Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Estimação Bayesiana da Conformidade de Produtos com Classificações Repetidas Imperfeitas

R. C. Quinino e E. Suyama

Relatório Técnico
RTP-10/2000

Relatório Técnico
Série Pesquisa

Estimação Bayesiana da Conformidade de Produtos com Classificações Repetidas Imperfeitas

Roberto da Costa Quinino
Departamento de Estatística – ICEX - UFMG
e-mail: roberto@est.ufmg.br

Emílio Suyama
Departamento de Estatística – ICEX - UFMG
e-mail: suyama@est.ufmg.br

Resumo

Neste trabalho estudamos o problema de estimar a proporção de itens conformes quando existe a possibilidade de realizar classificações independentes por unidade produzida. Estamos supondo que haja dois tipos de erro: o tipo I no qual classifica-se um produto como conforme quando ele na realidade não é e o tipo II no qual classifica-se um produto como não-conforme quando na realidade é conforme. Desenvolvemos um abordagem bayesiana para viabilizar a estimação.

1. Introdução

Um aspecto fundamental na avaliação da qualidade de produtos é a necessidade da estimação da proporção de produtos conformes. Esta estimação é freqüente após a recepção de matérias primas ou no controle de processos. Todo o procedimento é normalmente realizado supondo-se que o sistema de classificação é perfeito.

Burke, J. R. at alli (1995) argumentam que os erros estão longe de serem considerados desprezíveis em muitas tarefas de classificação. Mesmo sob condições ideais de classificação, técnicos altamente experientes podem cometer erros cujas proporções podem superar 25% [Jacobson, H. J. (1952)]. Tudo isto pode comprometer seriamente o processo de estimação da proporção de conformes e consequentemente o controle da qualidade.

Harris & Chaney (1969) e Johnson, Kotz & Wu (1991) apresentam numerosos estudos onde verificam que os erros de classificação podem comprometer seriamente o processo de inferência estatística e, consequentemente, o processo de controle estatístico da qualidade. Sendo assim, torna-se evidente que a avaliação da qualidade tem considerável dependência de como é realizada a estimação de produtos não-conformes quando a classificação não é perfeita.

Gaba&Winkler (1992) e Viana et al. (1993) apresentam um modelo bayesiano para estimar a proporção de interesse com a presença de erros de classificação para dados dicotômicos. O modelo proposto por estes autores não consideram a possibilidade de realizar classificações repetidas. A proposta deste trabalho é desenvolver um modelo geral e mais informativo através da incorporação de classificações repetidas. Assim sendo, os modelos anteriores tornam-se casos particulares menos informativos.

2. Modelo Probabilístico

Considere uma classificação de n produtos em conformes e não-conformes. Seja p a probabilidade de que um produto qualquer seja fabricado conforme, e_1 a probabilidade de que um produto conforme seja classificado como não-conforme em uma única classificação e e_2 a probabilidade de que um produto não-conforme seja classificado como conforme em uma única classificação. A figura 1 mostra uma árvore de eventos para facilitar a visualização destas probabilidades.

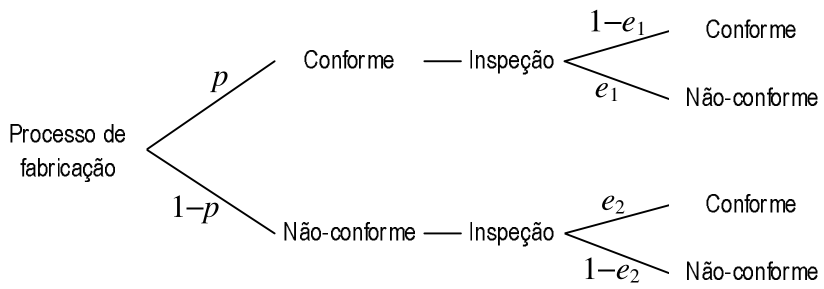

Figura 1: árvore de eventos das probabilidades consideradas

Suponha que cada um dos n produtos sejam classificados independentemente m vezes. Considere C_{ij} (i=1,...,n; j=1,...,m) como uma variável 0-1, correspondente ao ij-ésimo evento, isto é, a j-ésima classificação do i-ésimo produto. Observe a tabela 1 para uma melhor visualização.

