tau^2 \rightarrow \infty$.

Na literatura, quando se usa uma “distribuição a priori” que, como no caso anterior, tem integral infinita, fala-se de uma “*distribuição a priori imprópria*”. As aspas foram usadas acima pois, pelo menos na opinião do autor destas notas, distribuições a priori impróprias não são distribuições a priori. De qualquer forma, distribuições impróprias ou com variância infinita podem ser apropriadas em situações nas quais deseja-se usar uma distribuição a priori relativamente “achatada” (*flat*) na região onde a verossimilhança é significativamente diferente de zero. No caso particular desta seção, a verossimilhança $p(x_1, \dots, x_n | \mu) \propto \exp\{-n(\mu - \bar{x})^2\}$ comporta-se semelhante a uma densidade Normal e decresce muito rapidamente a zero quando μ se afasta de \bar{x} . Na maioria das situações práticas, o valor da densidade a priori $p(\mu)$ para valores de μ que distam de \bar{x} (digamos) mais do que 3 ou 4 vezes o desvio padrão σ/\sqrt{n} é irrelevante. Nessas situações uma distribuição a priori imprópria $p(\mu) \propto 1$ pode ser uma aproximação conveniente e mais fácil de especificar que uma distribuição própria.

Seja fazendo $\tau^2 \rightarrow \infty$ ou usando $p(\mu) \propto 1$, a seguinte observação merece ser feita. Uma vez que chegamos no resultado $\mu | x_1, \dots, x_n \sim \text{Normal}(\bar{x}, \sigma^2/n)$, se denotarmos por $z_{\alpha/2}$ o quantil $(1 - \alpha/2)$ da distribuição Normal Padrão, podemos escrever que

$$\mathbb{P}\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \mid x_1, \dots, x_n\right) = 1 - \alpha. \quad (4)$$

Em comparação, na estatística clássica, quando calcula-se o intervalo de confiança usual nesta situação, tem-se que

$$\mathbb{P}_\mu\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \quad (5)$$

Note a diferença na interpretação dessas duas afirmações. Em (4), a variável aleatória é μ , enquanto \bar{x} está fixa no valor que foi efetivamente observado. Em (5), a variável aleatória é \bar{X} , enquanto μ está fixa num valor desconhecido.

1.1.2 Distribuições preditivas

O problema do cálculo de distribuições preditivas no caso desta Seção pode tomar a seguinte forma, semelhante à que foi discutida na Seção anterior. Considere

$X_1, \dots, X_n, X_{n+1}, \dots, X_{n+m} \mid \mu \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ com a distribuição a priori $\mu \sim \text{Normal}(\mu_0, \tau^2)$. Com base na observação de $\bar{x}_{1:n} = n^{-1} \sum_{i=1}^n x_i$ desejamos prever $\bar{X}_{(n+1):(n+m)} = m^{-1} \sum_{i=n+1}^{n+m} X_i$ ou, em outras palavras, procuramos a distribuição condicional de $\bar{X}_{(n+1):(n+m)}$ dado $\bar{x}_{1:n}$. (se for desejado somente a distribuição preditiva de uma única observação futura X_{n+1} , tome $m = 1$ nos resultados a seguir).

Essa distribuição pode ser achada diretamente da distribuição conjunta de $(\theta, \bar{X}_{1:n}, \bar{X}_{(n+1):(n+m)})$ integrando primeiro com respeito a θ e depois usando a definição de densidade condicional. Neste caso, porém, é mais fácil proceder da seguinte forma. Primeiro, veja que a distribuição conjunta do vetor $(X_1, \dots, X_{n+m}, \theta)$ é Normal $(n+m+1)$ -variada. Logo, todas as distribuições marginais e condicionais também serão Normais multivariadas. Em particular, $X_{n+1}, \dots, X_{n+m} \mid \bar{x}_{1:n}$ segue uma distribuição Normal m -variada e, sendo $\bar{X}_{(n+1):(n+m)}$ uma combinação linear de X_{n+1}, \dots, X_{n+m} , concluimos que a distribuição preditiva de $\bar{X}_{(n+1):(n+m)}$ dado $\bar{x}_{1:n}$ deve ser Normal univariada. Assim, é suficiente achar os dois primeiros momentos da distribuição preditiva, o que pode ser feito facilmente com as propriedades (??)–(??). Isto é, lembrando que os dois primeiros momentos da distribuição a posteriori de $\mu \mid \bar{x}_{1:n}$ são dados pelas equações (2)–(3),

$$\begin{aligned} \mathbb{E}(\bar{X}_{(n+1):(n+m)} \mid \bar{x}_{1:n}) &= \mathbb{E}_{\mu \mid \bar{x}_{1:n}} \mathbb{E}(\bar{X}_{(n+1):(n+m)} \mid \bar{x}_{1:n}, \mu) \\ &= \mathbb{E}_{\mu \mid \bar{x}_{1:n}} \mathbb{E}(\bar{X}_{(n+1):(n+m)} \mid \mu) = \mathbb{E}_{\mu \mid \bar{x}_{1:n}}(\mu) = \frac{\frac{1}{\tau^2} \mu_0 + \frac{n}{\sigma^2} \bar{x}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} = \mu^*, \end{aligned}$$

