Estatística Bayesiana com Aplicações em Dados Categóricos e de Sobrevivência

Carlos Alberto de Bragança Pereira

Departamento de Estatística, Universidade de São Paulo

Adriano Polpo

Departamento de Estatística, Universidade Federal de São Carlos

Campina Grande, 22 de julho de 2013.

Prefácio

A abordagem Bayesiana, na estatística, tem cada vez mais adeptos e tem despertado o interesse de cientistas de áreas como Biologia, Medicina, Economia, Engenharia, Direito, entre outras. Neste livro, apresentamos os fundamentos básicos do paradigma Bayesiano e algumas ferramentas necessárias para seu amplo uso na estatística. Discutimos a formulação Bayesiana dos problemas estatísticos, a definição de modelo estatístico, o subjetivismo necessário e presente em qualquer pesquisa científica e a visão prática de estatísticos Bayesianos quando enfrentam os desafios da trindade estatística: estimação pontual, estimação intervalar e testes de hipótese. De forma genérica, técnicas para solução de problemas inerentes à trindade serão apresentadas. O foco principal de nossas aplicações estará na análise de dados categóricos e de sobrevivência. Os temas discutidos serão motivados por exemplos práticos que fizeram parte dos trabalhos publicados pelos autores. Todos os problemas discutidos têm suas soluções apresentadas, analisadas e implementadas no software `R`.

Sumário

Capítulo 1

Introdução a inferência Bayesiana

O trabalho principal do estatístico é o de fazer avaliações sobre quantidades ou estados da natureza que são invisíveis para o próprio estatístico e/ou para toda uma comunidade de interessados. Evidentemente, quanto menos incerteza se tem sobre o invisível, melhor pode ser a avaliação. Por melhor, entendemos afirmações mais precisas com relação às alternativas possíveis das quantidades ou estados da natureza de interesse.

Note o leitor que normalmente, para o trabalho do estatístico, procuramos qualificativos como melhor, mais preciso ou mesmo menos incerto, mesmo não se tendo claro o significado destes qualificativos. Com esta preocupação de linguagem, nestas notas consideramos, como em Basu (1975) e DasGupta (2011), que a palavra chave em estatística é informação. Informação sobre o invisível, quantidades ou estados da natureza representados por θ. Informação é o que ela provoca: mudança de opinião. Se ao observar uma nova situação, sua opinião ou avaliação sobre θ não se alterou, consideramos que esta observação, em relação à θ, é não informativa pelo menos para você. Por outro lado, outro individuo com diferente background, na mesma situação observacional, pode muito bem mudar de opinião e assim considerar que houve informação e que foi incorporada, pelo menos para este indivíduo. Eis aqui nossa posição subjetivista sobre probabilidade e estatística: diferentes estatísticos podem produzir diferentes avaliações mesmo em presença da mesma situação observacional. Este aspecto subjetivista ficará claro em nosso discurso ao longo destas notas.

Deve ficar claro para o leitor que o trabalho do estatístico perde o valor mágico da inferência quando as quantidades ou estados da natureza, θ, são totalmente conhecidos. É claro que a arte da estatística descritiva para a ilustração de situações como a de um censo ou mesmo de uma série temporal passada é bem apreciada.

Contudo este tipo de trabalho descritivo, não inferencial, não é foco destas notas.

Na nossa visão Bayesiana da estatística, e obviamente da probabilidade como disciplina, possíveis e distintos graus de incerteza sobre θ são descritos ou avaliados por distintas afirmações ou modelos probabilísticos relativos às possíveis alternativas que θ pode assumir. Desta forma, pensamos que o trabalho de um estatístico consiste nas seguintes etapas: (*i*) identificar quantidades observáveis, digamos X, Y, etc., que, na opinião do estatístico, relacionam-se com θ; (*ii*) construir modelos probabilísticos, respeitando as leis de probabilidade, que descrevam opiniões e incertezas relativas à θ, incluindo o tipo de possíveis relações (na opinião do analista) entre as observáveis e θ; e (*iii*) desenvolver métodos inferenciais, relativos à θ, baseados nestes modelos e nas observações efetivas das quantidades observáveis. O objetivo principal do uso dessas observações é a diminuição da incerteza – ou aumento de informação – que o estatístico carrega sobre θ. Neste sentido, todas as nossas afirmações probabilísticas são fruto de nossos julgamentos!

Em nossa idealização de probabilidade, deveria existir um instrumento que escolhesse aquele modelo probabilístico, dentre os alternativos possíveis, que capturasse toda a incerteza do estatístico (ou do cientista) sobre o desconhecido "estado da natureza" ou quantidade de interesse θ. A falta deste instrumento objetivo transforma o estatístico no responsável por encontrar um modelo probabilístico que melhor captura e represente sua incerteza ou informação sobre θ. Não consideramos a definição deste modelo uma tarefa fácil. Assim, além de ter uma boa formação em estatística, o estatístico precisa se envolver, dedicadamente, com a área do problema científico que exige inferências sobre θ. Lembremos que o conceito de informação aqui considerado é o operacional, no qual informação é o que ela produz: a mudança de opinião. Em Stern and Pereira (2012) o leitor irá encontrar as várias implicações deste conceito. Contudo, todas as definições de informação apresentadas na literatura, têm em comum o aspecto operacional da mudança de opinião como objetivo principal do seu uso. Operacional é também o objetivo principal deste livro, que terá um foco nas aplicações.

Na sequência, o leitor deve entender que o termo quantidade (variável) aleatória será usado para qualquer quantidade de interesse (número, vetor etc.) invisível, isto é, de valor desconhecido. Quando o invisível for não observável, será denominado parâmetro e, no caso dos observáveis, simplesmente quantidades (variáveis ou vetores etc.) aleatórias. Por exemplo, pensando na qualidade das peças produzidas por uma fábrica, a proporção de peças defeituosas que a fábrica irá produzir até o final de um determinado mês é um parâmetro, π, com relação ao final da primeira semana do mês. Por outro lado, o número de peças defeituosas, X, nesta primeira semana é uma variável aleatória cujo valor, x, será conhecido no momento da predição de π. Claro que observaremos X por estar disponível e por pensarmos que deve

estar associado ao valor desconhecido de π, no momento da predição. Note que, ao final do mês, o valor de π passa de desconhecido para conhecido, permitindo uma avaliação da predição feita na primeira semana do mês.

Ressaltamos o fato de observarmos quantidades, que de desconhecidas passaram a ser conhecidas, que pensamos serem informativas. A todo mecanismo que transforma quantidades desconhecidas em conhecidas, damos o título de *experimento*. Isto é, definimos como experimento aleatório qualquer mecanismo que nos permite observar o *valor* de uma quantidade aleatória: transforma invisíveis em visíveis.

1.1 Probabilidade

Ao observar o mundo, cientistas fazem uso de quantidades ou coisas que podem ser avaliadas numericamente. Para entender o funcionamento de aspectos da vida, quase sempre fazemos medições e coletamos observações que, em conjunto, denominamos banco de dados. Evidentemente só construímos bancos de dados que acreditamos estar associados ao parâmetro θ, o qual nos interessa descrever probabilisticamente. Se nossa descrição de incerteza é probabilística, necessariamente precisamos observar e seguir rigorosamente os axiomas da probabilidade, os quais lembramos na sequência.

Como nos cursos de probabilidade, consideremos os eventos associados a um espaço amostral, constituído pelas possíveis alternativas de um experimento aleatório. Se E é um evento definido neste espaço, $\Pr(E)$ é a probabilidade de E (ser observado) e $\Pr(E \mid \theta)$ é a probabilidade condicional de E fosse θ conhecido. De fato, $\Pr(E \mid \theta)$ é uma função de dois argumentos:

i. Fixado um valor q de θ, $\Pr(\cdot \mid q)$ é uma função de probabilidades definida no espaço dos eventos, condicional a $\theta = q$; e

ii. Observado um evento e, $\Pr(e \mid \cdot)$ é a função de verossimilhança no espaço paramétrico Θ (o conjunto das alternativas possíveis de θ) associada à ocorrência de e.

A seguir os axiomas da probabilidade são apresentados no contexto da função $\Pr(E \mid \theta)$, considerando dois eventos E_1 e E_2 definidos no espaço amostral de um estudo:

AC. Convexidade:
$0 \leq \Pr(E_i \mid \theta) \leq 1$. As igualdades ocorrem nos caso do evento ser impossível, $= 0$, ou certo, $= 1$;

AA. Adição:
$\text{Pr}(E_1 \text{ ou } E_2 \mid \theta) = \text{Pr}(E_1 \mid \theta) + \text{Pr}(E_2 \mid \theta) - \text{Pr}(E_1 \text{ e } E_2 \mid \theta)$; e

AM. Multiplicação:
$\text{Pr}(E1 \text{ e } E_2|\theta) = \text{Pr}(E_1|\theta)\text{Pr}(E_2|E_1 \text{ e } \theta) = \text{Pr}(E_2|\theta)\text{Pr}(E_1|E_2 \text{ e } \theta)$.

Evidentemente, em analogia com a teoria dos conjuntos, "ou" corresponde à união de conjuntos e "e" à intersecção de conjuntos. À esquerda da barra estão as quantidades que estão sendo probabilizadas, enquanto à direita estão as condicionantes. Diz-se que dois eventos são independentes quando a ocorrência de um não altera a probabilidade da ocorrência do outro. Isto é, dois eventos E_1 e E_2 são independentes quando as seguintes igualdades equivalentes são válidas:

i. $\text{Pr}(E_2 \mid E_1 \text{ e } \theta) = \text{Pr}(E_2 \mid \theta)$ ou

ii. $\text{Pr}(E_1 \mid E_2 \text{ e } \theta) = \text{Pr}(E_1 \mid \theta)$ ou

iii. $\text{Pr}(E_1 \text{ e } E_2 \mid \theta) = \text{Pr}(E_1 \mid \theta)\text{Pr}(E_2 \mid \theta)$.

É um bom exercício para o leitor verificar a equivalência dessas três igualdades; as duas primeiras formam a definição intuitiva de independência e a última é a definição simétrica.

Propositalmente, não estamos preocupados com a precisão nem da linguagem nem da notação. Lembramos o leitor que um curso básico de probabilidades é pré-requisito para o entendimento dessas notas. Note que, se o valor de θ é fixado e conhecido, não se necessita do condicionante e as fórmulas continuam válidas incondicionalmente – apenas um caso particular.

Por simplicidade, escrevemos $p(x \mid \theta)$ ou $p(\theta \mid x)$ sempre que estiver implícito o papel de θ como parâmetro, e de X como variável observável cuja observação é representada por x. Assim, temos duas funções de θ:

i. $p(x \mid \theta) = \text{Pr}(X = x \mid \theta)$ é a probabilidade do evento $\{X = x\}$ para cada possível valor de θ e

ii. $p(\theta \mid x) = \text{Pr}(\theta \mid X = x)$ é a função de probabilidade de θ avaliada para todo θ no espaço paramétrico, Θ, condicionado ao evento observável $\{X = x\}$.

Podemos agora recordar os teoremas da probabilidade total e de Bayes. De fato, com o uso dos axiomas acima enunciados, substituímos por fórmulas a palavra teoremas. Tanto a fórmula da probabilidade total, como a de Bayes, envolve duas variáveis aleatórias, X e θ, por exemplo.

FÓRMULA DA PROBABILIDADE TOTAL:

A probabilidade do evento $\{X = x\}$ pode ser escrita como

$$p(x) = \sum_{\theta \in \Theta} p(x, \theta) = \sum_{\theta \in \Theta} p(x \mid \theta) p(\theta)$$

FÓRMULA DE BAYES:

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)}$$

Podemos interpretar o teorema da probabilidade total como sendo o cálculo da probabilidade de uma consequência, x, vislumbrando todas as causas (θ) possíveis e o teorema de Bayes como o cálculo da probabilidade de uma possível causa θ para a consequência observada x.

Como ilustração considere a seguinte situação:

1.1 Exemplo. Joe mostra a Ed duas moedas, uma com duas faces distintas, cara e coroa, e a outra com duas caras. Coloca as duas em seu bolso e então retira uma delas e lança a moeda duas vezes, resultando em duas caras. Joe desafia Ed a dizer qual das duas moedas ele lançou. Representando θ_1 e θ_2 as moedas com uma e duas caras, respectivamente, e por x_0, x_1 e x_2 os resultados possíveis de zero, uma ou duas caras, consideramos as seguintes probabilidades como modelo para Ed:

$$\Pr(\theta_1) = \Pr(\theta_2) = \frac{1}{2}; \Pr(x_0 \mid \theta_1) = \Pr(x_2 \mid \theta_1) = 1/4,$$

$$\Pr(x_1 \mid \theta_1) = 1/2 \text{ e } \Pr(x_2 \mid \theta_2) = 1.$$

Com estas probabilidades podemos obter

$$\Pr(x_2) = \Pr(x_2 \mid \theta_1)\Pr(\theta_1) + \Pr(x_2 \mid \theta_2)\Pr(\theta_2) = \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} = \frac{5}{8}.$$

Com a fórmula de Bayes obtemos $\Pr(\theta_2 \mid x_2) = \left(\frac{1}{2}\right) / \left(\frac{5}{8}\right) = \frac{4}{5}$. Assim, se realmente Ed acredita que Joe foi honesto em escolher ao acaso a moeda que foi lançada, ele deve apostar na moeda com duas caras cuja probabilidade é quatro vezes maior do que a probabilidade da moeda comum. A consequência do lançamento (visível para ambos) foi x_2 e as causas (invisível para Ed) podem ter sido tanto θ_1 como θ_2.

■

Um conjunto de afirmações probabilísticas, como acima, é *coerente* se obedece aos axiomas AC, AA e AM. O conjunto coerente de afirmações probabilísticas de uma variável aleatória é chamado de *distribuição* desta variável.

É claro que nem sempre podemos realizar o experimento para a inferência sobre θ. Em muitas situações encontramos o banco de dados já definido e então o estatístico, "transfere" para o passado o seu olhar e idealiza um modelo gerador daqueles dados. Então, o estatístico constrói suas ferramentas ligando o invisível, objeto do estudo, com as observações que estão disponíveis e que em suas conjecturas foram fruto de observáveis ideais. Vejam o seguinte exemplo:

1.2 Exemplo. Um roubo foi cometido e uma quantidade apreciável de sangue do tipo A, presumidamente do infrator, foi encontrado no local do crime, perto de uma janela cujo vidro estava quebrado. Por alguma razão, uma pessoa foi denunciada e tornou-se um suspeito em relação àquela infração. Além de apresentar uma cicatriz em seu braço, o sangue do suspeito também era do tipo A. O parâmetro, neste caso, pode ser representado por θ e, como suas alternativas, assumimos $\theta = 1$ se o suspeito é culpado e $\theta = 0$ se é inocente. Um juiz, depois de analisar o caso e as possíveis motivações que teria o suspeito, considera que, a priori, $q = 0,3$, em que $q = \text{Pr}(\theta = 1)$. O estatístico então define duas variáveis aleatórias imaginando o passado, antes de tudo acontecer:

i. $X = 1$ se o sangue do suspeito é A e $X = 0$ caso contrário; e

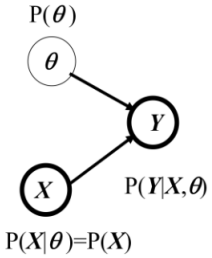
ii. $Y = 1$ se o sangue da cena do crime é do tipo A e $Y = 0$ caso contrário.

O interesse do Juiz é calcular a probabilidade de um estado invisível dado a evidência obtida, $\text{Pr}(\theta = 1 \mid X = 1; Y = 1)$. Considere então que a proporção de pessoas na população com sangue do tipo A é p, por exemplo, $p = 1/4$. Temos três variáveis aleatórias e podemos escrever a probabilidade conjunta como

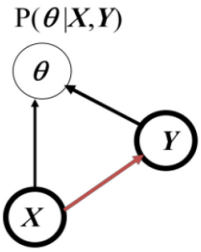
$$\text{Pr}(\theta, X, Y) = \text{Pr}(\theta)\text{Pr}(X \mid \theta)\text{Pr}(Y \mid \theta, X)$$

e, como se acredita que tipo de sangue e caráter de um indivíduo não são associados, $\text{Pr}(X \mid \theta) = \text{Pr}(X)$, simplificamos esta expressão por $\text{Pr}(\theta, X, Y) = \text{Pr}(\theta)\text{Pr}(X)\text{Pr}(Y \mid \theta, X)$. A Figura 1.3 apresenta o diagrama de influência que traduz visualmente esta modelagem. A Figura 1.4 é o diagrama de influência que representa a operação diagramática que deve ser realizada para responder ao juiz. Isto é, escrevemos a probabilidade conjunta como

$$\text{Pr}(\theta, X, Y) = \text{Pr}(X)\text{Pr}(Y \mid X)\text{Pr}(\theta \mid X, Y).$$



1.3 Figura. *Diagrama de influência para o modelo inicial.*



1.4 Figura. *Diagrama de influência final*

A operação gráfica de inversão de arcos é realizada pelo uso das duas fórmulas estabelecidas acima. Os círculos mais fortes são aqueles que representam as variáveis que são observadas, X e Y.

Usando as regras probabilísticas descritas anteriormente, podemos responder ao Juiz:

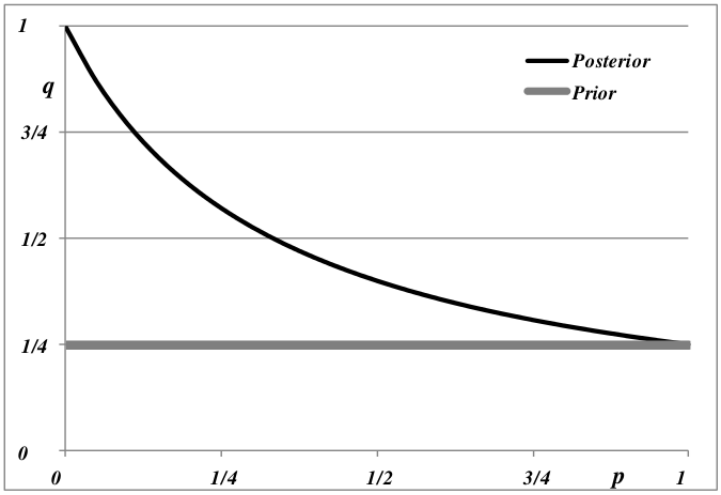
$$\begin{aligned}
&\Pr(\theta = 1 \mid X = 1; Y = 1) = \\
&= \frac{\Pr(\theta{=}1)\Pr(X{=}1)\Pr(Y{=}1 \mid X{=}1; \theta{=}1)}{\Pr(\theta{=}1)\Pr(X{=}1)\Pr(Y{=}1 \mid X{=}1; \theta{=}1) + \Pr(\theta{=}0)\Pr(X{=}1)\Pr(Y{=}1 \mid X{=}1; \theta{=}0)} \\
&= \frac{qp1}{qp1 + (1-q)p^2} = \frac{q}{q + (1-p)q} = \frac{q}{p + (1-p)q} > q.
\end{aligned}$$

Chamamos atenção para o fato de considerarmos que a chance do sangue do local do crime ser do tipo A, incondicionalmente aos outros eventos, é igual à proporção de pessoas com sangue do tipo A na população, p. No caso do suspeito não ser culpado, $\theta = 0$, X e Y não estão associados e sim independentes. Notamos também que o denominador da última fração é uma combinação convexa entre 1 e q e assim um número menor do que 1, garantindo a validade da desigualdade. O fato relevante é que se um indivíduo possui o mesmo tipo de sangue do criminoso, a chance de o indivíduo ser o criminoso aumenta e pode aumentar drasticamente se o tipo de sangue for raro. Consideramos que o juiz assumiu que $q = 1/4$ e que tivemos $X = Y = 1$. A Figura 1.5 ilustra a influência da proporção populacional de pessoas com determinado tipo sanguíneo na probabilidade (a posteriori) de um indivíduo ser culpado quando o sangue deste indivíduo é do mesmo tipo daquele encontrado na cena do crime.

