Revista Colombiana de Estadística
January 2015, Volume 38, Issue 1, pp. 93 to 105
DOI: http://dx.doi.org/10.15446/rce.v38n1.48803

Simulation Studies of a Hölder Perturbation in a New Estimator for Proportion Considering Extra-Binomial Variability

Estudios de simulación de una perturbación Hölder en un nuevo estimador de proporción considerando la variabilidad extra-binomial

Augusto Maciel da Silva[1,a], Marcelo Angelo Cirillo[2,b]

[1]Departamento de Estatística, Centro de Ciências Naturais e Exatas, Universidade Federal de Santa Maria, Santa Maria, Brasil

[2]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, Brasil

Abstract

This present work aims to propose an estimator in order to estimate the probability of success of a binomial model that incorporates the extra-binomial variation generated by zero-inflated samples. The construction of this estimator was carried out with a theoretical basis given by the Holder function and its performance was evaluated through Monte Carlo simulation considering different sample sizes, parametric values (π), and excess of zero proportions (γ). It was concluded that for the situations in ($\gamma = 0.20$) and ($\gamma = 0.50$) that the proposed estimator presents promising results based on the specified margin of error.

***Key words*:** Binomial Distribution, Monte Carlo simulation, Robust Estimator, Robustness.

Resumen

El presente trabajo tiene como objetivo proponer un estimador para estimar la probabilidad de éxito de un modelo binomial que incorpora la variación extra-binomial generada por muestras cero-inflados. La construcción de este estimador se llevó a cabo con una base teórica dada por la función Holder y su desempeño fue evaluado a través de la simulación de Monte Carlo considerando diferentes tamaños de muestra, valores paramétricos (π), y el exceso de proporciones cero (γ). Se concluyó que para las situaciones en ($\gamma = 0{,}20$) y ($\gamma = 0{,}50$) que el estimador propuesto presenta resultados prometedores basados en el margen de error especificado..

***Palabras clave*:** distribución binomial, estimador robusto, simulación Monte Carlo, robustez.

[a]Professor. E-mail: augusto.silva@ufsm.br
[b]Professor. E-mail: macufla@dex.ufla.br

1. Introduction

The inference on the parameter of a binomial population proportion, in general, is carried out considering sampling units are independent and provenient from a single population.

However, there are situations in certain data sets where the sampling variance may be superior in relation to the expected variability in the binomial model. Uncountable factors may cause overdispersion, among them, we can mention the existence of a correlation among the individual responses, data clustering and outliers.

Starting from the assumption that individuals belonging to the same population are more likely to provide correlated responses, meaning that an individual response depends on the previous response, consequently excess of zeros in a sample may occur, and in these cases, an alternative is given for modeling through the binomial model correlated and proposed by Kupper & Haseman (1998), in order to adjust extra-binomial variance caused by overdispersion or subdispersion (Achcar & Junqueira 2002).

In the presence of covariates, Hinde & Demetrio (1978) studied binary responses with overdispersion assuming random variables Y_i represent the success number of samples of size $m_i (i = 1, \ldots, n)$, where n is the n^{th} element of each sample. Thus, writing $E[Y_i] = \mu_i = m_i \pi_i$ through the generalized linear model, the proportion π_i is modulated assuming the explanatory variables X_i with a fitting link function.

Regarding robust inferential methods, which in general attenuate the present effects of outliers in the samples, we can mention some kinds of estimators, as M-estimators (Huber 1964) and minimum disparity estimators (Lindsay 1994). Specifically in the case of discrete data we can refer to M-estimators (Simpson 1987), minimum disparity estimators (Simpson, Carrol & Ruppert 1987) and E-estimators (Ruckstuhl & Welsh 2001).

As we know zero-inflated binomial samples in general exhibit an asymmetric form, thus explaining E-estimator use in the π proportion estimation of a binomial population, which model is described by

$$p_\pi(y) = \left(\begin{array}{c} m \\ y \end{array} \right) \pi^y (1-\pi)^{m-y} \text{with } y = 0, \ldots, m. \tag{1}$$

According to Ruckstuhl & Welsh (2001), the E-estimator is derived from a modification in the likelihood in order to reduce the effect of observations in the tails of the distributions. A brief presentation of the construction of this estimator is given below.