Tabela 1: classificação repetida de produtos

Produto	Classificações (C_{ij})				
	1	2	3	...	m
1	C_{11}	C_{12}	C_{13}	...	C_{1m}
2	C_{21}	C_{22}	C_{23}	...	C_{2m}
3	C_{31}	C_{32}	C_{33}	...	C_{3m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	C_{n1}	C_{n2}	C_{n3}	...	C_{nm}

Neste sentido, $C_i = \sum_{j=1}^{m} C_{ij}$ corresponde ao número de vezes que o i-ésimo produto foi classificado conforme nas m classificações. Logo, o vetor aleatório $\boldsymbol{C} = (C_1, C_2, ..., C_n)$ é distribuído segundo uma mistura de Binomiais dada por:

$$P[\boldsymbol{C} = \boldsymbol{c}] = \prod_{i=1}^{n} \binom{m}{c_i} \left\{ p(1-e_1)^{c_i} e_1^{m-c_i} + (1-p) e_2^{c_i} (1-e_2)^{m-c_i} \right\} \qquad (2.1)$$

A expressão (2.1) pode ser rescrita como:

$$P[\boldsymbol{C}=\boldsymbol{c}]=\prod_{i=1}^{n}\binom{m}{c_i}\sum_{x=0}^{1}p^{x}(1-e_1)^{c_i x}e_1^{(m-c_i)x}(1-p)^{(1-x)}e_2^{c_i(1-x)}(1-e_2)^{(m-c_i)(1-x)}$$
$$=\sum_{i=1}^{n}\sum_{x_i=0}^{1}\binom{m}{c_1}\ldots\ldots\binom{m}{c_n}p^{y_1}(1-p)^{y_2}e_1^{y_3}(1-e_1)^{y_4}e_2^{y_5}(1-e_2)^{y_6}$$

(2.2)

onde

$$y_1=\sum_{i=1}^{n}x_i\ ;\ y_2=\sum_{i=1}^{n}(1-x_i)\,.$$

$$y_3=\sum_{i=1}^{n}(m-c_i)x_i\ ;\ y_4=\sum_{i=1}^{n}c_i x_i\ .$$

$$y_5=\sum_{i=1}^{n}c_i(1-x_i)\ ;\ y_6=\sum_{i=1}^{n}(m-c_i)(1-x_i)\,.$$

Pode-se observar que se $m=1$ a expressão (2.2) pode ser reescrita como:

$$P[\boldsymbol{C}=\boldsymbol{c}]=[p(1-e_1)+(1-p)e_2]^{r}\ [pe_1+(1-p)(1-e_2)]^{n-r} \quad (2.3)$$

onde $r=\sum_{i=1}^{n}C_i$.

A expressão (2.3) é idêntica a expressão utilizada por GABA; WINKLER (1992) e VIANA et al. (1993) (2.2) e como conseqüência a análise bayesiana é realizada conforme este autores. Trata-se do caso particular para o caso $m=1$.

Considere $m\geq 1$ e adote uma função densidade conjunta a a priori representada pela expressão:

$$f(p,e_1,e_2)=f_\beta(p/\alpha,\beta)f_\beta(e_1/\alpha_1,\beta_1)f_\beta(e_2/\alpha_2,\beta_2) \quad (2.4)$$

onde $f_\beta(\,p/\alpha,\beta\,)$, $f_\beta(\,e_1/\alpha_1,\beta_1\,)$ e $f_\beta(\,e_2/\alpha_2,\beta_2\,)$ denotam funções densidade Beta respectivamente de p , e_1 e e_2 de parâmetros $(\alpha,\beta\,)$, $(\alpha_1,\beta_1\,)$ e $(\alpha_2,\beta_2\,)$ respectivamente.