É importante que o leitor entenda o caráter subjetivo das probabilidades neste exemplo. O suspeito, com probabilidade um, sabe do verdadeiro valor do estado da natureza θ. Por outro lado, o Juiz, mesmo depois de obter a evidência do mesmo tipo de sangue do local do crime e do suspeito, continua sem atingir a certeza. Assim, podemos entender a sentença de DeFinetti (1970) que afirma que *probabilidade não existe*: Probabilidade é a medida pessoal da incerteza sobre um estado da natureza

invisível e de interesse. Assim, a probabilidade de um indivíduo depende do conhecimento e do envolvimento que este tem sobre a área de conhecimento relacionada ao estado de natureza de interesse. Um juiz e os advogados provavelmente assumiriam probabilidades diferentes entre eles e entre um leigo e/ou indivíduos da família do réu ou suspeito.

■



1.5 Figura. *Influência da proporção populacional de um tipo sanguíneo na passagem priores/posteriores.*

O exemplo a seguir chama atenção para a forma sequencial que a metodologia Bayesiana pode seguir. A probabilidade posterior de θ, após a observação x de X, pode ser usada como a priori (antes da observação y de Y) para uma nova calibração da probabilidade de θ. Este processo pode seguir sequencialmente até a evidência mais recente ser usada para o ajuste da probabilidade de θ.

Por simplicidade, se θ é dicotômico (por exemplo, $\theta = 0$ e $\theta = 1$) a fórmula de Bayes pode ser escrita como $(1 + rR)^{-1}$ para r sendo a razão de probabilidades a priori (prior odds) e R a razão de Bayes. Isto é,

$$r = \frac{\text{Pr}(\theta = 0)}{\text{Pr}(\theta = 1)} \text{ e } R = \frac{\text{Pr}(X = x \mid \theta = 0)}{\text{Pr}(X = x \mid \theta = 1)},$$

$$p_1(x) = \text{Pr}(\theta = 1 \mid X = x) =$$
$$= \frac{\text{Pr}(X = x \mid \theta = 1)\text{Pr}(\theta = 1)}{\text{Pr}(X = x \mid \theta = 1)\text{Pr}(\theta = 1) + \text{Pr}(X = x \mid \theta = 0)\text{Pr}(\theta = 0)} = \frac{1}{1 + rR(x)}.$$

1.6 Exemplo. Antônio, um jovem filho de Maria, decide seguir o conselho de sua mãe e entra na justiça para que John, um rico empresário, reconheça a paternidade de Antônio. O Juiz então decide que, tanto o demandado, John, quanto os demandantes, Antônio e Maria, se submetam ao exame de DNA em amostras de seus respectivos sangues. A Tabela 1.7 mostra os genótipos em cada um dos três locos estudados. Os dados são provenientes de análise de material genético sob a técnica de Microsatélites pela Reação de Cadeia da Polimerase (PCR).

Por frequência alélica, f_i, entendemos a proporção em que o alelo i aparece na população de alelos da população: $2N$ é o total de alelos na população de tamanho N. A hipótese de nosso problema é a de que John é o pai biológico ($\theta = 1$) de Antônio, que sabemos ser filho de Maria. A hipótese alternativa ($\theta = 0$) é, logicamente, a que John não é o pai biológico de Antônio. A razão de verossimilhanças é a razão entre a probabilidade de observarmos o genótipo de Antônio quando John não é o pai, dividido pela probabilidade do genótipo sob a condição de que John seja o pai. Tomando as frequências alélicas do banco de dados como probabilidades dos alelos, as seguintes razões para cada loco são obtidas:

i. Loco 1 seria $1/4$ no denominador e $(1/2)f_{16}$ no numerador. Assim, $R_1 = 2f_{16}$;

ii. Para o Loco 2, continuamos a ter $1/4$ no denominador, caso de John ser o pai, e $(1/2)(g_{29} + g_{33})$ no numerador. Então, $R_2 = 2(g_{29} + g_{33})$.

iii. Respectivamente, para o loco 3, teríamos agora $1/2$ e $(1/2)(h_{19} + h_{20})$ no denominador e no numerador. Portanto, $R_3 = (h_{19} + h_{20})$.

1.7 Tabela. *Genótipos de três locos observados: Demandantes e Demandado.*

	Genótipos e Alelos			Frequência
Loco	Maria	Antônio	John	Alélica
L1	11 | 13	11 | 16	12 | 16	f_{11}; f_{12}; f_{13}; f_{16}
L2	29 | 33	29 | 33	29 | 35	g_{29}; g_{33}; g_{35}
L3	19 | 20	19 | 20	19 | 19	h_{19}; h_{20}

Suponha que q_1, q_2 e q_3 sejam as probabilidades posteriores de John ser o pai, depois de analisados os Locos 1, 2 e 3, com q_0 sendo a probabilidade inicial de John ser o pai antes do uso do Loco 1. Analogamente, as razões de chances (ou odds) são:

$$r_0 = (1 - q_0)/q_0;\ r_1 = (1 - q_1)/q_1;\ r_2 = (1 - q_2)/q_2;\ r_3 = (1 - q_3)/q_3.$$

Usando, sequencialmente, apenas a fórmula de Bayes, obtemos as seguintes probabilidades posteriores:

$$q_1 = (1 + r_0R_1)^{-1},\ q2 = (1 + r_1R_2)^{-1} \text{ e } q_3 = (1 + r_2R_3)^{-1}.$$

Como nossos juízes, em geral, não permitem o uso de opiniões informativas a priori, usamos a probabilidade $q_0 = 1/2$ no início.

Aqui são usadas as seguintes frequências relativas:

$$f_{16} = 0,05;\ g_{29} = 0,10;\ g_{33} = 0,04;\ h_{19} = 0,15 \text{ e } h_{20} = 0,08.$$

Estas frequências são provenientes de um grande banco de dados coletados por um laboratório com amostras de sangue analisadas. A consequência desses números é a seguinte sequência de probabilidades

$$p_0 = 0,50;\ p_1 = 0,9756;\ p_2 = 0,9931 \text{ e } p_3 = 0,9984.$$

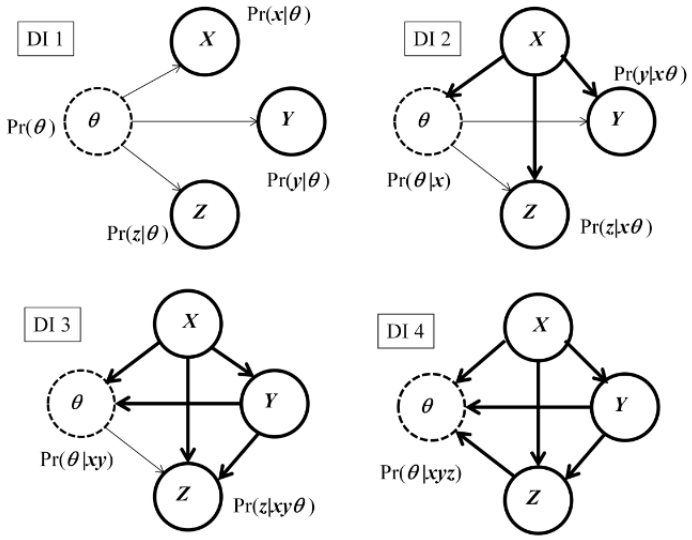
Notamos que, mesmo começando com chances iguais para John ser ou não o pai, após o uso de três Locos analisados, fica praticamente evidente (com $99,84\%$ de chances) que John seja o pai.

A Figura 1.8 ilustra o trabalho sequencial por meio de diagramas de influência em cada etapa de calibração de probabilidades. Sendo X, Y e Z as variáveis que representam as observações dos três locos estudados.

Novamente chamamos atenção para o fato de a decisão final sobre a paternidade ficar por conta do Juiz do caso. Há alguns deles que só considerariam paternidade no caso de não haver alguma chance de o demandado não ser o pai: $\theta = 1$ ser um evento certo. Contudo, a maioria dos juízes iria dar a causa em favor dos demandantes quando a probabilidade é alta, como no exemplo acima. Novamente, nosso papel como estatístico é o técnico e não o decisório.

■

O objetivo até este ponto foi o de mostrar o caráter indutivo e subjetivista do nosso trabalho como estatístico. Usamos apenas estados da natureza ou experimentos dicotômicos. A partir deste ponto passamos a usar a ferramenta padrão usada nos cursos básicos de estatística. Por espaço amostral entendemos o conjunto das alternativas que um experimento pode assumir e, por espaço paramétrico, as alternativas que pensamos ter um estado da natureza ou parâmetro θ. Ambos os espaços podem ser discretos ou contínuos e limitados ou ilimitados. O trabalho irá se restringir aos casos de modelos dominados e, dessa forma, às variáveis aleatórias estão associadas funções de probabilidade (no caso discreto) ou funções de densidade de probabilidade (no caso contínuo).



1.8 Figura. *Diagramas de Influência ilustrando as etapas do sequenciamento da calibração da probabilidade do estado da natureza θ.*

Lembramos aos leitores que fazer uso da fórmula (ou operador) de Bayes no caso de variáveis e parâmetros discretos é natural e simples. Contudo, quando temos variáveis contínuas, é necessário um estudo mais profundo dos fundamentos. Isto porque todo ponto do espaço amostral tem probabilidade zero de ocorrer e então a fórmula de Bayes não pode ser aplicada como foi descrita acima. Por sorte, podemos provar, usando a teoria da medida, que a densidade condicional da forma que é usada é sim uma esperança condicional na linguagem mais teórica de probabilidades. Assim, se X e Y são duas variáveis aleatórias contínuas com densidade conjunta $f(x, y)$ e densidades marginais $f(x)$ e $f(y)$, para quaisquer valores possíveis (x, y) de (X, Y); a densidade condicional de X dado Y e as fórmulas da probabilidade total e de Bayes recebem a seguinte versão para continuidade:

Densidade Condicional:

$$f(x \mid y) = \frac{f(x, y)}{f(y)};$$

Probabilidade Total:

$$f(x) = \int f(x, y)\mathrm{d}y; \text{ e}$$

Fórmula de Bayes:

$$f(y \mid x) = \frac{f(x \mid y)f(y)}{f(x)} = \frac{f(x \mid y)f(y)}{\int f(x, y)\mathrm{d}y}.$$

As integrais são calculadas sobre o espaço amostral de Y.

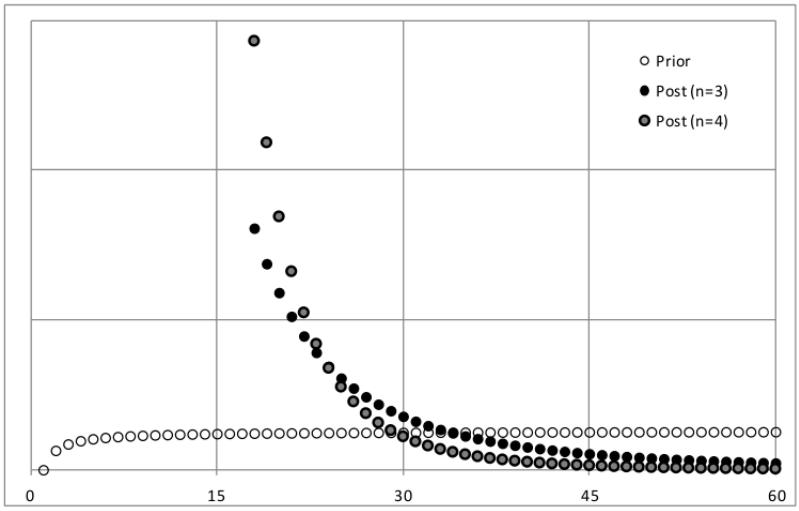
O leitor deve procurar entender bem essas fórmulas estudando um pouco da teoria da medida, de forma a se sentir confortável ao usá-las. Muitas vezes uma das variáveis é contínua e a outra discreta. Contudo se as duas forem dominadas, não temos algum problema teórico com o uso destas fórmulas.

Procurando a simplicidade, o leitor deve entender que as constantes envolvidas nas funções de densidade e de verossimilhança podem ser desconsideradas no início dos cálculos. Uma normalização ao final define a função de densidade a posteriori. Por exemplo, note que $f(y \mid x) \propto f(x \mid y)f(y)$. O símbolo indica proporcionalidade.

O Exemplo abaixo ilustra um caso onde, tanto a observável X, quanto o parâmetro θ, são variáveis aleatórias discretas. No Exemplo 1.11 ambas são contínuas, enquanto no Exemplo 1.12, a observável é discreta e o parâmetro é contínuo.

1.9 Exemplo. Um indivíduo deve estimar o número de bolas θ de uma urna e, para isso, pode selecionar ao acaso e com reposição, três bolas desta urna. As bolas são idênticas e numeradas (sem repetição) de 1 a θ. As bolas sorteadas foram as de números 12, 9 e 18. A informação disponível foi a de que o formador da urna não devia colocar mais de 60 bolas na urna e que podia dificultar o trabalho de estimação que o estatístico devia realizar. O estatístico por sua vez entendeu que a dificuldade aumentaria se o número de bolas fosse grande. Contudo, sabia também, que as bolinhas do jogo são caras. Usou assim uma priori que não privilegiava apenas valores altos de θ. Considerou uma função de probabilidade a priori proporcional a $1-\theta^{-1}$. Vamos lembrar que, a função de verossimilhança – a função de probabilidade avaliada na amostra observada, como função de θ – neste caso, é $I \times \theta^{-3}$, com a função de conjuntos I sendo a função indicadora do conjunto $\{18 \leq \theta \leq 60\}$; o máximo da amostra é menor ou igual a θ. Usando a fórmula de Bayes, podemos concluir que a função de probabilidade posterior é proporcional a $I \times (\theta-1)\theta^{-4}$. A Figura 1.10 ilustra as funções de probabilidade a priori e a posteriori, incluindo a posteriori se tivéssemos tomado uma amostra de tamanho quatro com o mesmo máximo 18. As médias obtidas com estas três funções de probabilidade, para θ, foram: 32 para a priori; $27,27$ para a posteriori sob $n=3$; e $22,95$ no caso de $n=4$. Fica claro, neste caso, que a informação relevante da amostra é proveniente do seu tamanho e do seu máximo.

■



1.10 Figura. *Priori e Posteriores para o número θ de bolas de uma urna quando o máximo da amostra é* 18 *e com amostras de tamanho* $n = 3$ *e* $n = 4$.

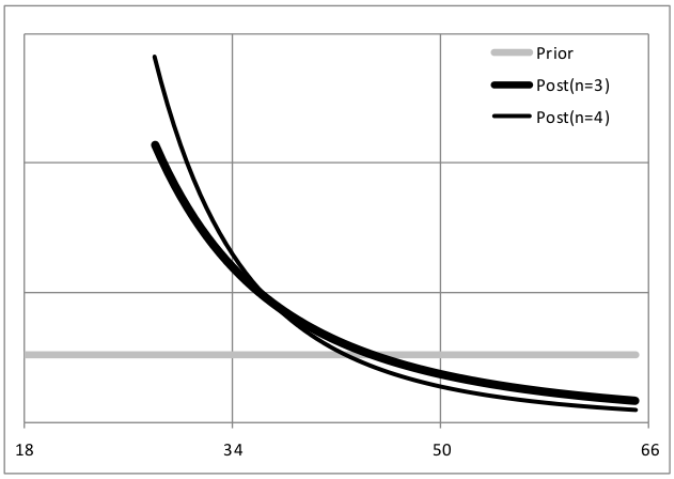
1.11 Exemplo. Imaginemos que um perito precisa decidir se uma empresa restringe de alguma forma a idade de seus empregados. Para isso, ele selecionou ao acaso 3 fichas de funcionários demitidos, anotando a idade na época da demissão. Os dados foram $x = 20$, $y = 24$ e $z = 28$. O perito, em sua primeira avaliação, considerou que a amostra é proveniente de uma distribuição uniforme no intervalo $[18; \theta]$ em que θ será no máximo 65 devido ao sistema de aposentadoria usado na classe de trabalhadores. Considerou então uma priori uniforme no intervalo $[18; 65]$. A função de verossimilhança é $I \times \theta^{-3}$, com a função de conjuntos I sendo a função indicadora do conjunto $\{28 \leq \theta \leq 60\}$; o máximo da amostra é menor ou igual a θ. Ao observar que o máximo foi 28 decidiu observar uma ficha adicional, verificando outra demissão ocorrida antes do funcionário completar 28 anos. A verossimilhança após essa observação seria então $I \times \theta^{-4}$. Por considerarmos a priori a distribuição uniforme, as duas verossimilhanças normalizadas produziriam as posteriores, ilustradas na Figura 1.13. As médias da priori e das posteriores com amostras de tamanho 3 e 4 seriam $41,50$; $38,94$; e $36,94$. Neste caso também o máximo e o tamanho da amostra são os únicos valores que influenciam o cálculo das posteriores.

