SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito.....14/05/99

Assinatura:.....*Paulo Sérgio Luri*.....

**Inferência Bayesiana para pesquisa
de mercado com erros de resposta
utilizando modelos mistos**

Vanda Donizetti Redondo Silveira

Orientador : Prof. Dr. Josemar Rodrigues

*Dissertação apresentada ao Instituto de Ciências Matemáticas de
São Carlos - USP, como parte dos requisitos para a obtenção
do título de Mestre em Ciências - Área de Ciências de Computação e
Matemática Computacional.*

USP - São Carlos

Maio de 1999

Errata

Na página 25, inserir no final do primeiro parágrafo após a equação (3.2.1): incluindo o vetor $p = (p_1, \dots, p_k)$ também desconhecido.

Na página 37, substituir no modelo misto considerado, o parâmetro $(1-\theta_2)$ da distribuição de Bernoulli, para a variável X_1 por $(1-\theta_1)$, isto é:

$$f(x | p, \theta_1, \theta_2) = p \text{ Bern}(x | 1 - \theta_1) + (1 - p) \text{ Bern}(x | \theta_2).$$

Na página 47, substituir $p^* = \{p^{(l)}, l = 1, \dots, 50\}$, por $p^* = \{p^{(l)}, l = 1, \dots, 300\}$; $q^* = \{p^{(l)}, l = 1, \dots, 50\}$, por $q^* = \{p^{(l)}, l = 1, \dots, 300\}$; e $p^{\text{rep},j} = \{p_1^{\text{rep},j}, \dots, p_{50}^{\text{rep},j}\}$, por $p^{\text{rep},j} = \{p_1^{\text{rep},j}, \dots, p_{300}^{\text{rep},j}\}$, respectivamente.

*"Se você pensa que pode ou sonha que pode,
comece. Ousadia tem genialidade, poder e
magia. Ouse fazer e o poder lhe será dado."
(Goethe)*

*"A maior revolução de todos os tempos é a
descoberta de que ao mudar as atitudes
internas de suas mentes, os seres humanos
podem mudar os aspectos externos de sua vida."
(William James)*

- *Ao meu pai que, com paciência, confiança e carinho, tornou meu caminho mais brando.*
- *À minha mãe, pela dedicação, amor, carinho e pelos dias a mais que passamos juntas nesta caminhada.*
- *Ao meu marido, pelo incentivo, apoio e confiança.*
- *À minha irmã Angélica, pela hospedagem, ajuda, apoio e sugestões.*
- *Ao professor Dr. Marinho Gomes de Andrade Filho que abriu as portas para a realização deste trabalho.*
- *Ao professor Dr. Josemar Rodrigues, pela orientação, estímulo e apoio no decorrer deste trabalho.*
- *Ao professor Dr. Jorge Alberto Achcar, pela oportunidade e apoio.*
- *À Maria Silvia Plepis Barros pela revisão deste trabalho.*
- *À Fátima Luiza Prince Antunes, pela amizade e consideração.*
- *Aos colegas que, ao folharem este exemplar, sintam o prazer de terem contribuído de alguma forma a realização deste trabalho.*

Meus Agradecimentos.

Sumário

Lista de Figuras	iii
Lista de Tabelas	iv
Resumo	v
Abstract	vi

Capítulo I

1. Introdução.....	1
1.1 Vantagens do modelo misto proposto	2
1.2 Uma abordagem geral e apresentação de alguns resultados	3

Capítulo II

2. Revisão da Literatura	6
2.1 Análise Bayesiana	9
2.2 Avaliando as distribuições a priori	14
2.3 Exemplos : Pesquisas realizadas por Wind e Lerner (1979) e Parfitt (1967).....	15

Capítulo III

3. Inferência Bayesiana para dados dicotômicos com erros de resposta	23
3.1 Modelo misto aplicado a pesquisa de mercado.....	23
3.2 Modelo misto proposto por Diebolt e Robert (1994).....	24
3.2.1 O algoritmo de Gibbs com amostra ampliada (Diebolt & Robert).....	28

3.3 O Algoritmo de Gibbs para o modelo com erros de respostas	29
3.4 Simulação para x , y e q utilizando o modelo misto proposto	33

Capítulo IV

4. Um Exemplo Ilustrativo - Resultados e Análises	36
4.1 Exemplo Ilustrativo	37
4.1.1 Seleção da amostra e estatísticas a posteriori.....	38
4.1.2 Análise de convergência	44
4.1.3 Verificação da adequabilidade da distribuição Beta para p e q	46
4.2 Comparação entre p e q	48
4.3 Comparação entre o modelo proposto e o modelo de Gaba e Winkler	49

Capítulo V

5. Conclusões	52
5.1 Conclusões	52
5.2 Resumo das Contribuições	53
5.3 Trabalhos Futuros	53
Anexo A – Programa em Matlab.....	54
Anexo B – Programa em Matlab	64
Referências Bibliográficas	68

Lista de Figuras

Figura 1 – Densidades Marginais a Posteriori para p e q	4
Figura 2 – Densidades Marginais a Priori e a Posteriori para p do Exemplo19 Café Instantâneo ($\alpha=11$, $\beta=9$, $\alpha_1=2$, $\beta_1=48$, $\alpha_2=10$, $\beta_2=10$).	
Figura 3 – Densidades Marginais a Posteriori Correspondentes a Seleção 21 das Distribuições a Priori para o Exemplo de Café Instantâneo.	
Figura 4 – Distribuição Marginal a Posteriori para p	41
Figura 5 – Distribuição Marginal a Posteriori para q	41
Figura 6 – Distribuição Marginal a Posteriori para θ_1	42
Figura 7 – Distribuição Marginal a Posteriori para x	42
Figura 8 – Distribuição Marginal a Posteriori para θ_2	43
Figura 9 – Distribuição Marginal a Posteriori para y	43
Figura 10 – Análise de Convergência para p	44
Figura 11 – Análise de Convergência para θ_1 e θ_2	44
Figura 12 – Densidades Marginais a Posteriori para p e q	48
Figura 13 – Distribuição Marginal a Posteriori para p	50

Lista de Tabelas

Tabela 1 – Estimativas para x , y , q , p , θ_1 e θ_2 por Marca ($n=450$).....	16
Tabela 2 – Proporções de Café Instantâneo ($n=940$)	17
Tabela 3 – Resumo das Distribuições a Posteriori para p no Exemplo..... Café Instantâneo sobre Diferentes Distribuições a Priori.	21
Tabela 4 – Distribuição Conjunta de (X_i, Y_i)	24
Tabela 5 – Estatísticas a Posteriori para p , θ_1 , θ_2 , x , y e q	40
Tabela 6 – Critério de Convergência de Gelman/Rubin para p , θ_1 e θ_2	46

Resumo

Inferência Bayesiana para pesquisa de mercado incluindo erros de resposta é estudado como uma mistura de duas distribuições de Bernoulli. Como a análise Bayesiana geralmente implica em cálculos complexos, o método de Monte Carlo com dados ampliados é desenvolvido para obter os resumos marginais a posteriori. Variáveis latentes foram introduzidas para indicar qual componente da mistura gerou a informação com erro de classificação. Também, um procedimento Bayesiano baseado no conceito de “p-value” e na distância de variação total foi introduzida para medir o efeito do erro na distribuição marginal a posteriori. É também realizado, uma comparação entre o modelo misto proposto e o modelo exato introduzido por Gaba e Winkler com o objetivo de verificar a eficiência da técnica aplicada. Uma ilustração com dados simulados é considerada.

Abstract

Bayesian Inference for market research including answer errors is studied as a mixture of two Bernoulli distributions. Since a formal Bayesian analysis leads to intractable calculations, a Markov Chain Monte Carlo method with data augmentation is developed to obtain the summary and marginal posteriors. Were introduced a latent variables that indicates which component of the mixture gives the information with classification errors. Also, a Bayesian procedure based on the Bayes p-value and the total variation distance to measure the effect of the errors on the marginal posterior distribution. It is also, accomplished a comparison between the proposed mixed model and the exact model introduced by Gaba and Winkler with the objective of verifying the efficiency of the applied technique. An illustration with simulated data is considered.

CAPÍTULO I

Introdução

*"Nenhum vento sopra a favor
de quem não sabe aonde ir."
(Sêneca)*

1. Introdução

O propósito deste projeto é formular um procedimento Bayesiano para pesquisa de mercado com erros de resposta utilizando um modelo misto e uma técnica de simulação (Amostrador de Gibbs com variáveis latentes).

Na prática há vários tipos de erros de resposta em pesquisa de mercado nas diversas áreas que envolvem a proporção de consumidores que compram ou não determinado produto ou marca. Por vários motivos, incluindo erros de resposta ou erros de codificação, os dados obtidos muitas vezes não expressam a realidade. Por exemplo, suponhamos que estamos interessados na proporção de consumidores que compram uma certa revista. Dois tipos de erros são possíveis: um indivíduo que não comprou a revista informa que comprou e um indivíduo que comprou a revista informa que não comprou. O problema fundamental é como utilizar este tipo de dado para fazer inferência a respeito da proporção real de consumidores que compram o referido produto ou marca (proporção de interesse).

Na maioria das pesquisas de mercado, nenhuma tentativa é feita para regularizar ou medir estes tipos de erros e, conseqüentemente, as taxas de erros não são conhecidas. Considerar o fato de que o erro adicional reduzirá a precisão do estimador da proporção de interesse, verificando assim sua importância. Com isso, a tomada de decisão sobre os resultados obtidos para as proporções de interesse será mais realística. Com a grande competitividade do mercado, as empresas têm a necessidade de buscar informações com maior confiabilidade sobre os dados de pesquisa, justificando assim a importância do tratamento deste tema.

Existe uma vasta literatura sobre este tema, onde destacamos Clancy, Ostlund e Wyner (1979), Singh e Churchill (1986), Wind e Lerner (1979), e outros. Como a teoria clássica apresenta um sério problema de identificabilidade, Gaba e Winkler (1992)

introduziram um procedimento Bayesiano conjugado, descrito no Capítulo 2, item 2.1. Com a finalidade de reformular esse procedimento Bayesiano, utilizando os modelos mistos propostos por Diebolt e Robert (1994), descritos no Capítulo 3 item 3.2, e colocados de forma didática no livro de Gelman, Carlin, Stern e Rubin (1995), exemplos simulados foram estudados via amostrador de Gibbs, programados via Matlab e comparados com o procedimento de Gaba e Winkler.

1.1 Vantagens da modelo misto proposto

1. Elimina com facilidade a não identificabilidade dos parâmetros.
2. É uma mistura de dois modelos de Bernoulli.
3. Introduzindo uma variável indicadora que especifica a componente do modelo misto da qual a observação dada é proveniente, o modelo misto pode ser visto como um modelo hierárquico.
4. É fácil de implementar o algoritmo de Gibbs proposto por Diebolt e Robert e obter algumas medidas estatísticas para os parâmetros de interesse. Esta facilidade não ocorre com a Inferência Bayesiana exata de Gaba e Winkler.
5. Não exige a necessidade de painel para analisar os parâmetros de interesse como sugerido por Gaba e Winkler.
6. Caracteriza com simplicidade os dois tipos de erros na resposta do consumidor.
7. A distribuição marginal a posteriori de p (proporção de consumidores que realmente compraram o produto) aproxima-se a uma densidade Beta; enquanto que no modelo de Gaba e Winkler ela é desconhecida quanto a sua forma.

1.2 Uma abordagem geral e apresentação de alguns resultados

De acordo com o propósito deste projeto, o modelo considerado é uma mistura de duas distribuições de Bernoulli que envolve dois tipos básicos de erros, como mencionados anteriormente na Introdução. Devido a facilidade de trabalharmos com essa distribuição, pois envolve variáveis dicotômicas, a inferência Bayesiana aplicada ao problema levou-nos a distribuições condicionais conhecidas. Por isso, para inferir sobre os parâmetros de interesse, introduzimos o Amostrador de Gibbs (Gelman, et al. 1995). O objetivo deste procedimento é gerar dados para os parâmetros de interesse por meio de iterações segundo distribuições condicionais a posteriori para obter estatísticas para cada parâmetro tais como: medidas de tendência central (média, mediana e moda), medidas de dispersão (desvio padrão, desvio médio), separatrizes (quartis) e intervalos de credibilidade. Os resultados e análises deste estudo apresentam-se no Capítulo 4, item 4.1.1.

Para garantir a convergência por meio do Amostrador de Gibbs, realizamos duas análises: o método gráfico e a utilização de uma medida chamada Critério de Convergência de Gelman e Rubin que estão descritos no Capítulo 4, item 4.2.

A grande vantagem do Amostrador de Gibbs é a possibilidade de gerar uma amostra de uma densidade, sem necessidade de conhecê-la, que são as distribuições marginais a posteriori. No nosso estudo, estas distribuições foram analisadas para os parâmetros p (proporção de consumidores que **realmente** compraram o produto ou marca) e q (proporção de consumidores que compraram o produto ou marca, **com possível erro**). As Figuras 4 e 5, do Capítulo 4, item 4.1.1 mostram as distribuições marginais a posteriori para p e q e um ajuste da densidade Beta. Utilizando um teste estatístico, sugerido por Gelman et al., descrito no item 4.1.3, comprovamos a adequabilidade desta distribuição para p e q .

Introdução

Uma das vantagens do modelo misto é que a distribuição marginal exata a posteriori para p , de Gaba e Winkler, pode ser aproximada por uma distribuição Beta, como será verificado por meio de estudos descritos no Capítulo 4, item 4.3.

Para analisarmos a qualidade dos dados quanto aos erros de respostas, realizamos uma comparação entre p e q por meio de uma medida calibrada, sugerida por Peng e Dey (1995), descrita no item 4.2. O procedimento para a comparação é o seguinte:

- segundo dados gerados pelo Amostrador de Gibbs para p , gerou-se q através da seguinte expressão: $q = p(1 - \theta_1) + (1 - p)\theta_2$, onde θ_1 e θ_2 são os erros de respostas condicionados, também gerados pelo Amostrador de Gibbs. O resultado deste estudo é tão importante que não poderíamos deixar de apresentar na Introdução a Figura abaixo que mostra claramente a influência dos erros nos dados.

Figura 1 – Densidades Marginais a Posteriori para p e q .

Este gráfico provem de estudos que foram realizados para os parâmetros p e q por meio do modelo misto proposto, utilizando o Amostrador de Gibbs e ajustado segundo a densidade Beta. A apresentação e o procedimento computacional para obtê-lo estão descritos no Capítulo 4, item 4.2.

Realizamos outra comparação importante envolvendo o parâmetro p , entre o modelo proposto e o modelo exato de Gaba e Winkler, com o objetivo de verificar a eficiência da técnica aplicada em estudo, isto é, comprovamos que os dois modelos aproximam-se. O procedimento aplicado está descrito no Capítulo 4, item 4.3, e o programa executado em Matlab encontra-se no Anexo B.

Consideramos também, outro estudo envolvendo os parâmetros de erros de respostas no modelo que são: θ_1 (proporção de **um comprador** ser registrado **erroneamente** como **não comprador** do produto) e θ_2 (proporção de **um não comprador** ser registrado **erroneamente** como **comprador** do produto). Da mesma forma como para p , atribuímos valores para θ_1 e θ_2 , com a restrição $(1-\theta_1) > \theta_2$, para evitar a não identificabilidade dos parâmetros no modelo e por meio do Amostrador de Gibbs, simulamos os dados. Os resultados e análises encontram-se no Capítulo 4, item 4.1.1 representados nas Figuras 6 e 8, assim como algumas estatísticas a posteriori na Tabela 5.