Assuming the disparity function $H(\pi, f_n)$ defined by

$$H(\pi, f_n) = \sum_{y=0}^{m} \rho(x) p_\pi(y), \text{ where } x = \frac{f_n(y)}{p_\pi(y)} \tag{2}$$

and $f_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i = y), y = 0, \ldots, m$, correspondent to the proportion of observations equal y in a sample of size n and $\mathrm{p}_\pi(\mathrm{y})$ to the probability of success of π of the binomial model, and

$$\rho(x) = \begin{cases} (\ln(c_1) + 1)x - c_1, & \text{if } x < c_1 \\ x \ln(x), & \text{if } c_1 \leq x \leq c_2 \\ (\ln(c_2) + 1)x - c_2, & \text{if } x > c_2 \end{cases} \tag{3}$$

where c_1 and c_2 are tuning constants from which the estimator depends. Based on these specifications, the estimator $\hat{\pi}$ that minimizes H is given by:

$$\hat{\pi} = \underset{\pi}{\operatorname{argmin}} \mathrm{H}(\pi, f_n) \tag{4}$$

The choice of tuning constants acts directly on the robustness of the estimators, providing them with full asymptotic efficiency, and good robustness properties (Basu, Shiyoa & Park 2011). In this way the accuracy of this estimator is given by the choice of tuning constants c_1 and c_2. In the case of binomial mixture, Silva & Cirillo (2010) concluded that the appropriate value for the tuning constant c_1 depends on the degree of contamination of the sample and, therefore, it is desirable that the researcher have some prior information about the probability of mixture.

Some recommendations made by Ruckstuhl & Welsh (2001) were mentioned, in a way that, when assuming $c_1 = 0$ and $c_2 \to \infty$, $\hat{\pi}$ will correspond to the minimum relative entropy estimator which is identical to the maximum likelihood estimator (MLE) of the binomial model. The authors also point out that the estimates become more robust assuming $c_1 < c_2 = 1$. Determination of the values that will guarantee better accuracy and precision is still a matter of study. Ruckstuhl & Welsh (2001) have mentioned that when f_n is a finite-sample realization of a binomial distribution and $c_1 < c_2 = 1$, the E-estimator may be biased and the substitution of c_2 by another value will be discussed in future work.

As a result of the above motivation, the present work aims to construct a new estimator in order to estimate the probability of success of a binomial distribution given a zero-inflated sample.

The main advantage provided by this estimator is highlighted in the computational aspect once the estimator (4) that minimizes (2) is obtained assuming infinite values belonging to the interval [0,1]. Therefore, we understand that a problem of a continuous nature that is treated in a discretized way, depending on the algorithm to be used, or even, in an application with real data may occasionally cause a non-fitted estimate.

Due to this fact, the estimator proposed in this work is shown in Section 2. Basically with of a modification in (3), in such manner that the researcher may fix an only value for the constants c_1 and c_2, based on a single point represented by the maximum likelihood estimator and not on every point of the parametric dominium, as suggested in (4).

2. Methodology

The binomial samples with different zero percentages (γ) were generated via Monte Carlo method according to the zero-inflated binomial model (ZIB):

$$f(Y=y)=\begin{cases}\gamma+(1-\gamma)(1-\pi)^m, y=0\\ (1-\gamma)\begin{pmatrix} m \\ y \end{pmatrix}\pi^y(1-\pi)^{m-y}, y=1,2,\ldots,m\end{cases} \tag{5}$$

According to the model mentioned and because it is an empirical study, arbitrarily, the parametric values were determined deliberately to represent different situations of sample of sizes ($n = 20$, 30, 50, 70, 90), extracted from a population of $m = 100$ elements, zero percentages ($\gamma = 0.2$, 0.5 and 0.7) and parametric values at "small", "medium" and "large" proportions ($\pi = 0.3$, 0.5, 0.8). Therefore, the resultant combination from these factors provided different configurations, in which the estimator $Pzib$ was evaluated by 10000 Monte Carlo simulations using the `R` software (R Development Core Team 2013).

