5 A abordagem bayesiana à inferência estatística

5.1 O teorema de Bayes

A abordagem bayesiana à inferência estatística é baseada numa interpretação particular do conhecido teorema de Bayes:

Teorema 5.1 *Seja* $\{A_i, i = 1, \ldots, n\}$ *uma partição do espaço amostral* Ω *tal que* $P(A_i) > 0$, $i = 1, \ldots, n$. *Seja* B *um acontecimento tal que* $P(B) > 0$. *Então, para* $i = 1, \ldots, n$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)\ P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)\ P(A_j)}$$

A utilização de este teorema num contexto dedutivo — o da teoria da probabilidade — não é controverso. Neste caso, $P(B \mid A_i)$ e $P(A_i)$ são conhecidos e pretende-se simplesmente calcular $P(A_i \mid B)$

A possível controvérsia surge num contexto indutivo, o da Estatística, em que o teorema de Bayes é utilizado como um instrumento inferencial:

- A_i denota uma hipótese ou um modelo que nós utilizamos para explicar um fenómeno; uma teoria à qual o investigador atribui *a priori* um grau de credibilidade dado por $P(A_i)$ — informação *a priori*
- B representa o resultado de observar tal fenómeno — os dados amostrais
- $P(B \mid A_i)$ representa a verosimilhança dos dados amostrais quando a explicação A_i é assumida como a correcta — informação amostral que os dados encerram sobre cada hipótese

Neste contexto,

- As probabilidades *a priori* de cada possível explicação, $P(A_i)$, são convertidas em probabilidades *a posteriori*: $P(A_i \mid B)$
- o teorema de Bayes é a forma natural de actualização, através do uso da informação amostral, da informação *a priori* existente antes da observação dos dados
- Esta utilização do teorema de Bayes levanta problemas quanto à interpretação do conceito de probabilidade envolvido nas probabilidades *a priori*, e consequentemente no das probabilidades *a posteriori*
- A interpretação frequentista não é suficientemente abrangente; temos que recorrer à interpretação subjectivista como medida de credibilidade

É preciso estender a noção de modelo estatístico que introduzimos no âmbito da inferência clássica para podermos descrever a metodologia bayesiana.

Em Estatística paramétrica, $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ é uma colecção de distribuições de probabilidade para os dados observáveis; no entanto,

- em Estatística frequencista, θ é desconhecido mas *fixo*
- em Estatística bayesiana, todas as quantidades desconhecidas são tratadas como aleatórias porque tudo o que é desconhecido é incerto, e toda a incerteza deve ser quantificada usando a linguagem da probabilidade

  $\longrightarrow$ distribuição de probabilidade no espaço paramétrico Θ, denotada por $\pi(\theta)$ e designada por distribuição *a priori*

$$\pi(\theta) - \text{distribuição } a\ priori$$

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) - \text{função de verosimilhança}$$

$$\Downarrow$$

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\ \pi(\theta)}{\int_\Theta f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta}, \quad \theta \in \Theta - \text{distribuição } a\ posteriori$$

Observações:

- $f(\boldsymbol{x} \mid \theta)\ \pi(\theta) = \pi(\boldsymbol{x}, \theta)$ define uma distribuição conjunta no espaço $(\mathcal{X}, \Theta)$
- $m(\boldsymbol{x}) = \int_\Theta f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta$ é a denominada distribuição preditiva *a priori* dos dados $\boldsymbol{X}$
- Uma outra forma de escrever o teorema de Bayes é

  $$\pi(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta)\ \pi(\theta)$$

  onde a constante de normalização $m(\boldsymbol{x})$ é omitida

Exemplo 5.1 *Suponhamos que* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$ *e que a priori* $\theta \sim Be(a, b)$, $a, b > 0$ *conhecidos.*

Distribuição beta: se $Y \sim \mathrm{Be}(a, b)$, então

$$f(y) = \frac{1}{B(a,b)}\ y^{a-1}\ (1-y)^{b-1}\ , \quad 0 < y < 1$$

onde $B(a, b) = [\Gamma(a)\ \Gamma(b)]/\Gamma(a + b)$ designa a função beta.

Então, com $t = \sum_{i=1}^{n} x_i$,

$$f(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t}$$

e

$$\pi(\theta) = \frac{1}{B(a,b)}\ \theta^{a-1}\ (1-\theta)^{b-1}, \quad 0 < \theta < 1\ .$$

