DOI: 10.4025/actascitechnol.v32i3.4145

Estudo por simulação Monte Carlo de um estimador robusto utilizado na inferência de um modelo binomial contaminado

Augusto Maciel da Silva[1*] e Marcelo Angelo Cirillo[2]

*[1]Programa de Pós-graduação em Estatística e Experimentação Agropecuária, Departamento de Ciências Exatas, Universidade Federal de Lavras, Cx. Postal 37, 37200-000, Lavras, Minas Gerais, Brasil. [2]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil. *Autor para correspondência. E-mail: augustolavras@gmail.com*

RESUMO. A inferência estatística em populações binomiais contaminadas está sujeita a erros grosseiros de estimação, uma vez que as amostras não são identicamente distribuídas. Por esse problema, este trabalho tem por objetivo determinar qual a melhor constante de afinidade (c_1) que proporcione melhor desempenho em um estimador pertencente à classe dos estimadores-E. Com esse propósito, neste trabalho, foi utilizada a metodologia, considerando-se o método de simulação Monte Carlo, no qual diferentes configurações descritas pela combinação de valores paramétricos, níveis de contaminação e tamanhos de amostra foram avaliados. Concluiu-se que, para alta probabilidade de mistura ($\gamma = 0{,}40$), recomenda-se assumir $c_1 = 0{,}1$ nas situações de grandes amostras ($n = 50$ e $n = 80$).

Palavras-chave: distribuição binomial, binomiais contaminadas, Monte Carlo, robustez.

ABSTRACT. A Monte Carlo simulation study of a robust estimator used in the inference of a contaminated binomial model. The statistical inference in binomial population is subject to gross errors of estimate, as the samples are not identically distributed. Due to this problem, this work aims to determine which is the best affinity constant (c_1) that provides the best performance in the estimator, belonging to the class of E-estimators. With that purpose, the methodology used in this work was applied considering the Monte Carlo simulation method, in which different configurations described by combination of parametric values, levels of contamination and sample sizes were appraised. It was concluded that for the high probability of contamination ($\gamma = 0.40$), $c_1 = 0.1$ is recommended in cases with large samples ($n = 50$ and $n = 80$).

Key words: binomial distribution, contaminated binomial, Monte Carlo, robustness.

Introdução

Em pesquisas científicas sempre há necessidade de se fazer inferência sobre o parâmetro da proporção de populações binomiais. Essas inferências são realizadas sobre a suposição de que os dados amostrais sejam independentes e provenientes de uma mesma população. Em se tratando de amostras contaminadas, geradas pela mistura de populações binomiais, essa suposição não é satisfeita e, portanto, os estimadores usuais não podem ser utilizados, o que necessariamente faz requerer a pesquisa de outros métodos inferenciais, mais especificamente os métodos robustos de estimação.

Outra situação em que se enquadra a estimação robusta se dá pelo excesso de observações. Copas (1988) advertiu que, mesmo se assumindo o modelo adequado, certo número de observações poderá vir a ser detectado como *outlier* pelo fato de que o valor da proporção π encontra-se próximo a 0 ou 1. O autor ressalta também que o método de estimação de máxima verossimilhança é sensível à presença de *outliers* e sugere correção nas estimativas, visto que diferentes modelos utilizados para resposta apresentam sensibilidade variada na detecção de *outliers*.

Victoria-Feser e Ronchetti (1997), em um estudo com dados agrupados, propuseram uma classe de estimadores MGP otimizados, que são estimadores de máxima verossimilhança generalizados e comprovaram que, sobre determinada função de influência e sobre contaminação, esses estimadores robustos são mais estáveis do que os estimadores clássicos para dados agrupados.

Wood et al. (2005) também propuseram duas alternativas para se estimar a probabilidade de sucesso em amostras binomiais contaminadas, nas quais cada unidade amostral foi determinada por meio de uma contagem das observações classificadas como sucesso, considerando-se diferentes números de ensaios (n). Essas alternativas referiram-se a dois estimadores, sendo estes diferenciados pela média aritmética e média ponderada das proporções

observadas. Comparando as variâncias dos estimadores, os autores concluíram que a recomendação de um estimador far-se-á em diferentes situações, caracterizadas pela distribuição das proporções e pelo número de ensaios (n) realizados.

