

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/390943588>

Advances and Applications in Statistics ON THE ZERO AND k -INFLATED NEGATIVE BINOMIAL DISTRIBUTION WITH APPLICATIONS

Article in *Advances and Applications in Statistics* · May 2023

CITATION

1

READS

34

2 authors, including:



[Ian Jay Serra](#)

University of the Philippines Cebu

2 PUBLICATIONS 1 CITATION

SEE PROFILE



ON THE ZERO AND k -INFLATED NEGATIVE BINOMIAL DISTRIBUTION WITH APPLICATIONS

Ian Jay A. Serra* and Daisy Lou L. Polestico

Mathematics and Statistics Programs

University of the Philippines Cebu

Cebu City, Philippines

e-mail: iaserra@up.edu.ph

Department of Mathematics and Statistics

Mindanao State University - Iligan Institute of Technology

Iligan City, Philippines

e-mail: daisylou.polestico@g.msuiit.edu.ph

Abstract

In the literature, there are a significant number of studies on mixtures and compound probability distributions used for count data with inflated frequencies. This study extended some existing zero-inflated distributions, by considering the flexibility of peaks in the data with excessive counts other than zeros and handled an overdispersion in the data. Moreover, this study formulated a proposed zero- and k -inflated

Received: August 30, 2022; Revised: March 3, 2023; Accepted: May 18, 2023

2020 Mathematics Subject Classification: 62E10.

Keywords and phrases: inflated count data, overdispersion, Zk INB distribution, excessive counts, negative binomial distribution, count distributions.

*Corresponding author

How to cite this article: Ian Jay A. Serra and Daisy Lou L. Polestico, On the zero and k -inflated negative binomial distribution with applications, *Advances and Applications in Statistics* 88(1) (2023), 1-23. <http://dx.doi.org/10.17654/0972361723037>

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Published Online: May 30, 2023

negative binomial ($ZkINB$) distribution which is a mixture of a multinomial logistic and negative binomial distribution. The multinomial logistic component captures the occurrence of excessive counts, at zero and at $k > 0$, while the negative binomial component captures the counts that are assumed to follow a negative binomial distribution. The probability mass function (pmf) and the moment generating function (mgf) of the distribution were derived in order to compute some vital structural properties of the formulated distribution, such as the mean and the variance. Examples showed that the formulated $ZkINB$ seems to capture better distributions as compared with other existing distributions for inflated count data.

1. Introduction

Count data refers to the number of times an event occurs within a fixed period of time. It is encountered in different areas of research and in many practical problems. There is now a great deal of interest in the literature on investigating the relationship between a count variable and other variables. One of the commonly used distributions for analyzing the count data is the Poisson (POI) distribution which is known to have the same mean and variance. But in many applications, this restriction is violated due to the overdispersion in count data (Sakthivel et al. [21]). The most popular method of dealing with Poisson overdispersion is to analyze the data using a negative binomial (NB) distribution. The NB distribution has an extra parameter, the negative binomial dispersion parameter, which helps to detect the overdispersion or underdispersion in the population.

The cause of overdispersion is the presence of excess number of zero counts in the data, known as zero inflation. The POI and NB distributions both assume that the count data being modeled have zero counts. In fact, in some data sets, there may be more zeros occurring than expected by the underlying probability distribution. This is zero inflated, that is, the abundance of zeros is more than what is theoretically expected. Moreover, there have been many studies on the fitting of zero-inflated data; these models heavily depend on the special features of the individual data. In such

cases, the zero-inflated Poisson (ZIP) models are extensively studied in the literature. Arora and Chaganty [4] stipulated that the ZIP models have been shown to perform better than the traditional count models and have been applied across a wide spectrum of academic disciplines; including biology, ecology, psychology, education, economics; industries such as in manufacturing, transportation, and insurance; and recently, for monitoring social networks. The other ZIP-like models are; zero-inflated geometric (ZIG), zero-inflated binomial (ZIB), and zero-inflated negative binomial (ZINB).

The ZINB distribution has been widely used for count data analyses in various biomedical settings due to its capacity of modeling excess zeros and overdispersion. The NB distribution, of which the POI distribution is a limiting case, has a capacity of modeling overdispersion that accounts for heterogeneity of the incidence processes and thus is widely used in practice. In cases where the data exhibits overdispersion, the ZINB is preferred to the POI or ZIP since it is a model that is used to address the issue of overdispersion in count data with excess zeros (Odhiambo et al. [19]).

In addition to the presence of zero counts, some data sets may have inflated counts of additional value $k > 0$ (Hilbe [13]). This violation may be due to extra zeros not expected from a POI distribution. If this is ignored, biased estimates will result, defeating the purpose of analysis which is to properly model and predict future events. In many situations, besides zero, the frequency of another count k tends to be higher in the data. The zero- and k -inflated Poisson ($ZkIP$) distribution model is appropriate in such situations. The $ZkIP$ distribution essentially is a mixture distribution of Poisson and degenerate distributions at points zero and k (Arora and Chaganty [4]). While the $ZkIP$ model accounts for dispersion in count data due to the presence of two inflated count frequencies, there remains the potential presence of underlying dispersion that needs to be addressed. More broadly, data dispersion of varying types can exist depending on what data features are likewise addressed during data modelling (Sellers and Shmueli [22]). In most cases, the $ZkIP$ captures the inflation at zero and $k \in \mathbb{Z}^+$, $k \neq 0$, but

fails to capture the underlying (over) dispersion in the data. In this study, the existing zero-inflated distributions are extended through the flexibility of peaks in the data with excessive counts other than zeros, while accounting for overdispersion in a data set. Moreover, this study constructs a zero- and k -inflated NB distribution which is a mixture of multinomial logistic and NB distribution. The multinomial logistic component captures the occurrence of excessive counts, including zero and k , while the NB component captures the counts that are assumed to follow a NB distribution. Specifically, this study aimed to develop some vital structural properties of the $ZkINB$, such as the mean and the variance.