É fácil de ver que

$$m(\boldsymbol{x}) = \frac{B(t+a, n-t+b)}{B(a,b)}$$

Então,

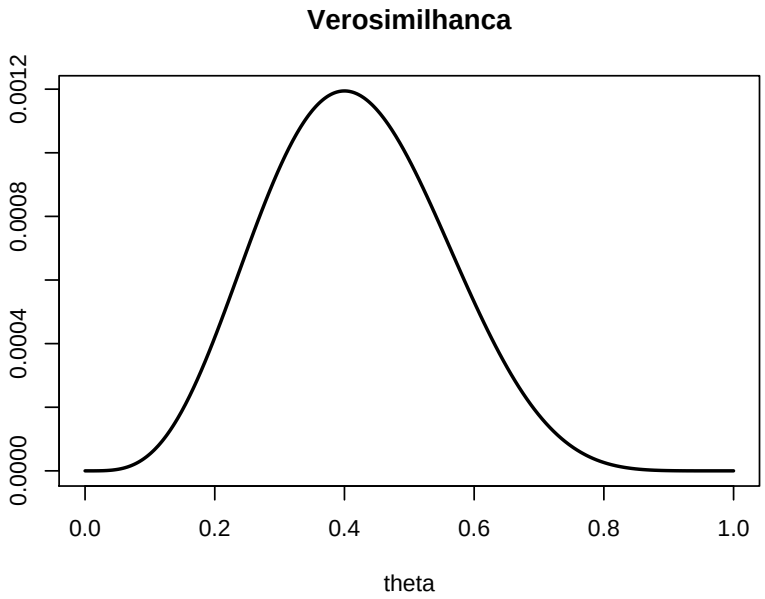
$$\pi(\theta \mid \boldsymbol{x}) = \frac{1}{B(t+a, n-t+b)}\ \theta^{t+a-1}\ (1-\theta)^{n-t+b-1}\ , \quad 0 < \theta < 1$$

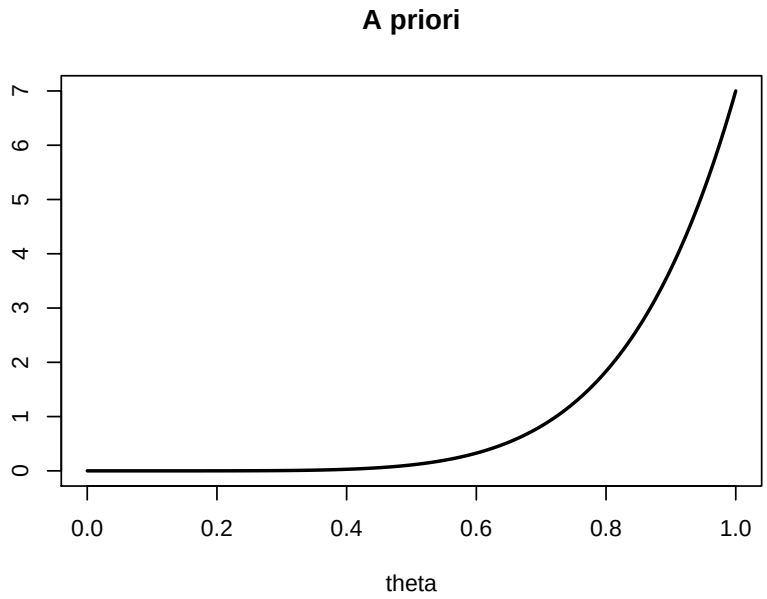
ou seja,

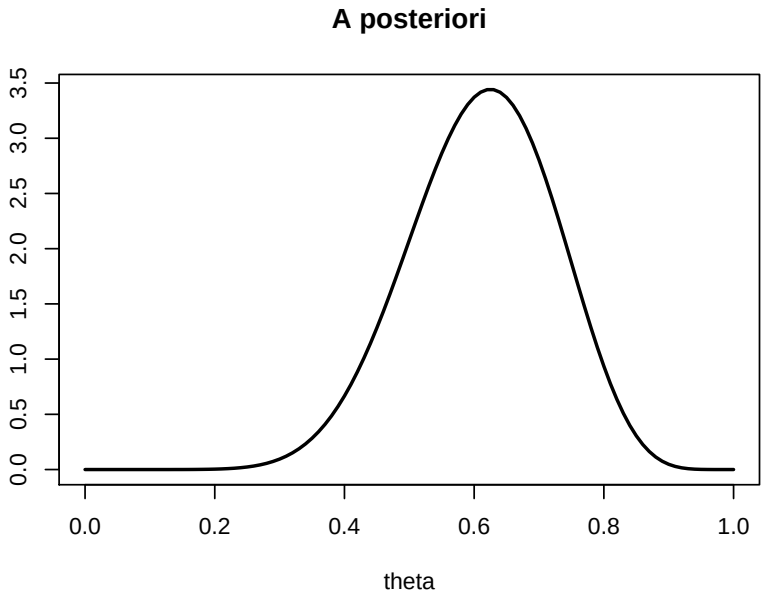
$$\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t+a, n-t+b)$$

Exemplo: $n = 10$, $t = 4$, $a = 7$, $b = 1$,

$$\theta \sim \mathrm{Be}(7, 1)$$
$$\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(11, 7)$$


Verosimilhanca
0.0000
0.0004
0.0008
0.0012
0.0
0.2
0.4
0.6
0.8
1.0
theta


A priori
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8
1.0
theta


A posteriori
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
theta

Observações:

1. Se duas funções de verosimilhança são proporcionais (como funções de θ) elas dão origem à mesma distribuição *a posteriori*
   (a) A inferência bayesiana só depende dos dados observados através do valor observado de uma estatística suficiente
   (b) $\pi(\theta \mid \boldsymbol{x}) = \pi(\theta \mid \boldsymbol{T}(\boldsymbol{x}))$ se $\boldsymbol{T}$ for suficiente para θ
2. $\pi(\theta \mid \boldsymbol{x})$, $\theta \in \Theta$, contém toda a informação sobre θ disponível, combinando a informação amostral (via $L(\theta \mid \boldsymbol{x})$) com a informação apriorística (via $\pi(\theta)$)