Baseando-se no exemplo apresentado por Gaba e Winkler, Capítulo 2, item 2.3 sobre pesquisa de marcas de margarina, incluímos também em nosso estudo os parâmetros x (proporção de consumidores que **realmente compraram** a marca, entre aqueles que informaram que **não compraram** a marca com **possível erro**) e y (proporção de consumidores que **realmente compraram** a marca, entre aqueles que informaram que **compraram** a marca com **possível erro**). O procedimento matemático e computacional para obtê-los, estão descritos no Capítulo 3, item 3.4. Os resultados e análises encontram-se no Capítulo 4, item 4.1.1 representados nas Figuras 7 e 9, assim como algumas estatísticas a posteriori na Tabela 5.

No Capítulo 5, apresenta-se as conclusões, que consta de três seções. Na primeira seção, apresenta-se as conclusões; na segunda seção, resumo das contribuições e na terceira seção, sugestões para trabalhos futuros. Após, seguem os anexos que trazem os programas em Matlab, e as referências bibliográficas.

CAPÍTULO II

Revisão da Literatura

***"Crie um caminho. O que
passou é referência."
(Benvenutti)***

2. Revisão da Literatura

Segundo Gaba e Winkler (1992), são coletados dados de pesquisa de consumidores para fazer vários tipos de inferência, com o propósito de estudar o comportamento de compra de um produto ou marca. É reconhecido na literatura que tais dados podem estar com erro. Por esse motivo, Gaba e Winkler introduziram um procedimento, que leva em consideração dois tipos de erros que podem ocorrer em dados de pesquisa, como por exemplo: em uma pesquisa sobre anúncio de um certo produto, o indivíduo que não viu o anúncio poderia informar ter visto, e um indivíduo que viu o anúncio poderia informar não ter visto. Portanto, o modelo a ser estudado, contém dois parâmetros de taxa de erro além do parâmetro de interesse sobre a pesquisa. O modelo probabilístico proposto, revela um problema de identificação dos parâmetros, que pode ser evitado quando uma abordagem Bayesiana é introduzida. Para melhor entendermos este problema, Gaba e Winkler consideraram o seguinte:

–seja p representando a proporção de consumidores que compram determinado produto ou marca. Se as informações são registradas com precisão em uma pesquisa, então os dados dicotômicos seguem um processo de Bernoulli com parâmetro p ; porém erros nos dados modificam este processo. Seja:

θ_1 = a probabilidade de um comprador ser registrado erroneamente como não comprador

θ_2 = a probabilidade de um não comprador ser registrado erroneamente como comprador

Utilizando as propriedades de probabilidade, podemos expressar as seguintes probabilidades dos dados sujeitos a erro:

$$q = p(1 - \theta_1) + (1 - p) \theta_2 : \quad (2.1)$$

é a probabilidade do consumidor ser registrado como **comprador**.

$$1 - q = p\theta_1 + (1 - p)(1 - \theta_2) : \quad (2.2)$$

é a probabilidade do consumidor ser registrado como **não comprador**.

O registro de compradores e não compradores seguem um processo de Bernoulli, com parâmetro q ao invés de p .

Suponha que numa amostra aleatória de n consumidores, são registrados r como compradores e $(n - r)$ como não compradores. Então, podemos construir a função de verossimilhança que será expressa como:

$$L(p, \theta_1, \theta_2 | r, n) = [p(1 - \theta_1) + (1 - p)\theta_2]^r [p\theta_1 + (1 - p)(1 - \theta_2)]^{(n-r)}. \quad (2.3)$$

Se as taxas de erros θ_1 e θ_2 são conhecidas, então o estimador de máxima verossimilhança para p baseado em (2.3), é dado pelo seguinte resultado:

Resultado 1 – O estimador de Máxima Verossimilhança para p com θ_1 e θ_2 conhecidos é dado por:

$$\hat{p} = \begin{cases} 1 & \text{se } [(r/n) - \theta_2] / (1 - \theta_1 - \theta_2) > 1 \\ [(r/n) - \theta_2] / (1 - \theta_1 - \theta_2) & \text{se } 0 \leq [(r/n) - \theta_2] / (1 - \theta_1 - \theta_2) \leq 1 \\ 0 & \text{se } [(r/n) - \theta_2] / (1 - \theta_1 - \theta_2) < 0 \end{cases} \quad (2.4)$$

Demonstração : Aplicando logarítmo em $L(p, \theta_1, \theta_2 | r, n)$, temos:

$$\ln L(p, \theta_1, \theta_2 | r, n) = r \ln [p(1 - \theta_1) + (1 - p)\theta_2] + (n - r) \ln [p\theta_1 + (1 - p)(1 - \theta_2)]$$

Derivando em relação a p e igualando a zero, temos:

$$\frac{\partial \ln L(p, \theta_1, \theta_2 | r, n)}{\partial p} = \frac{r[1 - \theta_1 - \theta_2]}{p(1 - \theta_1) + (1 - p)\theta_2} + \frac{(n - r)[\theta_1 + \theta_2 - 1]}{p\theta_1 + (1 - p)(1 - \theta_2)} = 0$$

$$\begin{aligned}\frac{\partial \ln L(p, \theta_1, \theta_2 | r, n)}{\partial p} &= r [1 - \theta_1 - \theta_2] \{ p (1 - \theta_1) + (1 - p) \theta_2 + p \theta_1 + (1 - p) (1 - \theta_2) \} + \\ &\quad [p (1 - \theta_1) + (1 - p) \theta_2] [n (\theta_1 + \theta_2 - 1)] = 0 \\ \Rightarrow (1 - \theta_1 - \theta_2) [r + n p ((\theta_1 + \theta_2 - 1) - n \theta_2)] &= 0 \\ \Rightarrow -p (1 - \theta_1 - \theta_2) + r/n - \theta_2 &= 0\end{aligned}$$

Logo, temos que:

$$\hat{p} = \frac{(r/n) - \theta_2}{1 - \theta_1 - \theta_2}.$$

Geralmente não conhecemos θ_1 e θ_2 e devemos considerar a função de verossimilhança como uma função de três parâmetros (p, θ_1, θ_2) . A estimativa de máxima verossimilhança de (p, θ_1, θ_2) não é única, isto é, diferentes combinações de valores entre p , θ_1 e θ_2 poderá conduzir ao mesmo valor de q e, conseqüentemente, a mesma função de verossimilhança. Aqui surge um problema de identificação dos parâmetros, e como resultado, a função de máxima verossimilhança atinge seu valor de máximo em todos os pontos (p, θ_1, θ_2) , tal que $p(1 - \theta_1) + (1 - p) \theta_2 = r/n$ (estimativa de q).

A incerteza introduzida pelos parâmetros de erro, pode ser vista mais claramente reescrevendo a função de verossimilhança dada em (2.3) em forma de mistura, aplicando Binômio de Newton, e que é dada pelo seguinte resultado:

Resultado 2 – A função de Verossimilhança de (2.3) pode ser dada por:

$$L(p, \theta_1, \theta_2) = \sum_{j=0}^r \sum_{t=0}^{n-r} \binom{r}{j} \binom{n-r}{t} p^{n-j-t} (1-p)^{j+t} \theta_1^{n-r-t} (1-\theta_1)^{r-j} \theta_2^j (1-\theta_2)^t. \quad (2.5)$$

onde

n = tamanho da amostra

j = n.º de **não compradores** registrados **erroneamente** como *compradores*

t = n.º de **não compradores** registrados **corretamente** como *não compradores*

r = n.º de compradores

Demonstração : Utilizando o Binômio de Newton, $(a + b)^n = \sum_{j=0}^n \binom{n}{j} a^{n-j} b^j$, temos que:

$$\begin{aligned} L(p, \theta_1, \theta_2 | r, n) &= [p(1-\theta_1) + (1-p)\theta_2]^r [p\theta_1 + (1-p)(1-\theta_2)]^{(n-r)} \\ &= \sum_{j=0}^r \binom{r}{j} [p(1-\theta_1)]^{r-j} [(1-p)\theta_2]^j \sum_{t=0}^{n-r} \binom{n-r}{t} [p\theta_1]^{n-r-t} [(1-p)(1-\theta_2)]^t. \end{aligned}$$

Logo, temos que:

$$L(p, \theta_1, \theta_2 | r, n) = \sum_{j=0}^r \sum_{t=0}^{n-r} \binom{r}{j} \binom{n-r}{t} p^{n-j-t} (1-p)^{j+t} \theta_1^{n-r-t} (1-\theta_1)^{r-j} \theta_2^j (1-\theta_2)^t.$$

2.1 Análise Bayesiana

Com erros em dados dicotômicos na área de pesquisa de mercado, esperamos obter informações adicionais a priori, evitando assim, problemas de identificação dos parâmetros que surgiram na função de verossimilhança. Portanto, a função de verossimilhança dada em (2.5), sugere uma densidade a priori proporcional a um produto de densidades Beta para os parâmetros p , θ_1 e θ_2 . A distribuição Beta é frequentemente utilizada em modelos Bayesianos para interpretar as informações que envolvem proporções. Uma constante normalizadora será introduzida para garantir que a priori é uma função densidade, isto é, sua integral é igual a um. Além disso, uma função $h(p, \theta_1, \theta_2)$ será incluída com o propósito de interpretar a dependência entre os três parâmetros (p, θ_1, θ_2) . Portanto, a densidade a priori para (p, θ_1, θ_2) é dada por:

$$f(p, \theta_1, \theta_2) = c(\alpha, \beta, \alpha_1, \beta_1, \alpha_2, \beta_2) h(p, \theta_1, \theta_2) f_\beta(p | \alpha, \beta) f_\beta(\theta_1 | \alpha_1, \beta_1) f_\beta(\theta_2 | \alpha_2, \beta_2) \quad (2.1.1)$$

com $\alpha, \beta, \alpha_1, \beta_1, \alpha_2$ e $\beta_2 > 0$, onde $c(\alpha, \beta, \alpha_1, \beta_1, \alpha_2, \beta_2)$ é a constante normalizadora, isto é:

$$c(\alpha, \beta, \alpha_1, \beta_1, \alpha_2, \beta_2) = \left[\int_0^1 \int_0^1 \int_0^1 h(p, \theta_1, \theta_2) f_\beta(p | \alpha, \beta) f_\beta(\theta_1 | \alpha_1, \beta_1) f_\beta(\theta_2 | \alpha_2, \beta_2) \partial p \partial \theta_1 \partial \theta_2 \right]^{-1}, \text{ onde}$$

$$f_{\beta}(a|c, d) = \frac{a^{c-1} (1-a)^{d-1}}{B(c, d)} \quad (2.1.2)$$

é a função densidade Beta com parâmetros c e d , e

$$B(c, d) = \frac{\Gamma(c) \Gamma(d)}{\Gamma(c+d)}.$$

Qualquer julgamento a priori sobre a relação entre p , θ_1 e θ_2 é resumido em h , e suposições de independência poderão simplificar a priori. Por exemplo, se p é independente de θ_1 e θ_2 , h pode ser fatorado em um termo que envolve p e um outro termo que envolve as taxas de erros. Além disso, se $h(p, \theta_1, \theta_2) = 1$, a densidade a priori em (2.1.1), simplifica em um produto de densidades Beta.

Multiplicando a função de verossimilhança, dada em (2.5), pela densidade a priori, dada em (2.1.1), e normalizando, como requer o teorema de Bayes, obtém-se a **densidade conjunta a posteriori exata para (p, θ_1, θ_2)** , dada pelo Resultado 3. Essa densidade conjunta a posteriori pode ser expressa como uma mistura de densidades, da mesma forma como em (2.1.1); mas com novos parâmetros para as densidades Beta.

Resultado 3 – A densidade conjunta a posteriori para (p, θ_1, θ_2) é dada por:

$$f(p, \theta_1, \theta_2 | r, n) = \sum_{j=0}^r \sum_{t=0}^{n-r} w_{jt} f(p, \theta_1, \theta_2 | r, n, j, t), \quad (2.1.3)$$

onde

$$w_{jt} = \frac{a_{jt}}{\sum_{j=0}^r \sum_{t=0}^{n-r} a_{jt}},$$

$$a_{jt} = \frac{\binom{r}{j} \binom{n-r}{t} B(\alpha^*, \beta^*) B(\alpha_1^*, \beta_1^*) B(\alpha_2^*, \beta_2^*)}{c(\alpha^*, \beta^*, \alpha_1^*, \beta_1^*, \alpha_2^*, \beta_2^*)},$$

$$f(p, \theta_1, \theta_2 | r, n, j, t) = c(\alpha^*, \beta^*, \alpha_1^*, \beta_1^*, \alpha_2^*, \beta_2^*) h(p, \theta_1, \theta_2) f_\beta(p | \alpha^*, \beta^*) f_\beta(\theta_1 | \alpha_1^*, \beta_1^*) f_\beta(\theta_2 | \alpha_2^*, \beta_2^*), \quad (2.1.4)$$

$$\begin{aligned} \alpha^* &= \alpha + n - j - t, \\ \beta^* &= \beta + j + t, \\ \alpha_1^* &= \alpha_1 + n - r - t, \\ \beta_1^* &= \beta_1 + r - j, \\ \alpha_2^* &= \alpha_2 + j, \\ \beta_2^* &= \beta_2 + t \end{aligned} \quad e$$

Demonstração: Aplicando o termo de proporcionalidade em $f(p, \theta_1, \theta_2 | r, n)$ e supondo $h(p, \theta_1, \theta_2) = 1$, temos:

$$\begin{aligned} f(p, \theta_1, \theta_2 | r, n) &\propto L(p, \theta_1, \theta_2 | r, n) f(p, \theta_1, \theta_2) \\ &\propto p^{n-j-t} (1-p)^{j+t} \theta_1^{n-r-t} (1-\theta_1)^{r-j} \theta_2^j (1-\theta_2)^t \cdot \\ &\quad p^{\alpha-1} (1-p)^{\beta-1} \theta_1^{\alpha_1-1} (1-\theta_1)^{\beta_1-1} \theta_2^{\alpha_2-1} (1-\theta_2)^{\beta_2-1} \\ &\propto p^{(\alpha+n-j-t)-1} (1-p)^{(j+t+\beta)-1} \theta_1^{(\alpha_1+n-r-t)-1} \cdot \\ &\quad (1-\theta_1)^{(\beta_1+r-j)-1} \theta_2^{(\alpha_2+j)-1} (1-\theta_2)^{(\beta_2+t)-1} \\ &\propto f_\beta(p | \alpha^*, \beta^*) f_\beta(\theta_1 | \alpha_1^*, \beta_1^*) f_\beta(\theta_2 | \alpha_2^*, \beta_2^*). \end{aligned}$$

Logo, de (2.1.4), temos que:

$$f(p, \theta_1, \theta_2 | r, n) \propto f(p, \theta_1, \theta_2 | r, n, j, t).$$

Considerando que o objetivo principal é fazer inferência sobre p , podemos obter a **distribuição marginal a posteriori exata de p** da seguinte forma: Se $h(p, \theta_1, \theta_2)$ é constante em p , isto é, a distribuição a priori de p é independente das taxas de erros, temos o seguinte resultado:

Resultado 4 – A distribuição marginal a posteriori exata para p é dada por:

$$f(p | r, n) = \sum_{j=0}^r \sum_{t=0}^{n-r} w_{jt} f_\beta(p | \alpha^*, \beta^*). \quad (2.1.5)$$