The main part of this study includes an exploration of the $ZkINB$ distribution. Its practical use will not only bring a significant improvement relative to the NB distribution, but also a wider flexibility due to its main properties, as for instance, its overdispersion. In the discussions, this study covered and only focused on the differences between the four inflated count distributions namely, the formulated $ZkINB$, $ZkIP$, $ZINB$, and ZIP distributions. Fitting more than one model to a given dataset is common to establish the best model for a given situation. To determine which distribution better captures the distribution of the data, the Absolute Error (ABE) is used to measure the differences. There are some caveats for the distribution. Firstly, the so-called “excessive value” can be subjectively considered by different researchers, but in this study, it is defined as the value having the larger frequencies in the data. Secondly, this study only considered two excessive values in the data, name zero- and a k -inflated value, $k \geq 1, k \in \mathbb{Z}^+$.

2. Methodology

A distribution that accounts for the inflated probability at zero is obtained by mixing the NB distribution with a point mass ϕ . Consider an experiment resulting to two processes as follows:

Case 1. With probability ϕ , $0 < \phi < 1$, the only response of the first process is zero counts.

Case 2. With probability $(1 - \phi)$, the response of the second process is governed by a negative binomial with mean $\mu \geq 0$.

Also, assume that the experiment is repeated independently a number of times. Assume that Case 1 occurs with a probability ϕ , and the corresponding Case 2 occurs with probability $(1 - \phi)$. Now, consider a count variable, say Y , and consider some distribution that allows for frequent zero-valued observations. When Case 1 occurs, Y is set at $Y = 0$ and when Case 2 occurs, Y is set at $Y > 0$, that is, the counts are generated according to the NB random variable. Thus, for $Y = 0$, which could be from the occurrence of either Case 1 with probability ϕ , or Case 2 with probability $(1 - \phi)$, we have

$$P(Y = 0) = \phi + (1 - \phi) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}}. \quad (2.1)$$

For $Y > 0$, the probability mass function (pmf) of Y follows the NB distribution written as

$$P(Y = y) = (1 - \phi) \binom{y + \alpha^{-1} - 1}{y} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^y. \quad (2.2)$$

Hence, the pmf of the count variable Y is given by (Greene [11]) and (Yau et al. [25]), that is,

$$P(Y = y) = \begin{cases} \phi + (1 - \phi) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}}, & y = 0, \\ (1 - \phi) \binom{y + \alpha^{-1} - 1}{y} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^y, & y > 0. \end{cases} \quad (2.3)$$

Equation (2.3) is known as a *zero-inflated negative binomial distribution*. According to Cameron and Trivedi [8], the mean and variance are given by

$$E(Y) = (1 - \phi)\mu \quad \text{and} \quad \text{Var}(Y) = (1 - \phi)\mu[1 + \mu(\phi + \alpha)].$$

Now, observe that as $\alpha \rightarrow 0$, $\text{Var}(Y) \rightarrow (1 - \phi)\mu(1 + \mu\phi)$; so, the ZINB distribution converges to the ZIP distribution (Moghimbeigi et al. [17]).

The NB distribution that is mixed with two-point masses ϕ and ψ , at 0 and $k > 0$, respectively, can be considered if the probability is also inflated at another count value k . When there are excessive values of k other than zero, we can extend the ZINB distribution to include the possibility of data with excessive zeros and k .

Remark 2.1. Suppose that there are three distinct data-generated processes similar to the cases mentioned above. The result of Bernoulli trial is used to determine which of the three processes is used. Consider a count variable Y . For each observation, with probability

(i) $P(Y = 0) = \phi$, the only possible response of the first process is zero count. The probability ϕ , $0 < \phi < 1$, is the proportion of zeros that does not follow a negative binomial distribution;

(ii) $P(Y = k) = \psi$, the only possible response of the second process is k -count. The probability ψ , $0 < \psi < 1$, is the proportion of k 's that does not follow a negative binomial distribution; and

(iii) $P(Y = y|\mu) = 1 - \phi - \psi$, the response of the third process is generated by a negative binomial with mean μ . The probability $1 - \phi - \psi$, where $0 < \phi$ and $0 < \phi + \psi < 1$, represents the proportion of zeros and k counts that belong to the true underlying negative binomial distribution.

The zero and k counts are generated from the first, second and third processes, where probabilities are estimated for whether zero counts and k counts are from the first, or the second, or the third process. The overall probability of the zero and k counts are combined probabilities of zeros and k 's from the three processes. Distributions which are zero and k inflated following a NB distribution, are said to follow a zero and k inflated negative binomial (ZkINB) distribution.