■

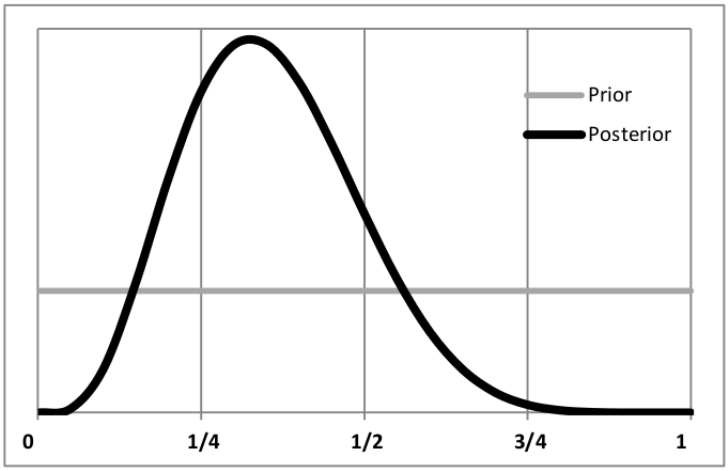
1.12 Exemplo. Em uma fábrica de peças, unidades são selecionadas de uma linha de produção com o intuito de estudar a proporção π de peças imperfeitas. Foram selecionadas 10 peças, das quais se observou três imperfeitas. A função de probabilidade amostral é a binomial com parâmetros $n = 10$ e π desconhecido e de interesse. A função de verossimilhança normalizada pela sua integral em $[0; 1]$ é

uma densidade beta com parâmetros $a = 4$ e $b = 8$. Na verdade esta função normalizada é a densidade a posteriori, se a priori fosse uma uniforme em $[0; 1]$, a Figura 1.14 mostra tanto a priori (média $1/2$) quanto a posteriori (média $1/3$).

■



1.13 Figura. *Priori e Posteriores para θ no caso das fixas de empregados demitidos. A máxima idade encontrada foi 28 anos nas amostras de tamanho $n = 3$ e $n = 4$.*



1.14 Figura. *Priori e Posterior para π, a proporção de peças imperfeitas que uma empresa produz, no caso de uma amostra de tamanho 10 contendo 3 imperfeitas.*

Os exemplos apresentados ilustram bem a forma probabilística de descrever as preferências relativas de cada valor possível do parâmetro. Nas próximas seções

mostramos como usar distribuições para os três tipos de padrões de inferência: estimação pontual, estimação intervalar e testes de significância e/ou testes de hipótese. A próxima seção é voltada aos pormenores da função de verossimilhança e de distribuições a priori especiais e úteis. Serão apresentadas distribuições que estão *conjugadas* com as distribuições amostrais.

1.2 Verossimilhanças e Classes Conjugas

Nas seções anteriores já utilizamos o termo verossimilhança por acreditar que o leitor já está acostumado com a linguagem estatística dos cursos básicos. No entanto, é nossa opinião que esta função é o recipiente de todas as informações experimentais que estão disponíveis para o trabalho estatístico. Por informação experimental entendemos aquela proveniente do experimento utilizado para a calibração das informações culturais que o cientista carrega. A descrição da incerteza sobre θ, feita pela sua distribuição de probabilidade, é fruto do conhecimento e de experimentos realizados (ou não) anteriormente. O processo de calibração da incerteza ou da informação é dinâmico: a cada etapa provoca (ou não) modificações nas distribuições de probabilidade de θ, o alvo de nossas indagações.

Ao definirmos o que será observado (X, por exemplo), definimos a distribuição amostral, que na verdade é um conjunto de distribuições de probabilidade de X, indexada por θ. Após a observação de x de X, para cada valor de θ, temos o valor da probabilidade ou da densidade avaliada em x (agora fixado). Isto é, temos uma função apenas de θ, a qual não é, em geral, função de densidade ou de probabilidade de θ. Nos exemplos apresentados nas seções anteriores isto fica bem claro.

Note que, ao usarmos o operador de Bayes para o cálculo da distribuição posteriori de θ, todas as constantes que podem aparecer na função de probabilidade ou de densidade amostral (de x) se tornam irrelevantes. Este fato nos permite estabelecer que, pontos amostrais com verossimilhanças proporcionais, necessariamente, devem produzir verossimilhanças proporcionais e assim posteriores iguais. Dessa forma, o espaço amostral inicial também fica irrelevante após a observação da amostra, x. Esta afirmação é relacionada ao Princípio da Verossimilhança que pode ser reescrito como:

Princípio da Verossimilhança: Funções de verossimilhança proporcionais devem produzir necessariamente a mesma inferência sobre θ sob a mesma distribuição a priori.

Da forma como foi definido o princípio, concluímos que não estamos restritos ao mesmo espaço amostral. Um exemplo clássico é o de uma amostra de um

processo de Bernoulli (variáveis binárias permutáveis) com parâmetro π, onde se observou 4 sucessos e 8 fracassos. O uso da operação Bayesiana só exige o fato da verossimilhança ser proporcional a

$$L(\pi) = \pi^4(1-\pi)^8.$$

Notamos aqui que esta seria a mesma nas seguintes situações: (*i*) Binomial; no início a observadora fixou $n = (x+y)$; (*ii*) Binomial Negativa; no início a observadora fixou x; e (*iii*) Indeterminado; a observadora parou de anotar quando foi chamada para almoçar. Somente em casos raros o espaço amostral influencia a operação de Bayes priori/posteriori. De fato, além da definição do modelo estatístico, a verossimilhança é o elemento que faz a ligação entre o visível e o invisível. Se o leitor lembrar-se do conceito de estatística suficiente, poderá entender facilmente que a função de verossimilhança é uma estatística suficiente mínima. Para cada observação possível podemos desenhar a verossimilhança correspondente e esta relação é a estatística suficiente mínima.

Outro conceito de muita importância para os estatísticos é o de classe conjugada de distribuições. Uma classe de distribuições $\mathcal{C}$ é conjugada ao modelo $P_\theta(X)$ se, ao tomarmos as observações x e um elemento dessa classe como priori, então a posteriori resultante também pertence a $\mathcal{C}$. Contudo, não é sempre que existe tal classe operacional. Outro problema é o fato desta definição de conjugada ser muito imprecisa. Notem que, se tomarmos uma classe de apenas uma distribuição de probabilidade degenerada em um valor de θ, digamos q, a posteriori continua sendo a mesma, independente das observações. No caso de considerarmos a classe de todas as distribuições, esta também seria conjugada. A definição, que em nossa opinião é a mais adequada operacionalmente, é a seguinte:

Classe Conjugada Natural: $\mathcal{CN}$ é uma classe conjugada natural de distribuições para um modelo estatístico $P_\theta(\cdot)$ se para cada elemento de $\pi(\theta)$ de $\mathcal{CN}$ existir um possível resultado experimental y tal que sua verossimilhança normalizada $v_y(\theta)$ seja igual a $\pi(\theta)$.

O leitor irá notar que classes conjugadas naturais podem restringir um pouco as nossas escolhas de priori. Uma classe rica em opiniões deve possuir representantes para a maioria das possíveis opiniões de usuários. No caso dos processos de Bernoulli, por exemplo, a classe das betas teria de ser restrita aos parâmetros pertencentes aos números inteiros. Notem também que esta classe obedece ao que denominamos processo sequencial do método Bayesiano: A distribuição a posteriori de hoje é a priori de amanhã, sendo assim um método sequencial de aprendizado.

Como exemplos de classes conjugadas temos as derivadas das distribuições padrões: Betas para as distribuições amostrais oriundas de processos de Bernoulli;

gamas para as distribuições de Poisson e exponencial; Normais para normal com variância fixa; gamas para normais com média fixada e outras situações mais complexas.

O seguinte exemplo mostra que há casos em que a modificação do sentido de conjugação necessita de alteração.

1.15 Exemplo. Ao comprar um estoque de lâmpadas de uma fábrica, um empresário pode ir para a linha de produção e testar peças que estão sendo produzidas. Pode assim ter uma ideia da qualidade do produto produzido estudando a proporção π de peças defeituosas da fábrica por meio de uma amostra (x, y) observada. Se sua priori fosse uma uniforme teria ao final da amostra uma beta com parâmetros $(x + 1, y + 1)$. Por outro lado, pode estar interessado apenas no lote que está comprando com 100 peças. O parâmetro de interesse neste caso seria o total de defeituosas na amostra.

Consideremos os seguintes fatos: na sua observação da linha de produção, o empresário não encontrou alguma peça defeituosa nas quatro peças inspecionadas e assim sua posteriori ficou sendo Beta$(1; 5)$ que agora serve para definir a priori para o parâmetro τ, o total de defeituosas do lote comprado.

Representemos o lote por $(U_1, \ldots, U_{30})$, com $U_i = 1$ se a peça i for defeituosa e $U_i = 0$ se perfeita. Uma amostra de tamanho cinco deste lote foi selecionada e testada. Sem perda de generalidade, representamos a amostra pelas cinco primeiras peças $(U_1, \ldots, U_5)$ e por X o total da amostra.

As premissas usadas aqui são: a de que o processo de fabricação está sobre controle e assim segue um processo de Bernoulli com parâmetro π desconhecido. Isto quer dizer que peças são produzidas de forma estatisticamente independentes com a mesma distribuição. De forma que, se soubéssemos o valor de π, o processo seria de variáveis independentes e identicamente distribuídas. Isto é,

$$U_1 \perp U_2 \perp \cdots \perp U_i \perp \cdots \perp \mid \pi \text{ e } \Pr(U_i = 1 \mid \pi) = \pi, \ \forall i, i = 1, 2, 3, \ldots$$

A principal consequência destas premissas é a de que X e $\Psi = \tau - X$ são condicionalmente independentes dado π. Em outras palavras, como a amostra é estatisticamente independente da "não amostra", então $X \perp \psi \mid \pi$. A primeira observação a partir destas considerações é a de que ao conhecer o valor $x = 1$ de X, o parâmetro de interesse passa a ser $\Psi = \tau - 1$. Note que o número de peças perfeitas na "não amostra" é $25 - \Psi$. Passemos então ao cálculo da probabilidade a posteriori do parâmetro de interesse, lembrando que a priori para π passou a ser uma beta com

parâmetros $(1; 5)$.

$$\begin{aligned}
\Pr(\Psi = y \mid X = 1) &= \frac{\Pr(\Psi = y \wedge X = x)}{\Pr(X = x)} \\
&= \frac{\int_0^1 \Pr(\Psi = y \wedge X = x \mid \pi) g(\pi) \mathrm{d}\pi}{\int_0^1 \Pr(X = x \mid \pi) g(\pi) \mathrm{d}\pi} \\
&= \frac{\binom{25}{y}\binom{4}{1} \int_0^1 \pi^y (1-\pi)^{25-y} \pi^1 (1-\pi)^4 (1-\pi)^4 \mathrm{d}\pi}{\binom{4}{1} \int_0^1 \pi^1 (1-\pi)^4 (1-\pi)^4 \mathrm{d}\pi} \\
&= \frac{\binom{25}{y} \int_0^1 \pi^{y+1} (1-\pi)^{33-y} \mathrm{d}\pi}{\int_0^1 \pi^1 (1-\pi)^8 \mathrm{d}\pi} \\
&= \frac{\binom{25}{y} \Gamma(y+2)\Gamma(34-y)\Gamma(11)}{\Gamma(36)\Gamma(2)\Gamma(9)} \\
&= \frac{\binom{25}{y}\binom{9}{1} 10}{\binom{34}{y+1} 35}
\end{aligned}$$

$$\therefore \Pr(\Psi = y \mid X = 1) = \frac{\binom{25}{y} 90}{\binom{34}{y+1} 35} I(0 \leq y \leq 25).$$

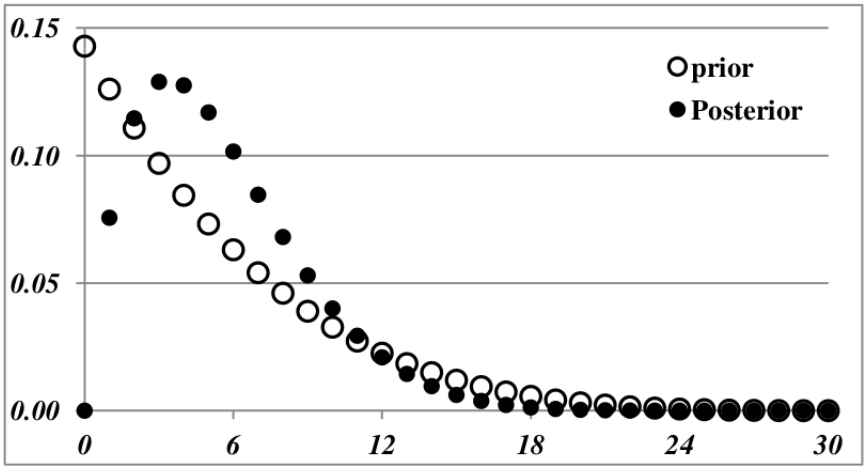
Lembrando que a priori de π é beta $(1; 5)$, podemos obter a priori para y da seguinte forma:

$$\begin{aligned}
\Pr(\tau = t) &= \int_0^1 \binom{30}{t} \pi^t (1-\pi)^{30-t} \frac{\Gamma(6)}{\Gamma(1)\Gamma(5)} (1-\pi)^4 \mathrm{d}\pi \\
&= \frac{5\binom{30}{t}\Gamma(t+1)\Gamma(35-t)}{\Gamma(36)}
\end{aligned}$$

$$\therefore \Pr(\tau = t) = \frac{5\binom{30}{t}}{35\binom{34}{t}} I(0 \leq t \leq 30).$$

Como $\tau = \Psi + 1$, a Figura 1.16 ilustra as distribuições a priori e a posteriori de τ. O vetor de medidas centrais (média, mediana, moda) para estas distribuições seriam, respectivamente, $(5; 4; 0)$ e $(5{,}5; 5{,}0; 3{,}0)$. Temos também que o evento $\{1 \leq \tau \leq 10\}$ tem probabilidade a posteriori igual a $0,91$. Usando apenas a priori o evento $\{\tau \leq 10\}$ tem probabilidade $0,87$. Lembremos que para a posteriori utilizamos uma amostra de tamanho 5 e nossa incerteza passou de 30 unidades para 25 unidades.

■



1.16 Figura. *Priori e Posterior para τ, o número de peças imperfeitas de um lote de 30 peças, onde se utilizou uma amostra de 5 peças com apenas uma imperfeita.*

1.3 Inferência Bayesiana

As seções anteriores focaram na descrição probabilística subjetiva do invisível de interesse, o parâmetro θ. Antes de iniciarmos a presente seção, gostaríamos de fazer alguns esclarecimentos importantes. O elemento *Modelo Estatístico* é universal e não pertence a nenhuma linha de trabalho estatístico. A diferença entre um modelo probabilístico e um modelo estatístico é o fato de o último ser um conjunto de modelos probabilísticos. Para a definição de um modelo estatístico, consideramos um espaço amostral, $\mathcal{X}$, que é constituído dos resultados possíveis de um experimento. Define-se também uma sigma-álgebra, $\mathcal{A}$ de subconjuntos (que se quer probabilizar) do espaço amostral. Em seguida, define-se uma classe de distribuições de probabilidade. Olhando do ponto de vista teórico, o trabalho do estatístico é o de selecionar um dos modelos probabilísticos, ou uma composição destes, para que possa melhor

representar o modelo ideal de geração dos dados observados. Nós Bayesianos usamos uma distribuição de probabilidades para descrever nossas preferências dentro desta classe de modelos. Isto foi justamente o que fizemos nas seções anteriores. A distribuição a priori (a posteriori) representa nossas preferências antes (depois) das observações experimentais. Não há dúvida que outros usam outras técnicas de descrição de preferências ou mesmo de escolhas de distribuições. A trinca formada pelo espaço amostral, pela sigma-álgebra de seus subconjuntos e a família de probabilidades é o arcabouço teórico do modelo estatístico. Nossa posição é a de que devemos descrever probabilisticamente nossas preferências sobre os elementos desta família de probabilidades. Uma das funções mais importantes para um Bayesiano é a probabilidade preditiva do valor da variável x de uma unidade populacional que não fez parte das observações já obtidas. Vamos supor que $v = (x_1, \ldots, x_n)$ é o vetor das observações de uma variável X que devem ser usadas na predição de x_{n+1} (invisível). Nosso desafio é definir a distribuição de probabilidades do estado da natureza. No caso de estarmos trabalhando com densidades, nosso objetivo seria descrever a função $f(x_{n+1}|v)$. Este cálculo, no contexto Bayesiano, é baseado nas fórmulas de probabilidade total e de Bayes. Se as observações são estatisticamente independentes temos o seguinte cálculo:

$$\begin{aligned} f(x_{n+1} \mid v) &= \int_{\Theta} f(x_{n+1}, \theta \mid v) \mathrm{d}\theta \\ &= \int_{\Theta} f(x_{n+1} \mid \theta, v) g(\theta \mid v) \mathrm{d}\theta \\ &= \int_{\Theta} f(x_{n+1} \mid \theta) g(\theta \mid v) \mathrm{d}\theta \end{aligned}$$

Isto é, uma média ponderada pela posteriori, das densidades possíveis de x_{n+1}. Este foi justamente o processo que usamos no último exemplo, bastando entender que no lugar de uma nova observação tínhamos um grande número de unidades em estudo e estávamos interessados na soma destes. Se naquele contexto tivéssemos um lote de 6 peças e observássemos 5 delas, caíríamos no caso acima descrito.

Nosso esforço em descrever o processo Bayesiano teve a intenção de mostrar que, ao definir a distribuição a posteriori, nos apossamos da ferramenta mais poderosa de nosso trabalho. Com esta ferramenta em mãos temos o poder de escolher apropriadamente um valor possível de θ e considerá-lo como sua estimativa. Podemos evidentemente construir conjuntos com as mais altas densidades de forma que as probabilidades desses conjuntos estejam próximas dos níveis que se definiu previamente. Esses são os conjuntos de credibilidade. Ao desejarmos comparar

hipóteses sobre a posição do parâmetro θ, **H** versus **A**, podemos simplesmente encontrar o conjunto tangente e calcular sua probabilidade. Por conjunto tangente (à hipótese **H**) entendemos o conjunto de pontos cujas densidades ou probabilidades sejam maiores do que qualquer ponto do conjunto que define a hipótese **H**. Claro que podemos calcular o conjunto tangente à hipótese **A**. Contudo, no caso de a dimensão de uma hipótese ser inferior a da outra – hipóteses precisas –, o problema deixa de ser simétrico e então o conjunto tangente refere-se apenas a hipótese precisa.

Nesta seção não desejamos desenvolver teoria para cada um dos elementos da trindade estatística; estimação pontual e intervalar e testes de significância ou de hipótese. Vamos apenas mostrar através de exemplos como podemos proceder em cada um dos casos.

Como estimativas pontuais, consideraremos os três tipos padrões, média, mediana e moda da distribuição a posteriori. Para a construção de conjuntos de credibilidade podemos considerar o espaço paramétrico ordenado pelas probabilidades de cada ponto paramétrico. Iniciando a construção do nosso conjunto de credibilidade pelo ponto de maior probabilidade, vamos incluindo os de maior valor na sequência até obtermos um conjunto com a probabilidade mais próxima da credibilidade estabelecida. Dessa forma, garantimos que estamos encontrando o menor conjunto com a credibilidade pretendida. A mesma ideia é usada para a definição de um conjunto tangente. Devemos considerar agora o maior conjunto de pontos paramétricos com densidades ou probabilidades superiores a de todos os pontos que compõem a hipótese: O maior conjunto de credibilidade fora do subespaço da hipótese.

