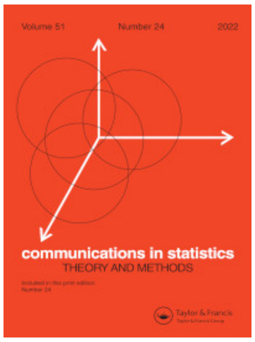

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: www.tandfonline.com/journals/lsta20

Zero-one inflated negative binomial - beta exponential distribution for count data with many zeros and ones

Chanakarn Jornsatian & Winai Bodhisuwan

To cite this article: Chanakarn Jornsatian & Winai Bodhisuwan (2022) Zero-one inflated negative binomial - beta exponential distribution for count data with many zeros and ones, Communications in Statistics - Theory and Methods, 51:24, 8517-8531, DOI: 10.1080/03610926.2021.1898642

To link to this article: https://doi.org/10.1080/03610926.2021.1898642

Published online: 14 Apr 2021.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2022, VOL. 51, NO. 24, 8517–8531
https://doi.org/10.1080/03610926.2021.1898642

Taylor & Francis
Taylor & Francis Group

Zero-one inflated negative binomial - beta exponential distribution for count data with many zeros and ones

Chanakarn Jornsatian and Winai Bodhisuwan

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand

ABSTRACT

The characteristic of count data that have a high frequency of zeros and ones can be considered under a zero-one inflated distribution. In this article, we present a zero-one inflated negative binomial - beta exponential distribution to analyze for such data. This distribution shows that it is extended the mixture negative binomial with beta exponential distributions, was proposed by Pudprommarat, Bodhisuwan, and Zeephongsekul (2012). Some important properties of this distribution are discussed, which include probability mass function, moment generating function, moment about the origin, mean and variance. Additionally, some sub-models are presented. Its parameters are also derived based on the maximum likelihood estimation procedure. The applicability of the proposed distribution is demonstrated for fitting to three real data sets. We also evaluate the abilities of model selection relying on the negative log-likelihood, Akaike information criterion, mean absolute error, root mean squared error, discrete Kolmogorov–Smirnov test and Anderson-Darling tests. Results from this study indicate the zero-one inflated negative binomial - beta exponential distribution has shown the best fit for these data sets when it is compared with some sub-models and the zero-one inflated of traditional distributions.

ARTICLE HISTORY
Received 9 July 2020
Accepted 26 February 2021

KEYWORDS
Count data; maximum likelihood estimator; negative binomial - beta exponential distribution; over-dispersed; zero-one inflated distribution

MATHEMATICS SUBJECT CLASSIFICATIONS
60E05; 62E15; 62F10; 62P99

1. Introduction

As found in the literature of count data have been prominent in various fields, such as accident analysis (Irwin 1968), manufacturing defects, medicine, and species luxuriance (Tang, Liu, and Xu 2017). The most commonly used model to analyze such data is developed using a Poisson probability distribution. Especially, the assumption of Poisson distribution implies equality of the mean and variance, called equidispersion (Cameron and Trivedi 1998). However, the Poisson distribution is no longer appropriate to be used in count data analysis, which is overdispersion or underdispersion (Winkelmann 2000).

In practice, count data generally exhibits the overdispersion. According to the overdispersion that is leading to introduce other distributions, which play an important role for count data modeling (Winkelmann 2000), such as the negative binomial, mixed Poisson, and mixed negative binomial distributions.

CONTACT Winai Bodhisuwan ✉ fsciwnb@ku.ac.th Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand.

There could be several reasons that cause overdispersion in the data. The most common cause of overdispersion for counts results from an inflated number of zeros. Indeed, the excess zeros decrease the mean of count data. That implies the sample mean is smaller than sample variance, as inflating the dispersion index (Sellers and Raim 2016).

Count data with excessive zeros, is also known as the zero-inflation of data, have gained popularity in described data by the zero-inflated distribution. One well-known zero-inflated model proposed by Lambert (1992), which is a zero-inflated Poisson (ZIP) regression model, is used to fit manufacturing defects. Furthermore, many researchers have been proposed the mixture distributions of zero-inflated model such as zero-inflated negative binomial (Greene 1994), zero-inflated negative binomial - generalized exponential (Aryuyuen, Bodhisuwan, and Supapakorn 2014), zero-inflated negative binomial - Crack (Saengthong, Bodhisuwan, and Thongteeraparp 2015), and zero-inflated negative binomial - Sushila (Yamrubboon *et al.* 2018) distributions.

Another characteristic of over-dispersed data may result from some phenomena that have high zero and one frequency simultaneously. One of the alternative models for fitting such data is called zero-one inflated model. In the unprinted manuscript, Melkersson and Olsson (1999) extended the ZIP distribution to a zero-one inflated Poisson (ZOIP) distribution. Additionally, some structural properties of ZOIP distribution are extensively studied, see Alshkaki (2016c). Alshkaki (2016a) proposed zero-one inflated negative binomial (ZOINB) distribution and its characteristics with the probability generating function.

Other zero-one inflated distributions are introduced. Alshkaki (2016b) suggested the zero-one inflated power series distributions are the extension of power series distributions that have members class of distributions such as; binomial, geometric, Poisson, negative binomial and Logarithmic. Some recent researches involving the zero-one inflated with mixed distribution are the zero–one inflated Poisson - Sushila (Pudprommarat 2018), zero–one inflated negative binomial - Crack (Tlhaloganyang and Nokwane 2019), zero–one inflated negative binomial – Sushila distributions (Pudprommarat 2020) to give for example.

As mentioned above, the objective of this work to develop the new zero-one inflated arises from the mixture negative binomial distribution, which is called the zero-one inflated negative binomial - beta exponential (ZOINB-BE) distribution. The proposed distribution is entailed a negative binomial - beta exponential (NB-BE) distribution initiated by Pudprommarat, Bodhisuwan, and Zeephongsekul (2012). The NB-BE distribution is derived from mixing between negative binomial and beta exponential (BE) distributions. This distribution is more flexible for some count data with overdispersion than the traditional Poisson and negative binomial distributions. For that reason, this work also aims to develop an alternative distribution that is expected to analyze the variability of excessive zeros and ones in count data. The proposed distribution appears to be performed effectively presenting excess zeroes and ones in modeling of counts that cannot be handled by classical distributions or zero-inflated distributions. Moreover, some interesting special cases of the proposed distribution are presented in terms of zero-one inflated, zero-inflated and baseline distributions.