3. Results and Discussions

Consider a count variable Y . Suppose $Y = 0$. Then zero counts are generated either from the first or third process. For the first process, $P(Y = 0) = \phi$, $0 < \phi < 1$, is the proportion of zeros that does not follow a NB distribution. If zero counts are generated from the third process, that is from a NB distribution, with mean μ , $P(Y = y)$ is given by

$$P(Y = y) = \left[\phi + (1 - \phi - \psi) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \right], \quad (3.1)$$

for $y = 0$.

Now, consider the case of repeated k counts, that is, when $Y = k$, $k > 0$, the k counts are either generated from the second process, which does not follow a NB distribution or from the third process which follows a NB distribution. These are given in Remark 2.1(ii) and (iii). Thus, the probability of k counts is given as

$$P(Y = y) = \left[\psi + (1 - \phi - \psi) \binom{k + \alpha^{-1} - 1}{k} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^k \right], \quad (3.2)$$

for $y = k$, $k > 0$.

Furthermore, for the other counts that belong or follow a NB distribution, we have Case 3 that follows. For each $Y = y \neq 0$, $y \neq k$, y is generated from a NB distribution with mean $\mu > 0$. Hence,

$$P(Y = y) = \left[(1 - \phi - \psi) \binom{y + \alpha^{-1} - 1}{y} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^y \right], \quad (3.3)$$

which is the probability of the other counts, that is, $y \neq 0$ and $y \neq k$.

It follows that the count variable Y , that considers frequent zeros, frequent $k > 0$ counts, and other values aside from 0 and k , has a ZkINB distribution with the pmf given in the following proposition.

Proposition 3.1. *Let Y be a count variable. Suppose that ϕ and ψ are proportions of zero and $k > 0$ counts, respectively, and α is the dispersion parameter. Then for $y \in \mathbb{Z}$, the pmf of Y with inflated values at 0 and $k > 0$, is given by*

$$\begin{aligned}
 P(Y = y) = & \left[\phi + (1 - \phi - \psi) \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \right] I_{(y=0)}(y) \\
 & + \left[\psi + (1 - \phi - \psi) \binom{k + \alpha^{-1} - 1}{k} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^k \right] I_{(y=k, k \neq 0)}(y) \\
 & + \left[(1 + \phi - \psi) \binom{y + \alpha^{-1} - 1}{k} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^y \right] I_{(y \neq 0, y \neq k)}(y).
 \end{aligned} \tag{3.4}$$

Equation (3.4) can be written as

$$\begin{aligned}
 P(Y = y) = & [\phi + \omega \theta^\kappa] I_{(y=0)}(y) \\
 & + \left[\psi + \omega \binom{k + \kappa - 1}{k} \theta^\kappa (1 - \theta)^k \right] I_{(y=k, k \neq 0)}(y) \\
 & + \left[\omega \binom{y + \kappa - 1}{k} \theta^\kappa (1 - \theta)^y \right] I_{(y \neq 0, y \neq k)}(y),
 \end{aligned} \tag{3.5}$$

where $\kappa = \alpha^{-1}$, $\theta = \frac{\alpha^{-1}}{\alpha^{-1} + \mu}$, $1 - \theta = \frac{\mu}{\alpha^{-1} + \mu}$, and $\omega = 1 - \phi - \psi$.

The following results give the distribution properties of the formulated distribution.

Proposition 3.2. *If a count variable Y is zero- and k -inflated negative binomial distributed having a probability mass function given in (3.4), then*

the moment generating function (mgf) of the $ZkINB$ distribution is given by

$$M_Y(t) = \sum_{y=0}^{\infty} e^{ty} \left\{ (\phi + \psi) + \omega\theta^{\kappa} + \omega\theta^{\kappa} \left[(-1)^k \binom{-\kappa}{k} (1-\theta)^{\kappa} \right] \right\} + \left[\frac{\theta}{1-(1-\theta)e^t} \right]^{\kappa}. \quad (3.6)$$

This follows that the first derivative of (3.6) is

$$\begin{aligned} M'_Y(t) &= \sum_{y=0}^{\infty} \left(\sum_{n=1}^{\infty} y^n \frac{t^{n-1}}{(n-1)!} \right) \left\{ (\phi + \psi) + \omega\theta^{\kappa} + \omega\theta^{\kappa} \left[(-1)^k \binom{-\kappa}{k} (1-\theta)^{\kappa} \right] \right\} \\ &\quad + \frac{\omega\theta^{\kappa}\kappa(1-\theta)e^t}{[1-(1-\theta)e^t]^{\kappa+1}}, \end{aligned} \quad (3.7)$$

and evaluating (3.7) at $t = 0$, we obtain the following proposition:

Proposition 3.3. *If Y has mgf given in (3.6), then the first raw moment of Y is given by*

$$E(Y) = (1 - \theta - \psi)\mu, \quad (3.8)$$

where $\mu > 0$ is the mean of the negative binomial distribution.

Taking the second derivative of (3.6), we have

$$\begin{aligned} M''_Y(t) &= \sum_{y=0}^{\infty} \left(\sum_{n=2}^{\infty} y^n \frac{t^{n-2}}{(n-2)!} \right) \left\{ (\phi + \psi) + \omega\theta^{\kappa} + \omega\theta^{\kappa} \left[(-1)^k \binom{-\kappa}{k} (1-\theta)^{\kappa} \right] \right\} \\ &\quad + \omega\theta^{\kappa}\kappa(1-\theta)e^t \frac{[1-(1-\theta)e^t]^{\kappa}[1+\kappa e^t(1-\theta)]}{[(1-(1-\theta)e^t)^{\kappa+1}]^2}, \end{aligned} \quad (3.9)$$

and evaluating (3.9) at $t = 0$, we obtain the following proposition.