Vamos olhar para os exemplos anteriores e construir essas inferências. No Exemplo 1.9, considerando apenas o caso de $n = 4$, a média é igual a 22,95, a mediana 21 e a moda 18. O conjunto $\{18 \leq \theta \leq 36\}$ tem credibilidade 95,4%. Por outro lado considere a hipótese **H**: $\theta = 30$; o conjunto tangente para esta hipótese é $T_H = \{18 \leq \theta \leq 29\}$ cuja probabilidade é 88,2%. Considerando índice de evidência desenvolvido em Pereira and Stern (1999), $Ev(H) = 1 - \Pr(T_H) = 11,8\%$, que muitas vezes pode ser considerado baixo para aceitar a nossa hipótese e não tanto para rejeitá-la. Note que, se tivéssemos 37 no lugar de 30, o conjunto tangente seria $\{18 \leq \theta \leq 36\}$ e a evidência seria 4,6%, abaixo do nível de significância canônico de 5%. Temos assim o mesmo problema que se enfrenta com o uso do valor-p da estatística frequentista.

Relembrando o Exemplo 1.11, novamente olhemos o caso de $n = 4$. A média, a mediana e a moda da posteriori assumem, respectivamente, os valores 36,57, 34,16 e 28. O conjunto $\{28 \leq q \leq 48,58\}$ tem credibilidade 90%. Para **H**: $q = 50$ a evidência seria $Ev(H) = 1 - \Pr(T_H) = 0,0824$.

No Exemplo 1.12 temos uma posteriori Beta com parâmetros 4 e 8 e assim a

média é 1/3 e a moda é 0,3. A mediana, usando-se o cálculo da beta incompleta, é igual a 0,3238. O intervalo [0,094; 0,588] é um intervalo de 95% de credibilidade para π. Por outro lado, a hipótese **H**: $\pi < 0,1$ tem probabilidade 0,0185 de ser verdadeira. Esta informação vai contra os interesses do vendedor. Para a hipótese pontual $\pi = 0,1$ a evidência a favor é apenas 0,06, o que também não ajuda muito o vendedor.

Finalmente com a posteriori do Exemplo 1.15, calculamos as estimativas padrões: média = 5,55; mediana = 5; e moda = 3. O conjunto $\{1 \leq \tau \leq 10\}$ tem credibilidade igual a 0,91. Por outro lado a chance do lote ter menos de seis peças imperfeitas é $\Pr(\tau < 6) = 0,56$. A evidência em favor de **H**: $\tau = 1$, é 0,51. Embora a probabilidade de a peça encontrada na amostra ser a única imperfeita do lote, a probabilidade deste evento é apenas 0,08. Contudo, com os outros dois indicadores, há uma crença de o comprador estar levando um bom lote.

Concluímos este capítulo apresentando mais alguns exemplos que devem esclarecer melhor a elegância e a simplicidade do método Bayesiano.

1.17 Exemplo. O raio de um círculo tem comprimento ρ e, ao escolher 5 pontos ao acaso dentro do círculo, observamos os seguintes pontos $(x; y)$'s: (0,5; -1,4); (1,3; 2,1); (-1,2; -2,4); (-0,7; -2,3); (2,5; 2,2). Considere o centro do círculo como o ponto $(x; y)$ = (0; 0) e a seguinte densidade como a priori para ρ: $f(\rho) = (2\rho^2)^{-1} \therefore \rho > 0,5$. Nosso objetivo é fazer inferências sobre o valor de ρ. Sendo a área do círculo $\pi\rho^2$, a verossimilhança associada a ρ é a seguinte: $L(\rho \mid dados) = (\pi\rho^2)^{-5} I(m \leq \rho)$. Dessa forma, ao usarmos a priori indicada, a distribuição a posteriori será proporcional a ρ^{-12}. Novamente, I é a função indicadora e m é o máximo das distâncias dos pontos ao centro (0; 0). No nosso caso, $m = 3,4$ e assim a moda sendo este valor, a média calculada como

$$M = \frac{\int_{3,4}^{\infty} \rho^{-11} \mathrm{d}\rho}{\int_{3,4}^{\infty} \rho^{-12} \mathrm{d}\rho} = \frac{11(3,4)^{-11}}{10(3,4)^{-10}} = 1,1 \times 3,4 = 3,74$$

e finalmente a mediana sendo o valor de z que satisfaz a seguinte igualdade

$$\frac{\int_{z}^{\infty} \rho^{12} \mathrm{d}\rho}{\int_{3,4}^{\infty} \rho^{12} \mathrm{d}\rho} = \frac{1}{2};$$

mediana igual a 3,62. Por outro lado, o intervalo [3,397;4,464] tem 95% de credibilidade. Por fim, a evidência a favor de **H**: $\rho = 4,5$ é Ev=0,0458.

■

1.18 Exemplo. Substitua no Exemplo 1.17 o círculo por um quadrado de lado 2ρ, onde o centro do quadrado é o ponto (0; 0). Com exatamente os mesmos dados e mesma priori, vamos realizar as inferências sobre o parâmetro de interesse ρ. Lembrando que a área do quadrado é $4\rho^2$, novamente a posteriori é proporcional a ρ^{12}, mas agora com o suporte sendo o conjunto dos reais não inferiores ao máximo dos valores absolutos das coordenadas dos pontos amostrais. Isto é, $\rho \geq 2,5$. Com apenas a mudança de suporte, os cálculos anteriores se repetem. A média igual a $1,1 \times 2,5 = 2,75$, mediana igual a 2,66 e a moda evidentemente é igual a 2,5. O intervalo (2,5; 3,3) tem credibilidade de 95%. A evidência a favor da hipótese **H**: $\rho = 4$ é Ev=0,0057.

■

Com os exemplos apresentados, provamos que a ferramenta mais importante do Bayesiano é sua posteriori, embora em alguns casos, sem algoritmos eficientes podemos não conseguir soluções para os problemas estatísticos. Ao longo de nosso livro veremos aplicações importantes e que nos ajudarão a resolver a maioria dos problemas estatísticos. A Figura 1.19 é simplesmente a ilustração do quadrado e do círculo com os pontos que foram escolhidos. O leitor irá perguntar "posso decidir de onde vieram os pontos amostrais que foram apresentados?". Essa é a questão fundamental da estatística: Olhar os dados e tentar adivinhar qual o modelo gerador. Vamos representar por **O** a hipótese de que a amostra veio de um círculo e por **D** do quadrado. Nosso objetivo é o de calcular a probabilidade de **O**, dado as observações. Aqui, como vimos no caso da paternidade, trabalhar com os odds pode ser adequado. Representemos por X os dados e por r e R os odds a priori e a razão de Bayes. Lembremos que a probabilidade a posteriori tem a seguinte representação: $\Pr(\boldsymbol{O} \mid X) = (1 + rR)^{-1}$.

A primeira questão a ser lembrada é sobre as estatísticas suficientes nos dois casos: círculo e quadrado. Percebam que, no caso do círculo, devemos olhar para a distância do ponto observado ao centro (0; 0) e assim a estatística suficiente C é o máximo das distâncias. No caso do quadrado, a estatística suficiente Q é o máximo dos máximos das coordenadas absolutas de todos os pontos amostrais. Assim, não é difícil encontrarmos as funções de distribuição dessas variáveis. Dizer que $C \leq c$, é dizer que todas as distâncias da amostra são iguais ou menores que c. Por outro lado, dizer que $Q \leq q$, é dizer que todas as coordenadas absolutas de todos os pontos amostrais são iguais ou menores que q. Podemos agora encontrar as densidades

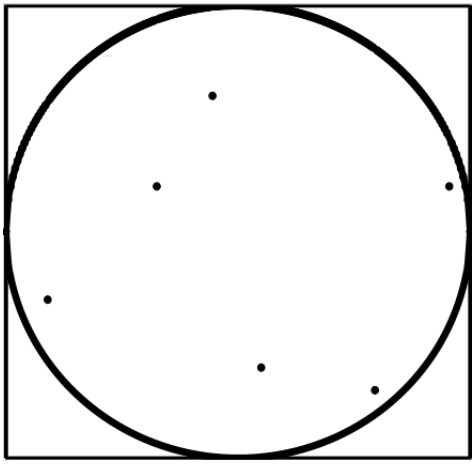
amostrais dos dois modelos no caso das 5 observações:

$$\begin{cases} \Pr(C \leq c \mid \rho) = \left(\frac{\pi c^2}{\pi \rho^2}\right)^5 I(\rho \geq c) = \left(\frac{c}{\rho}\right)^{10} I(\rho \geq c) \\ \qquad\qquad \Rightarrow f(c \mid \rho) = 10 c^9 \rho^{10} I(\rho \geq c) \\ \\ \Pr(Q \leq q \mid \rho) = \left(\frac{4q^2}{4\rho^2}\right)^5 I(\rho \geq q) = \left(\frac{q}{\rho}\right)^{10} I(\rho \geq q) \\ \qquad\qquad \Rightarrow g(q \mid \rho) = 10 q^9 \rho^{10} I(\rho \geq q) \end{cases}$$

$$\Rightarrow \begin{cases} f(c \mid \boldsymbol{O}) = 5c^9 \int_c^\infty \rho^{-12} \mathrm{d}\rho = \frac{5c^9}{11c^{11}} = \frac{5}{11c^2} I(c \geq 0,5) \\ g(q \mid \boldsymbol{D}) = 5q^9 \int_c^\infty \rho^{-12} \mathrm{d}\rho = \frac{5q^9}{11q^{11}} = \frac{5}{11q^2} I(q \geq 0,5) \end{cases}$$

$$\Rightarrow R = \left(\frac{c}{q}\right)^2$$

Para o cálculo das posteriores da escolha entre círculo e quadrado, temos de considerar priores para estas figuras. Imaginemos inicialmente a probabilidade simétrica: $\Pr(\boldsymbol{O}) = \Pr(\boldsymbol{D}) = 1/2$.



1.19 Figura. *Círculo de raio ρ e quadrado de lado 2ρ com as observações das escolhas aleatórias.*

Lembrando nossos resultados amostrais: $c = 3,4$ e $q = 2,5$ resulta em $R = 1,36$. Assim, $r = 1$ e $\Pr(O \mid dados) = (1 + 1,8496)^{-1} = 0,3509$. Evidentemente sempre ocorrerá $c \geq q$, pois a distância de um ponto ao centro é sempre maior que o valor absoluto de qualquer de suas coordenadas. A probabilidade a priori irá exercer um forte papel na decisão da escolha do modelo neste caso.

Caso semelhante a este ocorrerá se tivermos de decidir se uma amostra de peças, contando as falhas ocorridas, é proveniente de uma binomial ou de uma binomial negativa. Estes são casos de modelos separados. A Figura 1.19 é aqui colocada para a apreciação do leitor de forma que ele possa vislumbrar as razões da preferência sobre o quadrado. Sendo este assunto interessante para muitas discussões quanto ao uso de nossas preferências metodológicas.

Caso semelhante a este ocorrerá se tivermos de decidir se uma amostra de peças, contando as falhas ocorridas, é proveniente de uma binomial ou de uma binomial negativa. Estes são casos de modelos separados. A Figura 1.19 é aqui colocada para a apreciação do leitor de forma que ele possa vislumbrar as razões da preferência sobre o quadrado. Sendo este assunto interessante para muitas discussões quanto ao uso de nossas preferências metodológicas.

O Exemplo 1.20 a seguir complementa este capítulo apresentando um problema clássico da estatística.

1.20 Exemplo. Um funcionário de uma indústria deixou anotados os resultados realizou durante um dia de trabalho. Os números das peças imperfeitas e perfeitas, x e y, eram os únicos números anotados. Considerando um processo de Bernoulli com parâmetro π – a probabilidade de uma peça ser produzida com imperfeição – como modelo estatístico básico, ficou a questão de qual regra de parada foi usada no processo de observação. A estimação de p, certamente não é um problema para a inferência Bayesiana. Normalizando a verossimilhança teríamos uma densidade beta com parâmetros $x+1$ e $y+1$ como posteriori de p. Teríamos simplesmente a média $m = x/(x+y)$, a média, como uma possível estimativa de Bayes. cuja imprecisão pode ser avaliada pela variância $v = m(1-m)/(x+y+1)$. Contudo, para uma escolha entre o modelo binomial ou binomial negativa, tem-se o mesmo problema do Exemplo 1.18. A razão de Bayes seria igual a

$$R(x;y) = \frac{\binom{x+y-1}{y}}{\binom{x+y}{y}} = \frac{x}{x+y} \leq 1,$$

favorecendo, evidentemente, a Binomial. Senão vejamos: consideremos a priori probabilidade $1/2$ para cada uma das hipóteses, **H**: Binomial e **A**: Binomial Negativa. Com as observações $x = 5$ e $y = 13$, teríamos o seguinte:

$$\Pr\{H \mid (x;y) = (5;13)\} = \frac{1}{1+R(x;y)} = \frac{17}{22} = 0,7826,$$

passando de $1/2$ para esta nova probabilidade. Muitos argumentos podem ser usados para uma melhor escolha da probabilidade a priori da regra de parada usada

pelo funcionário. Dependendo da escolha a binomial negativa pode ser a de maior probabilidade a posteriori.

Para mais detalhes sobre teste de hipótese Bayesiano sugerimos a leitura de Pereira and Stern (1999), Pereira et al. (2008) e Diniz et al. (2012).

Capítulo 2

Dados categóricos

Neste capítulo apresentamos o teste de hipótese de homogeneidade para tabelas de contingências 2×2 e modelos de regressão para dados binários.

2.1 Teste de hipótese para tabelas de contingências 2×2.

Nesta seção, descrevemos o teste de hipótese de homogeneidade para tabelas de contingência 2×2 (Tabela 2.1). Note que, na Tabela 2.1, X e Y são duas variáveis aleatórias as quais temos o interesse de testar se a distribuição delas, referente às características C_1 e C_2, são homogêneas. Desta forma, os valores observados da variável X é a e da variável Y é c. Inicialmente, consideramos que n e m são valores fixos. Portanto, segue naturalmente a suposição de que X tem distribuição Binomial(n, p_x) e Y tem distribuição Binomial(m, p_y).

2.1 Tabela. *Tabela de contingência* 2×2.

	C_1	C_2	total
X	a	b	n
Y	c	d	m

A hipótese **H** de homogeneidade é escrita como

$$\mathbf{H} : p_x = p_y.$$

Para cálculo do e-valor, é necessário encontrar primeiro a distribuição a posteriori de p_x e p_y. Assume-se que a distribuição a priori de p_x é uniforme em $(0; 1)$ e

para p_y também. Como X e Y têm distribuição Binomial, a distribuição a posteriori p_x e p_y é dada por

$$p_x \mid a, n \sim \text{Beta}(a+1, n-a+1) \tag{2.2}$$

$$p_y \mid c, m \sim \text{Beta}(c+1, m-c+1). \tag{2.3}$$

Neste caso o $\sup_{p_x,p_y \in H} \pi(p_x, p_y \mid a, c, m, n)$ pode ser calculado pela maximização em p de

$$\pi(p \mid a, c, m, n) = \frac{p^{a+c}(1-p)^{n+m-a-c}}{\mathcal{B}(a+1, n-a+1)\mathcal{B}(c+1, m-c+1)},$$

em que $\mathcal{B}(\cdot, \cdot)$ é a função beta. Como a, c, n e m são número inteiros, temos que

$$\pi(p \mid a, c, m, n) = \binom{n}{a}\binom{m}{c}(n+1)(m+1)p^{a+c}(1-p)^{n+m-a-c}$$

e

$$\begin{aligned}\sup_p \pi(p \mid a, c, m, n) &= \sup_{p_x,p_y \in H} \pi(p_x, p_y \mid a, c, m, n) \\ &= \frac{(n+1)!(m+1)!}{a!c!(n-a)!(m-c)!}\left(\frac{a+c}{n+m}\right)^{a+c}\left(\frac{n+m-a-c}{n+m}\right)^{n+m-a-c}.\end{aligned}$$

Então, temos que T, conjunto tangente a hipótese **H**, é dado por

$$T(a, c, m, n) = \{p \in (0,1) : \pi(p \mid a, c, m, n) \geq \sup_p \pi(p \mid a, c, m, n)\}$$

e o e-valor é dado por

$$\text{e-valor} = 1 - \Pr(p \in T(a, c, m, n)).$$

Para calcular o e-valor utilizamos o seguinte algoritmo:

1. Geramos uma amostra de tamanho k da distribuição a posteriori de p_x, p_y, obtendo $\{p_{x1}, p_{y1}\}, \ldots, \{p_{xk}, p_{yk}\}$.

2. Calculamos o e-valor por

$$1 - \frac{1}{k}\sum_{i=1}^{k} I\left(\pi(p_{xi}, p_{yi} \mid a, c, m, n) \geq \sup_p \pi(p \mid a, c, m, n)\right),$$

em que $I(A)$ é a função indicadora do conjunto A.

2.4 Exemplo. Com o objetivo de entender melhor se psicóticos tinham tendências suicidas maiores do que neuróticos, um pesquisador fez um estudo avaliando estas condições. Os dados foram coletados e organizados em uma tabela de contingência (Tabela 2.5). A hipótese estatística referente ao interesse do pesquisador pode ser escrita como uma hipótese de homogeneidade entre Psicose e Neurose.