A classe dos estimadores-E, como é citado por Lindsay (1994), é construída a partir de uma modificação na função de verossimilhança com a finalidade de reduzir o efeito das observações *outliers* na 'cauda' da distribuição. Assim sendo, sugere-se realizar uma estimação, considerando-se uma escala de frequência e não as observações individuais. A forma mais natural para realização desse procedimento é considerar as frequências relativas. Por meio da função de probabilidade assumida no modelo, este conduz à distância mínima da estimativa sobre a frequência relativa, por isso esse estimador é conhecido também como estimador de mínima disparidade (SIMPSON, 1987). A principal característica desses estimadores é verificada na facilidade com que o pesquisador tem em manipular os 'erros grosseiros' a serem cometidos na estimação.

Para exemplificar algumas situações, Ruckstuhl e Welsh (2001) afirmam que o desempenho dos estimadores que pertencem a essa classe depende de determinados valores assumidos pelas constantes de afinidade, referenciadas por c_1 e c_2. Assim, os autores descreveram que, quando $c_1 = 0$ e $c_2 \rightarrow \infty$, o estimador pertencente à classe E apresenta mínima entropia relativa. Em particular, assintoticamente recomendam $c_1 < c_2 = 1$, porém nada restringem aos valores assumidos para c_1. É importante ainda ressaltar que, no presente estudo, o estimador será considerado robusto quando apresentar estimativas próximas ao parâmetro π_1. Isso posto, a motivação para a realização deste trabalho se deu com o objetivo de se avaliar, por meio do método Monte Carlo, o comportamento de um estimador proposto e utilizado para se estimar a probabilidade de sucesso de uma população binomial contaminada mediante os diferentes valores assumidos para c_1.

Material e métodos

A metodologia proposta neste trabalho consistiu em um estudo de simulação referente a um estimador para proporção de sucessos (π) de um modelo binomial formado pela mistura de duas binomiais.

Com esse propósito, considerou-se, no processo de simulação, uma variável aleatória Y_i (i = 1,2). Com isso, $Y_i = k / m, \pi_1, \pi_2, \gamma \sim \gamma$ Binomial (m, π_1) + (1-γ) Binomial (m, π_2) e sua função de distribuição foram definidos por:

$$f(k/m, \pi_1, \pi_2, \gamma) = (1-\gamma)\binom{m}{k}\pi_1^{\,k}(1-\pi_1)^{m-k} + \gamma\binom{m}{k}\pi_2^{\,k}(1-\pi_2)^{m-k} \quad (1)$$

em que: 1-γ correspondeu à probabilidade de k ter sido gerada da primeira distribuição Binomial; com probabilidade de sucesso π_1 e γ, referiu-se à probabilidade de k ter sido gerada da segunda distribuição Binomial com probabilidade de sucesso π_2. Dado esse modelo, no processo de simulação, considerou-se uma variável latente (u) distribuída por uma uniforme (0,1). Dessa forma, a unidade amostral k foi simulada por

$$k = \begin{cases} (1-\gamma)\binom{m}{k}\pi_1^{k}(1-\pi_1)^{m-k} & u \le \gamma \\ \gamma\binom{m}{k}\pi_2^{k}(1-\pi_2)^{m-k} & u > \gamma \end{cases} \quad (2)$$

Assim, a população de referência, cujas unidades amostrais não foram consideradas como *outliers*, foi caracterizada pela binomial (m, π_1). Os valores paramétricos assumidos na simulação encontram-se descritos na Tabela 1. No caso da população binomial (m, π_2), arbitrariamente se fixou o valor paramétrico em $\pi_2 = 0{,}1$. Convém ressaltar que as observações geradas dessa população foram consideradas como *outliers* uma vez que estas apresentaram menor probabilidade de ocorrência (γ) quando comparada com a população binomial (m, π_1), além do fato de que $\pi_1 \neq \pi_2$.