Proposition 3.4. *The second raw moment of Y is given by*

$$E(Y^2) = (1 - \phi - \psi)\mu(1 + \alpha\mu + \mu), \quad (3.10)$$

where $\mu > 0$ is the mean of the negative binomial distribution and $\alpha \geq 0$ is the dispersion parameter.

Proposition 3.5. *Let Y be a count variable with mgf given in (3.6). Then the mean and the variance of the ZkINB distribution are given by*

$$\mu_Y = (1 - \phi - \psi)\mu \quad (3.11)$$

and

$$\sigma_Y^2 = (1 - \phi - \psi)\mu + (1 - \phi - \psi)(\phi + \psi + \alpha)\mu^2, \quad (3.12)$$

where $\mu > 0$ is the mean of the negative binomial distribution and $\alpha \geq 0$ is the dispersion parameter.

Corollary 3.6. *Let count variable Y be ZkINB distributed with mean, and variance given in (3.11) and (3.12), respectively. Then,*

(i) *When $\psi = 0$, the mean and variance of the ZkINB are both reduced to $\mu_Y = (1 - \phi)\mu$ and $\sigma_Y^2 = (1 - \phi)\mu + (1 - \phi)(\alpha + \phi)\mu^2$; which are the mean and variance of the ZINB distribution, respectively.*

(ii) *When $\phi = 0$, the mean and variance of the ZkINB are both reduced to $\mu_Y = (1 - \psi)\mu$ and $\sigma_Y^2 = (1 - \psi)\mu + (1 - \psi)(\alpha + \psi)\mu^2$; which are the mean and variance of the k-INB distribution, respectively.*

(iii) *When $\psi = \phi = 0$, the mean and variance of the ZkINB are both simplified to $\mu_Y = \mu$ and $\sigma_Y^2 = \mu + \alpha\mu^2$; the mean and variance of the NB distribution, respectively.*

Estimation of the ZkINB Distribution Parameters

The Maximum Likelihood technique involves optimizing the likelihood or the log-likelihood function with respect to the unknown parameters. Let n_0 , n_k , and n_j be the numbers of Y_i s that are equal to 0, k , and j ($j \neq 0$, $j \neq k$), respectively, $j = 1, 2, \dots, k, \dots, K$, where $K = \max\{y_i\}$. Then the likelihood function of the ZkINB distribution is

$$\begin{aligned}
L(\phi, \psi, \mu, \alpha | j) \propto & \left[\phi + \omega \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \right]^{n_0} \\
& + \left[\psi + \omega \binom{k + \alpha^{-1} - 1}{k} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^k \right]^{n_k} \\
& + \prod_{j \neq 0, j \neq k}^K \left[\omega \binom{j + \alpha^{-1} - 1}{j} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^j \right]^{n_j},
\end{aligned} \tag{3.13}$$

where

$$\omega = (1 - \phi - \psi), \quad \text{and} \quad n_j = n - n_0 - n_k.$$

The maximum likelihood estimates are obtained by maximizing the log of the likelihood function. Using (3.13) and taking the natural log on both sides, we get the log-likelihood \mathcal{L} of the observed data. Let $\Theta = (\phi, \psi, \mu, \alpha)$. Then we have

$$\begin{aligned}
\mathcal{L}(\Theta | j) \geq & n_0 \ln \left[\phi + \omega \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \right] \\
& + n_k \ln \left[\psi + \omega \binom{k + \alpha^{-1} - 1}{k} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^k \right] \\
& + \sum_{j \neq 0, j \neq k}^K n_j \ln \left[\omega \binom{j + \alpha^{-1} - 1}{j} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^j \right].
\end{aligned} \tag{3.14}$$

The log-likelihood in (3.14) can be rewritten as

$$\mathcal{L}(\Theta | j) \geq n_0 \ln[\phi + \omega p_0] + n_k \ln[\psi + \omega p_k] + \sum_{j \neq 0, k}^K n_j \ln[\omega p_j], \quad (3.15)$$

where

$$p_0 = \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}},$$

$$p_k = \binom{k + \alpha^{-1} - 1}{k} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^k, \quad k \neq 0,$$

and

$$p_j = \binom{j + \alpha^{-1} - 1}{j} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^j, \quad j \neq 0, j \neq k.$$

Now, a score is a vector of the first partial derivatives of the log-likelihood function, one for each parameter, that is, $u(\Theta) = \frac{\partial \mathcal{L}}{\partial \Theta}$, where $\Theta = (\phi, \psi, \mu, \alpha)$. The score equations are given as follows:

$$(1) \quad \frac{\partial \mathcal{L}}{\partial \phi} = \frac{n_0(1 - p_0)}{\phi + \omega p_0} - \frac{n_k p_k}{\psi + \omega p_k} - \frac{n_j}{\omega};$$

$$(2) \quad \frac{\partial \mathcal{L}}{\partial \psi} = \frac{-n_0 p_0}{\phi + \omega p_0} + \frac{n_k(1 - p_k)}{\psi + \omega p_k} - \frac{n_j}{\omega};$$

$$(3)$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &\leq \frac{n_0 \omega p_0}{\phi + \omega p_0} \left(\frac{1}{1 + \mu \alpha} \right) + \frac{n_k \omega p_k}{\psi + \omega p_k} \left[\frac{k - \mu}{\mu(1 + \alpha \mu)} \right] \\ &\quad + \frac{n_j}{1 + \alpha \mu} \left(\frac{1}{\mu} \sum_{j \neq 0, k}^K j - 1 \right); \text{ and} \end{aligned}$$