A função de verossimilhança é dada por:

$$L[\mathbf{c} \mid n,m,p,e_1,e_2] = \sum_{x_1=0}^{1} \cdots \sum_{x_n=0}^{1} \binom{m}{c_1} \cdots \binom{m}{c_n} p^{y_1}(1-p)^{y_2} e_1^{y_3}(1-e_1)^{y_4} e_2^{y_5}(1-e_2)^{y_6}$$

(2.5)

A função densidade conjunta a posteriori de p, e_1 , e_2 é

$$f(p,e_1,e_2 / \mathbf{c},n,m) = \sum_{x_1=0}^{1} \cdots\cdots \sum_{x_n=0}^{1} W_{x_1 \ldots x_n} f_\beta(p / \alpha^\bullet, \beta^\bullet) f_\beta(e_1 / \alpha_1^\bullet, \beta_1^\bullet) f_\beta(\alpha_2^\bullet, \beta_2^\bullet)$$

(2.6)

onde $W_{x_1 \ldots x_n} = \dfrac{a_{x_1 \ldots x_n}}{\sum_{x_1=0}^{1} \cdots \sum_{x_n=0}^{1} a_{x_1 \ldots x_n}}$,

com

$$a_{x_1 \ldots x_n} = \binom{m}{c_1} \cdots \binom{m}{c_n} B(\alpha^\bullet, \beta^\bullet) B(\alpha_1^\bullet, \beta_1^\bullet) B(\alpha_2^\bullet, \beta_2^\bullet),$$

$\alpha^{\bullet} = y_1 + \alpha;$

$\beta^{\bullet} = y_2 + \beta;$

$\alpha_1^{\bullet} = y_3 + \alpha_1;$

$\beta_1^{\bullet} = y_4 + \beta_1;$

$\alpha_2^{\bullet} = y_5 + \alpha_2;$

$\beta_2^{\bullet} = y_6 + \beta_2.$

As funções densidades marginais a posteriori de p, e_1 e e_2 são obtidas integrando a expressão (2.6), respectivamente a (e_1 e e_2), (p e e_2) e (p e e_1), cujos resultados são dados respectivamente pelas expressões (2.7), (2.8) e (2.9).

$$f(p/\boldsymbol{c}, n, m) = \sum_{x_1=0}^{1} \cdots\cdots \sum_{x_n=0}^{1} W_{x_1 \ldots x_n} f_{\beta}(p/\alpha^{\bullet}, \beta^{\bullet}) \tag{2.7}$$

$$f(e_1/\boldsymbol{c}, n, m) = \sum_{x_1=0}^{1} \cdots\cdots \sum_{x_n=0}^{1} W_{x_1 \ldots x_n} f_{\beta}(e_1/\alpha^{\bullet}, \beta^{\bullet}) \tag{2.8}$$

$$f(e_2/\boldsymbol{c}, n, m) = \sum_{x_1=0}^{1} \cdots\cdots \sum_{x_n=0}^{1} W_{x_1 \ldots x_n} f_{\beta}(e_2/\alpha^{\bullet}, \beta^{\bullet}) \tag{2.8}$$

As esperanças a posteriori de p, e_1 e e_2 são obtidas diretamente das expressões (2.7), (2.8) e (2.9), sendo dadas respectivamente por:

$$E(p/\boldsymbol{c},n,m)=\sum_{x_1=0}^{1}\ldots\ldots\sum_{x_n=0}^{1}W_{x_1\ldots x_n}\frac{\alpha^{\bullet}}{\alpha^{\bullet}+\beta^{\bullet}} \tag{2.9}$$

$$E(e_1/\boldsymbol{c},n,m)=\sum_{x_1=0}^{1}\ldots\ldots\sum_{x_n=0}^{1}W_{x_1\ldots x_n}\frac{\alpha_1^{\bullet}}{\alpha_1^{\bullet}+\beta_1^{\bullet}} \tag{2.10}$$

$$E(e_2/\boldsymbol{c},n,m)=\sum_{x_1=0}^{1}\ldots\ldots\sum_{x_n=0}^{1}W_{x_1\ldots x_n}\frac{\alpha_2^{\bullet}}{\alpha_2^{\bullet}+\beta_2^{\bullet}} \tag{2.11}$$

Observe que o aumento de *n* torna o número de cálculos necessários para a solução exata extremamente grande, exigindo o uso intensivo de recursos computacionais.