Tabela 1. Valores paramétricos assumidos no modelo binomial (1).

n	π_1	γ
10	0,20	0,20
	0,80	0,40
50	0,20	0,20
	0,80	0,40
80	0,20	0,20
	0,80	0,40

Estimadores para o parâmetro π sobre contaminação

Seguindo as especificações descritas na Tabela 1 e a metodologia utilizada para gerar as amostras, o estimador de máxima verossimilhança (EMV) para o parâmetro π_1 do modelo binomial (1) sobre as amostras contaminadas foi obtido neste trabalho por,

$$\hat{\pi}_{EMV} = 1/m \sum_{k=0}^{m} k\, f_n(k) \quad (3)$$

em que: m representou o número de ensaios, previamente fixado em m = 100; $f_n(k)$ representou a proporção de observações iguais a k em uma amostra de tamanho n, conforme é verificado em (4):

$$f_n(k)=n^{-1}\sum_{i=1}^{n} I(Y_i=k), \quad k=0,...,m \tag{4}$$

Considerando-se o estimador $\hat{\pi}_{EMV}$ (3) e incorporando-se a função $\rho(z)$ (7) inerente aos estimadores pertencentes à classe E (RUCKSTUHL; WELSH, 2001), construiu-se o estimador $\hat{\pi}_E$ cuja expressão é dada por:

$$\hat{\pi}_E = \sum_{k=0}^{m} \rho(z) p_{\hat{\pi}_{EMV}}(k) \tag{5}$$

em que: $p_{\hat{\pi}_{EMV}}$ é a probabilidade de um modelo binomial, utilizando o estimador de verossimilhança como probabilidade de sucessos e z correspondeu ao argumento da função ρ, sendo este definido por

$$z=\frac{f_n(k)}{p_{\hat{\pi}_{EMV}}(k)} \tag{6}$$

sendo que

$$\rho(z)=\begin{cases} (\log(c_1)+1)z-c_1 & \text{se } z<c_1 \\ z\log(z) & \text{se } c_1<z<c_2 \\ (\log(c_2)+1)z-c_2 & \text{se } z>c_2 \end{cases} \tag{7}$$

em que: c_1 e c_2 são definidas como constantes de afinidade.

Pela recomendação de Ruckstuhl e Welsh (2001) de que $c1 < c2 = 1$, e tendo em vista o objetivo proposto, o desempenho do estimador robusto estudado neste trabalho foi avaliado, considerando-se os seguintes valores para c_1= 0,10; 0,20; 0,30; 0,40; 0,50; 0,60; 0,70; 0,80 e 0,90.

Para efeito de comparação, estimou-se o viés relativo (8) para as estimativas de $\hat{\pi}_{EMV}$ (3) e $\hat{\pi}_E$ (5). Assim sendo, realizaram-se 2.000 simulações Monte Carlo utilizando o *software* R (R DEVELOPMENT CORE TEAM, 2007).

$$v_E = \frac{\hat{\pi}_E - \pi_1}{\pi_1} \text{ e } v_{EMV} = \frac{\hat{\pi}_{EMV} - \pi_1}{\pi_1} \tag{8}$$

De acordo com os resultados obtidos, o valor adequado para c_1 em cada situação avaliada foi determinado pela análise da acurácia e precisão do estimador $\hat{\pi}_E$. No que se refere aos resultados do estimador $\hat{\pi}_{EMV}$, foi realizado apenas um estudo comparativo com π_1.

Resultados e discussão

Em consonância com o objetivo proposto, os resultados encontrados nas Tabelas 2 a 5 e ilustrados pela média das estimativas robustas $\hat{\pi}_E$ obtidas por meio da distribuição empírica gerada por Monte Carlo e viés (8) forneceram indicativo do desempenho do estimador robusto, pertencente à classe E, cuja expressão se encontra descrita na metodologia (5). Diferentes valores da constante de afinidade c_1 foram avaliados.