(4)

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha} \leq & \frac{n_0 \omega p_0}{\phi + \omega p_0} \left[\frac{(1 + \alpha \mu) \ln(1 + \alpha \mu) - \alpha \mu}{\alpha^2 (1 + \alpha \mu)} \right] \\
& + \frac{n_k \omega p_k}{\psi + \omega p_k} \left[\frac{(1 + \alpha \mu) \ln(1 + \alpha \mu) - \alpha \mu (1 + \alpha k)}{\alpha^2 (1 + \alpha \mu)} + \frac{k(k-1)}{\alpha(k-1)+1} \right] \\
& + \sum_{j \neq 0, k}^K n_j \left[-\frac{j}{\alpha(\alpha j - \alpha + 1)} + \frac{(1 + \alpha \mu) \ln(1 + \alpha \mu) - \alpha \mu}{\alpha^2 (1 + \alpha \mu)^{1+\alpha-1}} + \frac{j}{\mu} + \frac{j\mu}{1 + \alpha \mu} \right].
\end{aligned}$$

Goodness-of-fit

For count data, the most used statistics for testing the goodness-of-fit is the Pearson chi-square statistic $\chi^2 = \sum_{i=0}^c \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(c-1)$, where o_i is the observed frequency and e_i is the expected frequency of the i th category, and c is the total number of categories. However, Arora [3] noted that this test is not robust when there are inflated frequencies.

An alternate and a simpler measure for checking the goodness-of-fit among competing distributions is the absolute error (ABE), which is defined as

$$ABE = \sum_{i=0}^c |o_i - e_i|.$$

Furthermore, the distribution that has the minimum ABE has the least deviation between the observed and expected frequencies. Hence, the distribution with minimum ABE captures the data the best.

Example 3.1. Seizure data

A seizure, technically known as an *epileptic seizure*, is a period of symptoms due to abnormally excessive or synchronous neuronal activity in the brain. Outward effects vary from uncontrolled shaking movements involving much of the body with loss of consciousness (tonic-clonic

seizure), to shaking movements involving only part of the body with variable levels of consciousness (focal seizure), to a subtle momentary loss of awareness (absence seizure). Most of the time, these episodes last less than 2 minutes and it takes some time to return to normal.

Thall and Vail [23] gave a data set on two-week seizure counts for 59 epileptics. The number of seizures was recorded for a baseline period of 8 weeks, and then patients were randomly assigned to a treatment group or a control group. Counts were then recorded for four successive two-week periods. The subject's age is the only covariate. There were a total of 236 seizures recorded from the 59 epileptics. The mean number of seizures is 8.254237 and the variance is 152.4457. The percentage (count) of patients who did not experience seizure for four successive two-week periods is 9.75% (23) and the percentage (count) of patients who experienced three-time seizure for four successive two-week periods is 13.98% (33).

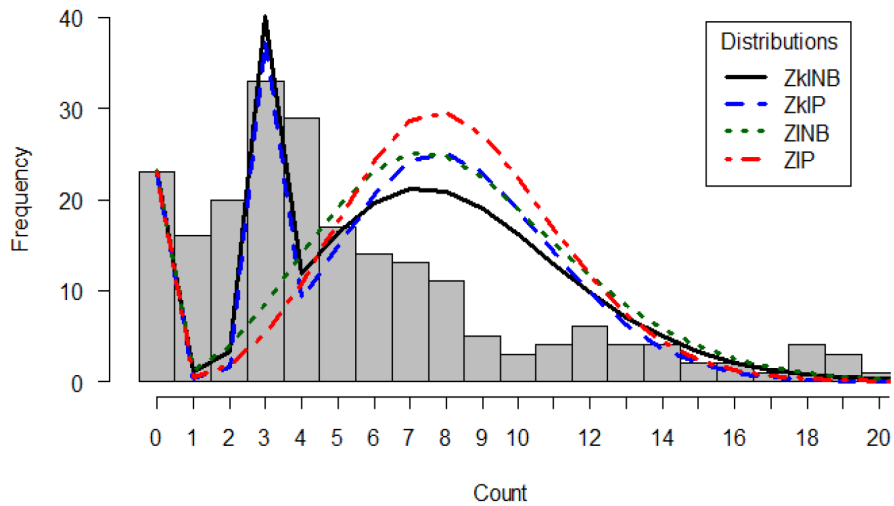


Figure 3.1. Count distributions of seizure counts for epileptics.

Figure 3.1 shows the different inflated count data distributions for the number of seizures for epileptics. It can be observed that the zero-inflated count distributions failed to capture the three inflation points in the data. Moreover, the distributions show peaks at count eight. Recall that the mean

number of seizures is eight, that is, the data is concentrated around the eight-count.

