Modelos bayesianos para dados categorizados com censura

Rogério Antonio de Oliveira

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, maio de 2009

Modelos bayesianos para dados categorizados com censura

Este exemplar corresponde à redação
final da tese devidamente corrigida
e defendida por Rogério Antonio de Oliveira
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof. Dr. Carlos Alberto de Bragança Pereira (orientador) - IME-USP.
- Prof. Dr. Fábio Prates Machado - IME-USP.
- Prof. Dr. Marcelo de Souza Lauretto - EACH - USP.
- Profa. Dra. Verônica Andrea Gonzalez Lopez - UNICAMP.
- Prof. Dr. Francisco Louzada Neto - UFSCar.

Agradecimentos

Agradeço a Deus por ter sempre me proporcionado forças nos momentos em que precisei, permitindo que eu tivesse acesso a chave do conhecimento.

Agradeço a minha família por estar ao meu lado em todos os momentos, incentivando-me a seguir sempre em frente apesar dos obstáculos. Agradeço especialmente a minha mãe Elisabete por fazer de tudo para minimizar as dificuldades encontradas durante esta longa caminhada.

Agradeço imensamente ao Prof. Carlinhos por seus ensinamentos e por ter me ajudado a vencer mais esta etapa da minha vida.

Agradeço a todos os professores da Banca por contribuirem na conclusão deste trabalho.

Agradeço a todos os professores do IME que colaboraram para minha formação, de forma direta ou indireta, por meio de incentivos para continuar a caminhada.

Agradeço a todos os amigos que me acompanharam durante essa etapa da minha vida, principalmente Graciella, Rosemeire, Miriam, Clécio, Raydonal, Luz e as novas amizades: Mirtes, pelos momentos de descontração, felicidade e também pelas trocas de conhecimento, favorecendo meu crescimento e amadurecimento profissional; Carol, por ter acreditado em mim e também pelas palavras de encorajamento nos momentos que necessitei. Agradeço também ao César por ter colaborado com algumas dicas para resolver os problemas computacionais.

Agradeço aos funcionários do IME que sempre me auxiliaram nas minhas atividades, aos funcionários da secretaria de pós-graduação, especialmente ao Pinho, por me ajudar com as documentações necessárias em cada etapa.

Agradeço ao Capes por ter me concedido a bolsa para a execução deste trabalho.

Muito obrigado, Rogério Antonio de Oliveira

Resumo

Dados categorizados com censura são comuns em diferentes áreas do conhecimento. A análise estatística para esse tipo de dados era um desafio para muitos estatísticos. Atualmente há algumas abordagens Bayesianas e frequentistas para resolver este problema. Paulino e Pereira (1995) apresentaram um modelo Bayesiano para dados categorizados com padrão de censura. Este modelo utiliza todas as informações da amostra, inclusive as observações com informações parciais ou totalmente faltantes. Também serão apresentados o modelo de Paulino, Soares e Neuhaus (2003) que emprega modelos lineares generalizados para modelar o efeito da variável explicativa, discreta ou contínua, na variável resposta discreta que está sujeita a erro e um modelo de regressão logística utilizando a idéia da partição. Para este tipo de análise, pode-se aplicar uma visão estatística por meio do *the Full Bayesian Significance Test* ($FBST$) como um teste de significância Bayesiano coerente, que foi originalmente proposto por Pereira e Stern (1999). O teste $FBST$ é um teste baseado no valor do conceito de evidência. O objetivo deste trabalho é apresentar a aplicação do $FBST$ para a análise de dados discretos com censura. Alguns exemplos ilustrativos são apresentados como aplicações da metodologia Bayesiana.

Palavras-chave: dados discretos, análise Bayesiana, regressão logística.

Abstract

The categorical data with missing values are common in different area of knowledge. The statistical analyses of those data were a challenge to many statisticians. Nowadays there are some Bayesian and frequentist approaches to solve it. Paulino and Pereira (1995) presented a Bayesian model for categorical data with missingness. That model uses all information from the sample, including the observations that have partial or completely missing. The generalized linear model given by Paulino, Soares e Neuhaus (2003) and a new logistic model using the principle of partial likelihood will be also presented. Using the Bayesian view, the Full Bayesian Significance Test ($FBST$), presented by Pereira and Stern (1999), can be applied as a coherent Bayesian significance test. The $FBST$ is a test based on a value of evidence concept. The goal of this work is to present the application of the $FBST$ for categorical data with missingness. Some examples are given as illustration for the Bayesian analysis.

Keywords: discrete data, Bayesian analysis, logistic regression.

Sumário

Lista de Abreviaturas ix

Lista de Símbolos xi

Lista de Figuras xiii

Lista de Tabelas xv

1 Introdução 1

1.1 Considerações Preliminares 1

1.2 Objetivos 2

1.3 Contribuições 2

1.4 Organização do Trabalho 2

1.5 Exemplo - Teste de Coloração Dentária 2

1.6 Enfoques Estatísticos para o Problema 3

1.7 O uso do $FBST$ 5

1.8 Dados Categorizados com Censura 5

1.8.1 Modelo Bayesiano de Paulino e Pereira (1995) 5

1.9 Modelo Bayesiano Particionado 8

1.10 Full Bayesian Significance Test ($FBST$) 10

1.11 Definição do $FBST$ 11
1.12 Calculando o $FBST$ 12
1.12.1 Otimização 12
1.12.2 Integração 12
1.12.3 Motivação 15
1.13 Aplicação do FBST 16
1.13.1 Testes de Coloração Dentária 16

2 Modelo de Regressão Logística **21**
2.1 Regressão Logística 21
2.2 Regressão Logística com censura 24
2.2.1 Modelo de Paulino, Soares e Neuhaus (2003) 24
2.2.2 Modelo Bayesiano particionado com censura 28
2.2.3 Aplicação dos Modelos Bayesianos 29
2.3 Modelo Dose Resposta 31

3 Conclusões **39**
3.1 Considerações Finais 39
3.2 Sugestões para Pesquisas Futuras 39

A Programas **41**

Referências Bibliográficas **49**

Lista de Abreviaturas

FBST	Full Bayesian Significance Test ;
ev	Valor aproximado da evidência do FBST;
T_p	Teste padrão;
T_s	Teste simplificado;
FB	Fator de Bayes.

Lista de Símbolos

θ	Vetor de probabilidades;
λ	Vetor de probabilidades condicionais das censuras;
$M_m(t, \theta)$	Distribuição Multinomial m-variada de tamanho t e vetor de probabilidades θ;
$DM_m(a)$	Distribuição Dirichlet m-variada com parâmetros a.

Lista de Figuras

2.1 Proporção de células sem micronúcleos. 36
2.2 Proporção de células com um micronúcleo. 37
2.3 Proporção de células com dois ou mais micronúcleos. 38

Lista de Tabelas

1.1 Freqüências observadas da suceptibilidade dentária de 97 voluntários 3

1.2 Dados hipotéticos de presença de uma doença. 15

1.3 Valor aproximado da evidência *ev* de um estudo de simulação 18

2.1 Dados Infecção do HPV 30

2.2 Valor aproximado da evidência *ev* dos modelos logísticos 31

2.3 Frequências de micronúcleos em doses de radiação 32

2.4 Parâmetros da distribuição *a priori* e *a posteriori* 33

2.5 Parâmetros da distribuição a posteriori de $(\theta_{i1}, \theta_{i2})$ 34

2.6 Parâmetros da distribuição a posteriori de $(\theta_{i1}, \theta_{i2})$ 35

Capítulo 1

Introdução

1.1 Considerações Preliminares

Dependendo da área do conhecimento e do objetivo da pesquisa científica, pode ser necessário coletar algumas informações relacionadas à classificação de unidades amostrais de acordo com os valores ou intervalos de valores de uma ou mais variáveis aleatórias, discretas ou contínuas. Os dados relativos a estas variáveis aleatórias formam um conjunto de dados, que são freqüentemente denominados dados discretos ou categorizados. Esta denominação é utilizada porque geralmente existe a contagem de eventos ou de unidades amostrais, cujas características são determinadas por categorias de interesse. As freqüências encontradas devido a classificação cruzada destas informações são geralmente apresentadas na forma de tabelas, as chamadas tabelas de contingência.

Durante a coleta de dados, podem ocorrer algumas perdas de informação relacionadas às unidades amostrais estudadas de forma aleatória, pois os valores de uma ou mais variáveis de categorização não são registradas. As tabelas de contingência portadoras de dados omissos apresentam informações completamente coletadas e, dependendo da variável estudada, apresentam também uma ou mais marginais de categorias com dados incompletos, pois algumas informações foram parcialmente coletadas ou totalmente omissas. Estes problemas podem ocorrer quando o pesquisador, por desconhecimento ou descuido, registra apenas o subconjunto de valores da varíavel ao invés de registrar os valores das variáveis discretas, pois a característica ou atributo estudado pode não ser distinguível entre si.

1.2 Objetivos

Este trabalho tem por objetivo propor novos modelos Bayesianos para análise de dados categorizados com censuras e utilizar o procedimento do *FBST* para apresentar a aplicação da análise estatística em conjunto de dados reais.

1.3 Contribuições

As principais contribuições deste trabalho estão discriminadas abaixo:

- Um modelo Bayesiano particionado é proposto para utilizar inclusive as informações das censuras observadas nos dados coletados;
- Um novo modelo Bayesiano de regressão logística é apresentado para avaliar os efeitos das variáveis explicativas, discretas ou contínuas, na variável com resposta binárias, na presença de censuras.

1.4 Organização do Trabalho

Nas próximas seções deste capítulo, alguns conceitos importantes serão abordados sobre o procedimento do *FBST*. Apresenta-se o modelo Bayesiano de Paulino e Pereira e também o modelo Bayesiano particionado. Um exemplo clássico de dados deficientemente categorizados é apresentado como motivação na próxima seção.

No Capítulo 2, abordam-se alguns modelos de regressão logística para análise Bayesiano de dados categorizados. Apresenta-se um novo modelo Bayesiano para avaliar os efeitos das variáveis explicativas, discretas ou contínuas, na variável com resposta binárias, na presença de censuras.

Finalmente, no Capítulo 3, são discutidas algumas conclusões sobre o trabalho e também são apresentadas algumas perspectivas de trabalhos futuros.

1.5 Exemplo - Teste de Coloração Dentária

Como motivação da pesquisa, considere os dados da Tabela 1.1 apresentados e analisados por Paulino e Pereira (1995). Este conjunto de dados também foi analisado nos trabalhos de Soares

(2004) e Poleto (2006). Os dados coletados são referentes a uma amostra composta por 97 voluntários. Aplicou-se dois testes de susceptibilidade à cárie dentária categorizada em três níveis: baixa, média ou alta. Os dois testes estudados apresentam algumas características diferentes: um representa o teste padrão e que é bastante confiável porém caro e o outro é um teste simplificado, mais barato, que se baseia na observação da coloração obtida na reação de um produto com a saliva do paciente. Em alguns casos, existe a possibilidade de dúvidas em relação a coloração do teste simplificado, o profissional pode não ser capaz de discernir entre níveis adjacentes de susceptibilidade dentária. Essa dúvida é capaz de criar um mecanismo de omissão, que pode ocorre porque existe um certo confundimento entre as categorias próximas. O principal objetivo de realizar este experimento consistia em avaliar a possibilidade da substituição do teste padrão (T_P) pelo teste simplificado (T_S).

Tabela 1.1: Freqüências observadas da suceptibilidade dentária de 97 voluntários

Teste	Teste Padrão (T_P)		
Simplificado (T_S)	Baixa(1)	Média(2)	Alta(3)
Baixa(1)	4	10	0
Média(2)	5	9	3
Alta(3)	2	11	7
Baixa(1)/Média(2)	7	14	7
Média(2)/Alta(3)	3	7	8

1.6 Enfoques Estatísticos para o Problema

Existem basicamente dois enfoques estatísticos para analisar os dados discretos com padrão de censura: o Freqüentista e o Bayesiano, sendo que este último será visto com mais detalhes no decorrer dos próximos capítulos neste trabalho. Dentre os trabalhos com enfoque frequentista, destacam-se Hartley (1958), Chen e Fienberg (1974, 1976), Dempster, Laird e Rubin (1977), Little e Rubin (1987), Poleto (2006).

Na abordagem Bayesiana do problema, existem duas linhas de pesquisa distintas: uma que considera tabelas de contingência 2 x 2 com informações faltantes nas linhas ou colunas, ou seja, dados categorizados com padrão de censura sob um mecanismo de censura não informativo e a outra corresponde aos estudos em que são considerados os dados categorizados com padrão de censura ignorável,

sob um mecanismo de registro informativo. A denominação de registro informativo está relacionada ao processo de coleta dos dados. É provável que possam existir casos em que a informação observada na unidade experimental é difícil de ser determinada. Desta forma, tem-se uma certa dúvida em registrar qual categoria esta unidade amostral pertença. Por exemplo, pode-se citar a classificação da susceptibilidade dentária nos níveis adjacentes: baixo e médio ou médio e alto. É importante frisar que este mecanismo de censura é informativo, pois fornece uma certa informação relacionada aos níveis adjacentes em que os voluntários possam estar.

Para a primeira linha de pesquisa que utiliza as tabelas de contingência 2 x 2, é importante ressaltar os trabalhos de Karson e Wrobleski (1970), Antelman (1972), Kaufman e King (1973), Albert e Gupta (1983), Gunel (1984), Smith, Choi e Gunel (1985). No entanto, Dickey, Jiang e Kadane (1987) apresentaram uma extensão considerando o modelo multinomial generalizado.