3. Modelo com caracterização final do produto

Suponha que cada uma das n peças sejam caracterizadas conformes ou não conformes após as *m* classificações. A tabela 2 ilustra o procedimento.

Tabela 2 : classificação dos *n* produtos

Produto	Classificações (C_{ij})					Classificação final (F_i)
	1	2	3	...	m	
1	C_{11}	C_{12}	C_{13}	...	C_{1m}	F_1
2	C_{21}	C_{22}	C_{23}	...	C_{2m}	F_2
3	C_{31}	C_{32}	C_{33}	...	C_{3m}	F_3
⋮	⋮	⋮	⋮		⋮	⋮
N	C_{n1}	C_{n2}	C_{n3}	...	C_{nm}	F_n

Neste ambiente, estaremos considerando que o i-ésimo produto será julgado conforme, F_i=1 , se e somente se $C_i = \sum_{j=1}^{m} C_{ij} > 0{,}5m\,(i = 1,...,n)$.

Os resultados tipo Bernoulli C_i são independentes e identicamente distribuídos. Definindo E_i (i=1,...,n) como uma variável 0-1, correspondente ao estado real de fabricação do i-ésimo produto e utilizando o teorema da probabilidade total podemos expressar Pr[C_i=1] da seguinte forma:

$$Pr[F_i = 1] = Pr[E_i = 1]\,Pr[F_i = 1/E_i = 1] + Pr[E_i = 0]\,Pr[F_i = 1/E_i = 0]$$
$$= p\left[\sum_{x=0}^{0{,}5m}\binom{m}{x} e_1^x (1-e_1)^{m-x}\right] + (1-p)\left[1 - \sum_{x=0}^{0{,}5m}\binom{m}{x} e_2^x (1-e_2)^{m-x}\right]$$

onde o termo Pr[a/b] indica a probabilidade condicional de a dado b.

Utilizando a notação $B[a;b;c]$ para uma função distribuição Binomial, com parâmetros a e b, calculada no ponto c, temos

$$Pr[F_i = 1] = pB[m;e_1;0{,}5m] + (1-p)(1 - B[m;e_2;0{,}5m]) \tag{3.1}$$

Para situações reais é razoável assumir que $e_1 < 0{,}5$ e $e_2 < 0{,}5$. Assim sendo, o crescimento de m decresce a probabilidade de julgamentos equivocados sobre a conformidade ou não dos produtos. Neste sentido, com o aumento de m, o número de produtos considerados conformes converge para a proporção de produtos fabricados conformes, como podemos observar na proposição 1.

Proposição 1. Com o aumento de m, o número de produtos classificados conformes converge para a proporção de produtos fabricados conformes. Matematicamente, a expressão (2.2) converge para p, quando $m \to \infty$, $e_1,e_2<0{,}5$.

Demonstração. A demonstração pode ser realizada através da aproximação da distribuição Binomial pela distribuição Normal. Assim, temos:

$$\lim_{m\to+\infty} \Pr[C_i = 1] = p\Phi\left[\frac{\sqrt{m}(0,5-e_1)}{\sqrt{e_1(1-e_1)}}\right] + (1-p)\left(1-\Phi\left[\frac{\sqrt{m}(0,5-e_2)}{\sqrt{e_2(1-e_2)}}\right]\right)$$
$$= p\Phi[+\infty] + (1-p)(1-\Phi[+\infty]) = p$$

onde $\Phi[\cdot]$ indica função distribuição Normal padrão.