enquanto

$$\begin{aligned} \text{Var}(\bar{X}_{(n+1):(n+m)} \mid \bar{x}_{1:n}) &= \mathbb{E}_{\mu \mid \bar{x}_{1:n}} \text{Var}(\bar{X}_{(n+1):(n+m)} \mid \bar{x}_{1:n}, \mu) + \text{Var}_{\mu \mid \bar{x}_{1:n}} \mathbb{E}(\bar{X}_{(n+1):(n+m)} \mid \bar{x}_{1:n}, \mu) \\ &= \mathbb{E}_{\mu \mid \bar{x}_{1:n}} \text{Var}(\bar{X}_{(n+1):(n+m)} \mid \mu) + \text{Var}_{\mu \mid \bar{x}_{1:n}} \mathbb{E}(\bar{X}_{(n+1):(n+m)} \mid \mu) \\ &= \mathbb{E}_{\mu \mid \bar{x}_{1:n}} \left(\frac{\sigma^2}{m} \right) + \text{Var}_{\mu \mid \bar{x}_{1:n}}(\mu) = \frac{\sigma^2}{m} + \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2}{m} + \tau^{*2}. \end{aligned}$$

Veja que está última variância reflete duas fontes de incerteza com respeito a $\bar{X}_{(n+1):(n+m)}$: o primeiro termo σ^2/m tem a ver com a variabilidade inerente de $\bar{X}_{(n+1):(n+m)}$, enquanto o segundo termo τ^{*2} reflete a incerteza com respeito à média μ .

1.2 Inferência para a média de uma distribuição Normal com variância desconhecida: O modelo Normal-Normal-Gama

Nesta seção seguimos essencialmente a apresentação no texto ?.

Considere como antes $X_1, \dots, X_n \mid \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2)$, mas agora tanto a média μ quanto a variância σ^2 são desconhecidas, i.é. o parâmetro $\theta = (\mu, \sigma^2)$ tem dimensão 2. Usualmente o foco da inferência é o parâmetro μ e σ^2 é chamado um parâmetro *nuissance* ou *de estorvo*.

Como vimos na Seção anterior, a notação fica mais simples quando o problema é parametrizado em termos da precisão $\phi = \sigma^{-2}$, de forma que pensamos numa amostra $X_1, \dots, X_n \mid \mu, \phi \stackrel{iid}{\sim} N(\mu, \phi^{-1})$.

É natural que especificar uma distribuição a priori bivariada $p(\mu, \phi)$ é mais difícil que especificar uma univariada, como foi o caso de $p(\theta)$ na Seção ?? ou de $p(\mu)$ na Seção 1.1. Uma opção conveniente é partir o problema em “pedazos” univariados, especificando sucessivamente $p(\mu \mid \phi)$ e $p(\phi)$ (essa ideia de “*dividir para conquistar*” é muito útil em modelagem e, de fato, vamos usar ela extensivamente para abordar o problema de *modelos hierárquicos* nas próximas unidades).

Começamos assim pela distribuição a priori condicional de $\mu \mid \phi$. Quando condicionamos no valor de ϕ (ou, equivalentemente, no valor de $\sigma^2 = \phi^{-1}$), o problema passa a ser o que estudamos na Seção anterior. Dessa forma, sabemos que a distribuição conjugada $p(\mu, \phi)$ deve satisfazer que a condicional de $\mu \mid \phi$ deve ser Normal e também que a distribuição a posteriori condicional de $\mu \mid x_1, \dots, x_n; \phi$ é Normal com média e variância dadas pelas equações (2) e (3).

Suponha que parametrizamos a distribuição a priori condicional de forma que $\mu \mid \phi \sim \text{Normal}(\mu_0, (\lambda_0 \phi)^{-1})$. Substituindo σ^2 por ϕ^{-1} e τ^2 por $(\lambda_0 \phi)^{-1}$ nas equações (2) e (3), o comentário anterior implica que, a posteriori, $\mu \mid \phi, x_1, \dots, x_n \sim \text{Normal}(\mu^*, (\lambda^* \phi)^{-1})$, onde

$$\mu^* = \frac{\lambda_0 \mu_0 + n \bar{x}}{\lambda_0 + n}, \quad (6)$$

$$\lambda^* = \lambda_0 + n. \quad (7)$$

Dessa forma, resolvemos o problema da distribuição a priori condicional de $\mu \mid \phi$ e também o de calcular a posteriori condicional de $\mu \mid \phi, x_1, \dots, x_n$.

Resta especificar a distribuição a priori marginal de ϕ e calcular a distribuição a posteriori também marginal de $\phi \mid x_1, \dots, x_n$. Por enquanto, considere o caso com uma priori $p(\phi)$ geral.

Lembre primeiro que $[\bar{X} = \sum_{i=1}^n X_i, S^2 = \sum_{i=1}^n (X_i - \bar{X})^2]$ é um estatístico suficiente e que, dado o par (μ, ϕ) , \bar{X} e S^2 são independentes com $\bar{X} \mid \mu, \phi \sim$

$\text{Normal}[\mu, (n\phi)^{-1}]$ e $\phi S^2 \mid \mu, \phi \sim \chi_{n-1}^2 \equiv \text{Gama}(\frac{n-1}{2}, \frac{1}{2})$. Logo, a verossimilhança é

$$\begin{aligned} p(x_1, \dots, x_n \mid \mu, \phi) &\propto p(\bar{x}, S^2 \mid \mu, \phi) \propto p(\bar{x} \mid \mu, \phi) p(S^2 \mid \mu, \phi) \\ &\propto \phi^{1/2} \exp \left\{ -\frac{n\phi}{2} (\bar{x} - \mu)^2 \right\} \times \phi^{(n-1)/2} \exp \left\{ -\frac{\phi}{2} s^2 \right\} \\ &\propto \phi^{n/2} \exp \left\{ -\frac{\phi}{2} [s^2 + (\mu - \bar{x})^2] \right\}. \end{aligned} \quad (8)$$

