Open Access Library Journal
2018, Volume 5, e4735
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Application of the Power Series Probability Distributions for the Analysis of Zero-Inflated Insect Count Data

Remi Mrume Sakia

Department of Statistics, University of Botswana, Gaborone, Botswana
Email: sakiar@ub.ac.bw

How to cite this paper: Sakia, R.M. (2018) Application of the Power Series Probability Distributions for the Analysis of Zero-Inflated Insect Count Data. *Open Access Library Journal*, **5**: e4735.
https://doi.org/10.4236/oalib.1104735

Received: June 22, 2018
Accepted: October 20, 2018
Published: October 23, 2018

Abstract

Excess number of zeros (zero inflation, ZI) in count data is a common phenomenon which must be addressed in any analysis. The extra zeros may be a result of over-dispersion in the data. Ignoring zero-inflation can result in biased parameter estimates and standard errors. Over-dispersion is also associated with a zero-inflated data. Depending on the selected model, different results and conclusions may be reached. In this paper two commonly encountered models in count data are considered, namely, the Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) probability distributions. Emphasis is placed on the Maximum Likelihood (ML) estimation of the model parameters. Specifically of interest was to estimate the zero-inflation parameter and hence, the corrected frequencies. It was found that for the Poisson model, the zero-inflation parameter estimate was considerably higher than that from the Negative Binomial model. From the results however, it is suspected that the effectiveness of adjusting for the high number of zeros in both models might have been greatly affected by the inherent high variability between sites. It is then proposed that in future research, the problem of heterogeneity in count data be addressed before any further analysis.

Subject Areas

Statistics

Keywords

Count Data, Poisson, Negative Binomial, Over-Dispersion, Zero-Inflation

1. Introduction

In most cases, analysis of insect counts data has been modelled by the three dis-

tributions, namely Poisson (θ), Binomial (n, θ) and Negative Binomial (k, θ). These discrete distributions fall under the category of power series distributions. These distributions have been generalized to what is referred to as the generalized power series distribution (GPSD). By expressing these probability distributions explicitly in the form of power series distributions, this has greatly simplified the derivation of the explicit form of the moments of these distributions. The power series distributions have been widely studied. Noak [1] specifically investigated its moments and cumulants properties which were later extended by Khatri [2] to multivariate distributions. The properties of the generalized power series distributions were later studied by Patil [3] who also investigated estimation and other properties of these models. In more recent studies, a wider interest has been focused on modifications of the power series probability distributions to address the problem of zero-inflation, *i.e.* observing more zeroes than what would be expected for example, in a Poisson or Negative binomial distribution. Count data has been noted to consist of more zeroes than expected if the data were generated by a simple Poisson or negative Binomial process. Ignoring zero-inflation can have two consequences: bias may be induced in the estimation of model parameters and standard errors may be inflated thus leading to erroneous statistical inferences. In addition, zero-inflated data may as a consequence, cause over-dispersion which is defined as having a variance that may be larger than the mean. If this data characteristic is left unaddressed, this will render the model specification for the Poisson distribution of equal mean and variance null and void. Cameron and Trivedi [4] have outlined some common departures from the standard Poisson regression that should be addressed before actual analysis. Dean [5] outlined a method of testing for over-dispersion in Poisson and Binomial regression models. Lately however, the problem of zero inflation has been widely researched in the context of zero-inflated count regression. See for example, Lambert [6] and Gupta, *et al.* [7] [8]. Most of the cited research however, has implemented count regression models.

2. Power Series Distributions

A discrete random variable X will have a power series distribution given as:

$$P(X=x)=\frac{a_x\theta^x}{f(\theta)},\quad x=0,1,2,\cdots,\ a_x>0;\ \theta>0 \tag{1}$$

The distribution belongs to the exponential family of distributions and can generally be expressed in the form:

$$P(X=x)=\mathrm{e}^{\left[xa(\theta)+c(x)+g(\theta)\right]} \tag{2}$$

where a and g are functions of the unknown parameter θ and c is a function of x.

This property has been exploited in the derivation of the moments and other properties of the distribution. It can be shown (see Edwin [9]) that the central moments for the power distribution are given as:

$$E(X)=\frac{\theta f'(\theta)}{f(\theta)}$$
$$V(X)=\frac{\theta^2}{f(\theta)}f''(\theta)+\frac{\theta}{f(\theta)}f'(\theta)-\left[\frac{\theta}{f(\theta)}f'(\theta)\right]^2 \quad (3)$$

Table 1 gives a summary of some power series distributions represented in the form given in (1).