A segunda abordagem Bayesiana de dados discretos com padrão de censura, sob um mecanismo de registro informativo, pode ser encontrada em Basu e Pereira (1982). Realizando uma extensão deste trabalho, Paulino e Pereira (1992) apresentaram um modelo Bayesiano que não assume um determinado processo de padrão de censura ignorável, pois considera um mecanismo de registro informativo para os dados, em que os registros podem ser estruturados em partições dos conjuntos das categorias amostradas. No entanto, em outro trabalho, Paulino e Pereira (1995) apresentaram uma solução Bayesiana para os dados deficientemente categorizados, que incorpora um padrão genérico de omissão, sob um mecanismo de registro informativo. No Capítulo 1, será abordado o modelo Bayesiano proposto por Paulino e Pereira (1995). Para esta linha de pesquisa, pode-se encontrar os trabalhos de Soares e Paulino (2001) e Soares (2004).

Devido à complexidade do problema estudado e ao tamanho amostral reduzido, o modelo de Paulino e Pereira (1992, 1995) pode não ser muito adequado. Para resolver este problema, pode-se empregar um modelo estrutural reduzido para analisar um conjunto de dados categorizados com padrão de censuras. Outro modelo interessante foi apresentado por Paulino, Soares e Neuhaus (2003), que utiliza modelos lineares generalizados para modelar o efeito de variavéis explicativas, discretas ou contínuas, na variável resposta discreta sujeita a erro.

1.7 O uso do *FBST*

O *Full Bayesian Significance Test* (*FBST*) proposto por Pereira e Stern (1999) também pode ser utilizado na análise de dados discretos com censura, pois o *FBST* é um teste de significância genuinamente Bayesiano para avaliar algumas hipóteses estatísticas. Como em toda análise bayesiana, o *FBST* utiliza a função de verossimilhança como uma forma apropriada de representar a informação estatística, para simplificar e unificar a análise estatística (Good, 1983). O termo genuinamente (*Fully*) é utilizado porque o método consiste em analisar os conjuntos de credibilidade, para isto basta conhecer o espaço paramétrico representado pela distribuição *a posteriori*. A maior crítica contra o *FBST* foi em relação a falta de invariância com respeito a reparametrização. Entretanto, Madruga *et al.* (2003) apresentaram uma versão invariante para esse método. No Capítulo 2, alguns conceitos importantes relacionados à aplicação do *FBST* serão brevemente apresentados.

O *FBST* tem sido aplicado em alguns problemas estatísticos, por exemplo, pode-se citar os testes de homogeneidade e independência em tabelas de contingência, seleção de modelos, comparação de coeficientes de variação, teste de equilíbrio de Hardy-Weinberg (Madruga et al., 2003; Pereira e Stern, 1999 e 2001) e Faria Junior (2006).

O objetivo deste trabalho é apresentar uma proposta de análise de dados categorizados com padrão de censura utilizando o *FBST*, fornecendo mais uma ferramenta estatística útil no teste de algumas hipóteses de interesse. Portanto, um exemplo ilustrativo da aplicação do *FBST* para o modelo de Paulino e Pereira (1995) será apresentado nas Seções 1.13 e 2.2.3. Algumas considerações finais e propostas de pesquisas futuras serão apresentadas no Capítulo 3.

1.8 Dados Categorizados com Censura

Nesta seção, será apresentado o modelo Bayesiano proposto por Paulino e Pereira (1995), cujo trabalho incorpora um padrão genérico de omissão, sob um mecanismo de registro informativo. Este modelo pode ser bastante útil para analisar vários problemas reais.

1.8.1 Modelo Bayesiano de Paulino e Pereira (1995)

Considere uma amostra de tamanho n retirada de uma população particionada em m categorias. Seja $\theta'=(\theta_1, ..., \theta_m)$ um vetor de probabilidades positivas para as categorias, de tal forma que

$\sum_i \theta_i = 1$. O processo amostral pode ser definido pela variável aleatória $\mathbf{W} = W_k, k = 1, 2, ..., n$ em que $W_k = i$ se a k–ésima unidade amostral pertence a i–ésima categoria, $i = 1, 2, ..., n$. Desta forma, tem-se como resultado uma seqüência finita de variáveis aleatórias independentes e identicamente distribuídas com distribução Bernoulli parametrizada por $\boldsymbol{\theta}$.

No caso de dados incompletos, $W_k, k = 1, 2, ..., n$ podem não ser totalmente observados para algumas unidades da amostra. Para cada unidade amostral, conhece-se apenas se esta pertence a algum subconjunto d não vazio de $1, 2, ..., m$ categorias. Na presença de algum mecanismo de censura um subconjunto de $\mathbf{W}$ não é completamente observado, ou seja, para algumas unidades amostrais um único elemento do espaço amostral Ω não é registrado, porém pertence a um dos eventos de Ω. Desta forma, é registrado que $W_k \in d$ para algum $d \subseteq \Omega$. Portanto, os dados observados podem ser representados por um vetor de subconjuntos registrados, $\mathbf{R} = R_k, k = 1, 2, ..., n$. Considere D como a classe de todos os possíveis subconjuntos registrados e denote $\boldsymbol{\lambda}$=$\lambda_i, i = 1, 2, ..., m$ como o vetor de probabilidades condicionais em que $\lambda_i = \lambda_{id}, d \in D_i$, $\lambda_{kd} = P(R_i = d|W_i = k)$ e $D_i = \{d \in D : i \in d\}$. Note que $d \in D$, se e somente se, $\lambda_{dk} > 0$ para algum $k \in \Omega$. Considere D_c como a classe de subconjuntos registrados em que ocorrem as censuras e $D_0 = \{i, i = 1, 2, ..., m\}$ como a classe do subconjunto dos dados completos, portanto $D = D_0 \cup D_c$. Logo, tem-se o vetor dos dados $\mathbf{N}' = (\mathbf{N}_0', \mathbf{N}_c')$, em que $\mathbf{N}_0' = (n_i, i = 1, 2, ..., m)$ contem a freqüência das observações completamente classificadas em cada uma das categorias e $\mathbf{N}_c'$ tal que as coordenadas são dadas por n_d com $d \in D_c$, em que as freqüências dos subconjuntos de Ω são as classificações incompletas. A função de verossimilhança de $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ é dada por

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}|N) \propto \prod_{d \in D} \left(\sum_{i=1} \theta_i \lambda_{di} \right)^{n_d} = \prod_{i=1}^{m} (\theta_i \lambda_{ii})^{n_i} \prod_{d \in D_c} \left(\sum_{i \in d} \theta_i \lambda_{di} \right)^{n_d}. \tag{1.1}$$

Uma parametrização alternativa pode ser apresentada em termos de $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$, em que γ' =($\gamma_0', \gamma_d', d \in D_c$) representa as probabilidades marginais das observações completamente categorizadas,

$$\gamma_0 = \sum_{i=1}^{m} \mu_{ii} = \sum_{i=1}^{m} \theta_i \lambda_{di}, \tag{1.2}$$

e as probabilidades marginais das observações com informações parciais em cada $d \in D_c$,

$$\gamma_d = \sum_{i \in d} \mu_{di} = \sum_{i \in d} \theta_i \lambda_{di}; \tag{1.3}$$

o vetor $\boldsymbol{\alpha}'=(\alpha_0', \alpha_d', d \in D_c)$ contem as probabilidades condicionais para cada categoria dada o tipo de registro,

$$\alpha_0 = (\alpha_{ii}, i = 1, 2, ..., m)', \alpha_{ii} = \mu_{ii}/\gamma_0, \sum_{i=1}^{m} \alpha_{ii} = 1, \tag{1.4}$$

$$\alpha_d = (\alpha_{di}, i \in d)', \alpha_{id} = \mu_{di}/\gamma_d, \sum_{i \in d} \alpha_{id} = 1, d \in D_c. \tag{1.5}$$

Considere $\mathbf{P}$ como a matriz de partição tal que $\boldsymbol{\gamma}=\mathbf{P}\boldsymbol{\mu}$. Então, tem-se que $\theta_i = \sum_{d \in D_i} \gamma_d \alpha_{id}, (i = 1, 2, ..., m)$. Considerando a nova reparametrização, obtem-se a função de verossimilhança,

$$L(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{N}) = \left(\gamma_0^{n_0} \prod_{x \in D_c} \gamma_d^{n_d} \right) \left[\prod_{i=1}^{m} \alpha_{ii}^{n_i} \prod_{d \in D_c} \left(\sum_{i \in d} \alpha_{id} \right)^{n_d} \right], \tag{1.6}$$

em que $n_0 = \sum_{i=1}^{n} n_i$ e $\sum_{i \in d} \alpha_{id} = 1$, $\forall d \in D_c$

Como conseqüência das propriedades da família das distribuições Dirichlet, tem-se que os parâmetros $\boldsymbol{\gamma}$, α_0 e α_d, $d \in D_c$ apresentam distribuição *a posteriori* Dirichlet independentes e identicamente distribuídas,

$$\boldsymbol{\gamma}|\mathbf{N} \sim D(\mathbf{P}'a + x), x = (n_0, N_c'), \boldsymbol{\alpha}|\mathbf{N} \sim D(a_0 + N_0), \alpha_d|\mathbf{N} \sim D(a_d), d \in D_c. \tag{1.7}$$

em que $a' = (a_0', a_d', d \in D_c)$, $a_0 = (a_{ii}, i = 1, 2, ..., m)'$ e $a_d = (a_{id}, i \in d)$.

Portanto, utilizando a distribuição *a posteriori*, tem-se que o vetor de probabilidades pode ser escrito como

$$\theta_i = \gamma_0 \alpha_{ii} + \sum_{d \in D_i \cap D_c} \gamma_d \alpha_{id}, i = 1, 2, ..., m. \tag{1.8}$$

Considerando este modelo Bayesiano, Soares e Paulino (2001) e Soares (2004) realizaram alguns estudos de simulação de Monte Carlo e também apresentaram algumas análises Bayesianas de

conjunto de dados reais para ilustrar a aplicação.

1.9 Modelo Bayesiano Particionado

Utilizando as idéias apresentadas nos trabalhos de Cox (1975), Basu e Pereira (1982), apresenta-se um modelo Bayesiano particionado para análise de dados categorizados com censura. Este modelo utiliza a informação dos dados censurados de forma diferente do modelo anterior, utilizando o particionamento da distribuição multiminomial dos dados completos e censurados.

Considere um exemplo simples em que uma pesquisa de opinião pergunta-se: "Você já usou algum tipo de droga ?". O entrevistado pode responder Sim, Não ou simplesmente não opiniar. Neste caso, tem-se uma distribuição de Bernoulli com censura, que será representado pelo vetor aleatório $(\boldsymbol{S}, \boldsymbol{N}, \boldsymbol{C})$ com distribuição trinomial com parâmetros $\boldsymbol{t}$ e $\boldsymbol{p}$=(p_1, p_2, p_3), em que $p_1 + p_2 + p_3 = 1$, $0 < p_i < 1$, $i = 1, 2, 3$ e $t = s + n + c$.

A função de probabilidade é dada por

$$f(s, n, c|\boldsymbol{t}, \boldsymbol{p}) = \begin{pmatrix} t \\ s, n, c \end{pmatrix} p_1^s p_2^n p_3^c = \frac{t!}{s!n!c!} p_1^s p_2^n p_3^c. \tag{1.9}$$

Para realizar a fatoração é necessário utilizar a seguinte parametrização:

$$\pi = p_1; \quad 1 - \pi = p_2 + p_3; \quad \theta = \frac{p_3}{1 - \pi}; \quad 1 - \theta = \frac{p_2}{1 - \pi}. \tag{1.10}$$

Considere também que $b = n + c = t - s$. Logo, tem-se que

$$\begin{pmatrix} t \\ s, n, c \end{pmatrix} = \frac{t!}{s!n!c!} = \frac{t!}{s!b!} \frac{b!}{c!n!} = \begin{pmatrix} t \\ s \end{pmatrix} \begin{pmatrix} b \\ n \end{pmatrix}. \tag{1.11}$$

Portanto, a função distribuição (1.9) pode ser escrita como

$$f(s, n, c|\boldsymbol{t}, \boldsymbol{p}) = f(s, n|\boldsymbol{t}, \boldsymbol{p}) = \begin{pmatrix} t \\ s \end{pmatrix} \pi^s (1 - \pi)^{t-s} \begin{pmatrix} b \\ n \end{pmatrix} \theta^n (1 - \theta)^{b-n}. \tag{1.12}$$

Utilizando distribuições Beta para π e θ iguais a Beta(x, y) e Beta (u, v), respectivamente e as distribuições *a priori* Beta(a_1, a_2) e Beta(b_1, b_2), tem-se que as distribuições *a posteriori* são

$$\pi|[x, y] \sim Beta(x + a_1, y + a_2); \qquad \theta|[u, v] \sim Beta(u + b_1, v + b_2). \tag{1.13}$$

É importante ressaltar que π e θ assumem valores no intervalo $[0, 1]$ e são de variação independente, ou seja, conhecido o valor de um parâmetro, o segundo continua incerto exatamente como antes, variando no mesmo espaço paramétrico. Desta forma, a distribuição Multinomial pode ser particionada em produtos de distribuições Binomiais com parâmetros de variação independente.

Para realizar a fatoração de acordo com o interesse da análise estatística de modelos Multinomiais de maior dimensão, é importante considerar dois resultados importantes: um da partição de classes da Multinomial e outro, do condicionamento na soma parcial.