Por outra parte, como especificamos que $\mu \mid \phi \sim \text{Normal}(\mu_0, (\lambda_0 \phi)^{-1})$, segue que a densidade a priori conjunta é

$$p(\mu, \phi) = p(\phi) p(\mu \mid \phi) \propto p(\phi) \phi^{1/2} \exp \left\{ -\frac{\phi}{2} \lambda_0 (\mu - \mu_0)^2 \right\}. \quad (9)$$

Multiplicando (8) e (9) temos então que

$$\begin{aligned} p(\mu, \phi \mid x_1, \dots, x_n) &\propto p(\mu, \phi) p(x_1, \dots, x_n \mid \mu, \phi) \\ &\propto p(\phi) \phi^{(n+1)/2} \exp \left\{ -\frac{\phi}{2} [s^2 + (\mu - \bar{x})^2 + \lambda_0 (\mu - \mu_0)^2] \right\}. \end{aligned} \quad (10)$$

Finalmente, usando o fato que $\mu \mid \phi, x_1, \dots, x_n \sim \text{Normal}(\mu^*, (\lambda^* \phi)^{-1})$,

$$\begin{aligned} p(\phi \mid x_1, \dots, x_n) &= \frac{p(\mu, \phi \mid x_1, \dots, x_n)}{p(\mu \mid \phi, x_1, \dots, x_n)} \\ &\propto p(\phi) \frac{\phi^{(n+1)/2} \exp \left\{ -\frac{\phi}{2} [s^2 + (\mu - \bar{x})^2 + \lambda_0 (\mu - \mu_0)^2] \right\}}{\phi^{1/2} \exp \left\{ -\frac{\phi}{2} \lambda^* (\mu - \mu^*)^2 \right\}} \\ &\propto p(\phi) \phi^{n/2} \exp \left\{ -\frac{\phi}{2} [s^2 + (\mu - \bar{x})^2 + \lambda_0 (\mu - \mu_0)^2 - \lambda^* (\mu - \mu^*)^2] \right\} \\ &\propto p(\phi) \left[\phi^{n/2} \exp \left\{ -\frac{\phi}{2} \left[s^2 + \frac{n\lambda_0}{\lambda_0 + n} (\bar{x} - \mu_0)^2 \right] \right\} \right] \end{aligned} \quad (11)$$

(para o cálculo acima, da equação é útil observar que o termo da esquerda depende somente de ϕ , enquanto o da direita depende também de μ . A equação 11 é portanto uma identidade que vale para todo μ . Em termos práticos, na direita é possível substituir μ por qualquer valor e o resultado final deveria ser o mesmo. Fazer a substituição $\mu = \bar{x}$ ou $\mu = \mu_0$ facilita o cálculo).

Veja que, quando considerado como função somente de ϕ , o termo entre colchetes no último termo à direita da equação 11 é o núcleo de uma distribuição Gama, o que sugere então especificar $p(\phi)$ também como uma densidade Gama.

Proposição 1. *Suponha que $X_1, \dots, X_n \mid \mu, \phi \stackrel{iid}{\sim} \text{Normal}(\mu, \phi^{-1})$ e, a priori, $\mu \mid \phi \sim \text{Normal}(\mu_0, (\lambda_0 \phi)^{-1})$ e $\phi \sim \text{Gama}(\alpha_0, \beta_0)$. Então, a posteriori, $\mu \mid \phi, x_1, \dots, x_n \sim$*

$\text{Normal}(\mu^*, (\lambda^* \phi)^{-1})$ e $\phi | x_1, \dots, x_n \sim \text{Gama}(\alpha^*, \beta^*)$, onde μ^* e λ^* são dados nas equações (6) e (7) e

$$\alpha^* = \alpha_0 + \frac{n}{2}, \quad (12)$$

$$\beta^* = \beta_0 + \frac{1}{2} s^2 + \frac{n \lambda_0}{2(\lambda_0 + n)} (\bar{x} - \mu_0)^2. \quad (13)$$

Demonstração. Como foi explicado acima das equações (6) e (7), a parte referente à $p(\mu | \phi, x_1, \dots, x_n)$ é consequência da Seção 1.1. Para achar a distribuição a posteriori marginal de ϕ , substitua $p(\phi) \propto \phi^{\alpha-1} e^{-\beta_0 \phi}$ na equação (11) para obter que

$$p(\phi | x_1, \dots, x_n) \propto \phi^{\alpha_0+n/2-1} \exp \left\{ -\phi \left[\beta_0 + \frac{1}{2} s^2 + \frac{n \lambda_0 (\bar{x} - \mu_0)^2}{2(\lambda_0 + n)} \right] \right\}, \quad (14)$$

o que implica que $\phi | x_1, \dots, x_n \sim \text{Gamma}(\alpha^*, \beta^*)$ \square

1.2.1 A distribuição a posteriori marginal de μ

Como mencionamos no início desta Seção, usualmente o interesse é no parâmetro μ . Nesse caso, precisamos da distribuição a posteriori marginal $p(\mu | x_1, \dots, x_n)$.