Especificamente, em relação aos resultados encontrados nas Tabelas 2 e 3, foi observado que, para diferentes probabilidades de mistura (γ = 0,20 e γ = 0,40), o estimador $\hat{\pi}_E$ apresentou desempenho diferenciado, tendo em vista que a probabilidade de mistura é dada como índice do grau de contaminação a que as populações simuladas foram submetidas. Portanto, há evidências estatísticas de que, em ambas as situações, a robustez desse estimador é mais pronunciada para determinados valores de c_1 escolhidos apropriadamente para cada grau de contaminação. Nesse contexto, observou-se que, dada à baixa contaminação (γ = 0,20), os resultados descritos na Tabela 2 revelaram diferenças das estimativas robustas em relação aos valores de constantes de afinidade. Tal fato ocorreu para todos tamanhos amostrais avaliados. Exemplificando essa situação, pode-se perceber que, para o valor paramétrico assumido em π_1 = 0,20 e tamanho da amostra n = 10, considerando-se o valor de c_1 = 0,6, o estimador proporcionou menor viés. Para as amostras maiores de tamanho n = 50 e n = 80, sobre o efeito de baixa contaminação nas amostras, esse estimador não apresentou resultados promissores que justifiquem a sua robustez. Quanto ao estimador de máxima verossimilhança, esse apresentou estimativa robusta, uma vez que esta foi acurada em relação ao parâmetro π_1. Essa robustez do estimador de máxima verossimilhança aconteceu pela baixa taxa de contaminação a que a amostra foi submetida (γ = 0,2).

Aumentando o grau de contaminação (γ = 0,40), os resultados encontrados na Tabela 3 confirmaram a robustez quando o valor de c_1 foi igual a 0,7. Tal fato foi detectável por meio do valor do viés calculado em -0,024, sendo este um valor considerável em relação à acurácia do estimador. Vale ressaltar que essa situação ocorreu para n = 10. No caso de amostras maiores n = 50 e 80, a robustez no estimador $\hat{\pi}_E$ foi verificada para c_1 = 0,1. Quanto ao tamanho amostral n = 50, foi observado que os valores dos vieses obtidos, mantendo-se as

constantes, revelaram que o estimador $\hat{\pi}_E$, considerando-se o grau de contaminação $\gamma = 0,40$, foi bem acurado e preciso, diferindo apenas na segunda casa decimal. O mesmo fato não ocorre com o $\hat{\pi}_{EMV}$, indicando que seu uso é inapropriado para altas probabilidades de mistura.

Tabela 2. Valores médios dos estimadores robustos e viés com taxa de mistura igual 0,20 e valores amostrais fixados em n = 10, 50 e 80, considerando-se o valor paramétrico $\pi_1 = 0,20$.

	$\pi_1 = 0,20$ n = 10		$\pi_1 = 0,20$ n = 50		$\pi_1 = 0,20$ n = 80	
	$\gamma = 0,20$		$\gamma = 0,20$		$\gamma = 0,20$	
c_1	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	viés
0,1	0,493	1,467	0,156	-0,218	0,119	-0,405
0,2	0,435	1,178	0,139	-0,303	0,107	-0,464
0,3	0,382	0,913	0,139	-0,303	0,105	-0,472
0,4	0,327	0,638	0,128	-0,358	0,099	-0,503
0,5	0,274	0,370	0,115	-0,425	0,089	-0,553
0,6	0,219	0,097	0,098	-0,506	0,080	-0,597
0,7	0,164	-0,180	0,079	-0,603	0,065	-0,671
0,8	0,109	-0,452	0,056	-0,718	0,048	-0,759
0,9	0,055	-0,724	0,030	-0,850	0,026	-0,869
$\hat{\pi}_{EMV}$	0,179	-0,105	0,179	-0,105	0,179	-0,105

Tabela 3. Valores médios dos estimadores robustos e viés com taxa de mistura igual 0,40 e valores amostrais fixados em n = 10, 50 e 80, considerando-se o valor paramétrico $\pi_1 = 0,20$.

