

MAGDA CARVALHO PIRES

ANÁLISE BAYESIANA EMPÍRICA DE DADOS DICOTÔMICOS
COM ERROS E CLASSIFICAÇÕES REPETIDAS

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Roberto da Costa Quinino

BELO HORIZONTE, 01 DE MARÇO 2006

Agradecimentos

- Agradecimento maior ao meu orientador Roberto Quinino pelo projeto proposto, a confiança depositada, a atenção dispensada, o acompanhamento contínuo, as noites sem dormir analisando cada detalhe desse trabalho.
- Agradecimentos ao Anderson Laécio pela disponibilidade e presteza, pelo auxílio inestimável no programa desenvolvido e na editoração do texto. Aos Profs. Emílio Suyama, Cibele Queiroz e Marta Afonso Freitas pelas ótimas sugestões e correções apresentadas durante o estudo, e pela atenção dispensada na avaliação dessa dissertação.
- Aos meus pais, pelo amor e apoio incondicionais durante esses anos de estudo. À minha irmã Erika, amiga e iniciadora nas artes da Estatística. Ao grande irmão Magno pela amizade e descontração nas horas de stress. Ao pequeno irmão Douglas pelo carinho e doçura inspiradores.
- Ao Vitor, pelo amor e companheirismo em todas as horas, compreensão nos momentos de ausência e incentivo renovador.
- Agradecimentos a todos os amigos e familiares que acreditaram nessa conquista, e principalmente a Deus, por proporcionar-me vencer mais esta jornada, crescendo espiritual e intelectualmente com saúde e na presença de tantas pessoas especiais em minha vida.

Análise bayesiana empírica de dados dicotômicos com erros e classificações repetidas

Magda Carvalho Pires

Roberto da Costa Quinino

Orientador

Departamento de Estatística - ICEX - UFMG

31270-901 - Belo Horizonte - MG - Brazil

Março - 2006

Resumo

Considera-se o problema da estimação bayesiana de uma proporção p de interesse onde a classificação das unidades está sujeita a erros de diagnóstico. Na abordagem Bayesiana, a utilização de distribuições *a priori* Uniforme com parâmetros zero e um para os erros de classificação e para proporção de interesse geram uma média *a posteriori* para proporção igual 0,50 independentemente do resultado amostral, além de grande variabilidade. É necessário, portanto, que a distribuição *a priori* seja informativa, o que nem sempre é possível. Neste trabalho, utiliza-se classificações repetidas e distribuição *a priori* empírica para apresentar uma solução ao problema. Resultados de simulação indicam que a metodologia desenvolvida apresenta uma boa estimativa da proporção de interesse quando o número de classificações repetidas é igual ou superior a três.

Palavras-Chave: Análise Bayesiana, Erros de Classificação, Classificações Repetidas, Método Bayes Empírico, Distribuição Binomial

1 Introdução

Na implementação do controle de qualidade de atributos, a eficiência do sistema que classifica os itens manufaturados como conforme ou não-conforme precisa ser considerada. Dois tipos de erros podem ocorrer durante a inspeção: o primeiro, conhecido como tipo I, ocorre quando um item conforme é classificado como não-conforme; e o segundo, denotado por tipo II, quando um item é dito conforme quando é, na verdade, não-conforme.

Pioneiramente, Bross (1954) mostrou que, na presença de erros de classificação, os estimadores obtidos por uma abordagem estatística clássica são extremamente viciados. Outros autores, como Johnson e Kotz (1988), Johnson *et al.* (1991), Evans *et al.* (1996), Viana (1994), Gustafson (2003) enfatizaram que os erros de classificação, quando ignorados, podem comprometer todo o processo de inferência e, conseqüentemente, o controle de qualidade.

Suponha que, numa amostra aleatória de n unidades, um número X de itens conformes é observado. Essa variável aleatória X tem distribuição binomial com parâmetros (n, p) , ou seja, $X \sim Bin(n, p)$. Contudo, a presença de erros de classificação no sistema implica numa modificação dessa função de probabilidade. Seja e_1 a probabilidade de que um item conforme seja erroneamente classificado como não-conforme, e seja e_2 a probabilidade de que um item não-conforme seja classificado como conforme. Então, a probabilidade de que um item seja classificado como conforme é $q = p(1 - e_1) + (1 - p)e_2$, definindo uma variável aleatória X que tem distribuição binomial com parâmetro q ao invés de p .

A dificuldade de análise pode ser melhor compreendida através da determinação do estimador de máxima verossimilhança. A função de verossimilhança para o caso com erros de classificação pode ser expressa como $L(x|n, q) = q^x(1 - q)^{n-x}$. Esta é maximizada para todos os pontos (p, e_1, e_2) tais que $p(1 - e_1) + (1 - p)e_2 = x/n$ (GABA; WINKLER, 1992). Portanto, o estimador de máxima verossimilhança não é único.

