Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Mestrado em Estatística
Inferência Bayesiana
1^o semestre de 2020

Fundamentos de Inferência Bayesiana

(Versão em preparação)

Gustavo L. Gilardoni

8 de Março de 2022

Nas seções anteriores discutimos a importância de usar um paradigma inferencial que, a diferença dos métodos clássicos, respeitem o *Princípio de Verossimilhança*. Nesse contexto, vimos o *Princípio da Probabilidade Inversa* de Laplace. Nestas notas vamos ver primeiro que o resultado da aplicação desse princípio não é invariante sob transformações do espaço paramétrico (ou reparametrizações) e como a introdução do conceito de *distribuição a priori* permite resolver esse problema, formalizando assim o *Paradigma Bayesiano*. Posteriormente, serão introduzidos todos os conceitos fundamentais associados a um modelo Bayesiano: além das já mencionadas distribuição a priori e verossimilhança, esses incluem os conceitos de distribuições a posteriori e preditivas, funções de perda (ou, equivalentemente, de utilidade), estimadores Bayesianos, regiões de credibilidade e particularmente aquelas com máxima densidade a posteriori (denominadas regiões HPD, por Highest Posterior Density) e, finalmente, abordaremos o problema de teste de hipótesis.

0 Introdução

Paradigmas inferenciais que não respeitam o Princípio de Verossimilhança (PV) levam a situações não desejáveis, por exemplo fazendo com que experimentos não realizados tenham influência nas inferências. Por outro lado, um paradigma que concorde com o PV deve ser tal que o resultado do experimento x somente influi na inferência a través da função de verossimilhança $\theta \mapsto L_x(\theta) = f_\theta(x)$. O paradigma da *Probabilidade Inversa*, enunciado por Laplace em 1874, parece respeitar o PV, embora também pode levar a resultados incoerentes (veja, por exemplo, o Exercício 5 da Lista 1).

Existem vários paradigmas inferenciais que respeitam o PV. Podemos citar a *Inferência Fiducial* preconizada por Fischer, a Inferência Pivotal de Barnard e a Inferência Estrutural de Fraser (veja, por exemplo, **?**). O mais aceito desses é indiscutivelmente o Paradigma Bayesiano, cujas origens Dentre os paradigmas que respeitam o PV, remontam a **?** e a Laplace (veja, por exemplo, **??**)

Nessa perspectiva, estas notas tem por objetivo apresentar uma breve introdução aos elementos fundamentais da Inferência Bayesiana e da Teoria de Decisão Bayesiana. Muitos desses conceitos serão depois aprofundados ao longo do semestre.

Na Seção 2 é introduzido o conceito fundamental de distribuição *a posteriori* junto com os de distribuições *a priori* e *preditiva*. A Seção 5 trata do problema de estimação Bayesiana e introduz o conceito de *função de perda* ou de *utilidade*. No contexto Bayesiano, os intervalos de confiança da inferência clássica são chamados de *intervalos críveis*, apresentados na Seção 6. Testes de hipóteses e o conceito de *Fator de Bayes* são discutidos na Seção 7.

1 Motivação

Na unidade anterior discutimos o *principio da probabilidade inversa* (PPI) de Laplace. Basicamente, após observar x da densidade $f_\theta(x)$, Laplace sugiere tratar o parâmetro θ como uma variável aleatória com densidade

$$\pi_x(\theta) = \frac{f_\theta(x)}{\int f_\theta(x)\, d\theta}\,. \tag{1}$$

O PPI está baseado numa idéia mais geral conhecido como *Princípio de Indiferença* ou *Princípio da Razão Insuficiente*, devido entre outros a Jacob Bernoulli e ao próprio Laplace. O último escreveu, por exemplo, "*A teoría da probabilidade consiste em reducir todos os eventos do mesmo tipo a um certo número de casos igualmente possíveis, isto é, a aqueles que podemos estar igualmente incertos no*

que diz respeito à sua existência, e em determinar o número de casos favoráveis ao evento cuja probabilidade é procurada." É esse princípio que justifica o cálculo de probabilidades como *casos favoráveis sobre casos possíveis* usado em cursos introdutórios para o cálculo, por exemplo, de probabilidades referentes a lançamentos de uma moeda ou um dado. Porém, na aplicação do Princípio de Indiferença a variáveis contínuas (como a maior parte dos parâmetros nos nossos modelos!) aparecem uma série de problemas tanto epistemológicos quanto práticos—veja, por exemplo, o *Paradoxo de Bertrand*. O seguinte exemplo mostra uma incoerência mais prática do PPI.

Exemplo 1. Suponha uma moeda lançada $n = 12$ vezes, de forma que o número de caras $X \sim \text{Binomial}(n = 12,\theta)$, e que foi observado $X = 9$. Se aplicamos o PPI obtemos

$$\pi_{x=9}(\theta) = \frac{220\,\theta^9\,(1-\theta)^3}{\int_0^1 220\,\theta^9\,(1-\theta)^3\,d\theta} = 2860\,\theta^9\,(1-\theta)^3\,, \tag{2}$$

de forma que $\theta \sim \text{Beta}(\alpha = 10,\beta = 4)$. Porém, se trabalhamos com a reparametrização $\eta = \theta^2$ (i.e. η é a probabilidade de duas caras consecutivas) e aplicamos o PPI na nova parametrização, teríamos

$$\begin{aligned}\pi_{x=9}(\eta) &= \frac{220\,(\sqrt{\eta})^9\,(1-\sqrt{\eta})^3}{\int_0^1 220\,(\sqrt{\eta})^9\,(1-\sqrt{\eta})^3\,d\eta} \\ &= \frac{\eta^{9/2}\,(1-\sqrt{\eta})^3}{\int_0^1 u^9\,(1-u)^3\,2u\,du} = \frac{1}{2}\,\frac{\Gamma(15)}{\Gamma(11)\,\Gamma(4)}\,\eta^{9/2}\,(1-\sqrt{\eta})^3\,, \end{aligned} \tag{3}$$

e, aplicando as regras do cálculo de probabilidade, deveríamos ter que

$$\pi_{x=9}(\theta) = \pi_{x=9}(\eta)\,\left|\frac{d\eta}{d\theta}\right| = \frac{1}{2}\,\frac{\Gamma(15)}{\Gamma(11)\,\Gamma(4)}\,\theta^9\,(1-\theta)^3\,2\,\theta = \frac{\Gamma(15)}{\Gamma(11)\,\Gamma(4)}\,\theta^{10}\,(1-\theta)^3\,,$$

de forma que agora $\theta \sim \text{Beta}(\alpha = 11,\beta = 4)$. □

Uma forma de contornar a dificuldade do PPI em tratar parâmetros contínuos seria introduzir uma função de pesos $w(\theta) \geq 0$ e substituir a equação (1) por

$$\pi_x(\theta) = \frac{w(\theta)\,f_\theta(x)}{\int w(\theta)\,f_\theta(x)\,d\theta}\,, \tag{4}$$

onde, é claro, assumimos que a função $w(\theta)$ satisfaz que $\int w(\theta)\,f_\theta(x)\,d\theta < \infty$. Por enquanto, podemos pensar que a função $w(\theta)$ codifica algum tipo de importância dos valores do parâmetro θ.

Uma observação importante é que na equação (4) $\pi_x(\theta)$ não muda se multiplicamos a função $w(\theta)$ por uma constante, . Por exemplo, se $\int w(\theta)\,d\theta = W < \infty$,

podemos substituir $w(\theta)$ por $w^*(\theta) = W^{-1}\, w(\theta)$, de forma que $\int w^*(\theta)\, d\theta = 1$. Em outras palavras, se a integral da função de pesos for finita, podemos assumir sem perda de generalidade que ela é igual a um.

Com essa observação, para ter uma interpretação precisa de $w(\theta)$, podemos ver que, se tratamos (θ, X) como um par aleatório e denotamos por $p(\theta, x)$ a sua densidade conjunta, por $p(\theta)$ e $p(x)$ as correspondentes densidades marginais e por $p(\theta \,|\, x)$ e $p(x \,|\, \theta)$ as densidades conjuntas, é bem conhecido do cálculo de probabilidades que

$$p(\theta \,|\, x) = \frac{p(\theta)\, p(x \,|\, \theta)}{p(x)} = \frac{p(\theta)\, p(x \,|\, \theta)}{\int p(\theta)\, p(x \,|\, \theta)\, d\theta}\,. \tag{5}$$

Comparando as equações (4) e (5) vemos que é possível interpretar respectivamente (i) $\pi(\theta)$ como $p(\theta \,|\, x)$, (ii) $f_\theta(x)$ como $p(x \,|\, \theta)$ e (iii) $w(\theta)$ como $p(\theta)$. Em outras palavras, a função de pesos $w(\theta) = p(\theta)$ é a densidade marginal do parâmetro θ, a verossimilhança $f_\theta(x) = p(x \,|\, \theta)$ é a densidade condicional do x efetivamente observado dado o parâmetro θ e, finalmente, a densidade $\pi_x(\theta) = p(\theta \,|\, x)$. Essa analogia permite agora trabalhar com a equação (4) ou, equivalentemente, (5) usando o poder de todo o ferramental do cálculo de probabilidade.

