Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1806-2563
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v37i1.21250

Improvement of the Wald method applied to the evaluation of zero-inflated binomial linear functions

Cleide Silveira Brasil Peixoto[1*], Marcelo Angelo Cirillo[2] and Augusto Maciel da Silva[1]

*[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, Brazil. [2]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil. *Author for correspondence. E-mail: cleidebpeixoto@gmail.com*

ABSTRACT. The Wald method is grounded on a statistic based on the asymptotic approximation of normal distribution. The method has shown incoherent values at a nominal level of confidence for the probability of coverage in intervallic estimates, mainly in small samples, noticeable in linear functions formed by binomial proportions. Current analysis improves this method used in inferring from binomial linear functions, taking into consideration zero-inflated samples. Improvement was assessed by Monte Carlo simulation techniques within different scenarios. Results show that the improvement proposed is recommended in situations in which sampling proportions are close to 0,5 and produce a maximum variance of the binomial proportions involved in the composition of the linear function.

Keywords: binomial families, probability of coverage, simulation.

Aprimoramento do método de Wald aplicado a estimação de funções lineares binomiais com excesso de zeros

RESUMO. O método de Wald é fundamentado em uma estatística que tem por base a aproximação assintótica da distribuição normal. Este método apresenta valores incoerentes de probabilidade de cobertura em estimativas intervalares em relação ao nível nominal de confiança, principalmente para pequenas amostras. Tal fato é perceptível em funções lineares formadas por proporções binomiais. O objetivo do trabalho consiste em aprimorar este método utilizado na inferência sobre funções lineares binomiais considerando amostras com excesso de zeros. Avaliou-se este aprimoramento utilizando técnicas de simulação Monte Carlo em diferentes cenários. Concluiu-se que o aprimoramento proposto é recomendável nas situações em que as proporções amostrais são próximas a 0.5 resultando em uma variância máxima das proporções binomiais envolvidas na composição da função linear.

Palavras-chave: famílias binomiais, probabilidade de cobertura, simulação.

Introduction

The Wald method is highly relevant among the known procedures in literature for inference from binomial proportions. The method, widely used to compare two binomial proportions, is characterized essentially for being asymptotic, where the distribution of the estimator is approximately normal. Due to this approximation, numerous studies show that the method presents shortcomings with regard to results of coverage probability and its use in small samples. Alternative methods are proposed to correct this deficiency. An improvement to the Wald method, proposed by Agresti and Coull (1998), briefly consists of adding four pseudo-observations, two successes and two failures, in the expression of the proportion estimator. This procedure is known as the 'add – 4 method'. However the more general problem of interval estimation for a linear function of binomial proportions mentioned by Price and Bonett (2004), including pairwise comparisons, complex contrasts, interaction effects and simple main effects (BONETT; WOODWARD, 1987), are factors that influence the probability coverage estimate.

Studying the Wald method and comparing it to other methods using the bootstrap approach, Carari et al. (2010) came to the conclusion that the Wald method presented probabilities of coverage with rates lower than the confidence coefficient´s nominal rates, thus jeopardizing its practical application to small samples. Regard to the add-4 method, the study showed that it stood out by producing adequate results for probabilities of coverage and intervals with shorter lengths.

The Wald method has also been used in dealing with linear functions which involve binomial proportions, also known as binomial families. A generalization of this method with its approach is stated by Price and Bonett (2004) as a confidence interval for the parameter rate $F=\sum_{i=1}^{q}\delta_i\pi_i$. As such, the confidence interval approximately 100(1-α)% from Wald to F is given (1)

$$\sum_{i=1}^{q}\delta_i\hat{\pi}_i \pm z_{\alpha/2}\sqrt{\sum_{i=1}^{q}\frac{\delta_i^2\hat{\pi}_i\left(1-\hat{\pi}_i\right)}{n_i}} \qquad (1)$$

where:

n_i is the reference sample size for i-th binomial population;

$\hat{\pi}_i = Y_i/n_i$; δ_i is a known co-efficient and specified by researcher;

q is the number of coefficients involved in the function. Even with the above generalization, the Wald method still presents the flaws mentioned and in this context alternative methods have emerged. More details may be found in Price and Bonett (2004), Tebbs and Roths (2008) and Cirillo et al. (2009).

It is worth mentioning that the Wald method applied to the comparison of two binomial proportions or generalized for binomial linear functions put forth in the literature does not consider zero-inflated binomial (ZIB) samples. In this case, the use of these methods would certainly exacerbate the deficiencies mentioned previously with regard to coverage probability and applications to small samples. Silva and Cirillo (2010) warn that, even assuming the adequacy of the model, some zeros may be considered outliers, and thus different methods of assessment are sensitive to this anomaly.