The remainder of work is arranged as follows. We propose the new zero-one inflated distribution, some properties and some sub-models in Section 2. The estimated parameters using the maximum likelihood estimation method are derived as in Section 3. Some application studies with the real dataset by applying the proposed distribution are studied and illustrated results follow in Section 4. Finally, Section 5 is presented the concluding remarks.

2. The zero-one inflated negative binomial - beta exponential distribution and its properties

This section is divided into three parts: the first part describes the zero-one inflated of NB-BE distribution and follows some interesting sub-models. The last part details some statistical properties.

2.1. The zero-one inflated negative binomial - beta exponential distribution

The zero-one inflated model can be regarded as a finite mixture with three count components: a zero, a one, and larger counts. Before explaining the ZOINB-BE distribution, we briefly outline the theory and its properties in accord with the NB-BE distribution.

The NB-BE distribution was first introduced in 2012. It can be derived as a mixing of the negative binomial with BE distributions. This obtained mixture distribution provided a better fit for fitting the over-dispersed accident data and it was compared with the Poisson and negative binomial distributions (Pudprommarat, Bodhisuwan, and Zeephongsekul 2012).

Theorem 1. *A random variable X follows NB-BE distribution with the parameter vector of $\Theta = (r, a, b, c)$, if its probability mass function (pmf) is represented below*

$$f_{NB-BE}(x;\Theta) = \binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\mathrm{B}\left(b+\frac{r+j}{c},a\right)}{\mathrm{B}(a,b)} \tag{1}$$

where $x = 0, 1, 2, \ldots$ with r, a, b and $c > 0$ and the factorial moment of order k is in Equation (2),

$$\mu_{[k]}(X) = \frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\frac{\mathrm{B}\left(b-\frac{k-j}{c},a\right)}{\mathrm{B}(a,b)} \tag{2}$$

for $k = 1, 2, \ldots$ and $\mathrm{B}(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$ is the beta function, where $\Gamma(\cdot)$ is the gamma function.

Proof. See Pudprommarat, Bodhisuwan, and Zeephongsekul (2012).

For more detailed information on NB-BE distribution, see Pudprommarat, Bodhisuwan, and Zeephongsekul (2012). For the convenience refer to (2) is the relations to derive the mean and variance, as follows

$$\mathbb{E}_{NB-BE}(X) = r(\delta_1 - 1) \tag{3}$$

$$\mathbb{VAR}_{NB-BE}(X) = r[(r+1)\delta_2 - (2r+1)\delta_1 + r] - [r(\delta_1 - 1)]^2, \tag{4}$$

where $\delta_m = \mathrm{B}\left(b - \frac{m}{c}, a\right)/\mathrm{B}(a,b), \quad b > m/c.$

In this work, we study the moment generating function (mgf) of NB-BE distribution in Proposition 1.

Proposition 1. *If a random variable X follows NB-BE distribution, then the mgf is*

$$M_{X_{NB-BE}}(t)=\sum_{\forall x}\sum_{j=0}^{x}\exp(tx)\binom{r+x-1}{x}\binom{x}{j}(-1)^{j}\frac{\mathrm{B}\left(b+\frac{r+j}{c},a\right)}{\mathrm{B}(a,b)} \tag{5}$$

where $x=0,1,2,\ldots$ with r, a, b and $c>0$.

Proof. For the mgf of random variable X is the expectation of random variable $\exp(tX)$, that is

$$M_{X_{NB-BE}}(t)=\mathbb{E}_{NB-BE}[\exp(tX)]=\sum_{\forall x}\exp(tx)f_{NB-BE}(x;\Theta)$$

by substituting the pmf of NB-BE random variable in Equation (1). It can be seen that the mgf of X is in Equation (5).

Since the NB-BE distribution can be viewed in form of the mixed negative binomial distribution. A random variable X distributed as NB-BE is presented stochastic representation:

$$X|\lambda\sim NB(r,p=\exp(-\lambda))$$
$$\lambda\sim BE(a,b,c).$$

The pmf of NB-BE will be expressed as $f_{NB-BE}(x)=\int_0^\infty f_{NB}(x|\lambda)g_{BE}(\lambda)\mathrm{d}\lambda$. Then, the mgf of NB-BE distribution can be written

$$\begin{aligned}M_{X_{NB-BE}}(t)&=\sum_{\forall x}\exp(tx)\int_0^\infty f_{NB}(x|\lambda)g_{BE}(\lambda)\mathrm{d}\lambda\\&=\int_0^\infty\sum_{\forall x}\exp(tx)\binom{r+x-1}{x}\exp(-\lambda r)(1-\exp(-\lambda))^{x}g_{BE}(\lambda)\mathrm{d}\lambda\\&=\int_0^\infty M_{X_{NB}}(t)g_{BE}(\lambda)\mathrm{d}\lambda\end{aligned}$$

where $M_{X_{NB}}(t)$ is the mgf of negative binomial distribution and $g(\lambda)$ is the probability distribution function of BE distribution (see, Nadarajah and Kotz (2006)). Next, we present a new discrete distribution that is a ZOINB-BE distribution.

Theorem 2. *A random variable X follows ZOINB-BE distribution with the parameter vector of $\Theta=(r,a,b,c,\pi_0,\pi_1)$, and the corresponding pmf is defined as follows*

$$f_{ZOINB-BE}(x;\Theta)=\begin{cases}\pi_0+\pi_2\dfrac{\mathrm{B}\left(b+\dfrac{r}{c},a\right)}{\mathrm{B}(a,b)}, & \text{for } x=0,\\ \pi_1+\pi_2 r\displaystyle\sum_{j=0}^{1}(-1)^{j}\dfrac{\mathrm{B}\left(b+\dfrac{r+j}{c},a\right)}{\mathrm{B}(a,b)}, & \text{for } x=1,\\ \pi_2\dbinom{r+x-1}{x}\displaystyle\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}\dfrac{\mathrm{B}\left(b+\dfrac{r+j}{c},a\right)}{\mathrm{B}(a,b)}, & \text{for } x=2,3,4,\ldots,\end{cases} \tag{6}$$

with $r,a,b,c>0, 0<\pi_0<1$, $0<\pi_1<1$ and $\pi_2=1-\pi_0-\pi_1, 0<\pi_2<1$.