Para resolver essa questão, muitos métodos clássicos foram sugeridos e uma revisão pode ser encontrada em Johnson *et al.* (1991). Em geral, os métodos propostos utilizam-

se de planos amostrais alternativos para estimação preliminar dos erros de classificação. Numa ótica bayesiana, Gaba e Winkler (1992) consideraram uma abordagem que requer a utilização de uma distribuição *a priori* informativa. Isto pode ser uma restrição considerável, pois em muitos casos essa informação não está disponível. Constataram que a utilização de distribuições *a priori* não informativas independentes e uniformes entre zero e um para os parâmetros (p, e_1, e_2) gera uma média *a posteriori* de p igual a $\frac{1}{2}$, independentemente do resultado amostral e , além disso, todos os pontos (p, e_1, e_2) tais que $p(1 - e_1) + (1 - p)e_2 = x/n$ eram modas *a posteriori*.

Em trabalhos sobre tamanho amostral bayesiano para dados dicotômicos na presença de erros de classificação, Dendukuri *et al.* (2004) e Rahme *et al.* (2000) também observaram a necessidade primordial de uma distribuição *a priori* informativa.

Neste artigo, propõe-se um modelo em que o processo de inferência Bayesiana para proporção na presença de erros de classificação incorpora a realização de classificações repetidas tanto para elicitare uma distribuição *a priori* empírica como para minimizar o impacto desses erros. A classificação final de um item será aquela que apresentar maioria nas classificações repetidas. Em termos práticos, considera-se que realizar classificações repetidas pode ser mais fácil e operacional do que obter distribuições *a priori* informativas.

A seção 2 apresenta um esquema para incorporar classificações repetidas com respectiva determinação da função de verossimilhança. Na seção 3 uma análise Bayesiana empírica para a proporção de interesse é apresentada, com exemplos numéricos descritos na seção 4. Conclusões são apresentadas na seção 5.

2 Função de verossimilhança

Suponha que cada item de uma amostra aleatória de tamanho n seja classificado m vezes, m ímpar, independentemente como conforme ou não-conforme. Seja C_{ij} ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, m$) uma variável aleatória Bernoulli correspondente à j -ésima classificação do i -ésimo item. Assim, $C_{2,3} = 1$ significa que o segundo item foi classificado como

conforme na terceira classificação. Seja F_i uma variável aleatória Bernoulli que denota a classificação final do i -ésimo item após as m classificações. Considere que $F_i = 1$ se, e somente se, $\sum_{j=1}^m C_{ij} > 0,5m$. A Tabela 1 apresenta a descrição desse procedimento de classificação.

Seja ainda E_i outra variável aleatória Bernoulli, que denota o estado real da i -ésima peça, de tal forma que o interesse seja estimar $P(E_i = 1) = p$. Desta forma, temos que $e_1 = P(C_{ij} = 0 \mid E_i = 1)$ e $e_2 = P(C_{ij} = 1 \mid E_i = 0)$. Então, a probabilidade de que uma peça seja classificada como conforme é dada por

$$P(F_i = 1) = p \text{Bin}(m; e_1; 0,5m) + (1 - p) [1 - \text{Bin}(m; e_2; 0,5m)] \quad (1)$$

em que $\text{Bin}(m; e_k; 0,5m)$ denota a função de distribuição acumulada Binominal definida no ponto $0,5m$. Observe que se $m \rightarrow \infty$ e as probabilidades associadas aos erros de classificação forem menores do que $0,5$ então (1) converge para p , corroborando o benefício da utilização de classificações repetidas.

Supondo agora uma amostra aleatória de n itens com r deles considerados conformes, a função de verossimilhança pode ser expressa por

Tabela 1: Classificações repetidas de n itens m vezes cada

Item	Classificações (C_{ij})					Classificação Final
	1	3	5	\dots	m	
1	C_{11}	C_{13}	C_{15}	\dots	C_{1m}	F_1
2	C_{21}	C_{23}	C_{25}	\dots	C_{2m}	F_2
3	C_{31}	C_{33}	C_{35}	\dots	C_{3m}	F_3
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
n	C_{n1}	C_{n2}	C_{n3}	\dots	C_{nm}	F_n

$$L(r|n, m, p, e_1, e_2) = \{p \text{Bin}(m; e_1; 0, 5m) + (1-p)[1 - \text{Bin}(m; e_2; 0, 5m)]\}^r \times \\ \{1 - [p \text{Bin}(m; e_1; 0, 5m) + (1-p)[1 - \text{Bin}(m; e_2; 0, 5m)]]\}^{n-r} \quad (2)$$