Na amostra aleatória de n elementos, supondo que r elementos são registrados como conforme e os restantes $(n - r)$ como não-conformes, a função de verossimilhança será:

$$L(r \mid n,m,p,e_1,e_2) =$$

$$\{pB[m;e_1;0,5m] + (1-p)(1-B[m;e_2;0,5m])\}^r \{p(1-B[m;e_1;0,5m]) + (1-p)B[m;e_2;0,5m]\}^{n-r}$$
(3.2)

Em situações reais os valores de e_1 e e_2 são desconhecidos. Em conseqüência disso a função de verossimilhança é função de três parâmetros: p, e_1 , e_2. Neste caso o estimador de máxima verossimilhança não é único, ou seja, a função de verossimilhança é maximizada para todos os pontos $(p\ ,\ e_1\ ,\ e_2)$ tal que $p(1-e_1)+(1-p)e_2 = r/n$.

Toda incerteza introduzida pelos erros de classificação pode ser vista mais claramente reescrevendo a expressão (2.2), utilizando o binômio de Newton:

$$L(r / n, p, e_1, e_2) = \sum_{j=0}^{r}\sum_{t=0}^{n-r} \binom{r}{j}\binom{n-r}{t} p^{n-j-t}(1-p)^{j+t} e_1^{n-r-t}(1-e_1)^{r-j} e_2^{j}(1-e_2)^{t}$$
(3.3)

Considerando a distribuição a priori conjunta de p , e_1 e e_2 expressa em (2.4) temos que a função densidade conjunta a posteriori de e_1, e_2 e p é proporcional ao produto da função de verossimilhança pela função densidade conjunta a priori que pode ser expressa como:

$$f(p,e_1,e_2/r,n)=\sum_{j=0}^{r}\sum_{t=0}^{n-r}w_{jt}f(p,e_1,e_2/r,n,j,t),$$

(3.4)

A função densidade marginal a posteriori de p, considerando a função densidade conjunta a priori representada pela expressão (2.5) pode ser expressa como:

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

(3.5)

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

com $a_{jt}^{*}=\binom{r}{j}\binom{n-r}{t}B(\alpha^{\circ},\beta^{\circ})K_1(j,t)K_2(j,t)$,

e

$\alpha^{\circ}=\alpha+n-j-t,$

$\beta^{\circ}=\beta+j+t,$

$$K_1(j,t)=\frac{1}{\int_0^1 e_1^{\alpha_1-1}(1-e_1)^{\beta_1-1}\left(B[m;e_1,0.5m]\right)^{r-j}\left(1-B[m;e_1,0.5m]\right)^{n-r-t}de_1}$$

$$K_2(j,t)=\frac{1}{\int_0^1 e_2^{\alpha_2-1}(1-e_2)^{\beta_2-1}\left(B[m;e_2,0.5m]\right)^{j}\left(1-B[m;e_2,0.5m]\right)^{t}de_2}$$

A esperança a posteriori de p pode ser expressa como

$$E(p/r,n,m)=\sum_{j=0}^{r}\sum_{t=0}^{n-r}w_{jt}^{*}\frac{\alpha^{\circ}}{\alpha^{\circ}+\beta^{\circ}},$$

(3.6)

A expressão (3.6) converge para a função densidade a posteriori de p quando não existe erros de classificação, $m\to\infty$, $e_1,e_2<0{,}5$. O modelo não é recomendável para estimação de e_1 e e_2 quando m cresce. Isto se deve ao fato que com o aumento de m a distribuição a posteriori de e_1 e e_2 converge para distribuição a priori de e_1 e e_2. Em outras palavras, o experimento não é informativo para atualização da distribuição a priori de e_1 e e_2 uma vez que o aumento de m minimiza a influência dos erros de classificação. Como forma de contornar este problema, propomos as seguintes expressões assintóticas:

$$e_1 \cong 1 - \frac{\sum_{i=1}^{n} C_i F_i}{\sum_{i=1}^{n} m F_i} \doteq 1 - \frac{\sum_{i=1}^{n} C_i F_i}{nmp}$$

$$e_2 \cong \frac{\sum_{i=1}^{n} C_i (1 - F_i)}{nm(1-p)} \doteq \frac{\sum_{i=1}^{n} C_i (1 - F_i)}{nm(1-p)} \tag{3.7}$$

Como forma de operacionalizar estas expressões podemos utilizar weighted Likelihood Bootstrap [Newton; Raftery (1994)] para geração de uma amostra a posteriori de p e através das expressões dadas em (3.7) gerar estimativas de e_1 e e_2.