The disparity function as defined by Lindsay (1994), such discrepancy between the data and the model density, based on the function $G(\cdot)$ and considering the sample space as $\Omega = \{0, 1, 2, \ldots\}$, is given by

$$\rho_G\left(d, f_\theta\right) = \sum_{x\in\Omega} G\left(\delta\left(x\right)\right) f_\theta\left(x\right)$$

where $G(\cdot)$ is a thrice differentiable convex function on $[-1, \infty)$ with $G(0) = 0$ and δ the Pearson residual at x, given by

$$\delta\left(x\right) = \frac{d\left(x\right) - f_\theta\left(x\right)}{f_\theta\left(x\right)}$$

with $f_\theta\left(x\right)$ representing a density function and $d\left(x\right)$ the empirical density at x.

According to Park, Basu & Lindsay (2002), the range of the Pearson residual is $[-1, \infty)$ and under certain regularity conditions, all minimum disparity estimators are first order efficient; in addition many of them have attractive robustness properties.

Due to these considerations, based on a function belonging to the Hölder class, a new estimator was constructed, modifying the function $\rho(x)$, given in (3).

The construction of the estimator denominated $Pzib$ was based on a modification in the disparity function $\mathrm{H}(\pi, f_n)$, given in (2), in order to reduce the speed $\rho(x) = xln(x)$ tends towards the infinite. This modification was made considering the following definitions:

Definition 1. We say that a function $f : X \subset \Re \to \Re$ is a Lipschitz continuous function in X if there is L>0 so that

$$|f(x_0) - f(x_1)| < L\,|x_0 - x_1|, \ \forall\ x_0, x_1 \in X.$$

Definition 2. We say that a function $f : X \subset \Re \to \Re$ is Hölder continuous with exponent $0< \alpha <1$ if there is H>0 so that

$$|f(x_0) - f(x_1)| \ < \ H\,|x_0 - x_1|^{\alpha}\,, \ \forall \ x_0, x_1 \ \in \ X.$$

Example 1. Consider the function:

$$\begin{array}{c} f : (0, +\infty) \ \subset \Re \to \Re \\ x \mapsto \ f(x) = \sqrt{x} \end{array} \tag{6}$$

The function described in (6) is Hölder continuous with exponent $\alpha = \frac{1}{2}$.

In fact:

For $x_0 > x_1$ we have:

$$\begin{array}{rl} \sqrt{x_0} = \sqrt{x_0 - x_1 + x_1} \ \leq \ \sqrt{x_0 - x_1} + \sqrt{x_1} = \sqrt{|x_0 - x_1|} + \sqrt{x_1} & \\ \Rightarrow \ \sqrt{x_0} - \sqrt{x_1} \ \leq \ \sqrt{|x_0 - x_1|} & (*) \end{array}$$

For $x_1 > x_0$ we have:

$$\begin{array}{rl} \sqrt{x_1} = \sqrt{x_1 - x_0 + x_0} \ \leq \ \sqrt{x_1 - x_0} + \sqrt{x_0} = \sqrt{|x_1 - x_0|} + \sqrt{x_0} & \\ \Rightarrow \ \sqrt{x_1} - \sqrt{x_0} \ \leq \ \sqrt{|x_1 - x_0|} & (**) \end{array}$$

In (*) and (**) we conclude that:

$$|\sqrt{x_0} - \sqrt{x_1}| \ \leq \ \sqrt{|x_0 - x_1|}, \text{ that is } |f(x_0) - f(x_1)| \ \leq \ |x_0 - x_1|^{\frac{1}{2}}$$

In a general way, given $0< \alpha <1$ the function

$$\begin{array}{c} f : (0, +\infty) \ \subset \Re \to \Re \\ x \mapsto \ f(x) = x^{\alpha} \end{array} \tag{7}$$

is Hölder continuous with exponent α (Begehr 1994).

