ON ZERO AND k INFLATED NEGATIVE BINOMIAL FOR COUNT DATA WITH INFLATED FREQUENCIES

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

DOI: 10.17654/0972361723037

Daisy Lou Lim Polestico
Mindanao State University – Iligan Institute of Technology

Advances and Applications in Statistics
© 20.. Pushpa Publishing House, Prayagraj, India
http://www.pphmj.com
http://dx.doi.org/10.17654/
Volume …, Number …, 20…, Pages … **ISSN: 0972-3617**

ON ZERO AND *k* INFLATED NEGATIVE BINOMIAL FOR COUNT DATA WITH INFLATED FREQUENCIES

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 is 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 handle an overdispersion in the data. Moreover, this study formulated a proposed zero and k inflated 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 are derived in order to compute some vital structural properties of the formulated distribution, such as the mean and the variance. Examples show that the formulated

Received: …….; Accepted: ………..
2020 Mathematics Subject Classification: …….
Keywords and phrases: ………………..
*Corresponding author
How to cite this article: ………………..

Published Online:…..20..

Z*k*INB seems to capture better the distributions as compared with other existing distributions for inflated count data.

1. Introduction

Count data refers to the number of times or 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 over dispersion in count data (Sakthivel, Rajitha, & Alshad [18]). 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 addition to the presence of zero counts, some data sets may have inflated counts of additional value k (Hilbe [11]). 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 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 ($k \neq 0$), 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 Z*k*INB, such as the mean, and the variance.

The main part of this study includes an exploration of the Z*k*INB 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 covers four inflated-count distributions namely, the formulated Z*k*INB, zero and k-inflated Poisson (Z*k*IP), zero-inflated negative binomial (ZINB), and zero-inflated (ZIP) distributions. This study only focused on the differences between the Z*k*INB, Z*k*IP, ZINB, and ZIP distributions. To determine which distribution better captured 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, namely zero and a k-inflated value, $k \geq 1$.

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:

- with probability ϕ, $0 < \phi < 1$, the only response of the first process is zero counts;
- with probability $(1 - \phi)$, the response of the second process is governed by a negative binomial with mean μ.

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 - ϕ*). 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, 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 - ϕ*). We could have

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

For *Y > 0*, the pdf of Y follows the NB distribution written as

$$P(Y=y) = (1-\phi)\binom{y+\alpha^{-1}-1}{y}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \tag{2}$$

Hence, the pdf of the count variable Y is given by (Greene [9]) and (Yau, Wang, & Lee [19])

$$P(Y=y) = \begin{cases} \phi + (1-\phi)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}, & y=0 \\ (1-\phi)\binom{y+\alpha^{-1}-1}{y}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}, & y>0 \end{cases} \tag{3}$$

The equation (3) is known as a zero-inflated negative binomial distribution. This ZINB distribution has (Cameron and Trivedi [6]) $E(Y) = (1-\phi)\mu$ and $Var(Y) = (1-\phi)\mu\left[1+\mu(\phi+\alpha)\right]$. As $\alpha \to 0$ (Moghimbeigi, Eshraghian, Mohammad, and Mcardle [15]), $Var(Y) \to (1-\phi)\mu(1+\mu\phi)$; then ZINB converges to ZIP.

The NB distribution that is mixed with two-point masses ϕ and ψ, at 0 and $k \neq 0$, 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 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,

i. *with probability $P(Y = 0) = \phi$, 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. *with probability $P(Y = k) = \psi$, 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. *with probability $P(Y = y|\mu) = 1 - \phi - \psi$, 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's 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, $P(Y = 0)$ is given by

$$\phi + (1 - \phi - \psi)\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}}.$$

It follows that the probability of zero counts is given by

$$P(Y=y) = \left[\phi + (1-\phi-\psi)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\right] I_{(y=0)}(y) \tag{4}$$