Segue da proposição 1 que, dado x_1, \dots, x_n , $Z = (\lambda^* \phi)^{1/2} (\mu - \mu^*)$ segue uma distribuição Normal Padrão. Considere então a distribuição a posteriori conjunta de (Z, ϕ) . O Jacobiano da transformação de $(Z, \phi) \mapsto (\mu = \mu^* + (\lambda^* \phi)^{-1/2} Z; \phi)$ é $(\lambda^* \phi)^{-1/2} \propto \phi^{-1/2}$. Logo, a densidade a posteriori de (Z, ϕ) é

$$\begin{aligned} p(Z, \phi | x_1, \dots, x_n) &\propto p(\mu, \phi | x_1, \dots, x_n) \left| \frac{\partial(\mu, \phi)}{\partial(Z, \phi)} \right| \\ &\propto p(\mu | \phi, x_1, \dots, x_n) p(\phi, x_1, \dots, x_n) \phi^{-1/2} \\ &\propto \phi^{1/2} \exp \left\{ -\frac{\lambda^* \phi}{2} (\mu - \mu^*)^2 \right\} \times \phi^{\alpha^*-1} \exp\{-\beta^* \phi\} \times \phi^{-1/2} \\ &\propto \exp \left\{ -\frac{1}{2} z^2 \right\} \times \phi^{\alpha^*-1} \exp\{-\beta^* \phi\}. \end{aligned}$$

Logo, a posteriori, Z e ϕ são independentes com $Z | x_1, \dots, x_n \sim N(0,1)$ e, como já sabíamos, $\phi | x_1, \dots, x_n \sim \text{Gama}(\alpha^*, \beta^*)$. Das propriedades da distribuição Gama sabemos que isso implica que $2\beta^* \phi | x_1, \dots, x_n \sim \text{Gama}(\alpha^*, 1/2)$, também denominada de distribuição χ^2 com $(2\alpha^*)$ graus de liberdade. Da definição da distribuição t de Student como a razão entre duas variáveis aleatórias independentes, uma com distribuição Normal(0,1) e a outra sendo a raiz quadrada de uma χ_m^2 dividida pelos

graus de liberdade m , segue que

$$\sqrt{\frac{\lambda^* \alpha^*}{\beta^*}} (\mu - \mu^*) = \frac{Z}{\sqrt{2 \beta^* \phi / (2 \alpha^*)}} \mid x_1, \dots, x_n \sim t_{2 \alpha^*}. \quad (15)$$

Note que, como a distribuição a priori e a posteriori pertencem a mesma família Normal-Gama, um resultado semelhante deve valer também para a distribuição a priori marginal de μ . É conveniente escrever o resultado para uso futuro (veja, por exemplo, a subseção 1.2.3).

Lema 2. *Suponha que $\mu \mid \phi \sim \text{Normal}(\mu_0, (\lambda_0 \phi)^{-1})$ e $\phi \sim \text{Gama}(\alpha_0, \beta_0)$, onde α, β, λ_0 e ϕ são parâmetros positivos. Então*

$$\sqrt{\frac{\lambda_0 \alpha}{\beta}} (\mu - \mu_0) \sim t_{2 \alpha}. \quad (16)$$

1.2.2 Distribuição a priori não informativa

O intervalo de intervalo de confiança clássico para μ é

$$\bar{x} \pm t_{n-1; a/2} \frac{s}{\sqrt{n(n-1)}}, \quad (17)$$

onde $t_{n-1; a/2}$ é o percentil $100(1-a)\%$ da distribuição t de Student com $(n-1)$ graus de liberdade [note que, por consistência da notação, continuamos definindo $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2$, sem dividir por $(n-1)$].

Com base na distribuição (15), o equivalente bayesiano desse intervalo de confiança seria

$$\mu^* \pm t_{2 \alpha^*; a/2} \sqrt{\frac{\beta^*}{\lambda^* \alpha^*}}, \quad (18)$$

no sentido que segue de (15) que

$$\mathbb{P} \left(\mu^* - t_{2 \alpha^*; a/2} \sqrt{\frac{\beta^*}{\lambda^* \alpha^*}} < \mu < \mu^* + t_{2 \alpha^*; a/2} \sqrt{\frac{\beta^*}{\lambda^* \alpha^*}} \mid x_1, \dots, x_n \right) = 1 - a.$$

Cabe então perguntar se existem valores dos hiperparâmetros $(\mu_0, \lambda_0, \alpha_0, \beta_0)$ para os quais os dois intervalos (17) e (18) coincidem. Dos graus de liberdade obtemos que $2 \alpha^* = 2 \alpha_0 + n$ teria que ser igual a $(n-1)$, isto é $\alpha_0 = -(1/2)$. Por outro lado, como os dois intervalos teriam que estar centrados em \bar{x} , deveríamos ter que $\mu^* = \bar{x}$, e da equação (2) obtemos então $\lambda_0 = 0$ (observe que o valor de μ_0 resulta irrelevante aqui após fazermos $\lambda_0 = 0$). Finalmente, precisaríamos ter que

$$\frac{s}{\sqrt{n(n-1)}} = \sqrt{\frac{\beta^*}{\lambda^* \alpha^*}} = \sqrt{\frac{\beta_0 + \frac{1}{2} s^2 + \frac{n \lambda_0}{2(\lambda_0 + n)} (\bar{x} - \mu_0)^2}{(\lambda_0 + n)(\alpha_0 + n/2)}},$$

que uma vez que substituímos $\alpha_0 = -(1/2)$ e $\lambda_0 = 0$ dá $\beta_0 = 0$.