Considere uma amostra de tamanho t com distribuição Multinomial originada de um processo de Bernoulli m-variado. Logo, a distribuição multinomial pode.ser denotada como

$$\boldsymbol{x}|[\boldsymbol{t}, \boldsymbol{\theta}] \sim M_m(\boldsymbol{t}, \boldsymbol{\theta}). \tag{1.14}$$

Seja $\{i : i = 1, 2, ..., m\}$ o domínio dos índices que denotam as classes de uma distribuição Multinomial de ordem m e $\mathbf{A}_{s\times m}$ uma matriz de partição de m classes em s super-classes, ou seja, em partições menores com $a_{ij} \in \{0, 1\}$ e linhas ortogonais. Desta forma, se $\mathbf{x} \sim M_m(\mathbf{t}, \boldsymbol{\theta})$, então $\mathbf{y} = \mathbf{Ax} \sim M_s(\mathbf{t}, \mathbf{A}\boldsymbol{\theta})$. Maiores detalhes destes resultados podem ser vistos em Pereire e Stern (2008).

Outro interesse está ligado à distribuição resultante do vetor $\mathbf{x}$ obtida por meio do condicionamento na soma parcial. A distribuição do vetor $\mathbf{x}$ condicionada a soma a sua soma apresenta distribuição Multinomial com parâmetros correspondentes normalizados. Por exemplo, o condicionamento nas l primeiras componentes resulta em

$$\boldsymbol{x}_{1:l}|[\mathbf{1}'\boldsymbol{x}_{1:l} = j] \sim M_l\left(j, \frac{1}{\mathbf{1}'\boldsymbol{\theta}_{1:l}}\boldsymbol{\theta}_{1:l}\right). \tag{1.15}$$

em que $j \in \{0, 1, ..., t\}$.

Suponha o interesse de analisar uma distribuição Multinomial $\mathbf{x} \sim M_m(\mathbf{t}, \theta)$, com censuras nas últimas componentes, decomposta em duas Multinomiais e uma Binomial.

$$P(\boldsymbol{x}|t,\theta) = \sum_{j=0}^{t} P\left(\boldsymbol{x}_{1:q}|j, \tfrac{1}{1'\theta_{1:q}}\theta_{1:q}\right) P\left(\boldsymbol{x}_{q+1:m}|t-j, \tfrac{1}{1'\theta_{q+1:m}}\theta_{q+1:m}\right) \tag{1.16}$$

$$P\left(\begin{bmatrix} j \\ (t-j) \end{bmatrix} | t, (1'\boldsymbol{\theta}_{1:q}, 1'\boldsymbol{\theta}_{q+1:m})'\right) \tag{1.17}$$

O problema de não idenficabilidade dos parâmetros é um problema geralmente encontrado nos modelos para análise de dados categorizados com censuras nas literaturas não-Bayesianas. No entanto, o uso de distribuições *a priori* para os parâmetros estudados permite que este problema não exista na análise Bayesiana.

Considerando distribuição Dirichlet $DM_m(\mathbf{a})$ como distribuição *a priori* para os parâmetros, pode-se obter que as distribuições *a posteriori* são dadas por

$$\frac{1}{1'\boldsymbol{\theta}_{1:q}}\boldsymbol{\theta}_{1:q}|[x_{1:q}, j] \sim DM_q(\boldsymbol{x_{1:q}} + \boldsymbol{a}); \quad \frac{1}{1'\boldsymbol{\theta}_{q+1:r}}\boldsymbol{\theta}_{q+1:m}|[x_{q+1:m}, n-j] \sim DM_{m-(q+1)}(\boldsymbol{x}_{q+1:m} + \boldsymbol{a}); \tag{1.18}$$

$$1'\boldsymbol{\theta}_{1:t}|[j,t] \sim Beta(j + a_1, t - j + a_2). \tag{1.19}$$

Na próxima seção, serão apresentados alguns conceitos importantes do procedimento do $FBST$ de Pereira e Stern (1999), que serão utilizados para a análise dos modelos de dados deficientemente categorizados.

1.10 Full Bayesian Significance Test ($FBST$)

A versão original do *Full Bayesian Significance Test* ($FBST$) foi introduzido por Pereira e Stern (1999). Este teste foi apresentado como um teste de significância genuinamente Bayesiano para avaliar hipóteses precisas. O $FBST$ é intuitivo e se baseia num valor do conceito de evidência. A maior crítica enfrentada estava relacionada a possível falta de invariância em relação à reparametrização. No entanto, este problema já foi resolvido, pois Madruga *et al.* (2003) apresentaram uma versão

invariante para o $FBST$.

1.11 Definição do $FBST$

Considere $X_1, X_2, ..., X_n$ como sendo variáveis aleatórias com função densidade conjunta dada por $\prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$, em que $\boldsymbol{\theta}$ é o vetor de parâmetros, definido no espaço paramétrico $\Theta \subseteq R^p (p \geq 1)$. Tem-se interesse em testar a hipótese nula $H_0 : \boldsymbol{\theta} \in \Theta_0$, $\Theta_0 \subset \Theta$, com dimensão de $\Theta_0 \leq$ dimensão de Θ. Geralmente, Θ_0 é um vetor de restrições escritas em termos de igualdades e inequações. Considere uma hipótese com, no mínimo, uma restrição de igualdade

$$\Theta_0 = \{\boldsymbol{\theta} \in \Theta | c(\boldsymbol{\theta}) \leq 0 \wedge h(\boldsymbol{\theta}) = 0\}. \tag{1.20}$$

Considere um modelo estatístico, para p inteiro, $\boldsymbol{\theta} \in \Theta \subset R^p$ é um vetor de parâmetro, $g(\boldsymbol{\theta})$ como a densidade *a priori*, x é uma observação (escalar ou vetorial) e $L(\boldsymbol{\theta}; x)$ representa a função de verossimilhança de $\boldsymbol{\theta}$ em Θ. Logo após observar os dados, a próxima etapa é avaliar o valor da evidência bayesiana. A densidade *a posteriori* para $\boldsymbol{\theta}$ dado x pode ser escrita

$$g_x(\boldsymbol{\theta}) = g(\boldsymbol{\theta}|x) \propto g(\boldsymbol{\theta}) L(\boldsymbol{\theta}; x). \tag{1.21}$$

Considere

$$g* = \sup_{\boldsymbol{\theta} \in \Theta_0} g_x(\boldsymbol{\theta}) \quad e \quad T = \{\boldsymbol{\theta} \in \Theta : g_x(\boldsymbol{\theta}) > g*\}. \tag{1.22}$$

O valor da evidência Bayesiana contra a hipótese nula H_0 é definido como a probabilidade a posteriori do conjunto tangencial T, isto é,

$$\overline{ev} = P(\boldsymbol{\theta} \in T|x) = \int_T g_x(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1.23}$$

Portanto, o valor da evidência em favor da hipótese nula H_0 é definida como $ev = 1 - \overline{ev}$. É importante esclarecer que a ev não é uma evidência contra a hipótese alternativa H_1. Da mesma forma, $\overline{ev}$ não é evidência em favor de H_1, embora esta seja contra a hipótese nula H_0.

Para minimizar as fortes críticas relacionadas a invariância do $FBST$, Madruga *et al.* (2003) apresentaram uma versão invariante com respeito a parametrizações alternativas do espaço para-

metrico. Considere uma função de referência $r(\theta)$ sobre Θ, que é definida no espaço paramétrico original em que as prioris são definidas. Por exemplo, $r(\theta)$ pode ser uma densidade não-informativa, possivelmente imprópria, sobre Θ. Considere agora a função:

$$s_x(\theta) = \frac{g_x(\theta)}{r(\theta)} \text{ e } s^* = \sup_{H_0} s_x(\theta). \tag{1.24}$$

A forma invariante do conjunto tangencial T e o valor da evidência contra a hipótese H_0 é definida como

$$T = \{\theta \in \Theta_H : s_x(\theta) > s_*\} \text{ e } \overline{ev} = \int_T g_x(\theta) d\theta. \tag{1.25}$$

Portanto, o procedimento do $FBST$ consiste em rejeitar H_0 sempre que o valor encontrado para $ev = 1 - \overline{ev}$ for relativamente pequeno.

1.12 Calculando o $FBST$

Para se calcular o valor da evidência do $FBST$ é necessário basicamente cumprir duas etapas: a da otimização da função a posteriori sob a hipótese H_0 e a da integração da posteriori restrita a região tangencial T.

1.12.1 Otimização

A etapa da Otimização consiste em encontrar g^* que maximiza a densidade a posteriori sob a hipótese nula H_0. Existem vários algoritmos de otimização. No entanto, é importante lembrar que deve ser levado em consideração também outras formas de otimização, de acordo com a conveniência e as necessidades do problema analisado.

1.12.2 Integração

Após encontrar o ponto de máximo sob a hipótese H_0, é necessário calcular o valor da integral em 1.23. Esta etapa consiste em integrar a posteriori sobre toda a região em que for maior que g^*. Para calcular o valor da integral, pode-se utilizar algum método numérico determinístico, por exemplo, o *Monte Carlo Importance Sampling*, que pode ser utilizado para encontrar o valor de ev.

Monte Carlo Importance Sampling aplicado ao $FBST$

A apresentação descrita a seguir foi apresentada por Stern e Zacks (2002) e Stern (2003). Inicialmente, escolhe-se uma densidade de probabilidade q sobre Θ adequada para aplicar o *Importance Sampling*. Pode-se reescrever a $\overline{ev}$ como

$$\overline{ev} = \frac{\int_\Theta \mathbf{1}_{\{\theta \in T\}}(\theta) L(\theta; x) g(\theta) d\theta}{\int_\Theta L(\theta; x) g(\theta) d\theta} = \frac{\int_\Theta \frac{p(\theta)}{q(\theta)} q(\theta) d\theta}{\int_\Theta \frac{h(\theta)}{q(\theta)} q(\theta) d\theta}. \tag{1.26}$$

Agora considere as variáveis aleatórias $Y_1, Y_2, ..., Y_n$ com densidade q. Logo, tem-se que o estimador de $\overline{ev}$ é dado por

$$\hat{\psi} = \frac{\overline{Z}_n^*}{\overline{Z}_n}. \tag{1.27}$$

em que

$$\overline{Z}_n^* = \frac{1}{n} \sum_{i=1}^{n} \frac{g(Y_i)}{q(Y_i)} \text{ e } \overline{Z}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{h(Y_i)}{q(Y_i)}. \tag{1.28}$$

Utilizando a Lei Forte dos Grandes Números, tem-se que $\overline{Z}_n^*$ e $\overline{Z}_n$ converge para $\int_{\Theta_0} L(\theta; x) g(\theta) d\theta$ e $\int_\Theta L(\theta; x) g(\theta) d\theta$, respectivamente.

Stern e Zacks (2003) demonstraram que o estimador $\hat{\psi}$ é consistente. Considere $Z_i = \frac{g(Y_i)}{q(Y_i)}$ e $Z_i^* = \frac{h(Y_i)}{q(Y_i)}$, então as componentes do vetor (Z_i, Z_i^*) são dependentes, mas os vetores (Z_i, Z_i^*) e (Z_j, Z_j^*) são independentes para $i \neq j$. Desta forma, pela Lei Forte dos Grandes Números, o estimador $\hat{\psi}$ converge para $\overline{ev}$.

Considere a quantidade pivotal dada por

$$U_n = \overline{Z_n}\psi - \overline{Z_n^*}. \tag{1.29}$$

Tem-se que $E(U_n) = 0$ e a variância é igual a

$$V(U_n) = \frac{1}{n}\left(V(\overline{Z_n}^*) + \psi^2 V(\overline{Z_n}) - 2\psi Cov(\overline{Z_n^*}, \overline{Z_n})\right) < \infty, \tag{1.30}$$

Os estimadores consistentes para as variâncias e covariância são, respectivamente, iguais a

$$\hat{\sigma}_n^{2*} = \frac{1}{n}\sum_{i=1}^{n}(Z_i^* - \overline{Z_n^*}), \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z_n}), \tag{1.31}$$

$$\hat{\sigma}_{g,h,n}^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i^* - \overline{Z_n}^*)(Z_i - \overline{Z_n}), \tag{1.32}$$

Quando $n \to \infty$, $V(U_n) \xrightarrow{p} \sigma^2$ e $U_n \xrightarrow{D} N(0, \sigma^2)$. Portanto, utilizando o Teorema de Slutsky e o Teorema Central do Limite, tem-se que

$$\frac{nU^2}{(\hat{\sigma}_n^{2*} + \psi^2\hat{\sigma}_n^2 - 2\psi\hat{\sigma}_{g,h,n}^2)} \sim \chi^2(1). \tag{1.33}$$

Pode-se aplicar o Teorema de Filler para encontrar as raízes da equação quadrática para ψ de

$$(\overline{Z}_n\psi - \overline{Z}_n^*)^2 = \frac{\chi_{1-\beta}^2(1)}{n}(\hat{\sigma}_n^{2*} + \psi^2\hat{\sigma}_n^2 - 2\psi\hat{\sigma}_{g,h,n}^2). \tag{1.34}$$

Desta forma, pode-se construir um intervalo de confiança de $(1-\beta)$x100% para ψ

$$IC(n, 1-\beta) = \hat{\psi} \pm \triangle_{n,1-\beta}, \tag{1.35}$$

em que

$$\triangle_{n,1-\beta} = \frac{\chi_{1-\beta}^2(1)}{n}(\hat{\sigma}_n^{2*} + \psi^2\hat{\sigma}_n^2 - 2\psi\hat{\sigma}_{g,h,n}^2). \tag{1.36}$$

Portanto, como critério de parada para o algoritmo de integração do método de *Importance Sampling*, pode-se utilizar o erro $\triangle_{n,1-\beta}$ de ψ.