Consequently, robust assessment methods must be arrived at which will consider the presence of divergent data and provide a coherent estimate of the parameter required. Faced with this problem, methods which deal with the effect of outliers on estimates is still the focus of research. Andrade et al. (2014) have proposed a bootstrap algorithm which looks at the effect of divergent observations and/or influential on estimates for non-linear parameter models.

While keeping focus on tallying data, Silva et al. (2012) studied the zero-inflated effect on a Poisson model according to sampling size and different parametric rates inferring from a zero-inflated Poisson (ZIP) model. The authors reached the conclusion that discrimination of ZIP and Poisson through a score test was recommended on the basis of a sampling size greater than n = 40 in samples with a high proportion of null rates.

Wood et al. (2005) proposed two alternatives to estimate the probability of success in binomial samples tainted with divergent observations. These alternatives referred to two estimators differentiated by arithmetic average and rationalized means of the proportions observed.

After comparing estimators variances, the authors come to the conclusion that an estimator's recommendation will apply at different situations characterized by the distribution of proportions and the number of trials (n) performed.

In view of a scarcity of robust, zero-inflated methods to estimate binomial linear functions, current research is characterized by the proposal for an improvement of the Wald method applied to the intervallic binomial linear functions. The above turns the method robust to zero-inflated binomial samples and replaces the maximum likelihood estimates by robust estimates. Several scenarios among different parametric configurations are assessed via Monte Carlo to validate the method.

Material and methods

Following the objectives proposed, the method was performed in two steps, specified in sections 2.1 and 2.2, with details below.

Simulation of zero-inflated binomial samplings.

Using Monte Carlo simulation techniques, the zero-inflated binomial samples were generated while taking into account the ZIB model (RUCKSTUHL; WELSH, 2001), characterized by the mixture of two components in such a way that one component presumes that the occurrence of zero is defined by a γ probability, while the other component represents a binomial distribution with a (1-γ) probability. The ZIB model is thus defined by the following expression (2)

$$P(Y=y)=\begin{cases}\gamma+(1-\gamma)(1-\pi)^{m}, & \text{if} \quad y=0\\ (1-\gamma)\begin{pmatrix}m\\ y\end{pmatrix}\pi^{y}(1-\pi)^{m-y}, & \text{if}\ y=1,2,\ldots,m\end{cases} \qquad (2)$$

with the expectation that

$$E(Y) = (1 - \gamma)\ (1 - \pi)\ m\pi$$

and the variance defined as

$$Var\ (Y) = [(1-\gamma)\ m\pi][(1 - \pi)(1 - \gamma m)]$$

where γ is a probability of zero occurrence and m the number of Bernoulli experiments. Using the model given in (2), set m = 100 Bernoulli experiments for n samples sizes, the parametric rates assumed in the Monte Carlo simulation process are described in Table 1.

Table 1. Parametric rates to generate zero-inflated binomial samples.

$\gamma = 0.2$		$\gamma = 0.3$	
n	π	n	π
30	0.5 and 0.7	30	0.5 and 0.7
40	0.5 and 0.7	40	0.5 and 0.7
50	0.5 and 0.7	50	0.5 and 0.7
60	0.5 and 0.7	60	0.5 and 0.7
70	0.5 and 0.7	70	0.5 and 0.7
80	0.5 and 0.7	80	0.5 and 0.7
90	0.5 and 0.7	90	0.5 and 0.7

Keeping the parametric rate configurations, estimators for robust to zero-inflated binomial proportions are defined by π_{zib}. This estimator was obtained as a combination of estimators found in Ruckstuhl and Welsh (2001) and Silva and Cirillo (2010) .

$$\hat{\pi}_{zib} = \sum_{y=0}^{m} \rho_s(x)\hat{\pi}_{mle} \tag{3}$$

where $\hat{\pi}_{mle}$ is the maximum likelihood estimator of π given in (4)

$$\hat{\pi}_{mle} = \frac{1}{m}\sum_{i=1}^{n} y_i f_n(y) \tag{4}$$

where:

$$f_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y=y_i) \tag{5}$$

The expression presented in (3) is based on the likelihood disparity of E-estimators (RUCKSTUHL; WELSH, 2001) and $\rho_s(x)$ represents a function that minimizes the disparity.

$$\rho_s(x) = \begin{cases} \left\{\left[\ln(c_1) + \frac{(1-u)\ln(c_1)+1}{u}\right]c_1^{1-u}\right\}x^u - [(1-u)\ln(c_1)+1]\frac{c_1}{u}, \; x<c_1 \\ x\ln(x), \text{ if } c_1 \leq x \leq c_2 \\ \left\{\left[\ln(c_2) + \frac{(1-u)\ln(c_2)+1}{u}\right]c_2^{1-u}\right\}x^u - [(1-u)\ln(c_2)+1]\frac{c_2}{u}, \; x>c_2 \end{cases} \tag{6}$$

where c_1 and c_2 are affinity constants.