2.5 Tabela. *Tabela de contingência* 2×2 *referente aos dados observados do Exemplo 2.4.*

Pacientes	Suicida	Não-Suicida	total
Psicose (X)	4	16	20
Neurose (Y)	12	18	30

As distribuições a posteriori p_x e p_y são

$$p_x \mid 4; 20 \sim \text{Beta}(5; 17) \tag{2.6}$$

$$p_y \mid 12; 30 \sim \text{Beta}(13; 19), \tag{2.7}$$

e

$$\pi(p \mid 4; 12; 20; 30) = 2,728078 \times 10^{14} p^{16}(1-p)^{34}.$$

Assim,

$$\sup_p \pi(p \mid 4; 12; 20; 30) = \sup_{p_x, p_y \in H} \pi(p_x, p_y \mid 4; 12; 20; 30) = 6,661308$$

e

$$T(4; 12; 20; 30) = \{p \in (0; 1) : \pi(p \mid 4; 12; 20; 30) \geq 6,661308\}.$$

Logo, o e-valor é dado por

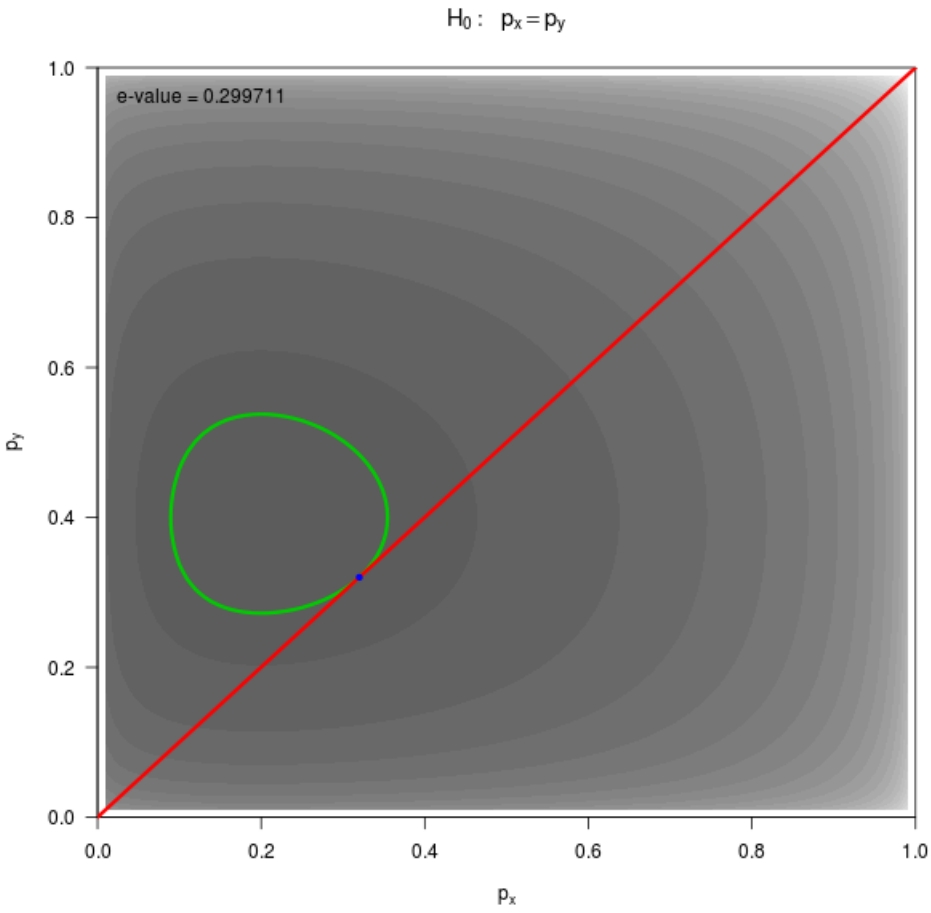
$$\text{e-valor} = 1 - \Pr(p \in T(4; 12; 20; 30)) = 0,2997.$$

A Figura 2.8 ilustra o FBST neste problema, em que a distribuição a posteriori conjunta, a região da hipótese e o espaço tangente, estão desenhadas.

■

2.2 Modelos de regressão para dados binários

Dados binários são aqueles que admitem dois resultados possíveis para a variável resposta. Estes são utilizados em diversas áreas do conhecimento. Alguns exemplos práticos em que este tipo de resposta aparece são: (*i*) concessões de crédito de



2.8 Figura. *Curvas de nível da posteriori de p_x, p_y (em cinza). A linha vermelha é a região da hipótese **H**, a elipse em verde determina a fronteira do espaço tangente T e o ponto em azul é o ponto em que **H** tangencia T.*

um banco, aprovado ou não aprovado; (*ii*) resultado do diagnóstico de um exame laboratorial, positivo ou negativo; (*iii*) intenção de voto de um eleitor em relação ao candidato A, vota ou não vota; (*iv*) inspeção de uma peça recém-fabricada, defeituosa ou não defeituosa; (*v*) teste da publicidade de um novo produto, vendeu ou não vendeu, etc. Consideramos esses casos como um problema de sucesso e fracasso.

A pesquisa com dados binários intensificou-se a partir da década de 50. Um dos primeiros estudos abordava um problema de uma tabela de contingência 2×2 aplicado em epidemiologia. Uma revisão sobre o tema está em Richardson (1994). Muitos trabalhos desenvolvidos nas décadas de 50 e 60 são utilizados até hoje, principalmente na análise descritiva dos dados.

Por muitos anos, a regressão linear normal foi usada para explicar a maioria dos fenômenos aleatórios. Mesmo quando não era razoável assumir normalidade, utilizava-se algum tipo de transformação para alcançar a normalidade desejada. Um

dos métodos mais utilizados para este fim é a transformação de Box-Cox (Box and Cox, 1964).

Com o desenvolvimento computacional a partir da década de 70, alguns modelos que exigiam a utilização de processos iterativos para a estimação dos parâmetros começaram a ser mais utilizados. Nelder and Wedderburn (1972) propuseram os modelos lineares generalizados (MLGs), cuja ideia básica consiste em abrir o leque de opções para a distribuição da variável resposta, incluindo as distribuições que pertençam à família exponencial, bem como dar maior flexibilidade para a relação funcional entre a média da variável resposta e a parte linear do modelo.

Sob a perspectiva Bayesiana, Dey et al. (2000) editaram um livro com trabalhos sobre MLGs. Existem também alguns pacotes para `R` (R Core Team, 2013) que abordam este assunto, como por exemplo Martin et al. (2011) e Gelman and Su (2013). Aqui, estamos interessados na parte de regressão binária. Sendo assim, apresentamos na sequência algumas funções de ligação e mostramos como fazer a estimação utilizando o pacote `binreg` (dos Santos et al., 2013).

2.2.1 Funções de Ligação

Seja $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{X_1}, \ldots, \boldsymbol{X_r})'$ a matriz de experimento, em que $\mathbf{1}$ é um vetor com todos seus valores iguais a 1. Denotamos a variável resposta binária pelo vetor $\boldsymbol{Y}$. O interesse consiste em modelar a $\Pr[Y_i = 1 \mid \eta_i] = \mu(\eta_i) = \mathrm{E}(Y_i)$ por $\Pr[Y_i = 1 \mid \eta_i] = g_{\boldsymbol{\theta}}^{-1}(\eta_i)$, $i = 1, \ldots, n$, em que $\boldsymbol{\eta} = \boldsymbol{\beta X}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_r)$ são os coeficientes lineares, $g_{\boldsymbol{\theta}}(\cdot)$ é a função de ligação e $\boldsymbol{\theta}$ é um vetor de parâmetros da função de ligação para as ligações que contém parâmetros. A função de ligação relaciona as covariáveis $\boldsymbol{X}$ com a média da resposta $\mu = \mathrm{E}(Y \mid X)$. Neste caso, $g_{\boldsymbol{\theta}}^{-1}$ é uma função de distribuição na reta real. A seguir apresentamos as funções de ligação tratadas neste livro.

LOGITO: A distribuição logística tem densidade dada por

$$f(y) = \frac{\exp(y)}{(1 + \exp(y))^2} \tag{2.9}$$

em que -$\infty < y < \infty$. A função de distribuição é dada por

$$F(y) = \frac{\exp(y)}{1 + \exp(y)}. \tag{2.10}$$

O modelo logístico binomial é obtido substituindo a notação $F(y)$ pela representação da proporção μ e y pela representação do componente linear ($X\beta$ =

η) na Equação (2.10). Note que, para qualquer valor de η no intervalo $(-\infty; \infty)$, existe um valor de μ em (0;1). Assim se $\eta \to -\infty$ temos que $\mu \to 0$, se $\eta \to \infty$ temos que $\mu \to 1$ e para η = 0, μ = 0,5.

O modelo para dados binários com ligação logito é definido por

$$\mu = g_{\boldsymbol{\theta}}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{2.11}$$

ou, equivalentemente, por

$$\eta = g_{\boldsymbol{\theta}}^{-1}(\mu) = \log\left(\frac{\mu}{1-\mu}\right). \tag{2.12}$$

PROBITO: A ligação probito é definida por

$$\eta = g_{\boldsymbol{\theta}}^{-1}(\mu) = \Phi^{-1}(\mu) \tag{2.13}$$

em que $\Phi(.)$ é a função de distribuição da normal padrão. Vale ressaltar que, para qualquer valor de η no intervalo $(-\infty; \infty)$, há um valor da função probito de μ no intervalo (0; 1). Observe que para η = 0 temos μ = 0,5.

COMPLEMENTAR LOG-LOG: A função de ligação complementar log-log é derivada da distribuição do valor extremo e dada por

$$\mu = g_{\boldsymbol{\theta}}(\eta) = 1 - \exp(-\exp(\eta)). \tag{2.14}$$

PRENTICE: A função de ligação proposta por Prentice (1976) abrange os modelos logito, probito e algumas ligações assimétricas como casos limites (por exemplo, complementar log-log). Prentice utilizou a função de distribuição do $\log(F_{2m_1,2m_2})$ como função de ligação, em que $F_{2m_1,2m_2}$ é uma variável aleatória com distribuição F-Snedecor com parâmetros $2m_1$ e $2m_2$, dada por

$$f(y) = \frac{\exp(ym_1)(1+\exp(y))^{-(m_1+m_2)}}{\mathcal{B}(m_1, m_2)}, \tag{2.15}$$

em que, $\theta = (m_1, m_2)$, $\mathcal{B}$ representa a função beta. Para $m_1 = m_2 = 1$ obtemos a ligação logito, $m_1 \to \infty$ e $m_2 \to \infty$ obtemos a ligação probito, $m_1 = 1$ e $m_2 \to \infty$ obtemos a ligação do valor mínimo extremo e $m_1 \to \infty$ e $m_2 = 1$ obtemos a ligação do valor máximo extremo.

ARANDA-ORDAZ: Uma outra transformação importante, proposta por Aranda - Ordaz (1981), é uma função de ligação uni-paramétrica assimétrica que tem como casos particulares os modelos logito e complementar log-log, dada por

$$\eta = g_{\boldsymbol{\theta}}^{-1}(\mu) = \log\left[\frac{(1-\mu)^{-\alpha}-1}{\alpha}\right], \tag{2.16}$$

em que, $0 < \mu < 1$, $\theta = \alpha$ e α é uma constante desconhecida. Quando α = 1 temos a ligação logito e $\alpha \to 0$ temos a ligação complementar log-log.

STUKEL: Stukel (1988) definiu uma classe de ligações bi-paramétricas que generaliza o modelo logístico. O modelo proposto por Stukel aproxima várias distribuições importantes, como a probito, a complementar log-log e outras funções de ligação assimétrica. A generalização proposta é

$$\log\left(\frac{\mu}{1-\mu}\right) = h(\eta), \tag{2.17}$$

em que $\theta = (a_1, a_2)$ e $h(\eta)$ é uma função não linear estritamente crescente indexada por dois parâmetros de forma a_1 e a_2. Esta função é definida a seguir.

Para $\eta > 0$

$$h(\eta) = \begin{cases} \left[\exp\left(a_1 \mid \eta \mid\right) - 1\right]/a_1, & \text{para } a_1 > 0 \\ \eta, & \text{para } a_1 = 0 \\ -\left[\log\left(1 - a_1 \mid \eta \mid\right)\right]/a_1, & \text{para } a_1 < 0 \end{cases}. \tag{2.18}$$

Para $\eta < 0$

$$h(\eta) = \begin{cases} -\left[\exp\left(a_2 \mid \eta \mid\right) - 1\right]/a_2, & \text{para } a_2 > 0 \\ \eta, & \text{para } a_2 = 0 \\ \left[\log\left(1 - a_2 \mid \eta \mid\right)\right]/a_2, & \text{para } a_2 < 0 \end{cases}. \tag{2.19}$$

Um detalhe importante sobre o modelo proposto pela Stukel é que, sob prioris uniformes impróprias para os parâmetros β, a distribuição a posteriori será imprópria (Chen et al., 1999).

WEIBULL: A função de ligação Weibull, proposta por Caron and Polpo (2009), pode ser simétrica ou assimétrica. Além disso, as funções de ligação logito, probito e complementar log-log podem ser obtidas como casos limites.

A função de ligação Weibull é definda por

$$\begin{array}{rll} \eta = & g_{\boldsymbol{\theta}}(\mu) & = [-\log(1-\mu)]^{\frac{1}{\gamma}}, \\ \mu = & g_{\boldsymbol{\theta}}^{-1}(\eta) & = 1-\exp\{-\eta^{\gamma}\}, \end{array} \tag{2.20}$$

em que $\theta = \gamma$, $\gamma > 0$ e $\eta > 0$.

2.2.2 Estimação

Considere uma amostra de tamanho n de uma varável binária Y, com $\Pr[Y_i = 1] = p_i$ para $i = 1, \ldots, n$. Denotamos os dados observados por $\mathcal{D} = \{n, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X} = \boldsymbol{x}\}$, em que $\boldsymbol{y} = (y_1, \ldots, y_n)$ é o vetor observado de $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ e $\boldsymbol{x} = (\boldsymbol{1}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_r})'$ é a matriz de experimento observada de $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X_1}, \ldots, \boldsymbol{X_r})'$. A função de verossimilhança pode ser escrita como

$$L(\boldsymbol{\beta}, \theta \mid \mathcal{D}) \propto \prod_{i=1}^{n} {p_i}^{y_i}(1-p_i)^{1-y_i} \tag{2.21}$$

e a função de log-verossimilhança por

$$l(\boldsymbol{\beta}, \theta \mid \mathcal{D}) \propto \sum_{i=1}^{n} \left[y_i \log(p_i) + (1-y_i)\log(1-p_i) \right].$$

Tomando $g_{\boldsymbol{\theta}}^{-1}(\eta_i) = p_i$ temos

$$l(\boldsymbol{\beta}, \theta \mid \mathcal{D}) \propto \sum_{i=1}^{n} \left[y_i \log(g_{\boldsymbol{\theta}}^{-1}(\eta_i)) + (1-y_i)\log(1-g_{\boldsymbol{\theta}}^{-1}(\eta_i)) \right] \tag{2.22}$$

em que η_i é o i-ésimo elemento do vetor $\boldsymbol{\eta} = \boldsymbol{\beta X}$ e os parâmetros a serem estimados são $\boldsymbol{\beta}$ e θ. A densidade a posteriori é dada por

$$p(\boldsymbol{\beta}, \theta \mid \mathcal{D}) \propto L(\boldsymbol{\beta}, \theta \mid \mathcal{D}) p(\boldsymbol{\beta}, \theta), \tag{2.23}$$

em que $p(\boldsymbol{\beta}, \theta)$ é a priori conjunta. Um procedimento de MCMC é utilizado para gerar uma amostra da distribuição a posteriori. O pacote `binreg` tem implementado o modelo estatístico para as funções de ligação descritos na Seção 2.2.1. Para a geração da distribuição a posteriori, é utilizado o pacote `LaplacesDemon` (Statisticat, 2013). O uso destes pacotes é apresentado na sequência.

2.24 Exemplo. O objetivo original desse estudo era obter um inseticida eficaz contra besouros (Bliss, 1935). Para isso, 481 besouros foram expostos a diferentes concentrações de dissulfeto de carbono (CS_2) durante cinco horas e contou-se o número de insetos mortos. Esse conjunto de dados é conhecido por não ser bem ajustado por modelos simétricos, em particular logito e probito. Por conta disso, é amplamente citado em trabalhos que buscam alternativas a esses modelos. Os dados são apresentados na Tabela 2.25.

2.25 Tabela. *Mortalidade de besouros expostos a CS_2.*

log(Dose)	N^o de besouros	
CS_2	Expostos	Mortos
1,6907	59	6
1,7242	60	13
1,7552	62	18
1,7842	56	28
1,8113	63	52
1,8369	59	53
1,8610	62	61
1,8839	60	60

Um ponto importante para a análise Bayesiana é a especificação da distribuição a priori dos parâmetros. Aqui, utilizamos como distribuição a priori aquelas especificadas no pacote `binreg`. Na sequência do exemplo, desenvolvemos a estimação do modelo Aranda-Ordaz, apresentando os códigos em `R`. A estimação dos outros modelos seguem de forma similar.

Desta forma, primeiro precisamos "carregar" o pacote `binreg` e digitar os dados:

```
library(binreg)

# Dados referentes a mortalidade de besouros
# (Bliss, 1935)
ldose    <- c(1.6907,1.7242,1.7552,1.7842,
              1.8113,1.8369,1.8610,1.8839)
expostos <- c(59,60,62,56,63,59,62,60)
mortos   <- c( 6,13,18,28,52,53,61,60)
```

O pacote `binreg` foi elaborado para variáveis binárias. Assim, é necessário organizar os dados para que a variável resposta seja zero ou um. Além disso, precisamos de uma formatação especial para utilizar o `LaplacesDemon`. Abaixo,

apresentamos o R para tratar os dados segundo as características das funções que utilizamos.

```
#Formatando os dados
y <- rep(0,sum(expostos))
x <- rep(0,sum(mortos))
k <- 0
for (j in 1:length(expostos)) {
  for (i in 1:expostos[j]) {
    k <- k+1
    if (i <= mortos[j])
      y[k] <- 1
    x[k] <- ldose[j]
  }
}

#Modelo Aranda-Ordaz:
dados <- dataLD(y ~ x, "Aranda")
```

A função dataLD é parte do pacote binreg. Para facilitar o usuário, esta função organiza os dados e apresenta uma sugestão de código para fazer a estimação do modelo.

Agora, temos a variável resposta y, a covariável x e dados formatados, como o necessário, em dados. Para obter a distribuição a posteriori do modelo Aranda-Ordaz via MCMC (para mais detalhes sobre métodos de MCMC veja Gamerman, 1997), primeiro, calculamos o máximo a posteriori dos parâmetros para utilizá-los como um valor inicial da cadeia de Markov:

```
set.seed(666) # recomenda-se o uso do set.seed para o
              # leitor obter resultados similares ao
              # do livro.

#Calculando a moda (máximo) a posteriori
fit.mode <- LaplaceApproximation(Model = ModelAranda(),
                      parm = GIV(ModelAranda(), dados),
                      Data = dados,
                      Iterations = 1000,
                      Method = "NM")
```

em que Model = ModelAranda() indica o modelo a ser estimado, Data = dados são os dados no padrão da função LaplacesApproximation,

`parm = GIV(ModelLogit(), data)` é o valor inicial para os parâmetros do modelo, `GIV(ModelLogit(), data)` gera pontos iniciais a partir da distribuição a priori, `Iterations = 1000` é a quantidade máxima de iterações do algoritmo de maximização escolhido e `Method = "NM"` indica o algoritmo de maximização escolhido. Neste caso, foi escolhido o método de Nelder-Mead (`NM`). Mais detalhes podem ser obtidos no `R` executando `help(LaplaceAprroximation)`.