que pode ser reescrita como

$$L(r|n, m, p, e_1, e_2) = \sum_{j=0}^r \sum_{t=0}^{n-r} \binom{r}{j} \binom{n-r}{t} p^{n-j-t} (1-p)^{j+t} \times \\ [\text{Bin}(m; e_1; 0, 5m)]^{r-j} [1 - \text{Bin}(m; e_1; 0, 5m)]^{n-r-t} \times \\ [\text{Bin}(m; e_2; 0, 5m)]^t [1 - \text{Bin}(m; e_2; 0, 5m)]^j \quad (3)$$

Note que se $m = 1$, então (2) é igual a

$$L[r|n, p, e_1, e_2] = [p(1 - e_1) + (1 - p)e_2]^r [pe_1 + (1 - p)(1 - e_2)]^{n-r} \quad (4)$$

A expressão (4) é exatamente a função de verossimilhança utilizada por Gaba e Winkler (1992) e Viana *et al.* (1993), indicando que a expressão (3) é uma generalização destes modelos obtida através da introdução de classificações repetidas.

3 Análise Bayesiana Empírica

Considere uma distribuição *a priori* conjunta de (p, e_1, e_2) dada por:

$$f(p, e_1, e_2) = f_\beta(p|\alpha, \beta) f_\beta(e_1|\alpha_1, \beta_1) f_\beta(e_2|\alpha_2, \beta_2) \quad (5)$$

em que $f_\beta(a | b, c)$ é função densidade de uma distribuição Beta para variável aleatória a com parâmetros b e c . Distribuições Beta são amplamente utilizadas em modelos Bayesianos para descrever informações sobre proporções (GUPTA; NADARAJAH, 2004).

Neste artigo, considera-se que as variáveis aleatórias (p, e_1, e_2) são mutuamente independentes *a priori*.

A densidade conjunta *a posteriori* de (p, e_1, e_2) é obtida multiplicando-se a distribuição *a priori* (5) pela verossimilhança (3) e normalizando como requerido pelo teorema de Bayes (WINKLER, 2003). Integrando-se em relação a e_1 e e_2 , obtém-se a função densidade marginal *a posteriori* de p , que pode ser expressa como:

$$f(p|r, n, m) = \sum_{j=0}^r \sum_{t=0}^{n-r} w_{jt}^* f_{\beta}(p|\alpha^*, \beta^*) \quad (6)$$

em que $w_{jt}^* = \frac{a_{jt}^*}{\sum_{j=0}^r \sum_{t=0}^{n-r} a_{jt}^*}$, com $a_{jt}^* = \binom{r}{j} \binom{n-r}{t} B(\alpha^*, \beta^*) k_1(j, t) k_2(j, t)$ e

$$k_1(j, t) = \int_0^1 e_1^{\alpha_1-1} (1 - e_1)^{\beta_1-1} [\text{Bin}(m; e_1; 0, 5m)]^{r-j} [1 - \text{Bin}(m; e_1; 0, 5m)]^{n-r-t} de_1;$$

$$k_2(j, t) = \int_0^1 e_2^{\alpha_2-1} (1 - e_2)^{\beta_2-1} [\text{Bin}(m; e_2; 0, 5m)]^t [1 - \text{Bin}(m; e_2; 0, 5m)]^j de_2;$$

e $B(\alpha^*, \beta^*)$ o valor da função Beta calculada no ponto (α^*, β^*) com $\alpha^* = \alpha + n - j - t$ e $\beta^* = \beta + j + t$.

A ausência de informações suficientes para definir distribuições *a priori* informativas para os erros de classificação implica, por exemplo, na utilização de distribuições $U(0, 1)$, caso particular da distribuição Beta, para os parâmetros (p, e_1, e_2) . Conseqüentemente, a distribuição marginal *a posteriori* para p pode ser multi-modal e/ou apresentar grande variabilidade. A realização de classificações repetidas não é garantia de minimização deste problema, e a situação é ainda mais grave quando apenas uma classificação é realizada, gerando uma média *a posteriori* de p igual a 0,5 independentemente do resultado amostral (GABA; WINKLER, 1992).

Assim, a distribuição *a posteriori* obtida pode ser de pouca utilidade para gerar informações necessárias sobre a proporção de interesse, ficando evidente a necessidade de obtenção de uma informação adicional sobre os erros de classificação.