Observações (cont.):

3. A operação bayesiana de actualização de conhecimento tem uma natureza sequencial: Suponha-se que $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ with $\boldsymbol{X}_1 \amalg \boldsymbol{X}_2 \mid \theta$. Então,

$$\begin{aligned}\pi(\theta \mid \boldsymbol{x}) &= \frac{f(\boldsymbol{x} \mid \theta)\ \pi(\theta)}{\int f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta} \\ &= \frac{f(\boldsymbol{x}_2 \mid \theta)\ \pi(\theta \mid \boldsymbol{x}_1)}{\int f(\boldsymbol{x}_2 \mid \theta)\ \pi(\theta \mid \boldsymbol{x}_1)\ d\theta}\end{aligned}$$

   Ou seja: $\pi(\theta \mid \boldsymbol{x})$ pode ser também visto como o resultado de combinar a distribuição "*a priori*" $\pi(\theta \mid \boldsymbol{x}_1)$ com a verosimilhança $f(\boldsymbol{x}_2 \mid \theta)$

Exemplo 5.2 *Suponha que* $X_1, \ldots, X_n \mid \lambda \overset{iid}{\sim} Po(\lambda)$ *e que* a priori $\lambda \sim Ga(a, b)$, *com* $a, b > 0$ *conhecidos, ou seja,*

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)}\ \lambda^{a-1}\ e^{-b\lambda}\ , \quad \lambda > 0.$$

Então, com $t = \sum x_i$, tem-se

$$L(\lambda \mid \boldsymbol{x}) \propto \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \propto e^{-n\lambda}\lambda^t$$

$$\begin{aligned} \pi(\lambda \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \lambda)\ \pi(\lambda) \\ &\propto e^{-n\lambda}\lambda^t \times \lambda^{a-1}e^{-b\lambda} \\ &\propto \lambda^{t+a-1}\ e^{-(n+b)\lambda} \\ &\propto \mathrm{Ga}(\lambda \mid t+a, n+b) \end{aligned}$$

e como consequência $\lambda \mid \boldsymbol{x} \sim \mathrm{Ga}(t+a, n+b)$.

Note-se que (*Candidate's formula*)

$$m(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\ \pi(\theta)}{\pi(\theta \mid \boldsymbol{x})} \quad \forall \theta \in \Theta$$

e portanto a distribuição preditiva *a priori* de $\boldsymbol{X}$ é

$$m(\boldsymbol{x}) = b^a\ \frac{\Gamma(t+a)}{\Gamma(a)}\ \prod_{i=1}^{n}(x_i!)^{-1}\ (n+b)^{-(t+a)}$$

Inferência:

Como se atacam os problemas de inferência no âmbito da abordagem bayesiana?

- A resposta completa a esta questão requer a introdução de conceitos de Decisão Estatística
- No entanto, a distribuição *a posteriori* contém toda a informação disponível acerca do parâmetro desconhecido, pelo que o que é necessário é encontrar o sumário dessa distribuição mais apropriado
- Se o problema é o de produzir uma estimativa pontual de θ, podemos calcular
  - a moda de $\pi(\theta \mid \boldsymbol{x})$, a moda *a posteriori*
  - a média *a posteriori*, $E[\theta \mid \boldsymbol{x}]$
  - a mediana *a posteriori*
- se o objectivo é estimar θ por um intervalo, podemos obter $(a(\boldsymbol{x}), b(\boldsymbol{x}))$ tal que $P(\theta \in (a(\boldsymbol{x}), b(\boldsymbol{x})) \mid \boldsymbol{x}) = 0.95$
- se se pretende confrontar as hipóteses estatísticas $H_0 : \theta \in \Theta_0$ e $H_1 : \theta \in \Theta_1$, há que comparar $P(\Theta_0 \mid \boldsymbol{x})$ com $P(\Theta_1 \mid \boldsymbol{x})$

5.2 A distribuição "a priori"

A abordagem bayesiana à inferência estatística é conceptualmente simples e particularmente intuitiva. No entanto, a implementação prática destas ideias é muitas vezes difícil:

- $\pi(\theta)$ deve reflectir informação sobre θ disponível antes dos dados $\boldsymbol{x}$ serem recolhidos. Sumariar informação que em geral existe apenas de forma não organizada numa distribuição de probabilidades não é uma tarefa trivial
- O que podemos fazer quando a informação *a priori* é apenas vaga ou difusa?
- E se o objectivo for o de produzir uma análise estatística tão "objectiva" quanto possível, que use o mínimo possível de informação *a priori* sobre θ?
- Cálculos: só em casos excepcionais é que $\pi(\theta \mid \boldsymbol{x})$ se consegue calcular explicitamente, porque em geral o integral $m(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\ d\theta$ não é calculável analiticamente
- A resposta a muitas questões inferenciais envolve o cálculo de integrais do tipo $E[\psi(\theta) \mid \boldsymbol{x}]$ para diferentes $\psi(\theta)$

"Soluções":

- distribuições *a priori* que permitem cálculos analíticos
- distribuições *a priori* "não-informativas"
- simulação, aproximações analíticas, cálculos numéricos

5.2.1 Distribuições "a priori" conjugadas naturais

Famílias de distribuições *a priori* que permitem cálculos explícitos.