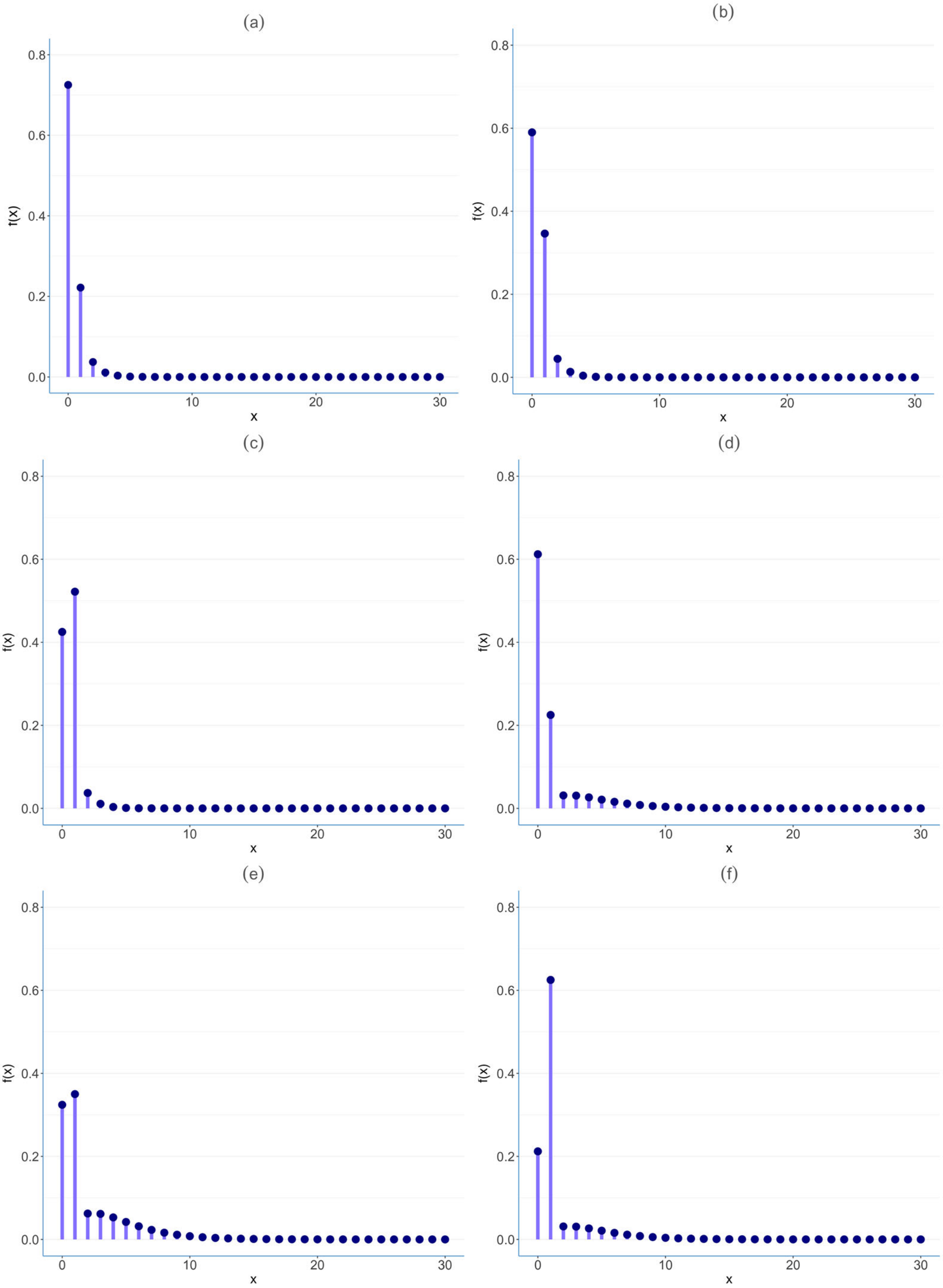

Figure 1. Some pmf plots of ZOINB-BE random variable for some parameter values (a) $r=3$, $a=4$, $b=5$, $c=5$, $p_0=0.4, p_1=0.15$, (b) $r=3$, $a=4$, $b=5$, $c=5$, $p_0=0.2, p_1=0.2$, (c) $r=3$, $a=4$, $b=5$, $c=5$, $p_0=0.1, p_1=0.4$, (d) $r=10$, $a=8$, $b=5$, $c=3$, $p_0=0.6, p_1=0.2$, (e) $r=10$, $a=8$, $b=5$, $c=3$, $p_0=0.3, p_1=0.3$ and (f) $r=10$, $a=8$, $b=5$, $c=3$, $p_0=0.2, p_1=0.6$.

Table 1. Some sub-models in the type of zero-one inflated distribution.

	Parameters					
Distribution	r	a	b	c	π_0	π_1
zero-one inflated generalized Waring	r	a	b	1	π_0	π_1
zero-one inflated negative binomial-generalized exponential	r	a	1	c	π_0	π_1
zero-one inflated negative binomial-Kumaraswamy	r	a	1	c	π_0	π_1
zero-one inflated Waring	r	1	$m-r$	1	π_0	π_1
zero-one inflated negative binomial-exponential	r	1	1	c	π_0	π_1
zero-one inflated Yule	1	1	b	1	π_0	π_1

Table 2. Some sub-models in the type of zero-inflated distribution.

	Parameters					
Distribution	r	a	b	c	π_0	Reference
zero-inflated negative binomial-beta exponential	r	a	b	c	π_0	Jornsatian (2016)
zero-inflated generalized Waring	r	a	b	1	π_0	–
zero-inflated negative binomial-generalized exponential	r	a	1	c	π_0	Aryuyuen, Bodhisuwan, and Supapakorn (2014)
zero-inflated negative binomial-Kumaraswamy	r	a	1	c	π_0	–
zero-inflated Waring	r	1	$m-r$	1	π_0	Bodhisuwan (2011)
zero-inflated negative binomial-exponential	r	1	1	c	π_0	–
zero-inflated Yule	1	1	b	1	π_0	–

Proof. The zero-one inflated model be ancillary proportions with an additional proportion of zero-valued and one-valued, then the pmf of zero-one inflated model can be elucidated as follows

$$f_{ZOINB-BE}(x;\Theta)=\begin{cases}\pi_0+\pi_2 f_{NB-BE}(x=0), & \text{for} \quad x=0,\\ \pi_1+\pi_2 f_{NB-BE}(x=1), & \text{for} \quad x=1,\\ \pi_2 f_{NB-BE}(x), & \text{for} \quad x=2,3,4,\ldots,\end{cases} \tag{7}$$

then, the pmf of ZOINB-BE distribution can be derived by substituting the pmf of NB-BE random variable as in Equation (1) into zero–one inflated model as in Eq. (7). Consequently, we obtain the pmf of ZOINB-BE distribution is shown in Eq. (6).

