Journal of Probability and Statistical Science
16(2), 151-163, Aug. 2018

Zero Inflated Negative Binomial-Sushila Distribution: Some Properties and Applications in Count Data with Many Zeros

Darika Yamrubboon / *Kasetsart University*
Ampai Thongteeraparp / *Kasetsart University*
Winai Bodhisuwan / *Kasetsart University*
Katechan Jampachaisri / *Naresuan University*
Andrei Volodin / *University of Regina*

ABSTRACT In this work, a new zero inflated distribution is proposed which is called Zero Inflated Negative Binomial-Sushila distribution. A method of constructing this distribution is presented. Some properties of the proposed distribution are derived including probability mass function, moments about origin, variance, skewness and kurtosis. Furthermore, its special case is discussed. The maximum likelihood method is also implemented for parameter estimation of the proposed distribution. In addition, the Zero Inflated Negative Binomial-Sushila distribution is applied for some real data sets. The results show that the proposed distribution can be used as an alternative model for count data with too many zeros and over-dispersion.

Keywords Excessive zero counts; Negative binomial-Sushila distribution; Overdispersion; Zero inflated distribution.

1. Introduction

Modeling of count data is used to explain a random phenomenon in many fields. The Poisson distribution has been often used for modeling the distribution of the count data observations, see, for example [1]. The main problem with an application of the Poisson distri-

Received April 2018, revised June 2018, in final form July 2018.
Darika Yamrubboon, Ampai Thongteeraparp, and Winai Bodhisuwan (corresponding author) are affiliated to the Department of Statistics, Faculty of Science, Kasetsart University, Chatuchak, Bangkok, Thailand; emails: darika.y@ku.th, fsciamu@ku.ac.th, and fsciwnb@ku.ac.th. Katechan Jampachaisri is affiliated to the Department of Mathematics, Faculty of Science, Naresuan University, Phitsanulok, Thailand; email: katechanj@nu.ac.th. Andrei Volodin is affiliated to the Department of Mathematics and Statistics, Faculty of Science, University of Regina, Regina Saskatchewan, Canada; email: andrei@uregina.ca.

 JPSS: ISSN 1726-3328

bution is that it has the property that the variance is equal to the mean. However, in real life, most of the collected count data are obey the property that the mean of data is less than the variance. This is called over-dispersion. Therefore, the Poisson distribution is not suitable for this case, see [2]. One of the contributing factors to the extra variability is the occurrence of extra zeros. Other statistical models that overcome this problem are Negative Binomial, Mixed Negative Binomial, and zero inflated. In particular, the zero inflated count models are commonly employed in modeling count data with too many zeros, see [3].

Count data with an excessive number of zeros, called zero inflated, are generally encountered in many areas, including insurance, public health, agriculture, psychology, and econometrics, see [1, 4]. The zero inflated, a frequent manifestation of over-dispersion, implies that the incidence of zero counts is higher than expected, see [5]. The zero inflated models are generally used for taking into consideration the excess of zeros. The zero inflated models take into consideration of the structural zeros which are inevitable and sampling zeros which occur by chance, see [6]. In such situations, the Zero Inflated Poisson (ZIP), see [7, 8], and Zero Inflated Negative Binomial (ZINB) [9] models have been commonly employed in modeling count data with too many zeros. The ZIP model is applied when the count data have the variance and mean equality. Therefore, the ZINB model is suggested for over-dispersed data, see [3]. In practice, the zero inflated and the problem with over-dispersion have often occurred, thus many researchers have developed new zero inflated distributions which may be more appropriate to deal with extra zeros by applying a mixture of some distributions. Recently, a mixture of two distributions: Bernoulli and Negative Binomial distributions have been introduced to describe of count data with excess zeros. Some recent studies of the mixed distributions fitting the count data by the Zero Inflated Negative Binomial-Generalized Exponential (ZINB-GE) can be found in [10] and studies of Zero Inflated Negative Binomial-Crack (ZINB-CR) distributions can be found in [11]. These distributions can be regarded as an alternative model for excessive zero and over-dispersed count data.

In this work, the new zero inflated distribution is introduced based on the Negative Binomial-Sushila (NB-S) distribution. The NB-S distribution has been recently proposed in [12] in 2017. It is a new mixed Negative Binomial distribution is obtained by mixing the Negative Binomial distribution with the Sushila distribution, see [13]. In some cases, the NB-S distribution can be used as an alternative model for count data, especially for over-dispersion phenomena. We are interested in developing the NB-S distribution to model count data with excessive zero counts.