Resumindo, para os intervalos (17) e (18) coincidir, precisamos tomar na distribuição a priori $\lambda_0 = 0$, $\alpha_0 = -(1/2)$ e $\beta_0 = 0$ (e qualquer μ_0). Esse tipo de distribuições a priori que fazem os intervalos clássicos coincidir com os Bayesianos são chamadas na literatura de distribuições a priori *matching*. Elas estão associadas usualmente com distribuições a priori não informativas, o que é o caso desta. Um problema, porém, é que nem a distribuição Normal está bem definida para variância infinita (i.é. precisão $\lambda_0 \phi = 0$), nem a distribuição Gama está definida quando $\alpha_0 = -(1/2) \leq 0$ e/ou $\beta_0 = 0 \leq 0$. Nesse sentido, veja que da equação (9) temos que

$$\begin{aligned} p(\mu, \phi) &\propto \phi^{\alpha_0-1} \exp\{-\beta_0 \phi\} \times \phi^{1/2} \exp\left\{-\frac{\phi}{2} \lambda_0 (\mu - \mu_0)^2\right\} \\ &\propto \phi^{\alpha_0-1/2} \exp\left\{-\frac{\phi}{2} \lambda_0 (\mu - \mu_0)^2 - \beta_0 \phi\right\}, \end{aligned}$$

de forma que substituindo (formalmente) pelos valores achados acima teríamos que

$$p(\mu, \phi) \propto \frac{1}{\phi},$$

que é uma distribuição a priori imprópria, no sentido que $\int_0^\infty \int_{-\infty}^\infty \phi^{-1} d\mu d\phi = \infty$. Independentemente disso, ela é considerada a distribuição a priori não informativa para o problema desta Seção. Veja que pode ser pensada como $p(\mu, \phi) \propto p(\mu) \times p(\phi) \propto 1 \times \phi^{-1}$, isto é, a priori μ e ϕ seriam independentes com $p(\mu) \propto 1$, que já vimos na Seção 1.1, e $p(\phi) \propto \phi^{-1}$.

Exemplo 2. (Os experimentos de Newcomb-Michelson sobre a velocidade da luz, adaptado de Gelman, Carlin, Stern e Rubin). Por volta de 1880 os físicos Simon Newcomb e Albert Michelson idealizaram e posteriormente realizaram o que é considerado o primeiro experimento que mediu com bastante precisão a velocidade da luz. Sucintamente, eles emitiram luz desde uma fonte situada na margem de Virgínia do Rio Potomac até um espelho situado a 3721.21 metros ao pé do Monumento a Washington em Washington, D.C. Dessa forma, o experimento tentava medir o tempo que a luz demoraria em viajar $7442.42 = 2(3721.21)\text{m}$ (Veja a Figura 2) Após emitida, a luz era refletida em um espelho menor que estava girando muito rapidamente, de forma que ao retornar, era deslocada por duas vezes o ângulo que o espelho menor rotou durante o tempo que a luz demorou para ir e voltar até o espelho maior. Dessa forma, conhecendo (i) a velocidade de rotação do espelho menor, (ii) o deslocamento da imagem e (iii) a distância entre o espelho menor e o maior, era possível determinar o tempo que a luz demorou para fazer o percurso. As 66 medições realizadas por Newcomb em 1882 são mostradas no boxplot da Figura 3. Os dados coletados por Newcomb foram registrados como desvios de 24800

nanossegundos ($= 24800 \times 10^{-9}$ segundos). No **R**, esses dados estão disponíveis sob o nome **light** no pacote **BayesDA** (veja o script que acompanha estas notas). O boxplot mostra duas medições discordantes no sentido de serem muito pequenas. Considerando essas medições, não parece razoável assumir que os dados seguem uma distribuição Normal e portanto para o resto do exemplo foram removidos da análise, ficando assim com uma amostra de tamanho $n = 64$ (seria certamente mais correto usar um modelo capaz de explicar as duas observações discordantes, usando por exemplo distribuições com caudas mais pesadas que a Normal, tais como a distribuição de Cauchy ou t de Student, mas fazer isso foge da matéria desta unidade). A média e o desvio padrão das 64 observações consideradas foram $\bar{x} = 27.75$ e $\sqrt{s^2/(n-1)} = 5.08$. Usando uma distribuição a priori não informativa e $100(1-a) = 95\%$, o intervalo (18) ou (17) para μ é

$$27.75 \pm t_{63,0.025} \frac{5.08}{64^{1/2}} \doteq (26.48; 29.02).$$

Com base nesse tempo, considerando que μ é a média do tempo que a luz demora para viajar 7442.42m, o tempo médio que a luz viajaria em 1s seria $10^9(7442.42)/(\mu + 24800)$, de forma que o correspondente intervalo para a velocidade da luz seria

$$\left(\frac{10^9 (7442.42)}{26.48 + 24800}; \frac{10^9 (7442.42)}{29.02 + 24800} \right) \doteq (299,729,915; 299,760,576)m/s.$$

[O valor aceito atualmente para a velocidade da luz no vacuo é 299,792,458m/s, fora do intervalo que calculamos acima; a análise de Newcomb dos dados foi diferente da mostrada aqui, ele chegou a um valor de 299,860km/s. Simon Newcomb (1835–1909) foi um físico canadense que viveu e trabalhou a maior parte da sua vida nos Estados Unidos. Albert Michaelson (1852–1931) era um físico americano que colaborou com Newcomb na construção do instrumento experimental. Recebeu o Prêmio Nobel em Física em 1907.]