Exemplo 1 (Continuação). (Continuação). Quando usamos o PPI na parametrização original para obter (2), é como se estivéssemos afirmando que $p(\theta) = 1$ para $0 < \theta < 1$, isto é, que marginalmente $\theta \sim$ Uniforme(0,1). Por outra parte, quando usamos o PPI na parametrização original para obter (3), estamos afirmando que $p(\eta) = 1$ para $0 < \eta < 1$, isto é que $\eta \sim$ Uniforme(0,1). Esses dois supostos são incompatíveis, pois sabemos do cálculo de probabilidade que, se $\theta \sim$ Uniforme(0,1) e $\eta = \theta^2$, então $p(\eta) = p(\theta)\, d\theta/d\eta = 1 \times (1/2\sqrt{\eta}) = (1/2)\, \eta^{-1/2}$. (Verifique que, se no exemplo usamos a função de pesos $w(\eta) = p(\eta) = (1/2)\, \eta^{-1/2}$ e calcularmos primeiro a densidade $p(\eta \,|\, x = 9)$ e depois a densidade $p(\theta \,|\, x = 9)$ usando a regra do Jacobiano, o resultado é $\theta \sim$ Beta($\alpha = 10$,$\beta = 4$), igual ao obtido quando assumimos que $\theta \sim$ Uniforme(0,1) na parametrização original.) □

Na literatura Bayesiana, $p(\theta)$ é chamada pelo nome de densidade *a priori* e deve refletir a incerteza sobre θ *antes de* observar o valor de X. Já $p(\theta \,|\, x)$ recebe o nome de densidade *a posteriori* e reflete a incerteza sobre θ *após* observar $X = x$. O paradigma Bayesiano, i.e. a equação (5) especifica a regra que relaciona essas duas incertezas (a priori ou antes de observar X, a posteriori ou após observar $X = x$).

Existe uma longa discussão de natureza filosófica sobre se é possível ou conveniente tratar um parâmetro desconhecido como uma quantidade aleatória. Essa discussão tem a ver, é claro, com a interpretação epistemológica do conceito de probabilidade. Embora não é o nosso objetivo aprofundar nessa discussão, deve ser dito que a inferência Bayesiana está baseada na interpretação da probabilidade como medida de incerteza sobre a ocorrência de um evento. Dessa forma,

posto por DeFinetti, "**A** *probabilidade não existe*", no sentido que não existe uma única probabilidade para um evento mas muitas probabilidades, que podem depender tanto do sujeito, pois indivíduos diferentes tem incertezas diferentes, quanto do momento, pois em momentos diferentes podemos ter informação diferente que afetam a nossa incerteza (por exemplo, antes e depois de observar que $X = x$).

2 Distribuições a priori, a posteriori e preditiva

Na inferência parâmetrica clássica, um modelo estatístico é especificado por uma familia de distribuições amostrais $\{p(x \mid \theta) : \theta \in \Theta \subset \mathbb{R}^p\}$. No paradigma Bayesiano acrescenta-se ao modelo a distribuição a priori $p(\theta)$. Para fazer qualquer tipo de inferência ou previsão sobre uma quantidade desconhecida, é necessário calcular a distribuição dessa quantidade **dado** tudo o que é conhecido. No caso que a inferência é sobre todo o parâmetro θ, isso implica no cálculo do que temos chamado a distribuição a posteriori (5). Porém, algumas vezes estamos interessados somente numa parte do parâmetro θ. Por exemplo, no caso da distribuição Normal(μ,σ^2) com média e variância desconhecida, é comum que estamos interessados somente em μ e σ (ou σ^2) é um parâmetro *de estorvo* (*nuissance parameter* em inglês). Nesse caso $\theta = (\mu, \sigma^2)$ e, para se fazer inferência para μ, será necessário calcular primeiro a densidade a posteriori *conjunta*

$$p(\mu, \sigma^2 \mid x) = \frac{p(\mu, \sigma^2)\, p(x \mid \mu, \sigma^2)}{\int_0^\infty \int_{-\infty}^\infty p(\mu, \sigma^2)\, p(x \mid \mu, \sigma^2)\, d\mu\, d\sigma^2}$$

para depois calcular a densidade a posteriori *marginal* de μ,

$$p(\mu \mid x) = \int_0^\infty p(\mu, \sigma^2 \mid x)\, d\sigma^2 \,.$$

Uma situação semelhante ocorre quando o interesse é na previsão de uma quantidade aleatória y relacionada com x e θ a partir de uma distribuição conjunta $p(x, y \mid \theta)$, como, por exemplo, uma observação futura. Nesse caso, para *prever* y precisamos calcular a distribuição condicional de y dado tudo o que foi observado, i.é., $p(y \mid x)$, o que implica no modelo conjunto marginalizar (integrar) com respeito ao parâmetro θ. Essa distribuição é denominada de *preditiva* e é dada por

$$p(y \mid x) = \int p(y, \theta \mid x)\, d\theta = \int p(y \mid x, \theta)\, p(\theta \mid x)\, d\theta = \mathbb{E}_{\theta \sim p(\theta \mid x)}[p(y \mid x, \theta)] \,, \qquad (6)$$

i.é. a média de $p(y \mid x, \theta)$ com respeito à distribuição a posteriori (com respeito a x!) de θ. Frequentemente, y e x são condicionalmente independentes dado θ, isto é, $p(x, y \mid \theta) = p(x \mid \theta)\, p(y \mid \theta)$ e $p(y \mid x, \theta) = p(y \mid \theta)$, de forma que a equação (6) fica

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta \mid x)\, d\theta = \int p(y \mid \theta)\, p(\theta \mid x)\, d\theta = \mathbb{E}_{\theta \sim p(\theta \mid x)}[p(y \mid \theta)] \,. \qquad (7)$$

Exemplo 2. Suponha a situação do Exemplo 1 com a distribuição a priori Uniforme para θ, isto é, $X_1, \ldots, X_{12} \,|\, \theta \overset{iid}{\sim}$ Bernoulli(θ) e a priori $\theta \sim$ Uniforme(0,1) $\equiv$ Beta($\alpha = 1$,$\beta = 1$). Quando $s = \sum_{i=1}^{n} x_i = 9$, vimos anteriormente que a distribuição a posteriori é Beta($\alpha = 10$,$\beta = 4$). Suponha agora que estamos interessados no resultado do próximo lançamento da moeda, que assumimos ser (condicionalmente) independente dos $n = 12$ lançamentos já observados, isto é, $X_1, \ldots, X_{12}, X_{13} \,|\, \theta \overset{iid}{\sim}$ Bernoulli(θ). Usando a equação (7) temos que

$$\begin{aligned} p(X_{13} = 1 \,|\, x_1, \ldots, x_{12}) &= \int_0^1 p(X_{13} = 1 \,|\, \theta)\, p(\theta \,|\, x_1, \ldots, x_{12})\, d\theta \\ &= \int_0^1 \theta\, p(\theta \,|\, x_1, \ldots, x_{12})\, d\theta = \mathbb{E}_{\theta \sim \text{Beta}(10,4)}(\theta) = \frac{10}{14} \end{aligned}$$

e, é claro, $p(X_{13} = 0 \,|\, x_1, \ldots, x_{12}) = 1 - p(X_{13} = 1 \,|\, x_1, \ldots, x_{12}) = \frac{4}{14}$. □

Em todos os exemplos mencionados pode ser visto que, no paradigma Bayesiano, um parâmetro ou uma observação futura recebem basicamente o mesmo tratamento. O motivo é que os dois são quantidades aleatórias que não foram observadas. Por esse motivo, seria mais correto distinguir somente dois tipos de quantidades: *observadas* [$(x_1, \ldots, x_{12})$ no exemplo acima] é *não observadas* (X_{13} e θ). Pode ser que uma quantidade *não observada* nunca será observada (θ), e nesse caso falamos que é um parâmetro, ou poderá vir a ser observada no futuro, e nesse caso falamos de previsão.

O cálculo de distribuições a posteriori resulta mais simples quando a priori é escolhida de uma familia dita *conjugada*. Dada uma distribuição amostral $X \,|\, \theta \sim p(x \,|\, \theta)$, uma familia de distribuições a priori é chamada *conjugada* se, sempre que a distribuição a priori $p(\theta)$ é um membro da família, a distribuição a posteriori também pertence à essa família. As próximas seções abordam alguns casos importantes.

3 Inferência para uma probabilidade: O modelo Beta-Binomial

Considere uma amostra aleatória $X_1, \ldots, X_n$ da distribuição de Bernoulli com parâmetro $\theta \in (0, 1)$, isto é, $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = \theta$. Suponha que a distribuição a priori de θ é Beta(α,β) para algum $\alpha > 0$ e $\beta > 0$. Então, para

calcular a densidade a posteriori, defina $s = \sum_{i=1}^{n} x_i$ de forma que

$$
\begin{aligned}
p(\theta \,|\, x_1, \ldots x_n) &= \frac{p(\theta)\, p(x_1, \ldots, x_n \,|\, \theta)}{\int_0^1 p(\theta)\, p(x_1, \ldots, x_n \,|\, \theta)\, d\theta} \\
&= \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^s\,(1-\theta)^{n-s}}{\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^s\,(1-\theta)^{n-s}\, d\theta} \\
= \frac{\theta^{\alpha+s-1}\,(1-\theta)^{\beta+n-s-1}}{\int_0^1 \theta^{\alpha+s-1}\,(1-\theta)^{\beta+n-s-1}\, d\theta} &= \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\,\Gamma(\beta+n-s)}\, \theta^{\alpha+s-1}\,(1-\theta)^{\beta+n-s-1}\,,
\end{aligned}
\tag{8}
$$

onde a última integral é imediato do fato que $\theta^{\alpha+s-1}\,(1-\theta)^{\beta+n-s-1}$ é o núcleo de uma densidade Beta($\alpha+s$,$\beta+n-s$). Segue de (8) que, a posteriori, $\theta \,|\, x_1, \ldots, x_n \sim$ Beta($\alpha+s$,$\beta+n-s$).