Table 3.1. The mean, median and variance of the count distributions of seizure counts for epileptics with inflations at 0 and 3

μ	Median	Variance	Distribution	μ_Y	Med_Y	Var_Y
8.25	4	152.4	ZkINB	6.30	6.71	26.15
			ZkIP	6.30	6.71	22.46
			ZINB	7.45	7.45	17.79
			ZIP	7.45	7.45	14.09

Table 3.2. Observed and expected frequencies of the count distributions of seizure counts for epileptics with inflations at 0 and 3

Count	Observed	ZkINB	ZkIP	ZINB	ZIP
0	23	23	23	23	23
1	16	1	0	1	0
2	20	3	2	4	2
3	33	40	37	8	5
4	29	12	9	14	11
5	17	16	15	19	18
6	14	20	21	23	24
7	13	21	24	25	29
8	11	21	25	25	30
9	5	19	23	22	27
10	3	16	19	19	22
11	4	13	14	15	17
12	6	10	10	12	12
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
101	0	0	0	0	0
102	1	0	0	0	0
ABE		153.2	172.8	194.4	218.9

Table 3.1 reports the means, medians, and variances of the inflated count data distributions of seizure counts for epileptics. The zero-count comprises

about 9.7% of the observations, while the k -count comprises about 14% of the observations. The mean and median of the data are 8.25 and 4, respectively, with a variance of 152.4. Furthermore, the data is inflated at 0 and 3 with a dispersion parameter value of 0.05. Moreover, the means and medians of the zero- and k -inflated count distributions are equal, with the variance of the $ZkINB$ distribution of 26.15 greater than the $ZkIP$ distribution variance of 22.46. It is simply because, the variance of the $ZkIP$ is underestimated since it is not a function of the dispersion parameter. It can be observed also that the descriptive measures for the zero- and k -inflated count distributions are smaller than that of the zero-inflated count distributions, since the latter does not account k -inflation.

Table 3.2 presents the observed and expected frequencies from the inflated count data distributions. Given the statistics of the data, the expected frequencies are obtained, and their respective absolute errors (ABEs) are calculated. A distribution that has a minimum ABE has the least deviation between the observed and expected frequencies. Hence, the distribution with minimum ABE captures the data well.

Among the four inflated count distributions, the $ZkINB$ distribution has the smallest absolute errors. Thus, the $ZkINB$ distribution seems to capture the data with zero and k inflations. It can be observed further that the mean of the $ZkINB$ and $ZkIP$ distributions are equal, their medians are both approximately equal to 7, and the variance of the $ZkINB$ is greater than the variance of the $ZkIP$ since the variability of the data is being considered in analyzing inflated count data with negative binomial distributions. Moreover, as compared to the $ZINB$ distributions, the $ZkINB$ distribution has smaller values of its statistics and a variance which is greater than that of $ZINB$ variance because the latter does not capture the k inflations in the data, where $k = 3$.

Example 3.2. Violent conflicts data

As a second example, we consider another count data that has inflated frequencies for two count values. Conflict alert (<https://conflictaalert.info/>) is

a subnational conflict monitoring system that tracks the incidence, causes, and human costs of violent conflict in the Philippines. It has data on violent conflict incidents from 2011 to 2016 in the Autonomous Region in Muslim Mindanao (ARMM) and from 2011 to 2015 in the Davao Region (except Davao City). Data from Caraga are currently being added to the database. The ARMM includes the provinces of Maguindanao, Lanao del Sur, Sulu, Basilan, and Tawi-tawi. Incidents in Cotabato City and in Isabela City are included in the database because spillovers in violence shape and are shaped by violence from these two urban centers. In the Davao Region, the provinces of Davao Oriental, Davao Occidental, Compostela Valley, Davao del Norte and Davao del Sur are covered.

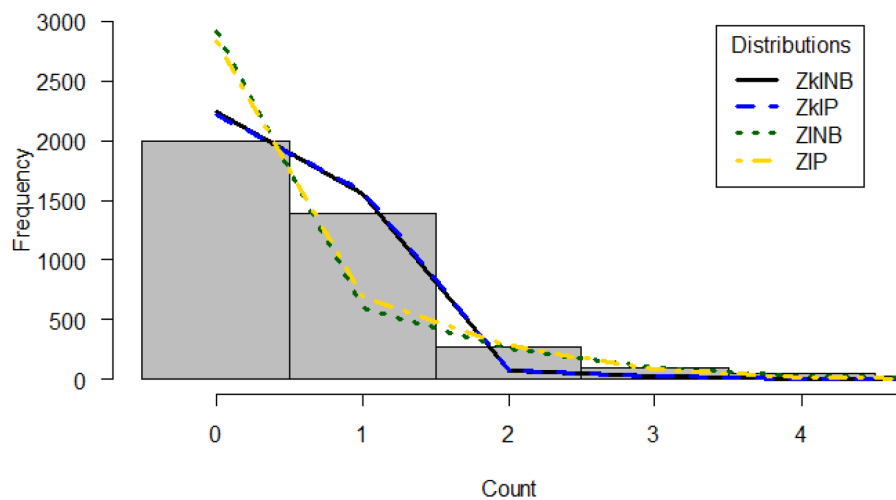


Figure 3.2. Total number of victims in violent conflicts from count distributions with 0 and 1 inflations.