Besides that, some pmf plots of ZOINB-BE random variable with several values of parameters r, a, b, c, π_0 and π_1 are depicted in Figure 1.

2.2. Some sub-models

The sub-models of this distribution present in this part. The proposed distribution contains several count distributions. Some special sub-models are categorized by the type of distribution. Table 1 shows the type of zero-one inflated distribution. As noted earlier, it is to be observed that when $\pi_1=0$ the zero-one inflated distribution reduces to the zero-inflated distribution, see Table 2. And if $\pi_0=\pi_1=0$, the zero-one inflated distribution reduces to the baseline count distribution, see Table 3.

2.3. Some statistical properties

This part presents some basic statistical properties of this distribution, especially the mgf and the kth moment about the origin.

Table 3. Some sub-models in the type of baseline count distribution.

	Parameters				
Distribution	r	a	b	c	Reference
negative binomial-beta exponential	r	a	b	c	Pudprommarat, Bodhisuwan, and Zeephongsekul (2012)
generalized Waring	r	a	b	1	Irwin (1968)
negative binomial-generalized exponential	r	a	1	c	Aryuyuen and Bodhisuwan (2013)
negative binomial-Kumaraswamy	r	a	1	c	Rashid, Ahmad, and Jan (2016)
Waring	r	1	$m-r$	1	Irwin (1975)
negative binomial-exponential	r	1	1	c	Panjer and Willmot (1981)
Yule	1	1	b	1	Xekalaki (1983)

Proposition 2. *For a random variable X follows ZOINB-BE distribution, its mgf is*

$$M_X(t) = \pi_0 + \pi_1 \exp(t) + \pi_2 M_{X_{NB-BE}}(t) \tag{8}$$

with $0 < \pi_0 < 1,\ 0 < \pi_1 < 1,\ \pi_2 = 1 - \pi_0 - \pi_1, 0 < \pi_2 < 1$ *and* $M_{X_{NB-BE}}(t)$ *is the mgf of NB-BE random variable in (5).*

Proof. Straightforward requires expression, that is obtained

$$\begin{aligned} M_X(t) &= \mathbb{E}[\exp(tX)] \\ &= \sum_{\forall x} \exp(tX) f(x) \\ &= \left[\exp(t \times 0) f(x=0)\right] + \left[\exp(t \times 1) f(x=1)\right] + \left[\exp(t \times 2) f(x=2)\right] + \ldots \\ &= \left[\pi_0 + \pi_2 f_{NB-BE}(x=0)\right] + \exp(t)\left[\pi_1 + \pi_2 f_{NB-BE}(x=1)\right] \\ &\quad + \exp(t \times 2)\left[\pi_2 f_{NB-BE}(x=2)\right] + \ldots \\ &= \pi_0 + \pi_1 \exp(t) + \pi_2 M_{X_{NB-BE}}(t). \end{aligned}$$

Alternatively, Zhang, Tian, and Ng (2016) indicated that the mixed zero-one inflated distribution is a mixture of Bernoulli distribution and baseline count distribution. Let Y be a non-negative integer-valued random variable that follows ZOINB-BE distribution which can be represented from the following expression $Y = (1-Z)\eta + ZX$, where Z is a Bernoulli random variable with probability of success $1-\pi$, η has a Bernoulli distribution with probability of success p, and X follows NB-BE(r, a, b, c) distribution. Assume (Z, η and X) be mutually independent. To construct the mgf of ZOINB-BE distribution by using $\mathbb{E}(W_1) = \mathbb{E}[\mathbb{E}(W_1|W_2)]$, as yields

$$M_Y(t) = \pi(p \exp(t) + 1 - p) + (1-\pi) M_{X_{NB-BE}}(t)$$

where $\pi = 1 - \pi_2$ and $p = \pi_1/(\pi_0 + \pi_1)$.

Proposition 3. *Suppose that X is a ZOINB-BE random variable with following the kth moment about origin is*

$$\mathbb{E}(X^k) = \pi_1 + \pi_2 \mathbb{E}_{NB-BE}(X^k) \tag{9}$$

with $0 < \pi_0 < 1,\ 0 < \pi_1 < 1,$ *and* $\pi_2 = 1 - \pi_0 - \pi_1, 0 < \pi_2 < 1.$

Proof. The kth moment about the origin of ZOINB-BE distribution is derived by utilizing the pmf of this distribution in Eq. (1). Its follows that

$$\begin{aligned}\mathbb{E}(X^k) &= \sum_{\forall x} x^k f(x)\\ &= 0^k\left[\pi_0 + \pi_2 f_{NB-BE}(x=0)\right] + 1^k\left[\pi_1 + \pi_2 f_{NB-BE}(x=1)\right]\\ &\quad + 2^k\left[\pi_2 f_{NB-BE}(x=2)\right] + \ldots\\ &= \pi_1 + \sum_{x=0}^{\infty} x^k\left[\pi_2 f_{NB-BE}(x)\right] = \pi_1 + \pi_2\mathbb{E}_{NB-BE}(X^k)\end{aligned}$$

Consequently, we get the following results for the mean and variance.

Proposition 4. *For X be a random variable has ZOINB-BE distribution, some characteristics of X are derived in Equations (10) and (11). Its mean*

$$\mathbb{E}(X) = \pi_1 + \pi_2 r(\delta_1 - 1) \tag{10}$$

and also, its variance

$$\mathbb{VAR}(X) = \{\pi_1 + \pi_2 r[(r+1)\delta_2 - (2r+1)\delta_1 + r]\} - [\pi_1 + \pi_2 r(\delta_1 - 1)]^2, \tag{11}$$

where $\delta_m = \mathrm{B}\left(b - \frac{m}{c}, a\right)/\mathrm{B}(a,b), \quad b > \frac{m}{c}.$

Proof. Both properties follow directly from the definition of expectation and variance. Equivalently, setting $k=1$ in Eq. (9), we get the mean as

$$\mathbb{E}(X) = \sum_{\forall x} x f(x) = \pi_1 + \pi_2\mathbb{E}_{NB-BE}(X)$$

Accordingly, the mean of a ZOINB-BE random variable can be obtained by substituting the mean of NB-BE random variable in Eq. (3). As with variance, since $\mathbb{VAR}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$, the proof of its variance demonstrates that it is similar to proof of the mean.