Revisão da Literatura

Se considerarmos $s = j + t$, a expressão em (2.1.5) pode ser expressa da seguinte forma:

$$f(p|r,n) = \sum_{s=0}^n v_s f_{\beta}(p | \alpha + n - s, \beta + s), \quad (2.1.6)$$

onde
$$v_s = \sum_{(j,t) : j+t=s} w_{j,t}.$$

Demonstração: Integrando a distribuição a posteriori conjunta $f(p, \theta_1, \theta_2 | r, n)$ em relação a θ_1 e θ_2 temos:

$$\begin{aligned} f(p | r, n) &= \int_0^1 \int_0^1 f(p, \theta_1, \theta_2 | r, n) d\theta_1 d\theta_2 \\ &\propto \int_0^1 \int_0^1 h(p, \theta_1, \theta_2) f_{\beta}(p | \alpha^*, \beta^*) f_{\beta}(\theta_1 | \alpha_1^*, \beta_1^*) f_{\beta}(\theta_2 | \alpha_2^*, \beta_2^*) d\theta_1 d\theta_2. \end{aligned}$$

Como p é independente de θ_1 e θ_2 , temos:

$$f(p | r, n) \propto f_{\beta}(p | \alpha^*, \beta^*).$$

Logo, temos que:

$$f(p | r, n) \propto f_{\beta}(p | \alpha + n - s, \beta + s).$$

Considerando as propriedades da distribuição Beta, a média e a variância a posteriori de p são expressas pelo seguinte resultado:

Resultado 5 – A média a posteriori e a variância a posteriori para p são dados por:

$$E(p|r,n) = \sum_{s=0}^n v_s \frac{(\alpha + n - s)}{(\alpha + \beta + n)} \quad e \quad (2.1.7)$$

$$V(p|r,n) = E(p^2|r,n) - E^2(p|r,n), \quad \text{onde}$$

$$E(p^2|r,n) = \sum_{s=0}^n v_s \frac{(\alpha + n - s + 1)(\alpha + n - s)}{(\alpha + \beta + n + 1)(\alpha + \beta + n)}. \quad (2.1.8)$$

Revisão da Literatura

Demonstração: Considerando as propriedades de esperança matemática, distribuição Beta e distribuição Gama, temos:

$$\begin{aligned} E(p | r, n) &= \int_0^1 p f(p | r, n) dp \\ &= \int_0^1 p \sum_{s=0}^n v_s f_p(p | \alpha + n - s, \beta + s) dp \\ &= \sum_{s=0}^n v_s \int_0^1 p \frac{p^{(\alpha + n - s) - 1} (1 - p)^{(\beta + s) - 1}}{B(\alpha + n - s, \beta + s)} dp \\ &= \sum_{s=0}^n v_s \frac{1}{B(\alpha + n - s, \beta + s)} \int_0^1 p^{(\alpha + n - s)} (1 - p)^{(\beta + s) - 1} dp \\ &= \sum_{s=0}^n v_s \frac{B(\alpha + n - s + 1, \beta + s)}{B(\alpha + n - s, \beta + s)} \\ &= \sum_{s=0}^n v_s \frac{\Gamma(\alpha + \beta + n) \Gamma(\alpha + n - s + 1)}{\Gamma(\alpha + n - s) \Gamma(\alpha + \beta + n + 1)} \\ &= \sum_{s=0}^n v_s \frac{(\alpha + \beta + n - 1)! (\alpha + n - s)!}{(\alpha + n - s - 1)! (\alpha + \beta + n)!} . \end{aligned}$$

Logo, temos que:

$$E(p | r, n) = \sum_{s=0}^n v_s \frac{(\alpha + n - s)}{(\alpha + \beta + n)} .$$

Utilizando o mesmo procedimento, podemos calcular $E(p^2 | r, n)$ como:

$$\begin{aligned} E(p^2 | r, n) &= \int_0^1 p^2 f(p | r, n) dp \\ &= \int_0^1 p^2 \sum_{s=0}^n v_s f_p(p | \alpha + n - s, \beta + s) dp . \end{aligned}$$

Por outro lado,

$$\begin{aligned} E(p^2 | r, n) &= \sum_{s=0}^n v_s \frac{1}{B(\alpha + n - s, \beta + s)} \int_0^1 p^{(\alpha + n - s + 1)} (1 - p)^{(\beta + s) - 1} dp \\ &= \sum_{s=0}^n v_s \frac{B(\alpha + n - s + 2, \beta + s)}{B(\alpha + n - s, \beta + s)} \\ &= \sum_{s=0}^n v_s \frac{\Gamma(\alpha + \beta + n) \Gamma(\alpha + n - s + 2)}{\Gamma(\alpha + n - s) \Gamma(\alpha + \beta + n + 2)} \\ &= \sum_{s=0}^n v_s \frac{(\alpha + \beta + n - 1)! (\alpha + n - s + 1)!}{(\alpha + n - s - 1)! (\alpha + \beta + n + 1)!}, \text{ concluindo que :} \\ E(p^2 | r, n) &= \sum_{s=0}^n v_s \frac{(\alpha + n - s + 1)(\alpha + n - s)}{(\alpha + \beta + n + 1)(\alpha + \beta + n)}. \end{aligned}$$

Logo, temos que:

$$V(p | r, n) = \sum_{s=0}^n v_s \frac{(\alpha + n - s + 1)(\alpha + n - s)}{(\alpha + \beta + n + 1)(\alpha + \beta + n)} - \left[\sum_{s=0}^n v_s \frac{(\alpha + n - s)}{(\alpha + \beta + n)} \right]^2.$$

Podemos também fazer inferência sobre θ_1 e θ_2 . Suas densidades marginais a posteriori exatas podem ser obtidas integrando a densidade conjunta a posteriori adequadamente.

2.2 Avaliando as distribuições a priori

A distribuição a priori assume um papel importante na análise Bayesiana. Aqui podemos discutir algumas distribuições a priori que são utilizadas em modelos Bayesianos, quando não temos nenhuma informação adicional além dos dados. Uma escolha comum é considerar uma priori difusa ou "não informativa" para (p, θ_1, θ_2) , como por exemplo,

uma densidade localmente uniforme, isto é, $f(p, \theta_1, \theta_2) = 1$. Este é um caso especial de (2.1.1) com $h(p, \theta_1, \theta_2) = 1$ e $\alpha = \beta = \alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1$. Assim, a densidade a posteriori dada em (2.1.3) é proporcional à função de verossimilhança dada em (2.5). O problema é que as densidades a priori não informativas são geralmente impróprias, não garantindo a existência da posteriori.

Na maioria das aplicações, temos mais informações sobre os parâmetros, antes dos dados da pesquisa serem coletados. Neste caso, essas informações são expressas por meio de distribuições de probabilidade sobre os parâmetros de interesse. Por exemplo, se p representa a proporção de consumidores que leram uma determinada revista, podemos considerar uma estimativa de p entre 0,10 a 0,20, mais provável de ocorrer do que valor de p superior a 0,40, e para θ_1 uma taxa de erro mais baixa do que para θ_2 .

A distribuição conjunta a posteriori exata, dada em (2.1.3), pode ser simplificada se os parâmetros p , θ_1 e θ_2 forem independentes, isto é, se $h(p, \theta_1, \theta_2)$ for igual a um em (2.1.1). Caso contrário, devemos avaliar a função h por meio da distribuição condicional para p , dado θ_1 e θ_2 e da distribuição condicional para θ_1 , dado θ_2 .

2.3 Exemplos: (Pesquisas realizadas por Wind e Lerner (1979), e Parfitt (1967)).

Exemplo 1 – Em Wind e Lerner (1979), foram selecionados 450 respondentes de uma sub-amostra nacional de registros de painel, e de dados de pesquisa do mesmo respondente durante os meses de junho a novembro de 1976, sobre 7 marcas de margarina. Para cada marca de margarina registraram:

[^]
 x = proporção de consumidores que **realmente** compraram a marca, entre aqueles que informaram que não compraram a marca com possível erro;

[^]
 y = proporção de consumidores que **realmente** compraram a marca, entre aqueles que informaram que compraram a marca com possível erro, e

$\hat{q}$ = proporção de consumidores que compraram a marca com possível erro.

Utilizando as propriedades de probabilidade condicional, podemos expressar as estimativas de p , θ_1 e θ_2 como:

$$\hat{p} = \hat{x}(1 - \hat{q}) + \hat{y}\hat{q};$$

$$\hat{\theta}_1 = \frac{\hat{x}(1 - \hat{q})}{\hat{x}(1 - \hat{q}) + \hat{y}\hat{q}}; \quad e$$

$$\hat{\theta}_2 = \frac{\hat{q}(1 - \hat{y})}{(1 - \hat{x})(1 - \hat{q}) + (1 - \hat{y})\hat{q}}.$$

As estimativas de x , y e q e de p , θ_1 e θ_2 , estão ilustradas por marca na Tabela 1. Observe que, a estimativa de p (probabilidade dos que informaram que realmente compraram a marca) está abaixo da estimativa de q (probabilidade dos que informaram que compraram a marca **com possível erro**) para todas as marcas.

Tabela 1

Estimativas para x , y , q , p , θ_1 e θ_2 por Marca (n=450)

Marca	$\hat{x}$	$\hat{y}$	$\hat{q}$	$\hat{p}$	$\hat{\theta}_1$	$\hat{\theta}_2$
Bluebonnet	0,20	0,60	0,72	0,49	0,11	0,56
Chiffon	0,14	0,36	0,44	0,24	0,32	0,37
Fleischmann	0,10	0,50	0,54	0,32	0,15	0,39
Imperial	0,14	0,34	0,56	0,25	0,24	0,49
Mazola	0,06	0,27	0,45	0,15	0,21	0,39
Mrs. Filbert's	0,10	0,56	0,38	0,27	0,23	0,23
Parkay	0,39	0,63	0,65	0,55	0,25	0,53

A relação de $\hat{p}$ para $\hat{q}$ varia de 0,33 (para Mazola) e 0,85 (para Parkay). Observe também estimativas mais altas para θ_2 (variando de 0,23 a 0,56) do que para θ_1 (variando de 0,11 a 0,32). Estes resultados indicam que a incidência de erros nos dados é bastante alta e levam em conta particularidades importantes nos estudos de pesquisa de mercado.

Exemplo 2 – Em Parfitt (1967), foi selecionada uma amostra de 940 donas de casa do “Attwood Consumer Panel” na Grã Bretanha e também entrevistou-as, sobre o café instantâneo. As proporções conjuntas e marginais estão ilustradas na Tabela 2.

Tabela 2
Proporções de Café Instantâneo (n=940)

Painel			
	Compradores	Não compradores	Total
Comprou	0,55	0,21	0,76
Não comprou	0,02	0,22	0,24
Total	0,57	0,43	100,00

Baseando-se nestes resultados, foram obtidas as estimativas de $\hat{p} = 0,57$ e de $\hat{q} = 0,76$. Com isto, verificamos a importância de considerar erros nos dados. Através desses resultados, podemos obter também as estimativas das taxas de erros, isto é:

$$\hat{\theta}_1 = \frac{0,02}{0,57} = 0,04 \quad \text{e} \quad \hat{\theta}_2 = \frac{0,21}{0,43} = 0,49.$$

Além desse estudo, Parfitt selecionou outra amostra de 940 donas de casa, não pertencente ao painel, e entrevistou-as obtendo 724 donas de casa que compraram o café instantâneo. O objetivo desse estudo foi utilizar as estimativas de θ_1 e de θ_2 obtidas

pela Tabela 2 e o resultado acima, para estimar o parâmetro p por meio do estimador de máxima verossimilhança, dado em (2.4).

Assim, de (2.3) a função de verossimilhança é:

$$L(p, \theta_1, \theta_2) = [p(1 - \theta_1) + (1 - p)\theta_2]^{724} [p\theta_1 + (1 - p)(1 - \theta_2)]^{216}.$$

Se tratarmos as taxas de erros obtidas por meio dos dados de painel como sendo θ_1 e θ_2 , a estimativa de máxima verossimilhança de p , de (2.4) é dada por:

$$\hat{p} = \frac{[(724/940) - 0,49]}{(1 - 0,04 - 0,49)} = 0,59.$$

Por outro lado, se não assumirmos que θ_1 e θ_2 são conhecidos, o valor de máximo da função de máxima verossimilhança é dado pelos parâmetros (p, θ_1, θ_2) , satisfazendo,

$$p(1 - \theta_1) + (1 - p)\theta_2 = \frac{724}{940} = 0,79.$$

Parfitt considerou outro estudo utilizando o modelo Bayesiano apresentado aqui, para inferir sobre o parâmetro p . Vamos assumir que os parâmetros p , θ_1 e θ_2 são independentes, isto é, $h(p, \theta_1, \theta_2) = 1$, simplificando a densidade a priori dada em (2.1.1) como um produto de densidades Beta.

Baseado nas informações da Tabela 2, vamos considerar os seguintes valores para os parâmetros de cada densidade Beta a priori: $\alpha=11, \beta=9$; $\alpha_1=2, \beta_1=48$; $\alpha_2=10, \beta_2=10$. Assim, a média a priori para p , θ_1 , e θ_2 (0,55, 0,04, e 0,50, respectivamente) estão próximas das estimativas obtidas na Tabela 2, com seus respectivos desvios padrões a priori de 0.11, 0.03, e 0.11.

Revisão da Literatura

A distribuição marginal a posteriori de p que segue sem painel para uma amostra com $r = 724$ e $n = 940$, está ilustrada na Figura 2. É apresentada também a distribuição marginal a priori de p , assim como, a distribuição marginal a posteriori de p livre de erro.

Observe na Figura 2, que a distribuição a posteriori com erro mudou muito comparada com a distribuição a posteriori sem erro, mas não tanto em relação a priori. A distribuição a priori admite valores de p entre 0,25 a 0,85.

O desvio padrão a priori é maior do que o desvio padrão a posteriori que são de 0.11 e 0.08, respectivamente. Observe também que, as distribuições a posteriori com e sem erros no modelo diferem muito em termos de variabilidade. Há também, uma diferença grande na localização, com uma média a posteriori de 0,56 para o modelo com erro e 0,77 para o modelo sem erro. Com isso, ignorando os erros, podemos conduzir grosseiramente a resultados enganosos em termos de localização e variabilidade da distribuição a posteriori.

Figura 2 - Densidades Marginais a Priori e a Posteriori para p do Exemplo Café Instantâneo ($\alpha=11$, $\beta=9$, $\alpha_1=2$, $\beta_1=48$, $\alpha_2=10$, $\beta_2=10$).

Revisão da Literatura

Para investigar a influência da densidade a priori sobre a densidade a posteriori, Parfitt (1967) considerou algumas distribuições a priori para p , θ_1 , e θ_2 . Esta análise é desejável devido a importância da distribuição a priori, e o propósito aqui é fazer uma breve ilustração da sua sensibilidade. As cinco distribuições a priori consideradas são as seguintes:

caso 1 : A distribuição a priori utilizada na análise anterior, é dada pelos seguintes parâmetros: ($\alpha=11$, $\beta=9$, $\alpha_1=2$, $\beta_1=48$, $\alpha_2=10$, $\beta_2=10$);

caso 2 : A distribuição a priori utilizada no caso 1 considerando um aumento do desvio padrão para θ_2 ($\alpha=11$, $\beta=9$, $\alpha_1=2$, $\beta_1=48$, $\alpha_2=2$, $\beta_2=2$);

caso 3 : A distribuição a priori utilizada no caso 1 considerando uma redução do desvio padrão para θ_2 ($\alpha=11$, $\beta=9$, $\alpha_1=2$, $\beta_1=48$, $\alpha_2=50$, $\beta_2=50$);

caso 4 : A distribuição a priori utilizada no caso 1 considerando uma priori difusa (uniforme) para p ($\alpha=1$, $\beta=1$, $\alpha_1=2$, $\beta_1=48$, $\alpha_2=10$, $\beta_2=10$);

caso 5 : A distribuição a priori considerada difusa (uniforme) para os três parâmetros ($\alpha=1$, $\beta=1$, $\alpha_1=1$, $\beta_1=1$, $\alpha_2=1$, $\beta_2=1$).