The count variable being considered for this example is the number of total victims in the violent conflicts. This data consists of 3,906 observations with 20 variables. The inflated number of zeros has a frequency of 1995, and account for 51.1% of the sample. The frequency and proportion of ones are 1398 and 35.8%, respectively. The dispersion is 0.32 and the mean number of injured is 0.826, with variance of 2.559. Figure 3.2 displays the inflated count data distributions for the number of victims from the violent conflict

data. It can be observed that all distributions capture the zero inflation while only the zero- and k -inflated count distributions capture the one-inflation in the data.

Table 3.3. The mean, median and variance of the count distributions of victims in violent conflicts with inflations at 0 and 1

μ	Median	Variance	Distribution	μ_Y	Med_Y	Var_Y
0.83	0	2.56	ZkINB	0.109	0.466	0.922
			ZkIP	0.109	0.466	0.702
			ZINB	0.404	0.404	0.947
			ZIP	0.404	0.404	0.753

Table 3.4. Observed and expected frequencies of the count distributions of victims in violent conflicts with inflations at 0 and 1

Count	Observed	ZkINB	ZkIP	ZINB	ZIP
0	1995	2242	2219	2914	2831
1	1398	1559	1584	599	691
2	275	69	77	259	286
3	94	25	21	93	79
4	52	8	4	30	16
5	27	2	1	9	3
6	20	1	0	3	0
7	8	0	0	1	0
8	4	0	0	0	0
9	5	0	0	0	0
10	4	0	0	0	0
11	3	0	0	0	0
12	2	0	0	0	0
13	3	0	0	0	0
14	2	0	0	0	0
15	3	0	0	0	0
16	4	0	0	0	0
17	2	0	0	0	0
18	2	0	0	0	0
19	1	0	0	0	0
20	2	0	0	0	0
ABE		814.9	820.0	1837.2	1693.7

Table 3.3 reports the means, medians, and the variances of the zero-inflated count distributions for the number of victims. The zero-count comprised about 51.1% of the observations, while the k -count comprises 35.8% of the observations. The mean and median of the data are 0.826 and 0, respectively, with variance of 2.559.

Moreover, the data is inflated at 0 and 1 with a dispersion parameter value 0.32. Thus, the means and medians of the zero and k -inflated count distributions are equal with the $ZkINB$ variance of 0.922 which is greater than the $ZkIP$ variance of 0.702. It can be observed then that the statistics for the zero and k -inflated count distributions are smaller than the statistics of the zero-inflated count distributions since the latter does not account k -inflations.

Table 3.4 presents the observed and expected frequencies from the inflated count data distributions. It further shows that among the four inflated count distributions, the $ZkINB$ distribution has the smallest absolute errors. Thus, the $ZkINB$ distribution well-captured the data with zero and k inflations.

4. Conclusions and Recommendations

This paper proposes a $ZkINB$ distribution given in Proposition 3.1 as a tool to analyze count data with zero- and k -inflated frequencies. It is an extension of the $ZINB$ and/or the $ZkIP$ distributions. It is more flexible than either $ZINB$ or $ZkIP$ distribution as it not only captures inflation at zero and $k > 0$ but also the overdispersion that may be present in a count data. Proposition 3.2 gives the moment generating function of the $ZkINB$ distribution and used it to derive the structural or distributional properties of the $ZkINB$ distribution such as the mean and the variance, as given in Proposition 3.5. From Corollary 3.6, it is observed that as the proportion of zero counts approaches to zero, the $ZkINB$ distribution is reduced to the $ZINB$ distribution; as well as both zero and k proportions approach to zero, the $ZkINB$ distribution is simplified to the standard negative binomial distribution.

This paper illustrated the application of the $ZkINB$ distribution on two count data examples. We observed that, the $ZkINB$ seems to have a better capture of the inflations in the data as compared to the existing inflated count distributions. Moreover, on both examples, considering the absolute errors between the observed frequencies from the data and the expected frequencies from the count distributions, the $ZkINB$ distribution has the smallest error value. Thus, it seems that the $ZkINB$ distribution better captured the count data with zero and k inflations.

There are many possible extensions of this paper that one could pursue. The following are recommended for future studies:

- (1) This study focused only on zero and positive k counts. It is recommended to research in cases where there are more than one high-frequent positive k 's in the data.
- (2) Maximum likelihood estimation of the parameters for $ZkINB$ distribution could pose potential convergence problems, and the standard errors could be difficult to obtain. Thus, it is recommended to consider into consideration in using the expectation maximization (EM) algorithm to get the ML estimates for $ZkINB$ distribution.
- (3) Derive the Fisher information matrix to get the standard errors of the unknown parameters.
- (4) The focus of this study has been on formulating a proposed $ZkINB$ distribution. It is recommended to use the results in this study to proceed in the modelling procedures for the zero and k -inflated count data.

Acknowledgements

This study has been supported by the Philippine Department of Science and Technology - Science Education Institute under the Accelerated Science and Technology Human Resource Development Program, the Applied Mathematics and Statistics Research Group of the Mindanao State University - Iligan Institute of Technology Premier Research Institute of Science and Mathematics, and the Mathematics and Statistics Programs Office of the University of the Philippines Cebu College of Science.

The authors are highly grateful to the referee for his careful reading, valuable suggestions and comments, which helped to improve the presentation of this paper.