3. Parameter estimation

The most frequently used technique of parameter estimate is maximum likelihood estimation or MLE. This technique is a method to find the parameter values of a probability distribution by maximizing a likelihood or log-likelihood function.

Suppose that $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random variables, each random variable has ZOINB-BE distribution. We have a sample $x_1, x_2, \ldots, x_n$, is that an independent random sample of size n from the ZOINB-BE distribution. If $\Theta = (r, a, b, c, \pi_0, \pi_1)^T$ be the vector of the ZOINB-BE parameters, then the likelihood function is as follows

$$L(\Theta;x)=\prod_{i=1}^{n}\left[I_{\{0\}}(x_i)\left(\pi_0+\pi_2\frac{\mathrm{B}\left(b+\frac{r}{c},a\right)}{\mathrm{B}(a,b)}\right)\right]$$
$$\times\prod_{i=1}^{n}\left[I_{\{1\}}(x_i)\left(\pi_1+\pi_2 r\sum_{j=0}^{1}(-1)^j\frac{\mathrm{B}\left(b+\frac{r+j}{c},a\right)}{\mathrm{B}(a,b)}\right)\right]$$
$$\times\prod_{i=1}^{n}\left[I_{\{1,2,3,\dots\}}(x_i)\left(\pi_2\binom{r+x_i-1}{x_i}\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\frac{\mathrm{B}\left(b+\frac{r+j}{c},a\right)}{\mathrm{B}(a,b)}\right)\right].$$

We define, ℓ, as log-likelihood function and denote it as

$$\ell(\Theta;x)=\sum_{i=1}^{n}[I_{\{0\}}(x_i)\log\left(\pi_0+\pi_2\frac{\mathrm{B}\left(b+\frac{r}{c},a\right)}{\mathrm{B}(a,b)}\right)$$
$$+I_{\{1\}}(x_i)\log\left(\pi_1+\pi_2 r\sum_{j=0}^{1}(-1)^j\frac{\mathrm{B}\left(b+\frac{r+j}{c},a\right)}{\mathrm{B}(a,b)}\right)$$
$$+I_{\{2,3,\dots\}}(x_i)(\log\pi_2+\log\Gamma(r+x_i)-\log\Gamma(r)-\log\Gamma(x_i+1)$$
$$+\log\left(\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\mathrm{B}\left(b+\frac{r+j}{c},a\right)\right)$$
$$-\log\Gamma(a)-\log\Gamma(b)+\log\Gamma(a+b))].$$

The maximum likelihood estimates of the ZOINB-BE distribution, $\hat{\Theta}=(\hat{r},\hat{a},\hat{b},\hat{c},\hat{\pi}_0,\hat{\pi}_1)^T$, can be obtained by solving the system of non-linear equations and setting the score functions equal to zero, $\partial\ell(\Theta;x)/\partial\Theta=0$. As the mentioned score functions are non-linear, they could not be solved using explicit solutions. We are using the optimization method, `optim` function that are available in R programming language (R Core Team 2019), is used to compute the maximum likelihood estimates $(\hat{\theta})$.

4. Application study and illustrations

In the following section, three real data sets have been considered to compare the efficacy of fitting among competing distributions. We select candidate probability distributions composed of some special cases relative to the proposed distribution and the zero-one inflated of traditional distributions. These distributions are the NB-BE, zero-inflated negative binomial - beta exponential (ZINB-BE), ZOIP, and ZOINB-BE distributions. The parameters of these distributions can be estimated by MLE. For model selection, we use some selection criteria which are the minus of log-likelihood $(-\log L)$, Akaike's information criterion (AIC), mean absolute error (MAE), root mean squared error

Table 4. Summary descriptive statistics for three data sets.

Data set	Min	Mode	Max	Mean	ID	% zero	% one
Data 1: No. doctor visits from German health	0	0	40	2.353	5.092	31.943	19.698
Data 2: No. accident injuries in United Stated	0	0	11	0.707	1.419	53.600	30.910
Data 3: No. monthly crimes in Greece	0	1	9	2.241	1.520	14.482	28.275

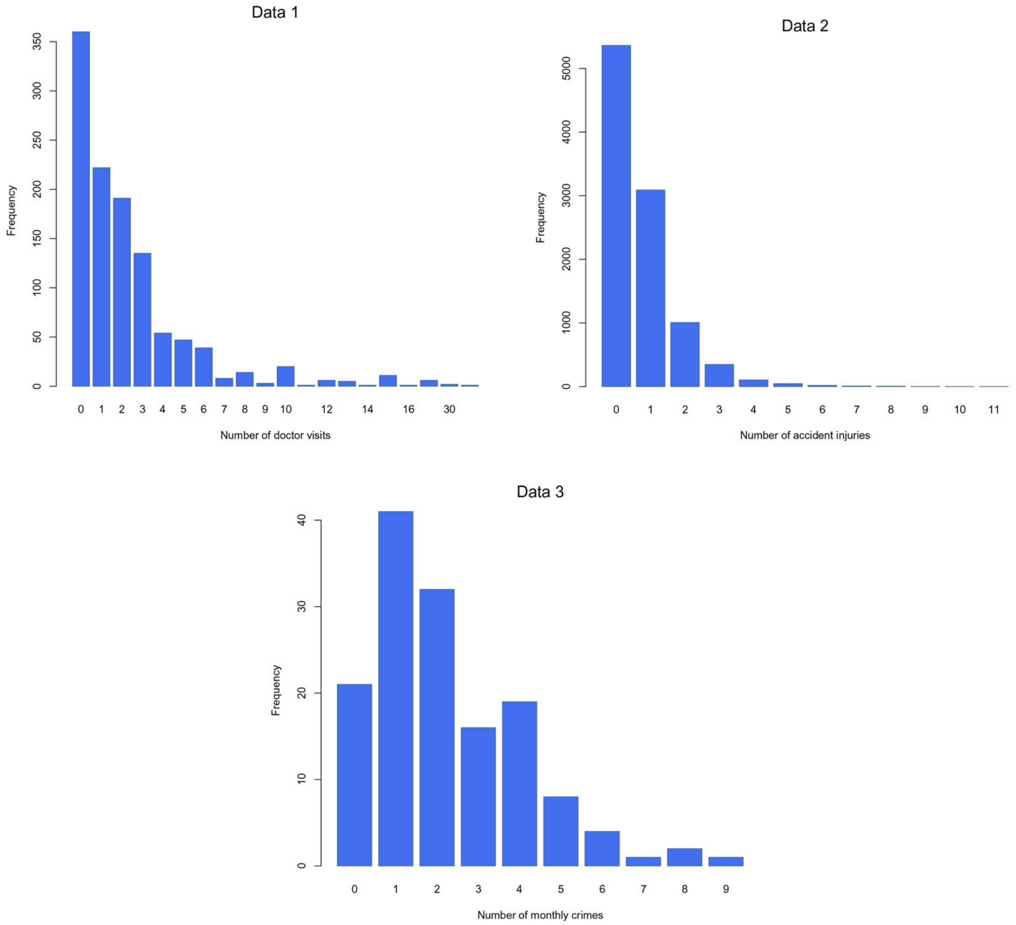

Figure 2. Histogram for three data sets.