1.12.3 Motivação

Como exemplo ilustrativo da aplicação do $FBST$, considere os dados hipotéticos apresentados na Tabela 1.2. O objetivo é testar se as proporções populacionais da presença da doença são iguais para ambos os sexos. As proporções para os sexos feminino e masculino são denotadas como π_1 e π_2, respectivamente. Portanto, tem-se a hipótese de interesse $H_0 : \pi_1 = \pi_2$. A densidade a posteriori para o problema estudado é

$$f(\pi_1, \pi_2) \propto \pi_1^{m_1}(1-\pi_1)^{n_1-m_1}\pi_2^{m2}(1-\pi_2)^{n_2-m_2}. \tag{1.37}$$

em que m_i e n_i representam, respectivamente, o número de pessoas com doença e o número total de pessoas do sexo i.

Tabela 1.2: Dados hipotéticos de presença de uma doença.

Sexo	Doença Presente	Doença Ausente	Total
Feminino	3	10	13
Masculino	7	6	13
Total	10	16	26

Com intuito de realizar uma simples comparação do resultado do $FBST$, considere o valor p obtido pelo teste χ^2, o Fator de Bayes (FB) e a probabilidade posteriori para testar a hipótese de homogeneidade. Considerando a priori $P(H) = P(\pi_1 = \pi_2) = 1/2$, Irony e Pereira (1995) obtiveram que o Fator de Bayes para o teste de homogeneidade de uma tabela de contingência 2 x 2 é dado por

$$FB = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1+n_2}{m_1+m_2}} \frac{(n_1+1)(n_2+1)}{(n_1+n_2+1)}. \tag{1.38}$$

A expressão do fator de Bayes acima é encontrada a partir de uma aproximação de Taylor de segunda ordem usando uma distribuição *a priori* uniforme, ou seja, uma distribuição a priori Dirichlet com parâmetros iguais a $(1,1,1,1)$. Logo, a probabilidade posteriori (PP) de H pode ser calculada como $PP = (1-(FB)^{-1})^{-1}$.

Para o problema hipotético apresentado, tem-se que o valor aproximado da evidência (ev) do $FBST$ é igual a $0,874$ ($\overline{ev} = 0,126$). O valor p do teste χ^2 para a tabela acima é de $0,223$. O valor do Fator de Bayes e a probabilidade posteriori são iguais a $FB = 0,671$ e $PP = 0,401$, respectivamente. Portanto, baseando-se nos resultados encontrados, a hipótese de homogeneidade dos dados da tabela acima não é rejeitada.

Na próxima seção, para facilitar o entendimento da aplicação do $FBST$, são apresentados alguns exemplos da análise bayesiana de dados categorizados com omissão, utilizando o cálculo aproximado da ev do $FBST$.

1.13 Aplicação do FBST

Considere o exemplo a seguir como aplicação do $FBST$, cujos dados são frequentemente analisados em diferentes trabalhos de Paulino e Pereira (1995), Soares (2004) e Poleto (2006).

1.13.1 Testes de Coloração Dentária

Considere os dados da Tabela 1.1 apresentados em Paulino e Pereira (1995). Os dois testes estudados apresentam algumas características diferentes: um representa o teste padrão, que é bastante confiável, porém caro, e o outro é um teste simplificado, mais barato, que se baseia na observação da coloração obtida na reação de um produto com a saliva do paciente. Portanto, existe um certo interesse econômico na possibilidade de trocar por um teste mais confiável e viável economicamente. No entanto, em alguns casos, existe a possibilidade de dúvidas em relação à coloração do teste simplificado. Isto acontece quando o profissional não é capaz de se decidir entre níveis adjacentes de escala de coloração dentária. Um fator importante a ser condiderado é que existe a possibilidade de dúvida, que é capaz de criar um mecanismo de omissão. Isto pode ocorrer devido um certo confundimento entre as categorias próximas. O principal objetivo de realizar este tipo de experimento está relacionado na possibilidade de avaliar uma possível substituição do teste padrão (T_P) pelo teste simplificado (T_S).

Para facilitar a compreensão, denotam-se a classe Baixa por 1, Média por 2 e Alta por 3. Agora considere o vetor de probabilidades das categorias amostrais ordenadas lexicograficamente, $\theta=(\theta_1, \theta_2, ..., \theta_9)'$, em que $\theta_1 = P(T_S = 1, T_P = 1)$, $\theta_2 = P(T_S = 1, T_P = 2)$,..., $\theta_9 = P(T_S = 3, T_P = 3)$. O experimento estudado estabelece a existência de 15 classes possíveis: sendo que as 9 primeiras

correspondem aos dados amostrais completos e as 6 últimas restantes correspondem aos dados deficientemente categorizados, resultantes dos agrupamentos das classes $D_c = \{(1,4),(2,5),(3,6),(4,7),(5,8),(6,9)\}$. Considere o interesse de testar a hipótese de homogeneidade marginal das proporções dos dois testes dentários.

Um dos problemas enfrentados na simulação da distribuição a posteriori (1.7), utilizando diretamente o método de Monte Carlo, é a verificação de uma certa concentração da posteriori em uma região relativamente pequena do espaço amostral. Para resolver este problema, aplicou-se a relação das distribuições Dirichlet com as distribuições Normais Logísticas proposta por Aitchison (2003). Outra vantagem é que os parâmetros da distribuições Normais evitam os problemas de cálculos númericos para maximização. Utilizando a transformação logística de $\boldsymbol{y} \in R^d$ para $\boldsymbol{x} \in R^d$, definida como

$$x_i = \frac{\exp(y_i)}{(\sum_{i=1}^{d} \exp(y_i) + 1)}; i = 1, 2, ..., d, \tag{1.39}$$

$$x_D = 1 - \sum_{i=1}^{d} x_i = \frac{1}{(\sum_{i=1}^{d} \exp(y_i) + 1)}, \tag{1.40}$$

com a transformação $y_i = \ln(x_i/x_D)$, $i = 1, 2, ..., d$ e jacobiano $J = (\prod_{i=1}^{d})^{-1}$.

Logo, após a transformação logística, o vetor de médias e a matriz de variância e covariância da distribuição Normal são respectivamente iguais a

$$\mu_i = \Psi(\alpha_i) - \Psi(\alpha_D); i = 1, 2, ..., d; \tag{1.41}$$

$$\sigma_{ii} = \Psi'(\alpha_i) + \Psi'(\alpha_D); i = 1, 2, ..., d; \tag{1.42}$$

$$\sigma_{ij} = \Psi'(\alpha_D); i \neq j = 1, 2, ..., d. \tag{1.43}$$

em que $\Psi(x)$ representa a função digama,

$$\Psi(x) = \frac{d}{dx} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}, \tag{1.44}$$

e $\Psi'(x)$ denota a segunda derivada da função trigama,

Tabela 1.3: Valor aproximado da evidência ev de um estudo de simulação

Classes D_c	$\{(1,4),(2,5),(3,6)\}$	$\{(4,7),(5,8),(6,9)\}$	$\{(1,4),(4,7)\}$	$\{(2,5),(5,8)\}$	$\{(4,7)\}$
(ev)-Modelo Paulino e Pereira	0,365	0,347	0,348	0,353	0,356
(ev)-Modelo Particionado	0,425	0,287	0,268	0,210	0,237

$$\Psi'(x) = \frac{\partial^2}{\partial x^2} ln\Gamma(x) = \frac{d}{dx}\Psi(x). \tag{1.45}$$

Abramowitz e Stegun(1972) e Devroye (1992) apresentam fórmulas para calcular os valores aproximado das funções $\Psi(x)$ e $\Psi'(x)$, como apresentadas a seguir:

$$\Psi(a) - \Psi(b) \approx \frac{1}{b} - \frac{1}{a}, \tag{1.46}$$

$$\psi'(x) = \sum_{n=0}^{\infty} \left(\frac{1}{(x+n)^2} \right) \tag{1.47}$$

Para calcular o valor aproximado da evidência ev, foi utilizado um procedimento numérico de *Importance Sampling*, implementado em linguagem *R*. O valor aproximado da evidência para as hipóteses de homogeneidade foi de 0.626, considerando as classes de censuras $D_c = \{(1,4),(2,5),(3,6),(4,7),(5,8),(6,9)\}$. No entanto, para avaliar outros resultados, um estudo de simulação foi realizado para os modelos Bayesianos de Paulino e Pereira e o particionado. Os resultados encontrados estão apresentados na Tabela 1.3.

Utilizando o mesmo conjunto de dados para o modelo Bayesiano particionado em (1.18), calculou-se o valor aproximado da evidência por meio de simulação e obteve-se ev igual a $0,427$. Baseando-se nos valores calculados da evidência da Tabela, não se rejeita a hipótese nula de homogeneidade marginal dos dois testes odontológicos. Portanto, conclui-se pelo $FBST$ que o teste padrão pode ser substituído pelo teste simplificado.

No próximo capítulo, um modelo Bayesiano para estudar a relação entre o vetor de variáveis explicativas e a varíavel resposta binária deficientemente categorizada será apresentado, com o intuito

de estudar sua aplicação em pesquisas nas quais existem dados categorizados com censuras.

Capítulo 2

Modelo de Regressão Logística

Dentre os métodos de modelagem estatística de dados, a regressão logística é a modelagem que ocupa uma certa posição de importância para análise de conjunto de dados com variável resposta binária. Atualmente alguns modelos Bayesianos tem sido estudados e grandes contribuições nesta área tem sido apresentadas, especialmente na modelagem de dados com respostas binárias. Neste capítulo, uma breve comparação é realizada entre o modelo de regressão logística clássico e o modelo Bayesiano, utilizando um conjunto de dados reais. Um modelo logístico bayesiano para dados categorizados com censura é apresentado utilizando o conceito de partição. Será também apresentado o modelo Bayesiano proposto por Paulino, Soares e Neuhaus (2003) que utiliza modelos lineares generalizados para modelar os efeitos das variáveis explicativas, discretas ou contínuas, na variável resposta deficientemente categorizada.

2.1 Regressão Logística

Considere o conjunto de dados do estudo de Henrietta Cedergren sobre a eliminação do s da língua espanhola panamenha. O arquivo dos dados está disponível na página http://www-npl.stanford.edu/~manning/courses/ling236/handouts/panama-win.tkn. A pesquisadora entrevistou falantes nativos da cidade do Panamá, que costumam eliminar o s do final das palavras, como em vários dialetos da Espanha. O objetivo do estudo era investigar se a mudança observada é significativa para entender o dinamismo regional do espanhol panamenho. A selecão dos entrevistados foi realizada em várias regiões da cidade com várias classes sociais para verificar como a variação é estruturada na comunidade. Também houve a preocupação de estudar se havia alguma restrição para a eliminação do s, por exemplo, se o próximo determinante linguístico é uma consoante, vogal ou uma pausa e a

classe gramatical das palavras: advérbios (por exemplo, menos), verbos, (tu tienes, el tiene), artigo determinante (los, las), adjetivo (buenos) e substantivos (amigos).

As variáveis foram codificadas como: $P1$: 1=eliminação s, 0= não eliminação s; $P2$- categoria gramatical: m=advérbio, v=verbo (2^a pessoa do singular), d=artigo determinante no plural; a=flexão de plural nos adjetivos e n=flexão de substantivos; $P3$- próximo determinante linguístico: c=consoante; v=vogal e p=pausa e P-classe social: 1=alta; 2=média alta; 3=média baixa e 4=baixa.

Este conjunto de dados será analisado utilizando o modelo de regressão logística clássica e a Bayesiana. Uma breve comparação entre os dois modelos é apresentada como ilustração da aplicação.

Considere um conjunto de dados obtidos para cada unidade amostral de uma varíável resposta binária e de variáveis explicativas, discretas ou contínuas, $(n_k, N_k, \mathbf{x}_k), k = 1, 2, ..., N$, em que os valores de n_k representam os números de sucessos num total de N_k observações com um padrão comum de p variáveis explicativas agrupadas pelo vetor $\mathbf{x}_k$. O número de n_k de sucessos representam distribuições Binomiais independentes, $B(N_k, \phi_k)$.

Considere o modelo de regressão logística

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta}, \tag{2.1}$$

em que $\mathbf{x}$ é o vetor dos valores observados das (p-1) variáveis explicativas. A estimativa do vetor de parâmetros $\boldsymbol{\beta}$ é obtida por meio de processo iterativo de mínimo quadrados reponderados.

Na análise de regressão logística clássica, a qualidade de ajuste do modelo ajustado pode ser avaliada por meio da função desvio, geralmente denominada *deviance*. Esta medida é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de máxima verossimilhança de $\boldsymbol{\beta}$. Um pequeno valor para a função desvio mostra que, para um número menor de parâmetros, obtém-se um bom ajuste quanto o ajuste do modelo saturado. Geralmente se compara os valores da função desvio com os percentis de uma distribuição qui-quadrado com $n - p$ graus de liberdade.