Uma alternativa para minimizar esse problema é utilizar os resultados das classificações repetidas ($m > 1$) para estimar os hiperparâmetros (α_1, β_1) e (α_2, β_2) da distribuição *a priori* Beta dos erros de classificação e continuar utilizando a distribuição $U(0, 1)$ para p . Este procedimento pode ser descrito como um processo de estimação Bayes empírico paramétrico como apresentado por Carlin e Louis (1996), Gupta e Nadarajah (2004), Morris (1983) e Gelman (2004).

A estimação dos hiperparâmetros (α_1, β_1) e (α_2, β_2) foi realizada pelo método dos momentos. Primeiramente, a amostra aleatória de tamanho n foi dividida em duas sub-amostras: uma constituída pelos itens com classificação final conforme ($F_i = 1$) e a outra com itens com classificação final não-conforme ($F_i = 0$). Para cada unidade da primeira sub-amostra foi calculada a proporção de classificações repetidas não-conformes, sendo que a média e a variância destas proporções estimam, respectivamente, a média e a variância da distribuição *a priori* Beta de e_1 . Na segunda sub-amostra, calculou-se a proporção de classificações conforme para cada unidade. A média e a variância destas proporções estimam, respectivamente, a média e a variância da distribuição *a priori* Beta de e_2 . Finalmente, através das formas fechadas da média e da variância da distribuição Beta, foi possível estimar (α_1, β_1) e (α_2, β_2) resolvendo sistemas de duas equações e duas incógnitas. As estimativas para (α_1, β_1) e (α_2, β_2) podem ser expressas, respectivamente, por:

$$\hat{\alpha}_1 = k_3(k_4^2 + k_3^2 - k_3)/k_4^2 \quad (7)$$

$$\hat{\beta}_1 = (k_4^2 + k_3^2 - k_3)(k_3 - 1)/k_4^2 \quad (8)$$

$$\hat{\alpha}_2 = k_5(k_6^2 + k_5^2 - k_5)/k_6^2 \quad (9)$$

$$\hat{\beta}_1 = (k_6^2 + k_5^2 - k_5)(k_5 - 1)/k_6^2 \quad (10)$$

em que

$$k_3 = \sum_{i=1}^n \sum_{j=1}^m \frac{(1 - C_{ij}) I_{\{F_i=1\}}}{m \sum_{s=1}^n I_{\{F_s=1\}}}; \quad k_4 = k_3 (1 - k_3) \sum_{i=1}^n I_{\{F_i=1\}};$$

$$k_5 = \sum_{i=1}^n \sum_{j=1}^m \frac{(1 - C_{ij}) I_{\{F_i=0\}}}{m \sum_{s=1}^n I_{\{F_s=0\}}}; \quad k_6 = k_5 (1 - k_5) \sum_{i=1}^n I_{\{F_i=0\}}.$$

A utilização da distribuição *a priori* empírica para $m = 1$ não é viável através do método proposto, haja visto a impossibilidade de estimar (α_1, β_1) e (α_2, β_2) através das proporções de classificações equivocadas. Quando em uma sub-amostra todas as classificações repetidas gerarem resultados idênticos, será necessário aumentar n ou m de tal forma a captar o efeito dos erros de classificação e tornar possível estimar (α_1, β_1) e (α_2, β_2) pelo método dos momentos.

4 Exemplo numérico e discussões

A avaliação do desempenho numérico da metodologia proposta neste artigo foi realizada através de uma simulação considerando todas as combinações decorrentes dos seguintes valores de parâmetros: $p = 0,55; 0,75$ ou $0,9$; $e_1 = 0,05$ ou $0,15$; $e_2 = 0,05$ ou $0,15$; $n = 250$ ou 500 . Além disso, foram utilizadas distribuições *a priori* empírica e $U(0, 1)$ para os erros, considerando também a possibilidade de até sete classificações repetidas. A distribuição *a priori* de p foi $U(0, 1)$ em todos os casos simulados. Foi desenvolvido um programa no software Matlab de tal forma a calcular (6) e gerar graficamente a distribuição *a posteriori*

de p com respectiva média, moda, mediana e intervalo de credibilidade. O programa e os resultados de todas as simulações podem ser obtidos em www.est.ufmg.br/~roberto. As Figuras de 1 a 7 foram simuladas com os parâmetros $p = 0,55$; $e_1 = 0,15$; $e_2 = 0,15$; $n = 500$, correspondendo às sete classificações repetidas e são representativas dos resultados obtidos em todo o processo de simulação.