Exemplo 5.3 *Suponhamos que* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$*; a priori* $\theta \sim Be(a, b)$*,* $a, b > 0$ *conhecidos.*

Vimos que

$$\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t + a, n - t + b)$$

ou seja, a actualização é feita dentro da mesma família de distribuições de probabilidade:

$$(a, b) \longrightarrow (t + a, n - t + b)$$

Definição 5.1 *A família* $\Pi = \{\pi(\cdot \mid \tau) : \tau \in \Gamma\}$ *diz-se conjugada natural do modelo estatístico* $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ *se*

1. $\forall \tau_0, \tau_1 \in \Gamma \; \exists \tau_2 \in \Gamma$*:*

$$\pi(\theta \mid \tau_0) \; \pi(\theta \mid \tau_1) \propto \pi(\theta \mid \tau_2)$$

2. $\exists \tau_0 \in \Gamma : f(\boldsymbol{x} \mid \theta) \propto \pi(\theta \mid \tau_0)$

Consequência:

$$\begin{aligned} \pi(\theta \mid \boldsymbol{x}) &\propto f(\boldsymbol{x} \mid \theta) \; \pi(\theta \mid \tau_1) \\ &\propto \pi(\theta \mid \tau_0) \; \pi(\theta \mid \tau_1) \\ &\propto \pi(\theta \mid \tau_2) \in \Pi \end{aligned}$$

ou seja, a actualização da distribuição *a priori* faz-se dentro da mesma família de distriuições!

Exemplo 5.4 *Suponha-se que* $X_i,\ i = 1, \ldots, n \overset{iid}{\sim} Po(\lambda)$.

Então, com $t = \sum x_i$,

$$\begin{aligned} f(\boldsymbol{x} \mid \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &\propto \lambda^t e^{-n\lambda} \\ &\propto \mathrm{Ga}(\lambda \mid t+1, n) \end{aligned}$$

Adicionalmente, $\mathrm{Ga}(\lambda \mid a, b) \times \mathrm{Ga}(\lambda \mid c, d) \propto \mathrm{Ga}(\lambda \mid a + c - 1, b + d)$.

Portanto, a família gama é a conjugada natural do modelo Poisson. A actualização faz-se da forma $(a, b) \rightarrow (a + t, b + n)$.

Escolher (a, b):

- Especificar $E(\theta) = \mu_0$ e $\mathrm{Var}(\theta) = \sigma_0^2$ usando informação *a priori*. Depois resolver em ordem a (a, b) as equações $a/b = \mu_0$ e $a/b^2 = \sigma_0^2$.
- $\mathrm{Ga}(a, b)$ contém a mesma informação que uma amostra de "tamanho" b e total a:

$$(a, b) \rightarrow (a + t, b + n)$$

- Considerar a, b como parâmetros descohecidos e associar-lhes uma distribuição *a priori* $\pi(a, b)$ - distribuição *a priori* especificada de forma hierárquica

Desvantagens:

- Nem sempre existem famílias conjugadas naturais
- A forma funcional da distribuição *a priori* é escolhida por conveniência matemática. Esta forma pode ter consequências importantes ao nível da inferência...

5.2.2 Distribuições a priori "não-informativas"

- Situações onde não há informação *a priori* substancial
- Pretende-se obter uma distribuição a posteriori onde a informação amostral deve ter um peso muito superior ao da informação apriorística
- Obter uma análise estatística de referência, que pode ser comparada com outras análises que incorporem informação apriorística considerável
- Área de investigação activa: "Objective Bayes" — métodos e estratégias para obter distribuições *a priori* "não-informativas" ou "objectivas"

Método de Bayes-Laplace

Princípio da razão insuficiente de Bayes-Laplace: quando em presença de $n > 1$ possibilidades mutuamente exclusivas e exaustivas que sã permutáveis, então a cada possibilidade deve ser atribuída a probabilidade $1/n$

Consequências:

- Θ finito, $\Theta = \{\theta_1, \ldots, \theta_k\}$, então $\pi(\theta_i) = 1/k$, $i = 1, \ldots, k$
- Se Θ for contável, não existe nenhuma distribuição de probabilidade em Θ compatível com este princípio: $\pi(\theta) = c$, $\theta \in \{\theta_1, \ldots, \theta_k, \ldots\}$ implica que $\sum_{\theta \in \Theta} \pi(\theta) = +\infty$: trata-se de uma distribuição *a priori* **imprópria**
- A utilização formal do teorema de Bayes com uma distribuição *a priori* imprópria é um assunto controverso; no entanto, é prática corrente fazê-lo desde que a distribuição *a posteriori* resultante seja própria
- Θ infinito não contável: $\pi(\theta) \propto c$, $\theta \in \Theta$ é imprópria a não ser que Θ seja limitado

Mais importante objecção ao uso de distribuições *a priori* uniformes:

Exemplo 5.5 $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$.