The function argument $x = \frac{f_n(y)}{p_n(y)}$ is fixed, where $p_n(y)$ is the probability for a Binomial distribution, considering the estimate of π given by (4). The rates for s are set in 1 and 2, defining the estimator $\hat{\pi}_{zib}$ in two approaches mentioned in current research as the incorporation of ρ_1 and ρ_2 components.

We would like to emphasize that the structure of ρ_1 and ρ_2 in the estimation process is understood as a systematic component taking into consideration that the researcher may choose which function will be assumed. Note that by assuming $u = 1$, $\rho_2 = \rho_1$ suggests that ρ_2 is a generalization of ρ_1 differing only in the asymptotic properties.

In this context, the rates for affinity constants c_1 and c_2 are defined on the basis of the component in such a way that, upon assuming the component ρ_1, the coefficients $u = c_2 = 1$ are fixed and a value for $c_1 < c_2 = 1$ is investigated. Thus, $\rho_1(x)$ is prone to a greater increase when $x \to \infty$.

Keeping the $c_1 < c_2 = 1$ inequality in mind, according to Ruckstuhl and Welsh (2001), the maximum likelihood estimates tend to be more robust. Taking into consideration ρ_2, it is assumed that $c_1 = 0.1$ keeping the $c_1 < c_2 = 1$ restriction, whereas the rate of *u* is examined so as to reduce the increase of $\rho_2(x)$ when $x \to \infty$.

It is worth underscoring that the accuracy and precision of the estimator (3) depend on the rates of the affinity constants c_1 and c_2 which make it robust to expected numbers of null values. Consequently, the research for these constants was carried out by a computer routine.

The intention of Silva and Cirillo (2010) was to reproduce Tables for rates of *u* and c_1 in two situations of $\rho_s(x)$ described in (6). Thus the researcher may use the estimator π_{zib} in a statistical inference based on the maximum likelihood estimate of π on a zero inflated sample, observing the deviations between π_{mle} and π_{zib} estimates, given by $|\pi_{zib} - \pi_{mle}| < k$, where k indicates a tolerable rate for this difference. The first step is to evaluate the maximum likelihood estimate (π_{mle}), proceeding the evaluation of $p_n(y)$ for application on $\rho_s(x)$ and evaluation of π_{zib}, according to expression (4). Thus foregrounded, Tables presented by Silva and Cirillo (2010) may be helpful for the choice of *u* and c_1.

So that zero-inflated binomial estimates could be compared and validated, the relative bias for π_{mle} and $\hat{\pi}_{zib}$ estimates were valued according to expression (7)

$$V_{mle} = \frac{\hat{\pi}_{mle} - \pi}{\pi} \text{ and } V_{zib} = \frac{\hat{\pi}_{zib} - \pi}{\pi} \tag{7}$$

Definition and estimation of linear functions of binomial proportions taking into consideration the Wald method

After generating the binomial samples, the structure of the binomial linear functions was represented by the parametric rate, as shown in (8)

$$F=\sum_{i=1}^{q}\delta_i\pi_i \quad , \tag{8}$$

where q is the total number of binomial populations, the i-th coefficient associated with the success proportion regarding to the i-th binomial population is expressed as δ_i, following specifications shown in Table 2.

Table 2. Coefficients used for linear function specifications.

Family	*q*	Coefficient vector used in composition of F
F1	3	$\tilde{\delta}_1 = (2, -1, -1)$
F2	5	$\tilde{\delta}_2 = (4, -1, -1, -1, -1)$
F3	7	$\tilde{\delta}_3 = (6, -1, -1, -1, -1, -1, -1)$
F4	10	$\tilde{\delta}_4 = (9, -1, -1, -1, -1, -1, -1, -1, -1, -1)$

For each F linear function representing a binomial family, the intervallic estimates for F were numerated, taking into account Wald's confidence intervals according to expression (1). Maximum likelihood estimates were replaced by π_{zib} estimates with the systematic ρ_1 and ρ_2 component.

Finally, according to assessment scenario (Table 1), the intervals adapted for robust zero-inflated proportions were compared by a 100 (1-α)% interval for exact probability of coverage for a fixed value of F(8) defined by (9)

$$\sum_{y_1=0}^{n_1}\cdots\sum_{y_q}^{n_q}\prod_{i=1}^{q}\binom{n_i}{y_i}p_i^{y_i}(1-p_i)^{n_i-y_i}\,I(y_1,\dots,y_q), \tag{9}$$

where I $(y_1,\dots,y_q)$ equals 1 if the intervals contains F (8), when $Y_1 = y_1,\dots,Y_g = y_g$ equals zero if the interval does not contain F (8). An approximation is obtained from 2000 Monte Carlo simulations by means of estimated interval percentages which include the F parameter calculated from a program developed by R 3.00 software (R DEVELOPMENT CORE TEAM, 2011).