The property below will be important for the following:
Property 1: Given $0 < \alpha < \beta \leqslant 1$, x^{α} tends to infinite when x tends to infinite less rapidly than x^{β}. We note $\beta - \alpha >0$ and therefore

$$\lim_{x \to \infty} \frac{\mathrm{x}^{\beta}}{\mathrm{x}^{\alpha}} = \lim_{x \to \infty} \mathrm{x}^{\beta - \alpha} = \infty \tag{8}$$

In order to reduce the speed, the function, $\mathrm{g}\,(x) \ = \ x \ln(x)$ tends to infinite when x tends to infinite, the proposal of the $Pzib$ estimator implies modifying the function $\rho(x)$ given in 3, in order to reduce its growth when x tends to infinite.

Define:

$$\rho_1(x) = \begin{cases} \begin{cases} \left\{ c_1^{1-\alpha} \ln(c_1) + [(1-\alpha)\ln(c_1)+1] \dfrac{c_1^{1-\alpha}}{\alpha} \right\} x^{\alpha} \\ - [(1-\alpha)\ln(c_1)+1] \dfrac{c_1}{\alpha} \end{cases} & \text{if } x < c_1 \\ x \ln(x) & \text{if } c_1 \leqslant x \leqslant c_2 \\ \begin{cases} \left\{ c_2^{1-\alpha} \ln(c_2) + [(1-\alpha)\ln(c_2)+1] \dfrac{c_2^{1-\alpha}}{\alpha} \right\} x^{\alpha} \\ - [(1-\alpha)\ln(c_2)+1] \dfrac{c_2}{\alpha} \end{cases} & \text{if } x > c_2 \end{cases} \quad (9)$$

$$\{x$$

Note that for $\alpha = 1$, $\rho_1(x)$ is equivalent to $\rho(x)$ given in (3).

We point out that the function $\rho_1(x)$ has the same kind of differentiability as the function $\rho(x)$, and $\rho(x)$ and $\rho_1(x) \in C^1(\Re^+)$. Based on the foregoing, the constructed estimator $Pzib$ is given by:

$$Pzib = \sum_{y=0}^{m} \rho_1(x) p_{\hat{\pi}_{mle}}(Y = y) \quad (10)$$

Note that this estimator will depend only on α parameter, once the constants c_1 and c_2 were fixed in 0.1 and 1 , which makes it accurate with a tolerable margin of error defined by $|\hat{P}zib - \hat{\pi}_{mle}|$ where k indicates the tolerable value for this difference, being interpreted as a deviation resulting from the incorporation of the extra-binomial variability in the estimation of π when compared with the maximum likelihood estimator.

3. Results and Discussion

As we know, the usual maximum likelihood estimator submitted to zero-inflated samples may present gross errors along with the $\hat{P}zib$ estimate, the researcher may use it as a reference of Table consulting, since through the deviation it is possible to verify error magnitude.

Starting from a situation where the fitting of $Pzib$ estimator is verified only for a simple value of c_1, maintaining $c_2 = 1$, as well as the study of the function $\rho_1(x)$ in relation to the speed of convergence when $x \to \infty$, the results found in the Tables 1-3 correspond to the maximum likelihood estimates, $\hat{P}zib$ and deviations.

The results shown in Table 1 were obtained from a study via Monte Carlo method which goal was to verify whether the choice of α could be made correctly with the knowledge of $\hat{\pi}_{mle}$. Fixing $k = 0.20$ for this purpose, we noted that the values of deviation satisfied the tolerable margin of error. Naturally, $\hat{\pi}_{mle}$ may be obtained given a zero-inflated sample. Therefore, taking this estimate

as a reference, the $\hat{P}zib$ estimate within the tolerable margin of error will provide information to verify whether the assumed value of α is in fact the value to be used in the $Pzib$ estimator to render accurate estimates. In this context, the results found in Table 1, confirm that given the low concentration of zeros ($\gamma = 0.20$) and different samples of size (n), the estimates of $\hat{P}zib$ were considered reasonable. Notice that the deviations obtained were inferior to $k = 0.20$ and the estimates $\hat{\pi}_{mle}$ were not so accurate when compared to the estimates obtained in the $Pzib$ estimator.