A distribuição de Bayes-Laplace seria $\pi(\theta) = 1$, $\theta \in (0, 1)$. Uma parametrização alternativa do modelo Bernoulli é feita em termos de $\psi = \ln[\theta/(1-\theta)]$. A distribuição uniforme em θ induz a distribuição em ψ dada por

$$\pi(\psi) = \frac{e^{\psi}}{(1+e^{\psi})^2}, \ \psi \in \mathbb{R}$$

Ou seja: ignorância sobre θ implica alguma informação sobre ψ!

Em geral, com $\theta = g(\psi)$,

$$\pi(\psi) = |g'(\psi)| \ \pi(g(\psi))$$

Método de Jeffreys

Idea: invariância com respeito a reparametrizações

Seja $\theta = g(\psi)$ e denote-se por $I_X(\theta)$ a quantidade de informação de Fisher sobre θ em X. Então, a quantidade de informação sobre ψ em X é

$$I_X^*(\psi) = [g'(\psi)]^2 \; I_X(g(\psi)) \; .$$

Se *a priori*

$$\boxed{\pi(\theta) \propto \sqrt{I_X(\theta)}}$$

então a distribuição induzida em ψ é

$$\begin{aligned} \pi(\psi) &= |g'(\psi)| \; \pi(g(\psi)) \\ &\propto |g'(\psi)| \sqrt{I_X(g(\psi))} \\ &\propto \sqrt{I_X^*(\psi)} \end{aligned}$$

Ou seja, é indiferente qual a parametrização a que aplicamos a regra!

Exemplo 5.6 $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$

Relembremo-nos que $I_X(\theta) = E_\theta[-d^2 \ln f(X \mid \theta)/d\theta^2]$. Portanto,

$$I_X(\theta) = E_\theta[X/\theta^2 - (1-X)/(1-\theta)^2] = \theta^{-1}(1-\theta)^{-1}$$

e logo

$$\pi^J(\theta) \propto \sqrt{I_X(\theta)} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto \mathrm{Be}(\theta|1/2, 1/2)$$

ou seja, a distribuição *a priori* de Jeffreys para θ é $\mathrm{Be}(1/2, 1/2)$, que não corresponde a uma distribuição uniforme.

Exemplo 5.7 $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$

É fácil de ver que $I_X(\mu) = 1$, so,

$$\pi^J(\mu) \propto c, \quad \mu \in \mathbb{R}$$

o que corresponde a uma distribuição imprópria. No entanto, a utilização formal do teorema de Bayes leva a

$$\mu \mid x_1, \ldots, x_n \sim \mathrm{N}(\bar{x}, 1/n)$$

5.3 Estimação pontual bayesiana

O problema consiste em produzir um sumário pontual da distribuição *a posteriori*. Escolhas possíveis: moda, média e mediana *a posteriori*

- moda *a posteriori*

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \pi(\theta \mid \boldsymbol{x}) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ f(\boldsymbol{x} \mid \theta)\ \pi(\theta)\end{aligned}$$

Observações:

1. não é necessário conhecer $m(\boldsymbol{x})$ para calcular $\hat{\theta}$
2. Se $\pi(\theta)$ é (aproximadamente) constante, $\hat{\theta}$ coincide (aproximadamente) com a estimativa de máxima verosimilhança de θ
3. Assim, a estimativa de máxima verosimilhança pode ser pensada como uma estimativa bayesiana, mas a sua interpretação é distinta: aqui é o valor de θ mais credível *a posteriori*, e não o valor de θ que torna mais plausível os dados observados
4. Se $\hat{\theta}$ é a moda *a posteriori* de θ e $\psi = g(\theta)$, então a moda *a posteriori* de ψ não é em geral $g(\hat{\theta})$, já que (com $h = g^{-1}$ biunívoca)

$$\pi^{\star}(\psi \mid \boldsymbol{x}) = |h'(\psi)|\ \pi(h(\psi) \mid \boldsymbol{x})$$

2. média *a posteriori*:

$$\hat{\theta} = E[\theta \mid \boldsymbol{x}] = \int_{\Theta} \theta \; \pi(\theta \mid \boldsymbol{x}) \; d\theta$$

3. mediana *a posteriori*:

$$\hat{\theta} : P(\theta \geq \hat{\theta} \mid \boldsymbol{x}) \geq 1/2 \text{ e } P(\theta \leq \hat{\theta} \mid \boldsymbol{x}) \geq 1/2$$

o que no caso contínuo significa

$$\hat{\theta} : P(\theta \leq \hat{\theta} \mid \boldsymbol{x}) = 1/2$$

Numa situação concreta, como escolher entre estas estimativas e outros sumários de $\pi(\theta \mid \boldsymbol{x})$?