□

1.2.3 Distribuição preditiva para uma observação futura

O problema do cálculo de distribuições preditivas no caso desta Seção pode tomar a seguinte forma, semelhante à que foi discutida na Seção 1.1. Considere $X_1, \dots, X_n, X_{n+1}, \dots, X_{n+m} | \mu, \phi \stackrel{iid}{\sim} \text{Normal}(\mu, (\lambda_0 \phi)^{-1})$ com a distribuição a priori $\mu | \phi \sim \text{Normal}(\mu_0, (\lambda_0 \phi)^{-1})$ e $\phi \sim \text{Gama}(\alpha_0, \beta_0)$. Com base na observação de x_1, \dots, x_n desejamos prever $\bar{X}_{(n+1):(n+m)} = m^{-1} \sum_{i=n+1}^{m+n} X_i$ ou, em outras palavras, procuramos a distribuição condicional de $\bar{X}_{(n+1):(n+m)}$ dado x_1, \dots, x_n .

Da subseção 1.1.2 sabemos que $\bar{X}_{(n+1):(n+m)} | x_1, \dots, x_n; \phi \sim \text{Normal}[\mu^*, (\lambda^* \phi)^{-1}]$. Por outro lado, da subseção ??, sabemos também que $\phi | x_1, \dots, x_n \sim \text{Gama}(\alpha^*, \beta^*)$.

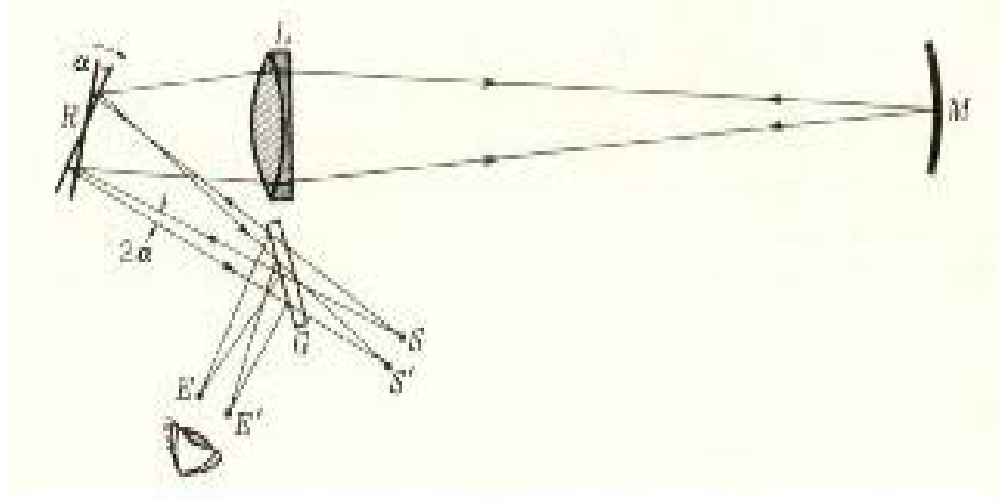


Figura 2: Esboço do experimento de Newcomb-Michelson. S é uma fonte de luz, G é divisor de feixe, R é um espelho rotando muito rapidamente, L é uma lente objetiva, M é um espelho esférico fixo e E o ocular do observador. A imagem que retorna do espelho fixo M é deslocada duas vezes o ângulo que R rotou durante o tempo que a luz demorou para viajar de R até M . Portanto, sabendo a velocidade de rotação de R , o desvio angular da luz que retorna de M e a distância RM , é possível calcular a velocidade da luz.

Portanto, se olharmos o par $(\bar{X}_{(n+1):(n+m)}, \phi)$ condicionado a x_1, \dots, x_n , temos uma distribuição Normal-Gama e o Lema 2 implica que

$$\sqrt{\frac{(\frac{1}{m} + \frac{1}{\lambda^*})}{\beta^*}} \alpha^* (\bar{X}_{(n+1):(n+m)} - \mu^*) \mid x_1, \dots, x_n \sim t_{2\alpha^*}.$$

Exemplo 2 (Continuação). Como a distribuição a posteriori de μ [equação (15)] é simétrica com respeito a μ^* , segue que a estimativa bayesiana com respeito a PQ, PA ou perda zero-um é $\mu^* = 27.75$.

2 Prioris impróprias e não informativas

2.1 Prioris impróprias

No paradigma bayesiano temos que

$$p(\theta \mid x) = \frac{p(\theta) p(x \mid \theta)}{\int p(\theta) p(x \mid \theta) d\theta} \propto p(\theta) p(x \mid \theta). \quad (19)$$