Now, consider the case of repeated k-counts, that is, when $Y = k,\ k \neq 0$, 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

$$P(Y=y) = \left[\psi + (1-\phi-\psi)\binom{k+\alpha^{-1}-1}{k}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{k}\right] I_{(y=k,\ k>0)}(y) \tag{5}$$

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,

$$P(Y=y) = \left[(1-\phi-\psi)\binom{y+\alpha^{-1}-1}{y}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}\right] I_{(y\neq 0,\ y\neq k)}(y) \tag{6}$$

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

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

$$\begin{aligned} P(Y=y) = &\left[\phi + (1-\phi-\psi)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\right] I_{(y=0)}(y) \\ &+ \left[\psi + (1-\phi-\psi)\binom{k+\alpha^{-1}-1}{k}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{k}\right] I_{(y=k,\ k>0)}(y) \\ &+ \left[(1-\phi-\psi)\binom{y+\alpha^{-1}-1}{y}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}\right] I_{(y\neq 0,\ y\neq k)}(y) \end{aligned} \tag{7}$$

Equation (7) can be written as

$$P(Y=y) = (\phi + \omega\theta^{\kappa})\, I_{(y=0)}(y) + \left[\psi + \omega\binom{k+\kappa-1}{k}\theta^{\kappa}(1-\theta)^{k}\right] I_{(y=k,\,k>0)}(y) + \left[\omega\binom{y+\kappa-1}{y}\theta^{\kappa}(1-\theta)^{y}\right] I_{(y\neq 0,\, y\neq k)}(y) \tag{8}$$

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

The following results are about 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 (7), then the moment generating function of the ZkINB distribution is given by*

$$M_Y(t) = \sum_{y=0}^{\infty} e^{ty}\left\{(\phi+\psi) + \omega\theta^{\kappa} + \omega\theta^{\kappa}\left[(-1)^{k}\binom{-\kappa}{k}(1-\theta)^{k}\right]\right\} + \omega\left[\frac{\theta}{1-(1-\theta)e^{t}}\right]^{\kappa} \tag{9}$$

This follows that the first derivative of (9) is

$$M'_Y(t) = \sum_{y=0}^{\infty}\left(\sum_{n=1}^{\infty} y^{n}\frac{t^{n-1}}{(n-1)!}\right)\left\{(\phi+\psi) + \omega\theta^{\kappa} + \omega\theta^{\kappa}\left[(-1)^{k}\binom{-\kappa}{k}(1-\theta)^{k}\right]\right\} + \frac{\omega\theta^{\kappa}\kappa(1-\theta)e^{t}}{\left[1-(1-\theta)e^{t}\right]^{\kappa+1}} \tag{10}$$

and evaluate (10) at $t = 0$, we obtain the following proposition.

Proposition 3.3. *If Y has mgf given in (9), then the first raw moment of Y is given by,*

$$E(Y) = (1-\phi-\psi)\,\mu, \tag{11}$$

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

Taking the second derivative of (9), we have

$$M''_Y(t) = \sum_{y=0}^{\infty}\left(\sum_{n=2}^{\infty} y^{n}\frac{t^{n-2}}{(n-2)!}\right)\left\{(\phi+\psi) + \omega\theta^{\kappa} + \omega\theta^{\kappa}\left[(-1)^{\kappa}\binom{-\kappa}{k}(1-\theta)^{k}\right]\right\} + \omega\theta^{\kappa}\kappa(1-\theta)e^{t}\frac{\left[1-(1-\theta)e^{t}\right]^{\kappa}\left[1+\kappa e^{t}(1-\theta)\right]}{\left[(1-(1-\theta)e^{t})^{\kappa+1}\right]^{2}} \tag{12}$$

and evaluate (12) at $t = 0$, we obtain the following proposition.