TABLE 1: Values of α for $\hat{P}zib$ estimates with $\rho_1(x)$ approach minimizing the difference in relation to the parameter of reference π with proportion of zeros $\gamma = 0.2$, $m = 100$, c_1 fixed in 0.1 and c_2 fixed in 1.

n	Estimates	$\pi = 0.3$	$\pi = 0.5$	$\pi = 0.8$
	α	**0.2600**	**0.1700**	**0.1200**
20	$\hat{\pi}_{mle}$	0.2374	0.3958	0.6515
	$\hat{P}zib$	0.2897	0.5268	0.8106
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.0523	0.1310	0.1591
	α	**0.1900**	**0.1600**	**0.1200**
30	$\hat{\pi}_{mle}$	0.2418	0.3986	0.6479
	$\hat{P}zib$	0.2827	0.4955	0.8118
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.0409	0.0968	0.1639
	α	**0.1300**	**0.1300**	**0.1200**
50	$\hat{\pi}_{mle}$	0.2382	0.3990	0.6336
	$\hat{P}zib$	0.3063	0.5124	0.8156
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.0681	0.1135	0.1819
	α	**0.1000**	**0.1200**	**0.1200**
70	$\hat{\pi}_{mle}$	0.2411	0.4027	0.6419
	$\hat{P}zib$	0.2957	0.4621	0.7958
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.0546	0.0594	0.1540
	α	**0.0900**	**0.1100**	**0.1200**
90	$\hat{\pi}_{mle}$	0.2387	0.3993	0.6387
	$\hat{P}zib$	0.2985	0.4895	0.7803
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.0598	0.0902	0.1417

Increasing the concentration of zeros to $\gamma = 0.50$, given the searched values of α the results shown in Table 2 provided fitted estimates. However, we can note that in a general way, the $\hat{\pi}_{mle}$ estimates resulted in inappropriate values, leading to an increase of k value in the tolerable margin of error. For high concentrations of zero (Table 3), the same behavior of the $\hat{\pi}_{mle}$ estimates in relation to the estimator $Pzib$ was observed.

According to Ruckstuhl & Welsh (2001) The choice $c_2 = 1$ gives improved first order robustness against gross error contamination and the choice $c_l = 1$ gives improved robustness against truncation. Under the binomial model the asymptotic distribution of the E-estimator is Gaussian for $c_1 < 1 < c_2$ and non-Gaussian for $c1 < c2 = 1$.

Choosing $c1 < c2$ we cannot treat both types of contamination simultaneously. The study of the properties for other c_2 values will be covered in future work.

TABLE 2: Values of α for $\hat{P}zib$ estimates with $\rho_1(x)$ approach minimizing the difference in relation to the parameter of reference π with proportion of zeros $\gamma = 0.5$, $m = 100$, c_1 fixed in 0.1 and c_2 fixed in 1.

n	Estimates	$\pi = 0.3$	$\pi = 0.5$	$\pi = 0.8$
	α	**0.2700**	**0.1800**	**0.1300**
20	$\hat{\pi}_{mle}$	0.1521	0.2537	0.3895
	$\hat{P}zib$	0.2800	0.5012	0.7720
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1279	0.2475	0.3825
	α	**0.2600**	**0.1800**	**0.1300**
30	$\hat{\pi}_{mle}$	0.1471	0.2442	0.4008
	$\hat{P}zib$	0.2925	0.5001	0.7720
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1454	0.2558	0.3711
	α	**0.2500**	**0.1800**	**0.1300**
50	$\hat{\pi}_{mle}$	0.1492	0.2527	0.3922
	$\hat{P}zib$	0.3053	0.4979	0.7717
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1562	0.2452	0.3795
	α	**0.2500**	**0.1800**	**0.1300**
70	$\hat{\pi}_{mle}$	0.1494	0.2537	0.4084
	$\hat{P}zib$	0.2922	0.4986	0.7718
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1428	0.2449	0.3633
	α	**0.2400**	**0.1800**	**0.1300**
90	$\hat{\pi}_{mle}$	0.1478	0.2452	0.3981
	$\hat{P}zib$	0.3068	0.4969	0.7717
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1590	0.2517	0.3736