Utilizar o máximo a posteriori como valor inicial dos parâmetros muitas vezes reduz o tamanho do *burn-in* necessário, mas não é uma receita que deva ser sempre seguida. Depois, gera-se primeiro o *burn-in* (descarte das primeiras amostras da cadeia), que neste caso será de 100.000 (cem mil pontos):

```
#Gerando o burn-in
Initial.Values <- as.initial.values(fit.mode)
fit.burn <- LaplacesDemon(Model = ModelAranda(),
                      Data = dados,
                      Initial.Values = Initial.Values,
                      Iterations = 100000,
                      Thinning = 1,
                      Algorithm = "AM",
                      Specs = list(Adaptive = 1000,
                                   Periodicity = 1000))
```

em que `Model = ModelAranda()` indica o modelo a ser estimado, `Data = dados` são os dados no padrão da função `LaplacesDemon`, `Initial.Values = Initial.Values` é o valor inicial para os parâmetros do modelo, `Iterations = 100000` é o tamanho da amostra da posteriori (neste caso o *burn-in*), `Thinning = 1` é o salto entre cada ponto gerado da posteriori, `Algorithm = "AM"` indica o uso do algoritmo de Metropolis Adaptativo e `Specs = list( Adaptive = 1000, Periodicity = 1000)` indica que o processo adaptativo será realizado após os primeiros `1000` pontos e será atualizado a cada `1000` pontos gerados.

Na segunda etapa, geramos uma amostra de tamanho 1000 da distribuição a posteriori dos parâmetros do modelo, considerando um salto de tamanho 500 entre cada ponto gerado:

```
#Gerando amostra da posteriori
Initial.Values <- as.initial.values(fit.burn)
fit <- LaplacesDemon(ModelAranda(), Data=dados,
   Initial.Values, Covar=fit.burn$Covar,
```

```
   Iterations=500000, Thinning=500, Algorithm="AM",
   Specs=list(Adaptive=1000, Periodicity=1000))
```

2.26 Observação. *Para a estimação dos outros modelos, basta substituir o nome do modelo na função* `dataLD` *(isto é,* `"Aranda"` *por* `"logit"`, `"probit"`, `"cloglog"`, `"Weibull"`, `"Prentice"` *ou* `"Stukel"`*) e na função* `"LaplacesDemon"` *(isto é,* `"ModelAranda"` *por* `"ModelLogit"`, `"ModelProbit"`, `"ModelCloglog"`, `"ModelWeibull"`, `"ModelPrentice"` *ou* `"ModelStukel"`*).*

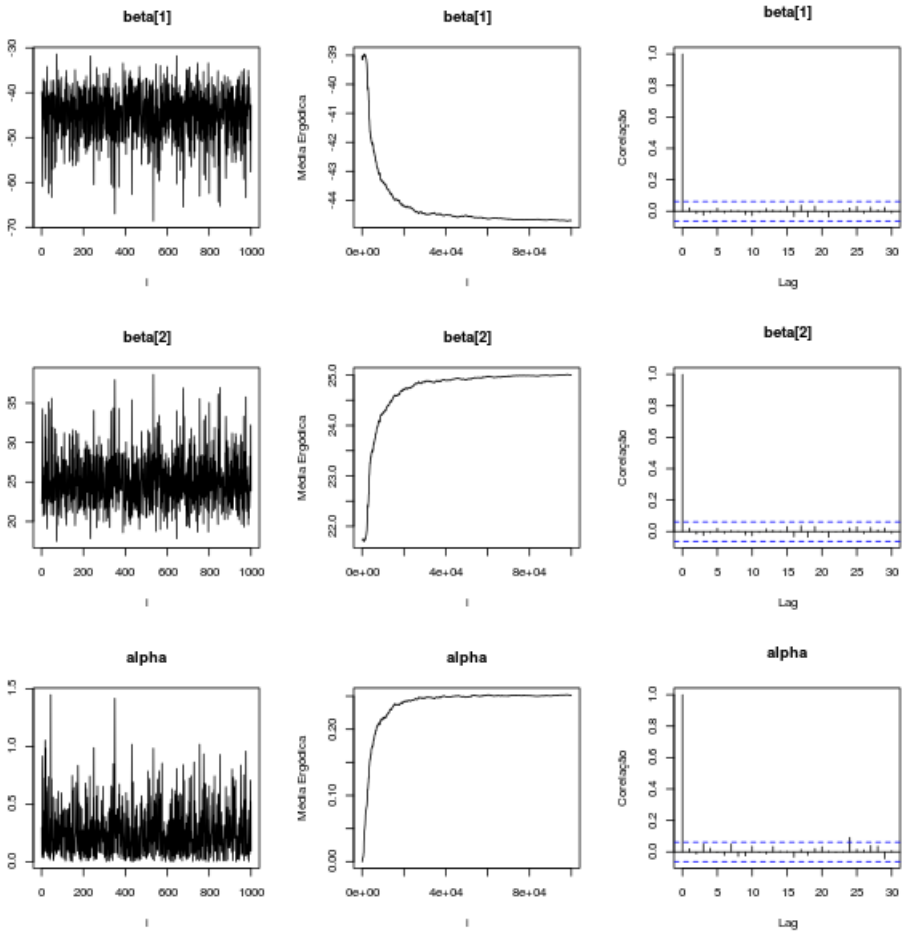
A função `Consort(fit)` do pacote `LaplacesDemon` auxilia a verificação da qualidade da geração da amostra da posteriori (isto é, convergência do algoritmo de MCMC), em que podemos verificar a taxa de aceitação, DIC, medidas descritivas dos parâmetros e uma sugestão de comando para utilizar novamente, a fim de obter uma melhor amostra da distribuição a posteriori.

Técnicas gráficas também são importantes para verificar a convergência do algoritmo de MCMC. Por exemplo, o gráfico das médias ergódicas para visualizar se o *burn-in* foi grande o suficiente (atingiu a medida estacionária da cadeia de Markov), o gráfico das auto-correlações para verificar se as amostras geradas são independentes e o gráfico da "série de tempo" das amostras geradas, verificando se existe uma aleatoriedade na geração. Aqui, analisamos a convergência do algoritmo de MCMC apenas através de técnicas gráficas.

```
#Gráficos de convergência
par(mfrow=c(3,3))
for (i in 1:length(fit$Posterior1[1,])) {
  plot(fit$Posterior1[,i],type="l",xlab="i",ylab="",
       main=colnames(fit$Posterior1)[i])
  plot(cumsum(fit.burn$Posterior1[,i])/
              seq(1,length(fit.burn$Posterior1[,i]),1),
       type="l",xlab="i",ylab="Média Ergódica",
       main=colnames(fit.burn$Posterior1)[i])
  acf(fit$Posterior1[,i],ylab="Corelação",
      main=colnames(fit$Posterior1)[i])
}
par(mfrow=c(1,1))
```

A Figura 2.27 apresenta os gráficos para a verificação da convergência do método. Pelos gráficos, entendemos que a amostra gerada é uma amostra aleatória da distribuição a posteriori dos parâmetros. Note que é importante gerar amostras distintas

com valores iniciais diferentes, para verificar se a convergência ocorre para as diferentes amostras, isto é, se a geração não é dependente da escolha do valor inicial. Além disso, com várias amostras distintas, é possível calcular algumas medidas da qualidade da convergência, como por exemplo, a estatística de Gelman-Rubin.



2.27 Figura. *Gráficos da qualidade da convergência do MCMC para o modelo Aranda-Ordaz.*

Uma vez gerada uma amostra da distribuição a posteriori, nós podemos estimar a proporção de besouros mortos por log(Dose). É comum o uso de estimador *plug-in* na inferência Bayesiana, entretanto, entendemos que esta não é a forma mais correta. O estimador *plug-in* é similar ao feito na análise frequentista com os estimadores de máxima verossimilhança. A ideia por trás do *plug-in* é de tomar uma estimativa pontual dos parâmetros (por exemplo, média a posteriori) e "plugar" os valores destas no modelo estatístico. Do ponto de vista do estimador de máxima verossimilhança este procedimento é correto, dada a propriedade da invariância do estimador. Entretanto, no caso Bayesiano, ao tomarmos o *plug-in* estamos deixando de lado toda a distribuição a posteriori, não aproveitando todo o potencial

do método. No código R abaixo, nós mostramos como calcular a média a posteriori da proporção de besouros mortos, bem como uma região de credibilidade (que é facilmente obtida). O resultado do código é dado na Figura 2.28.

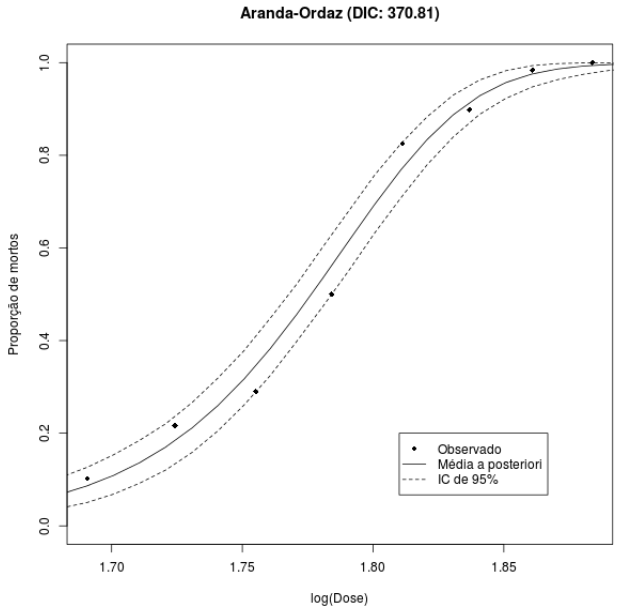
```
#Estimando a propoção de besouros mortos
aux <- seq(min(ldose)-0.5,max(ldose)+0.5,0.01)
eixo.y <- matrix(NA,length(fit$Posterior1[,1]),
                    length(aux))
for (i in 1:length(fit$Posterior1[,1])) {
  eixo.y[i,] <- pprentice(fit$Posterior1[i,1]+
                            fit$Posterior1[i,2]*aux,
                            c(fit$Posterior1[i,3],
                              fit$Posterior1[i,4]))
}

# Gráfico do modelo estimado
plot(ldose,mortos/expostos,ylim=c(0,1),pch=18,
     xlab="log(Dose)",ylab="Proporção de mortos",
     main=paste("Prentice (DIC: ",
                round(fit$DIC1[3],2),")",sep=""))
#média a posteriori
lines(aux,apply(eixo.y,2,mean))
#região de credibilidade
lines(aux,apply(eixo.y,2,quantile,probs=0.025),lty=2)
lines(aux,apply(eixo.y,2,quantile,probs=0.975),lty=2)

legend(1.81,0.2,lty=c(0,1,2),pch=c(18,NA,NA),
       c("Observado","Média a posteriori","IC de 95%"))
```

Note que, para utilizarmos a mediana a posteriori como estimador pontual, basta substituir `mean` por `median`, no comando `lines(aux, apply(eixo.y, 2, mean))`.

Como dado na Observação 2.26, a estimação dos outros modelos é bem simples. Em quase todos os casos, não foi necessário nenhuma alteração no tamanho do *burn-in* ou do salto, exceto no caso do modelo do Prentice. A Figura 2.29 mostra que o método de MCMC não convergiu. Os gráficos das médias ergódicas não estabilizam (ficam próximos de uma reta constante), a cadeia gerada apresenta uma tendência e existe uma alta correlação entre os pontos amostrais. Tentamos diferentes tamanho e saltos entre as amostras mas não conseguimos convergência, para este conjunto de dados, em nossas tentativas. Um dos motivos para isso é que, apesar



2.28 Figura. *Estimativa da proporção de mortes por log(Dose) utilizando o modelo Aranda-Ordaz.*

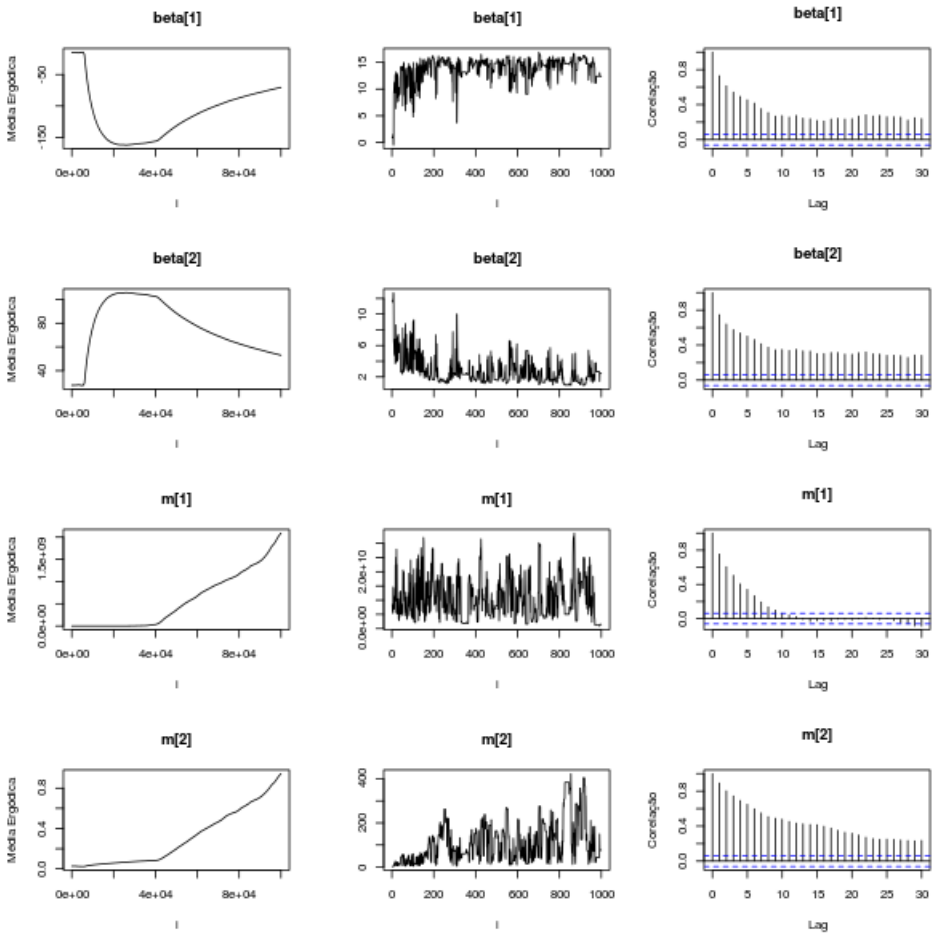
de ser um modelo muito interessante, a implementação computacional do modelo proposto por Prentice depende de algumas aproximações o que dificulta a estimação (inclusive a estimação por máxima verossimilhança).

A Tabela 2.30 apresenta o DIC dos modelos estimados, sendo que os modelos Weibull e Aranda-Ordaz foram os que mais próximos ficaram do modelo complementar log-log. Vale ressaltar que é conhecido na literatura que o modelo complementar log-log é um dos melhores, se não o melhor, modelo para estes dados. As figuras 2.31 a 2.35 apresentam as estimativas dos modelos complementar log-log, Weibull, Stukel, probito, logito.

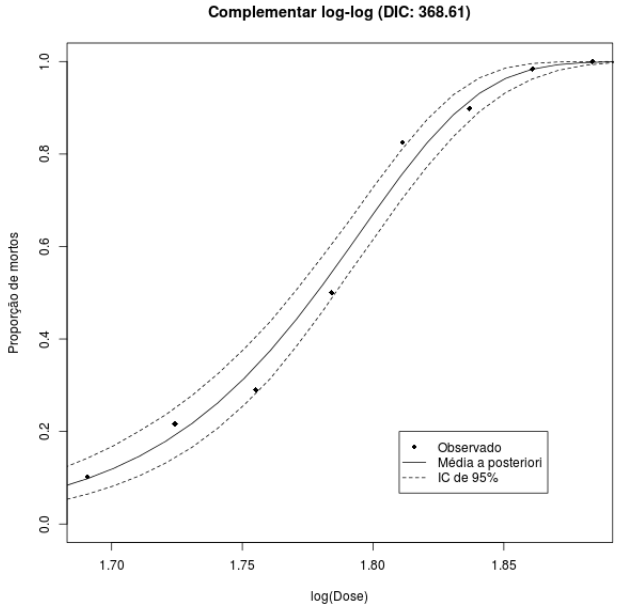
■

2.30 Tabela. *DIC dos modelos estimados.*

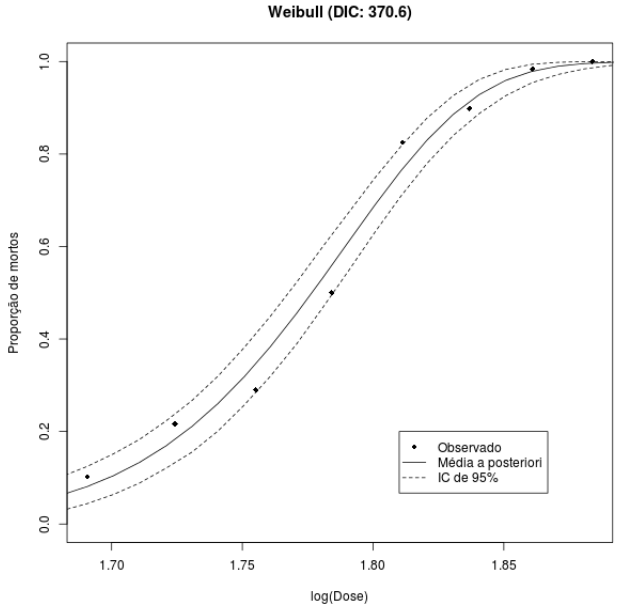
Modelo	DIC
Complementar log-log	368,61
Weibull	370,60
Aranda-Ordaz	370,81
Stukel	373,24
Probito	375,08
Logito	376,58



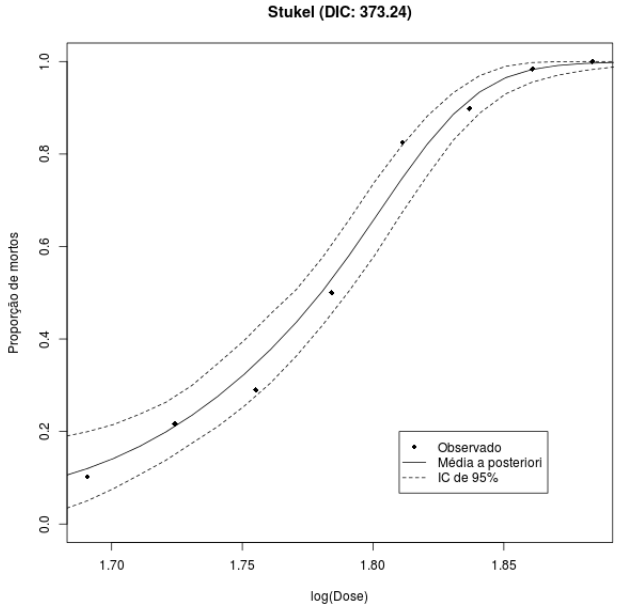
2.29 Figura. *Gráficos da qualidade da convergência do MCMC para o modelo Prentice (brun-in de* 100.000 *e salto de* 500*).*