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

$$E\left(Y^2\right) = (1-\phi-\psi)\,\mu(1+\alpha\mu+\mu)\,, \tag{13}$$

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 (9). Then the mean and the variance of the ZkINB distribution are given by,*

$$\mu_Y = (1-\phi-\psi)\,\mu, \text{ and} \tag{14}$$

$$\sigma_Y{}^2 = (1-\phi-\psi)\,\mu + (1-\phi-\psi)\,(\phi+\psi+\alpha)\,u^2 \tag{15}$$

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 (14) and (15), respectively. Then,*

i. *when $\psi = 0$, the mean and variance of the ZkINB are both reduced to*

$\mu = (1-\phi)\,\mu$

$\sigma^2 = (1-\phi)\,\mu + (1-\phi)\,(\alpha+\phi)\,\mu^2$

the mean and variance of the ZINB distribution.

ii. *when $\phi = 0$, the mean and variance of then ZkINB are both reduced to*

$\mu = (1-\psi)\,\mu$

$\sigma^2 = (1-\psi)\,\mu + (1-\psi)\,(\alpha+\psi)\,\mu^2$

iii. *the mean and variance of the kINB distribution.*

when $\psi = \phi = 0$, the mean and variance of the ZkINB are both simplified to

$\mu = \mu$

$\sigma^2 = \mu + \alpha\mu^2$

the mean and variance of the NB distribution.

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.

There was 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-times seizure for four successive two-week periods is 13.98% (33).

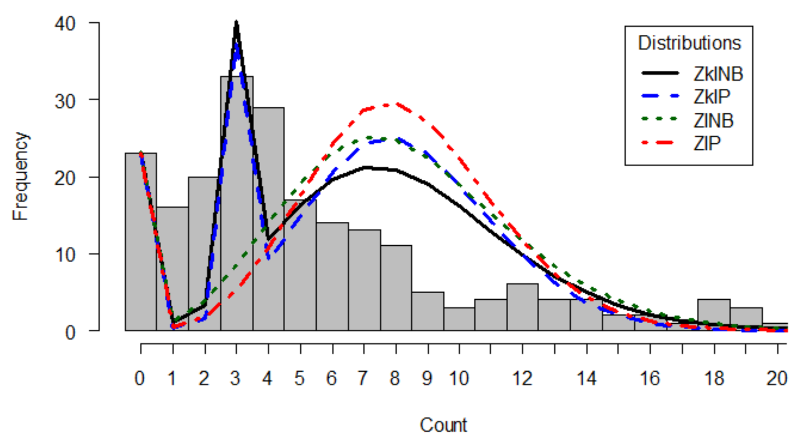

Figure 1. Count Distributions of Seizure Counts for Epileptics.

Figure 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 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	Z*k*INB	6.30	6.71	26.15
			Z*k*IP	6.30	6.71	22.46
			ZINB	7.45	7.45	17.79
			ZIP	7.45	7.45	14.09

Table 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 Z*k*INB distribution of 26.15 greater than the Z*k*IP distribution variance of 22.46. It is simply because, the variance of Z*k*IP is under-estimated 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 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.

Table 2. Observed and Expected Frequencies of the Count Distributions of Seizure Counts for Epileptics with Inflations at 0 and 3.

Count	Observed	Z*k*INB	Z*k*IP	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
⋮	⋮	⋮	⋮	⋮	⋮
101	0	0	0	0	0
102	1	0	0	0	0
ABE		**153.2**	**172.8**	**194.4**	**218.9**

Among the four inflated count distributions, the Z*k*INB distribution has the smallest absolute errors. Thus, the Z*k*INB distribution seems well-captured the data with zero and *k* inflations. It can be observed further that the mean of the Z*k*INB and Z*k*IP distributions are equal; their medians are both approximately equal to 7; and the variance of the Z*k*INB. is greater than the variance of the Z*k*IP 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, Z*k*INB distributions has smaller values of its statistics and a variance which is greater than that ZINB variance because the latter does not capture the k-inflations in the data, $k = 3$.