TABLE 3: Values of α for $\hat{P}zib$ estimates with $\rho_1(x)$ approach minimizing the difference in relation to the parameter of reference π with proportion of zeros $\gamma = 0.7$, $m = 100$, c_1 fixed in 0.1 and c_2 fixed in 1.

n	Estimates	$\pi = 0.3$	$\pi = 0.5$	$\pi = 0.8$
	α	**0.2600**	**0.1800**	**0.1300**
20	$\hat{\pi}_{mle}$	0.0949	0.1484	0.2485
	$\hat{P}zib$	0.2935	0.4995	0.7725
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1987	0.3511	0.5240
	α	**0.2600**	**0.1800**	**0.1300**
30	$\hat{\pi}_{mle}$	0.0900	0.1510	0.2343
	$\hat{P}zib$	0.2843	0.4945	0.7721
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.1943	0.3435	0.5378
	α	**0.2500**	**0.1800**	**0.1300**
50	$\hat{\pi}_{mle}$	0.0882	0.1452	0.2367
	$\hat{P}zib$	0.3021	0.4936	0.7717
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.2139	0.3484	0.5350
	α	**0.2500**	**0.1800**	**0.1300**
70	$\hat{\pi}_{mle}$	0.0890	0.1500	0.2468
	$\hat{P}zib$	0.2991	0.4936	0.7717
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.2101	0.3436	0.5250
	α	**0.2500**	**0.1800**	**0.1300**
90	$\hat{\pi}_{mle}$	0.0885	0.1506	0.2349
	$\hat{P}zib$	0.2981	0.4936	0.7717
	$\lvert\hat{P}zib - \hat{\pi}_{mle}\rvert$	0.2096	0.3430	0.5369

It must be noted that other values of c_1 and c_2 may be obtained in the function described in the Appendix, constructed in the `R` software (R Development Core Team 2013).

4. An Ilustrative Example

For didactic purposes, is shown, where the application of the $Pzib$ estimator considering the $\rho_1(x)$ in a sample of size n, each sampling unit was considered independent and identically distributed with a binomial (m, π). Given this specification and for comparison with a simulated sample from an inflated binomial model with $m = 100$ and $\pi = 0.3$ (Table 4).

TABLE 4: Values assumed to illustrate the estimation of π using the estimator $Pzib$.

Parameter	m	n	γ	Sample Units
$\pi = 0.3$	100	20	0.20	25, 30, 31, 25, 24, 27, 0, 26, 31, 32, 26, 28, 21, 27, 29, 30, 0, 0, 28, 32

Obtaining the maximum likelihood estimate:

$$\hat{\pi}_{mle} = 1/m \sum_{y=0}^{m} y f_n(y)$$

and

$$f_n(y) = n^{-1} \sum_{i=1}^{n} I(Y = y_i), y = 0, \ldots, m$$

$$f_n(0) = \frac{I(Y=0) + I(Y=0) + \cdots + I(Y=0)}{20} = \frac{3}{20} = 0.15$$

$$\vdots$$

$$f_n(32) = \frac{I(Y=32) + I(Y=32) + \cdots + I(Y=32)}{20} = \frac{2}{20} = 0.10$$

Thus the maximum likelihood estimator is given by

$$\hat{\pi}_{mle} = \frac{0 f_n(0) + 1 f_n(1) + \cdots + 100 f_n(100)}{100} = 0.236$$

Based on this estimate, the Table 1 can be used, to seek a value of π_{mle} near 0.236. Obeying the rule that deviations must be less than 0.20, provides an idea of parameter estimates, which could be 0.30 with α equal to 0.26.