3. Zero-Inflation and Over-Dispersion

The Poisson distribution theoretically specifies that the mean and variance are equal. However, it is quite common to have data for which the variance is far larger than the mean and the phenomenon is referred to as over-dispersion. In this case, Poisson GLM has been used to correct the anomaly. See for example, Cameron and Trivedi [4]. Over-dispersion is also common in a Negative Binomial distribution. When the distribution is parameterized in terms of its mean μ and variance σ^2 distribution, the mean and variance are given as;

$$E(X)=\mu \text{ and } \sigma^2=\mu+\frac{\mu^2}{k} \quad (4)$$

The distribution is reduced to equi-distribution as$k$ becomes large, implying convergence to the Poisson distribution. As k becomes small for a small μ, a zero-inflated Negative Binomial distribution is a consequence.

A zero-inflated statistical model is based on a zero-inflated probability distribution. It arises when probability mass at a point zero exceeds the one allowed under the particular family of distributions. These models have been widely studied. See for example, Jasankul [10], Ridout *et al.* [11] and Bohning [12]. They provide one method to explain the excess zeroes by modelling the data as a mixture of two separate distributions in which one is typically a Poisson or Negative Binomial distribution that can generate both zeroes and non-zero counts, and the second distribution is a constant distribution that generates only zero counts.

4. The Zero-Inflated Distribution

Count data that have an incidence of zeroes greater than expected for the underlying probability distribution is modelled as:

Table 1. Specification of some common power series distributions.

$Poisson(\theta)$	$\frac{1}{x!}$	θ	e^{θ}
$Binomial(n,\theta)$	$\binom{n}{x}$	$\frac{\theta}{1+\theta}$	$(1+\theta)^n$
$Logarithmic(\theta)$	$\frac{1}{x}$	θ	$-\ln(1-\theta)$
$Neg.\ Binomial$	$\binom{k+x-1}{x}$	θ	$(1-\theta)^{-k}$

$$P(X=x)=\begin{cases}\rho+(1-\rho)P(X=0) & \text{for } x=0\\ (1-\rho)P(X=x) & \text{for } x=1,2,\cdots\end{cases} \quad (5)$$

where $0<\rho<1$ is the zero-inflation parameter. $P(X=x)$ in this case represents any one of the count data distributions, e.g. Poisson, Negative Binomial, etc. using the format presented in (1). The means and variances of these distributions may then be obtained by using the expressions in (3) for the respective probability model. For example, for the Poisson (λ) distribution;

$$P(X=x)=\begin{cases}\rho+(1-\rho)\mathrm{e}^{-\theta}; & \text{for } x=0\\ (1-\rho)\dfrac{\mathrm{e}^{-\theta}\theta^{x}}{x!}; & \text{for } x=1,2,\cdots\end{cases} \quad (6)$$

And for a Negative Binomial distribution;

$$P(X=x)=\begin{cases}\rho+(1-\rho)p^{k}; & \text{for } x=0\\ (1-\rho)\binom{k+x-1}{x}p^{k}(1-p)^{x}; & \text{for } x=1,2,\cdots\end{cases} \quad (7)$$

Using (5) and (1) and applying the usual definitions for $E(X)$ and $V(X)$, the mean and variance of a zero-inflated distribution is given as:

$$\begin{aligned}E(X)&=(1-\rho)\theta\frac{f'(\theta)}{f(\theta)}\\ V(X)&=(1-\rho)\theta\left\{\theta\frac{f''(\theta)}{f(\theta)}+\frac{f'(\theta)}{f(\theta)}-(1-\rho)\theta\left[\frac{f'(\theta)}{f(\theta)}\right]^{2}\right\}\end{aligned} \quad (8)$$

where $f(\theta)$ is given in Table 1 for the respective probability distribution.