- Sem mais elementos, a resposta não é inequivoca. A escolha pode basear-se na facilidade de cálculo ou na relevância de cada uma das estimativas no problema em mão
- uma justificação formal de uma estimativa em detrimento de outras exige a introdução de elementos de decisão estatística
- Função perda: $L(a, \hat{\theta})$ representa a perda incorrida quando se estima θ por $\hat{\theta}$, e o verdadeiro valor de θ é a
- Escolhas mais frequentes: $L(a, \hat{\theta}) = (\hat{\theta} - a)^2$; $L(a, \hat{\theta}) = |\hat{\theta} - a|$
- critério mais usado: usar a estimativa que minimiza o risco *a posteriori*

$$r(\hat{\theta}) = E[L(\theta, \hat{\theta}) \mid \boldsymbol{x}] = \int_{\Theta} L(\theta, \hat{\theta})\ \pi(\theta \mid \boldsymbol{x})\ d\theta$$

  a chamada estimativa de Bayes, $\hat{\theta}^B$
- Se $L(a, \hat{\theta}) = (\hat{\theta} - a)^2$, então $\hat{\theta}^B$ é a média *a posteriori*; se $L(a, \hat{\theta}) = |\hat{\theta} - a|$, então $\hat{\theta}^B$ é a mediana *a posteriori*

Exemplo 5.8 *Suponhamos que* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$*; a priori* $\theta \sim Be(a, b)$*,* $a, b > 0$ *conhecidos.*

Vimos que $\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t + a, n - t + b)$ onde $t = \sum x_i$.

Logo, a média *a posteriori* de θ é

$$\begin{aligned}\hat{\theta} &= \frac{t + a}{t + a + n - t + b} = \frac{t + a}{a + b + n} \\ &= \frac{a + b}{a + b + n}\ \frac{a}{a + b} + \left(1 - \frac{a + b}{a + b + n}\right)\ \frac{t}{n}\end{aligned}$$

que corresponde à média ponderada entre a média de θ *a priori* (dada por $a/(a + b)$) e a média amostral (dada por t/n).

Observações:

- Note-se que quando $n \to +\infty$ com t/n fixo, então $\hat{\theta} \to t/n$, a estimativa de máxima verosimilhança de θ
- Nos casos extremos em que $t = 0$ ou $t = n$, as estimativas de MV de θ são respectivamente 0 e 1; tal não acontece com a estimativa bayesiana: $a/(a + b + n)$ e $(a + n)/(a + b + n)$, respectivamente

Exemplo 5.9 *Seja* $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$.

Vimos que a distribuição *a priori* de Jeffreys é neste caso $\pi^J(\mu) \propto 1$ e conduz a $\mu \mid \boldsymbol{x} \sim \mathrm{N}(\bar{x}, 1/n)$. Assim, a média, a mediana e a moda *a posteriori* de μ coincidem com a estimativa de MV de μ. Note-se a dualidade:

$$E[\mu \mid \boldsymbol{x}] = \bar{x}$$
$$E[\bar{X} \mid \mu] = \mu$$

A distribuição conjugada natural é neste caso a família normal, $\mu \sim \mathrm{N}(m_0, v_0^2)$ e tem-se

$$\mu \mid \boldsymbol{x} \sim \mathrm{N}(m_n, v_n^2)$$

onde $v_n^2 = (n + 1/v_0^2)^{-1}$ e

$$m_n = \frac{n}{n + 1/v_0^2}\ \bar{x} + \frac{1/v_0^2}{n + 1/v_0^2}\ m_0$$

Note-se que a média *a posteriori* é a média ponderada da média amostral e da média *a priori* de μ com pesos proporcionais às respectivas precisões (recíproco da variância). Quando $v_0^2 \to +\infty$, $E[\theta \mid \boldsymbol{x}] \to \bar{x}$, ou seja, a informação amostral é dominante.

5.4 Predição bayesiana

O objectivo aqui é o de predizer uma variável aleatória Y com distribuição de probabilidade dependente de θ com base em observações $x_1, \ldots, x_n$, concretização de amostra casual proveniente de $f(x \mid \theta)$

Como? Determinar a distribuição de $Y \mid x_1, \ldots, x_n$:

$$\begin{aligned} f(y \mid \boldsymbol{x}) &= \int_\Theta f(y, \theta \mid \boldsymbol{x}) \; d\theta \\ &= \int_\Theta f(y \mid \boldsymbol{x}, \theta) \; \pi(\theta \mid \boldsymbol{x}) \; d\theta \end{aligned}$$

conhecida como a distribuição preditiva *a posteriori* de Y.

Na maior parte dos casos de interesse, Y é independente de $X_1, \ldots, X_n$ dado θ, caso em que

$$f(y \mid \boldsymbol{x}) = \int_\Theta f(y \mid \theta) \; \pi(\theta \mid \boldsymbol{x}) \; d\theta$$

Observações

- Note-se a simplicidade conceptual e quão geral é a estratégia
- No âmbito da estatística frequencista, a solução mais frequente é utilizar $f(y \mid \hat{\theta})$, onde $\hat{\theta}$ é uma estimativa de θ: procede-se como se o parâmetro fosse conhecido e igual à estimativa.
- a solução bayesiana reconhece a incerteza associada a qualquer estimativa do parâmetro e incorpora-a na resposta

Exemplo 5.10 *Suponhamos que* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} B(1, \theta)$*; a priori* $\theta \sim Be(a, b)$*,* $a, b > 0$ *conhecidos.*

Vimos que $\theta \mid \boldsymbol{x} \sim \mathrm{Be}\,(t + a, n - t + b)$ onde $t = \sum x_i$.