Results and discussion

Taking into consideration the evaluation scenarios mentioned in Methodology (Section 2.1), the number of Bernoulli m = 100 trials in this first step was established when obtaining the study samples for the recommended methods.

With this specification, $\hat{\pi}_{mle}$ maximum likelihood estimates and zero-inflated robust as represented by $\hat{\pi}_{zib}$ were obtained in binomial samples generated via Monte Carlo with the null observations percentages nearing 20 and 30% as per the parametric values specified in the γ = 0.2 and 0.3 mixture probability. Results are shown in Tables 3-6.

Table 3. Comparative results of $\hat{\pi}_{mle}$ and π_{zib} estimators, taking into account the parametric rate π = 0.5 with c_2 = u = 1 restriction characterizing the systematic ρ_1 component.

n	γ	c_1	$\hat{\pi}_{mle}$	Bias	$\hat{\pi}_{zib}$	bias
30	0.2	0.2900	0.3995	-0.2010	0.5000	-0.0001
30	0.3	0.4300	0.3504	-0.2992	0.5018	0.0036
40	0.2	0.2600	0.4006	-0.1988	0.5007	0.0014
40	0.3	0.4300	0.3499	-0.3002	0.5001	0.0003
50	0.2	0.2600	0.4002	-0.1996	0.4941	-0.0119
50	0.3	0.4300	0.3508	-0.2984	0.4986	-0.0027
60	0.2	0.2500	0.3985	-0.2030	0.4983	-0.0034
60	0.3	0.4300	0.3506	-0.2988	0.4983	-0.0034
70	0.2	0.2500	0.3998	-0.2024	0.4913	-0.0173
70	0.3	0.4300	0.3504	-0.2992	0.4974	-0.0052
80	0.2	0.2400	0.3991	-0.2018	0.4962	-0.0076
80	0.3	0.4300	0.3497	-0.3006	0.4982	-0.0037
90	0.2	0.2400	0.4000	-0.2000	0.4914	-0.0171
90	0.3	0.4300	0.3499	-0.3002	0.4975	-0.0049

Table 4. Comparative results of $\hat{\pi}_{mle}$ and π_{zib} estimators taking into account the parametric rate = 0.5 with c_1 = 0.1 and c_2 = 1 restriction characterizing the systematic ρ_2 component.

n	γ	u	$\hat{\pi}_{mle}$	Bias	$\hat{\pi}_{zib}$	bias
30	0.2	0.1540	0.3995	-0.2010	0.5102	0.0203
30	0.3	0.1800	0.3514	-0.2972	0.4930	-0.0141
40	0.2	0.1400	0.4000	-0.2000	0.5026	0.0052
40	0.3	0.1730	0.3499	-0.3002	0.5020	0.0041
50	0.2	0.1310	0.4003	-0.1994	0.4997	-0.0006
50	0.3	0.1700	0.4002	-0.1995	0.4941	-0.0119
60	0.2	0.1240	0.4002	-0.1995	0.4952	-0.0096
60	0.3	0.1660	0.3507	-0.2986	0.4996	-0.0008
70	0.2	0.1170	0.3993	-0.2014	0.5039	0.0078
70	0.3	0.1640	0.3509	-0.2982	0.4970	-0.0061
80	0.2	0.1130	0.4003	-0.1994	0.4900	-0.0201
80	0.3	0.1610	0.3500	-0.3000	0.4982	-0.0037
90	0.2	0.1060	0.3999	-0.2002	0.5102	-0.0204
90	0.3	0.1590	0.3491	-0.3018	0.5000	-0.0010

Table 5. Comparative results of $\hat{\pi}_{mle}$ and π_{zib} estimators taking into account the parametric rate π = 0.7 with c_2 = u = 1 restriction characterizing the systematic ρ_1 component.

n	γ	c_1	$\hat{\pi}_{mle}$	Bias	$\hat{\pi}_{zib}$	Bias
30	0.2	0.0001	0.5580	-0.2029	0.8322	0.0188
30	0.3	0.2700	0.4889	-0.3016	0.7011	0.0015
40	0.2	0.1500	0.5611	-0.1984	0.6977	-0.0033
40	0.3	0.2700	0.4924	-0.2966	0.6994	-0.0008
50	0.2	0.1500	0.5610	-0.1996	0.6961	-0.0056
50	0.3	0.2700	0.4907	-0.1986	0.7051	0.0073
60	0.2	0.1500	0.5591	-0.2990	0.6979	-0.0030
60	0.3	0.2700	0.4899	-0.3001	0.7050	0.0072
70	0.2	0.1500	0.5597	-0.2004	0.6961	-0.0056
70	0.3	0.2800	0.4902	-0.2997	0.6951	-0.0069
80	0.2	0.1500	0.5605	-0.1990	0.6944	-0.0080
80	0.3	0.2800	0.4904	-0.2994	0.6978	-0.0012
90	0.2	0.1500	0.5597	-0.2004	0.6976	-0.0035
90	0.3	0.2800	0.4921	-0.2970	0.6949	-0.0073