5. Estimation of Model Parameters

In estimating the zero inflated parameters, some three methods seem to have been prominently used. The methods of moments (MM) is said to provide estimates which are not very accurate. Nanjundan and Naika [13], [14] applied the MM as well as the maximum likelihood (ML) estimation and concluded that both methods are asymptotically equally efficient. In some cases however, the MM is considered as appropriate for use as initial values for ML estimation in the case of a non-closed solution of the model parameters. The two methods are outlined below:

5.1. Method of Moments Estimation

Let $X_1, X_2, \cdots, X_n$ be a random sample from the probability distribution of the form in (5) for a p-parameter ZI model, the estimates are obtained by solving the equations:

$$E(X^{p})=\frac{\sum X_i^{p}}{n} \quad \text{for } i=1,2,\cdots,p \quad (9)$$

For a two-parameter ZI model, Equation (9) leads to the following two equa-

tions:

$$\begin{aligned}(1-\rho)\theta\frac{f'(\theta)}{f(\theta)} &= \bar{x} \\ (1-\rho)\theta\left\{\theta\frac{f''(\theta)}{f(\theta)}+\frac{f'(\theta)}{f(\theta)}\right\} &= \frac{\sum x_i^2}{n}\end{aligned} \tag{10}$$

Equation (10) may then be solved simultaneously to obtain estimates for ρ and θ.

5.2. Maximum Likelihood Estimation

The likelihood function for a ZI model given in Equation (5) may be written as:

$$L(\theta,\rho;\underline{x}) = \prod_{i=1}^{n}\left\{\rho+(1-\rho)\frac{a_0}{f(\theta)}\right\}^{1-\pi_i}\left\{(1-\rho)\frac{a_{x_i}\theta^{x_i}}{f(\theta)}\right\}^{\pi_i} \tag{11}$$

where $x_i = 1,2,\cdots$ and $\tau_i = \begin{cases}0 & \text{if } x_i = 0\\ 1 & \text{if } x_i \neq 0\end{cases}$.

The log-likelihood function is then given as:

$$\ell = n_0 \ln\left\{\rho+(1-\rho)\frac{a_0}{f(\theta)}\right\}+\sum_i \tau_i \ln(1-\rho)+\sum_i \tau_i \ln a_{x_i}+\sum_i \tau_i x_i \ln\theta-\sum_i \tau_i \ln f(\theta) \tag{12}$$

where n_0 is the number of zeros in the observed sample. Solving $\frac{\partial \ell}{\partial \rho}=0$ and $\frac{\partial \ell}{\partial \theta}=0$ for θ and ρ, the following expressions are obtained after some simplification:

$$\hat{\rho} = \frac{n_0 f(\hat{\theta}) - n a_0}{n\left(f(\hat{\theta})-a_0\right)} \tag{13}$$

$$\bar{x} = \frac{\hat{\theta} f'(\hat{\theta})}{f(\hat{\theta})-a_0} \tag{14}$$

The Newton-Raphson iterative method usually provide a solution of the form $f(\theta)=0$. For an initial value of θ_n, the next estimate is given as

$$\theta_{n+1} = \theta_n - \frac{f(\theta_n)}{f'(\theta_n)}.$$

Applying the Newton-Raphson iterative method to Equation (14) and using the MM estimates from Equation (10) as initial values then the improved estimate of θ is:

$$\hat{\theta}_{r+1} = \bar{x}\left\{\frac{f(\hat{\theta}_r)-a_0}{f'(\hat{\theta}_r)}\right\} \tag{15}$$

Finally, substituting the estimate $\hat{\theta}$ from Equation (15) into Equation (13),

an estimate of ρ is obtained. We shall now apply the estimation procedures outlined above to two probability distributions, namely ZIP and ZINB as given in Equation (6) and Equation (7) respectively.

5.3. Fitting the ZI-Poisson Model

For the ZI-Poisson distribution, the corresponding function $f(\theta)$ from Table 1 is substituted in Equation (10) and solved simultaneously to obtain the closed form MM estimates for ρ and θ as:

$$\hat{\theta} = \frac{\sum_i f_i x_i^2}{n\bar{x}} - 1 \tag{16}$$

$$\hat{\rho} = 1 - \frac{n\bar{x}^2}{\sum_i f_i x_i^2 - n\bar{x}} \tag{17}$$

Likewise, the ML estimates for ZIP model are obtained by substituting $f(\theta)$ in Equation (13)and Equation (14) resulting into:

$$\bar{x}\left(e^{\hat{\theta}} - 1\right) = \hat{\theta}e^{\hat{\theta}} \tag{18}$$

$$\hat{\rho} = \frac{n_0 e^{\hat{\theta}} - n}{n\left(e^{\hat{\theta}} - 1\right)} \tag{19}$$

Using the MM estimate for θ as the initial value, Equation (18) is solved iteratively and the estimate $\hat{\theta}$ is subsequently substituted in Equation (19) to obtain $\hat{\rho}$.