(RMSE), discrete Kolmogorov–Smirnov (KS) test and discrete Anderson-Darling (AD) test.

The first data set is available from COUNT package in R programming language (R Core Team 2019), under the dataset name as *badhealth*. This data set comes from the German health survey data for the year 1998 and the details about the dataset are found in the online help, as well as see Hilbe (2016). *Badhealth* data set provides 1,127 observations for the number of visits to doctor during 1998.

The second data set, a data from Kadane, Krishnan, and Shmueli (2006) and Low, Ong, and Gupta (2017) is used, consisting of 10,000 observations and describe the number of accident injuries in the United Stated in 2001.

Table 5. The maximum likelihood estimators of distributions for the real data sets.

		Maximum likelihood estimators							
Data	Distributions	$\hat{\pi}_0$	$\hat{\pi}_1$	$\hat{\lambda}$	$\hat{p}$	$\hat{r}$	$\hat{a}$	$\hat{b}$	$\hat{c}$
1	NB-BE	–	–	–	–	9.353	2.507	4.770	2.048
	ZINB-BE	0.027	–	–	–	8.567	1.304	3.082	1.874
	ZOIP	0.310	0.162	4.178	–	–	–	–	–
	ZOINB	0.306	0.138	–	0.670	7.877	–	–	–
	ZOINB-BE	0.050	0.014	–	–	10.048	1.473	3.764	1.904
2	NB-BE	–	–	–	–	4.607	4.814	5.151	5.043
	ZINB-BE	0.002	–	–	–	4.269	5.302	7.493	3.726
	ZOIP	0.440	0.176	1.383	–	–	–	–	–
	ZOINB	0.391	0.152	–	0.892	10.042	–	–	–
	ZOINB-BE	0.122	0.075	–	–	7.387	3.134	5.466	4.895
3	NB-BE	–	–	–	–	14.768	10.521	2.507	13.005
	ZINB-BE	0.001	–	–	–	16.034	10.320	2.906	12.771
	ZOIP	0.099	0.156	2.802	–	–	–	–	–
	ZOINB	0.053	0.094	–	0.805	10.279	–	–	–
	ZOINB-BE	0.027	0.079	–	–	12.545	13.732	8.703	5.609

Table 6. Some criteria of model fitting to the real data sets.

Data	Distributions	$-\log L$	AIC	MAE	RMSE	KS	AD
1	ZOIP	2636.169	5278.338	18.638	32.935	0.123	24.546
	ZOINB	2437.104	4882.208	15.556	25.750	0.088	11.598
	NB-BE	2263.125	4534.250	14.782	29.262	0.101	25.597
	ZINB-BE	2229.274	4468.548	9.343	13.145	0.035	1.803
	ZOINB-BE	**2217.446**	**4446.892**	**8.888**	**13.130**	**0.030**	**1.329**
2	ZOIP	11506.310	23018.620	19.829	35.765	0.008	1.342
	ZOINB	11485.470	22978.940	16.108	28.804	0.007	0.885
	NB-BE	11471.060	22950.120	38.391	71.643	0.012	3.661
	ZINB-BE	11471.760	22953.520	38.411	71.474	0.012	3.642
	ZOINB-BE	**11455.620**	**22923.240**	**8.369**	**14.761**	**0.003**	**0.230**
3	ZOIP	275.235	556.469	2.127	3.375	0.043	0.282
	ZOINB	274.105	**556.211**	1.825	2.839	0.027	0.182
	NB-BE	274.489	556.978	2.226	3.295	0.028	0.280
	ZINB-BE	274.514	559.028	2.284	3.356	0.032	0.295
	ZOINB-BE	**273.826**	559.652	**1.681**	**2.666**	**0.026**	**0.120**

The third set of data is appeared in Karlis (2005) and Gómez-Déniz and Calderín-Ojeda (2016). It reports the number of monthly crimes during the time period 1982–1993 in Greece and total number of observations is 145. This data had been fitted previously with the generalized Poisson-Lindley distribution, negative binomial distribution, and new generalization of geometric (NGG) distribution. The NGG distribution is a more appropriate model based on minus log-likelihood and Pearson's Chi-squared.

The descriptive summary statistics of these data including min, mode, max, sample mean, index of dispersion (ID), percentage of zero and one are summarized in Table 4. We can see that two sets of data correspond to the percentage of zero are greater than the percentage of one, but in the last data set is contrast, as illustrated in Figure 2. The dispersion of such data sets are more than ones, indicating overdispersion.

Summary results of fitting the existing distributions for such data sets are given in Table 5. As can be seen in Table 6, that presents the criteria for model selection which are the $-\log L$, AIC, MAE, RMSE values, discrete KS test and discrete AD test. In the mentioned above table, the most appropriate to fit three data is ZOINB-BE distribution

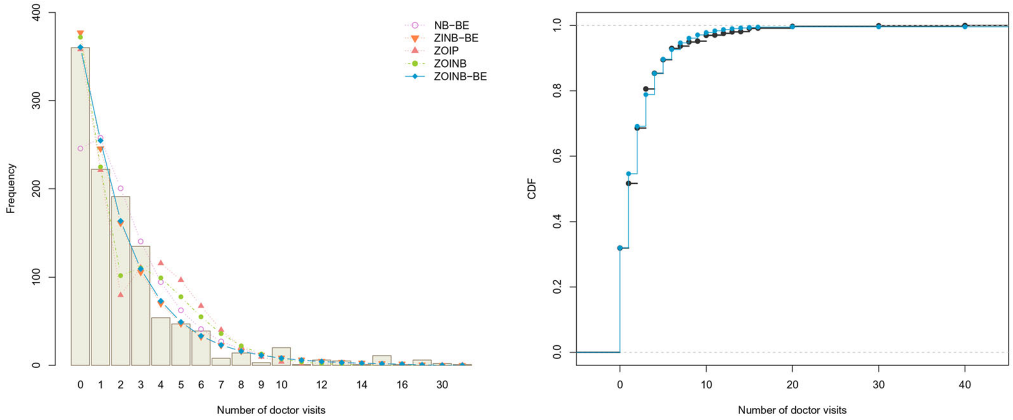

Figure 3. Plots of the observed and expected frequencies, and ZOINB-BE fitted cumulative distribution function from German health data (Data 1).