	$\pi_1 = 0,20$ n = 10		$\pi_1 = 0,20$ n = 50		$\pi_1 = 0,20$ n = 80	
	$\gamma = 0,40$		$\gamma = 0,40$		$\gamma = 0,40$	
c_1	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	Viés
0,1	0,524	1,624	0,204	0,023	0,163	-0,180
0,2	0,466	1,333	0,196	-0,019	0,161	-0,190
0,3	0,408	1,043	0,185	-0,074	0,154	-0,227
0,4	0,345	0,727	0,170	-0,146	0,147	-0,260
0,5	0,290	0,451	0,152	-0,237	0,134	-0,326
0,6	0,233	0,166	0,130	-0,349	0,116	-0,419
0,7	0,195	-0,024	0,103	-0,482	0,094	-0,529
0,8	0,116	-0,417	0,072	-0,636	0,067	-0,664
0,9	0,058	-0,708	0,038	-0,809	0,035	-0,825
$\hat{\pi}_{EMV}$	0,160	-0,200	0,160	-0,200	0,160	-0,200

Mantendo-se as mesmas configurações analisadas anteriormente, porém alterando apenas o valor de π_1, os resultados encontrados na Tabela 4 evidenciaram um comportamento do estimador $\hat{\pi}_E$ mais 'estabilizado', no sentido de que sua robustez, assumindo a constante de afinidade $c_1 = 0,1$, foi verificada em todos os tamanhos amostrais avaliados. Analogamente, percebe-se o mesmo comportamento ao se aumentar o grau de contaminação das amostras para $\gamma = 0,40$ (Tabela 5). Dessa forma, verifica-se que, para baixos valores de c_1, mais especificamente em $c_1 = 0,2$, é onde ocorre a estimativa que se encontra mais próxima do parâmetro. Quanto ao desempenho do estimador de máxima verossimilhança, as situações representadas nas Tabelas 4 e 5 indicaram o maior viés desse estimador perante as outras situações.

Tabela 4. Valores médios dos estimadores robustos e viés com taxa de mistura igual 0,20 e valores amostrais fixados em n = 10, 50 e 80, considerando-se o valor paramétrico $\pi_1 = 0,80$.

	$\pi_1 = 0,80$ n = 10		$\pi_1 = 0,80$ n = 50		$\pi_1 = 0,80$ n = 80	
	$\gamma = 0,20$		$\gamma = 0,20$		$\gamma = 0,20$	
c_1	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	Viés
0,1	0,781	-0,022	0,771	-0,035	0,769	-0,037
0,2	0,694	-0,131	0,688	-0,138	0,692	-0,134
0,3	0,607	-0,241	0,605	-0,242	0,606	-0,241
0,4	0,525	-0,343	0,521	-0,347	0,522	-0,347
0,5	0,437	-0,453	0,434	-0,454	0,438	-0,451
0,6	0,347	-0,565	0,350	-0,561	0,351	-0,560
0,7	0,262	-0,671	0,263	-0,670	0,263	-0,670
0,8	0,173	-0,783	0,176	-0,779	0,177	-0,778
0,9	0,087	-0,890	0,088	-0,889	0,088	-0,889
$\hat{\pi}_{EMV}$	0,650	-0,188	0,650	-0,188	0,650	0,188

Tabela 5. Valores médios dos estimadores robustos e viés com taxa de mistura igual 0,40 e valores amostrais fixados em n = 10, 50 e 80, considerando-se o valor paramétrico $\pi_1 = 0,80$.