4. Conclusão e Considerações Finais

Em dados dicotômicos, os erros de classificação podem causar um significativo impacto nas inferências e consequentemente gerar estimativas enganosas. A consideração formal dos erros de classificação através do procedimento tradicional da maximização da função de verossimilhança não é eficaz, uma vez que ela é maximizada para um conjunto infinito de pontos abrangendo, inclusive, resultados totalmente antagônicos.

O procedimento bayesiano, utilizado neste trabalho, permitiu considerar o impacto dos erros de uma maneira formal. Os julgamentos prévios, incorporados em uma distribuição a priori, proporcionou a informação necessária para evitar problemas de identificação, como o ocorrido na maximização da função de verossimilhança. Consideremos também a possibilidade de classificações repetidas por peça avaliada. Isto adicionou informação ao processo de decisão minimizando o impacto dos erros de classificação. De acordo com este trabalho o aumento de m melhora a estimativa de p quando existe erros de classificação.

Referências Bibliográficas:

1 - **Albert, J.H.:** "Teaching bayesian statistics using sampling methods and minitab". The American Statistician, 47 , p.182-91, 1993.

2 - **Berger, J. O.:** "Statistical decision theory and bayesian analysis". New York, Springer-Verlag, 1985.

3 - **Burke, J. R. at alli.:** "The effect of inspector errors on the true fraction non-conforming: an industrial experiment". Quality Engineering, 7(3), 543-550, 1995.

4 - **Case, K. E.: "**The effect of inspector errors on average outgoing quality". Journal of Quality Techonology, 7(1) , 28-33, 1975.

5 - **Fishburn, P. C.:** "The axioms of subjective probability". Statistical Science, 1(3), p.335-58, 1986.

6 - **Gaba, A. & Winkler, R. L.:** "Implications of errors in survey data: a bayesian model". Management Science, 38(7), p.913-25, 1992.

7 - **Gelfand, A. E. & Smith, A. F. M.:** "Sampling-based approaches to calculating marginal densities". Journal of The American Statistical Association", 85, p.398-409, 1990.

8 - **Gelfand, A. E. & Smith, A. F. M.:** "Bayesian statistics without tears: a sampling-resampling perspective". The American Statistician, 46, 84-8, 1992.

9 - **Gokhale, D. & Press, S. J.: "**Assessment a prior distribution for the correlation coefficient in a bivariate normal distribution". Journal of Royal Statistical Society. Series A. 145, 237-49, 1982.

10 - **Kotz, S.; Johnson, N. L.;Kemp, A. W.:** "Univariate discrete distributions". New York. 2nd ed., John Wiley & Sons, Inc., 1992.

11 - **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.

12 - **Rubin, D. B.:** “Using the SIR algorithm to simulate posterior distributions”. In Bayesian Statistics 3, Eds. J. M. Bernado, M. H. De Groot, D.V. Lindley, and A. F. M. Smith, New York: Oxford University Press, 1988.

13 - **Shachter, R.:** “Evaluating influence diagrams”. Operations Research, 34, p.871-82, 1986.

14 - **Spetzler, C. A. & Staël Von Holstein, C. -A. S.:** “Probability encoding in decision analysis”. Management Science, 22 , p.340-58, 1975.

15 - **Tierney, L. & Kardane, J. A.:** “Accurate approximations for posterior moments and marginal densities”. Journal of the American Statistical Association, 81 , p.82-6, 1986.

16 - **Winkler, R. L.:** “The assessment of prior distributions in bayesian analysis”. Journal of the American Statistical Association , 74 , p.776-800, 1967.