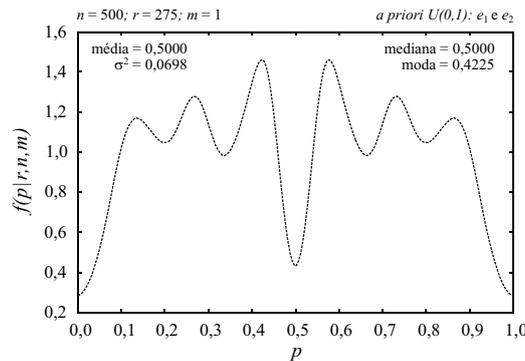


Figura 1: Distribuição *a posteriori* de p com $n=500$, $m=1$ e distribuição *a priori* $U(0,1)$ para e_1 e e_2 .

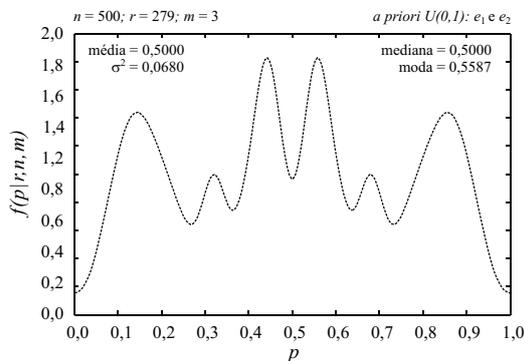


Figura 2: Distribuição *a posteriori* de p com $n=500$, $m=3$ e distribuição *a priori* $U(0,1)$ para e_1 e e_2 .

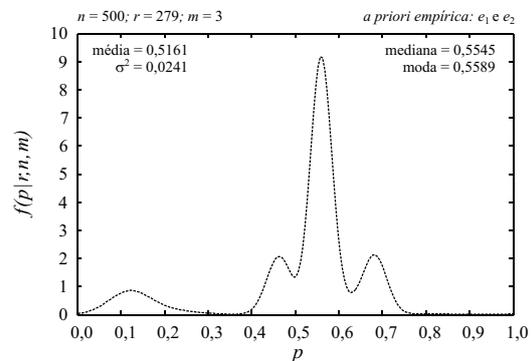


Figura 3: Distribuição *a posteriori* de p com $n=500$, $m=3$ e distribuição *a priori* empírica para e_1 e e_2 .

De maneira geral, o aumento das classificações repetidas reduziu o número de pontos extremos na função densidade *a posteriori* de p quando utiliza-se tanto a distribuição *a priori* $U(0,1)$ (Figuras 1, 2, 4, 6) quanto a distribuição *a priori* empírica (Figuras 3, 5, 7). Observa-se também que, considerando o mesmo número de classificações repetidas, as distribuições *a posteriori* de p obtidas através de *priori* empírica para os erros apresentam menor variabilidade e um número de pontos extremos menor ou igual ao das distribuições

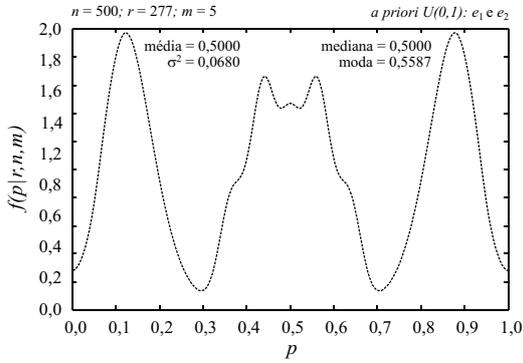


Figura 4: Distribuição *a posteriori* de p com $n=500, m=5$ e distribuição *a priori* $U(0, 1)$ para e_1 e e_2 .

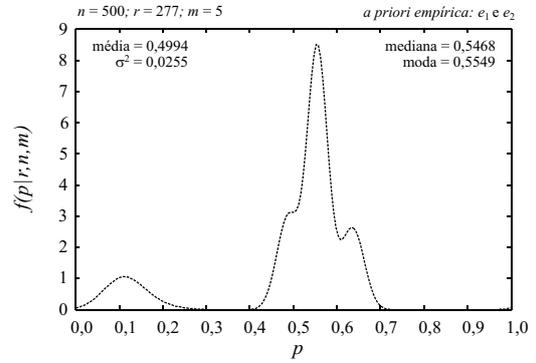


Figura 5: Distribuição *a posteriori* de p com $n=500, m=5$ e distribuição *a priori* empírica para e_1 e e_2 .

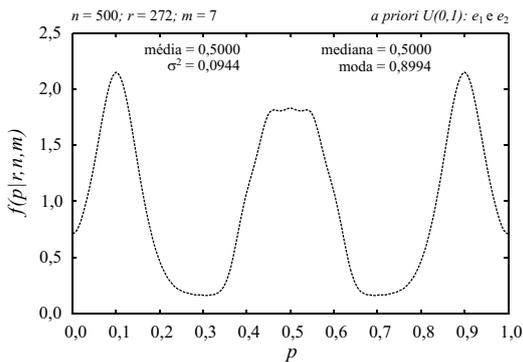


Figura 6: Distribuição *a posteriori* de p com $n=500, m=7$ e distribuição *a priori* $U(0, 1)$ para e_1 e e_2 .