As distribuições a posteriori correspondentes a seleção das distribuições a priori são ilustradas a seguir, na Figura 3, e algumas medidas são apresentadas na Tabela 3.

Figura 3 - Densidades Marginais a Posteriori correspondentes a Seleção das Distribuições a Priori para o Exemplo de Café Instantâneo.

Tabela 3

Resumo das Distribuições a Posteriori para p no Exemplo Café Instantâneo sobre Diferentes Distribuições a Priori

Priori	α	β	α_1	β_1	α_2	β_2	Média a post.	Desvio padrão a post.	Amostra livre de erro	$\mu_{com} - \mu_{sem}$ erro erro	$\frac{\sigma_{com}}{\sigma_{sem}}$ erro erro
Caso 1	11	9	2	48	10	10	0,56	0,08	16	-0,20	6,0
Caso 2	11	9	2	48	2	2	0,57	0,10	3	-0,20	7,3
Caso 3	11	9	2	48	50	50	0,57	0,06	53	-0,19	4,2
Caso 4	1	1	2	48	10	10	0,54	0,14	9	-0,23	10,3
Caso 5	1	1	1	1	1	1	0,50	0,29	0	-0,27	21,4

Observe na Tabela 3 e Figura 3, que a média a posteriori de p , considerando as quatro primeiras distribuições a priori, não diferem muito, exceto para a distribuição a priori considerada difusa para todos os parâmetros (caso 5). O desvio padrão a posteriori é mais sensível, porém, aumentando quando p é uma priori difusa (como seria esperado) e diminuindo quando o desvio padrão a priori de θ_2 é reduzido. Observe que no caso da priori com desvio padrão reduzido de 0,05 para θ_2 , o tamanho da amostra livre de erro para p é igual a 53, isto é, quase três vezes maior do que o tamanho da amostra do caso 1, mas equivalente a 6% do tamanho da amostra total.

Além da análise apresentada para p , podemos também considerar distribuições marginais a posteriori para os parâmetros de erro. Em nosso exemplo, as distribuições a posteriori estão muito próximas das distribuições a priori correspondentes. Para o caso 1, a diferença entre a média a posteriori e a priori é zero para θ_1 e 0,01 para θ_2 , e a relação do desvio padrão a posteriori e a priori é 0,96 para θ_1 e 0,76 para θ_2 , respectivamente.

Nossas análises indicam a importância da informação a priori, pois verificamos a influência da priori difusa (caso 5) sobre a distribuição a posteriori, proporcionando grande variabilidade e a ocorrência de picos, isto é, duas modas.

CAPÍTULO III

Inferência Bayesiana para Dados Dicotômicos com Erros de Respostas

*"Se você pode sonhá-lo,
você pode fazê-lo."
(Walt Disney)*

3. Inferência Bayesiana para dados dicotômicos com erros de respostas

Neste capítulo será proposto um modelo para dados dicotômicos, considerando os dois tipos de erros de respostas ou de registros θ_1 e θ_2 definidos no Capítulo 2; apresentação do Algoritmo de Gibbs para o modelo com erros de respostas e o procedimento para simular os parâmetros x , y e q descritos por Gaba e Winkler no Capítulo 2, item 2.3.

3.1 Modelo misto aplicado a pesquisa de mercado

Suponhamos que n consumidores de um determinado produto sejam entrevistados com o objetivo de verificar se compraram ou não a marca A. Para cada consumidor definamos a seguinte variável, X_i , sujeita a erros de registro:

$$\begin{cases} X_i = 1 & \Rightarrow \text{comprou a marca A} \\ X_i = 0 & \Rightarrow \text{não comprou a marca A} \end{cases}$$

Como a variável X_i está sujeita a erros de registro, definamos, Y_i , como sendo o valor real (sem erro) de X_i , e os seguintes parâmetros:

$$p = P[Y_i = 1] :$$

é a probabilidade do consumidor **realmente** comprar a marca A.

$$\theta_1 = P[X_i = 0 \mid Y_i = 1] :$$

é a probabilidade de um *comprador* ser registrado **erroneamente** como *não comprador* da marca A.

$$\theta_2 = P[X_i = 1 \mid Y_i = 0] :$$

é a probabilidade de um *não comprador* ser registrado **erroneamente** como *comprador* da marca A.

Assim, podemos construir a distribuição conjunta dos pares (X_i, Y_i) em relação aos parâmetros (p, θ_1, θ_2) , dada na tabela abaixo:

Tabela 4
Distribuição conjunta de (X_i, Y_i)

	$X_i = 0$	$X_i = 1$
$Y_i = 0$	$(1 - p)(1 - \theta_2)$	$(1 - p)\theta_2$
$Y_i = 1$	$p\theta_1$	$p(1 - \theta_1)$

A tabela acima sugere uma mistura de distribuições de Bernoulli para a variável X_i , definida da seguinte forma:

$$f(x \mid p, \theta_1, \theta_2) = p \text{ Bern}(x \mid 1 - \theta_1) + (1 - p) \text{ Bern}(x \mid \theta_2). \quad (3.1.1)$$

Para evitar a não identificação dos parâmetros, é necessário restringirmos o modelo para $(1 - \theta_1) > \theta_2$; assim a estimativa de máxima verossimilhança de (p, θ_1, θ_2) é única. Esta restrição é realística, pois esperamos que a probabilidade de *um comprador* ser registrado corretamente como *comprador* da marca, seja maior, do que a probabilidade de um *não comprador* ser registrado erroneamente como *comprador* da marca.

3.2 Modelo misto proposto por Diebolt e Robert (1994)

Consideremos o modelo misto da distribuição de n variáveis aleatórias independentes $X_1, X_2, \dots, X_n$, como uma mistura de k - componentes, dado por:

$$P(x_i | \theta, p) = \sum_{j=1}^k p_j f(x_i | \theta_j), \quad i = 1, \dots, n. \quad (3.2.1)$$

onde a forma das densidades $f(x_i | \theta_j)$, ($1 \leq j \leq k$) são conhecidas, o vetor $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ é desconhecido e as proporções $0 < p_j < 1$, satisfazendo $\sum_{j=1}^k p_j = 1$, também são desconhecidas.

Segundo Dempster et al. (1977), o modelo misto (3.2.1), pode ser expresso em termos de dados faltantes (ou incompletos). Isto é, vamos introduzir uma variável Z_i como sendo um vetor k - dimensional que indica a componente do modelo misto da qual a observação x_i é proveniente, tal que $Z_{ij} \in \{0, 1\}$ e $\sum_{j=1}^k Z_{ij} = 1$.

Assim, a **distribuição conjunta dos dados** (X_i, Z_i) é dada por:

$$P(x, z | \theta, p) = \prod_{i=1}^n \prod_{j=1}^k p_j^{z_{ij}} f(x_i | \theta_j). \quad (3.2.2)$$

Justificativa: Seja $L(\theta, p)$ a função de verossimilhança dada por:

$$L(\theta, p) = \prod_{i=1}^n P(X_i | \theta, p).$$

Seja a matriz Z ($n \times k$), de variáveis latentes Z_{ij} , $i=1, \dots, n$ e $j=1, \dots, k$ com a seguinte função densidade conjunta:

$$f(Z) \propto \prod_{i=1}^n \prod_{j=1}^k p_{ij}^{z_{ij}}.$$

Como as variáveis dicotômicas X e Z são independentes, temos que:

$$L(\theta, p) = \prod_{i=1}^n P(X_i / \theta, p) \prod_{i=1}^n \prod_{j=1}^k p_{ij}^{z_{ij}}.$$

Segundo o modelo em (3.2.1), e substituindo p_{ij} pela expressão que será dada em (3.2.3), temos:

$$\begin{aligned} L(\theta, p) &= \prod_{i=1}^n \sum_{j=1}^k p_{ij} f(x_i | \theta_j) \prod_{i=1}^n \prod_{j=1}^k \frac{p_{ij}^{z_{ij}} f^{z_{ij}}(X_i / \theta_j)}{[\sum_{l=1}^k p_{il} f(X_i / \theta_l)]^{z_{ij}}} \\ &= \prod_{i=1}^n \sum_{j=1}^k p_{ij} f(x_i | \theta_j) \frac{\prod_{i=1}^n \prod_{j=1}^k p_{ij}^{z_{ij}} f^{z_{ij}}(X_i / \theta_j)}{\prod_{i=1}^n \prod_{j=1}^k [\sum_{l=1}^k p_{il} f(X_i / \theta_l)]^{z_{ij}}} \\ &= \prod_{i=1}^n \sum_{j=1}^k p_{ij} f(x_i | \theta_j) \frac{\prod_{i=1}^n \prod_{j=1}^k p_{ij}^{z_{ij}} f^{z_{ij}}(X_i / \theta_j)}{\prod_{i=1}^n [\sum_{l=1}^k p_{il} f(X_i / \theta_l)]^{\sum_{j=1}^k z_{ij}}}. \end{aligned}$$

Como $\sum_{j=1}^k Z_{ij} = 1$, temos que:

$$L(\theta, p) = P(x, z | \theta, p) = \prod_{i=1}^n \prod_{j=1}^k p_{ij}^{z_{ij}} f^{z_{ij}}(x_i | \theta_j).$$

Desta perspectiva, o modelo misto é hierárquico, pois o interesse agora é verificar a que distribuição de probabilidade a variável X_i é proveniente, por meio de Z_i .

Inferência Bayesiana para dados dicotômicos com erros de respostas

Por exemplo, suponhamos que em uma pesquisa sobre consumo de cerveja por adultos do sexo feminino, estamos interessados em identificar a que distribuição de probabilidade a variável X_i é proveniente, na mistura entre as seguintes variáveis:

- 1 – Uma **consumidora** ser registrada **corretamente** como **consumidora** de cerveja e;
- 2 – Uma **não consumidora** ser registrada **erroneamente** como **consumidora** de cerveja.

Uma variável indicadora Z_i será introduzida, completando a observação x_i , tal que $Z_i = 1$ se a consumidora provém da distribuição (1) e $Z_i = 0$, se a consumidora provém da distribuição (2). Assim, podemos identificar a proveniência de x_i .

Uma maneira de estimar Z_i é por meio do método iterativo conhecido como algoritmo EM (ver Celeux & Diebolt), isto é, vamos substituir cada Z_{ij} pela sua média $p_{ij}^{(m)}$, dada por :

$$p_{ij}^{(m)} = \frac{p_j^{(m)} f^{(m)}(x_i | \theta_j)}{\sum_{j=1}^k p_j^{(m)} f^{(m)}(x_i | \theta_j)} \quad (3.2.3)$$

onde m indexa a iteração atual e $p_j^{(m)}$ e $f_j^{(m)}$ são as avaliações dos parâmetros.

Novas estimativas de máxima verossimilhança são então derivadas dos dados pseudo completados $(x_i, p_i^{(m)})$.

Para evitar alguns defeitos do método EM, Celeux e Diebolt (1985/1992) introduziram uma versão estocástica chamada SEM. Ao invés de estimar os dados completados, eles simularam os Z_i 's da distribuição condicional, isto é, distribuições

multinomiais com pesos, dado em (3.2.3). O procedimento Bayesiano utilizado para simular os Z_i 's será o Amostrador de Gibbs, descrito no item 3.2.1.

3.2.1 O Algoritmo de Gibbs com amostra ampliada (Diebolt & Robert)

Generalizando o Amostrador de Gibbs, vamos descrever uma aproximação para as **distribuições marginais a posteriori**, começando com um valor inicial para os subvetores $(\theta_1^{(0)}, \dots, \theta_s^{(0)})$, isto é:

Iniciando com valores $(\theta_1^{(0)}, \dots, \theta_s^{(0)})$, o algoritmo simula da seguinte forma com passo m ,

(a) gerar $z_i^{(m)} \sim \text{Multinomial}(p_i^{(m)})$ $i = 1, \dots, n$
 $p_i^{(m)} = (p_{i1}^{(m)}, \dots, p_{ik}^{(m)})$

onde

$$p_{ij}^{(m)} = \frac{p_j^{(m)} f^{(m)}(x_i | \theta_j)}{\sum_{j=1}^k p_j^{(m)} f^{(m)}(x_i | \theta_j)}$$

(1) na etapa $m + 1$ gerar:

$$\theta_1^{(m+1)} \sim \pi(\theta_1 \setminus x_i, z_i^{(m)}, \theta_2^{(m)}, \dots, \theta_s^{(m)})$$

.
.
.
.

$$\theta_s^{(m+1)} \sim \pi(\theta_s \setminus x_i, z_i^{(m)}, \theta_1^{(m+1)}, \dots, \theta_{s-1}^{(m+1)})$$

O vetor $(\theta_1^{(m+1)}, \dots, \theta_s^{(m+1)})$ é gerado pelas suas respectivas distribuições condicionais π . A questão principal a ser analisada no algoritmo de Gibbs é a convergência para as distribuições marginais de cada parâmetro.

3.3 O Algoritmo de Gibbs para o modelo com erros de respostas

Nesta seção será proposto um algoritmo para estimar os parâmetros do modelo misto introduzido em (3.1). Este algoritmo será totalmente baseado no algoritmo computacional proposto por Diebolt e Robert (1994), descrito no item 3.2.

Seja n variáveis aleatórias binárias independentes, $X_i, i=1,2,\dots,n$ com o seguinte modelo misto:

$$f(x | p, \theta_1, \theta_2) = p \text{ Bern}(x | 1 - \theta_1) + (1 - p) \text{ Bern}(x | \theta_2), \quad (3.3.1)$$

onde $(1 - \theta_1) > \theta_2$.