5.4. Fitting the ZI-Negative Binomial Model

To analyse the data using this model, we need to obtain a pooled estimate of k. Anscombe [15] has indicated that fitting a negative binomial distribution with a common value of k to sets of counts on the same species is a reasonable procedure. He then proposed such an estimate to be obtained from Equation (4) by substituting μ and σ^2 with the sample values $\bar{x}$ and s^2 and solve for k. For the count data the estimate of k is 0.12.

Again the corresponding $f(\theta)$ for the Negative binomial distribution from Table 1 is substituted in Equation (10) and solved simultaneously for $\hat{\theta}$ and $\hat{\rho}$ to obtain the MM estimates as:

$$\hat{\theta} = \frac{\sum_i f_i x_i^2 - n\bar{x}}{kn\bar{x} + \sum_i f_i x_i^2} \tag{20}$$

$$\hat{\rho} = 1 - \frac{\bar{x}\left(kn\bar{x} + \sum_i f_i x_i^2\right)}{k\left(\sum_i f_i x_i^2 - n\bar{x}\right)} + \frac{\bar{x}}{k} \tag{21}$$

Similarly, for the ML estimates, we substitute the respective $f(\theta)$ from Ta-

ble 1 in (13) and solve to obtain the following estimates:

$$\overline{x} = \frac{k\hat{\theta}\left(1-\hat{\theta}\right)^{-k-1}}{\left(1-\hat{\theta}\right)^{-k}-1} \tag{22}$$

$$\hat{\rho} = \frac{n_0\left(1-\hat{\theta}\right)^{-k}-n}{n\left\{\left(1-\hat{\theta}\right)^{-k}-1\right\}} \tag{23}$$

Equation (22) is solved iteratively for $\hat{\theta}$ and substituted in (23) to obtain $\hat{\rho}$.

6. Application

Data Description

The data to be used are some counts of eggs of ***Aphis fabae*** made by Dr. D. P. Jones in the course of a survey of the Eastern counties of England in 1947 and reproduced by Anscombe [15]. In this paper, the data has been sorted into a slightly different frequency distribution form but the original data remains the same. Ninety four hedgerow spindle sites were visited, that had been cut down the previous winter, so that the shoots were of one year growth. At each site ten shoots were removed and the ***A. fabae*** eggs on them subsequently counted. The counts are given in Table 2.

The mean and variance of the egg counts is 5.29 and 235.01 respectively which points to a strong evidence of over-dispersion. A frequency distribution for the counts is given in Figure 1. The large spike at zero gives early warning of possible zero inflation. Therefore, the ZIP and ZINB distributions are to be fitted.

7. Results and Conclusions

Table 3 gives the estimated parameters for the count data on the basis of fitting

Table 2. Egg counts.

Eggs (X):	0	1	2	3	4	5	6	7	8	9
Frequency:	589	62	48	30	24	15	20	5	5	6
Eggs (X):	10	11	12	13	14	15	16	17	18	19
Frequency:	11	10	7	7	6	4	2	8	4	4
Eggs (X):	20	21	22	23	24	25	26	27	29	31
Frequency:	1	3	2	3	6	2	2	1	1	5
Eggs (X):	32	33	34	35	36	38	39	40	42	43
Frequency:	2	3	2	5	1	1	2	1	1	1_
Eggs (X):	44	45	47	48	49	50	51	52	58	59
Frequency:	2	1	1	2	1	2	1	1	1	2
Eggs (X):	65	66	70	82	83	84	105	110	120	123
Frequency:	1	1	2	1	1	1	1	1	2	1
Eggs (X):	148	163								
Frequency:	1	1								

Table 3. Parameter estimates.

	ZIP	Model	ZINB	Model
Method	$\hat{\theta}$	$\hat{\rho}$	$\hat{\theta}$	$\hat{\rho}$
MME	48.67	0.89	0.98	0.12
MLE	5.26	0.62	0.92	0.42