A partir desse resultado, podemos enxergar o paradigma Bayesiano nesta situação como uma regra de *atualização* dos parâmetros da distribuição Beta: antes de observar s, temos parâmetros a priori α e β; após observar $S = s$, temos $\alpha^* = \alpha + s$ e $\beta^* = \beta + n - s$. Por outro lado, veja que

$$
\mathbb{E}(\theta \,|\, x_1, \ldots, x_n) = \frac{\alpha^*}{\alpha^* + \beta^*} = \frac{\alpha+s}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\,\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\,\frac{s}{n}\,.
$$

Em outras palavras, a média a posteriori $\mathbb{E}(\theta \,|\, x_1, \ldots, x_n)$ é uma média ponderada entre a média a priori, $\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$, e o estimador de máxima verossimilhança (EMV) de θ, $\frac{s}{n}$. Note observando os pesos $w = \frac{\alpha+\beta}{\alpha+\beta+n}$ e $1 - w = \frac{n}{\alpha+\beta+n}$ que, quanto maior for n com respeito a $(\alpha+\beta)$, mais perto do EMV $\frac{s}{n}$ vai estar a média a posteriori e viceversa, quanto maior for $(\alpha+\beta)$ com respeito a n, mais perto da média a priori $\frac{\alpha}{\alpha+\beta}$ vai estar a média a posteriori.

É importante observar que que, como o valor da integral no denominador de (8) não depende de θ, não era necessário calculá-la para concluir que a distribuição a posteriori é Beta($\alpha+s$,$\beta+n-s$) Efetivamente, descartando todos os termos que não dependem de θ em (8), podemos escrever

$$
p(\theta \,|\, x_1, \ldots x_n) \propto \theta^{\alpha+s-1}\,(1-\theta)^{\beta+n-s-1}\,,
$$

onde o símbolo "$\propto$" significa "*proporcional, a menos de uma constante que não depende de* θ". A partir dessa observação é imediato que a distribuição a posteriori tem que ser Beta($\alpha+s$,$\beta+n-s$) e que a constante de proporcionalidade tem que ser $\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+s)\,\Gamma(\beta+n-s)}$ (no final, essa constante está para asegurar que $\int_0^1 p(\theta \,|\, x_1, \ldots x_n)\, d\theta = 1$, da mesma forma que a constante que aparece na densidade da distribuição Beta também está para garantir que a integral é um, e por isso devem coincidir).

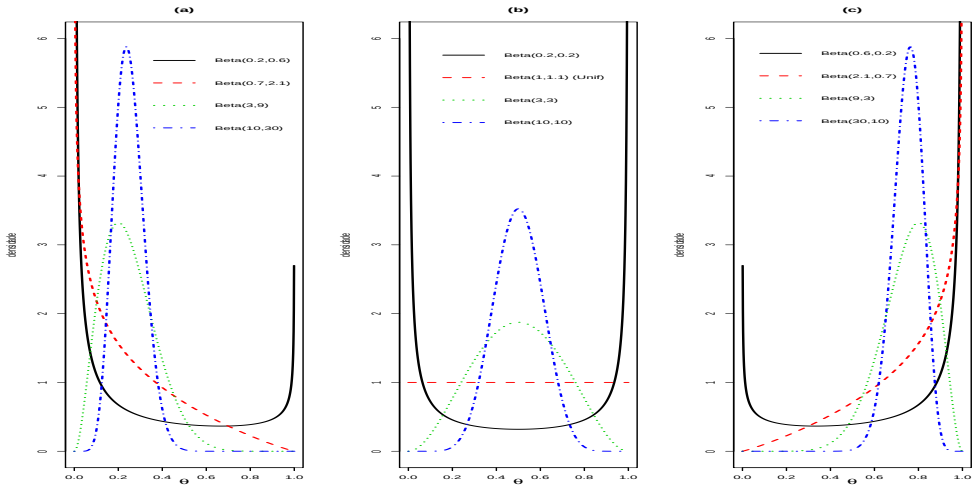

Figura 1: Densidades da distribuição Beta com média **(a)** 0.25, **(b)** 0.5 e **(c)** 0.75. Note os diferentes comportamentos que essas densidades podem assumir.

A família de distribuições Beta é muito flexível para modelar quantidades aleatórias, como o parâmetro θ, que tomam valores no intervalo $(0, 1)$. Por exemplo, se especificarmos dois valores $0 < \mu < 1$ e $\sigma^2 > 0$, sempre é possível achar valores de $\alpha > 0$ e $\beta > 0$ tais que a média da distribuição Beta(α,β) é igual a μ e a variância igual a σ^2 (verifique!). A família inclui distribuições com uma única moda entre zero e distribuições com cauda infinita perto do extremo zero e/ou do extremo um (veja a Figura 1). Note ainda que a distribuição Uniforme que usamos no exemplo 2 é um caso particular da distribuição Beta, quando $\alpha = 1$ e $\beta = 1$.

Exemplo 3. Suponha que uma moeda foi lançada $n = 200$ vezes obtendo-se $s = 93$ caras e $n - s = 107$ coroas. Antes de realizar o experimento, o experimentador estava 90% confiante que a diferença entre a probabilidade de cara θ e a de coroa $(1 - \theta)$ não devia ser superior a 20%.

Devido à simetria entre os rótulos *cara* e *coroa*, parece razoável usar uma distribuição a priori simétrica com respeito ao valor $\theta = 0.5$. No caso das distribuições Beta, isso significa que devemos tomar $\beta = \alpha$. Para garantir que a priori $\mathbb{P}(0.4 < \theta < 0.6) \approx 0.90$ podemos proceder por prova e erro atribuindo valores aos parâmetros α e $\beta = \alpha$ da distribuição Beta, usando algum algoritmo que calcula probabilidades dessa distribuição, tal como a função `pbeta` do **R**. No caso em tela, os valores $\alpha = \beta = 33.39$ satisfazem a restrição desejada (verifique!) Assim, usamos a distribuição a priori Beta($\alpha = 33.39$,$\beta = 33.39$) e, considerando que observamos $s = 93$ caras e $n - s = 107$ coroas, a distribuição a posteriori é Beta($\alpha = 33.39 + 93 = 126.39$,$\beta = 33.39 + 107 = 140.39$) (veja a Figura 2).

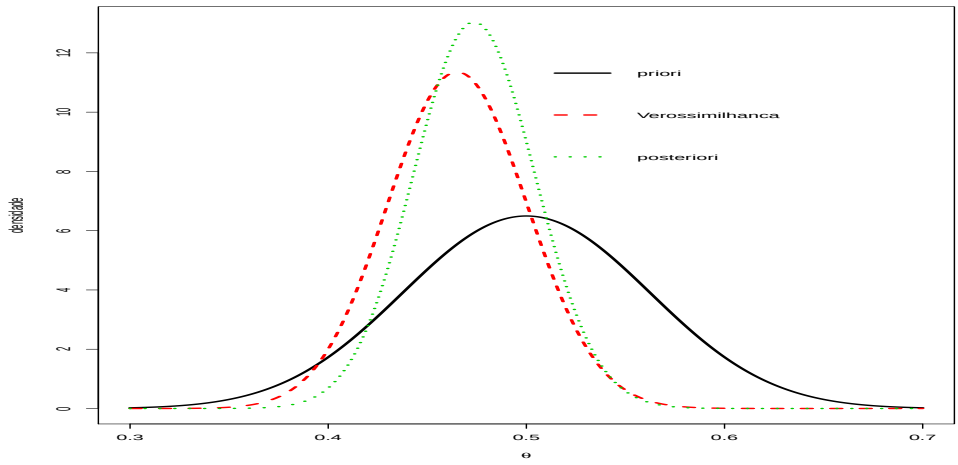

Figura 2: Densidades a priori e a posteriori e verossimilhança para o Exemplo 3 (a verossimilhança foi re-escalada para que sua integral seja igual a um).

Veja que a posteriori

$$\mathbb{E}(\theta \,|\, s = 93, n - s = 107) = \frac{126.39}{126.39 + 140.39} \doteq 0.474$$

(compare com $\mathbb{E}(\theta) = 0.5$ a priori e com a proporção amostral $s/n = 0.465$). Por outro lado, a probabilidade inicial de 90% para o intervalo $0.4 < \theta < 0.6$ passa a ser agora $\mathbb{P}(0.4 < \theta < 0.6 \,|\, s = 93, n - s = 107) \doteq 0.993$. Porém, como a média a posteriori não é mais 0.5, poderia ser mais interessante calcular um intervalo mais ou menos centrado na nova média. Assim, é possível que a afirmação $\mathbb{P}(0.424 < \theta < 0.524 \,|\, s = 93, n - s = 107) \doteq 0.90$ seria mais informativa (os valores 0.424 e 0.524 são respectivamente os percentis 5% e 95% da distribuição a posteriori Beta(126.39,140.39), calculados usando a função `qbeta` do **R**).