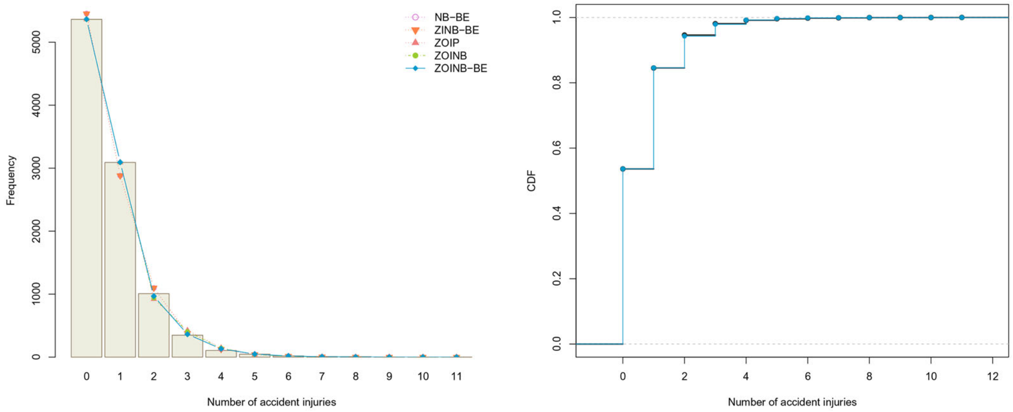

Figure 4. Plots of the observed and expected frequencies, and ZOINB-BE fitted cumulative distribution function for accident injuries data (Data 2).

describing in terms of their lowest LL, AIC, MAE, and RMSE values. For discrete goodness of fit tests, KS and AD statistics of ZOINB-BE distribution have the smallest values of these statistics and give the best fit for the data. However, Data 3 of Table 6 shows that the AIC value for ZOINB-BE distribution is roughly greater than the values for NB-BE, ZINB-BE, ZOIP, and ZOINB distributions because of more number of parameters in this distribution. Then, the proposed distribution provides a better fit for these data with excess zeroes and ones that cannot be handled by zero-inflated distribution and baseline count distributions.

We also compare the expected frequencies and observed frequencies that the fit of each data set. As can be observed, Figures 3–5 on the left-hand side reveal plots of the observed and expected frequencies for five distributions in such data. The results show that all plots of expected frequency for ZOINB-BE distribution is the closest to that observed frequency for three data. Futhermore, we demonstrate the graph of the empirical distribution functions and ZOINB-BE theoretical cumulative distribution function for each data set, as shown in Figures 3–5 on the right-hand side, respectively. The

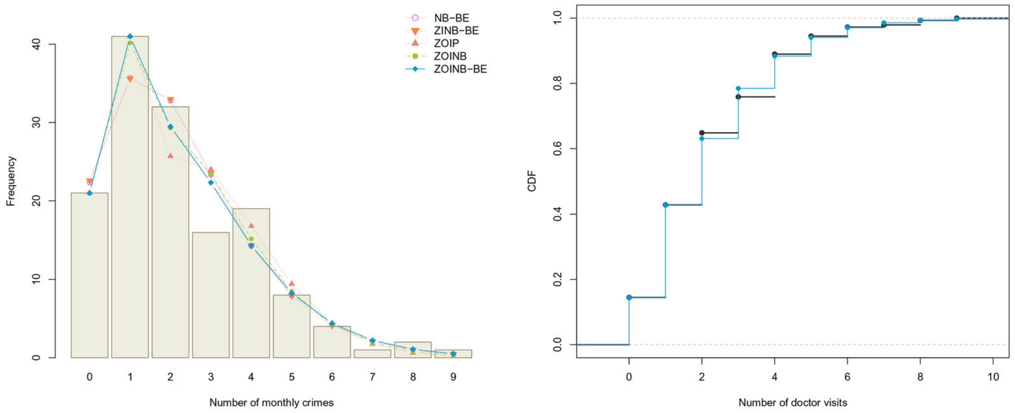

Figure 5. Plots of the observed and expected frequencies, and ZOINB-BE fitted cumulative distribution function for crimes in Greece data (Data 3).

black line is empirical cumulative distribution function whereas the blue one is ZOINB-BE theoretical cumulative distribution function. As it can be seen, the blue line is rather near the black line. Therefore, we firmly suggest the ZOINB-BE distribution can be an alternative distribution for count data in some applications, which have a high frequency of zeros and ones.

6. Conclusions

This article has introduced a new distribution, zero-one inflated negative binomial - beta exponential (ZOINB-BE) distribution. Some basic statistical properties of ZOINB-BE random variable are presented, inclusively the probability mass function, moment generating function, moment about the origin, mean and variance. We present some interesting sub-models of this distribution. Parameter estimation of ZOINB-BE distribution was also implemented by the method of maximum likelihood. This distribution was applied to three real data sets which illustrated the usefulness and compared among some special cases relative to this distribution and the zero-one inflated of traditional distributions. These distributions are the negative binomial - beta exponential, zero-inflated negative binomial - beta exponential, zero-one inflated Poisson, and zero-one inflated negative binomial distributions. The result for fitting performance of different distributions based on the minus log-likelihood, Akaike information criterion, mean absolute error, root mean squared error, discrete Kolmogorov–Smirnov test and discrete Anderson-Darling test reveal that the ZOINB-BE distribution was a flexible distribution. In addition, the proposed distribution was a particularly useful alternative distribution since most count data have excess of zeros and ones.

Acknowledgements

Authors are grateful to reviewers for helpful comments and suggestions that have substantial improvements of this paper. The first author would like to acknowledge the Science Achievement Scholarship of Thailand (SAST) for the scholarship. Also, the research work was supported by Department of Statistics, Faculty of Science, Kasetsart University.