Table 6. Comparative results of $\hat{\pi}_{mle}$ and π_{zib} estimators taking into account the parametric rate $\pi = 0.7$ with $c_1 = 0.1$ and $c_2 = 1$ restriction characterizing the systematic ρ_2 component.

n	γ	u	$\hat{\pi}_{mle}$	Bias	$\hat{\pi}_{zib}$	Bias
30	0.2	0.1300	0.5589	-0.2820	0.7078	0.0134
30	0.3	0.1400	0.4904	-0.2994	0.7025	0.0040
40	0.2	0.1270	0.5582	-0.2026	0.7020	-0.0019
40	0.3	0.1400	0.4920	-0.2971	0.6979	-0.0017
50	0.2	0.1240	0.5604	-0.1986	0.7001	0.0044
50	0.3	0.1400	0.4908	-0.2990	0.6972	-0.0014
60	0.2	0.1220	0.5605	-0.2013	0.6967	0.0017
60	0.3	0.1390	0.4895	-0.2986	0.7025	0.0028
70	0.2	0.1190	0.5604	-0.3001	0.7008	-0.0072
70	0.3	0.1390	0.4884	-0.2997	0.7015	0.0030
80	0.2	0.1180	0.5590	-0.1993	0.7002	-0.0008
80	0.3	0.1390	0.4899	-0.2994	0.6953	-0.0068
90	0.2	0.1160	0.5592	-0.2004	0.7048	-0.0038
90	0.3	0.1390	0.4897	-0.2970	0.6983	-0.0052

In short, results made it clear that, in fact, in zero-inflated contaminated binomials, estimates for maximum likelihood were not accurate. This statement might be confirmed from the bias results, including situations of greater size sampling. However, when taking into consideration $\hat{\pi}_{zib}$ estimates, it was noted that for almost all sample sizes and γ rates on an average the relative biases were less than 0.01, including small swings due to the Monte Carlo error in $\pi = 0.5$ Tables (3 and 4) and $\pi = 0.7$ rates (Tables 5 and 6).

Based on results on π_{zib} estimates accuracy, the composition of binomial linear functions for the Wald method was conducted and coverage probabilities were calculated. For comparison purpose, a 95% nominal confidence level was taken into consideration. Each binomial family was represented by F_1, F_2, F_3 and F_4, respectively with regard to $\tilde{\delta}_{1,}$ $\tilde{\delta}_{2,}$ $\tilde{\delta}_3$ and $\tilde{\delta}_4$ coefficient vectors, described in Table 2. Thus, the graphics with probability estimates are shown as follows in the Figures 1 - 8:

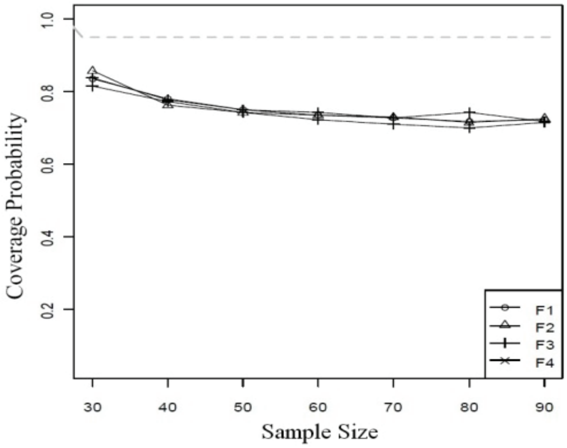

Figure 1. Probability of coverage assuming parameters γ = 0.2 and $\pi = 0.5$ and the systematic component ρ_1.

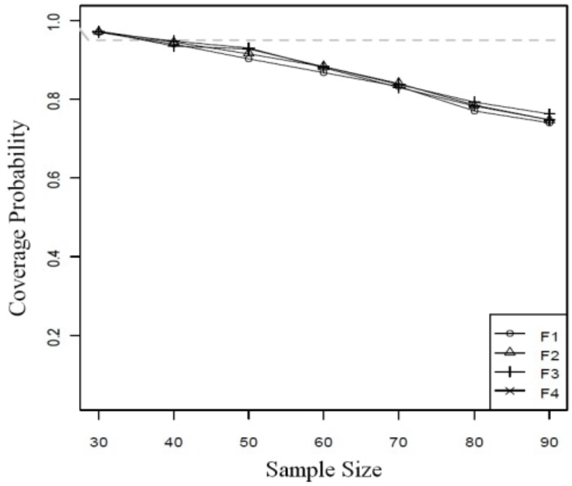

Figure 2. Probability of coverage assuming parameters γ = 0.2 and $\pi = 0.5$ and the systematic component ρ_2.