Finalmente, observe que a distribuição a priori Beta($\alpha = 33.39$,$\beta = 33.39$) contém bastante informação. Uma forma de pensar essa distribuição a priori surge de considerar que, se tivéssemos partido de uma distribuição a priori Uniforme para θ (i.e. uma Beta com $\alpha = \beta = 1$) e se tivéssemos observado 31 caras e 31 coroas em 62 lançamentos, a distribuição a posteriori seria Beta($\alpha = 33$,$\beta = 33$), muito parecida à que foi usada como priori no exemplo. □

3.1 Distribuição Preditiva

No modelo Beta-Binomial, o problema de previsão usualmente toma a seguinte forma. Temos $X_1, \ldots, X_n; X_{n+1}, \ldots, X_{n+m} \,|\, \theta \overset{iid}{\sim}$ Bernoulli(θ), a priori $\theta \sim$ Beta(α,β)

e, com base na observação do número de sucessos $s_{1:n} = \sum_{i=1}^{n} x_i$ nos primeiros n ensaios, queremos prever o número de sucessos $S_{(n+1):(n+m)} = \sum_{i=n+1}^{n+m} X_i$ nos próximos m ensaios. Usando a equação (7), a distribuição a posteriori $\theta \,|\, s_{1:n} \sim \text{Beta}(\alpha + s_{1:n}, \beta + n - s_{1:n})$ e o fato que $S_{(n+1):(n+m)} \,|\, \theta \sim \text{Binomial}(m, \theta)$, temos para $j = 1, \ldots, m$ que

$$\begin{aligned}
\mathbb{P}(S_{(n+1):(n+m)} = j \,|\, s_{1:n}) &= \int_0^1 \mathbb{P}(S_{(n+1):(n+m)} = j \,|\, \theta) p(\theta \,|\, s_{1:n})\, d\theta \\
&= \int_0^1 \binom{m}{j} \theta^j (1-\theta)^{m-j} \frac{1}{B(\alpha + s_{1:n}; \beta + n - s_{1:n})} \theta^{\alpha + s_{1:n} - 1} (1-\theta)^{\beta + n - s_{1:n} - 1}\, d\theta \\
&= \binom{m}{j} \frac{1}{B(\alpha + s_{1:n}; \beta + n - s_{1:n})} \int_0^1 \theta^{\alpha + s_{1:n} + j - 1} (1-\theta)^{\beta + n - s_{1:n} + m - j - 1}\, d\theta \\
&= \binom{m}{j} \frac{B(\alpha + s_{1:n} + j; \beta + n - s_{1:n} + m - j)}{B(\alpha + s_{1:n}; \beta + n - s_{1:n})}, \qquad (9)
\end{aligned}$$

onde $B(x; y) = \Gamma(x)\,\Gamma(y)/\Gamma(x+y)$ é a função *Beta*. A distribuição (9) é usualmente chamada na literatura pelo nome de *Beta-Binomial*.

Os momentos da distribuição preditiva podem ser calculados diretamente de (9), mas é mais fácil calcular usando as propriedades dos momentos condicionais, o que reduz o problema ao cálculo de momentos na distribuição a posteriori $\theta \,|\, s_{1:n} \sim \text{Beta}(\alpha + s_{1:n}, \beta + n - s_{1:n})$. Lembre do Cálculo de Probabilidades que, dado um vetor aleatório (U, V) com segundo momentos finitos,

$$\mathbb{E}(V) = \mathbb{E}_U\, \mathbb{E}(V \,|\, U) \qquad (10)$$

e

$$\text{Var}(V) = \mathbb{E}_U\, \text{Var}(V \,|\, U) + \text{Var}_U\, \mathbb{E}(V \,|\, U) \qquad (11)$$

Logo, considerando a distribuição condicional do vetor aleatório $(\theta, S_{(n+1):(n+m)})$ dado que $S_{1:n} = s_{1:n}$, temos que

$$\begin{aligned}
\mathbb{E}[S_{(n+1):(n+m)} \,|\, s_{1:n}] &= \mathbb{E}_{\theta \,|\, s_{1:n}}\, \mathbb{E}[S_{(n+1):(n+m)} \,|\, s_{1:n}, \theta] \\
&= \mathbb{E}_{\theta \,|\, s_{1:n}}\, \mathbb{E}[S_{(n+1):(n+m)} \,|\, \theta] = \mathbb{E}[m\,\theta \,|\, s_{1:n}] = m\, \frac{\alpha + s_{1:n}}{\alpha + \beta + n}\,.
\end{aligned}$$

e

$$\text{Var}[S_{(n+1):(n+m)} \,|\, s_{1:n}] = m\, \frac{(\alpha + s_{1:n})\,(\beta + n - s_{1:n})\,(\alpha + \beta + n + m)}{(\alpha + \beta + n)^2\,(\alpha + \beta + n + 1)}$$

Exemplo 4. Nos últimos dois semestres somente 6 dos 30 alunos matriculados na disciplina *Inferência Bayesiana* foram aprovados. Considerando que neste semestre tem 15 alunos matriculados, avalie a probabilidade de mais da metade da turma passar.

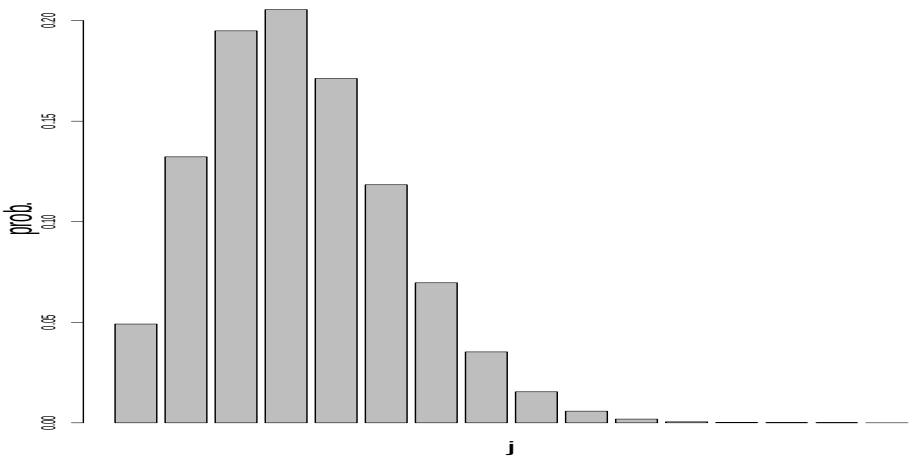

Figura 3: Probabillidades preditivas no Exemplo 4.

Seja θ a probabilidade de um aluno passar, e suponha que, condicionado ao valor de θ, podemos considerar que o total dos 45 alunos como ensaios de Bernoulli idênticos e independentes. Dessa forma temos acima $n = 30$, $s_{1:30} = 6$, $m = 15$ e queremos avaliar $\mathbb{P}(S_{31:45} > 8 \,|\, s_{1:30} = 6)$. Ainda, como não temos informação sobre a taxa de aprovação θ, podemos considerar uma distribuição a priori Uniforme (i.é. $\alpha = \beta = 1$; você teria outra sugestão?) Com essa informação, o gráfico de barras das probabilidades (9) são mostradas na Figura 3. Mais precisamente, temos que $\mathbb{P}(S_{31:45} > 8 \,|\, s_{1:30} = 6) \doteq 0.024$.

Observação. Quando algum dos inteiros considerados é muito grande, o cálculo das probabilidades (9) pode envolver fatoriais ou funções Gama muito grandes. Por isso, pode ser conveniente calcular primeiro o logaritmo das probabilidades, usando por exemplo as funções `lgamma` ou `lbeta` do **R**, e somente no final retornar para a escala original usando a função exponencial.

4 Inferência para dados politômicos: O modelo Dirichlet-Multinomial[1]

A seção anterior considerou o caso da inferência para dados *dicotômicos*, isto é, cujas respostas pertencem somente a duas categorias (cara ou coroa, sucesso ou fracasso, etc.) Nesse caso, vimos que a verossimilhança provenía da distribuição Binomial que depende de um único parâmetro θ (a probabilidade do sucesso de

[1]Antes da leitura desta seção recomenda-se revisar os exercícios referentes às distribuições Multinomial e de Dirichlet da Lista 0.

interesse), e que a família de distribuições a priori Beta possui a propriedade que, nesse caso, a distribuição a posteriori também será um membro dessa família. Nesta seção vamos generalizar esse modelo considerando respostas *politômicas*, isto é, respostas que pertencem a um número finito $k \geq 3$ categorias.

O modelo estatístico mais usado para dados politômicos é a *distribuição Multinomial*, que depende de um vetor de parâmetros $(\theta_1, \ldots, \theta_k)$ que representa as probabilidades das classes. Mais precisamente, dizemos que a observação $\boldsymbol{X} = (X_1, \ldots, X_k)$ segue uma distribuição Multinomial com tamanho n e vetor de probabilidades $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, que escrevemos $\boldsymbol{X} \,|\, \boldsymbol{\theta} \sim \text{Multi}(n; \boldsymbol{\theta})$, se $\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k \,|\, \boldsymbol{\theta}) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^k \theta_i^{x_i}$, onde os x_is são inteiros não negativos tais que $\sum_{i=1}^k x_i = n$ e os parâmetros $\theta_i \geq 0$ satisfazem que $\sum_{i=1}^k \theta_i = 1$.

Na seção anterior vimos o modelo Beta-Binomial. Lembrando que as distribuições Multinomial e de Dirichlet são generalizações respectivamente da Multinomial e da Beta, não é surprendente que Dirichlet é uma família de distribuições a priori atrativa para tratar dados com distribuição Multinomial. Suponha então que a priori $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, no sentido que $p(\theta_1, \ldots, \theta_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$, onde os hyperparametros $\alpha_i > 0$.

O cálculo da distribuição a posteriori é muito simples, pois

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

de forma que $\boldsymbol{\theta} \,|\, \boldsymbol{x} \sim \text{Dirichlet}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$. Se denotamos $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ e $\boldsymbol{x} = (x_1, \ldots, x_k)$, vemos que a distribuição de Dirichlet a priori tem parâmetros $\boldsymbol{\alpha}$ e, após observar $\boldsymbol{x}$, a posteriori os parâmetros são $\boldsymbol{\alpha} + \boldsymbol{x}$.

As propriedades da Distribuição de Dirichlet podem ser usadas para fazer inferências sobre funções do parâmetro $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Por exemplo, se estamos interessados numa componente específica θ_i, sabemos que a distribuição a posteriori marginal é $\theta_i \,|\, \boldsymbol{x} \sim \text{Beta}(\alpha_i + x_i, \sum_{j \neq i} \alpha_j + x_j)$. Por outra parte, se o interesse for em funções mais complicadas do parâmetro, tais como diferenças $\theta_j - \theta_i$ ou outras combinações lineares de $\boldsymbol{\theta}$, para as quais o cálculo explícito da distribuição a posteriori for impossível ou difícil, é possível obter aproximações baseadas em amostras simuladas da posteriori para todo o parâmetro $\boldsymbol{\theta}$. Existem varias formas de simular observações da distribuição de Dirichlet, mas possivelmente a mais fácil é usando a relação entre a distribuição Gamma e a Dirichlet. Nesse sentido, lembre que se $Y_1, \ldots, Y_k$ é um vetor de variáveis aleatórias independentes com $Y_i \sim \text{Gama}(\alpha_i, 1)$, então o vetor das proporções $R_i = Y_i / \sum_{j=1}^k Y_j$ segue uma Dirichlet$(\alpha_1, \ldots, \alpha_k)$. Assim, para simular da distribuição a posteriori basta simular variáveis independentes com distribuição Gama$(\alpha_i + x_i, 1)$ e depois calcular as respectivas razões R_i.