Como no modelo misto definido em (3.3.1) não informa qual componente do modelo gerou a observação x , vamos introduzir uma variável Z que indicará a proveniência da observação x , isto é:

$$Z = \begin{cases} 1 & \text{se a observação } x \text{ provem da distribuição Bernoulli com parâmetro } (1 - \theta_1) \\ 0 & \text{se a observação } x \text{ provem da distribuição Bernoulli com parâmetro } \theta_2 \end{cases}$$

Assim, podemos construir a **distribuição conjunta dos dados completados** $(X_i, Z_i), i = 1, \dots, n$ que será expressa por:

$$f(x, z | p, \theta_1, \theta_2) = \prod_{i=1}^n p_i^{z_i} (1 - p_i)^{(1-z_i)} f(x | p, \theta_1, \theta_2). \quad (3.3.2)$$

Substituindo p_i pela expressão dada em (3.2.3), temos o seguinte resultado:

Resultado 6 – A distribuição conjunta para (x_i, z_i) é dada por:

$$f(x, z | p, \theta_1, \theta_2) = \prod_{i=1}^n [p f(x_i | \theta = 1 - \theta_1)]^{z_i} [(1 - p) f(x_i | \theta = \theta_2)]^{(1-z_i)} \quad (3.3.3)$$

Inferência Bayesiana para dados dicotômicos com erros de respostas

Demonstração: Seja $L(p, \theta_1, \theta_2)$ a função de verossimilhança dada por:

$$L(p, \theta_1, \theta_2) = \prod_{i=1}^n f(x_i / p, \theta_1, \theta_2).$$

Seja o vetor latente Z , com a seguinte função densidade conjunta:

$$f(Z) \propto \prod_{i=1}^n p_i^{z_i} (1 - p_i)^{(1-z_i)}.$$

Como as variáveis dicotômicas X e Z são independentes, temos que:

$$L(\theta, p) = \prod_{i=1}^n f(x_i / p, \theta_1, \theta_2) \prod_{i=1}^n p_i^{z_i} (1 - p_i)^{(1-z_i)}.$$

Segundo o modelo em (3.3.1), e substituindo p_i pela expressão dada em (3.2.3), temos:

$$L(p, \theta_1, \theta_2) = \prod_{i=1}^n [p f(X_i | 1 - \theta_1) + (1 - p) f(X_i | \theta_2)] \left[\frac{p f(X_i | 1 - \theta_1)}{p f(X_i | 1 - \theta_1) + (1 - p) f(X_i | \theta_2)} \right]^{z_i} \cdot \left[\frac{(1 - p) f(X_i | \theta_2)}{p f(X_i | 1 - \theta_1) + (1 - p) f(X_i | \theta_2)} \right]^{(1-z_i)}.$$

Logo, temos que:

$$L(p, \theta_1, \theta_2) = f(x, z | p, \theta_1, \theta_2) = \prod_{i=1}^n [p f(x_i | \theta = 1 - \theta_1)]^{z_i} [(1 - p) f(x_i | \theta = \theta_2)]^{(1-z_i)}.$$

Inferência Bayesiana para dados dicotômicos com erros de respostas

Desenvolvendo a expressão dada em (3.3.3), obtemos a função de verossimilhança expressa pelo seguinte resultado:

Resultado 7 – A função de Verossimilhança baseada nos dados ampliados é dada por:

$$L(p, \theta_1, \theta_2 \mid x, z) = p^{\sum_{i=1}^n z_i} (1-p)^{n - \sum_{i=1}^n z_i} \theta_1^{\sum_{i=1}^n (1-x_i) z_i} (1-\theta_1)^{\sum_{i=1}^n x_i z_i} \theta_2^{\sum_{i=1}^n x_i (1-z_i)} (1-\theta_2)^{\sum_{i=1}^n (1-x_i) (1-z_i)} \quad (3.3.4)$$

Demonstração: Utilizando a expressão em (3.3.3), com $f(x_i \mid \theta) = \theta^{x_i} (1-\theta)^{(1-x_i)}$ temos:

$$L(p, \theta_1, \theta_2 \mid x, z) = \prod_{i=1}^n p^{z_i} (1-p)^{(1-z_i)} \theta_1^{x_i z_i} (1-\theta_1)^{(1-x_i) z_i} \theta_2^{x_i (1-z_i)} (1-\theta_2)^{(1-x_i) (1-z_i)}$$

Logo, temos que:

$$L(p, \theta_1, \theta_2 \mid x, z) = p^{\sum_{i=1}^n z_i} (1-p)^{n - \sum_{i=1}^n z_i} \theta_1^{\sum_{i=1}^n (1-x_i) z_i} (1-\theta_1)^{\sum_{i=1}^n x_i z_i} \theta_2^{\sum_{i=1}^n x_i (1-z_i)} (1-\theta_2)^{\sum_{i=1}^n (1-x_i) (1-z_i)}$$

A função de verossimilhança dada em (3.3.4), sugere as seguintes densidades a priori para os parâmetros p , θ_1 e θ_2 :

$$\begin{aligned} p &\sim \text{Beta}[a, b], \\ \theta_i &\sim \text{Beta}[a_i, b_i] \quad i = 1, 2, \quad \text{onde } a, b, a_i \text{ e } b_i \text{ são conhecidos.} \end{aligned} \quad (3.3.5)$$

Multiplicando a função de verossimilhança dada em (3.3.4) pelas densidades a priori dadas em (3.3.5) e normalizando como requer o teorema de Bayes, obtém-se a **densidade conjunta a posteriori para (p, θ_1, θ_2)** , dada pelo seguinte resultado:

Resultado 8 – A densidade conjunta a posteriori para (p, θ_1, θ_2) baseada nos dados ampliados é dado por:

$$\pi(\theta_1, \theta_2, p \mid x, z) \propto p^{a + \sum_{i=1}^n Z_i} (1-p)^{b + n - \sum_{i=1}^n Z_i} \theta_1^{a_1 + \sum_{i=1}^n (1-X_i) Z_i} (1-\theta_1)^{b_1 + \sum_{i=1}^n X_i Z_i} \theta_2^{a_2 + \sum_{i=1}^n X_i (1-Z_i)} (1-\theta_2)^{b_2 + \sum_{i=1}^n (1-X_i)(1-Z_i)} \quad (3.3.6)$$

Portanto de (3.3.6), podemos obter as seguintes **distribuições condicionais** :

$$p \mid \{x_i\}, \{z_i\} \sim \text{Beta} [a + \sum_{i=1}^n z_i ; n + b - \sum_{i=1}^n z_i]$$

$$\theta_1 \mid \{x_i\}, \{z_i\} \sim \text{Beta} [a_1 + \sum_{i=1}^n (1-x_i) z_i ; b_1 + \sum_{i=1}^n x_i z_i]$$

$$\theta_2 \mid \{x_i\}, \{z_i\} \sim \text{Beta} [a_2 + \sum_{i=1}^n x_i (1-z_i) ; b_2 + \sum_{i=1}^n (1-x_i)(1-z_i)]$$

$$Z_i \mid \theta_i, p \sim \text{Bernoulli} (p_i) \quad i = 1, \dots, n$$

onde

$$p_i = \frac{p f(x_i \mid \theta = 1 - \theta_1)}{p f(x_i \mid \theta = 1 - \theta_1) + (1-p) f(x_i \mid \theta = \theta_2)}$$

Baseado na seção 3.2.1, podemos formular o seguinte algoritmo para avaliar os parâmetro p , θ_1 e θ_2 .

Inferência Bayesiana para dados dicotômicos com erros de respostas

Iniciando com valores $(p^{(0)}, \theta_1^{(0)}, \theta_2^{(0)})$, o algoritmo simula com passo m , da seguinte forma:

(a) gerar $z_i^{(m)} \sim \text{Bernoulli}(p_i^{(m)})$, $i = 1, \dots, n$

$$p_i^{(m)} = \frac{p^{(m)} f(x_i | \theta = 1 - \theta_1^{(m)})}{p^{(m)} f(x_i | \theta = 1 - \theta_1^{(m)}) + (1 - p^{(m)}) f(x_i | \theta = \theta_2^{(m)})}$$

(b) na etapa $m + 1$ gerar:

$$\theta_1^{(m+1)} | \{x_i\}, \{z_i^{(m)}\} \sim \text{Beta} \left[a_1 + \sum_{i=1}^n (1 - X_i) Z_i^{(m)}; b_1 + \sum_{i=1}^n X_i Z_i^{(m)} \right],$$

$$\theta_2^{(m+1)} | \{x_i\}, \{z_i^{(m)}\} \sim \text{Beta} \left[a_2 + \sum_{i=1}^n X_i (1 - Z_i^{(m)}); b_2 + \sum_{i=1}^n (1 - X_i) (1 - Z_i^{(m)}) \right],$$

$$p^{(m+1)} | \{x_i\}, \{z_i^{(m)}\} \sim \text{Beta} \left[a + \sum_{i=1}^n Z_i^{(m)}; n + b - \sum_{i=1}^n Z_i^{(m)} \right].$$

3.4 Simulação para x , y e q utilizando o modelo misto proposto

De acordo com dados simulados obtidos para os parâmetros p , θ_1 e θ_2 via Amostrador de Gibbs, simulamos x , y e q , pelo desenvolvimento do seguinte resultado apresentado por Gaba e Winkler, Capítulo 2, item 2.3:

Resultado 9 – As expressões de p , θ_1 e θ_2 para simular x , y e q são dadas por:

$$p = x(1 - q) + yq;$$

Inferência Bayesiana para dados dicotômicos com erros de respostas

$$\theta_1 = \frac{x(1-q)}{x(1-q) + yq} ; \quad e$$

$$\theta_2 = \frac{q(1-y)}{(1-x)(1-q) + (1-y)q}.$$

Demonstração: De acordo com as definições em (3.1), dadas para as variáveis X , Y e para p , θ_1 e θ_2 , vamos considerar as seguintes probabilidades para x , y e q :

$$x = P(Y_i = 1 | X_i = 0) \Rightarrow P(Y_i = 0 | X_i = 0) = 1 - x$$

$$y = P(Y_i = 1 | X_i = 1) \Rightarrow P(Y_i = 0 | X_i = 1) = 1 - y$$

$$q = P(X_i = 1) \Rightarrow P(X_i = 0) = 1 - q$$

Podemos expressar p , θ_1 e θ_2 segundo a regra da probabilidade total da seguinte forma:

$$(1) - p = P(Y_i = 1) = P(Y_i = 1 | X_i = 0) P(X_i = 0) + P(Y_i = 1 | X_i = 1) P(X_i = 1),$$

portanto,

$$p = x(1-q) + yq.$$

$$(2) - \theta_1 = P(X_i = 0 | Y_i = 1) = \frac{P(X_i = 0, Y_i = 1)}{P(Y_i = 1)} = \frac{P(Y_i = 1 | X_i = 0) P(X_i = 0)}{P(Y_i = 1)},$$

Portanto, temos que:

$$\theta_1 = \frac{x(1-q)}{x(1-q) + yq}.$$

Inferência Bayesiana para dados dicotômicos com erros de respostas

$$(3) - \theta_2 = P(X_i = 1 \mid Y_i = 0) = \frac{P(X_i = 1, Y_i = 0)}{P(Y_i = 0)} = \frac{P(Y_i = 0 \mid X_i = 1) P(X_i = 1)}{P(Y_i = 0 \mid X_i = 0)P(X_i = 0) + P(Y_i = 0 \mid X_i = 1)P(X_i = 1)}.$$

Portanto, temos que:

$$\theta_2 = \frac{q(1-y)}{(1-x)(1-q) + (1-y)q}.$$

Como o objetivo é simular x , y e q por meio de p , θ_1 e θ_2 , as expressões acima em função de p , θ_1 e θ_2 são dadas por:

$$q = p(1 - \theta_1) + (1 - p)\theta_2;$$

$$x = \frac{p\theta_1}{1 - [p(1 - \theta_1) + (1 - p)\theta_2]}; \quad e$$

$$y = \frac{p(1 - \theta_1)}{p(1 - \theta_1) + (1 - p)\theta_2}.$$

A mesma análise que foi feita para p , θ_1 e θ_2 também será feita para x , y e q . Os resultados apresentam-se no Capítulo 4, item 4.1.1.

CAPÍTULO IV

Um Exemplo Ilustrativo – Resultados e Análises

*“Alguns vêem coisas que existem e
perguntam por quê? Eu vejo as coisas
que nunca existiram e pergunto por quê não?
(Robert Kennedy)”*

4. Um Exemplo Ilustrativo – Resultados e Análises

Neste capítulo será apresentado uma aplicação sobre o modelo proposto na área de pesquisa de mercado; análise de convergência pelo método gráfico para p , θ_1 e θ_2 e por meio do Critério de Gelman e Rubin utilizando o Amostrador de Gibbs; estudos sobre as distribuições marginais a posteriori para p e q e uma comparação entre os modelos de Gaba e Winkler e de misturas de componentes para p .

Após a aplicação do algoritmo de Gibbs, descrito no Capítulo 3, item 3.3, selecionamos uma amostra final de tamanho 300 para os parâmetros p , θ_1 , θ_2 , x , y e q . Os resultados e análises das estatísticas a posteriori apresentam-se na Tabela 5 e os dados estão ilustrados nas Figuras de 4 a 9.

A seguir apresentamos o estudo de convergência para p , θ_1 e θ_2 pelo método gráfico, ilustrados nas Figuras 10 e 11 e por uma medida (Critério de Convergência de Gelman e Rubin) apresentada na Tabela 6.

Além desta análise apresentada para p , houve interesse em avaliar a forma das distribuições marginais a posteriori para p e q , na aplicação de um teste estatístico apresentado e analisado no item 4.1.3. Após a identificação da forma destas distribuições marginais, comparamos p e q por meio de uma medida estatística (Distância da Variação Total), com o objetivo de analisar a influência dos erros nos dados. O critério está descrito no item 4.2, e os dados estão ilustrados na Figura 12. O procedimento computacional para obter a Figura 12 é o mesmo apresentado no item 4.1.3.

E finalmente no item 4.3, realizamos o último estudo na comparação entre as **distribuições marginais a posteriori para p** , considerando o modelo proposto e o modelo exato de Gaba e Winkler. A análise apresenta-se na Figura 13, assim como o procedimento computacional considerado para a comparação.

4.1 Exemplo Ilustrativo

De acordo com o modelo misto proposto em (3.3.1), simulamos uma amostra de tamanho $n=10$ para as variáveis X_1 e X_2 e para o parâmetro p , por meio das seguintes distribuições de probabilidade:

$$X_1 \sim \text{Bern}(x_1 | 1 - \theta_1)$$

$$X_2 \sim \text{Bern}(x_2 | \theta_2)$$

$$p \sim \text{Bern}(p), \text{ onde } \theta_1=0,04, \theta_2=0,60 \text{ e } p=0,25.$$

Com os valores simulados para X_1 , X_2 e p , simulou-se X por meio do seguinte modelo misto:

$$f(X | p, \theta_1, \theta_2) = p \text{ Bern}(X_1 | 1 - \theta_2) + (1 - p) \text{ Bern}(X_2 | \theta_2);$$

obtendo-se os seguintes valores: $X = [1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 1]$.

De acordo com os valores atribuídos para os parâmetros p , θ_1 e θ_2 acima, considerou-se as seguintes priori:

$$p \sim \text{Beta}[10, 30]$$

$$\theta_1 \sim \text{Beta}[2, 48]$$

$$\theta_2 \sim \text{Beta}[30, 20]$$

Os valores atribuídos para p , θ_1 e θ_2 de 0,25, 0,04 e 0,60, respectivamente, com $(1 - \theta_1) > \theta_2$ foram arbitrários, próximos de suas médias a priori considerando as densidades Beta acima.

Um Exemplo Ilustrativo - Resultados e Análises

Assim, por meio de $m = 5.000$ (iterações) gerou-se $p_i^{(m)}$ e $z_i^{(m)}$; $i = 1, \dots, 10$ segundo o algoritmo de Gibbs descrito anteriormente. Na etapa $(m+1)$ gerou-se $p^{(m+1)}$, $\theta_1^{(m+1)}$ e $\theta_2^{(m+1)}$ segundo distribuições condicionais a posteriori Beta com novos parâmetros.

Obtidos os valores simulados para p , θ_1 e θ_2 , simulou-se x , y e q por meio das expressões descritas no item 3.4.

4.1.1 Seleção da amostra e estatísticas a posteriori

Para garantir a independência entre os valores simulados para cada parâmetro, de $m = 5.000$ (iterações), considerou 60%, isto é, desprezou-se 2.000 valores e desses um salto de $K = 10$, selecionando assim, uma amostra final de 300 valores para p , θ_1 , θ_2 , x , y e q . Os resultados obtidos por esta amostra, apresentam-se na Tabela 5 abaixo, com as seguintes estatísticas a posteriori:

- Intervalo de credibilidade à 95%, 1º quartil (25%), mediana, 3º quartil (75%), média e desvio padrão para cada parâmetro acima.

Observe que, na Tabela 5, ocorre um deslocamento significativo entre p (proporção dos que realmente compraram a marca) e q (proporção dos que compraram a marca com possível erro), admitindo valores entre 0,1210 a 0,3879 para p e 0,5618 a 0,7747 para q ao nível de 95% de credibilidade.