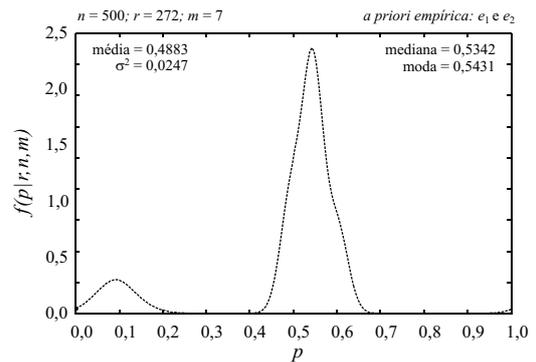


Figura 7: Distribuição *a posteriori* de p com $n=500, m=7$ e distribuição *a priori* empírica para e_1 e e_2 .

obtidas com distribuição *a priori* $U(0, 1)$.

As Tabelas 2 e 3 apresentam o vício médio (em percentual relativo ao valor real do parâmetro) para as combinações de erros simuladas e amostras de 500 e 250 elementos respectivamente. A utilização da distribuição *a priori* empírica demonstra um melhor desempenho, pois gera vícios, em módulo, menores do que a utilização da distribuição *a priori* uniforme. A média *a posteriori* apresenta os maiores vícios médios, não sendo esta, portanto, uma boa escolha para estimar p . Já a mediana e a moda, quando $n=500$ e 250 , apresentam valores de vícios médios absolutos inferiores a 5% para 3, 5 ou 7 classificações repetidas, tendo a moda um desempenho ligeiramente melhor.

As Tabelas 4 e 5 apresentam os vícios máximos obtidos para as combinações de erros simulados nas amostras de 500 e 250 elementos. Considerando 3, 5 ou 7 classificações repetidas, observa-se que a mediana apresenta um melhor desempenho, com valores de vícios máximos (em módulo) inferiores a 7,2%, enquanto a moda apresenta valores inferiores a 9,0%.

Quando a distribuição *a priori* empírica ou $U(0, 1)$ é utilizada para os erros, percebe-se freqüentemente a ocorrência de vícios negativos, ou seja, a proporção p de interesse está sendo subestimada. Isso pode ocorrer devido ao critério de decisão para classificação final em conforme ($F_i = 1$) ou não-conforme ($F_i = 0$). Como nos exemplos simulados $p > 0,5$ então em média a quantidade de itens realmente conformes é maior implicando que o número de ocorrências em que classifica-se um item como não-conforme quando é conforme é maior do que os casos onde classifica-se um item como conforme quando é na verdade não-conforme. Conseqüentemente a proporção tende a ficar sub-estimada. Caso $p < 0,5$ existirá uma tendência de super-estimação da proporção.

Tabela 2: Vício médio de estimativas *a posteriori* de p com $n=500$

m	p	Distribuição <i>a priori</i> Empírica			Distribuição <i>a priori</i> Uniforme		
		Média	Mediana	Moda	Média	Mediana	Moda
1	0,55	-	-	-	-9,10%	-9,10%	12,70%
	0,75	-	-	-	-33,30%	-33,30%	19,20%
	0,90	-	-	-	-44,40%	-44,40%	-71,00%
3	0,55	-5,20%	0,00%	0,80%	-9,10%	-9,10%	18,20%
	0,75	-2,80%	-1,80%	-1,20%	-33,30%	-33,30%	-61,10%
	0,90	-9,80%	-3,70%	-2,50%	-44,40%	-44,50%	-60,60%
5	0,55	-6,80%	-1,00%	3,90%	-9,10%	-9,10%	-62,00%
	0,75	-1,80%	-1,10%	-0,80%	-33,30%	-33,30%	-65,90%
	0,90	-7,40%	-1,40%	-0,70%	-44,40%	-44,50%	-72,50%
7	0,55	-5,70%	-1,00%	-0,30%	-9,10%	-9,10%	-79,90%
	0,75	-2,00%	-0,80%	-1,20%	-33,30%	-33,30%	-42,10%
	0,90	-2,30%	-0,40%	-0,40%	-44,50%	-44,40%	-44,40%