Não existe consenso na literatura sobre a forma da distribuição não informativa

no caso da priori Dirichlet. Uma primeira possibilidade surge de observar que, quando $\alpha_i = 1$ para $i = 1, \ldots, k$, a densidade a priori é constante, de forma que nesse caso temos uma distribuição Uniforme no simplexo definido pelas condições $\theta_i \geq 0$ e $\sum_{i=1}^{k} \theta_i = 1$. Como veremos na próxima unidade, a priori com $\alpha_1 = \cdots = \alpha_k = 1/2$, denominada de *priori de Jeffreys* e a priori imprópria correspondente ao caso $\alpha_1 = \cdots = \alpha_k = 0$ também tem sido propostas na literatura. Em todo caso, é importante observar que quanto maior o tamanho amostral n, menor será o impacto de qualquer uma dessas prioris "não informativas" nas inferências realizadas.

Exemplo 5. Numa pesquisa eleitoral realizada na Argentina na véspera do primeiro turno da eleição presidencial de 2015, os candidatos Scioli (Frente para la Victoria), Macri (Cambiemos) e Massa (Unidad para una Nueva Alternativa) receberam respectivamente a preferência de 486, 348 e 295 do total de 1250 eleitores entrevistados. Outros candidatos eram preferidos de 121 entrevistados. Supondo que a pesquisa seja uma amostra da população total de eleitores que concorreriam a votar, queremos obter intervalos que contenham o percentual da votação de cada um dos três candidatos principais com 95% de probabilidade a posteriori. Além disso, um problema importante era prever se haveria segundo turno. Nesse sentido, a legislação eleitoral na Argentina é mais complicada que no Brasil. O segundo turno é disputado se (i) o candidato mais votado não chegar aos 45% dos votos ou que (ii) superando o 40% dos votos, a diferença para o segundo mais votado seja menor que 10%.

Para a resolução defina $\boldsymbol{x} = (x_{Scioli}, x_{Macri}, x_{Massa}, x_{Outros}) = (486, 348, 295, 121)$ o número de eleitores observados, que modelamos de acordo com uma distribuição Multinomial com $n = 1250$ e parâmetro $\boldsymbol{\theta} = (\theta_{Scioli}, \theta_{Macri}, \theta_{Massa}, \theta_{Outros}) \sim$ Dirichlet(1,1,1,1) (outras prioris podem ser consideradas, mas devido ao n ser consideravelmente grande, o resultado final não deve diferir muito do que encontraremos aqui). Logo, a distribuição a posteriori é uma Dirichlet com $\alpha_{Scioli} = 486+1 = 487$, $\alpha_{Macri} = 349$, $\alpha_{Massa} = 296$, $\alpha_{Outros} = 122$. As distribuições marginais são Betas com parâmetros (α, β) respectivamente iguais a $(487, 767)$, $(349, 905)$, $(296, 958)$ e $(122, 1132)$ para Scioli, Macri, Massa e Outros. Os intervalos desejados podem ser calculados usando os percentis 2.5 e 97.5% das distribuições correspondentes. Assim, os intervalos são $(0.362, 0.415)$ para Scioli, $(0.254, 0.303)$ para Macri e $(0.213, 0.260)$ para Massa[2]

O cálculo da probabilidade exata a posteriori de segundo turno é relativamente difícil. Considerando que a chance dos outros candidatos ganharem no primeiro turno é claramente desprezível, deveríamos calcular a probabilidade de θ_{Scioli} ou θ_{Macri} ou θ_{Massa} ser maior de 45% mais a probabilidade de uma delas estar entre 40 e 45% enquanto as outras duas ficam mais de 10% menores do que a maior. Embora todos esses eventos são excludentes, o cálculo é consideravelmente tedioso. Porém,

[2] Na seção *Intervalos Críveis* veremos como obter intervalos com a mesma cobertura a posteriori mas com comprimento menor do que os achados aqui.

uma aproximação baseada em simulação a posteriori é fácil de se obter. Podemos simular M observações da distribuição a posteriori Dirichlet(487,349,296,122) e verificar para cada uma delas se haveria segundo turno checando as condições para o mesmo. Para M suficientemente grande, a Lei dos Grandes Números garante que a proporção de observações para as quais tem segundo turno deve estar muito próxima da probabilidade desejada. No caso do exemplo, usamos $M = 10.000$ e obtivemos uma probabilidade aproximada de segundo turno de 0.804. O código **R** usado é mostrado a seguir. (Na eleição de 2015, houve segundo turno entre Scioli e Macri e o último foi eleito com 51.34% dos votos).

```
> set.seed(4321)
> M<-10000
> alfa<-c(487,349,296,122)
> segundo<-rep(1,M)
> for(m in 1:M){
+   y<-rgamma(4,alfa,rate=1)
+   theta<-y/sum(y)
+   ordem<-order(theta)
+   if(theta[ordem==4]>=0.45) segundo[m]<-0
+   if((theta[ordem==4]>0.40)&((theta[ordem==4]-theta[ordem==3])>=0.10))
+     segundo[m]<-0
+ }
> sum(segundo)/M
[1] 0.8044
```

5 Estimação Bayesiana

Suponha um modelo $X \mid \theta \sim p(x \mid \theta)$ e $\theta \sim p(\theta)$, onde o espaço paramétrico Θ é um subconjunto de $\boldsymbol{R}^p$. O objetivo do problema de estimação é achar um valor $\hat{g}(x)$ que "estima" de alguma forma ótima, a ser definida a seguir, o valor de uma função $g(\theta) : \Theta \to \boldsymbol{R}^k$. A função $g(\theta)$ e o valor $\hat{g}(x)$ são chamados respectivamente *estimando* e *estimativa*[3].

Assim como na estatística clássica, a definição de um critério de otimalidade passa necessariamente pela definição de uma *função de perda* $L(g, \hat{g})$ que mede a

[3] A quantidade aleatória $\hat{g}(X)$ (isto é, a quantidade aleatória $x \to \hat{g}(x)$) é chamada de *estimador*. Na inferência clássica existe uma teoria relacionada com estimadores bayesianos que, embora o nome, não é relevante na apresentação do paradigma bayesiano abordada nestas notas. Mais precisamente, como toda a inferência bayesiana é condicional ao valor de x observado, o paradigma bayesiano está interessado em $\hat{g}(x)$ e não no valor da estimativa para x's que não ocorreram

discordância entre o estimando $g(\theta)$ e a estimativa $\hat{g}(x)$[4].

- **Problema**: Estimar uma função $\eta = g(\theta)$ do parâmetro θ.
- **Função de perda**: $L(\widehat{\eta}, \eta)$
- $\widehat{\eta} = \widehat{g}(x)$ é uma estimativa Bayesiana de $\eta = g(\eta)$ se, para qualquer outro valor $\tilde{\eta}$,
$$\mathbb{E}[L(\widehat{\eta}, \eta) \,|\, x] \leq \mathbb{E}[L(\tilde{\eta}, \eta) \,|\, x]$$
- Veja que da definição não segue necessariamente que o estimador Bayesiano deva ser único.
- Veja também que a condição é condicional no valor de x. Por isso, falamos de *estimativa* e não de *estimador*.

Funções de Perda usuais:

- **Função de perda quadrática** (PQ): $L(\widehat{\eta}, \eta) = (\widehat{\eta} - \eta)^2$.
- **Função de perda absoluta** (PA): $L(\widehat{\eta}, \eta) = |\widehat{\eta} - \eta|$.
- **Função de perda zero-um** (depende de um valor $\epsilon > 0$):
$$L(\widehat{\eta}, \eta) = 1 - I_{[\eta-\epsilon;\eta+\epsilon]}(\widehat{\eta}) = \begin{cases} 0 \text{ se } \eta - \epsilon < \widehat{\eta} < \eta + \epsilon \\ 1 \text{ caso contrario} \end{cases} .$$
- No caso de PQ e PA, os dois lemas a seguir mostram que calcular estimadores bayesianos reduz-se basicamente a calcular valores esperados e medianas das distribuições a posteriori do estimando $\eta = g(\theta)$.

Lema 1. *Seja Y uma v.a. com segundo momento $\mathbb{E}\, Y^2 < \infty$, e defina para a real a função $h(a) = \mathbb{E}(Y - a)^2$. Então $h(a)$ é mínima quando $a = \mathbb{E}\, Y$, isto é, $h(a) \geq h(\mathbb{E}\, Y) = \mathrm{Var}(Y)$.*

Demonstração. Denote $\mu = \mathbb{E}\, Y$. Então

$$\begin{aligned} h(a) = \mathbb{E}(Y - \mu + \mu - a)^2 &= \mathbb{E}(Y - \mu)^2 + 2\,(\mu - a)\ \mathbb{E}(Y - \mu) + (\mu - a)^2 \\ &= \mathbb{E}(Y - \mu)^2 + (\mu - a)^2 = h(\mu) + (\mu - a)^2 \geq h(\mu) \end{aligned}$$

(Note que $h(\mu) = \mathrm{Var}\, Y$). □

[4] Alguns autores preferem utilizar uma *função de utilidade* definida como o oposto da perda. A distinção é puramente semântica: Abaixo, onde minimizamos perda esperada, deveríamos maximizar utilidade ou, onde dizemos que a perda é convexa, diríamos que a utilidade é cóncava e assim por diante.