ORCID

Chanakarn Jornsatian http://orcid.org/0000-0003-0288-8413
Winai Bodhisuwan http://orcid.org/0000-0003-3207-9019

References

Alshkaki, R. S. A. 2016a. A characterization of the zero-one inflated negative binomial distribution. *Research Journal of Mathematical and Statistical Sciences* 4 (9):1–3.

Alshkaki, R. S. A. 2016b. An extension to the zero-inflated generalized power series distributions. *International Journal of Mathematics and Statistics Invention* 4 (9):45–48.

Alshkaki, R. S. A. 2016c. On the zero-one inflated Poisson distribution. *International Journal of Statistical Distributions and Applications* 2 (4):42–48.

Aryuyuen, S., and W. Bodhisuwan. 2013. The negative binomial - Generalized exponential (NB-GE) distribution. *Applied Mathematical Sciences* 7 (22):1093–105. doi: 10.12988/ams.2013.13099.

Aryuyuen, S., W. Bodhisuwan, and T. Supapakorn. 2014. Zero inflated negative binomial - Generalized exponential distribution and its applications. *Songklanakarin Journal of Science and Technology* 36 (4):483–91.

Bodhisuwan, W. 2011. Zero inflated Waring distribution and its application. In Proceedings of the 37th Congress on Science and Technology of Thailand, 1–5. Bangkok, Thailand: The Centara Grand Bangkok Convention Centre at Central World.

Cameron, A. C., and P. K. Trivedi. 1998. *Regression analysis of count data*. New York: Cambridge University Press.

Gómez-Déniz, E., and E. Calderín-Ojeda. 2016. The Poisson-conjugate Lindley mixture distribution. *Communications in Statistics - Theory and Methods* 45 (10):2857–72. doi: 10.1080/03610926.2014.892134.

Greene, W. 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. New York University Working Papers no. EC-94-10.

Hilbe, J. M. 2016. COUNT: Functions, data and code for count data. R package version 1.3.4.

Irwin, J. O. 1968. The generalized Waring distribution applied to accident theory. *Journal of the Royal Statistical Society. Series A (General)* 131 (2):205–25. doi: 10.2307/2343842.

Irwin, J. O. 1975. The generalized Waring distribution. Part I. *Journal of the Royal Statistical Society. Series A (General)* 138 (1):18–31. doi: 10.2307/2345247.

Jornsatian, C. 2016. Statistical inference of the zero inflated negative binomial - Beta exponential distribution, Master's thesis. Kasetsart University.

Kadane, J. B., R. Krishnan, and G. Shmueli. 2006. A data disclosure policy for count data based on the com-Poisson distribution. *Management Science* 52 (10):1610–17. doi: 10.1287/mnsc.1060.0562.

Karlis, D. 2005. EM algorithm for mixed Poisson and other discrete distributions. *ASTIN Bulletin* 35 (1):3–24. doi: 10.1017/S0515036100014033.

Lambert, D. 1992. Zero - Inflated Poisson regression with application to defects in manufacturing. *Technometrics* 34 (1):1–14. doi: 10.2307/1269547.

Low, Y. C., S.-H. Ong, and R. Gupta. 2017. Generalized Sichel distribution and associated inference. *Journal of Statistical Theory and Applications* 16 (3):322–36. doi: 10.2991/jsta.2017.16.3.4.

Melkersson, M., and C. Olsson. 1999. *Is visiting the dentist a good habit?: Analyzing count data with excess zeros and excess ones*. Umeå: University of Umeå.

Nadarajah, and S. M. Kotz. 2006. The beta exponential distribution. *Reliability Engineering & System Safety* 91 (6):689–97. doi: 10.1016/j.ress.2005.05.008.

Panjer, H. H., and G. E. Willmot. 1981. Finite sum evaluation of the negative binomial - Exponential model. *ASTIN Bulletin* 12 (2):133–7. doi: 10.1017/S0515036100007066.

Pudprommarat, C., W. Bodhisuwan, and P. Zeephongsekul. 2012. A new mixed negative binomial distribution. *Journal of Applied Sciences* 12 (17):1853–58. doi: 10.3923/jas.2012.1853.1858.

Pudprommarat, C. 2018. Zero-one inflated Poisson – Sushila distribution and its application. *International Journal of Advances in Science Engineering and Technology* 6 (2):40–44.

Pudprommarat, C. 2020. Zero-one inflated negative binomial – Sushila distribution and its application. In *International Academic Multidisciplinary Research Conference in Rome 2020*, eds. K. Heuer, C. Kerdpitak, N. P. Mahalik, B. Barrett, and V. Nadda, 20–28. Rome: ICBTS.

R Core Team. 2019. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rashid, A., Z. Ahmad, and T. R. Jan. 2016. A new count data model with application in genetics and ecology. *Electronic Journal of Applied Statistical Analysis* 9 (1):213–26.

Saengthong, P., W. Bodhisuwan, and A. Thongteeraparp. 2015. The zero inflated negative binomial - Crack distribution: Some properties and parameter estimation. *Songklanakarin Journal of Science and Technology* 37 (6):701–11.

Sellers, K. F., and A. Raim. 2016. A flexible zero - Inflated model to address data dispersion. *Computational Statistics & Data Analysis* 99:68–80. doi: 10.1016/j.csda.2016.01.007.

Tang, Y., W. Liu, and A. Xu. 2017. Statistical inference for zero-and-one-inflated Poisson models. *Statistical Theory and Related Fields* 1 (2):216–26. doi: 10.1080/24754269.2017.1400419.

Tlhaloganyang, B., and S. Nokwane. 2019. Structural properties of zero-one inflated negative-binomial Crack distribution. *Research Journal of Mathematical and Statistical Sciences* 7 (3): 1–12.

Winkelmann, R. 2000. *Econometric analysis of count data*. Berlin: Springer.

Xekalaki, E. 1983. A property of the Yule distribution and its applications. *Communications in Statistics - Theory and Methods* 12 (10):1181–89. doi: 10.1080/03610928308828523.

Yamrubboon, D., A. Thongteeraparp, W. Bodhisuwan, K. Jampachaisri, and A. Volodin. 2018. Zero inflated negative binomial - Sushila distribution: Some properties and applications in count data with many zeros. *Journal of Probability and Statistical Science* 16 (2):151–63.

Zhang, C., G.-L. Tian, and K.-W. Ng. 2016. Properties of the zero-and-one inflated Poisson distribution and likelihood-based inference methods. *Statistics and Its Interface* 9 (1):11–32. doi: 10.4310/SII.2016.v9.n1.a2.