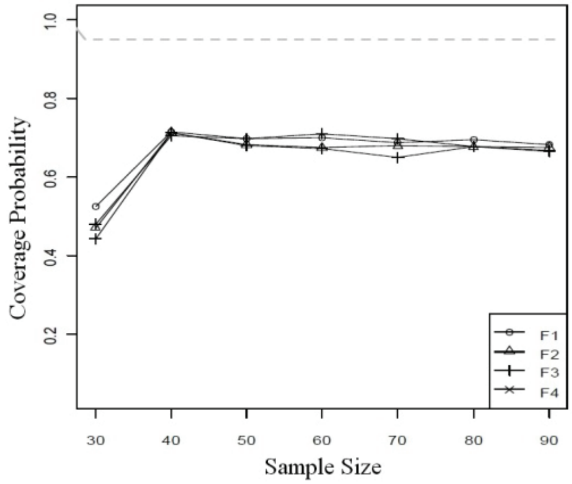

Figure 3. Probability of coverage assuming the parameters γ = 0.2 and $\pi = 0.7$ and the systematic component ρ_1.

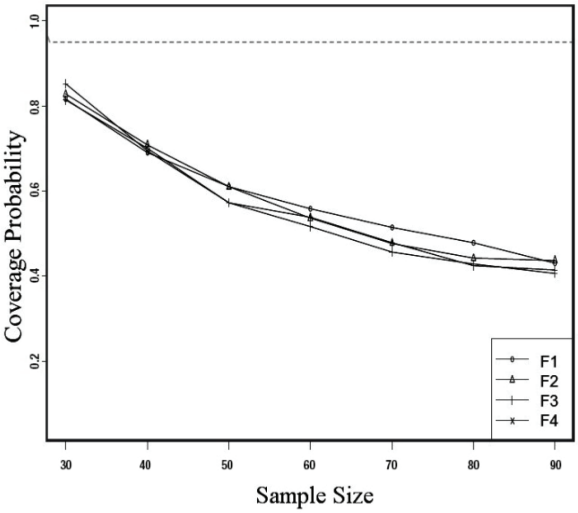

Figure 4. Probability of coverage assuming the parameters γ = 0.2 and $\pi = 0.7$ and the systematic component ρ_2.

Keeping a mean proportion of null values around 20% ($\gamma = 0.20$) of sampled observations, the results shown in Figures 1-4 made it clear that the increase in sampling size resulted in a decrease of coverage probability, with rates much lower than the nominal confidence level.

This was demonstrated by arranging the binomial families using π_{zib} estimates with the use of ρ_1 and ρ_2 components. However, when the null observation proportion was increased to about 30% of sample units ($\gamma = 0.30$), while taking into consideration the parametric values which maximize the variance of binomial proportions, that is, $\pi = 0.5$, the binomial families whose zero-inflated proportions were estimated with ρ_1 components showed probabilities of greater coverage at the nominal level of confidence (Figure 5). The same result for all sample sizes was observed when the parametric value increased, in situations where estimates were obtained using ρ_1 and ρ_2 systematic component (Figures 7-8).

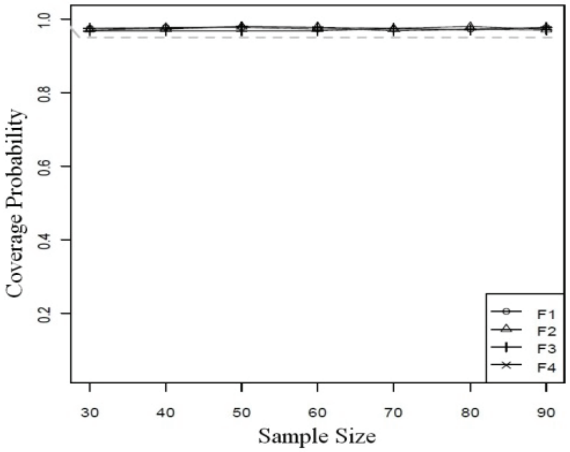

Figure 5. Probability of coverage assuming the parameters $\gamma = 0.3$ and $\pi = 0.5$ and the systematic component $\rho_{1.}$