Tabela 3: Vício médio de estimativas *a posteriori* de p com $n=250$

m	p	Distribuição <i>a priori</i> Empírica			Distribuição <i>a priori</i> Uniforme		
		Média	Mediana	Moda	Média	Mediana	Moda
1	0,55	-	-	-	-9,10%	-9,10%	14,30%
	0,75	-	-	-	-33,30%	-33,30%	8,10%
	0,90	-	-	-	-44,40%	-44,50%	1,00%
3	0,55	-4,50%	-0,30%	0,70%	-9,10%	-9,10%	25,60%
	0,75	-3,50%	-2,50%	-2,00%	-33,30%	-33,30%	-36,80%
	0,90	-10,10%	-4,70%	-3,60%	-44,40%	-44,50%	-94,20%
5	0,55	-7,20%	-2,20%	-1,00%	-9,10%	-9,10%	8,20%
	0,75	-2,10%	-1,20%	-1,40%	-33,30%	-33,30%	-19,30%
	0,90	-7,60%	-2,80%	-2,20%	-44,40%	-44,50%	-69,50%
7	0,55	-5,10%	-0,90%	-0,40%	-9,10%	-9,10%	-5,60%
	0,75	-2,80%	-1,30%	-1,70%	-33,30%	-33,30%	-33,30%
	0,90	-2,60%	-1,30%	-1,40%	-44,50%	-44,40%	9,10%

Tabela 4: Vício máximo de estimativas *a posteriori* de p com $n=500$

m	p	Distribuição <i>a priori</i> Empírica			Distribuição <i>a priori</i> Uniforme		
		Média	Mediana	Moda	Média	Mediana	Moda
1	0,55	-	-	-	-9,10%	-9,10%	81,80%
	0,75	-	-	-	-33,30%	-33,30%	33,30%
	0,90	-	-	-	-44,40%	-44,50%	-93,00%
3	0,55	-6,50%	4,20%	6,00%	-9,10%	-9,10%	81,80%
	0,75	-6,40%	-4,30%	-4,30%	-33,30%	-33,30%	-91,00%
	0,90	-15,40%	-7,20%	-5,70%	-44,40%	-44,50%	-82,30%
5	0,55	-12,90%	-2,10%	8,60%	-9,10%	-9,10%	-79,30%
	0,75	-5,10%	-3,10%	-2,50%	-33,30%	-33,30%	-77,00%
	0,90	-15,10%	-3,60%	-2,20%	-44,40%	-44,50%	-100,00%
7	0,55	-11,20%	-2,90%	-1,50%	-9,10%	-9,10%	-82,30%
	0,75	-2,90%	-1,60%	-2,70%	-33,30%	-33,30%	-76,00%
	0,90	-3,50%	-1,20%	-1,40%	-44,50%	-44,50%	-100,00%

Tabela 5: Vício máximo de estimativas *a posteriori* de p com $n=250$

m	p	Distribuição <i>a priori</i> Empírica			Distribuição <i>a priori</i> Uniforme		
		Média	Mediana	Moda	Média	Mediana	Moda
1	0,55	-	-	-	-9,10%	-9,10%	81,80%
	0,75	-	-	-	-33,30%	-33,30%	54,20%
	0,90	-	-	-	-44,40%	-44,50%	11,50%
3	0,55	-9,70%	-6,50%	8,90%	-9,10%	-9,10%	81,80%
	0,75	-7,20%	-6,10%	-6,20%	-33,30%	-33,30%	-91,60%
	0,90	-14,90%	-8,10%	-6,60%	-44,40%	-44,50%	-100,00%
5	0,55	-14,70%	-4,20%	-2,70%	-9,10%	-9,10%	54,70%
	0,75	-5,10%	-3,00%	-3,50%	-33,30%	-33,30%	-65,80%
	0,90	-10,40%	-5,20%	-4,70%	-44,40%	-44,50%	-100,00%
7	0,55	-9,40%	-3,40%	-2,40%	-9,10%	-9,10%	-9,10%
	0,75	-4,50%	-3,00%	-3,90%	-33,30%	-33,30%	-33,30%
	0,90	-3,30%	-2,20%	-2,80%	-44,50%	-44,40%	10,20%

Observa-se também que o vício é assintoticamente nulo com o crescimento das classificações repetidas. A Figura 9 ilustra esta situação através da determinação da distribuição *a posteriori* de p com as respectivas média, mediana e moda obtidas de uma simulação em que $n = 500$, $p = 0,75$, $m = 99$ e distribuição *a priori* empírica para os erros. Neste cenário, o vício máximo está em torno de $-0,1\%$ decorrente da média *a posteriori*. O gráfico também indica que quando m cresce a moda *a posteriori* tende a apresentar menores vícios.