A média a posteriori de q (0,6725) é aproximadamente o triplo de p (0,2441), mas não diferem muito em termos de variabilidade, sendo de 0,06981 para p e de 0,05558 para q . Estes resultados mostram a importância de considerar o erro de resposta nas pesquisas.

Um Exemplo Ilustrativo - Resultados e Análises

Observe também que, θ_1 (proporção de compradores serem registrados erroneamente como não compradores da marca) é maior do que x (proporção de consumidores que realmente compraram a marca, entre aqueles que são registrados como não compradores da marca), obtendo intervalo de credibilidade à 95% de 0,0055 a 0,1060 para θ_1 e 0,0034 a 0,1027 para x sendo a média a posteriori de θ_1 (0,0430) e de x (0,0331).

Para o parâmetro θ_2 (proporção de não compradores serem registrados erroneamente como compradores da marca), observam-se proporções altas com um intervalo de credibilidade ao nível de 95% de 0,4463 a 0,7096, uma média a posteriori de 0,5801 e mediana de 0,5790.

Os resultados de y (proporção de consumidores que realmente compraram a marca, entre aqueles que são registrados como compradores da marca), variam de 0,1740 a 0,5446 ao nível de 95% de credibilidade com uma média a posteriori de 0,3463 e desvio padrão de 0,0912.

As Figuras 4 e 5, apresentam os histogramas para p e q e um ajuste da densidade Beta.

Tabela 5
Estatísticas a posteriori para p , θ_1 , θ_2 , x , y e q

Estatísticas a posteriori							
	2,5%	25%	Mediana	75%	97,5%	Média	Desvio padrão
p	0,1210	0,1943	0,2383	0,3879	0,3879	0,2441	0,06981
θ_1	0,0055	0,0241	0,0387	0,0542	0,1060	0,0430	0,02556
θ_2	0,4463	0,5357	0,5790	0,6264	0,7096	0,5801	0,06688
x	0,0034	0,0169	0,0284	0,0436	0,1027	0,0331	0,02346
y	0,1740	0,2816	0,3397	0,4053	0,5446	0,3463	0,09122
q	0,5618	0,6362	0,6731	0,7148	0,7747	0,6725	0,05558

Um Exemplo Ilustrativo - Resultados e Análises

Observe na Figura 4, que o valor atribuído para p de 0,25 está em torno da média a posteriori obtida pelo algoritmo de Gibbs, apresentada na Tabela 5. O gráfico sugere para a densidade marginal a posteriori para p uma função densidade Beta.

Figura 4 – Distribuição Marginal a Posteriori para p .

Na Figura 5, observa-se claramente um deslocamento de q comparado com p da Figura 4, verificando a importância de considerar o erro de resposta. O gráfico sugere também para a densidade marginal a posteriori para q uma função densidade Beta.

Figura 5 – Distribuição Marginal a Posteriori para q .

Um Exemplo Ilustrativo - Resultados e Análises

Na Figura 6, observa-se que o valor atribuído para θ_1 de 0,04 está em torno da média a posteriori obtida pelo algoritmo de Gibbs, apresentada na Tabela 5. O gráfico sugere para a densidade marginal a posteriori para θ_1 uma função densidade Beta.

Figura 6 – Distribuição Marginal a Posteriori para θ_1 .

Observe na Figura 7 que, a maior concentração de dados está entre 0 a 0,05, comparado com θ_1 que concentra-se entre 0 a 0,07 da Figura 6.

Figura 7 - Distribuição Marginal a Posteriori para x .

Um Exemplo Ilustrativo - Resultados e Análises

Observe na Figura 8, que o valor atribuído para θ_2 de 0,60 está em torno da média a posteriori obtida pelo algoritmo de Gibbs, apresentada na Tabela 5. O gráfico sugere para a densidade marginal a posteriori para θ_2 uma função densidade Beta.

Figura 8 – Distribuição Marginal a Posteriori para θ_2 .

Observe na Figura 9, pouca dispersão entre os dados, pois a variação relativa entre o desvio padrão e a média a posteriori de y , segundo resultados apresentados na Tabela 5, é de 26,34%.

Figura 9 - Distribuição Marginal a Posteriori para y .

4.1.2 Análise de convergência

- **Método gráfico**

É uma verificação informal da convergência dos valores gerados pela simulação para p , θ_1 e θ_2 . A Figura 10, apresenta a análise de convergência para p e a Figura 11, para θ_1 e θ_2 . Por meio de $m=5.000$ (iterações), a olho nú, podemos concluir pela convergência, pois observa-se uma nuvem de pontos concentrados em torno dos valores atribuídos inicialmente para cada parâmetro.

Figura 10 – Análise de Convergência para p .

Figura 11 – Análise de Convergência para θ_1 e θ_2 .

• Critério de Gelman / Rubin (1992)

É uma verificação formal da convergência que mede a variabilidade dentro de cadeias e entre cadeias. O critério é baseado em uma medida que chamaremos de $\hat{R}$ que será definido a seguir.

Vamos considerar $J=5$ (cadeias) a partir da iteração $m = 5.000$, com saltos $K= [0 \ 100 \ 200 \ 300 \ 400 \ 500]$ e de comprimento $N=1000$. Seja θ_{ij} o elemento i ($i=1, \dots, 1000$) da sequência j ($j=1, \dots, 5$) para p , θ_1 e θ_2 .

Definamos as seguintes variâncias:

1 – Variância dentro das cadeias:

$$W = \frac{1}{J} \sum_{j=1}^J s_j^2,$$

$$\text{onde} \quad s_j^2 = \frac{1}{N-1} \sum_{i=1}^N (\theta_{ij} - \bar{\theta}_{.j})^2$$

$$\bar{\theta}_{.j} = \frac{\sum_{i=1}^N \theta_{ij}}{N}.$$

2 - Variância entre as cadeias:

$$B = \frac{N}{J-1} \sum_{j=1}^J (\bar{\theta}_{.j} - \bar{\theta}_{..})^2,$$

$$\text{onde} \quad \bar{\theta}_{..} = \frac{1}{J} \sum_{j=1}^J \bar{\theta}_{.j}.$$

Um Exemplo Ilustrativo - Resultados e Análises

- 3 - Variabilidade Global (média ponderada) envolvendo os JN valores (variância a posteriori marginal estimada):

$$\hat{Var}(\theta / y) = \frac{(N-1)}{N}W + \frac{1}{N}B .$$

- 4 - Critério: $\sqrt{\hat{R}} = \sqrt{\frac{\hat{Var}(\theta \setminus y)}{W}}$, quando $\sqrt{\hat{R}} \sim 1$, concluímos pela convergência.

A Tabela 6, apresenta os valores de $\hat{R}$ para p , θ_1 e θ_2 . Observe que os dados convergem para os três parâmetros.

Tabela 6

Critério de Convergência de Gelman/Rubin para p , θ_1 e θ_2

	$\hat{R}$
p	1,014
θ_1	1,012
θ_2	1,011

4.1.3 Verificação da adequabilidade da distribuição Beta para p e q

Um teste estatístico foi utilizado com o objetivo de verificar se os dados gerados (seleção) por meio do Algoritmo de Gibbs, apresentados nos itens 3.3 e 3.4, para os parâmetros p e q , ajustam-se a uma densidade Beta.

Um Exemplo Ilustrativo - Resultados e Análises

Este teste consiste em calcular os parâmetros da distribuição a posteriori marginal de p e de q , segundo um coeficiente chamado “p-value” Bayesiano, sugerido por Gelman et al. (1995), isto é, assumindo “p-value” = 0,50 (valor de aceitabilidade do modelo), podemos calcular os parâmetros da distribuição de p , dados por (a_3, b_3) e da distribuição de q , dados por (a_4, b_4) , supondo uma distribuição a posteriori marginal Beta.

• “p-value” Bayesiano:

Vamos assumir as seguintes distribuições marginais a posteriori para p e q , dado os dados:

$$\pi(p | x) \sim B(a_3, b_3) \quad \text{e} \quad \pi(q | x) \sim B(a_4, b_4)$$

Nosso propósito é encontrar (a_3, b_3) e (a_4, b_4) , tal que as densidades marginais Beta de p e q deverão ser ajustadas pelo Algoritmo de Gibbs para $p^* = \{p^{(l)}, l = 1, \dots, 50\}$ e $q^* = \{q^{(l)}, l = 1, \dots, 50\}$, respectivamente. Como sugerido por Gelman et al., vamos definir “p-value” para o parâmetro p , da seguinte forma:

$$B_p = \frac{\sum_{j=1}^J I[T_{a_3, b_3}(p^{\text{rep}, j}) \geq T_{a_3, b_3}(p^*)]}{J}, \quad \text{onde } p^{\text{rep}, j} = \{p_1^{\text{rep}, j}, \dots, p_{50}^{\text{rep}, j}\}$$

I = variável indicadora.

para J réplicas, gerada por $\pi(p | x)$ para um valor fixado de (a_3, b_3) , onde,

$$T_{a_3, b_3}(p^*) = \sum_{l=1}^m \frac{(p^{(l)} - \frac{a_3}{a_3 + b_3})^2}{\frac{a_3 b_3}{(a_3 + b_3)^2 (a_3 + b_3 + 1)}}.$$

Escolher o par (a_3, b_3) , tal que “p-value” seja igual a 0,50. Mesmo procedimento é aplicado para (a_4, b_4) .

4.2 Comparação entre p e q

Uma medida que chamaremos de D_v , que mede a distância entre duas distribuições, sugerida por Peng e Dey (1995), foi utilizada para verificar a influência dos erros nos dados. Esta medida consiste em comparar as distribuições marginais a posteriori obtidas pelo algoritmo de Gibbs, apresentadas nos itens 3.3 e 3.4, entre p e q , isto é, vamos verificar se os dados da Figura 4 estão significativamente distantes dos dados da Figura 5, representados na Figura 12 abaixo.

Figura 12 – Densidades Marginais a Posteriori para p e q .

• Distância da Variação Total (Pend & Dey, 1995):

Vamos representar a distribuição marginal Beta para p e q , por : $\pi_p(\theta) = B(\theta; a_3, b_3)$ e $\pi_q(\theta) = B(\theta; a_4, b_4)$, respectivamente, onde (a_k, b_k) , $k = 3, 4$, são obtidos no primeiro estágio usando o “p-value” Bayesiano.

A distância da variação total entre estas duas densidades marginais é definida como:

$$D_v = D[\pi_p(\theta), \pi_q(\theta)] = \frac{1}{2} \int \left| \frac{\pi_p(\theta)}{\pi_q(\theta)} - 1 \right| \pi_q(\theta) d\theta . \quad (4.2.1)$$

Como em Peng e Dey (1995), a seguinte escala será utilizada para verificar a influência dos erros nos dados:

$$\begin{aligned} D_v > 0,25 &\Rightarrow \text{forte influência dos erros nos dados} \\ 0,1 \leq D_v \leq 0,25 &\Rightarrow \text{razoável influência dos erros nos dados} \\ D_v < 0,10 &\Rightarrow \text{nenhuma influência dos erros nos dados} \end{aligned}$$

Para estimar D_v , consideramos o seguinte procedimento computacional baseado em (4.2.1):

– de acordo com as distribuições marginais Beta $\pi_p(\theta)$ e $\pi_q(\theta)$ definidas anteriormente, geramos uma amostra $\theta^{(i)}$, $i = 1, \dots, 300$ por meio da distribuição marginal a posteriori de q , e estimamos D_v da seguinte forma :

$$\hat{D}_v = \frac{\sum_{i=1}^{300} \frac{1}{2} \left| \frac{\pi_p(\theta^{(i)})}{\pi_q(\theta^{(i)})} - 1 \right|}{300} .$$

Portanto, a medida estimada entre estas duas distribuições marginais é, $D_v = 0,4998$ (ver Anexo A). Isto indica uma forte influência dos erros nos dados segundo a escala apresentada acima.

4.3 Comparação entre o modelo proposto e o modelo de Gaba e Winkler

De acordo com os dados simulados para p , obtidos por meio do algoritmo de Gibbs, realizamos uma comparação entre o modelo misto proposto em (3.1.1) e o modelo exato de Gaba e Winkler dado em (2.1.5). No caso do modelo exato, considerou-se os

Um Exemplo Ilustrativo - Resultados e Análises

mesmos dados de p que foram selecionados anteriormente segundo o algoritmo de Gibbs.

A Figura 13, apresenta a distribuição marginal a posteriori exata para p , de Gaba e Winkler e a distribuição marginal a posteriori para p , proveniente do Amostrador de Gibbs. Observe que as duas distribuições praticamente se igualam.

Figura 13 – Distribuição Marginal a Posteriori para p .

O procedimento para obter a curva da distribuição marginal a posteriori para p , considerando o modelo proposto, é o mesmo descrito no item 4.1.3. Para o modelo exato de Gaba e Winkler, além de considerarmos os mesmos dados de p gerados pelo algoritmo de Gibbs, a função distribuição marginal a posteriori para p , dada em (2.1.5), depende do tamanho da amostra e do número de compradores que foram provenientes da variável dicotômica X , definida e gerada inicialmente no item 4.1. Assim, obtém-se a curva da Figura 13. (ver Anexo B)

Para garantirmos que o modelo misto proposto aproxima-se do modelo exato de Gaba e Winkler, vários estudos foram realizados com o objetivo de verificar a influência

Um Exemplo Ilustrativo - Resultados e Análises

dos erros em estudo sobre p , alterando-se o tamanho da amostra (n) e os valores tanto da priori quanto o valor real de θ_1 ou de θ_2 .

Observamos que, fixando os valores das priori para p , θ_1 e θ_2 e alterando-se o tamanho da amostra (n), não houve influência nas estatísticas a posteriori para os mesmos, isto é, os valores reais de $p = 0,25$, $\theta_1 = 0,04$ e $\theta_2 = 0,60$ mantêm-se em torno de suas médias a posteriori. Os resultados apresentados na Tabela 5 não alteraram significativamente.

Observamos também que, o tempo de execução do programa em Matlab, considerando θ_2 próximo de $(1 - \theta_1)$ é maior, devido a restrição $(1 - \theta_1) > \theta_2$. Por exemplo, se $\theta_1 = 0,30$ e $\theta_2 = 0,60$, isto implica que, $(1 - \theta_1) = 0,70$, está mais próximo de θ_2 do que, por exemplo, se θ_1 for igual a 0,03.

CAPÍTULO V

Conclusões

*"A verdadeira alegria da vida é
ser usada em prol de algo que você
mesmo considera grandioso."*

(George Shaw)

5. Conclusões

5.1 Conclusões

Os resultados obtidos neste trabalho, por meio do Amostrador de Gibbs com dados ampliados, mostraram que:

- 1 – É fácil verificar a influência do erro a respeito da proporção real de consumidores que compraram o produto ou marca, por meio gráfico, analisando o comportamento das distribuições marginais a posteriori, e por meio de uma medida que mede a distância da variação total entre estas distribuições;
- 2 – Considerando os dois tipos de erros mais frequentes em pesquisa de mercado que são: a proporção de compradores serem registrados *erroneamente* como não compradores do produto ou marca, e a proporção de não compradores serem registrados *erroneamente* como compradores do produto ou marca, mantém-se o efeito realístico de que o primeiro tipo de erro, é menos provável de ocorrer, do que o outro;
- 3 – O pesquisador terá dados de melhor qualidade, a medida que a proporção de pessoas que *realmente* compraram o produto ou marca aproxima-se da proporção de pessoas que compraram o produto ou marca com possível erro e os dois tipos de erros acima, tornarem-se nulos;
- 4 - Não há necessidade de considerarmos dados de painel para analisar a influência dos erros nos dados pois o modelo proposto, aproxima-se do modelo exato de Gaba e Winkler;
- 5 – A distribuição marginal a posteriori de p (proporção de consumidores que *realmente* compraram o produto ou marca), aproxima-se a uma densidade Beta.