Pretendemos predizer o resultado da $(n+1)$-ésima prova de Bernoulli, independente das anteriores, que denotamos por Y

$$\begin{aligned} f(y \mid \boldsymbol{x}) &= \int_0^1 f(y \mid \theta)\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\ &= \int_0^1 \theta^y (1-\theta)^{1-y}\ \frac{1}{B(t+a, n-t+b)}\ \theta^{t+a-1}\ (1-\theta)^{n-t+b-1}\ d\theta \\ &= \frac{B(t+a+y, n-t+b+1-y)}{B(t+a, n-t+b)}\ , \quad y = 0, 1 \end{aligned}$$

(Distribuição Beta-Bernoulli.)

Cont:

Usando $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ e $\Gamma(x+1) = x\ \Gamma(x)$, sai que

$$P(Y = 1 \mid \boldsymbol{x}) = f(1 \mid \boldsymbol{x}) = \frac{t+a}{n+a+b}$$

Mais simples:

$$\begin{aligned} P(Y = 1 \mid \boldsymbol{x}) &= E[I_{\{1\}}(Y) \mid \boldsymbol{x}] \\ &= E[\ E[I_{\{1\}}(Y) \mid \theta, \boldsymbol{x}] \mid \boldsymbol{x}] \\ &= E[\ E[I_{\{1\}}(Y) \mid \theta] \mid \boldsymbol{x}] \\ &= E[\ P(Y = 1 \mid \theta) \mid \boldsymbol{x}] \\ &= E[\theta \mid \boldsymbol{x}] \\ &= \frac{t+a}{n+a+b} \end{aligned}$$

Exemplo 5.11 *Suponha-se que* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Ex(\theta)$ *e que* $\theta \sim G(a, b)$, $a, b > 0$ *conhecidos.*

É fácil de ver que $\theta \mid \boldsymbol{x} \sim G(n + a, b + t)$ onde $t = \sum x_i$. Suponha que se pretende predizer a próxima observação, $Y = X_{n+1}$, independente das restantes. Assim,

$$\begin{aligned} f(y \mid \boldsymbol{x}) &= \int_0^{+\infty} f(y \mid \theta)\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\ &= (n + a)\ \left(\frac{b + t}{b + y + t}\right)^{n+a}\ \left(\frac{1}{b + y + t}\right)\ , \quad y > 0 \end{aligned}$$

(Distribuição Gama-Gama.)

Nem sempre é necessário conhecer $f(y \mid \boldsymbol{x})$ para se determinarem estimativas pontuais de Y:

$$\begin{aligned} E[Y \mid \boldsymbol{x}] &= E[\ E[Y \mid \boldsymbol{x}, \theta] \mid \boldsymbol{x}] \\ &= E[\ E[Y \mid \theta] \mid \boldsymbol{x}] \\ &= E[1/\theta \mid \boldsymbol{x}] \\ &= \int_0^{+\infty} \frac{1}{\theta}\ \pi(\theta \mid \boldsymbol{x})\ d\theta \\ &= \cdots \\ &= \frac{b + t}{n + a - 1} \end{aligned}$$

Exemplo 5.12 *Seja* $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$.

Vimos que a distribuição *a priori* de Jeffreys é neste caso $\pi^J(\mu) \propto 1$ e conduz a $\mu \mid \boldsymbol{x} \sim \mathrm{N}(\bar{x}, 1/n)$. Suponhamos que pretendemos predizer a média das m observações seguintes, $\bar{Y} = \sum_{j=1}^{m} X_{n+j}/m$.

$$\begin{aligned} E[\bar{Y} \mid \boldsymbol{x}] &= E[\ E[\bar{Y} \mid \boldsymbol{x}, \mu] \mid \boldsymbol{x}] \\ &= E[\ E[\bar{Y} \mid \mu] \mid \boldsymbol{x}] \end{aligned}$$

Como $\bar{Y} \mid \mu \sim \mathrm{N}(\mu, 1/m)$, segue-se que

$$E[\bar{Y} \mid \boldsymbol{x}] = E[\mu \mid \boldsymbol{x}] = \bar{x}$$

5.5 Estimação intervalar bayesiana

O problema consiste em produzir um sumário intervalar da distribuição *a posteriori*.