References

- [1] M. A. Abdel-Aty and A. E. Radwan, Modeling traffic accident occurrence and involvement, *Accident Analysis & Prevention* 32(5) (2000), 633-642.
doi:10.1016/s0001-4575(99)00094-9.
- [2] A. Agresti, *Categorical Data Analysis*, 2nd ed., John Wiley & Sons, Inc., Publication, Hoboken, New Jersey, USA, 2002. DOI:10.1002/0471249688.
- [3] M. Arora, Extended Poisson models for count data with inflated frequencies, Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia, United States of America, 2018. DOI: 10.25777/nz1e-d763.
- [4] M. Arora and N. R. Chaganty, EM estimation for zero- and k -inflated Poisson regression model, *Computation* 9(9) (2021), 94.
<https://doi.org/10.3390/computation9090094>.
- [5] M. Arora, N. Rao Chaganty and K. F. Sellers, A flexible regression model for zero- and k -inflated count data, *J. Stat. Comput. Simul.* 91(9) (2021), 1815-1845. DOI: 10.1080/00949655.2021.1872077.
- [6] G. Baetschmann and R. Winkelmann, Modeling zero-inflated count data when exposure varies: with an application to tumor counts, *Biom. J.* 55(5) (2013), 679-686. doi:10.1002/bimj.201200021.
- [7] S. C. Barry and A. H. Welsh, Generalized additive modelling and zero inflated count data, *Ecological Modelling* 157(2-3) (2002), 179-188.
doi:10.1016/s0304-3800(02)00194-1.
- [8] A. C. Cameron and P. K. Trivedi, *Regression analysis of count data*, 2nd ed., Econometric Society Monograph No. 53, Cambridge University Press, 2013.
- [9] G. A. Dagne, Hierarchical Bayesian analysis of correlated zero-inflated count data, *Biom. J.* 46(6) (2004), 653-663. doi:10.1002/bimj.200310077.
- [10] W. Gardner, E. P. Mulvey and E. C. Shaw, Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models, *Psychological Bulletin* 118(3) (1995), 392-404.
doi: 10.1037/0033-2909.118.3.392.

- [11] W. H. Greene, Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, Technical Report No. EC-94-10, Department of Economics, Stern School of Business, New York University, 1994. SRRN: <https://ssrn.com/abstract=1293115>.
- [12] D. B. Hall, Zero-inflated Poisson and binomial regression with random effects: a case study, *Biometrics* 56(4) (2000), 1030-1039.
doi: 10.1111/j.0006-341x.2000.01030.x.
- [13] J. M. Hilbe, *Modeling Count Data*, Cambridge University Press, New York, New York, USA, 2014. <https://doi.org/10.1017/CBO9781139236065>.
- [14] M. Kong, S. Xu, S. M. Levy and S. Datta, GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries, *Comput. Statist. Data Anal.* 85 (2015), 54-66. doi:10.1016/j.csda.2014.11.014.
- [15] T. H. Lin and M. H. Tsai, Modeling health survey data with excessive zero and K responses, *Stat. Med.* 32(9) (2012), 1572-1583. doi:10.1002/sim.5650.
- [16] B. J. Park and D. Lord, Adjustment for maximum likelihood estimate of negative binomial dispersion parameter, *Transportation Research Record: Journal of the Transportation Research Board* 2061(1) (2008), 9-19. doi:10.3141/2061-02.
- [17] A. Moghimbeigi, M. R. Eshraghian, K. Mohammad and B. Mcardle, Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros, *J. Appl. Stat.* 35(10) (2008), 1193-1202.
doi:10.1080/02664760802273203.
- [18] J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, *Journal of the Royal Statistical Society, Series A (General)* 135(3) (1972), 370.
doi:10.2307/2344614.
- [19] C. Odhiambo, S. Kibika and E. Okango, The zero inflated negative binomial - Shanker distribution and its application to HIV exposed infant data, *International Journal of Probability and Statistics* 9(1) (2020), 7-13.
DOI: 10.5923/j.ijps.20200901.02.
- [20] F. B. Oppong, E. C. Mbukam and A. A. Agyapong, Statistical models for analyzing count data, *International Journal of Scientific and Engineering Research* 8(2) (2017), 454-460.
- [21] K. M. Sakthivel, C. S. Rajitha and K. B. Alshad, Zero-inflated negative binomial-Sushila distribution and its application, *Int. J. Pure Appl. Math.* 117(13) (2017), 117-126.

- [22] K. Sellers and G. Shmueli, Data dispersion: now you see it... now you don't, *Comm. Statist. Theory Methods* 42 (2013), 3134-3147.
- [23] P. F. Thall and S. C. Vail, Some covariance models for longitudinal count data with over-dispersion, *Biometrics* 46 (1990), 657-671.
- [24] D. Yamruboon, A. Thongteeraparp, W. Bodhisuwan and K. Jampachaisri, Zero inflated negative binomial-Sushila distribution and its application, *AIP Conf. Proc.* (2017). doi:10.1063/1.5012263.
- [25] K. K. W. Yau, K. Wang and A. H. Lee, Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, *Biom. J.* 45(4) (2003), 437-452. doi:10.1002/bimj.200390024.
- [26] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev and G. M. Smith, Zero-truncated and zero-inflated models for count data, *Mixed Effects Models and Extensions in Ecology with R*, 2009, pp. 261-293. doi:10.1007/978-0-387-87458-6_11.