Lema 2. *Seja Y uma v.a. tal que $\mathbb{E}|Y| < \infty$, e defina para a real a função $k(a) = \mathbb{E}|Y - a|$. Então, se m é uma mediana de Y, segue que $k(a) \geq k(m)$.*

Demonstração. Exercício (para provar, use a definição formal de *mediana*: m é uma mediana de Y se $\mathbb{P}(Y \leq m) \geq 1/2$ e $\mathbb{P}(Y \geq m) \geq 1/2$). □

No caso da perda zero-um o cálculo da estimativa bayesiana é um pouco mais complicada. Quando a densidade a posteriori de η é unimodal, é possível mostrar que a estimativa bayesiana é o ponto médio do intervalo com comprimento $2\,\epsilon$ que tem máxima probabilidade a posteriori. Nesse caso, o problema é semelhante ao cálculo de intervalos críveis HPD que discutiremos na próxima seção. Embora o cálculo para ϵ fixo seja mais complicado, observe que no caso de densidade unimodal, o intervalo com probabilidade a posteriori máxima deve conter a moda da distribuição. Portanto, quando $\epsilon \to 0$, o limite das estimativas bayesianas sob perda zero-um tende a essa moda.

Quando a densidade a posteriori de η é simétrica, segue das considerações anteriores que todas essas estimativas (sob PQ ou PA ou perda zero-um para ϵ fixo ou ainda no limite quando $\epsilon \to 0$) devem coincidir com o centro de simetria da densidade.

Quando a densidade a posteriori não é simétrica, pode ser necessário calcular numericamente todas ou algumas das estimativas anteriores.

5.1 Modelo Beta-Binomial

Suponha que o modelo da Seção 3 de forma que $\theta \,|\, s \sim \text{Beta}(\alpha^*,\beta^*)$ com $\alpha^* = \alpha + s$ e $\beta^* = \beta + n - s$

- A estimativa de θ sob PQ é $\mathbb{E}(\theta \,|\, s) = \alpha^*/(\alpha^* + \beta^*) = (\alpha + s)/(\alpha + \beta + n)$.
- Já para calcular a estimativa sob PA é usualmente necessário recorrer a uma aproximação numérica, pois não existe uma fórmula fechada para a mediana da distribuição Beta. Usando por exemplo a linguagem **R** podemos calcular a estimativa com `qbeta(0.5,`α^*`,`β^*`)` (a aproximação não é necessária quando $\alpha^* = \beta^*$; por que?)
- Mais geralmente, suponha que queremos estimar $\eta_m = \theta^m$. Por exemplo, se m é um inteiro, η_m é a probabilidade de obtermos m sucessos em m ensaios.

Nesse caso, sob PQ,

$$\widehat{\eta_m} = \mathbb{E}(\eta_m \,|\, s) = \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \int_0^1 \theta^m\, \theta^{\alpha^*-1}\,(1-\theta)^{\beta^*-1}\, d\theta$$
$$= \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)}\, \frac{\Gamma(\alpha^* + m)\,\Gamma(\beta^*)}{\Gamma(\alpha^* + \beta^* + m)} = \frac{\Gamma(\alpha^* + m)}{\Gamma(\alpha^*)}\, \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + m)}.$$

Por exemplo, para $m = 2$, temos que

$$\widehat{\eta_2} = \widehat{\theta^2} = \frac{\alpha^*\,(\alpha^* + 1)}{(\alpha^* + \beta^*)\,(\alpha^* + \beta^* + 1)}\,. \tag{12}$$

- Calcular um estimador de η_m sob PA pode ser um pouco mais complicado. Nesse caso é também possível usar, com um pouco de cuidado, a função `qbeta` do **R** (faça como exercício). Uma outra possibilidade é recorrer à simulação da distribuição a posteriori: pode-se gerar uma amostra $\theta_1, \ldots, \theta_R$ da distribuição a posteriori usando a função `rbeta`, calcular $\eta_{m,r} = \theta_r^m$ e depois a mediana amostral dos valores $\eta_{m,1}, \ldots, \eta_{m,R}$. Propriedades usuais das estatísticas de ordem garantem que, quando $R \to \infty$, a mediana amostral converge para a estimativa de η_m sob PA.

Exemplo 2 (Continuação). Temos que a distribuição a posteriori de θ é Beta($\alpha^* = 10$,$\beta^* = 4$). Portanto, a estimativa de θ sob PQ é $\mathbb{E}(\theta \,|\, s) = \frac{\alpha^*}{\alpha^*+\beta^*} = 10/14 \doteq 0.714$. A estimativa de θ sob PA é $\text{med}(\theta \,|\, \text{s}) \doteq 0.725$, que calculamos digitando `qbeta(0.5,10,14)` numa sessão do **R**. Embora a estimativa de θ sob perda zero-um seja um pouco mais difícil de calcular (veja a próxima seção), o limite quando $\epsilon \to 0$ é a moda da distribuição Beta, isto é, $\frac{\alpha^*-1}{\alpha^*+\beta^*-2} = 9/12 = 0.75$.

Se o interesse for em $\eta_2 = \theta^2$, a estimativa sob PQ é dada pela equação (12) e neste caso é $\frac{(10)(11)}{(14)(15)} \doteq 0.524$. Para calcular uma aproximação da estimativa de η_2 podemos proceder da seguinte forma: primeiro, geramos uma amostra de tamanho M grande da distribuição Beta(10,4), depois calculamos a mediana da amostra cujos elementos são os quadrados da amostra original. No nosso caso, usamos $M = 50000$ é a aproximação calculada foi 0.526 (o código **R** para obter esse valor é muito simples: `quantile(rbeta(50000,10,4)^2,probs=0.5)`; veja que cada vez ao simular repetidamente podemos obter valores ligeiramente diferentes).

6 Regiões Críveis

Na inferência clássica, uma região de confiança $100(1-\alpha)\%$ para uma função do parâmetro $\eta = \eta(\theta)$ é uma família de conjuntos $C(X)$ tal que $\mathbb{P}(\eta \in C(X) \,|\, \theta) \geq 1-\alpha$ para todo θ. Como todas as medidas da inferência clássica, essa probabilidade

é calculada com respeito à distribuição amostral, isto é, $\mathbb{P}[\eta \in C(X) \,|\, \theta] = \int I[\eta \in C(x)]\, p(x \,|\, \theta)\, dx$. Quando a região $C(x)$ é calculada para aquele x que realmente ocorreu, fala-se que temos $100(1-\alpha)\%$ de *confiança* que $\eta \in C(x)$.

Na inferência bayesiana, toda a inferência depende da distribuição a posteriori $p(\theta \,|\, x)$ e portanto o natural é falar de probabilidades no espaço paramétrico condicionais ao valor de x observado. Assim, no paradigma bayesiano, o análogo ao intervalo de confiança clássico é um conjunto $C(x)$ tal que $\mathbb{P}[\eta \in C(x) \,|\, x] = 1-\alpha$. Nesse caso, diz-se que a região $C(x)$ tem *credibilidade* $100(1-\alpha)\%$ para η. Veja que no paradigma bayesiano somente precisamos conhecer a forma de $C(x)$ para o x que efetivamente ocorreu.

Exemplo 6. Suponha uma amostra de tamanho $n = 10$ da distribuição de Poisson com $n = 10$, $s = n\bar{x} = 13$ e assuma que a priori a média $\theta \sim \text{Gama}(\alpha = 0.1, \beta = 0.1)$, de forma que a posteriori $\theta \,|\, x \sim \text{Gama}(13.1, 10.1)$. As seguintes regiões tem todas credibilidade 90% para θ (verifique!): $C_1 \doteq (0.611; 1.798)$, $C_2 \doteq (0.848; 2.273)$ e $C_3 \doteq (0; 1.220) \cup (1.309; \infty)$. Observe que C_1 parece ser mais preciso que C_2 no sentido que o seu comprimento é menor, enquanto C_3 tem comprimento infinito.

O exemplo anterior mostra que podemos construir infinitas regiões com uma credibilidade dada. Dessas, gostaríamos escolher aquela que é "*mais precisa*". Usualmente, considera-se mais precisa aquela que tem volume (comprimento, área etc. dependendo da dimensão) menor no espaço do estimando η. Parece bastante intuitivo que para minimizar a medida da região devemos incluir nela pontos que tem alta probabilidade a posteriori. Nesse sentido, uma região crível $C(x)$ para $\eta = g(\theta)$ é dita HPD (acrónimo do inglês Highest Posterior Density) se, dados dois pontos quaisquer $\eta_1 \in C(x)$ e $\eta_2 \notin C(x)$, segue que $p(\eta = \eta_1 \,|\, x) \geq p(\eta = \eta_2 \,|\, x)$. Em outras palavras, as regiões HPD incluem valores de η que tem densidade a posteriori grande até atingir o nível de credibilidade desejado.

No caso univariado, se a densidade a posteriori de η for unimodal e continua, é fácil verificar que o intervalo HPD deve satisfazer que a densidade a posteriori de η avaliada nos dois extremos do intervalo deve ser igual.

Exemplo 6 (Continuação). O intervalo HPD para θ é $(0.714; 1.862)$, que corresponde aos percentis 0.031 e 0.931 da distribuição Gama(13.1,10.1) e tem comprimento 1.148. A Figura 4 mostra intervalos HPD com vários níveis de credibilidade.