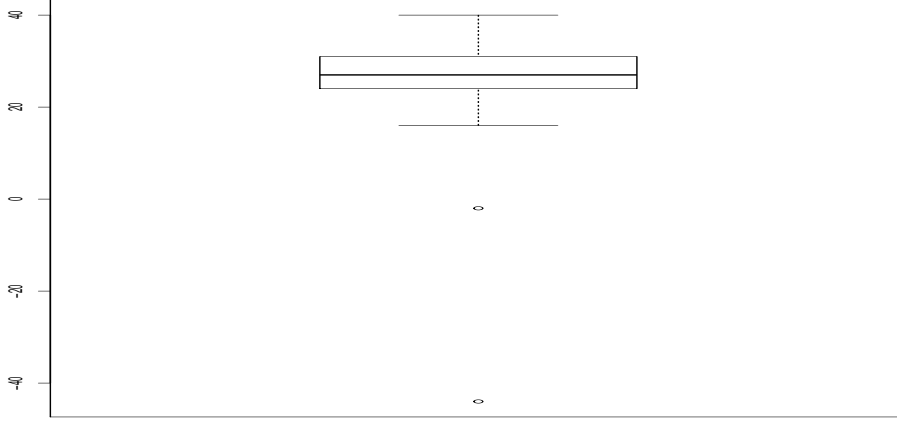


Figura 3: Boxplot das 66 medições realizadas por Newcomb em 1882. Os dados foram registrados como desvios de 24,800 nanosegundos. Observe as duas medições discordantes (*outliers*) com valores $x_6 = -44$ e $x_{10} = -2$. Elas foram descartadas para a análise mostrada no Exemplo 2.

Quando tanto $p(\theta)$ quanto $p(x|\theta)$ são densidades, resultados do cálculo de probabilidades mostram que a distribuição a posteriori está bem definida, no sentido que a integral no denominador acima é finita. Veja que nada muda se usarmos acima uma “priori” cuja integral é finita mas não necessariamente igual a um, pois se $\int p(\theta) d\theta = I \neq 1$, podemos definir a priori $p^*(\theta) = p(\theta)/I$ e a constante I cancelará no numerador e no denominador (19).

A situação é diferente se, formalmente, usarmos em (19) uma “densidade” (dita *imprópria* na literatura) cuja integral é infinita, pois nesse caso não é garantido que $\int p(\theta) p(x|\theta) d\theta < \infty$. Esse tipo de “prioris” podem ser pensadas como aproximações para prioris próprias, mas o argumento anterior implica que a propriedade da distribuição a posteriori deve ser analisada caso a caso.

Exemplo 3. Vimos na seção 1.1 o uso da priori “não informativa” $p(\mu) \propto 1$, que é imprópria. A partir da equação (19) temos que, nesse caso, $p(\mu|x_1, \dots, x_n) \propto \exp\{\frac{-n}{2\sigma^2}(\bar{x} - \mu)^2\}$. Como $\int_{-\infty}^{\infty} \exp\{\frac{-n}{2\sigma^2}(\bar{x} - \mu)^2\} d\mu < \infty$, a posteriori resultante será própria qualquer seja a amostra (x_1, \dots, x_n) observada. Pode-se verificar que o mesmo vale na seção 1.2 para a priori imprópria $p(\mu, \phi) \propto \phi^{-1}$.

Exemplo 4. Suponha que o modelo especifica $x_1, \dots, x_n|\theta \sim \text{Poisson}(\theta)$ e queremos usar uma distribuição a priori com média θ_0 e variância tão grande quanto possível. Usando a priori conjugada $\text{Gama}(\alpha, \beta)$ para esse problema, devemos fazer $\alpha = \beta\theta_0$ para a restrição da média, em cujo caso a variância será θ_0/β , que é maximizada quando $\beta \downarrow 0$. Na prática, isso significa que a priori deveria ser

uma “Gama(0,0)” com densidade proporcional a θ^{-1} . Substituindo em (19) obtemos $p(\theta | x_1, \dots, x_n) \propto e^{-n\theta} \theta^{s-1}$. Logo, quando $s = \sum_{i=1}^n x_i \geq 1$, a posteriori será uma Gama(s, n), que é própria. Porém, se observarmos $s = 0$, temos que $\int_0^\infty \theta^{-1} e^{-n\theta} d\theta = \infty$ e a posteriori não vai estar bem definida.

2.2 Os problemas de locação e de escala

Um modelo de locação específica a observação de uma quantidade X tal que $p(x | \theta) = f(x - \theta)$, onde f é uma densidade (própria!) definida em toda a reta real. Esse modelo tem uma simetria fundamental no sentido que é invariante a translações. Mais precisamente, se a é um real fixo e definirmos $Y = X - a$ e $\lambda = \theta - a$, segue do fato que $p_Y(y | \theta) = p_X(x = y + a | \theta) |dx/dy| = f(y + a - \theta) = f(y - \lambda)$ que o modelo $Y | \lambda$ também é um modelo de locação. Portanto, se existe nesse caso uma “distribuição a priori não informativa”, ela deveria tomar a mesma forma para os modelos $X | \theta$ e $Y | \lambda$. Em outras palavras, deveríamos ter que $p_\lambda(\lambda) = p_\theta(\lambda)$. Por outro lado, sabemos do cálculo de probabilidades e o fato que $\lambda = \theta - a$ que $p_\lambda(\lambda) = p_\theta(\lambda + a) |d\theta/d\lambda| = p_\theta(\lambda + a)$. Juntando os dois resultados, segue que $p_\theta(\lambda) = p_\theta(\lambda + a)$ deveria valer para todo λ e todo a , o que somente é possível se a densidade p_θ for constante (usualmente denotamos $p(\theta) \propto 1$).