Considerando o modelo Bayesiano, a distribuição *a priori* de (β), que representa o conhecimento *a priori* do pesquisador, pode ser bastante complicada devido às dificuldades na explicação dos coeficientes da regressão por sua própria estrutura e também por sua interpretação depender do modelo adotado. Bedrick *et al.* (1996) analisaram este problema e apresentaram um método na especificação da distribuição *a priori* para o vetor $\boldsymbol{x}_l$, $l = 1, 2, ..., p$ variáveis regressoras. As distribuições *a priori* de $(\boldsymbol{x}_l'\beta)$ são distribuições Beta (c_l, d_l) independentes. A distribuição induzida de β é

$$\pi(\beta) \propto \prod_{l=1}^{p}((\boldsymbol{x}_l'\beta))^{c_l-1}(1-(\boldsymbol{x}_l'\beta))^{d_l-1} f(\boldsymbol{x}_l'\beta). \tag{2.2}$$

em que $f(.)$ é a função densidade correspondente a função de distribuição acumulada $F(.)$ do modelo de regressão para dados binários. Os hiperparâmetros c_l e d_l são determinados a partir dos conhecimentos prévio das características da distribuição *a priori* para $(\boldsymbol{x}_l'\beta)$. Pereira e Stern (2001) apresentaram um critério de seleção de modelo utilizando o procedimento do $FBST$. Portanto, pode-se aplicar o $FBST$ para avaliar a qualidade do ajuste do modelo. A hipótese testada é $H_0 : \beta = 0$ vs. H_1 : pelo menos um $\beta_j \neq 0$, $j = 1, 2, ..., p$.

O primeiro modelo ajustado é

$$log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = -1.32 - 0.17d + 0.18m + 0.67n - 0.77v + 0.95P + 0.53V + 1.27c2 + 1.05c3 + 1.37c4, \tag{2.3}$$

O valor da função desvio e o valor aproximado da evidência do $FBST$ são 760.03 com 42 graus de liberdade e $0,428$, respectivamente. Note que o valor da função desvio está relativamente alto em relação aos graus de liberdade. Ajustando o modelo novamente, sem a variável de categoria d, obtem-se

$$log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = -1.43 + 0.30m + 0.78n - 0.65v + 0.96P + 0.53V + 1.26c2 + 1.04c3 + 1.37c4, \tag{2.4}$$

Calculando o valor da função desvio e o valor aproximado da evidência do $FBST$, pode-se obter os valores de 201.47 com 43 graus de liberdade e $0,381$, respectivamente. Note que o valor da função desvio apresentou uma certa diminuição e o valor da ev aumentou. Isto mostra que este modelo apresenta um ajuste melhor que o do modelo ajustado anteriormente. Portanto, a aplicação do

procedimento do $FBST$ pode ser bastante útil na seleção de modelos.

Uma investigação importante é verificar se a classe social apresenta alguma interação com a variável de categoria gramatical. Para testar a possível interação entre estas varíaveis, ajustou-se o modelo logístico apresentado a seguir.

$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = -1.85 + 2.47d + 7.63m + 1.63n - 1.65v + 9.57P + 5.39V + 1.95c2 + 1.37c3 + 2.08c4$ - 4.82c2*d-7.52c2*m-9.50c2*n+1.47c2*v+9.87c3*d-2.38c3*m-1.09c3*n+1.67c3*v -1.90c4*d-8.53c4*m-1.07c4*n+1.52c4*v

O valor de ev calculado para o modelo com as interações de classe social e classificação gramatical foi igual a 0,023. Logo, comparando este valor aos valores de ev dos modelos anteriores, tem-se evidência de que este modelo apresenta um melhor ajuste aos dados que os anteriores sem interação. Desta forma, pode-se concluir que o hábito de eliminar o s das classifações gramaticais ocorre diferentemente nas classes sociais.

Na próxima Seção, a possibilidade de censuras na variável resposta será considerada no ajuste e algumas abordagens em conjunto de dados que apresentam esta característica.

2.2 Regressão Logística com censura

Paulino, Soares e Neuhaus (2003) apresentaram uma abordagem interessante. Outra proposta de análise do modelo logístico Bayesiano também é apresentada utilizando a ideia de particionamento, como já foi abordada anteriormente.

2.2.1 Modelo de Paulino, Soares e Neuhaus (2003)

Assumindo a possibilidade de censuras na variável resposta, pode-se definir as variáveis R^o e R^v como sendo as variáveis da resposta observável e da verdadeira, respectivamente. Denote as probabilidade dos verdadeiros valores da variável resposta como $\theta_{ki} = P(R^v = i|\mathbf{x}_k)$, $k = 1, 2, ..., N$, $i = 1, 2$ com $\sum_i \theta_{ki} = 1$. Por sua vez, denota-se as probabilidades da resposta observada com possível censura como $\lambda_{ki} = P(R^o = i|\ R^v = i; \mathbf{x}_k)$, $k = 1, 2, ..., N$, $i = 1, 2$ com $\sum_i \lambda_{ki} = 1$. O modelo probabilístico para os dados é dado pela verossimilhança do produto de Binomiais.

$$L(\boldsymbol{\theta},\boldsymbol{\lambda};\, \boldsymbol{n}) = \prod_{k=1}^{N} \binom{N_k}{n_k} \left(\sum_i \lambda_{ki1}\theta_{ki}\right)^{n_k} \left(\sum_i \lambda_{ki0}\theta_{ki}\right)^{N_k - n_k}, \tag{2.5}$$

em que $\boldsymbol{\theta}$ e $\boldsymbol{\lambda}$ representam os vetores dos parâmetros θ_{ki} e λ_{kij}.

O efeito das variáveis explicativas na variável resposta pode ser estudado utilizando os modelos lineares generalizados

$$E\left(\frac{n_k}{N_k}\middle|\boldsymbol{\theta}\right) = \theta_{k1} = \theta_1(x_k'\beta) = f(x_k'\beta), \tag{2.6}$$

em que β é um vetor p x 1 de coeficientes do modelo linear e $f(x'\beta)$ pode ser a função logística, a normal ou a Gumbel, por exemplo,

$$F(x'\beta) = \begin{cases} e^{x'\beta}/(1+e^{x'\beta}), \\ \Phi(x'\beta), \\ 1 - e^{-x'\beta}. \end{cases} \tag{2.7}$$

A falta de identificabilidade deste modelo pode ser verificada em alguns casos. Por exemplo, considere $\alpha = (\beta, \lambda)$ e $\alpha^* = (\beta^*, \lambda^*)$ tais que $\beta^* = -\beta$, $\lambda_{01}^* = 1 - \lambda_{10}$ e $\lambda_{10}^* = 1 - \lambda_{01}$. A probabilidade de sucesso é igual a

$$\xi(\alpha^*) = F(x_k'\beta^*)(1-\lambda_{01}^*) + (1 - F(x_k'\beta^*))\lambda_{10}^* = F(-x_k'\beta^*)\lambda_{10} + (1 - F(-x_k'\beta^*))(1-\lambda_{01}).$$

Se é verdadeira a condição $F(-x) = 1 - F(x), \forall x \in \Re$, então

$$\xi(\alpha^*) = (1 - F(x_k'\beta^*))\lambda_{10} + F(x_k'\beta^*)(1-\lambda_{01}) = \xi(\alpha).$$

Assumi-se *a priori* que os parâmetros $\boldsymbol{\theta}$ e $\boldsymbol{\lambda}$ são independentes, pois as ocorrências de erros de classificação acontecem devido ao procedimento de coleta que se considera independente da proporção de sucessos da variável resposta.

A distribuição *a priori* de (β, λ), que representa o conhecimento *a priori* do pesquisador, pode ser bastante complicada devido às dificuldades na explicação dos coeficientes da regressão por causa

de sua própria estrutura e interpretação, dependendo do modelo adotado. Bedrick *et al.* (1996) analisaram este problema e apresentaram um método na especificação da distribuição *a priori* para as médias condicionais de $\theta_1(\boldsymbol{x}_l'\boldsymbol{\beta})$ baseadas no vetor $\boldsymbol{x}_l$, $l = 1, 2, ..., p$ das variáveis regressoras. As distribuições *a priori* de $\theta_1(\boldsymbol{x}_l'\boldsymbol{\beta})$ são distribuições Beta (c_l, d_l) independentes. A distribuição induzida de $\boldsymbol{\beta}$ é

$$\pi(\boldsymbol{\beta}) \propto \prod_{l=1}^{p} (\theta_1(\boldsymbol{x}_l'\boldsymbol{\beta}))^{c_l - 1}(1 - \theta_1(\boldsymbol{x}_l'\boldsymbol{\beta}))^{d_l - 1} f(\boldsymbol{x}_l'\boldsymbol{\beta}). \tag{2.8}$$

em que $f(.)$ é a função densidade correspondente a função de distribuição acumulada $F(.)$ do modelo escolhido.

Os hiperparâmetros c_l e d_l são determinados a partir dos conhecimentos dos pesquisadores sobre as características da distribuição *a priori* para $\theta_1(\boldsymbol{x}_l'\boldsymbol{\beta})$. Os hiperparâmetros da distribuição *a priori* de $\boldsymbol{x}_l'\boldsymbol{\beta}$ também são encontrados da mesma forma.

Utilizando estes resultados, Paulino, Soares e Neuhaus (2003) apresentaram uma abordagem que permite a fatorização da verossilhança de $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{m})$, como descrito a seguir.

Considere m_{kij} como sendo o número de observações com $R^v = i$ e $R^o = j$ entre aquelas observações com vetor de covariáveis definidos por $\boldsymbol{x_k}$. Para as quantidades não-observadas, tem-se $m_{ki+} = \sum_j m_{kij}$, $m_{k+1} = \sum_i m_{kij} = n_k$ e $m_{k+0} = N_k - n_k$. Os dados $\boldsymbol{m}$ representam uma amostra de um produto de distribuições multinomiais $M(n_{k,kij}\,\theta_{ki})$ com verossimilhança, sob a parametrização do modelo linear, igual a

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{m}) \propto \prod_{k,i} (\theta_1(\boldsymbol{x}_l'\boldsymbol{\beta}))^{m_{ki+}} \prod_{k,i,j} \lambda_{kij}^{m_{kij}}. \tag{2.9}$$

A função de verossimilhança permite a fatorização $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{m}) = L(\boldsymbol{\beta}|\boldsymbol{m})L(\boldsymbol{\lambda}|\boldsymbol{m})$. A distribuição *aposteriori* dos dados é

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{m}) \propto \pi(\boldsymbol{\beta}|\boldsymbol{m})\pi(\boldsymbol{\lambda}) \prod_{k,i,j} \lambda_{kij}^{m_{kij}}, \tag{2.10}$$

em que

$$\pi(\beta|\boldsymbol{m}) \propto \prod_{l=1}^{p}(\theta_1(\boldsymbol{x}_l'\beta))^{c_l-1}(1-\theta_1(\boldsymbol{x}_l'\beta))^{d_l-1}f(\boldsymbol{x}_l'\beta)\prod_{k,i}(\theta_i(\boldsymbol{x}_k'\beta))^{m_{ki+}}. \tag{2.11}$$

e $\pi(\boldsymbol{\lambda})$ é a distribuição *a priori* para $\boldsymbol{\lambda}$.

Para os parâmetros $\boldsymbol{\lambda}$, pode ser razoável assumir a independência a *priori* entre os conjuntos $(\lambda_{kij}, j = 0, 1), \forall k, i$ e usar a distribuição Beta para cada um.

Com o objetivo de melhorar a análise dos dados, pode-se propor algumas considerações sobre os mecanismos de censuras, que podem ser incorporadas no modelo Bayesiano de regressão apresentado em (2.9).

Hjort (1996) apresentou uma abordagem baseada na parametrização do mecanismo de censura e também no uso da família de distribuições de Dirichlet generalizadas. Posteriormente, Walker (1996) utilizou as idéias do trabalho de Hjort e as aplicou com um enfoque diferenciado em problemas de dados categorizados com censuras.

No trabalho de Hjort(1996), um vetor aleatório $\boldsymbol{x}=(x_1, x_2..., x_k)' \in S^{k-1}$ apresenta distribuição Dirichlet generalizada de parâmetros $(\alpha_1, \alpha_2, ..., \alpha_k)'$ se a sua função densidade de probabilidade for proporcional a

$$\prod_{i=1}^{k} x_i^{\alpha_i-1} g(x_1, x_2, ..., x_k), \tag{2.12}$$

em que a função $g(.)$ é uma função não negativa. Portanto, a escolha adequada de uma função $g(.)$ permite a modificação da distribuição Dirichlet. Uma opção bastante apropriada para o problema estudo é a função $g(\boldsymbol{x}) = exp\{-\delta\Delta(\boldsymbol{x})\}$, pois os valores pequenos de $\Delta(\boldsymbol{x})$ indicam a proximidade de uma determinada característica de interesse e o parâmetro de penalização δ determina o peso da importância da característica estudada.

De uma forma geral, pode-se escrever na forma de um conjunto de restrições $\Delta(\boldsymbol{\lambda}) = 0$, $i =$

$1, 2, ..., r$. Portanto, a escolha adequada de uma distribuição *a priori* é proporcional a

$$h(\boldsymbol{\lambda}) exp\left\{-\delta \sum_{i=1}^{r} \Delta_i^2(\boldsymbol{\lambda})\right\}, \tag{2.13}$$

em que $h(\boldsymbol{\lambda})$ representa o produto das funções densidade de probabilidade de Dirichlet utilizadas anteriormente como distribuições *a priori* para $\boldsymbol{\lambda}$.

O parâmetro de penalização δ permite atribuir um maior ou menor peso para a informação *a priori* e para as restrições de $\boldsymbol{\lambda}$. A escolha de diferentes valores de δ possibilitam investigar a qualidade do ajuste do modelo ao conjunto de dados. Um dos problemas encontrados é a dificuldade de encontrar os valores para δ, pois existe um critério prático para isto. Logo, o pesquisador necessita de uma certa sensibilidade e familiaridade com o problema para determinar quais os valores de δ podem ser importantes para a análise. Caso não haja muita conhecimento sobre alguns possíveis valores de δ, sugere-se o uso de valores bem diferentes com a finalidade de avaliar os resultados encontrados e, posteriormente, refinar a análise para os mais prováveis.