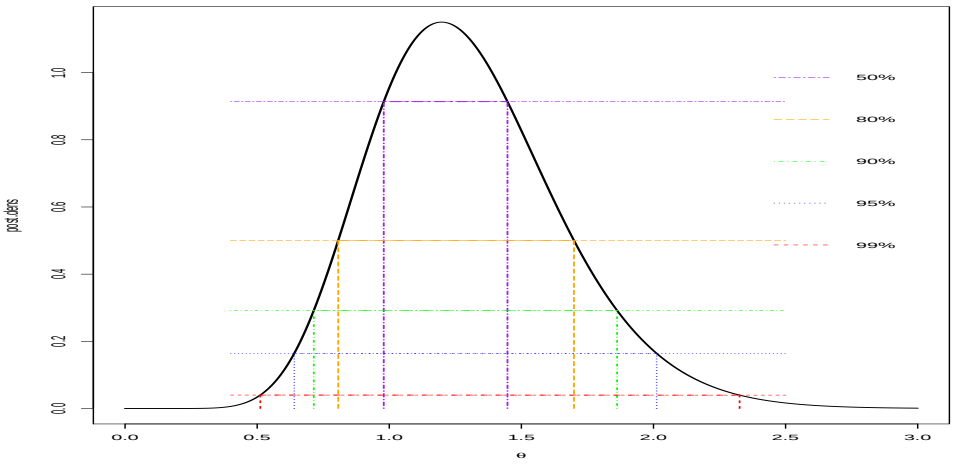

Figura 4: Intervalos HPD para $\theta \,|\, s \sim$ Beta(10,4). Observe que a densidade a posteriori toma o mesmo valor nos dois extremos de cada intervalo.

7 Testes de hipóteses Bayesianos e Fator de Bayes

Considere o modelo com distribuição a priori $p(\theta)$ e verossimilhança $p(x \,|\, \theta)$ ($\theta \in \Theta$). O problema de teste de hipóteses consiste em decidir se $\theta \in \Theta_0$ (a hipótese nula) ou se $\theta \in \Theta_1 = \Theta_0^c$ (a hipótese alternativa). Como vimos nos exemplos, na inferência bayesiana a decisão vai depender das probabilidades a posteriori $\mathbb{P}(\theta \in \Theta_0 \,|\, x)$ e $\mathbb{P}(\theta \in \Theta_1 \,|\, x) = 1 - \mathbb{P}(\theta \in \Theta_0 \,|\, x)$.

- *Chance relativa* ("odds" no inglês) de um evento A:

$$\mathcal{O}(A) = \frac{\mathbb{P}(A)}{\mathbb{P}(A^c)} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}$$

- É claro que

$$\mathbb{P}(A) = \frac{\mathcal{O}(A)}{1 + \mathcal{O}(A)}\,.$$

  Por exemplo, se a chance relativa de um evento A é 7/9, segue que a probabilidade de A é 7/16.

- Sejam A e B dois eventos. Segue imediatamente do Teorema de Bayes que

$$\mathcal{O}(A \,|\, B) = \frac{\mathbb{P}(A \,|\, B)}{\mathbb{P}(A^c \,|\, B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A^c)} \, \frac{\mathbb{P}(B \,|\, A)}{\mathbb{P}(B \,|\, A^c)}$$

ou, equivalentemente,

$$\frac{\mathcal{O}(A \,|\, B)}{\mathcal{O}(A)} = \frac{\mathbb{P}(B \,|\, A)}{\mathbb{P}(B \,|\, A^c)}$$

- Agora, voltando ao problema de teste de hipóteses, se associarmos na relação acima A com Θ_0, A^c com Θ_1 e B com uma observação x, temos que

$$\frac{\mathcal{O}(\Theta_0 \,|\, x)}{\mathcal{O}(\Theta_0)} = \frac{p(x \,|\, \theta \in \Theta_0)}{p(x \,|\, \theta \in \Theta_1)}\,.$$

  O termo da esquerda mede a mudança nas chance relativa da hipótese nula antes e após observar x. O da direita mede a verossimilhança relativa de x sob a hipótese nula com respeito à sua verossimilhança sob a hipótese alternativa. Ele é chamado de *Fator de Bayes* na literatura bayesiana. Nós vamos denotar ele por $B^p_{0,1}(x)$. De uma certa forma, ele mede quanto a observação x suporta a hipótese nula com respeito à hipótese alternativa. Deve ser interpretado com cuidado pois, como essa igualdade mostra, o fator de Bayes depende da distribuição a priori.

- Veja que

$$\begin{aligned} B^p_{0,1}(x) &= \frac{\mathbb{P}(\theta \in \Theta_1)}{\mathbb{P}(\theta \in \Theta_0)} \, \frac{\int_{\Theta_0} p(x \,|\, \theta)\, p(\theta)\, d\theta}{\int_{\Theta_1} p(x \,|\, \theta)\, p(\theta)\, d\theta} \\ &= \frac{\int_{\Theta_0} p(x \,|\, \theta)\, p(\theta) / \,\mathbb{P}(\theta \in \Theta_0)\, d\theta}{\int_{\Theta_1} p(x \,|\, \theta)\, p(\theta) / \,\mathbb{P}(\theta \in \Theta_1)\, d\theta} = \frac{\int_{\Theta_0} p(x \,|\, \theta)\, p(\theta \,|\, \theta \in \Theta_0)\, d\theta}{\int_{\Theta_1} p(x \,|\, \theta)\, p(\theta \,|\, \theta \in \Theta_1)\, d\theta} \end{aligned}$$

  é uma ração entre a verossimilhança média (com respeito à distribuição a priori) sob a hipótese nula e sob a hipótese alternativa.

- É interessante comparar a expressão anterior com a estatística do teste de razão de verossimilhança da estatística clássica;

$$\lambda_{0,1}(x) = \frac{\sup_{\theta \in \Theta_0} p(x \,|\, \theta)}{\sup_{\theta \in \Theta_1} p(x \,|\, \theta)} = \frac{p(x \,|\, \hat{\theta}_0)}{p(x \,|\, \hat{\theta}_1)}\,,$$

  onde $\hat{\theta}_0$ e $\hat{\theta}_1$ são as estimativas de Máxima Verossimilhança de θ respectivamente sob H_0 e sob H_1.

Exemplo 1 (Continuação) Suponha que $X \,|\, \theta \sim$ Binomial($n = 12$,θ) e que a priori $\theta \sim$ Uniforme(0,1). Para testar $H_0 : \theta \leq 1/2$ contra $H_a : \theta > 1/2$, suponha que foi observado $x = 9$. Então a priori $\mathbb{P}(\theta \leq 1/2) = \mathbb{P}(\theta > 1/2) = 1/2$, de forma que $\mathcal{O}(\theta \leq 1/2) = 1$. A posteriori $\theta \,|\, x = 9 \sim$ Beta($\alpha = 10$,$\beta = 4$) e $\mathbb{P}(\theta \leq 1/2 \,|\, x = 9) \doteq 0.046$ e $\mathbb{P}(\theta > 1/2 \,|\, x = 9) \doteq 1 - 0.046 = 0.954$. Logo

$\mathcal{O}(\theta \leq 1/2 \,|\, x = 9) \doteq 0.048$, ou alternativamente $\mathcal{O}(\theta > 1/2 \,|\, x = 9) \doteq 20.67$, indicando que a H_a é aproximadamente 20 vezes mais provável a posteriori do que a H_0. Neste caso o fator de Bayes é também $B^p_{0,1} \doteq 0.048$ (pois $\mathcal{O}(\theta \leq 1/2) = 1$).

Porém, suponha que ao invés da priori Uniforme(0,1) usamos uma Beta($\alpha = 1/2$,$\beta = 1/2$), denominada neste caso de priori de Jeffreys e que alguns autores preferem como priori "não-informativa". Como essa priori também é simétrica, teremos ainda que $\mathcal{O}(\theta \leq 1/2) = 1$, mas agora a posteriori $\theta \,|\, x = 9 \sim$ Beta($\alpha = 9.5$,$\beta = 3.5$), $\mathbb{P}(\theta \leq 1/2 \,|\, x = 9) \doteq 0.039$, $\mathbb{P}(\theta > 1/2 \,|\, x = 9) \doteq 1 - 0.039 = 0.961$ e Logo $\mathcal{O}(\theta \leq 1/2 \,|\, x = 9) \doteq 0.041$. Neste caso o fator de Bayes é também $B^p_{0,1} \doteq 0.041$ (pois $\mathcal{O}(\theta \leq 1/2) = 1$).

O exemplo mostra que, embora alguns autores referem ao fator de Bayes como uma medida bayesiana "objetiva", ele efetivamente depende da especificação da distribuição a priori. □

Observações:

- O problema pode ser visto do ponto de vista decisional. A função de perda usual, devida a Neyman é

$$L(d, \theta) = \begin{cases} 0 \text{ se } d = 0,\, \theta \in \Theta_0 \\ 0 \text{ se } d = 1,\, \theta \in \Theta_1 \\ a_0 \text{ se } d = 1,\, \theta \in \Theta_0 \\ a_1 \text{ se } d = 0,\, \theta \in \Theta_1 \end{cases} .$$

- Nesse caso a decisão ótima é rejeitar H_0 ($d = 1$) se

$$\mathbb{P}(\theta \in \Theta_1 | x) > \frac{a_0}{a_0 + a_1}$$

- Precisa ter muito cuidado quando queremos testar hipóteses "precisas" ("*sharp*"), como $H_0 : \theta = \theta_0$ ou $H_0 : \mu_y - \mu_x = d_0$. Em geral, se a probabilidade a priori de uma hipótese é nula, também será nula a probabilidade a posteriori. Assim, para testar hipóteses precisas é necessário especificar prioris com massa positiva nessa hipótese. Em particular, se a distribuição a priori de θ for continua, qualquer hipótese precisa terá probabilidade a posteriori nula.