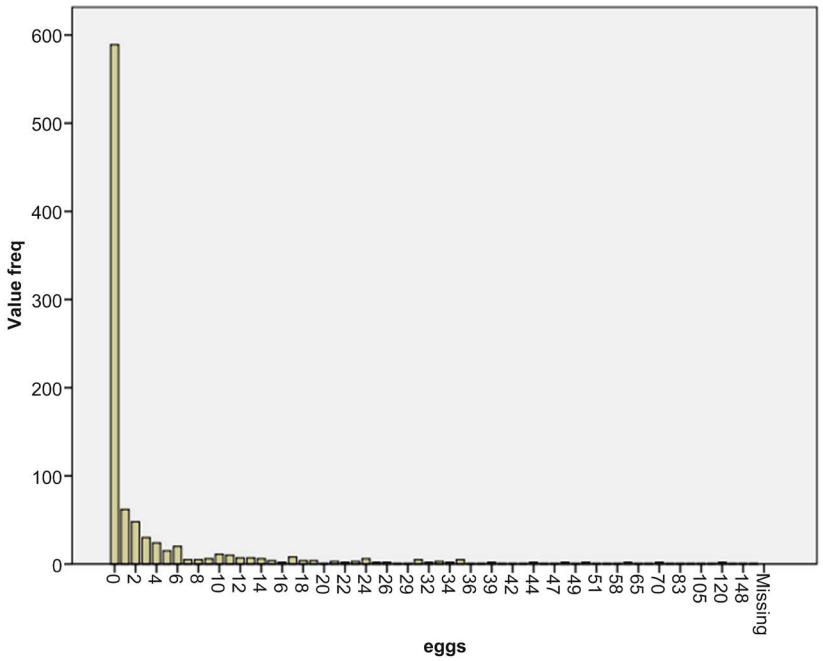

Figure 1. Frequency distribution.

the ZIP and ZINB distributions. It is noted that the estimated value of the exponent for ZINB model from the sample version of Equation (4) was $k = 0.12$ which in turn, resulted in a negative estimate of the inflation parameter ρ, that may be interpreted as a case of zero-deflation. This may have been a consequence of the observed heterogeneity between sites which was not addressed here and which ranged between 0 and 2032 within sites. As proposed by Anscombe [15], for a $k < 1$ a logarithmic transformation of the counts may be necessary. This problem may be addressed in a future analysis. However, the situation may not arise when count regression model is used.

It should be noted however, that possible design errors, such as sampling practices, might also have caused the excess zeros.

Below are charts representing the Poisson frequency distribution. **Figure 1** represents the observed frequencies of the count data which is highly positively skewed and **Figure 2** represents the Poisson expected frequencies which is centred at around five eggs. That means if we assume a Poisson distribution for this data, then we would expect approximately less than 10 zeros. **Figure 3** gives a plot of the ZIP fitted frequencies. The zero-inflation factor does not differ much from the original counts but overall, the counts are centred at five eggs. It is

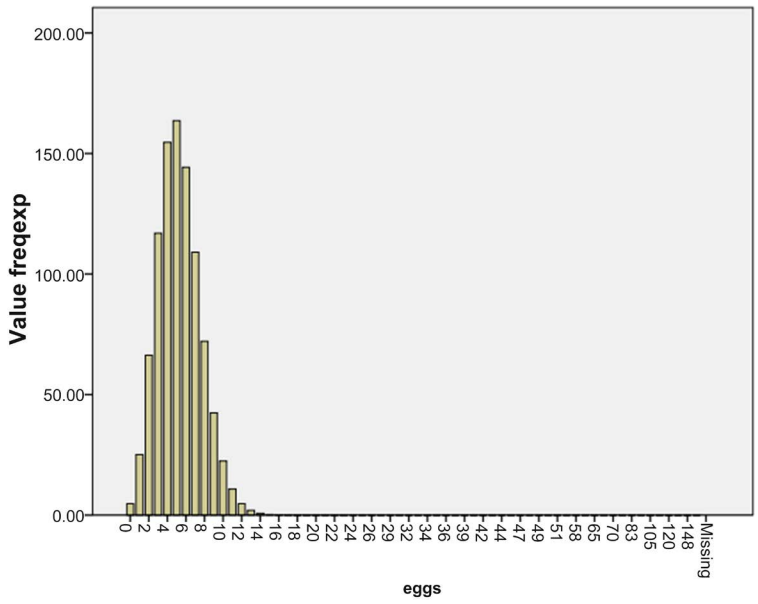

Figure 2. Expected frequencies.