5.2 Resumo das Contribuições

- 1 – Procedimento computacional baseado na mistura de duas distribuições de Bernoulli, incluindo erros de resposta, utilizando uma técnica de simulação: o Amostrador de Gibbs com variáveis latentes;
- 2 – Utilização de uma medida que mede, a distância da variação total, entre as distribuições marginais a posteriori para p (proporção de consumidores que *realmente* compraram o produto ou marca), e q (proporção de consumidores que compraram o produto ou marca com possível erro) com o objetivo de verificar a qualidade dos dados baseado nos erros;
- 3 – Pode-se dizer que, a distribuição marginal a posteriori exata para p , de Gaba e Winkler, comporta-se como uma densidade Beta.

5.3 Trabalhos Futuros

- 1 - Este trabalho pode ser aplicado em dados reais e em outras situações que envolvam variáveis dicotômicas com erros de classificação, como, por exemplo, pesquisas eleitorais (segundo turno) para Prefeito, Governador ou Presidente da República;
- 2 - O modelo misto proposto pode ser generalizado, por exemplo, em pesquisas eleitorais, se considerarmos mais de dois candidatos; em pesquisas de mercado, considerando vários produtos ou marcas e outros;
- 3 – Encontrar uma solução não Bayesiana, quando os parâmetros de erros, θ_1 e θ_2 , forem desconhecidos, utilizando restrições para a função de máxima verossimilhança.

A n e x o s

Anexo A – Programa em Matlab**INFERÊNCIA BAYESIANA PARA PESQUISA DE MERCADO
COM ERROS DE RESPOSTA UTILIZANDO MODELOS MISTOS.***Vanda D. R. Silveira**Orientador: Dr. Josemar Rodrigues*

clear;

% **Condições iniciais**

```
mint=5000; n0=0.4*mint; int=10; m=10; nint=50;
p(1) = 0.18; theta1(1)= 0.03; theta2(1) = 0.55;
    theta1r(1)= 1-theta1(1);
    theta2r(1)= theta2(1);
                                a = 10;  b = 30;
                                a1 = 2;  b1= 48;
                                a2 = 30; b2 = 20;
```

% **Dados : simulação**

for j=1:m

```
p_a(j) = binornd(1,0.25);
x1(j) = binornd(1,(1-0.04));
x2(j) = binornd(1,0.60);
end
x = p_a.*x1 + (1-p_a).*x2;
x =[ x]
```

% **Gerando Z, p, theta1 e theta2**

```
n=length(x); teste=0;
cont=0; i=1;
```

for i=1:mint

for k=1:n

$$p_1 = [p(i)*((1-\theta_1(i))^{x(k)}*\theta_1(i)^{(1-x(k))})]/[(p(i)*((1-\theta_1(i))^{x(k)}*\theta_1(i)^{(1-x(k))})+(1-p(i))*(\theta_2r(i)^{x(k)}*(1-\theta_2r(i))^{(1-x(k))}))];$$

```

        z(k)= binornd(1,p_1);
    end

    s1 =sum((1-x)'.*z');    s2 =sum(x'.*z');
    s3 =sum(x'.*(1-z)');    s4 =sum((1-x)'.*(1-z)');
    s5 =sum(z);

    theta1(i+1)= betarnd( a1 + s1, b1 + s2);
    theta1r(i+1)= 1 - theta1(i+1);
    p(i+1)= betarnd(a + s5,b + n - s5);

while teste==0
    theta2(i+1)=betarnd(a2 + s3,b2 + s4);

    if  theta1r(i+1) > theta2(i+1)

        theta2r(i+1)= theta2(i+1);
        cont=cont+1;
        teste=1;
    end
end
teste=0;
end
    for j=1:mint

        q(j) = p(j)*(1-theta1(j)) + (1-p(j))*theta2r(j);
        X(j) = (theta1(j)*p(j))/(1-q(j));
        Y(j) = (p(j) - X(j)*(1-q(j)))/q(j);

    end

%    Seleção da amostra

theta1s=theta1(n0+1:mint); theta1ss=theta1(n0+1:int:mint);
theta2s=theta2r(n0+1:mint); theta2ss=theta2r(n0+1:int:mint);
ps=p(n0+1:mint);          pss=p(n0+1:int:mint);
Xs=X(n0+1:mint);          Xss=X(n0+1:int:mint);
Ys=Y(n0+1:mint);          Yss=Y(n0+1:int:mint);
qs=q(n0+1:mint);          qss=q(n0+1:int:mint);

```

% Estatísticas

```
theta1pi= prctile(theta1ss,2.5); theta1ps= prctile(theta1ss,97.5);
theta1p1Q= prctile(theta1ss,25); theta1p3Q= prctile(theta1ss,75);
theta2pi= prctile(theta2ss,2.5); theta2ps= prctile(theta2ss,97.5);
theta2p1Q= prctile(theta2ss,25); theta2p3Q= prctile(theta2ss,75);
ppi= prctile(pss,2.5); pps= prctile(pss,97.5);
```

```
pp1Q= prctile(pss,25); pp3Q= prctile(pss,97.5);
Xpi= prctile(Xss,2.5); Xps= prctile(Xss,97.5);
Xp1Q= prctile(Xss,25); Xp3Q= prctile(Xss,75);
Ypi= prctile(Yss,2.5); Yps= prctile(Yss,97.5);
Yp1Q= prctile(Yss,25); Yp3Q= prctile(Yss,75);
qpi= prctile(qss,2.5); qps= prctile(qss,97.5);
qp1Q= prctile(qss,25); qp3Q= prctile(qss,75);
```

```
disp(['          Estatísticas a posteriori  ']);
disp(['-----']);
disp('          2.5%   25%   mediana   75%   97.5%   média   d_ padrão ');
disp(['-----']);
disp(['p      ',num2str(ppi),' ',num2str(pp1Q),' ',num2str(median(pss)),' ',num2str(pp3Q),' ',num2str(pps),' ',num2str(mean(pss)),' ',num2str(std(pss))]);
disp([' ']);
disp(['theta1 ',num2str(theta1pi),' ',num2str(theta1p1Q),' ',num2str(median(theta1ss)),' ',num2str(theta1p3Q),' ',num2str(theta1ps),' ',num2str(mean(theta1ss)),' ',num2str(std(theta1ss))]);
disp([' ']);
disp(['theta2 ',num2str(theta2pi),' ',num2str(theta2p1Q),' ',num2str(median(theta2ss)),' ',num2str(theta2p3Q),' ',num2str(theta2ps),' ',num2str(mean(theta2ss)),' ',num2str(std(theta2ss))]);
disp([' ']);
disp(['-----']);
disp(['x      ',num2str(Xpi),' ',num2str(Xp1Q),' ',num2str(median(Xss)),' ',num2str(Xp3Q),' ',num2str(Xps),' ',num2str(mean(Xss)),' ',num2str(std(Xss))]);
disp([' ']);
disp(['y      ',num2str(Ypi),' ',num2str(Yp1Q),' ',num2str(median(Yss)),' ',num2str(Yp3Q),' ',num2str(Yps),' ',num2str(mean(Yss)),' ',num2str(std(Yss))]);
disp([' ']);
disp(['q      ',num2str(qpi),' ',num2str(qp1Q),' ',num2str(median(qss)),' ',num2str(qp3Q),' ',num2str(qps),' ',num2str(mean(qss)),' ',num2str(std(qss))]);
disp(['-----']);
```


% Parâmetros :distribuição Beta para p,theta1 e theta2 - seleção

```
t_a1 = mean(theta1ss)*(mean(theta1ss)-(mean(theta1ss))^2 - (std(theta1ss))^2)/((std(theta1ss))^2);
t_b1 = mean(theta1ss)*((1 - mean(theta1ss))^2)/((std(theta1ss))^2 - (1 - mean(theta1ss)));
t_a2 = mean(theta2ss)*(mean(theta2ss)-(mean(theta2ss))^2 - (std(theta2ss))^2)/((std(theta2ss))^2);
t_b2 = mean(theta2ss)*((1 - mean(theta2ss))^2)/((std(theta2ss))^2 - (1 - mean(theta2ss)));
t_a = mean(pss)*(mean(pss)-(mean(pss))^2 - (std(pss))^2)/((std(pss))^2);
t_b = mean(pss)*((1 - mean(pss))^2)/((std(pss))^2 - (1 - mean(pss)));

X_a = mean(Xss)*(mean(Xss)-(mean(Xss))^2 - (std(Xss))^2)/((std(Xss))^2);
X_b = mean(Xss)*((1 - mean(Xss))^2)/((std(Xss))^2 - (1 - mean(Xss)));
Y_a = mean(Yss)*(mean(Yss)-(mean(Yss))^2 - (std(Yss))^2)/((std(Yss))^2);
Y_b = mean(Yss)*((1 - mean(Yss))^2)/((std(Yss))^2 - (1 - mean(Yss)));
q_a = mean(qss)*(mean(qss)-(mean(qss))^2 - (std(qss))^2)/((std(qss))^2);
q_b = mean(qss)*((1 - mean(qss))^2)/((std(qss))^2 - (1 - mean(qss)));
```

% Estatística Qui- Quadrado para p, theta1 e theta2 - seleção

```
soma1=0; soma2=0; soma3=0; i=0;
```

```
for i=1:length(pss)
```

```
    soma1=soma1 + (( pss(i) - (t_a/(t_a+t_b)))^2)/((t_a*t_b)/((t_a+t_b)^2*(t_a+t_b+1)));
    soma2=soma2 + (((theta1ss(i)- (t_a1/(t_a1+t_b1)))^2)/
        ((t_a1*t_b1)/((t_a1+t_b1)^2*(t_a1+t_b1+1)));
    soma3=soma3 + (((theta2ss(i)- (t_a2/(t_a2+t_b2)))^2)/
        ((t_a2*t_b2)/((t_a2+t_b2)^2*(t_a2+t_b2+1)));
```

```
end
```

```
    Tp=soma1; Ttheta1=soma2; Ttheta2=soma3;
```

% Réplicas para p,theta1 e theta2 - seleção

```
cont1=0; cont2=0; cont3=0; i=0;
```

```
for i=1:nint
```

```
    for j=1:length(pss)
```

```
        p_rep(i,j)=betarnd(t_a,t_b);
        theta1_rep(i,j)=betarnd(t_a1,t_b1);
        theta2_rep(i,j)=betarnd(t_a2,t_b2);
```

```
    end
```

```
end
```

```

P=[p_rep]; Q=[P - t_a/(t_a+t_b)];
E_p=t_a/(t_a+t_b); Var_p=(t_a*t_b)/((t_a+t_b)^2*(t_a+t_b+1));

R1=[theta1_rep]; R2=[R1 - t_a1/(t_a1+t_b1)];
E_theta1=t_a1/(t_a1+t_b1); Var_theta1=(t_a1*t_b1)/((t_a1+t_b1)^2*(t_a1+t_b1+1));

R3=[theta2_rep]; R4=[R3 - t_a2/(t_a2+t_b2)];
E_theta2=t_a2/(t_a2+t_b2); Var_theta2=(t_a2*t_b2)/((t_a2+t_b2)^2*(t_a2+t_b2+1));

Tp_rep=sum(Q'.^2)/Var_p;
Tth1_rep=sum(R2'.^2)/Var_theta1;
Tth2_rep=sum(R4'.^2)/Var_theta2;

% Cálculo do "p-Value"

for i=1:nint
    if Tp_rep(i) >= Tp
        cont1=cont1+1;
    end
    if Tth1_rep(i) >= Ttheta1
        cont2=cont2+1;
    end
    if Tth2_rep(i) >= Ttheta2
        cont3=cont3+1;
    end
end
pv_p=cont1/nint; pv_theta1=cont2/nint; pv_theta2=cont3/nint;

disp(['      ']);
disp([' p-value para p theta-1 e theta-2 : ']);
disp(['      ']);
disp([' pv_p : ', num2str(pv_p)]);
disp([' pv_theta1 : ', num2str(pv_theta1)]);
disp([' pv_theta2 : ', num2str(pv_theta2)]);

% Estatística Qui-Quadrado para x, y e q - seleção

soma4=0; soma5=0; soma6=0; i=0;

for i=1:length(Xss)

```

```

soma4=soma4 + ((Xss(i) - (X_a/(X_a+X_b)))^2)/
((X_a*X_b)/((X_a+X_b)^2*(X_a+X_b+1)));

soma5=soma5 + ((Yss(i) - (Y_a/(Y_a+Y_b)))^2)/
((Y_a*Y_b)/((Y_a+Y_b)^2*(Y_a+Y_b+1)));
soma6=soma6 + ((qss(i) - (q_a/(q_a+q_b)))^2)/ ((q_a*q_b)/((q_a+q_b)^2*(q_a+q_b+1)));
end
Tx = soma4; Ty = soma5; Tq = soma6;

```

% Réplicas para x, y e q - seleção

```

cont4=0; cont5=0; cont6=0; i=0;

for i=1:nint
    for j=1:length(Xss)

X_rep(i,j)=betarnd(X_a,X_b); Y_rep(i,j)=betarnd(Y_a,Y_b);
q_rep(i,j)=betarnd(q_a,q_b);

        end
    end
    G1=[X_rep]; G2=[G1 - X_a/(X_a+X_b)];
    E_X=X_a/(X_a+X_b); Var_X= (X_a*X_b)/((X_a+X_b)^2*(X_a+X_b+1));

    G3=[Y_rep]; G4=[G3 - Y_a/(Y_a+Y_b)];
    E_Y=Y_a/(Y_a+Y_b); Var_Y= (Y_a*Y_b)/((Y_a+Y_b)^2*(Y_a+Y_b+1));

    G5=[q_rep]; G6=[G5 - q_a/(q_a+q_b)];
    E_q=q_a/(q_a+q_b); Var_q= (q_a*q_b)/((q_a+q_b)^2*(q_a+q_b+1));

    Tx_rep=sum(G2'.^2)/Var_X;
    Ty_rep=sum(G4'.^2)/Var_Y;
    Tq_rep=sum(G6'.^2)/Var_q;

```

% Cálculo do "p-Value"

```

for i=1:nint
    if Tx_rep(i) >= Tx
        cont4=cont4+1;
    end
    if Ty_rep(i) >= Ty
        cont5=cont5+1;
    end
end

```

```

    if Tq_rep(i) >= Tq
        cont6=cont6+1;
    end
end
pv_x=cont4/nint; pv_y=cont5/nint; pv_q=cont6/nint;

disp(['      ']);
disp(['    p-value para x y e q : ']);
disp(['      ']);
disp(['    pv_x      : ', num2str(pv_x)]);
disp(['    pv_y      : ', num2str(pv_y)]);
disp(['    pv_q      : ', num2str(pv_q)]);

```

% Comparação entre p e q (seleção)

```

soma7=0; k=0; q_s=sort(qss); p_s=sort(pss);

for k=1:length(qss)

    B_p(k) = betapdf(qss(k),t_a,t_b);
    B_q(k) = betapdf(qss(k),q_a,q_b);

    soma7 = soma7 + (abs((B_p(k)/B_q(k)) - 1));
end
med = soma7/(2*length(qss));

disp(['      ']);
disp(['    Comparação entre p e q : ']);
disp(['      ']);
disp(['    D_v      : ', num2str(med)]);

b_p = betapdf(p_s,t_a,t_b);
b_q = betapdf(q_s,q_a,q_b);