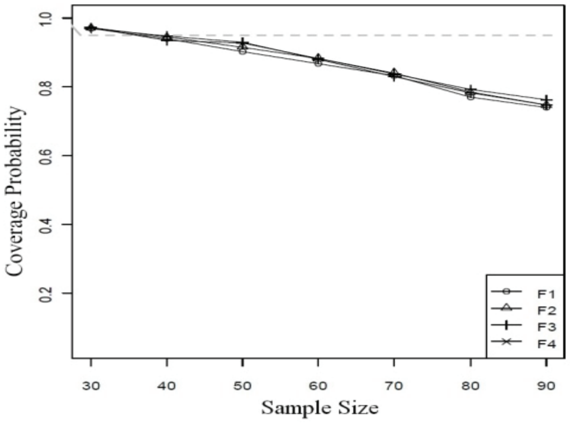

Figure 6. Probability of coverage assuming the parameters $\gamma = 0.3$ and $\pi = 0.5$ and the systematic component $\rho_{2.}$

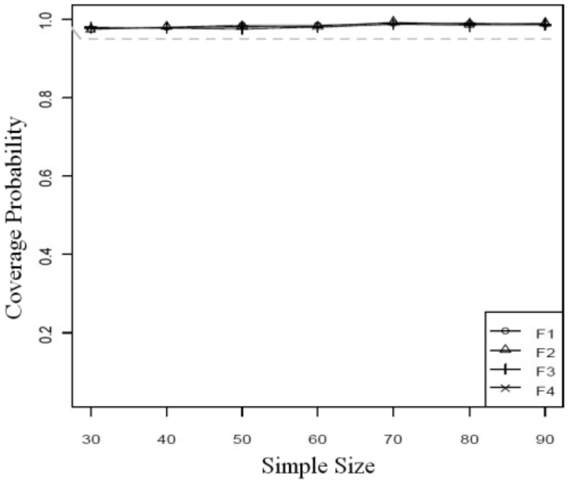

Figure 7. Probability of coverage assuming the parameters $\gamma = 0.3$ and $\pi = 0.7$ and the systematic component ρ_1.

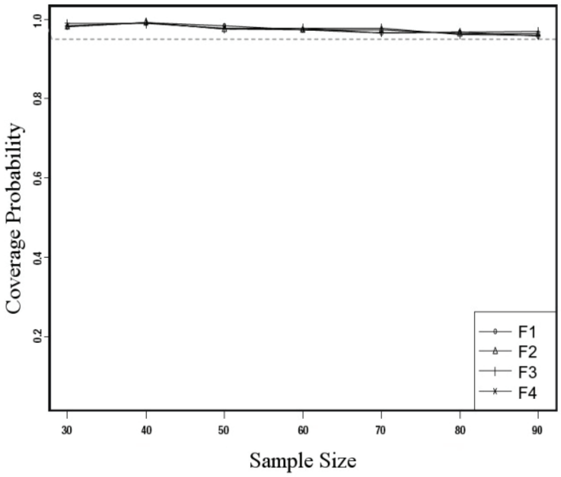

Figure 8. Probability of coverage assuming the parameters $\gamma = 0.3$ and $\pi = 0.7$ and the systematic component ρ_2.

It is worth mentioning that the Wald method, put into context for the obtainment of the estimates of binomial families, was assessed by Cirillo et al. (2009) for the use of the infinite bootstrap algorithm recommended by Conlon and Thomas (1990). Within this approach, authors of different assessment scenarios also concluded that results related to the probabilities of coverage were incoherent with the nominal level of confidence.

Silva and Cirillo (2010) produced studies related to the use of a robust estimator used in the inference of a binomial model contaminated by the mixture of binomial populations, when samples were obtained through Monte Carlo simulations. This study used an estimator belonging to the E estimator class (RUCKSTUHL; WELSH, 2001) incorporated into the $\rho_1(x)$ (8), a component which altered the E estimator. Several c_1 affinity constant rates were considered, specified in rates $0.1 \leq c_1 \leq 0.9$ sample sizes equal to 10, 50 and 80, besides the mixture rates equal to 0.20 and 0.40. The main conclusive results were illustrated in the recommendation to assume $c_1 = 0.1$ for samples greater than $n = 50$.

Already confirmed results were described regarding to flows noticed in the Wald method and the choice of c_1 constants based on sampling size and degree of contamination for the results listed in this work.

The Wald method, when using zero-inflated proportion estimates obtained by the $\hat{\pi}_{zib}$ estimator incorporated into the systematic ρ_2 component, may be recommended in situations with proportions which maximize the binomial family variance, that is π 0.7, since, for this parametric value, the scenarios evaluated led to coverage probabilities greater than 95%.

Conclusion

The use of the Wald method incorporated into estimates for zero-inflated binomial proportions using the ρ_2 component showed results in line with the nominal confidence level of binomial proportions. In practical terms, this method is recommended for samples in which proportions are close to 0.7 with proportions close to 0.3.