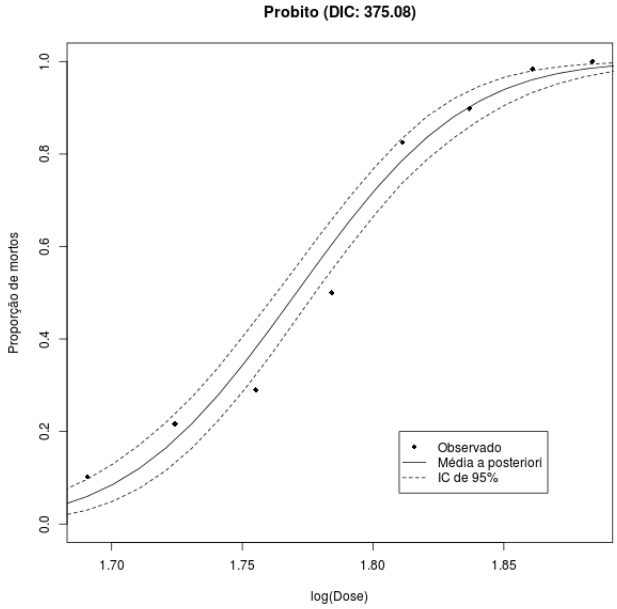
2.31 Figura. *Estimativa da proporção de mortes por log(Dose) utilizando o modelo complementar log-log.*



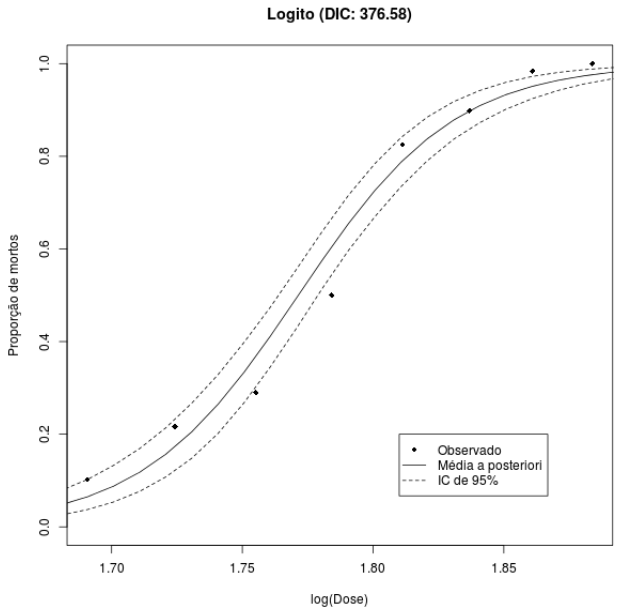
2.32 Figura. *Estimativa da proporção de mortes por log(Dose) utilizando o modelo Weibull.*



2.33 Figura. *Estimativa da proporção de mortes por log(Dose) utilizando o modelo Stukel.*



2.34 Figura. *Estimativa da proporção de mortes por log(Dose) utilizando o modelo probito.*



2.35 Figura. *Estimativa da proporção de mortes por log(Dose) utilizando o modelo logito.*

Capítulo 3

Análise de sobrevivência

Neste capítulo, apresentamos o modelo de regressão TBS (*Transform-Both-Sides*) paramétrico para dados não censurados ou com censura à direita (Seção 3.1). Na sequência, é apresentado um modelo de regressão para dados grupados (ou com censura intervalar) baseado na distribuição Weibull (Seção 3.2).

3.1 Modelo TBS

Seja T_i o tempo de sobrevivência do sujeito $i = 1, \ldots, n$ e seja $\boldsymbol{X_i}$ o vetor $(1, X_{1,i}, \ldots, X_{k,i})'$ de k covariáveis constantes no tempo mais o termo referente ao intercepto. O modelo TBS (Lin et al., 2012) assume que

$$g_\lambda(\log(T_i)) = g_\lambda(\boldsymbol{\beta X_i}) + \varepsilon_i, \tag{3.1}$$

em que $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ é um parâmetro de regressão, ε_i é um erro não especificado com densidade f_ε unimodal e simétrica, livre de covariáveis $\boldsymbol{X_i}$ e

$$g_\lambda(u) = \frac{\text{sign}(u)|u|^\lambda}{\lambda}, \tag{3.2}$$

para $\lambda > 0$, $\text{sign}(u) = 1$ se $u \geq 0$ e $\text{sign}(u) = -1$ se $u < 0$.

O modelo TBS é uma extensão da família poder de Box-Cox (Box and Cox, 1964), uma popular transformação para obter uma densidade simétrica e unimodal para a variável aleatória (transformada). Assume-se que a densidade f_ε dos erros ε_i na Equação (3.1) é centrada em zero. Lin et al. (2012) assumem uma distribuição a não-paramétrica para f_ε e um modelo semi-paramétrico para o tempo de sobrevivência. Aqui descrevemos apenas as distribuições paramétricas para f_ε, obtendo um modelo de regressão paramétrico. Na Tabela 3.3 são apresentadas algumas distribuições para o erro. Para cada distribuição, ξ denota seu parâmetro livre, a ser

estimado. Aqui foca-se nestas cinco distribuições. Entretanto, destaca-se que qualquer distribuição unimodal e simétrica em zero pode ser usada.

3.3 Tabela. *Distribuições para o erro.*

Distribuição	Parâmetro	Função de densidade ($f_\varepsilon(\epsilon \mid \xi)$)
Normal	$\xi = \sigma^2$	$(2\pi\sigma^2)^{-1/2} \exp\left\{-\epsilon^2/(2\sigma^2)\right\}$
ExpDupla	$\xi = b$	$(2b)^{-1} \exp\left\{-\lvert\epsilon\rvert/b\right\}$
t-Student	$\xi = \eta$ $(d.f.)$	$\frac{\Gamma((\eta+1)/2)}{\Gamma(\eta/2)\sqrt{\pi\eta}} \left(1 + \frac{\epsilon^2}{\eta}\right)^{-(\eta+1)/2}$
Cauchy	$\xi = c$	$\left[\pi c \left(1 + (\epsilon/c)^2\right)\right]^{-1}$
Logística	$\xi = s$	$\frac{\exp\{\epsilon/s\}}{s\left[(1+\exp\{\epsilon/s\})^2\right]}$

O espaço paramétrico para o parâmetro de todas as distribuições é $(0; +\infty)$.

Algumas características importantes do modelo são obtidas pelas funções de densidade, sobrevivência e taxa de falha. A Equação (3.1) pode ser reescrita como

$$\begin{aligned} \varepsilon_i &= g_\lambda(\log(T_i)) - g_\lambda(\boldsymbol{\beta}\boldsymbol{X_i}), \text{ ou equivalentemente como} \\ T_i &= \exp\left\{g_\lambda^{-1}\left[g_\lambda(\boldsymbol{\beta}\boldsymbol{X_i}) + \varepsilon_i\right]\right\}, \end{aligned} \tag{3.4}$$

em que a função inversa g_λ^{-1} é tal que $g_\lambda^{-1}(u) = \text{sign}(u)|\lambda u|^{\frac{1}{\lambda}}$. Note que esta formulação impede valores negativos para T_i, o que pode não ocorrer com uma transformação de Box-Cox se os estimadores de $(\boldsymbol{\beta}, \lambda, \xi)$ tiverem algum (amostra finita) viés (Fitzenberger et al., 2010). Considerando valores fixos para os parâmetros $(\boldsymbol{\beta}, \lambda, \xi)$, a distribuição de T_i é uma transformação da distribuição do erro. Pela Equação (3.4), nós obtemos as funções de densidade e sobrevivência de T_i, dadas por

$$f_T(t_i \mid \boldsymbol{X_i}, \lambda, \boldsymbol{\beta}, \xi) = t_i^{-1} |\log(t_i)|^{\lambda-1} f_\varepsilon\left(g_\lambda(\log(t_i)) - g_\lambda(\boldsymbol{\beta}\boldsymbol{X_i}) \mid \xi\right), \tag{3.5}$$

$$S_T(t_i \mid \boldsymbol{X_i}, \lambda, \boldsymbol{\beta}, \xi) = S_\varepsilon\left(g_\lambda(\log(t_i)) - g_\lambda(\boldsymbol{\beta}\boldsymbol{X_i}) \mid \xi\right). \tag{3.6}$$

Para o caso de dados não censurados ou com censura à direita, a verossimilhança do modelo é dada por

$$L(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D}) = \prod_{i=1}^{n} f_T(t_i \mid \boldsymbol{X_i}, \lambda, \boldsymbol{\beta}, \xi)^{\delta_i} S_T(t_i \mid \boldsymbol{X_i}, \lambda, \boldsymbol{\beta}, \xi)^{1-\delta_i} \tag{3.7}$$

em que δ_i é a indicadora de censura, isto é, $\delta_i = 1$ a quantidade t_i observada foi não censurada e $\delta_i = 0$ foi censurada à direita, $i = 1, \ldots, n$.

Outra função importante em análise de sobrevivência é a função quantílica. Defina $\epsilon_{(\alpha)}$ como o α-ésimo quantil de ε, isto é, $S_\varepsilon(\epsilon_{(\alpha)}) = 1-\alpha$. Para obter o α-ésimo quantil do tempo de sobrevivência, isto é, a função $t_{(\alpha)}$ tal que $S_T(t_{(\alpha)}) = 1-\alpha$, basta substituir ε_i na Equação (3.4) por $\epsilon_{(\alpha)}$ obtendo

$$t_{(\alpha)} = \exp\left\{g_\lambda^{-1}\left[g_\lambda(\boldsymbol{\beta X}) + \epsilon_{(\alpha)}\right]\right\}. \tag{3.8}$$

O tempo mediano de sobrevivência $t_{(0,5)}$ pode ser obtido da Equação (3.8) e pelo fato de que a distribuição do erro é simétrica em zero, isto é, $\epsilon_{(0,5)} = 0$. Logo,

$$\begin{aligned} t_{(0,5)} &= \exp\left\{g_\lambda^{-1}\left[g_\lambda(\boldsymbol{\beta X})\right]\right\} \\ &= e^{\boldsymbol{\beta X}} \ (= e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}). \end{aligned} \tag{3.9}$$

A Tabela 3.10 apresenta as funções quantis das cinco distribuições consideradas aqui.

3.10 Tabela. *Função quantil para a distribuição do erro.*

Distribuição	Parâmetros	$\epsilon_{(\alpha)}$
Normal	$\xi = \sigma^2$	$\sigma^2\Phi^{-1}(\alpha)$
ExpDupla	$\xi = b$	$-b\,\text{sign}(\alpha-1/2)\log(1-2\lvert\alpha-1/2\rvert)$
t-Student	$\xi = \eta\ (d.f.)$	$\Psi_\eta^{-1}(\alpha)$
Cauchy	$\xi = c$	$c\tan(\pi(\alpha-1/2))$
Logística	$\xi = s$	$s\log(\alpha/(1-\alpha))$

Φ^{-1} é a inversa da função distribuição da normal padrão, Ψ_η^{-1} é a inversa da distribuição t-Student (com η graus de liberdade).

Considere uma covariável binária X_1 tal que $X_1 = 1$ representa a presença de alguma característica, enquanto $X_1 = 0$ representa a ausência. Neste caso, $\boldsymbol{\beta}X = \beta_0 + \beta_1 X_1$, então podemos definir a *razão de medianas* O por

$$O = \frac{\text{mediana}(T \mid X_1 = 1)}{\text{mediana}(T \mid X_1 = 0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}, \tag{3.11}$$

sendo interpretada como "o tempo mediano de vida de sujeitos que apresentam a característica é O vezes maior do que o tempo mediano de vida dos sujeitos sem estas" (caso O seja menor que um, o tempo de vida será menor). Nesta situação, a

estimação do tempo mediano de vida não depende do parâmetro λ nem do parâmetro ξ da distribuição do erro. Esta é uma propriedade importante do modelo TBS, porque implica que, para qualquer escolha da distribuição do erro e do valor de λ, a interpretação dos parâmetros $\boldsymbol{\beta}$ é diretamente relacionada ao tempo mediano de sobrevivência. Na verdade, os parâmetros $\boldsymbol{\beta}$ podem ser vistos como logaritmos do tempo mediano de sobrevivência, o que facilita a inferência e ajuda na eliciação de prioris subjetivas.

Na estimação Bayesiana é necessário calcular a distribuição a posteriori de $(\lambda, \boldsymbol{\beta}, \xi)$. É considerado que os parâmetros são independentes a priori, isto é,

$$p(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D}) \propto L(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D}) p(\lambda) p(\boldsymbol{\beta}) p(\xi),$$

em que $L(\lambda, \boldsymbol{\beta}, \xi \mid \mathcal{D})$ é a função de verossimilhança dada na Equação (3.7) e $p(\lambda)$, $p(\boldsymbol{\beta})$ e $p(\xi)$ são as densidades a priori.

Baseado nas características da transformação g_λ, os valores mais razoáveis para λ estão no intervalo $(0; 3)$ e então sugere-se o uso de uma distribuição a priori com grande densidade para valores dentro deste intervalo e que seja decrescente para valores maiores que 3. Considerando que o parâmetro ξ da distribuição do erro é diretamente relacionado com a variância (segundo as cinco distribuições utilizadas aqui), sugere-se que a distribuição a priori favoreça pontos no intervalo $(0; 2)$, uma vez que não é esperado que a distribuição do erro tenha uma variância tão grande.

Como mencionando antes, uma propriedade importante do modelo TBS é a interpretação dos parâmetros $\boldsymbol{\beta}$ em termos de valores medianos e razões de medianas. Por esta razão, a eliciação de uma priori subjetiva para $\boldsymbol{\beta}$ é geralmente uma tarefa simples. Basta perguntar para um especialista alguns quantis do tempo de sobrevivência, por exemplo. Usando as equações (3.9) e (3.11), é possível "traduzir" a informação dos quantis para a priori. No pacote `TBSSurvival`, é utilizado a distribuição normal como priori para $\boldsymbol{\beta}$, mas é deixado para o usuário a opção de escolher a média e a variância.

O alvo aqui pode ser definido como `tempo`, um vetor de tempos de falha (sobrevivência) de componentes (pacientes) e `delta`, uma variável indicadora do evento, isto é, `delta` é igual a um se o evento ocorreu naquele tempo ou zero no caso do evento ter censura à direita.

3.12 Exemplo. Neste exemplo é gerado um conjunto de dados com tempos censurados à direita. O problema está relacionado a um experimento com 30 máquinas. Observa-se o tempo de falha destas máquinas, entretanto, o experimento tem um tempo total de 6 semanas. Depois disso, todas as máquinas que não haviam falhado ainda têm seus tempos de falha censurados. Geramos o tempo de falha de uma variável aleatória com distribuição gama.

```
set.seed(2360873) # recomenda-se o uso do set.seed para
                  # o leitor obter resultados similares
                  # aos do livro.

#gerando os dados
tempo <- pmin(rgamma(30,10,2),rep(6,30))
delta <- rep(1,30)
for (i in 1:30) {
  if (tempo[i] == 6) delta[i] <- 0
}
dados <- cbind(tempo,delta)
```

■

Quando se faz uma análise Bayesiana é necessário escolher a distribuição à priori, bem como a convergência do algoritmo de MCMC, no qual a estimação do modelo TBS é baseada. A estimação do modelo TBS está implementada no pacote `TBSSurvival` para `R`. Utilizamos a função `tbs.survreg.be` para a geração da amostra da distribuição a posteriori, que é baseada o algoritmo de Metropolis-Hasting. O modelo pode ser facilmente estimado por

```
#Estimacao do modelo
fit <- tbs.survreg.be(Surv(dados[,1],dados[,2]) ~ 1,
                      dist=dist.error("norm"),
                      burn=100000,
                      jump=1000,
                      size=1000,
                      scale=0.06)
```

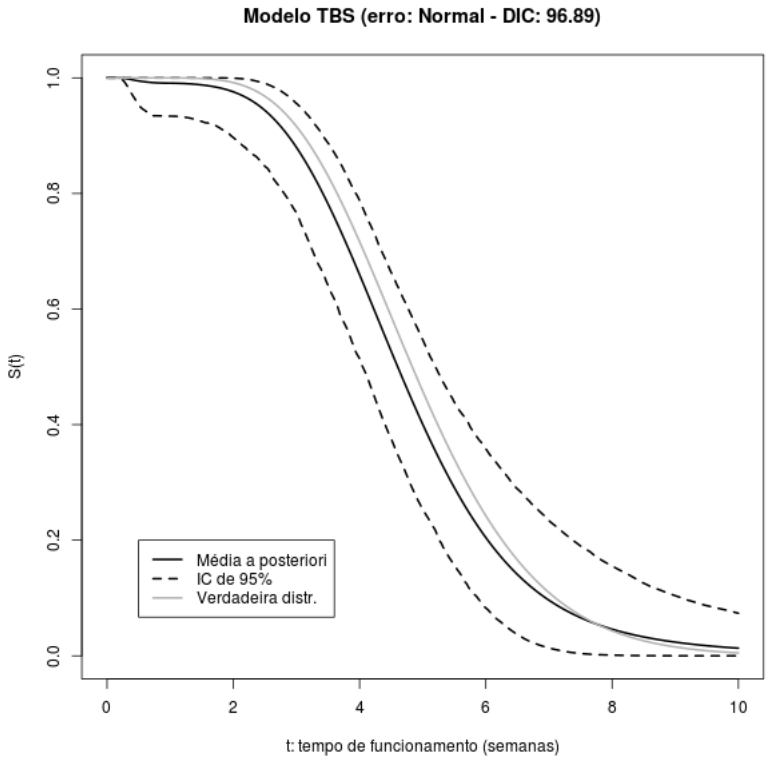
em que a estrutura da fórmula `Surv(data[,1],data[,2]) ~ 1` segue o mesmo padrão do pacote `survival` (Therneau, 2013), `dist = dist.error("norm")` é a distribuição do erro escolhida (para as outras opções basta trocar `"norm"` por `"t"`, `"doubexp"`, `"cauchy"` ou `"logistic"`), `burn = 1000` é o tamanho do *burn-in*, `jump = 10` é a quantidade de amostras descartadas a cada amostra "escolhida", `size=1000` é o tamanho da amostra final e `scale=0.06` é um fator de escala escolhido pelo usuário para controlar a taxa de aceitação do algoritmo de Metropolis-Hasting.