	$\pi_1 = 0,80$ n = 10		$\pi_1 = 0,80$ n = 50		$\pi_1 = 0,80$ n = 80	
	$\gamma = 0,40$		$\gamma = 0,40$		$\gamma = 0,40$	
c_1	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	viés	$\hat{\pi}_E$	Viés
0,1	0,884	0,105	0,897	0,122	0,898	0,123
0,2	0,786	-0,017	0,798	-0,002	0,798	-0,001
0,3	0,687	-0,141	0,698	-0,127	0,698	-0,126
0,4	0,590	-0,261	0,598	-0,251	0,599	-0,251
0,5	0,491	-0,386	0,498	-0,376	0,499	-0,376
0,6	0,391	-0,510	0,399	-0,501	0,399	-0,500
0,7	0,294	-0,631	0,299	-0,625	0,299	-0,625
0,8	0,196	-0,754	0,199	-0,750	0,199	-0,750
0,9	0,098	-0,877	0,099	-0,875	0,099	-0,875
$\hat{\pi}_{EMV}$	0,510	-0,362	0,510	-0,362	0,510	-0,362

De um modo geral, convém salientar que, em situações práticas, a probabilidade de mistura (γ) não é conhecida. Sendo assim, chama-se a atenção pelo fato de que as observações discrepantes também podem ser representadas em uma amostra pelo excesso de valores zero. Nota-se, assim, que o estimador pertencente à classe dos estimadores-E, avaliado neste trabalho, poderá ser aplicado sem nenhum problema, pois, para essa situação, a referida probabilidade de mistura (γ) poderá ser estimada como uma taxa. Assim, o pesquisador terá conhecimento sobre a taxa de contaminação, e a estimação de π poderá ser feita de acordo com o valor adequado para c_1, escolhido conforme recomendações propostas neste trabalho.

Uma alternativa mais atrativa para a utilização do estimador $\hat{\pi}_E$ seria nas situações em que a probabilidade de mistura γ fosse conhecida. Entretanto, esse fato não ocorre e, assim sendo, seguindo uma recomendação de Efron et al. (2001), a estimativa desta probabilidade poderá ser feita mediante a aproximação bayesiana empírica. Um exemplo de como se estimar esta probabilidade pode ser encontrado em Do et al. (2005).

Conclusão

O valor apropriado para a constante de afinidade c_1 depende do grau de contaminação da amostra e, portanto, é desejável que o pesquisador tenha alguma informação *a priori* sobre a probabilidade de mistura.

Para alta probabilidade de mistura ($\gamma = 0,40$), recomenda-se assumir $c_1 = 0,1$ nas situações de grandes amostras ($n = 50$ e $n = 80$).

Referências

COPAS, J. B. Binary regression models for contaminated data. **Journal of the Royal Statistical Society. Series B. Methodological**, v. 50, n. 2, p. 225-265, 1988.

DO, K. A.; MÜLLER, P.; TANG, F. A Bayesian mixture model for differential gene expression, **Journal of The Royal Statistical Society. Series C**, v. 54, n. 3, p. 627-644, 2005.

EFRON, B.; TIBSHIRANI, R.; STOREY, J. D.; TUSHER, V. Empirical Bayes analysis of a microarray experiment. **Journal of the American Statistical Association**, v. 96, n. 456, p. 1151-1160, 2001.

LINDSAY, B. G. Efficiency versus robustness: the case for minimum hellinger distance and related methods. **The Annals of Statistics**, v. 22, n. 4, p. 1081-1114, 1994.

R DEVELOPMENT CORE TEAM. **R:** A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2007.

RUCKSTUHL, A. F.; WELSH A. H. Robust fitting of the binomial model. **The Annals of Statistics**, v. 29, n. 4, p. 1117-1136, 2001.

SIMPSON, D. G. Minimum hellinger distance estimation for the analysis of count data. **Journal of the American Statistical Association**, v. 82, n. 399, p. 802- 807, 1987.

VICTORIA-FESER, M. P.; RONCHETTI, E. Robust Estimation for Grouped Data. **Journal of the American Statistical Association**, v. 92, n. 437, p. 333-340, 1997.

WOOD, G. R.; LAI, C. D.; QIAO, C. G. Estimation of a proportion using several independent samples of binomial mixtures. **The Australian and New Zealand Journal of Statistics**, v. 47, n. 4, p. 441-448, 2005.

Received on July 1, 2008.
Accepted on May 18, 2009.