Calculating probabilities considering the mle:

$$p_{\hat{\pi}_{mle}}(0) = \left(\begin{array}{c} 100 \\ 0 \end{array} \right) 0.236^0(1-0.236)^{100-0} = 2.04 \times 10^{-12}$$

$$p_{\hat{\pi}_{mle}}(1) = \left(\begin{array}{c} 100 \\ 1 \end{array} \right) 0.236^1(1-0.236)^{100-1} = 6.30 \times 10^{-11}$$

$$p_{\hat{\pi}_{mle}}(2) = \left(\begin{array}{c} 100 \\ 2 \end{array} \right) 0.236^2(1-0.236)^{100-2} = 9.63 \times 10^{-10}$$

$$\vdots$$

$$p_{\hat{\pi}_{mle}}(100) = \left(\begin{array}{c} 100 \\ 100 \end{array} \right) 0.236^{100}(1-0.236)^{100-100} = 1.96 \times 10^{-63}$$

Calculating the Estimator $Pzib$:

$$Pzib = \sum_{y=0}^{m} \rho_1(x) p_{\hat{\pi}_{mle}}(Y = y) \,\mathrm{e}\, x = \frac{f_n(y)}{p_{\hat{\pi}_{mle}}(y)}$$

Table 5: Values for $Pzib$.

m	$p_{\hat{\pi}_{mle}}$	x	$\rho_1(x)$	$\rho_1(x)p_{\hat{\pi}_{mle}}$
0	2.0386x10^{-12}	7.3579x10^{10}	2568.8177	5.2368x10^{09}
1	6.2970x10^{-11}	0	0.2707	1.7049x10^{11}
2	9.6280x10^{-10}	0	0.2707	2.6069x10^{10}
3	9.7163x10^{-09}	0	0.2707	2.6305x10^{09}
4	7.2783x10^{-08}	0	0.2707	1.9705x10^{08}
5	4.3167x10^{-07}	0	0.2707	1.1686x10^{07}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
21	0.0804	0.6214	-0.2956	-0.0237
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
100	1.9552x10^{-63}	0	0.2707	5.2935x10^{-64}
			$Pzib = \sum_{y=0}^{m} \rho_1(x) p_{\hat{\pi}_{MLE}}$	0.2900

Comparing the estimated $\hat{P}zib$ with the parametric value ($\pi = 0.30$), note that the value of α used resulted in an accurate estimate, following the criteria established by $|\hat{P}zib - \hat{\pi}_{mle}| < 0.20$ shows that the value α is suitable for performing this inference.

5. Considerations Regarding the Use of RAF (Residual Adjustment Function)

Park et al. (2002) develop another graphical representation to summarize the behavior of the minimum disparity estimators in relation to maximum likelihood.

For this purpose, the authors developed a function called RAF (residual adjustment function), which considers as input the Pearson residual, represented by $\delta(x)$.

Referring to this methodology, taking ρ_1 function (9) as a function of δ, redefined by $\delta = f_n(y)/p_\pi(y) - 1$ instead of $x = f_n(y)/p_\pi(y)$, the graphic procedure RAF used to evaluate the disparity is not suitable, since Park et al. (2002) mention a graphical interpretation of the robustness of the estimators, but this representation is not completely satisfactory since the domain of the RAF is infinite. With emphasis on the domain of ρ_1 function (9), we do not recommend the use of this procedure for the following reasons:

- The ρ_1 function (9) proposed for obtaining the estimator allows researchers to search by simulation, the value of α that defines ρ_1 function (9). That is, each value of α has corresponding a ρ_1 function (9). Therefore, we note a limitation in assuming $x < 0$, since x^α cannot be calculated. Such statement may be conjectured if we consider $\alpha = 1/2$, which implies that x^α can only be calculated for greater than or equal to zero values.