7.1 O Lema de Neyman-Pearson

Considere um teste da hipóteses simples $H_0 : \theta = \theta_0$ contra a alternativa também simples $H_a : \theta = \theta_1$. O Lema de Neyman-Pearson afirma que se a região crítica C é tal que existe um número real k para o qual

- $p(x \,|\, \theta_1) > k\, p(x \,|\, \theta_0)$ implica que $x \in C$ e

- $p(x \,|\, \theta_1) < k\, p(x \,|\, \theta_0)$ implica que $x \notin C$;

então para qualquer outra região crítica C_* tal que $\mathbb{P}(x \in C_* \,|\, \theta_0) \leq \mathbb{P}(x \in C \,|\, \theta_0)$, devemos ter que $\mathbb{P}(x \notin C_* \,|\, \theta_1) \geq \mathbb{P}(x \notin C \,|\, \theta_1)$. Em outras palavras, se denotamos por $\alpha = \mathbb{P}(x \in C \,|\, \theta_0)$ e $\alpha_* = \mathbb{P}(x \in C_* \,|\, \theta_0)$ e por $\beta = \mathbb{P}(x \notin C \,|\, \theta_1)$ e $\beta_* = \mathbb{P}(x \notin C_* \,|\, \theta_1)$ as respectivos probabilidades dos *erros tipo I* e *tipo II*, para qualquer região crítica C_* cujo probabilidade do erro tipo I é menor ou igual do que a da RC C, devemos ter que a probabilidade do erro tipo II para C_* é maior ou igual que para C.

É importante salientar que o Lemma de N-P não específica como os valores de α e β devem ser escolhidos, embora na prática seja usual escolher (fixar) um valor para o nível de significância α e trabalhar então com a região crítica C do Lema de N-P, qualquer que seja o seu valor de β.

Quando consideramos somente testes da forma de Neyman-Pearson, usualmente existe uma relação um-a-um entre os valores de α, β e k, de forma que dado quaisquer um deles, os outros dois ficam determinados.

Exemplo 7. Suponha uma amostra $X_1, \ldots, X_n \,|\, \mu \sim \text{Normal}(\mu, \sigma^2 = 1)$ e o teste de $H_0 : \mu = 0$ contra $H_1 : \mu = 1$. Um teste de N-P rejeita H_0 quando a razão de verossimilhança

$$\frac{p(x_1, \ldots, x_n \,|\, \mu = 1)}{p(x_1, \ldots, x_n \,|\, \mu = 0)} = \frac{\exp\{-\sum_{i=1}^n (x_i - 1)^2/2\}}{\exp\{-\sum_{i=1}^n x_i^2/2\}} = \exp\{n\,(\bar{x} - 1/2)\} > k$$

ou, equivalentemente, quando

$$\bar{x} > k'$$

onde $k' = \frac{1}{2} + \frac{1}{n} \log k$. Dado k (ou k') podemos calcular

$$\alpha = \mathbb{P}(\bar{X} > k' \,|\, \mu = 0) = \mathbb{P}(\sqrt{n}\,\bar{X} > \sqrt{n}\,k' \,|\, \mu = 0) = 1 - \Phi(\sqrt{n}\,k')\,,$$

onde $\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-1/2}\, e^{-u^2/2}\, du$ é a função de distribuição acumulada da Distribuição Normal Padrão, e

$$\beta = \mathbb{P}(\bar{X} \leq k' \,|\, \mu = 1) = \mathbb{P}(\sqrt{n}\,[\bar{X} - 1] \leq \sqrt{n}\,[k' - 1] \,|\, \mu = 1) = \Phi(\sqrt{n}\,[k' - 1])\,.$$

Analogamente, dado $\alpha = \mathbb{P}(\bar{X} > k' \,|\, \mu = 0)$ podemos achar $k' = (1/\sqrt{n})\, \Phi^{-1}(1-\alpha)$ e daí temos os valores de k e de β usando as equações anteriores. □

	$\alpha = 0.05$ fixo			$k = 2$ fixo		
n	β	k	k'	k'	α	β
5	0.277	3.2	0.736	0.638	0.077	0.210
10	0.065	1.2	0.520	0.569	0.036	0.087
20	0.002	0.07	0.368	0.535	0.008	0.019
30	6×10^{-05}	2×10^{-03}	0.300	0.523	0.002	0.0044
50	2×10^{-08}	2×10^{-06}	0.233	0.514	1×10^{-04}	3×10^{-04}
100	3×10^{-17}	3×10^{-15}	0.164	0.507	2×10^{-07}	4×10^{-07}

Tabela 1: Valores de β, k e k' para $\alpha = 0.05$ fixo e de α, β e k' para $k = 2$ fixo (Lema de N-P, exemplo 7). Veja como quando $\alpha = 0.05$ os valores de β são muito pequenos para n grandes, enquanto quando $k = 2$ fixo, os valores de α e β são comensuráveis quando n cresce.

7.2 Versão Bayesiana do Lema de N-P

Defina as probabilidades a priori $p_0 = \mathbb{P}(\theta = \theta_0)$ e $p_1 = \mathbb{P}(\theta = \theta_1)$. Logo as probabilidades a posteriori são

$$\mathbb{P}(\theta = \theta_i \,|\, x) \propto \mathbb{P}(\theta = \theta_i)\, p(x \,|\, \theta = \theta_i) \propto p_i\, p(x \,|\, \theta = \theta_i)\,,$$

isto é,

$$\mathbb{P}(\theta = \theta_i \,|\, x) = \frac{p_i\, p(x \,|\, \theta = \theta_i)}{p_0\, p(x \,|\, \theta = \theta_0) + p_1\, p(x \,|\, \theta = \theta_1)}$$

para $i = 0, 1$.

Como vimos ao longo desta unidade, os problemas de decisão requerem a introdução de funções de perda. Denote por $d = 1$ a decisão de rejeitar H_0 e por $d = 0$ a decisão de não rejeitar H_0. A perda de Neyman neste caso tem a forma

- $L(d = 1, \theta = \theta_0) = a_0$;
- $L(d = 0, \theta = \theta_1) = a_1$ e
- $L(d = 1, \theta = \theta_1) = L(d = 0, \theta = \theta_0) = 0$.

Agora as perdas esperadas a posteriori são

$$\begin{aligned}\mathbb{E}[L(d = 1, \theta) \,|\, x] &= \mathbb{P}(\theta = \theta_0 \,|\, x)\, L(d = 1, \theta = \theta_0) + \mathbb{P}(\theta = \theta_1 \,|\, x)\, L(d = 1, \theta = \theta_1) \\ &= \frac{p_0\, p(x \,|\, \theta = \theta_0)}{p_0\, p(x \,|\, \theta = \theta_0) + p_1\, p(x \,|\, \theta = \theta_1)}\, a_0 \quad (13)\end{aligned}$$

e

$$\begin{aligned}\mathbb{E}[L(d = 0, \theta) \,|\, x] &= \mathbb{P}(\theta = \theta_0 \,|\, x)\, L(d = 0, \theta = \theta_0) + \mathbb{P}(\theta = \theta_1 \,|\, x)\, L(d = 0, \theta = \theta_1) \\ &= \frac{p_1\, p(x \,|\, \theta = \theta_1)}{p_0\, p(x \,|\, \theta = \theta_0) + p_1\, p(x \,|\, \theta = \theta_1)}\, a_1\,. \quad (14)\end{aligned}$$

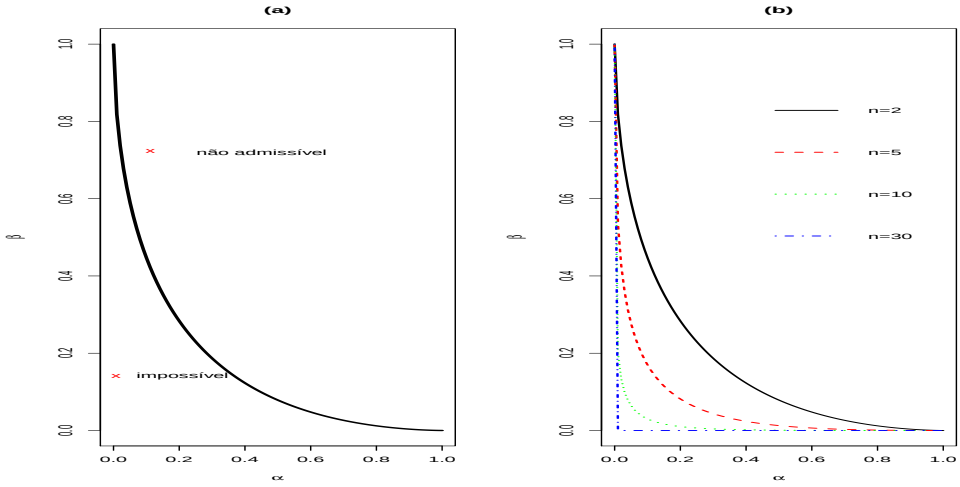

Figura 5: Valores de $\alpha = \mathbb{P}(\bar{X} > k' \,|\, \mu = 0)$ e $\beta = \mathbb{P}(\bar{X} \leq k' \,|\, \mu = 1)$ admissíveis (isto é, dos testes de Neyman-Pearson) no exemplo **??**. A curva da figura (a) mostra as combinações de valores de α e β admissíveis para o caso $n = 2$. Pares de valores (α, β) *acima* da curva são *não admissíveis*, no sentido que eles são dominados por algum teste na curva; pares de valores (α, β) *debaixo* da curva são impossíveis, no sentido que não existe nenhuma região crítica que possa atingir eles. As curvas da figura (b) mostra as curvas dos pares (α, β) para vários tamanhos amostrais.

Finalmente, a decisão $d = 1$ (rejeitar H_0) será preferida quando a sua perda esperada [equação (13)] for menor do que a da decisão $d = 0$ [equação (14)], ou equivalentemente, quando

$$\frac{p(x \,|\, \theta = \theta_1)}{p(x \,|\, \theta = \theta_0)} > \frac{p_0\, a_0}{p_1\, a_1}\,.$$

O resultado então é efetivamente um teste de N-P com $k = (p_0/p_1)\,(a_0/a_1)$, isto é, o produto das chances relativas de H_0 a priori vezes a razão dos custos de cometer um Erro Tipo I e um Erro Tipo II.