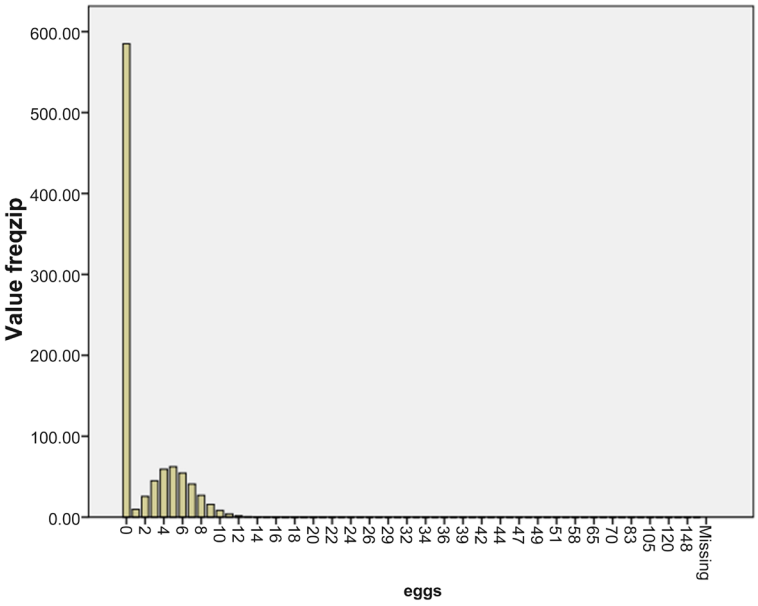

Figure 3. ZIP frequencies.

therefore recommended that before implementing procedures to address the zero-inflation, the inherent heterogeneity between sites be further examined.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Noak, A. (1950) A Class of Random Variable with Discrete Distribution. *Annals of*

the Institute of Statistical Mathematics, **21**, 127-132. https://doi.org/10.1214/aoms/1177729894

[2] Khatri, C.G. (1959) On Certain Properties of Power Series Distributions. *Biometrika*, **46**, 486-490. https://doi.org/10.1093/biomet/46.3-4.486

[3] Patil, M.M. (1962) On Certain Properties of the Generalized Power Series Distribution. *Annals of the Institute of Statistical Mathematics*, **14**, 179-182. https://doi.org/10.1007/BF02868639

[4] Cameron, A.C. and Trivedi, P.K. (2013) Regression Analysis of Count Data. 2nd Edition, Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139013567

[5] Dean, C.B. (1992) Testing for Over Dispersion in Poisson and Binomial Regression Models. *Journal of the American Statistical Association*, **87**, 451-457. https://doi.org/10.1080/01621459.1992.10475225

[6] Lambert, D. (1992) Zero. Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14. https://doi.org/10.2307/1269547

[7] Gupta, P.L. and Gupta, R.C. (1995) Inflated Modified Power Series Distribution with Applications. *Communication in Statistics-Theory and Methods*, **24**, 2355-2374. https://doi.org/10.1080/03610929508831621

[8] Gupta, P.L., Gupta, R.L. and Tripathi, R.C. (2004) Score Test for Zero Inflated Generalized Poisson Regression Model. *Communication in Statistics-Theory and Methods*, **33**, 47-64. https://doi.org/10.1081/STA-120026576

[9] Edwin, T.K. (2014) Power series Distributions and Zero Inflated Models. Unpublished Thesis, University of Nairobi, Nairobi.

[10] Jasankul, N. and Hinde, J.P. (2002) Score Tests for Zero Inflated Poisson Models. *Computational Statistics and Data Analysis*, **40**, 75-96. https://doi.org/10.1016/S0167-9473(01)00104-9

[11] Ridout, M., Hinde, J. and Demetrio, C.G.B. (2001) A Score Test for Testing Zero-Inflated Poisson Regression Model against Zero Inflated Negative Binomial Alternatives. *Biometrics*, **57**, 219-223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

[12] Bohning, D. (1998) Zero-Inflated Poisson Models and CAMAN. A Tutorial of Evidence. *Biometrical Journal*, **40**, 833-843. https://doi.org/10.1002/(SICI)1521-4036(199811)40:7<833::AID-BIMJ833>3.0.CO;2-O

[13] Nanjundan, G. and Naika, T.R. (2012) Asymptotic Comparison of Method of Moments Estimators and Maximum Likelihood Estimators of Parameters in Zero-Inflated Poisson Model. *Applied Mathematics*, **3**, 610-616. https://doi.org/10.4236/am.2012.36095

[14] Nanjundan, G. and Naika, T.R. (2013) Estimation of Parameters in a Zero-Inflated Power Series Model. *International Journal of Statistika and Mathematika*, **6**, 109-114.

[15] Anscombe, F.J. (1949) The Statistical Analysis of Insect Counts Based on the Negative Binomial Distribution. *Biometrics*, **5**, 165-173. https://doi.org/10.2307/3001918