Logo, incluindo o fator de penalização, a distribuição *a posteriori* em (2.10) pode ser escrita como

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{m}) \propto \pi(\boldsymbol{\beta}|\boldsymbol{m}) \prod_{k,i,j} \lambda_{kij}^{m_{kij}} h(\boldsymbol{\lambda}) exp\left\{-\delta \sum_{i=1}^{r} \Delta_i^2(\boldsymbol{\lambda})\right\}. \tag{2.14}$$

O próximo modelo proposto é inovador para analisar dados censurados com variáveis respostas binárias e varíaveis explicativas, sejam elas discretas ou contínuas.

2.2.2 Modelo Bayesiano particionado com censura

Considere uma amostra aleatória de tamanho t com variável resposta de distribuição Bernoulli com censura, cujos valores de sucessos (S), fracassos (N) e censuras (C) podem ser expressos pelo vetor aleatório $(\boldsymbol{S}, \boldsymbol{N}, \boldsymbol{C})$, que apresenta distribuição trinomial com parâmetros $\boldsymbol{t}$ e $\boldsymbol{p}$=(p_1, p_2, p_3), em que $p_1 + p_2 + p_3 = 1$, $0 < p_i < 1$, $i = 1, 2, 3$ e $t = n_1 + n_2 + n_3$.

Portanto, utilizando a função distribuição em (1.9), a função de verossimilhança pode ser escrita

como

$$\Pi(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{n}, \boldsymbol{x}) \propto \prod_{i=1}^{t} \pi_i^s (1-\pi_i)^{t-s} \theta_i^n (1-\theta_i)^{m-n}. \tag{2.15}$$

Logo, a distribuição *a posteriori* pode ser escrita como

$$\Pi(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{n}, \boldsymbol{x}) \propto \prod_{i=1}^{n} \left(x_i'\beta\right)^s \left(1 - x_i'\beta\right)^{t-s} \theta^n (1-\theta)^{m-n} f\left(x'\beta\right) h(\theta), \tag{2.16}$$

em que $f\left(x'\beta\right)$ representa a função logística, $h(\theta)$ representa a distribuição *a priori* para o parâmetro correspondente as censuras, que pode ter distribuição Beta $(0, 1)$, por exemplo. Para simplificar o modelo, considera-se que o parâmetro θ é o mesmo para todas as unidades experimentais, ou seja, a probabilidade de censura de todas as unidades experimentais são iguais.

Como motivação da análise Bayesiana do modelo de regressão, considere o exemplo de dados reais de um estudo clínico apresentado a seguir.

2.2.3 Aplicação dos Modelos Bayesianos

Um estudo da infecção por papillomavirus humano (HPV) foi realizado pela Universidade da Califórnia - São Francisco ($UCSF$). O HPV é formado por um grupo de vírus responsável por várias lesões epiteliais. Cerca de 30 subtipos têm preferências pelos tecidos genitais e estão frequentemente associados a presença de câncer uterino em mulheres. Como os testes de HPV são limitados a um grupo de vírus, portanto podem existir infecções não detectadas, afetando a variável resposta. Os dados da Tabela 2.1 foram originalmente apresentados por Moscicki *et al* (2001). O objetivo é examinar a possível relação em potenciais níveis de risco com a infecção cervical do HPV de mulheres que apresentaram teste negativo no ínicio do estudo. Cerca de 104 mulheres com idades entre 13 a 21 anos participaram da pesquisa clínica. O status da infecção no final do estudo foi registrado para cada mulher por meio de teste das amostras uterinas. Foram anotadas também se a mulher tinha um histórico de verrugas vulvares (VV), se havia tido um novo parceiro nos últimos dois meses (NP) e se tinha um histórico de herpes vaginal(HV). O tempo médio de acompanhamento foi de 26 meses para as mulheres que permaneciam negativas quanto ao HPV.

Tabela 2.1: Dados Infecção do HPV

$x_k = (VV, NP, HV)$*	número de casos com HPV	total de observações
(0,0,0)	12	44
(0,0,1)	1	2
(0,1,0)	29	40
(0,1,1)	3	3
(1,0,0)	6	9
(1,1,0)	1	4
(1,1,1)	2	2

*VV: verrugas vulgares; NP: número de parceiros; HV: herpes vaginal.

Para avaliar a importância de uma ou mais variáveis explicativas no modelo Bayesiano para analisar a variável resposta da presença ou ausência da infecção do HPV, considere o interesse em testar a hipótese nula $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. H_1 : pelo menos um $\beta_i \neq 0$, $i = 1, 2, 3$, utilizando o procedimento do $FBST$.

É importante lembrar que a presença de censuras está relacionada apenas a variável resposta e não as explicativas. Portanto, existem apenas dois parâmetros: λ_{01}=P(falso positivo)=1-especificidade e λ_{10}=P(falso negativo)=1-sensibilidade.

As informações sobre a especificidade e sensibilidade dos diagnósticos de HPV de uma amostra conhecida para as distribuições *a priori* podem ser encontradas em Moscicki et al. (2001) e Paulino et al.(2003).

O modelo logístico ajustado para os dados é

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = -1.06 + 0.372VV + 1.623NP + 0.327HV.$$

Aplicando o $FBST$ para testar as hipóteses e utilizando a tranformação logística, um algoritmo em R foi implementado para calcular o valor aproximado da evidência. Após realizar um estudo de simulação, encontrou-se o valor da evidência igual a $0,0265$ para o primeiro modelo logístico e $0,0157$ para o modelo Bayesiano particionado. Logo, pode-se concluir que existe pelo menos um $\beta_i \neq 0$ no modelo Bayesiano de regressão, ou seja, existe, no mínimo, uma covariável que apresenta um efeito

Tabela 2.2: Valor aproximado da evidência *ev* dos modelos logísticos

Modelo	VV, NP HV	VV NP	VV HV	NP HV	NP
Paulino, Soares e Neuhaus	0,0265	0,021	0,098	0,029	0,010
Particionado	0,0157	0,012	0,091	0,031	0,011

estatisticamente significativo na variável resposta referente ao diagnóstico positivo de HPV. A Tabela 2.2 apresenta os valores da evidência *ev* calculados para os dois modelos logísticos.

Considerando os resultados apresentados na Tabela

$$log\left(\frac{\pi(x)}{1-\pi(x)}\right) = -2,66 + 2,623NP. \tag{2.17}$$

Portanto, mulheres com trocas mais frequentes de parceiros tem mais chances em adquirir a infecção pelo HPV e podem posterioremente desenvolverem câncer uterino em virtudo do contagio. É importante ressaltar que as campanhas de saúde para a prevenção da infecção do HPV devem estar associadas a das doenças sexualmente transmissíveis.

A área de modelagem Bayesiana para dados categorizados com censura é bastante rica. Alguns modelos mais sofisticados podem ser propostos considerando também as censuras nas varíaveis independentes, com o intuito de estudar seus efeitos na variável resposta.

2.3 Modelo Dose Resposta

A avaliação da dose-resposta é caracterizada pela relação existente entre a dose de um agente administrado e a incidência de um efeito de interesse em uma amostra de indivíduos. O termo dose é utilizado para indicar a quantidade do agente, enquanto o termo resposta corresponde ao efeito do agente que foi administrado. Geralmente, o aumento da dose de um agente tóxico pode resultar em um aumento na incidência de um efeito de interesse, assim como na severidade do efeito.

Nas primeiras etapas de um estudo clínico, a variável observada Y tem resposta binária, por exemplo, para um agente com probabilidade de sucesso $p(d)$, em que d é o nível da dose do agente. O

Tabela 2.3: Frequências de micronúcleos em doses de radiação

i	Dose(cGy)	y_{i0}	y_{i1}	y_{i2}	n_i
0	5	481	17	2	500
1	10	477	19	4	500
2	25	471	24	5	500
3	50	450	44	6	500
4	100	431	59	10	500
5	200	339	140	21	500
6	300	304	132	64	500
7	400	240	189	72	501
8	500	174	197	129	500
9	600	122	173	211	506

sucesso do agente pode ser, por exemplo, a diminuição do número de celulas cancerígenas, ausência de dor, incidência de uma determinada doença. Ao determinar o número de sucessos em vários níveis de dose, pode-se estudar a relação existente entre a dose e as respostas observadas. Um dos interesses principais destes estudos é determinar a dose effetiva com probabilidade de sucesso igual a $\alpha\%$ (DE_α), $\alpha \in [0, 100]$. O problema da dose resposta inversa é a denominação dada para o estudo da DE_α. O modelo de regressão logística é frequentemente utilizado em Toxicologia no estudo de desenvolvimento de drogas para determinar níveis de doses que possuem toxidade e/ou efetividade. Hu, Ji e Tsui (2008); Madruga, Pereira e Rabello-Gay (1994), e Madruga, Okazaki, Pereira e Rabello-Gay (1996) abordam o assunto com um enfoque Bayesiano.

Como ilustração, será considerado o conjunto de dados apresentado em Balasem e Ali (1991) referente a um estudo de doses de radiação versus frequências de celulas com aberrações citogenéticas. Os dados do estudo citogenético são apresentados na Tabela 2.3.

O estudo de citogenética observou as frequências de micronúcleos obtidas em culturas de linfócitos expostas a certas doses de radiação. Após o processo de divisão celular, os micronúcleos podem conter um ou mais fragmentos cromossômicos, originados de uma simples quebra ou de um tipo de aberração. O experimento consistiu em observar as frequências de núcleos celulares divididas em três categorias: sem micronúcleos, com um micronúcleo ou com dois ou mais micronúcleos. O experimento foi planejado para dez níveis diferentes de radiação: $5, 10, 25, 50, 100, 200, 300, 400, 500$

e 600 cGy.

Tabela 2.4: Parâmetros da distribuição *a priori* e *a posteriori*

i	Dose(cGy)	a_{i0}	a_{i1}	a_{i2}	$a_{i.}$	A_{i0}	A_{i1}	A_{i2}	$A_{i.}$
0	5	12	2	1	15	493	19	3	515
1	10	11	2	1	14	488	21	5	514
2	25	10	2	1	13	481	26	6	513
3	50	9	2	1	12	459	46	7	512
4	100	8	2	1	11	439	61	11	511
5	200	7	2	1	10	346	142	22	510
6	300	6	2	1	9	310	134	65	509
7	400	5	2	1	8	245	191	73	509
8	500	4	2	1	7	178	199	130	507
9	600	3	2	1	6	125	175	212	512

Para cada nível i de dose de radiação, $i = 0, 1, 2, ..., 9$, as componentes do vetor $\boldsymbol{y_i}=(y_{i1}, y_{i2}, y_{i3})$ correspondem ao número de células sem micronúcleos, com um micronúcleo e com dois ou mais micronúcleos, respectivamente. Portanto, para cada dose i, pode-se assumir um modelo trinomial com parâmetros $(n_i, \pi_{i0}, \pi_{i1}, \pi_{i2})$, em que $n_i = y_{i0} + y_{i1} + y_{i2}$ e $\pi_{i0} + \pi_{i1} + \pi_{i2} = 1$ com $0 \leq \pi_{ij} \leq 1$, $j = 0, 1, 2$. A Tabela 2.4 apresenta a distribuição *a priori* para π_i, cuja escolha reflete o conhecimentos biológicos: π_{i0} decresce para doses elevadas, enquanto π_{i1} e π_{i2} aumentam ao longo das doses de radiação. A Tabela 2.4 também apresenta a distribuição *a posteriori* para π_i, $i = 1, 2, ..., 9$.

Para determinado valores de dose de radiação, observou-se uma proporção de respostas y_i. No entanto, pode ser necessário estimar uma determinada dose d para uma determinada proporção de interesse. Para estudar a relação entre dose e resposta, será utilizada a transformação

$$\theta_j = ln\frac{\pi_j}{\pi_0} \tag{2.18}$$

em que π_0 representa a proporção de células sem micronúcleos, $j = 1, 2$. A transformação inversa é dada por

$$\pi_j = \frac{exp(\theta_j)}{1 + exp(\theta_1) + exp(\theta_2)}. \tag{2.19}$$

Tabela 2.5: Parâmetros da distribuição a posteriori de $(\theta_{i1}, \theta_{i2})$

i	Dose(cGy)	m_{i1}	m_{i2}	s^2_{i1}	s^2_{i2}	c_i
0	5	-3.28	-5.28	0.06	0.4	0.002
1	10	-3.17	-4.68	0.05	0.22	0.002
2	25	-2.93	-4.46	0.04	0.18	0.0021
3	50	-2.31	-4.26	0.02	0.16	0.0022
4	100	-1.98	-3.73	0.02	0.10	0.0023
5	200	-0.89	-2.77	0.01	0.05	0.0029
6	300	-0.84	-1.56	0.01	0.02	0.0032
7	400	-0.25	-1.22	0.01	0.02	0.0041
8	500	0.11	-0.32	0.01	0.01	0.0056
9	600	0.34	0.53	0.01	0.01	0.008

Como a distribuição π tem distribuição Dirichlet, tem-se que a distribuição da transformação (θ_1, θ_2) é Normal bivariada com parâmetros calculados por 1.41. A Tabela 2.5 apresenta os parâmetros da distribuição*a posteriori* de (θ_1, θ_2). O modelo de dose-resposta para estes dados é

$$- m_{ij} = \alpha_j + \frac{\beta_j}{\gamma_j + d_i}, \tag{2.20}$$

em que d_i é a dose no nível i, (m_{i1}, m_{i2}) são as médias *a posteriori* de (θ_1, θ_2) para $i = 1, 2, ..., 9$. As estimativas dos parâmetros são apresentados na Tabela 2.6. Portanto, a relação entre doses d_i e as proporções π_{ij} pode ser dada pela curva

$$\pi_{ij} = \frac{exp(-\alpha_j + \frac{\beta_j}{\gamma_j + d_i})}{1 + exp(-\alpha_1 + \frac{\beta_1}{\gamma_1 + d_i}) + exp(-\alpha_2 + \frac{\beta_2}{\gamma_2 + d_i})}. \tag{2.21}$$

As Figuras 2.1 a 2.3 apresentam as curvas ajustados e os valores das proporções observadas das células considerando a presença de micronúcleos nas doses de radiação. Analisando o comportamento das três curvas, seria natural propor um teste para avaliar em qual nível de dose de radiação a proporção de células sem micronúcleos é igual aproporção de células com um micronúcleo e também a proporção de células com dois ou mais micronúcleos. Portanto, tem-se interesse em testar $H_0 : (\theta_1, \theta_2) = (0, 0)$. Utilizando os valores da distribuição *a posteriori* de (θ_1, θ_2), o procedimento do $FBST$ será aplicado para testar a hipótese nula. O valor aproximado da evidência para a dose 500

cGy é igual a $0,0213$ e, para as demais doses, o valor da evidência encontrado foi $< 0,001$. Portanto, pode-se concluir que a proporção de células sem micronúcleos são diferentes das de um micronúcleo e também de dois ou mais micronúcleos para as doses estudadas de radiação.