Para obter uma estimativa da função de sobrevivência, basta utilizar o comando `plot` do modelo estimado. Note que a região de credibilidade é automaticamente desenhada.

```
#gráfico da função de sobrevivência
plot(fit,ylab="S(t)",xlim=c(0,10),
     xlab="t: tempo de funcionamento (semanas)",
     main=paste("Modelo TBS (erro: Normal - DIC: ",
                round(fit$DIC,2),")",sep=""),col=1,
     lwd=2,lty=1,lwd.HPD=2,lty.HPD=2,col.HPD=1)

#distribuicao teórica
lines(seq(0,10,0.01),1-pgamma(seq(0,10,0.01),10,2),
      lwd=2,col="gray70")
legend(0.5,0.2,lty=c(1,2,1),col=c(1,1,"gray70"),
       lwd=c(2,2,2),c("Média a posteriori","IC de 95%",
                      "Verdadeira distr."))
```

A curva estimada do modelo TBS com distribuição do erro normal é dada na Figura 3.13.



3.13 Figura. *Estimativa do Modelo TBS com erro normal e curva teórica Gama(10; 2).*

Os códigos para a estimação em R do modelo TBS com as outras distribuições de erro seguem de forma análoga, sendo necessário apenas trocar o nome da dis-

tribuição. Na Tabela 3.14, apresentamos a comparação dos modelos, em que pelo critério DIC consideramos que o modelo TBS com distribuição do erro normal foi o que apresentou melhor ajuste.

3.14 Tabela. *DIC dos modelos estimados.*

Distr. do erro	DIC
Normal	96,89
Logística	97,20
ExpDupla	97,59
Cauchy	101,51
t-Student	105,90

Detalhes sobre as estimativas dos parâmetros podem ser obtidas executando o comando `summary(fit)`, obtendo a média, desvio padrão e intervalo de credibilidade a posteriori dos parâmetros do modelo, algumas medidas descritivas da média a posteriori do erro e as médias a posteriori dos percentis 5%, 25%, 50%, 75% e 95% do tempo de vida. Estes resultados são apresentados abaixo.

```
----------------------------------------------------
 TBS model with norm error distribution (BE).

              Mean  Std. Dev.      95% HPD CI
  lambda:    1.5321     0.8085 (0.0577, 2.8944)
      xi:    0.2173     0.1838 (0.0354, 0.5163)
    beta:    1.5271     0.0674 (1.4014, 1.6503)

Summary statistic of the posterior mean of the error
for the TBS model:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.83590 -0.36140 -0.02158 -0.04674  0.19460  0.37030

Estimated quantiles of time event:
    5%    25%    50%    75%    95%
2.2708 3.5381 4.6046 5.8602 8.0632
----------------------------------------------------
```

3.2 Dados com censura intervalar

Aqui, descrevemos o uso do modelo Weibull em análise de sobrevivência para dados com censura intervalar ou dados grupados (Sun, 2006). Uma boa referência sobre a distribuição Weibull é o livro do Rinne (2009). Seja $\boldsymbol{X} = (X_1, \ldots, X_k)'$ a matriz de experimento e T o tempo de vida (variável resposta). Aqui, temos a restrição de que a observação de $T \in (A, B)$. Uma amostra de T e $\boldsymbol{X}$ pode ser escrita como $T_i \in (A_i, B_i)$, $\boldsymbol{X}_i = (X_{11}, \ldots, X_{ki})'$, $i = 1, \ldots, n$. A função de verossimilhança para dados com censura intervalar, do modelo Weibull é

$$\begin{aligned} L(\boldsymbol{\beta}, \gamma \mid \mathcal{D}) &= \prod_{i=1}^{n} \Pr(T_i \in (A_i, B_i) \mid \mathcal{D}) \\ &= \prod_{i=1}^{n} \left\{ \exp\left[-\left(\frac{A_i}{\boldsymbol{\beta}\boldsymbol{X}_i} \right)^{\gamma} \right] - \exp\left[-\left(\frac{B_i}{\boldsymbol{\beta}\boldsymbol{X}_i} \right)^{\gamma} \right] \right\}, \end{aligned}$$

em que $\mathcal{D} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{X}\}$, $\boldsymbol{A} = (A_1, \ldots, A_n)$, $\boldsymbol{B} = (B_1, \ldots, B_n)$, $\gamma > 0$ (parâmetro de forma), $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$, $\boldsymbol{\beta}\boldsymbol{X}_i > 0$ (parâmetro de escala).

A função de sobrevivência para uma característica ($\boldsymbol{X}$) da população é dada por

$$S(t \mid \boldsymbol{X}, \boldsymbol{\beta}, \gamma) = \exp\left\{ -\left(\frac{t}{\boldsymbol{\beta}\boldsymbol{X}} \right)^{\gamma} \right\}.$$

Considerando que a esperança do tempo de sobrevivência é $\mathrm{E}(T \mid \boldsymbol{X}, \boldsymbol{\beta}, \gamma) = (\boldsymbol{\beta}\boldsymbol{X})\Gamma(1 + 1/\gamma)$, em que Γ é a função gama ($\Gamma(x) = \int_0^{\infty} u^{x-1} e^{u} \mathrm{d}t$), então a única diferença no valor da esperança do tempo de sobrevivência para diferentes valores de $\boldsymbol{X}$ é dada por $\boldsymbol{\beta}\boldsymbol{X}$. Neste caso, $\boldsymbol{\beta}\boldsymbol{X}$ pode ser visto como um regressor linear do tempo médio de sobrevivência.

A análise Bayesiana é baseada em um algoritmo de MCMC para geração de uma amostra da densidade a posteriori $p(\boldsymbol{\beta}, \gamma \mid \mathcal{D}) = L(\boldsymbol{\beta}, \gamma \mid \mathcal{D})\ p(\boldsymbol{\beta}, \gamma)$, em que $p(\boldsymbol{\beta}, \gamma)$ é a densidade conjunta a priori dos parâmetros do modelo.

3.15 Exemplo. Foi realizado um experimento com 60 caramujos, os quais foram pré expostos a quatro diferentes tratamentos (A, B, C e D), em que 30 caramujos foram pré expostos ao tratamento A (considerado como grupo controle), 10 ao tratamento B, 10 ao tratamento C e 10 ao tratamento D. Depois disso, os caramujos foram expostos à uma temperatura letal. O objetivo do estudo é de investigar qual tratamento de pré exposição resultará em maior tempo de vida dos caramujos após a exposição à temperatura letal.

Os caramujos foram observados a cada hora e contou-se quantos haviam morrido. Neste caso, o tempo exato de morte é desconhecido. A informação disponível é censurada por intervalos de uma hora. Os dados são apresentados na Tabela 3.16.

3.16 Tabela. *Número de caramujos sobreviventes após exposição a temperatura letal.*

	Número de caramujos vivos			
Horas	Trat. A	Trat. B	Trat. C	Trat. D
1	30	10	10	10
2	25	10	10	10
3	13	9	10	10
4	1	5	8	10
5	0	2	7	7
6	0	0	6	5
7	0	0	5	1
8	0	0	4	0
9	0	0	2	0
10	0	0	0	0

Temos quatro diferentes grupos para análise. Neste caso, construímos três variáveis binárias para representar estes grupos, $\boldsymbol{X} = (X_1; X_2; X_3)'$. O grupo pré exposto ao tratamento A (controle) é representado por $\boldsymbol{X} = (0; 0; 0)'$; o grupo pré exposto ao tratamento B é representado por $\boldsymbol{X} = (1; 0; 0)'$; o grupo pré exposto ao tratamento C é representado por $\boldsymbol{X} = (0; 1; 0)'$; e o grupo pré exposto ao tratamento D é representado por $\boldsymbol{X} = (0; 0; 1)'$.

O próximo passo consiste na escolha da distribuição a priori para os parâmetros. Aqui, nós consideramos prioris independentes para os parâmetros do modelo, sendo que a distribuição a priori para γ foi uma uniforme no intervalo $(0; 50)$ e para os parâmetros β_j, $j = 0, \ldots, 3$, escolhemos distribuições normais independentes com média zero e variância 400. Neste caso, não são consideradas prioris não-informativas, mas são prioris "vagas", isto é, estamos considerado "pouca" informação a priori, uma vez que a variância a priori é razoavelmente grande. A Tabela 3.17 apresenta as estimativas dos parâmetros e a evidência da hipótese nula (H_0) do parâmetro ser igual a zero. Note que, tanto o parâmetro γ quanto o parâmetro β_0 não podem ser iguais a zero, pela definição do modelo. Neste caso, não são calculadas as medidas de evidência referentes a estes dois parâmetros.

Considere que $E_0 = \mathrm{E}(T \mid \boldsymbol{X} = (0; 0; 0), \mathcal{D})$, $E_1 = \mathrm{E}(T \mid \boldsymbol{X} = (1; 0; 0), \mathcal{D})$, $E_2 = \mathrm{E}(T \mid \boldsymbol{X} = (0; 1; 0), \mathcal{D})$ e $E_3 = \mathrm{E}(T \mid \boldsymbol{X} = (0; 0; 1), \mathcal{D})$, isto é, E_0 é a esperança a posteriori do grupo pré exposto ao tratamento A, E_1 para o grupo pré exposto ao tratamento B, E_2 para o grupo pré exposto ao tratamento C e E_3 para o grupo pré exposto ao tratamento D. A Tabela 3.18 apresenta as estimativas para estas quantidades.

3.17 Tabela. *Estimativa dos parâmetros (medidas à posteriori)*

parâmetro	média	dp	L_I (95%)	L_S (95%)	$Ev(H_0)$
γ	4,507	0,520	3,501	5,534	–
β_0	3,100	0,145	2,821	3,387	–
β_1	1,388	0,384	0,668	2,151	$\approx$ 0,005
β_2	4,716	0,622	3,580	5,981	$<$ 0,001
β_3	3,100	0,497	2,117	4,068	$<$ 0,001

dp: desvio padrão;

Limite inferior (L_I) e superior (L_S) do intervalo de credibilidade de 95% de máxima densidade à posteriori;

$Ev(H_0)$ evidência de H_0: parâmetro igual a zero.

3.18 Tabela. *Esperança a posteriori do tempo de sobrevivência para cada grupo.*

	média	dp	L_I (95%)	L_S (95%)	L_I (99%)	L_S (99%)
E_0	2,828	0,136	2,547	3,080	2,489	3,205
E_1	4,094	0,329	3,439	4,729	3,292	5,012
E_2	7,132	0,560	6,005	8,195	5,827	8,833
E_3	5,656	0,433	4,798	6,492	4,648	6,873

dp: desvo padrão;

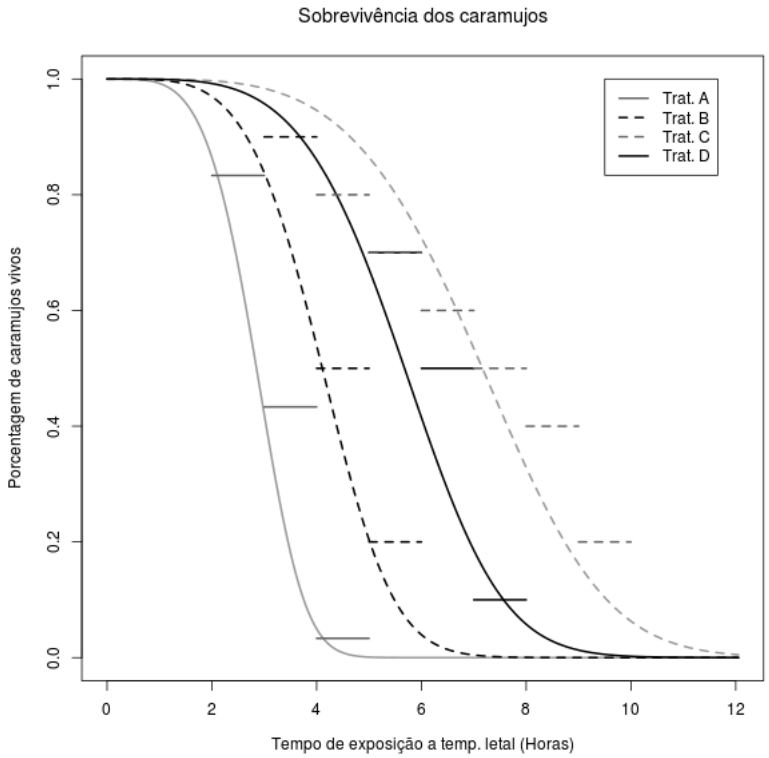
Limite inferior (L_I) e superior (L_S) do intervalo de credibilidade de máxima densidade a posteriori.

Também foram calculadas algumas probabilidades a respeito da ordenação das esperanças a posteriori, dadas na Tabela 3.19. Vemos que a alta probabilidade a posteriori de que $E_0 < E_1 < E_3 < E_2$ é de 0,982, o que significa que o grupo pré exposto ao tratamento A tem o menor tempo médio de sobrevivência, seguido em sequência pelo grupo pré exposto ao tratamento B, depois pelo grupo pré esposto ao tratamento D e, por fim, o grupo com maior esperança a posteriori é o grupo pré exposto ao tratamento C. Desta forma, conclui-se que todas os tratamentos (B, C e D) de pré exposição "aumentaram" o tempo de vida dos caramujos em relação ao grupo controle (tratamento A), sendo que o grupo com maior tempo de vida esperado é o grupo de pré exposição ao tratamento C. Na Figura 3.20, apresentamos as funções de sobrevivência estimada para cada tratamento.

■

3.19 Tabela. *Algumas probabilidades a posteriori referentes a ordenação das esperanças do tempo de sobrevivência para cada grupo.*

$\Pr[E_0 < \min(E_1, E_2, E_3)] \approx 1$	$\Pr[E_1 < \min(E_2, E_3)] \approx 0,999$
$\Pr[E_1 < \min(E_0, E_2, E_3)] \approx 0$	$\Pr[E_2 < \min(E_1, E_3)] \approx 0$
$\Pr[E_2 < \min(E_0, E_1, E_3)] \approx 0$	$\Pr[E_3 < \min(E_1, E_2)] \approx 0,001$
$\Pr[E_3 < \min(E_0, E_1, E_2)] \approx 0$	
$\Pr[E_2 < E_3] \approx 0,017$	$\Pr[E_0 < E_1 < E_3 < E_2] \approx 0,982$



3.20 Figura. *Estimativa do modelo Weibull da função de sobrevivência dos caramujos para cada um dos quatro tratamentos (as barras horizontais representam a proporção de caramujos ainda vivos para o intervalo de uma hora).*

Referências Bibliográficas

F. J. Aranda Ordaz. On two families of transformations to additivity for binary response data. *Biometrika*, 68:357–363, 1981.

D. Basu. Statistical information and likelihood. *Sankhya, Ser. A.*, 37:1–71, 1975.

D. Blackwell. *Estatística Básica (Basic Statistic)*. McGraw-Hill do Brasil, São Paulo, 1973. Traduzido por Pereira, C.A. de B. e Borges, W.

C.I. Bliss. The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1):134–167, 1935. doi: 10.1111/j.1744-7348.1935.tb07713.x.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Ser. B*, 26:211–243, 1964.

R. Caron and A. Polpo. Binary data regression: Weibull distribution. *AIP Conference Proceedings*, 1193(1):187–193, 2009. doi: 10.1063/1.3275613.

M-H. Chen, D. K. Dey, and Q-M. Shao. A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94(448): 1172–1186, December 1999. URL `http://www.jstor.org/stable/2669933`.

A. DasGupta, editor. *Selected works of Debabrata Basu*. Selected Works in Probability and Statistics. Springer, 2011.

M. DeGroot. *Probability and Statistics*. Addison-Wesley, New York, 2nd edition, 1975.

D. K. Dey, S. K. Sujit K. Ghosh, and B. K. Mallick, editors. *Generalized Linear Models: A Bayesian Perspective*. Biostatistics Series. CRC Press, 2000.

M. A. Diniz, C. A. B. Pereira, A. Polpo, J. Stern, and S. Wechesler. Relationship between Bayesian and frequentist significance indices. *International Journal for Uncertainty Quantification*, 2(2):161–172, 2012. doi: 10.1615/Int.J.UncertaintyQuantification.2012003647.

B. dos Santos, A. Polpo, and C. A. de B. Pereira. *binreg R package*, 2013. URL `http://code.google.com/p/binreg/`.

B. Fitzenberger, R. A. Wilke, and X. Zhang. Implementing Box-Cox Quantile Regression. *Econometric Reviews*, 29(2):158–181, 2010.

D. Gamerman. *Markov Chain Monte Carlo Stochastic Simulation for Baeysian Inference*. Chapman & Hall, London, 1997.

Andrew Gelman and Yu-Sung Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2013. URL `http://CRAN.R-project.org/package=arm`. R package version 1.6-06.01.

J. Lin, D. Sinha, S. Lipsitz, and A. Polpo. Semiparametric Bayesian survival analysis using models with log-linear median. *Biometrics*, 2012. *to appear*.

Andrew D. Martin, Kevin M. Quinn, and Jong Hee Park. MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, 42(9):22, 2011. URL `http://www.jstatsoft.org/v42/i09/`.

J. A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Ser. A*, 135:370–384, 1972.

C. A. de B. Pereira and M. A. G. Viana. *Elementos de inferência Bayesiana*. Associação Brasileira de Estatística, 1982. 5o. Simpósio Nacional de Probabilidade e Estatística.

C. A.B. Pereira, J.M. Stern, and S. Wechsler. Can a significance test be genuinely Bayesian? *Bayesian Analysis*, 3(1):19–100, 2008.

C.A.B. Pereira and J. Stern. Special characterizations of standard discrete models. *REVSTAT - Statistical Journal*, 6(3):199–230, 2008.

C.A.B. Pereira and J.M. Stern. Evidence and credibility: a full Bayesian test of precise hypothesis. *Entropy*, 1:104–115, 1999.

R.L. Prentice. Generalization of the probit and logit models. *Biometrics*, 32:761–768, 1976.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org`.

J. Richardson. The analysis of 2 x 1 and 2 x 2 contingency tables: an historical review. *Statistical Methods in Medical Research*, 3(2):107–133, 1994.

H. Rinne. *The Weibull Distribution: A Handbook*. CRC Press, 2009.

LLC. Statisticat. *LaplacesDemon: Complete Environment for Bayesian Inference*. CRAN, 2013. URL `http://cran.r-project.org/web/packages/LaplacesDemon/index.html`. R package version 13.03.04.

R.B. Stern and C. A. de B. Pereira. Statistical information: A bayesian perspective. *Entropy 14*, 14(11):2254–2264, 2012. doi: 10.3390/e14112254.

T.A. Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.

J. Sun. *The Statistical Analysis of Interval-Censored Failure Time Data*. Number XVI in Statistics for Biology and Health. Springer, New York, 2006.

Terry Therneau. *A Package for Survival Analysis in S*, 2013. URL `http://CRAN.R-project.org/package=survival`. R package version 2.37-4.