- Another point to be emphasized, refers to the fact that ρ_1 function (9), considered in its construction $\ln(x)$ with the R^+ domain, so we cannot allow negative or zero values in the evaluation.

6. Conclusions

The $Pzib$ estimator proposed in this work, fit the situations where the samples presented low ($\gamma = 0.20$) and medium ($\gamma = 0.50$) concentrations of zero. The accuracy and precision of the $\hat{P}zib$ estimates are flexible to computational improvement, adopting the criterion $|\hat{P}zib - \pi_{mle}| < k$, where k corresponds to a tolerable value subjectively specified by the researcher.

Acknowledgements

The authors would like to thank: FAPEMIG, CNPq and Capes for financial support; Paulo Leandro Dattori da Silva, "livre-docente" of ICMC - USP for the assistance with this article and the reviewers for suggestions and contributions given in the publication of this paper.

[Received: December 2013 — Accepted: October 2014]

References

Achcar, J. A. & Junqueira, J. G. (2002), 'Extra-binomial variability: A Bayesian approach', *Journal of Statistical Research* **36**, 1–14.

Basu, A., Shiyoa, H. & Park, C. (2011), *Statistical Inference: The Minimum Distance Approach*, Chapman and Hall.

Begehr, H. G. W. (1994), *Complex Analytic Methods for Partial Differential Equations: An Introductory Text*, World Scientific, Singapore.

Hinde, S. & Demetrio, G. G. B. (1978), 'Overdispersion models and estimation', *Computational Statistics & Data Analysis* **34**, 69–76.

Huber, P. (1964), 'Robust estimation of a location paramenter.', *Annals of Mathematical Statistics* **35**, 73–101.

Kupper, L. L. & Haseman, J. K. (1998), 'The use of a correlated binomial model for the analysis of certain toxicological experiments', *Biometrics* **27**, 151–170.

Lindsay, B. G. (1994), 'Efficiency versus robustness: The case for minimum Hellinger distance and related methods', *The Annals of Statistics* **22**, 1081–1114.

Park, C., Basu, A. & Lindsay, B. (2002), 'The residual adjustment function and weighted likelihood: a graphical interpretation of robustness of minimum disparity estimators', *Computational Statistics and Data Analysis* **39**, 21–33.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Ruckstuhl, A. F. & Welsh, A. H. (2001), 'Robust fitting of the binomial model', *The Annals of Statistics* **29**, 1117–1136.

Silva, A. M. & Cirillo, M. A. (2010), 'Estudo por simulação Monte Carlo de um estimador robusto utilizado na inferência de um modelo binomial contaminado', *Acta Scientiarum. Technology* **32**, 303–307.

Simpson, D. G. (1987), 'Minimum Hellinger distance estimation for the analysis of count data', *Journal of the American Statistical Association* **82**, 802–807.

Simpson, D. G., Carrol, R. J. & Ruppert, D. (1987), 'M-estimation for discrete data: Asymptoptic distribution theory and implications', *The Annals of Statistics* **15**(2), 657–669.

Appendix. R Function to Compute the Simulations

```
# definition of the fn(y) function
#m: Population size
#n: Sample size
fny<-function(m,n,data,vet)
{
for (a in 1:(m))
{
prop=0; aux=vet[a]
for (b in 1:n)
{
if (aux==data[b]) prop=prop+1
}
vcont[a]=(prop)/n
}
return(vcont)
}
# definition of the Pzib function
#p: probability of success
#c1,c2,alfa: constants to be inserted.
estimaPzib<-function(x,p,c1,c2,alfa)
{
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))*x[b]^u)
-(((1-u)*log(c1)+1)*c1/u)
if(x[b]>c2)               rho[b]=((c2^(1-u)*log(c2)
+((1-u)*log(c2)+1)
*(c2^(1-u)/u))*x[b]^u)
-(((1-u)*log(c2)+1)*c2/u)
auxPzib=rho[b]*p[b]
estPzib=auxPzib + estPzib
}
return (estPzib)
}
```