Seizure Data

As a second example, we consider another count data that has inflated frequencies for two count values. Conflict Alert 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.

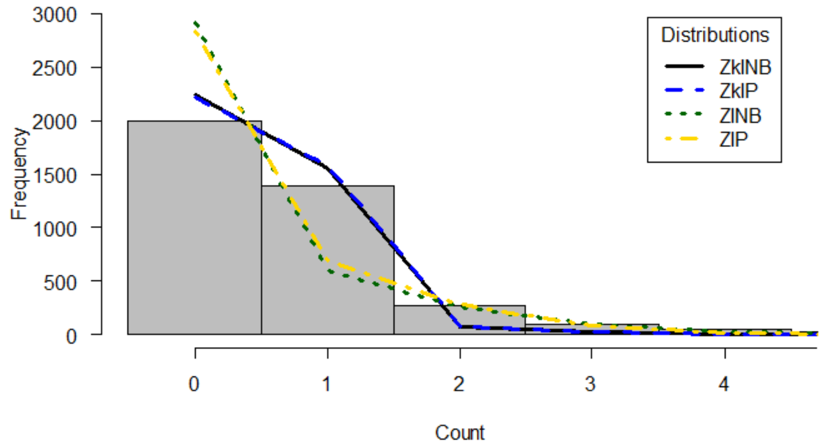

Figure 2. Total Number of Victims in Violent Conflicts from Count Distributions with 0 and 1 Inflations.

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 and 20 variables. The inflated number of zeros with 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 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 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 about 35.8% of the observations. The mean and median of the data are 0.826 and 0, respectively, with variance of 2.559.

Table 3. 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
0.83	0	2.56	Z*k*INB	0.109	0.466	0.922
			Z*k*IP	0.109	0.466	0.702
			ZINB	0.404	0.404	0.947
			ZIP	0.404	0.404	0.753

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 Z*k*INB variance is 0.922 which is greater than the Z*k*IP 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 2. Observed and Expected Frequencies of the Count Distributions of Seizure Counts for Epileptics with Inflations at 0 and 3.

Count	**Observed**	**Z*k*INB**	**Z*k*IP**	**ZINB**	**ZIP**
0	1995	2242	2219	2914	2831
1	1398	1559	1584	599	691
2	275	69	77	259	286
⋮	⋮	⋮	⋮	⋮	⋮
19	1	0	0	0	0
20	2	0	0	0	0
ABE		**814.9**	**820.0**	**1837.2**	**1693.7**

Table 4 presents the observed and expected frequencies from the inflated count data distributions. It further shows that among the four inflated count distributions, the Z*k*INB distribution has the smallest absolute errors. Thus, the Z*k*INB distribution well-captured the data with zero and *k* inflations.

4. Conclusions and Recommendations

This paper proposes a Z*k*INB 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 Z*k*IP distributions. It is more flexible than either ZINB or Z*k*IP distribution as it not only captures inflation at zero and $k \neq 0$ but also the overdispersion that may be present in a count data. Proposition (3.2) gives the moment generating function of the Z*k*INB distribution and used it to derive the structural or distributional properties of the Z*k*INB 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 Z*k*INB distribution is reduced to the ZINB distribution; as well as both zero and *k* proportions approach to zero, the Z*k*INB distribution is simplified to the standard negative binomial distribution.

This paper illustrated the application of the Z*k*INB distribution on two count data examples. We observed that, the Z*k*INB 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 Z*k*INB distribution has the smallest error value. Thus, it seems that the Z*k*INB distribution better-captured the count data with zero and *k*-inflations.

There are many possible extensions of the research in this paper that one could pursue. The researcher recommends the following for the future study:

1. This study focused only on zero and a positive *k*-counts, it is recommended to research in case there are more than one high-frequent positive *k* in the data.

2. Maximum likelihood estimation of the parameters for Z*k*INB distribution could pose convergence problems, and the standard errors could be difficult to obtain. Thus, it is recommended to take into consideration in using the expectation maximization (EM) algorithm to get the ML estimates for Z*k*INB 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 Z*k*INB 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.