Um modelo de escala específica que $p(x | \sigma) = \sigma^{-1} f(x/\sigma)$, onde o parâmetro $\sigma > 0$ e a densidade f é considerada fixa. Esse modelo também tem uma simetria fundamental, no sentido que, se definirmos $Y = X/a$ e $\phi = \sigma/a$ ($a > 0$), segue que $p_Y(y | \sigma) = p_X(ay | \sigma) |dx/dy| = \phi^{-1} f(y/\phi)$. Dessa forma, tanto o modelo $X | \sigma$ quanto o $Y | \phi$ são modelos de escala. Se repetimos o argumento que fizemos acima para o modelo de locação, chegaremos ao final que uma distribuição a priori não informativa para o problema de escala deveria satisfazer a condição $p(\sigma) = a^{-1} p(\sigma/a)$ para todo a e σ positivos. É fácil ver que essa condição somente é possível se tivermos que $p(\sigma) \propto \sigma^{-1}$.

(Uma outra forma de obter esse resultado é a seguinte: No modelo de escala com $X > 0$ defina $X^* = \log X$ e $\mu = \log \sigma$ e verifique que o modelo resultante é de locação, de forma que pelo argumento anterior deveríamos ter que $p(\mu) \propto 1$, o que é equivalente a assumir $p(\sigma) \propto \sigma^{-1}$.)

2.3 Distribuições a priori de Jeffreys

Vimos na unidade anterior que um problema fundamental com o Princípio da Probabilidade Inversa de Laplace era o fato de ele não ser invariante a reparametrizações. A partir de essa idéia o físico britânico sugeriu usar como distribuição a priori não informativa uma densidade baseada informação de Fisher. No caso

uniparamétrico, define-se

$$p_J(\theta) \propto I(\theta)^{1/2} = \sqrt{\mathbb{E}_{X|\theta} \left[\frac{\partial}{\partial \theta} \log p(X|\theta) \right]^2} = \sqrt{\mathbb{E}_{X|\theta} \left[-\frac{\partial^2}{\partial \theta^2} \log p(X|\theta) \right]} \quad (20)$$

Mais geralmente, quando o espaço paramétrico tem dimensão r maior do que um, a informação de Fisher é uma matriz quadrada com elementos

$$I_{i,j}(\theta) = \mathbb{E}_{X|\theta} \left(\frac{\partial}{\partial \theta_i} \log p(X|\theta) \cdot \frac{\partial}{\partial \theta_j} \log p(X|\theta) \right) = -\mathbb{E}_{X|\theta} \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X|\theta) \right).$$

Nesse caso a densidade a priori de Jeffreys é

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

- Se $x_1, \dots, x_n | \theta \stackrel{iid}{\sim} p(x_1 | \theta)$ e $I_n(\theta)$ é a informação de Fisher associada a essa amostra aleatória de tamanho n , como $I_n(\theta) = n I_1(\theta)$, não faz diferença se calcularmos a priori de Jeffreys para toda a amostra ou para uma única observação (i.é. $p_J(\theta) \propto I_n(\theta) \propto I_1(\theta)$).
- Se considerarmos uma reparametrização $\eta = \eta(\theta)$, como $I_\eta(\eta) = I_\theta[\theta(\eta)] (d\theta/d\eta)^2$, segue que $p_{J,\eta}(\eta) = p_{J,\theta}[\theta(\eta)] |d\theta/d\eta|$, que é a regra do Jacobiano usual e portanto a priori de Jeffreys é invariante sob reparametrizações.
- Alguns autores afirmam que o uso da priori de Jeffreys pode levar a situações anômalas no caso multiparamétrico e sugerem nesse caso usar prioris de Jeffreys calculadas para cada parâmetro por separado e assumir independência. Por exemplo, se $\theta = (\theta_1, \theta_2)$, a sugestão é calcular $p_{J,\theta_i}(\theta_i)$ assumindo o outro parâmetro fixo e depois usar a priori $p_{J,\theta_1}(\theta_1) p_{J,\theta_2}(\theta_2)$.
- No caso dos modelos de locação e de escala, a priori de Jeffreys coincide com as achadas acima (veja o exemplo 5 abaixo).
- Como o valor esperado na equação 20 é calculado com respeito à distribuição de $x | \theta$, o uso de prioris de Jeffreys viola o Princípio de Verossimilhança.

Exemplo 5. Suponha o modelo de locação $x | \theta \sim f(x - \theta)$. A Informação de Fisher é

$$\begin{aligned} I_1(\theta) &= \int_{-\infty}^{\infty} \left[\frac{d}{d\theta} \log f(x - \theta) \right]^2 f(x - \theta) dx \\ &= \int_{-\infty}^{\infty} \left[\frac{f'(x - \theta)}{f(x - \theta)} \right]^2 f(x - \theta) dx = \int_{-\infty}^{\infty} \left[\frac{f'(u)}{f(u)} \right]^2 f(u) du \propto 1. \end{aligned}$$

Portanto, nesse caso $p_J(\theta) \propto 1$. Um argumento semelhante no modelo de escala mostra que, nesse caso, $p_J(\sigma) \propto \sigma^{-1}$ (faça como exercício).

Exemplo 6. Suponha que $x_1, \dots, x_n \mid \theta \sim \text{Binomial}(n, \theta)$. Como nesse caso temos que $I_1(\theta) = [\theta(1 - \theta)]^{-1}$, segue que a priori de Jeffreys é $p_J(\theta) \propto [\theta(1 - \theta)]^{-1/2}$, isto é, a distribuição Beta($\alpha = 1/2, \beta = 1/2$).

Exemplo 7. Suponha que $x_1, \dots, x_n \mid \mu, \phi \sim \text{Normal}(\mu, \phi^{-1})$. Como nesse caso temos que $I_1(\mu, \phi) =$, segue que a priori de Jeffreys é $p_J(\mu, \phi) \propto$.