A análise de dose-resposta para dados na presença de censura pode ser realizada da mesma forma que o exemplo citogénetico apresentado anteriormente. Com esta abordagem, pode-se investigar inclusive as informações fornecidas pelas censuras e avaliar algumas hipóteses de interesse para o estudo.

Tabela 2.6: Parâmetros da distribuição a posteriori de $(\theta_{i1}, \theta_{i2})$

j	α_j	β_j	γ_j
1	-1,792	1303,348	254,149
11	-2.012	22033,864	1298,498

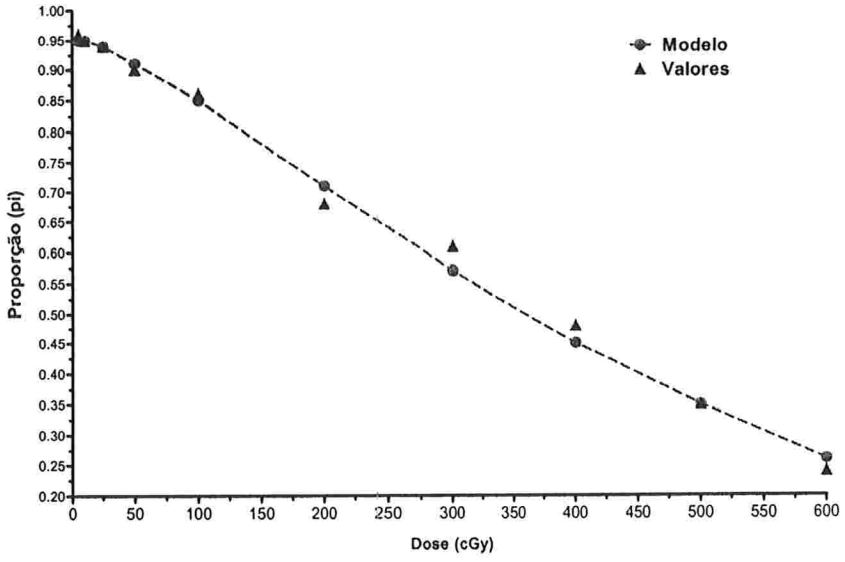

Figura 2.1: Proporção de células sem micronúcleos.

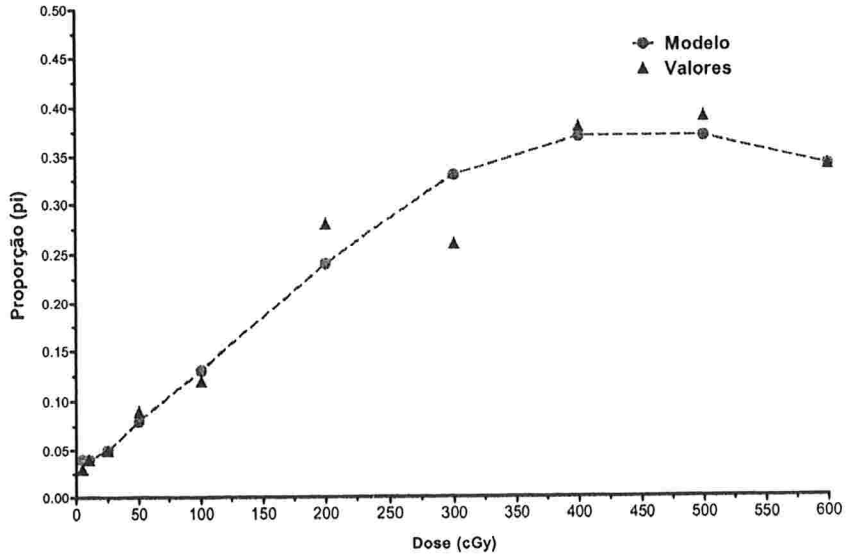

Figura 2.2: Proporção de células com um micronúcleo.

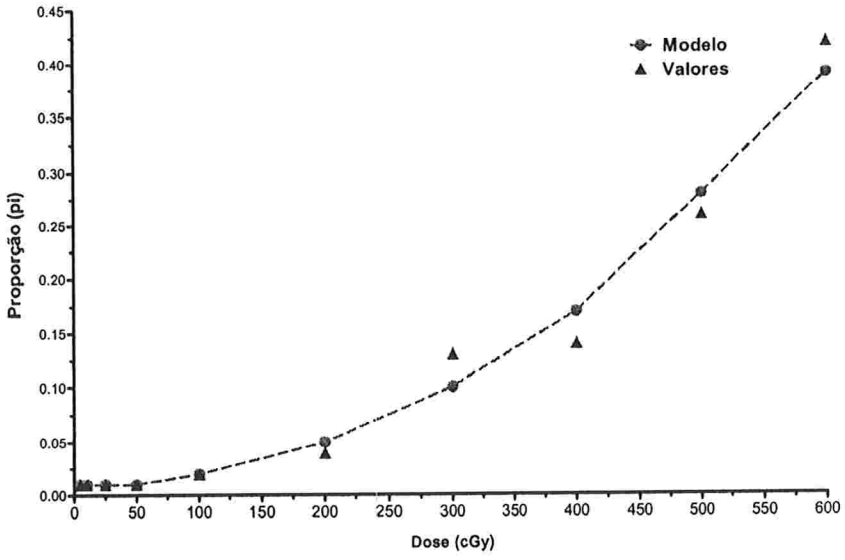

Figura 2.3: Proporção de células com dois ou mais micronúcleos.

Capítulo 3

Conclusões

3.1 Considerações Finais

O uso do $FBST$ pode ser bastante útil na análise estatística com o objetivo de testar algumas hipóteses de interesse em diferentes problemas estatísticos. Associado ao modelo Bayesiano para dados discretos com censura, tem-se uma metodologia bastante interessante porque a análise estatística se torna mais informativa na tomada de uma decisão.

O $FBST$ apresenta uma forma bastante intuitiva para o cálculo da evidência em algumas hipóteses estatísticas de interesse. Além disso, possui boas propriedades teóricas e de fácil interpretação. No entanto, a implementação computacional do $FBST$ pode ser vista como um pequeno problema a ser superado, pois pode exigir um determinado conhecimento de métodos de simulação.

3.2 Sugestões para Pesquisas Futuras

- Análise de dados categorizados com censuras fornece uma vasta área de pesquisa, pois a informação proveniente das censuras podem ser modeladas de diferentes formas;
- Novos modelos logísticos podem ser estudados realizando algumas restrições quanto a conjuntos de categorias em que as originam;
- Incorporar censuras nas varíaveis explicativas nos modelos de regressão logística também podem ser estudadas e adequar a alguns problemas práticos.
- Estudar o efeito de censuras em modelos log-lineares também pode ser bastante interessante.