References

[1] Abdel-Aty, M. A. and Radwan, A. E. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention, 32(5) (2000), 633-642. https://doi.org/10.1016/S0001-4575(99)00094-9.

[2] Agresti, A. Categorical Data Analysis Second Edition. Hoboken, New Jersey, USA: JohnWiley & Sons, Inc., Publication, 2002. DOI:10.1002/0471249688.

[3] Arora, M. 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] Baetschmann, G., & Winkelmann, R. Modeling zero-inflated count data when exposure varies: With an application to tumor counts. Biometrical Journal, 55(5) (2013), 679–686. doi:10.1002/bimj.201200021.

[5] Barry, S. C., & Welsh, A. H. Generalized additive modelling and zero inflated count data. Ecological Modelling, 157(2-3) (2002), 179–188. doi:10.1016/s0304-3800(02)00194-1

[6] Cameron, A. C. and Trivedi, P. K. Regression Analysis of Count Data, 2nd edition, Econometric Society Monograph No.53, Cambridge University Press, 2013.

[7] Dagne, G. A. Hierarchical Bayesian Analysis of Correlated Zero-inflated Count Data. Biometrical Journal, 46(6) (2004), 653–663. doi:10.1002/bimj.200310077

[8] Gardner, W., Mulvey, E. P., and Shaw, E. C. 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.

[9] Greene, W. H. 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.

[10] Hall, D. B. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. Biometrics, 56(4) (2000), 1030-9. doi: 10.1111/j.0006-341x.2000.01030.x.

[11] Hilbe, J.M. Modeling Count Data. 32 Avenue of the Americas, New York, New York, USA: Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139236065.

[12] Kong, M., Xu, S., Levy, S. M., and Datta, S. Gee type inference for clustered zero-inflated negative binomial regression with application to dental caries. Computational Statistics & Data Analysis, 85 (2015), 54-66.

[13] Lin, T. H., and Tsai, M.-H. Modeling health survey data with excessive zero and K responses. Statistics in Medicine, 32(9) (2012), 1572–1583. doi:10.1002/sim.5650.

[14] Lord, D. and Park B.J. Adjustment for the Maximum Likelihood Estimate of the

Negative Binomial Dispersion Parameter. Texas A&M University, Texas, United States of America, 2008.

[15] Moghimbeigi, A., Eshraghian, M. R., Mohammad, K., and Mcardle, B. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. Journal of Applied Statistics, 35(10) (2008), 1193-1202. doi:10.1080/02664760802273203.

[16] Nelder, J. A. and Wedderburn, R. W. Journal of the Royal Statistical Society, Series A (Statistics in Society). 135 (3) (1972), 370-384. https://doi.org/10.2307/2344614.

[17] Oppong, F.B., Mbukam, E.C., & Agyapong, A.A. Statistical Models for Analyzing Count Data. International Journal of Scientific & Engineering Research, Volume 8, Issue 2 (2017), 454-460. ISSN 2229-5518.

[18] Sakthivel, K.M., Rajitha, C.S., & Alshad, K.B. Zero-Inflated Negative Binomial-Sushila Distribution and Its Application. International Journal of Pure and Applied Mathematics, Volume 117 No. 13 (2017), 117-126.

[19] Yau, K. K. W., Wang, K., & Lee, A. H. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. Biometrical Journal, 45(4) (2003), 437452. doi:10.1002/bimj.200390024

[20] Zuur A.F., Walker, N.J., Saveliev, A.A. and Smith, G.M. Zero-Truncated and Zero-Inflated Models for Count Data, In: Mixed effects models and extensions in Ecology with T-Statistics for Biology and Health. Springer, New York, (2009), 261-293. DOI: http://dx.doi.org/10.1007%2F978-0-387-87458-6_11.