References

ANDRADE, L. R.; CIRILLO, M. A.; BEIJO, L. A. Proposal of a bootstrap procedure using measures of influence in non-linear regression models with outliers. **Acta Scientiarum. Technology**, v. 36, n. 1, p. 93-99, 2014.

AGRESTI, A.; COULL, B. A. Approximate is better than ‘exact’ for interval estimation of binomial proportions. **The American Statistician**, v. 52, n. 2, p. 119-126, 1998.

BONETT, D. G.; WOODWARD, J. A. Application of the Kronecker product and Wald test in log-linear models. **Computational Statistics Quartely**, v. 3, n. 1, p. 235-243, 1987.

CARARI, M. L.; LIMA, P. C.; FERREIRA, D. F.; CIRILLO, M. A. Estimação da diferença entre duas proporções binomiais via bootstrap. **Revista Brasileira de Biometria**, v. 28, n. 3, p. 112-134, 2010.

CIRILLO, M. A.; FERREIRA, D. F.; SAFÁDI, T. S. Avaliação de métodos de estimação intervalar para funções lineares binomiais via bootstrap infinito. **Ciência e Agrotecnologia**, v. 33, n. esp., p. 1741-1746, 2009.

CONLON, M.; THOMAS, R. G. A new confidence interval for the difference of two binomial proportions. **Computational Statistics and Data Analysis**, v. 9, n. 2, p. 237-241, 1990.

PRICE, M. R.; BONETT, D. G. An improved confidence interval for a linear function of binomial proportions. **Computational Statistics and Data Analysis**, v. 45, n. 3, p. 449-456. 2004.

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

SILVA, A. M.; CIRILLO, M. A. C. Estudo por simulação Monte Carlo de um estimador robusto utilizado na inferência de um modelo binomial contaminado. **Acta Scientiarum. Technology**, v. 32, n. 3, p. 303-307, 2010.

SILVA, V. S. P.; CIRILLO, M. A.; CESPEDES, J. G.; A study of the score test in discrimination poisson and zero-inflated poisson models. **Acta Scientiarum. Technology**, v. 35, n. 2, p. 333-337, 2012.

TEBBS, J. M.; ROTHS, S. A. New large-sample confidence intervals for a linear combination of binomial proportions. **Journal of Statistical Planning and Inference**, v. 138, n. 6, p. 1884-1893, 2008.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2011.

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 August 16, 2012.
Accepted on November 19, 2012.

APPENDIX 1

R function for evaluation of fn, described in (5)

```
fny = function (m, n, data, vet)
{
m: number of Bernoulli trials
n: sample size
data: binomial sample inflated by zeros
vet: vector formed by 0,1,…,m
for (a in 1:(m))
{
prop = 0;    aux = vet[a]
for (b in 1:n)
{
if (aux == data[b]) prop = prop + 1
}
v cont[a] = (prop) n^-1
}
return (v cont)
}
Obtaining MLE when the specifications of the function arguments are given
rfny = fny (m, n, data, vet)
est_MLE = sum (data*rfny) m^-1
d = d binom (y, m, est_MLE)                    input value on estimaPzib function
x = (rfny massa^-1)                            input value on estimaPzib function
```

APPENDIX 2

Function to estimate the robust binomial proportion inflated by zeroes

```
estimaPzib = function (x, d, c1, c2, u)
{
c1 and c2 : specification of constants to be used in ρ1
u: constant  to be used in ρ2
x: rate to be replaced on function p1 or p2
d: densities given the maximum likelihood estimates
estPzib = 0
for (b in 1 : length (x))
{
if (x[b] > = c1 && x[b] < = c2)     rho[b] = x[b]*log(x[b])
if (x[b] < c1)
rho[b] = ((c1 ^ (1-u) * log(c1) + ((1-u) * log(c1) + 1) * (c1 ^ (1-u) u^-1)) * x[b] ^ u)
- (((1 - u) * log(c1) + 1) * c1 u^-1)
if(x[b] > c2)
rho[b] = ((c2 ^ (1 - u) * log(c2) + ((1 - u) * log(c2) + 1) * (c2 ^ (1 - u) u^-1)) * x[b] ^ u)
- (((1 - u) * log(c2) + 1) * c2 u^-1)
auxPzib = rho[b]*d[b];     estPzib = auxPzib + estPzib
}

return (estPzib)

}
```

Robust Estimate to zero excess

```
With regard to this step, the researcher may choose between ρ1 or ρ2
u=0.13                  function considering ρ2 for any rate of u different from 1;
u = 1                                                                     function considering ρ1
Assume any value for c1, keeping the restriction c1 < c2 = 1 on ρ1 or ρ2
 c1 is researched, like to an example assumed as c1 = 0.1
c1 = 0.1; c2 = 1
```

Evaluation of Pzibestimator

```
Specify the functions arguments
Pzib = estimaPzib (x,d,c1,c2,u)
```