Apêndice A

Programas

```
##############################################################################
#programa de simulação da aplicação exemplo testes odontológicos
##############################################################################
library(MSBVAR)

n = 20
#dimensão da matriz dos parâmetros
VAR0 = matrix(c(0.12,0.10,0.10,0.10,0.10,0.10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.10,
0.20,0.10,
0.10,0.10,0.10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.10,0.10,0.20,0.10,0.10,0.10,0,0,0,
0,0,0,0,0,
0,0,0,0,0,0,0.10,0.10,0.10,0.20,0.10,0.10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.10,0.10,
0.10,0.10,
0.20,0.10,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.10,0.10,0.10,0.10,0.10,0.20,0,0,0,0,0,0,
0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0.31,0.16,0.16,0.16,0.16,0.16,0.16,0.16,0,0,0,0,0,0,0,0,0,0,
0,0.16,
0.31,0.16,0.16,0.16,0.16,0.16,0.16,0,0,0,0,0,0,0,0,0,0,0,0,0,0.16,0.16,0.31,0.16,
0.16,0.16,
0.16,0.16,0,0,0,0,0,0,0,0,0,0,0,0,0.16,0.16,0.16,0.31,0.16,0.16,0.16,0.16,0,0,
0,0,0,0,0,0,
0,0,0,0,0.16,0.16,0.16,0.16,0.31,0.16,0.16,0.16,0,0,0,0,0,0,0,0,0,0,0,0,0.16,
```

```
0.16,0.16,0.16,
0.16,0.31,0.16,0.16,0,0,0,0,0,0,0,0,0,0,0,0,0.16,0.16,0.16,0.16,0.16,0.16,0.31,
0.16,0,0,
0,0,0,0,0,0,0,0,0,0,0.16,0.16,0.16,0.16,0.16,0.16,0.16,0.31,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,
0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,0,
0,0,0,0,0,
0,0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,
0,0,0,0,0,
0,0,0,0,0,0,0,0,1.29,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.29), ncol=n,
 byrow=TRUE)
MAXI = (2*pi)^(-n/2)*(abs(det(VAR0))^(-1/2))
# acima sob espaco theta0

MED1 = c(0.0811,-0.0111,0.0375,-0.0111,-0.1000,-0,0111,-0.0985,0.0405,-4.4070,
-0.0531,0.0300,
-0.1721,-0.3119,0.0494,0,0,0,0,0)
VAR1 = matrix(c(0.1137,0.0951,0.0951,0.0951,0.0951,0.0951,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,
0.0951,0.2002,0.0951,0.0951,0.0951,0.0951,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.0951,0.0951,0.1556,0.0951,0.0951,0.0951,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.0951,0.0951,0.0951,0.2002,0.0951,0.0951,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.0951,0.0951,0.0951,0.0951,0.2764,0.0951,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.0951,0.0951,0.0951,0.0951,0.0951,0.2002,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0.3403,0.1293,0.1293,0.1293,0.1293,0.1293,0.1293,0.1293,0,0,
0,0,0,0,
0,0,0,0,0,0,0,0.1293,0.2226,0.1293,0.1293,0.1293,0.1293,0.1293,0.1293,0,0,
0,0,0,0,
0,0,0,0,0,0,0,0.1293,0.1293,1.3677,0.1293,0.1293,0.1293,0.1293,0.1293,0,0,
0,0,0,0,
0,0,0,0,0,0,0,0.1293,0.1293,0.1293,0.3036,0.1293,0.1293,0.1293,0.1293,0,0,
0,0,0,0,
0,0,0,0,0,0,0,0.1293,0.1293,0.1293,0.1293,0.2320,0.1293,0.1293,0.1293,0,0,
```

```
0,0,0,0,
0,0,0,0,0,0,0.1293,0.1293,0.1293,0.1293,0.1293,0.3965,0.1293,0.1293,0,0,
0,0,0,0,
0,0,0,0,0,0,0.1293,0.1293,0.1293,0.1293,0.1293,0.1293,0.4929,0.1293,0,0,
0,0,0,0,
0,0,0,0,0,0,0.1293,0.1293,0.1293,0.1293,0.1293,0.1293,0.1293,0.2145,0,0,
0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.2897,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.2897,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.2897,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.2897,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.2897,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.2897), ncol=n, byrow=TRUE)
#inicio da simulacao da amostras Normais multivariadas com matriz de
 medias e variancia
 de cov
definida acima
fdens = function(x, n, vmed, mvar) (2*pi)^(-n/2) * (abs(det(mvar))^(-1/2))
 * exp(-1/2 * matrix(x-vmed, nrow=1)%*%solve(mvar)%*%matrix(x-vmed, ncol=1))
# R é orientado a matriz (operações ponto a ponto num vetor e matriz e,
para fazer
operações matricias, precisa usar %*%)

set.seed(090) #semente aleatoria para geraçao numeros pseudo-aleatorios;
contador = 0
passo = 1
var.evid=10

repeat{
  y = rmultnorm(1, MED1, VAR1, tol = 1e-10)
  f = fdens(y, n, MED1, VAR1)

  if(f>MAXI) contador = contador+1
```

```
  evidencia = contador/passo
  var.evid = evidencia*(1-evidencia)/passo

  if(var.evid<10^(-5) & passo>10) break

  if(!(passo%%100)) print(passo)
  passo = passo+1
}
evidencia
var.evid
passo
```

```
##############################################################################
###############
# programa para simulacao da aplicação regressão logistica para exemplo HPV
##############################################################################
###############

library(MSBVAR)

n = 8
VAR0 = matrix(c(2.305,1.644,0.000,0.000,0.000,0.000,0.000,0.000,1.644,3.200,
0.000,0.000,
0.000,0.000,0.000,0.000,0.000,0.000,3.289,1.645,0.000,0.000,0.000,0.000,0.000,
0.000,
1.645,2.290,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,3.290,1.645,0.000,
0.000,
0.000,0.000,0.000,0.000,1.645,2.290,0.000,0.000,0.000,0.000,0.000,0.000,0.000,
0.000,
```

```
3.290,1.645,0.000,0.000,0.000,0.000,0.000,0.000,1.645,2.524), ncol=n,
 byrow=TRUE)
MAXI = (2*pi)^(-n/2)*(abs(det(VAR0))^(-1/2))
# sob H0

MED1 = c(0.961,0.038,0.000,1.000,0.000,1.000,0.000,1.000)
VAR1 = matrix(c(2.305,1.644,0.000,0.000,0.000,0.000,0.000,0.000,1.644,
3.200,0.000,0.000,
0.000,
0.000,0.000,0.000,0.000,0.000,3.289,1.645,0.000,0.000,0.000,0.000,0.000,
0.000,1.645,
2.290,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,3.290,1.645,0.000,
0.000,0.000,
0.000,0.000,0.000,1.645,2.290,0.000,0.000,0.000,0.000,0.000,0.000,0.000,
0.000,3.290,
1.645,0.000,0.000,0.000,0.000,0.000,0.000,1.645,2.524), ncol=n,
 byrow=TRUE)

fdens = function(x, n, vmed, mvar) (2*pi)^(-n/2) * (abs(det(mvar))^(-1/2))
 * exp(-1/2 * matrix(x-vmed, nrow=1)%*%solve(mvar)%*%matrix(x-vmed, ncol=1))
#perceba que o R é orientado a matriz (operações ponto as ponto num vetor
 e matriz e, para fazer operações matricias, precisa usar %*%, por exemplo)

set.seed(4756) #semente aleatoria do time do DOS; mude isso para cada
experiemnto seu; sorteie um numero aleatorio de 3 digitos (de uma tabela,
 por exemplo) e coloque aí
contador = 0
passo = 1

repeat{
  y = rmultnorm(1, MED1, VAR1, tol = 1e-10)
  f = fdens(y, n, MED1, VAR1)
  if(f>MAXI) contador = contador+1
```

```
  evidencia = contador/passo
  var.evid = evidencia*(1-evidencia)/passo

  if(var.evid<10^(-5) & passo>100) break

  if(!(passo%%100)) print(passo)
  passo = passo+1
}

evidencia

var.evid

passo



install.packages("MSBVAR") #faça isso uma soh vez, para instalar a library
 na sua maquina


library(MSBVAR)

n = 15
VAR0 = matrix(c(0.5,4.9,4.9,4.9,4.9,4.9,4.9,4.9,4.9,0,0,0,0,0,0,4.9,0.8,4.9,4.9,
4.9,4.9,4.9,
4.9,4.9,0,0,0,0,0,0,4.9,4.9,1.5,4.9,4.9,4.9,4.9,4.9,4.9,0,0,0,0,0,0,4.9,4.9,4.9,
6,4.9,4.9,
4.9,4.9,4.9,0,0,0,0,0,0,4.9,4.9,4.9,4.9,2.9,4.9,4.9,4.9,4.9,0,0,0,0,0,0,4.9,4.9,
4.9,4.9,
4.9,6.7,4.9,4.9,4.9,0,0,0,0,0,0,4.9,4.9,4.9,4.9,4.9,4.9,1.9,4.9,4.9,0,0,0,0,0,0,
4.9,4.9,
```

```
4.9,4.9,4.9,4.9,4.9,3.9,4.9,0,0,0,0,0,0,4.9,4.9,4.9,4.9,4.9,4.9,4.9,4.9,5.1,0,0,
0,0,0,0,
0,0,0,0,0,0,0,0,0,5.3,5.2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5.2,2.8,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,
0,3.8,2.7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2.7,9.7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.8,
2.1,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,2.1,2.7), ncol=n, byrow=TRUE)
MAXI = (2*pi)^(-n/2)*(abs(det(VAR0))^(-1/2))
#1.502303e-09 sob H0

MED1 = c(12,25,10,9,6,3,15,2,6,7,9,5,6,3.2,8.1)
VAR1 = matrix(c(0.3,3.8,3.8,3.8,3.8,3.8,3.8,3.8,3.8,0,0,0,0,0,0,3.8,0.6,3.8,
3.8,3.8,3.8,
3.8,3.8,
3.8,0,0,0,0,0,0,3.8,3.8,1.8,3.8,3.8,3.8,3.8,3.8,3.8,0,0,0,0,0,0,3.8,3.8,3.8,
5.7,3.8,3.8,
3.8,3.8,3.8,0,0,0,0,0,0,3.8,3.8,3.8,3.8,1.5,3.8,3.8,3.8,3.8,0,0,0,0,0,0,3.8,
3.8,3.8,3.8,
3.8,1.8,3.8,3.8,3.8,0,0,0,0,0,0,3.8,3.8,3.8,3.8,3.8,3.8,3.8,3.2,3.8,3.8,0,0,0,0,
0,0,3.8,3.8,
3.8,3.8,3.8,3.8,3.8,3.8,3.8,0,0,0,0,0,0,3.8,3.8,3.8,3.8,3.8,3.8,3.8,3.8,3.8,6.5,
0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,2.8,2.7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2.7,1.1,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,
0,2.8,2.9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2.9,3.7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1.6,1.7,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,1.7,2.8), ncol=n, byrow=TRUE)

fdens = function(x, n, vmed, mvar) (2*pi)^(-n/2) * (abs(det(mvar))^(-1/2))
 * exp(-1/2 * matrix(x-vmed, nrow=1)%*%solve(mvar)%*%matrix(x-vmed, ncol=1))
#perceba que o R é orientado a matriz (operações ponto as ponto num vetor e
 matriz e, para fazer operações matricias, precisa usar %*%, por exemplo)
```

```
set.seed(090) #semente aleatoria do time do DOS; mude isso para cada experiemnto
seu; sorteie um numero aleatorio de 3 digitos (de uma tabela, por exemplo) e coloque aí
contador = 0
passo = 1

repeat{
  y = rmultnorm(1, MED1, VAR1, tol = 1e-10)
  f = fdens(y, n, MED1, VAR1)
  if(f>MAXI) contador = contador+1

  evidencia = contador/passo
  var.evid = evidencia*(1-evidencia)/passo

  if(var.evid<10^(-5) & passo>10) break

  if(!(passo%%100)) print(passo)
  passo = passo+1
}

#evidencia

#var.evid
```

Referências Bibliográficas

[1] Agresti, A. e Hitchcook, D. B.(2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14, 297-330.

[2] Aitchison, J.(2003). The Statistical Annalysis of Compositional Data. *Chapter 3*, 126-128, Chapman and Hall.

[3] Albert, J. H. e Gupta, A. K.(1983). Bayesian estimation methods for 2 x 2 contigency tables using mixtures of Dirichlet distribuitions. *Journal of the American Statistical Association*, 78, 708-717.

[4] Antelman, G.R. (1972). Interrelated Bernoulli Processes. *Journal of the American Statistical Association* , 67, 831-841.

[5] Basu, D. e Pereira, C. A. B. (1982). On the bayesian analysis of categorical data: the problem of nonresponse. *Journal of Statistical Planning and Inference*, 6, 345-362.

[6] Bedrick, E. J.; Christensen, R. e Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of American Statistical Association*, 91, 1450-1460.

[7] Chen, T. e Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

[8] Cox (1975). Partial likelihood. *Biometrika*, 62, 269-276.

[9] Dempster, A. P, Laird, N. M. e Rubin, E. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39, 1-38.

[10] Devroye, L.(1992). Random variate generation for the Digamma e Trigamma distributions. *Journal Statist. Comput. Simul*, 43, 197-216.

[11] Dickey, J.M., Jiang, J. M. e Kadane, J. B. (1987). Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, 82, 773-781.

[12] Evans, M. (1997). Bayesian inference procedure derived via the concept of relative surprise. *Communications in Statistics*, 26, 1125-1143.

[13] Gunel, E. (1984). A bayesian analysis of the multinomial model for a dichotomous response with nonrespondents. *Communications in Statistics - Theory and Methods*, 13, 737-751.

[14] Hartley, H. O. (1958). Maximum Likelihood estimation from incomplete data. *Biometrics*, 14, 174-194.

[15] Hjort, N. L. (1996). Bayesian Approaches to non and semiparametric density estimation. *Bayesian Statistics*, 5, 223-253.

[16] Hu, B., Ji, Y. e Tsui, K. W. (2008). Bayesian estimation of inverse dose response. *Biometrics*, 68, 1223-1230.

[17] Irony, T. Z. e Pereira, C. A. B. (1995). Bayesian Hypothesis test: using surface integrals to distribute prior information among the hypotheses. *Resenhas IME-USP*, Vol.2, 1, 27-46.

[18] Karson, M. J. e Wrobleski, W. J. (1970). A bayesian analysis of binomial data with a partially information category. *In Proceeding of the Business and Economic Statistics Section, American Statistical Association*, 532-534.

[19] Kaufman, G. M. (1973). A bayesian analysis of nonresponse in dichotomous processes. *Journal of the American Statistical Association*, 68, 670-678.

[20] Lauretto, M., Pereira, C. A. B., Stern, J. M. e Zacks, S. (2003). Full bayesian significance test applied to multivariate normal structure models. *Brazilian Journal of probability and Statistics*, 17, 147-168.

[21] Little, R. J. A. e Rubin, D. B. (1987). *Statistical Analysis of Missing Data*, New York, John Wiley & Sons.

[22] Madruga, M. R., Pereira, C. A. B. e RAbello-Gay, M. N. (1994). Bayesian dosimetry: radiation dose versus frequencies of cells with aberrations.*Environmetrics*, 5, 47-56.

[23] Madruga, M. R., Ochi-Lohmann, T. H., Okazaki, K., Pereira, C. A. B. e RAbello-Gay, M. N. (1996). Bayesian dosimetry II: credibility intervals for radiation dose.*Environmetrics*, 7, 325-331.

[24] Madruga, M. R., Pereira, C. A. B. e Stern, J. M. (2003). Bayesian evidence test for precise hypotheses.*J. of Stat. Plann. and Inference*, 117, 185-198.

[25] Moscicki, A. B., Hills, N., Shiboski, S et al. (2001). Risks for incident human papillomavirus infection and low-grade squamous intraepithelial lesion development in young females.*Journal of American Medical Assoc.*, 285, 2995-3002.

[26] Paulino, C. D. M. e Pereira, C. A. B. (1995). Bayesian analysis of categorical data informatively censored.*Comm. Statist. -Theory method.*, 21, 2689-2705.

[27] Paulino, C. D. M. e Pereira, C. A. B. (1995). Bayesian methods for categorical data under informative general censoring.*Biometrika*, 82, 439-46.

[28] Paulino, C. D. M., Soares, P. e Neuhaus, J. (2003). Binomial Regression with misclassification.*Biometrics*, 59, 670-675.

[29] Pereira, C. A. B. e Barlow, R. E. (1990). Medical Diagnosis using influence diagrams.*Networks*, 20, 565-577.

[30] Pereira, C. A. B. e Pericchi, L. R. (1990). Analysis of Diagnosability.*Appl. Statist.*, 39, n.2, 189-204.

[31] Pereira, C. A. B. e Stern, J. M. (1999). Evidence and Credibility: full bayesian significance test for precise hypotheses.*Entropy*, 1, 69-80.

[32] Pereira, C. A. B. e Stern, J. M. (2001). Full bayesian significance tests for coefficients of variation.*Bayesian Methods with applications to Statistics*, 391-400, Monographs of Official Statistics, Eurostat.

[33] Pereira, C. A. B. e Stern, J. M. (2001). Model selection: Full Bayesian approach.*Environmetrics*, 12, 559-568.

[34] Pereira, C. A. B. e Stern, J. M. (2008). An essay on the role of Bernoulli and Poisson processes in Bayesian Statistics. To appear in Statistical Journal.

[35] Poleto, F. Z. (2006). Análise de Dados categorizados com omissão. *Dissertação de Mestrado*, Universidade de São Paulo, Instituto de Matemática e Estatística.

[36] Smith, P. J., Choi, S. C. e Gunel, E. (1985). Bayesian analysis of a 2 x 2 contingency table with both completely and partially cross-classified data. *Journal of Education Statistics*, 10, 31-43.

[37] Soares, P. e Paulino, C. D (2001). Incomplete Categorical data analysis:a bayesian perspective. *J. Statist. Comput. Simul.*, 69, 157-170.

[38] Soares, P. J. J. (2004). Análise Bayesiana de dados deficientemente categorizados. *Dissertação de Doutorado*, Universidade Técnica de Lisboa, Instituto Superior Técnico.

[39] Stern, J. M. e Zacks, S. (2002). Testing the independence of Poisson variates under the Holgate bivariate distribuition: the power of a new evidence test. *Statistical and Probability Letters*, 60, 313-320.

[40] Stern, J. M. (2003). Significance tests, belief calculi and burden of proof in legal and scientific discourse. Laptec - 2003. *Frontiers in artificial intelligence and its applications*, 101, 139-147.

[41] Walker, S. (1996). A bayesian maximum a posteriori algorithm for categorical data under informative general censoring. *The Statistician*, 45 (3), 293-298.

[42] Zacks, S. e Stern, J. M. (2003). Sequential estimation of ratios with application to bayesian analysis. *RT-MAC-IME-USP 2003-10.*