Definição 5.2 *Diz-se que* $R(\boldsymbol{x}) = (a(\boldsymbol{x}), b(\boldsymbol{x})) \subset \Theta \subset \mathbb{R}$ *é um intervalo de credibilidade* $(1-\alpha)$ *para* θ *se*

$$P(\theta \in R(\boldsymbol{x}) \mid \boldsymbol{x}) = P(a(\boldsymbol{x}) < \theta < b(\boldsymbol{x}) \mid \boldsymbol{x}) = 1 - \alpha$$

Observações:

- No caso em que $\pi(\theta \mid \boldsymbol{x})$ é contínua, tem-se

$$P(\theta \in R(\boldsymbol{x}) \mid \boldsymbol{x}) = \int_{a(\boldsymbol{x})}^{b(\boldsymbol{x})} \pi(\theta \mid \boldsymbol{x})\, d\theta = 1 - \alpha$$

- Recorde-se que $C(\boldsymbol{X})$ é um intervalo aleatório de confiança $(1-\alpha)$ para θ se

$$P(\theta \in C(\boldsymbol{X}) \mid \theta) = 1 - \alpha \quad \forall\, \theta \in \Theta$$

  mas relativamente ao intervalo de confiança observado, $C(\boldsymbol{x})$, apenas podemos dizer que

$$P(\theta \in C(\boldsymbol{x}) \mid \theta) = \begin{cases} 1 & \text{se } \theta \in (\boldsymbol{x}) \\ 0 & \text{caso contrário} \end{cases}$$

  daí a necessidade do conceito de "confiança"

Exemplo 5.13 *Considere-se que* $X_1, \ldots, X_n \mid \mu \overset{iid}{\sim} N(\mu, 1)$ *e que usamos a respectiva distribuição a priori de Jeffreys para* μ, $\pi^J(\mu) \propto 1$. *Sabemos que* $\mu \mid \boldsymbol{x} \sim N(\bar{x}, 1/n)$.

Obter intervalo de credibilidade $(1-\alpha)$ para μ:

- Existe uma infinidade de intervalos $(a(\boldsymbol{x}), b(\boldsymbol{x}))$ tal que $P(a(\boldsymbol{x}) < \mu < b(\boldsymbol{x}) \mid \boldsymbol{x}) = 1-\alpha$.
- Tipicamente opta-se por seleccionar um intervalo de credibilidade central, i.e. $P(\theta < a(\boldsymbol{x}) \mid \boldsymbol{x}) = P(\theta > b(\boldsymbol{x}) \mid \boldsymbol{x}) = \alpha/2$
- Ou por um intervalo de credibilidade HPD (highest posterior density): determinar c tal que $R(\boldsymbol{x}) = \{\theta : \pi(\theta \mid \boldsymbol{x}) \geq c\}$ e $P(\theta \in R(\boldsymbol{x})) = 1-\alpha$
- no caso vertente, estas duas estratégias coincidem e resultam no intervalo $\bar{x} \pm \frac{1}{\sqrt{n}}\, z_{\alpha/2}$, onde $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$
- note-se que este intervalo de credibilidade coincide numericamente com o intervalo de confiança clássico

Continuação:

- em geral, intervalos de credibilidade e intervalos de confiança não coincidem, e em todo o caso a interpretação do intervalo numérico é distinta nos dois casos
- Exemplo: $n = 30$, $\bar{x} = 25$ e $1 - \alpha = 0.9$, o que implica $z_{\alpha/2} = 1.64$. Assim, o intervalo de credibilidade é $(24.70, 25.30)$
- No caso do intervalo de credibilidade, podemos afirmar que

$$P(\mu \in (24.70, 25.30) \mid \boldsymbol{x}) = 0.9$$

enquanto que no caso do intervalo confiança apenas podemos dizer que

$$P(\mu \in (24.70, 25.30) \mid \mu) = I_{(24.70,25.30)}(\mu)$$

ou que temos uma confiança de 90% que o verdadeiro valor de μ esteja no intervalo $(24.70, 25.30)$

Exemplo 5.14 *Suponha-se que* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} Ex(\theta)$ *e que* $\theta \sim G(a, b)$, $a, b > 0$ *conhecidos.*

Sabemos que $\theta \mid \boldsymbol{x} \sim \mathrm{G}(a+n, b+t)$ onde $t = \sum x_i$.

- É fácil ver que o intervalo de credibilidade $(1-\alpha)$ HPD é $(\underline{\theta}, \bar{\theta})$ que verifica

$$\mathrm{G}(\underline{\theta} \mid a+n, b+t) = \mathrm{G}(\bar{\theta} \mid a+n, b+t)$$

$$\int_{\underline{\theta}}^{\bar{\theta}} \mathrm{G}(\theta \mid a+n, b+t)\, d\theta = 1-\alpha$$

- Já o intervalo de credibilidade central é $(\underline{\theta}, \bar{\theta})$ tal que

$$P(\theta > \bar{\theta} \mid \boldsymbol{x}) = \alpha/2$$
$$P(\theta < \underline{\theta} \mid \boldsymbol{x}) = \alpha/2$$

  o que significa que

$$\bar{\theta} = \frac{1}{2(b+t)} F^{-1}_{\chi^2(2(n+a))}(1-\alpha/2)$$
$$\underline{\theta} = \frac{1}{2(b+t)} F^{-1}_{\chi^2(2(n+a))}(\alpha/2)$$

- Exemplo: se $n = 10$, $t = 10$, $a = b = 1$, $1-\alpha = 0.99$, $F^{-1}_{\chi^2(22)}(0.005) = 8.643$, $F^{-1}_{\chi^2(22)}(0.995) = 42.796$, o que corresponde a um intervalo de credibilidade dado por $(0.39, 1.95)$