```

% Gráficos

```

whitebg('w');
figure (1)
histfit2('beta',30,theta1ss,t_a1,t_b1)
% title('      Histograma - densidade para theta1 ')
xlabel('theta1')
pause
figure (2)

```

```
histfit2('beta',30,theta2ss,t_a2,t_b2)
% title('      Histograma - densidade para theta2')
xlabel('theta2')
pause
figure(3)
histfit2('beta',30,pss,t_a,t_b)
% title('      Histograma - densidade para p ')
xlabel('p')
pause
t1=[theta1']; t2=[theta2r'];
figure (4)
plot(p,'*')
% title(' Análise de Convergência para p')
xlabel('p')
pause
figure(5)
subplot(1,2,1), plot(t1,'+')
% title('      Análise de Convergência ')
xlabel('theta1')
subplot(1,2,2), plot(t2,'+')
xlabel('theta2')
pause
figure(6)
histfit2('beta',30,Xss,X_a,X_b)
% title(' Histograma -densidade para x')
xlabel('x')
pause
figure(7)
histfit2('beta',30,Yss,Y_a,Y_b)
% title(' Histograma -densidade para y')
xlabel('y')
pause
figure(8)
histfit2('beta',30,qss,q_a,q_b)
% title(' Histograma -densidade para q')
xlabel('q')
pause
figure(9)
plot(p_s,b_p,'b',q_s,b_q,'r');
ylabel('Densidade de p e q')
pause
```

% Convergência - Critério de Gelman/Rubin

```
nc=5;
d=length(p)/nc;

n1=[0 100 200 300 400 500];

for i=1:5
    Mep(i)=[mean(p(n1(i)+1:n1(i+1)))];
    Vap(i)=[cov(p(n1(i)+1:n1(i+1)))];
end

QMp=Mep.*Mep;
MeMp=sum(Mep)/nc;

Wp=sum(Vap)/nc;

VarVp=(Vap*Vap'-nc*Wp^2)/(nc-1);
Up=d*(Mep*Mep'-nc*MeMp^2)/(nc-1);
VARp=((d-1)/d)*Wp+(1/d)*Up;
VHp=VARp+Up/(nc*d);
Mcovp=cov(Vap,Mep);
Qcovp=cov(Vap,QMp);
COVp=Qcovp(1,2)-2*MeMp*Mcovp(1,2);
VarVHp=((d-1)/d)^2*(1/nc)*VarVp+(((nc+1)/(nc*d))^2)*(2/(nc-1))*Up^2+2*COVp*(nc+1)*(d-1)/(d*nc^2);
dfp=2*(VHp^2)/VarVHp;

Rp=sqrt((VHp/Wp)*dfp/(dfp-2));

for i=1:5
    Met1(i)=[mean(theta1(n1(i)+1:n1(i+1)))];
    Vat1(i)=[cov(theta1(n1(i)+1:n1(i+1)))];
end

QMt1=Met1.*Met1;
MeMt1=sum(Met1)/nc;

Wt1=sum(Vat1)/nc;
VarVt1=(Vat1*Vat1'-nc*Wt1^2)/(nc-1);
Ut1=d*(Met1*Met1'-nc*MeMt1^2)/(nc-1);

VARt1=((d-1)/d)*Wt1+(1/d)*Ut1;
```

```

    VHt1=VARt1+Ut1/(nc*d);
    Mcovt1=cov(Vat1,Met1);
    Qcovt1=cov(Vat1,QMt1);
    COVt1=Qcovt1(1,2)-2*MeMt1*Mcovt1(1,2);
    VarVHt1=((d-1)/d)^2*(1/nc)*VarVt1+(((nc+1)/(nc*d))^2*(2/(nc-1))* Ut1^2+2*
        COVt1*(nc+1)*(d-1)/(d*nc^2);
    dft1=2*(VHt1^2)/VarVHt1;
    Rt1=sqrt((VHt1/Wt1)*dft1/(dft1-2));

for i=1:5
    Met2(i)=[mean(theta2r(n1(i)+1:n1(i+1)))];
    Vat2(i)=[cov(theta2r(n1(i)+1:n1(i+1)))];
end

    QMt2=Met2.*Met2;
    MeMt2=sum(Met2)/nc;
    Wt2=sum(Vat2)/nc;
    VarVt2=(Vat2*Vat2'-nc*Wt2^2)/(nc-1);
    Ut2=d*(Met2*Met2'-nc*MeMt2^2)/(nc-1);
    VARt2=((d-1)/d)*Wt2+(1/d)*Ut2;

    VHt2=VARt2+Ut2/(nc*d);
    Mcovt2=cov(Vat2,Met2);
    Qcovt2=cov(Vat2,QMt2);
    COVt2=Qcovt2(1,2)-2*MeMt2*Mcovt2(1,2);
    VarVHt2=((d-1)/d)^2*(1/nc)*VarVt2+(((nc+1)/(nc*d))^2*(2/(nc-1))* Ut2^2+2*
        COVt2*(nc+1)*(d-1)/(d*nc^2);
    dft2=2*(VHt2^2)/VarVHt2;
    Rt2=sqrt((VHt2/Wt2)*dft2/(dft2-2));

disp(['      ']);
disp(['      ']);

disp([' Critério de Convergência de G/Rubin para p theta-1 e theta-2 : ']);
disp(['      ']);
disp(['      Rp      : ', num2str(Rp)]);
disp(['      Rtheta1 : ', num2str(Rt1)]);
disp(['      Rtheta2 : ', num2str(Rt2)]);

```

Anexo B - Programa em Matlab

INFERÊNCIA BAYESIANA PARA PESQUISA DE MERCADO COM ERROS DE RESPOSTA UTILIZANDO MODELOS MISTOS.

COMPARAÇÃO ENTRE O (AG) E O MODELO DE GABA\WINKLER

Vanda D. R. Silveira

Orientador: Dr. Josemar Rodrigues

```
clear;

%   Condições iniciais

mint=5000; n0=0.4*mint; int=10; m=10;
p(1) = 0.20; theta1(1)= 0.03; theta2(1) = 0.55;
    theta1r(1)= 1-theta1(1);
    theta2r(1)= theta2(1);

                                a = 10;  b = 30;
                                a1 = 2;  b1 = 48;
                                a2 = 30; b2 = 20;

%   Dados : simulação

for j=1:m

    p_a(j) = binornd(1,0.25);
    x1(j) = binornd(1,(1-0.04));
    x2(j) = binornd(1,0.60);
    x(j) = p_a(j)*x1(j)+(1-p_a(j))*x2(j);
end

%   Gerando Z, p, theta1 e theta2

n=length(x); teste=0; cont=0; i=1;

    for i=1:mint
        for k=1:m
            p_1 = [p(i)*((1-theta1(i))^x(k))*theta1(i)^(1-x(k))]/[p(i)*((1-theta1(i))^x(k))*theta1(i)^(1-x(k)))+(1-p(i))*(theta2r(i)^x(k))*(1-theta2r(i))^(1-x(k))];
            z(k) = binornd(1,p_1);
```



```

end

s1=sum((1-x)'.*z');    s2=sum(x'.*z');
s3=sum(x'.*(1-z));    s4=sum((1-x)'.*(1-z));
s5=sum(z);

theta1(i+1)=betarnd( a1 + s1, b1 + s2);  theta1r(i+1)= 1 - theta1(i+1);

p(i+1)=betarnd(a + s5,b + n - s5);

while teste==0
    theta2(i+1)=betarnd(a2 + s3,b2 + s4);

    if  theta1r(i+1) > theta2(i+1)

        theta2r(i+1)= theta2(i+1);
        cont=cont+1;
        teste=1;
    end
end
teste=0;
end

%   Seleção da amostra

theta1s=theta1(n0+1:mint);  theta1ss=theta1(n0+1:int:mint);
theta2s=theta2r(n0+1:mint); theta2ss=theta2r(n0+1:int:mint);
ps=p(n0+1:mint);            pss=p(n0+1:int:mint);

%   Estatísticas

theta1pi= prctile(theta1ss,2.5);  theta1ps= prctile(theta1ss,97.5);
theta1p1Q= prctile(theta1ss,25); theta1p3Q= prctile(theta1ss,75);
theta2pi= prctile(theta2ss,2.5);  theta2ps= prctile(theta2ss,97.5);
theta2p1Q= prctile(theta2ss,25); theta2p3Q= prctile(theta2ss,75);
ppi= prctile(pss,2.5);            pps= prctile(pss,97.5);
pp1Q= prctile(pss,25);            pp3Q= prctile(pss,97.5);

disp(['          Estatísticas a posteriori  ']);
disp(['-----']);
disp('      2.5%    25%    mediana    75%    97.5%    média    d_padrão ');
disp(['-----']);
disp(['p      ',num2str(ppi),' ',num2str(pp1Q),' ', num2str(median(pss)),' ',num2str(pp3Q),' ',num2str(pps),' ',num2str(mean(pss)),' ',num2str(std(pss))]);

```

```

disp(['      ']);

disp(['theta1 ',num2str(theta1pi),' ',num2str(theta1p1Q),' ', num2str(median(theta1ss)),',
',num2str(theta1p3Q),' ',num2str(theta1ps),' ',num2str(mean(theta1ss)),',
',num2str(std(theta1ss))]);
disp(['      ']);
disp(['theta2 ',num2str(theta2pi),' ',num2str(theta2p1Q),' ', num2str(median(theta2ss)),',
',num2str(theta2p3Q),' ',num2str(theta2ps),' ',num2str(mean(theta2ss)),',
',num2str(std(theta2ss))]);
disp(['      ']);
disp(['-----']);

%      Parâmetros :distribuição Beta para p - seleção

t_a = mean(pss)*(mean(pss)-(mean(pss))^2 - (std(pss))^2)/((std(pss))^2);
t_b = mean(pss)*((1 - mean(pss))^2)/((std(pss))^2 - (1 - mean(pss)));

%      Densidade Beta(t_a,t_b) para p

p_s=sort(pss);      B_p=betapdf(p_s,t_a,t_b);

% modelo proposto por GABA & WINKLER- f(p|r,n)

cont=0; n2=length(p_s);

for i=1:m
    if x(i)==1
        cont=cont+1;
    end
end
r=cont;

for l=1:n2
    for j=1:r+1
        for t=1:m-r+1
            comb1(t) = gamma(m-r+1)/(gamma(m-r-t+2)*gamma(t));
            B_ab1(t) = gamma(a+m-t+1-j+1)*gamma(b+t-2+j)/gamma(a+m+b);
            B_alb11(t) = gamma(a1+m-r-t+1)*gamma(b1+r-j+1)/gamma(a1+b1+m-t-j+2);
            B_a2b21(t) = (gamma(a2+j-1)*gamma(b2+t-1)/gamma(a2+b2+t+j-2))*betapdf(p_s(l),a+m-j-
                t+2,b+j+t-2);

            T_1(t) = comb1(t)*B_ab1(t)*B_alb11(t)*B_a2b21(t);
            B_a2b21s(t) = gamma(a2+j-1)*gamma(b2+t-1)/gamma(a2+b2+t+j-2);
            T_1s(t) = comb1(t)*B_ab1(t)*B_alb11(t)*B_a2b21s(t);
        end
    end
end

```

```
end

    comb2(j) = gamma(r+1)/(gamma(r-j+2)*gamma(j));

    T_2s(j) = comb2(j)*sum(T_1s);
    T_2(j) = comb2(j)*sum(T_1);
end
    S1(l)=sum(T_2);
end
    S1s = sum(T_2s);

    f_p = S1./S1s;

% Gráficos

whitebg('w');
figure(1)
histfit2('beta',30,pss,t_a,t_b)
% title('      Histograma - densidade para p ')
xlabel('p')
pause
figure(2)
plot(p_s,f_p,'b');
title('  função densidade de Gaba\Winkler ')
xlabel(' p ')
ylabel(' Densidade de p')
pause
figure(3)
plot(p_s,B_p);
title('  função densidade de Gibbs Sampling')
xlabel(' p ')
ylabel(' Densidade de p ')
pause
figure(4)
plot(p_s,f_p,'b',p_s,B_p);
title('  função densidade de p ')
% xlabel(' p (Gaba/Winkler) e p(simulação) ')
ylabel(' Densidade ')
pause
```

Referências Bibliográficas

Referências Bibliográficas

- CELEUX, G.; DIEBOLT, J. *The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comput. Statist. Q.*, v. 2, p. 73-82, 1985.
- *A stochastic approximation type EM algorithm. Stochastics*, to be published, 1992 .
- CLANCY, K.; OSTLUND, L.E.; WYNER, G. A . *False reporting of Magazine Readership, J. Adverstising Res.*, v. 19, p. 23-30, 1979.
- DEMPSTER, A .P.; LAIRD, N.M.; RUBIN, D.B. *Maximum likelihood from incomplete data via the EM algorithm (with discussion). J.R. Statist. Soc. B*, v. 39, p. 1-38, 1977.
- DIEBOLT, J.; ROBERT, C.P. *Estimation of finite mixture distributions throught Bayesian Sampling. J.R. Statist. Soc. B*, v. 56, n. 2, p. 363-375, 1994.
- ESCOBAR, M.D.; WEST, M. *Bayesian prediction and density estimation. J. Am. Statist. Assoc.*, to be published, 1991.
- GABA, A .; WINKLER, R. *Implications of errors in survey data : A Bayesian model. Management Science*, v. 38, n. 7, p. 913-925, 1992.
- GELMAN, A .E.; RUBIN, D. *Inference from iterative simulation using multiple sequences. Satisfistical Science*, 7, p. 457-472, 1992.
- GELMAN, A .; CARLIN, J.B.; STERN, H.S.; RUBIN, D.B. *Bayesian Data Analysis. Chapman & Hall*, 1995.
- GELFAND, A .E. ; SMITH, A .F.M. . *Sampling - Based approaches to calculating marginal densities. J. Amer. Statist. Assoc.*, 85, p. 398-409, 1990.
- MORRISON, D. G. *Purchase intentions and purchase behavior. J. Marketing* , v. 43, p. 65-74, 1979.

Referências Bibliográficas

- PARFITT, J. *A comparison of purchase recall with diary panel records*. **J. Advertising Res.**, v. 7, p. 16-31, 1967.
- PENG, F.; DEY, D.K. *Bayesian analysis of outlier problems using divergence measures*. **The Canadian J. of Statistics**, v.23, n. 2, p. 199-213, 1995.
- PETERSON, R. A.; KERIN, R. A. *The quality of self-report data : Review and Synthesis*, in B. M. Ennis and K. J. Roering (Eds.), **Review of Marketing, American Marketing Association**. Chicago, p. 5-20, 1981.
- RODRIGUES, J.; SILVEIRA, V.D.R. *Bayesian computation for dichotomous variables with classification errors*. **Notas, Série Estatística do Instituto de Ciências Matemáticas e de Computação de São Carlos**, nº 55.
- ROSS, S. **A First Course in Probability**. *Prentice Hall*, 4th ed., 1994.
- SING, S.N.; CHURCHILL, G. A. *Using the theory of signal detection to improve ad recognition testing*. **J. Marketing Res.**, v. 23, p. 327-336, 1986.
- SMITH, A. F. M.; ROBERTS, G. O. *Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods*. **J. R. Statist. Soc. B**, v. 55, p. 3-23, 1993.
- TANNER, M.; WONG, W. *The calculation of posterior distributions by data augmentation*. **J. Amer. Statist. Assoc.**, 82, p. 528-550, 1987.
- WIND, Y.; LERNER, D. *On the measurement of purchase data : Surveys versus Purchase Diaries*. **J. Marketing Res.**, v. 16, p. 39-47, 1979.