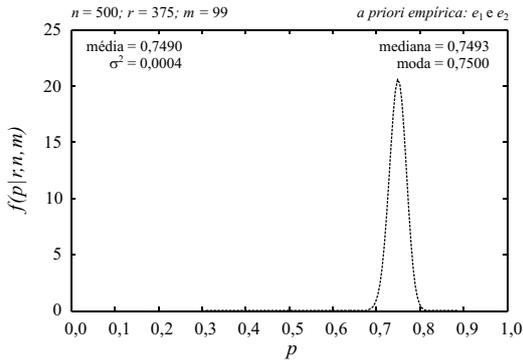


Figura 8: Distribuição *a posteriori* de p com $n=500$, $m=99$ e distribuição *a priori* empírica para e_1 e e_2 .

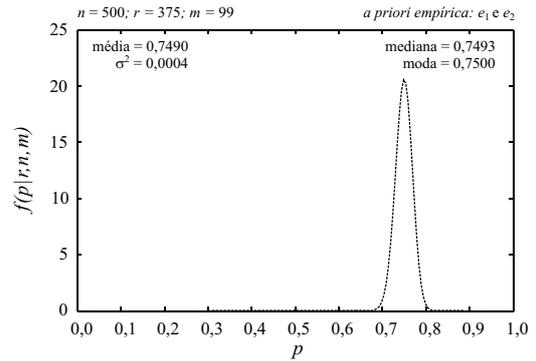


Figura 9: Distribuição *a posteriori* de p com $n=500$, $m=99$ e distribuição *a priori* empírica para e_1 e e_2 .

5 Conclusão

Este trabalho apresenta uma metodologia bayesiana empírica para estimar uma proporção quando as avaliações estão sujeitas a erros de classificação e informações *a priori* sobre tais erros não estão disponíveis. A proposta é realizar classificações repetidas e, através destas, elicitare distribuições *a priori* empíricas para os erros de classificação.

Um estudo de simulação demonstrou que a metodologia apresenta desempenho satisfatório, pois a utilização da distribuição *a priori* empírica, quando comparada com a distribuição *a priori* $U(0, 1)$, gera estimativas *a posteriori* com vícios absolutos menores e distribuições *a posteriori* com menor variabilidade.

Como estimativa *a posteriori* da proporção de interesse, considerando-se uma postura conservadora do pior caso, recomenda-se a mediana como melhor alternativa. Decidindo-

se por uma abordagem de vício médio, a melhor estimativa é fornecida pela moda. Em ambos os casos, recomenda-se pelo menos três classificações repetidas de tal forma a garantir um número menor de pontos extremos e um vício que não comprometa o processo de decisão.

Referências

- BROSS, I. Misclassification in 2×2 tables. *Biometrics*, v. 10, p. 478–486, 1954.
- CARLIN, B. P.; LOUIS, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall, 1996.
- DENDUKURI, N. *et al.* Bayesian sample size determination for prevalence and diagnostic test studies in the absense of a gold standard test. *Biometrics*, v. 60, p. 388–397, 2004.
- EVANS, M. *et al.* Bayesian analysis of binary data subject to misclassification. In: BERRY, D.; CHALONER, K.; GEWEKE, J. (Ed.). *Bayesian Analysis In Statistics and Econometrics: Essays In Honor Of Arnold Zellner*. New York: North Holland, 1996. p. 66–77.
- GABA, A.; WINKLER, R. L. Implications of errors in survey data: A bayesian model. *Management Science*, v. 38, n. 7, p. 913–925, 1992.
- GELMAN, A. *Bayesian Data Analysis*. 2. ed. London: Chapman & Hall, 2004.
- GUPTA, A. K.; NADARAJAH, S. *Handbook of Beta Distribution and Its Applications*. New York: Marcel Dekker, 2004.
- GUSTAFSON, P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. New York: Chapman & Hall, 2003.
- JOHNSON, N. L.; KOTZ, S. Estimation from binomial data with classifiers of known and unknown imperfections. *Naval Research Logistics*, v. 35, p. 147–156, 1988.

JOHNSON, N. L.; KOTZ, S.; WU, X. *Inspection Errors for Attributes in Quality Control*. London: Chapman & Hall, 1991.

MORRIS, C. N. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, v. 78, p. 47–65, 1983.

RAHME, E.; JOSEPH, L.; GYORKOS, T. W. Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. *Applied Statistics*, v. 49, n. 1, p. 119–128, 2000.

VIANA, M. A. G. Bayesian small-sample estimation of misclassification multinomial data. *Biometrics*, v. 50, p. 237–243, 1994.

VIANA, M. A. G.; RAMAKRISHNAN, V.; LEVY, P. S. Bayesian analysis of prevalence from the results of small screening samples. *Communications in statistics - V theory and methods*, v. 22, n. 2, p. 575–85, 1993.

WINKLER, R. L. *Bayesian Inference and Decisions*. 2. ed. London: Probabilistic Publishing, 2003. 384 p.