The objective of this research is to propose the Zero Inflated Negative Binomial-Sushila (ZINB-S) distribution. The methodology to construct the ZINB-S distribution is presented. Some mathematical properties of the proposed distribution are also derived. Furthermore, the parameters of the ZINB-S distribution are estimated by using a maximum likelihood method. The comparison of models based on ZIP, ZINB and ZINB-S for some real data sets have been presented. The ZINB-S distribution may be an appropriate alternative model to model count

data with too many zeros and over-dispersion in a correct way.

The content of this work is as follows. In Section 2, the new zero inflated distribution, called the ZINB-S distribution, is introduced. Some mathematical properties and a special case of the distribution are also derived in this section. We discuss the parameter estimation method of the proposed distribution in Section 3. The applications of ZIP, ZINB and ZINB-S distributions are illustrated with three data sets, as it is shown in Section 4. Finally, some of conclusions are drawn in Section 5.

2. A New Zero Inflated Distribution

In this section, a new zero inflated distribution is developed for modeling count data with excess of zeros. The zero inflated distribution assumes that the zero observations have two different components. The first component generates only zeros with probability ϕ. The second component generates count from a count distribution with probability $1-\phi$.

Let random variable X has a count distribution. The probability mass function (pmf) of a zero inflated distribution can be written as follows:

$$f(x)=\begin{cases}\phi+(1-\phi)p(0), & x=0\\ (1-\phi)p(x), & x=1,2,\cdots\end{cases} \tag{1}$$

where ϕ is the zero inflation parameter ($0<\phi<1$), see [14].

2.1 The Zero Inflated Negative Binomial-Sushila Distribution

The ZINB-S distribution is a new mixture of two distributions, the Bernoulli and NB-S distributions which is developed for count data with excesses of zeros. Before discussing the ZINB-S distribution, we firstly provide a definition and some mathematical properties belong to the NB-S distribution.

Definition 1 A random variable X is said to have NB-S distribution with parameters $r>0,\ \alpha>0,$ and $\theta>0$, denoted as $X\sim \text{NB-S}(r,\alpha,\theta)$, if X is distributed as the NB with parameter $r>0$ and $p=\exp(-\lambda)$, where λ has the Sushila distribution with parameter $\alpha>0$ and $\theta>0$, that is, $X\mid\lambda\sim \text{NB}(r,p=\exp(-\lambda))$ and $\lambda\sim\text{Sushila}(\alpha,\theta)$.

Theorem 1 If $X\sim \text{NB-S}(r,\alpha,\theta)$, then its pmf and the factorial moments of order *k* are, respectively,

$$p(x)=\frac{\theta^2}{\theta+1}\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{[\theta+\alpha(r+j)]^2} \tag{2}$$

where $x=0,1,2,\cdots,\ r>0,\ \alpha>0,$ and $\theta>0$. Next,

$$\mu_{[k]}=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\frac{\theta^2[\theta-\alpha(k-j)+1]}{(\theta+1)[\theta-\alpha(k-j)]^2}$$

where $\Gamma(\cdot)$ is the Gamma function.

Proof. See [12]. □

Let X be a random variable with NB-S distribution with parameters $r>0,\ \alpha>0,$ and $\theta>0$. The ZINB-S distribution with parameters $r, \alpha, \theta,$ and $\phi,$ written as

$$X \sim \text{ZINB-S}(r,\alpha,\theta,\phi)$$

is described in the following theorem.

Theorem 2 If $X \sim \text{ZINB-S}(r,\alpha,\theta,\phi)$, then the pmf of X is

$$f(x)=\begin{cases}\phi+(1-\phi)\dfrac{\theta^2(\theta+\alpha r+1)}{(\theta+1)(\theta+\alpha r)^2}, & x=0\\ (1-\phi)\dfrac{\theta^2}{\theta+1}\binom{r+x-1}{x}\displaystyle\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{[\theta+\alpha(r+j)]^2}, & x=1,2,\cdots\end{cases}$$

where $r>0,\ \alpha>0,\ \ \theta>0$ and $0<\phi<1$.

Proof. The pmf of $\text{ZINB-S}(r,\alpha,\theta,\phi)$ is obtained by substituting (2) into (1). □

Some plots of pmf of the ZINB-S random variable with specified parameters $r, \alpha, \theta,$ and ϕ are presented in Figure 1. It shows that the distribution has are positively skewed. Moreover, ϕ is the parameter whose effect on the shape of this distribution is obviously seen.

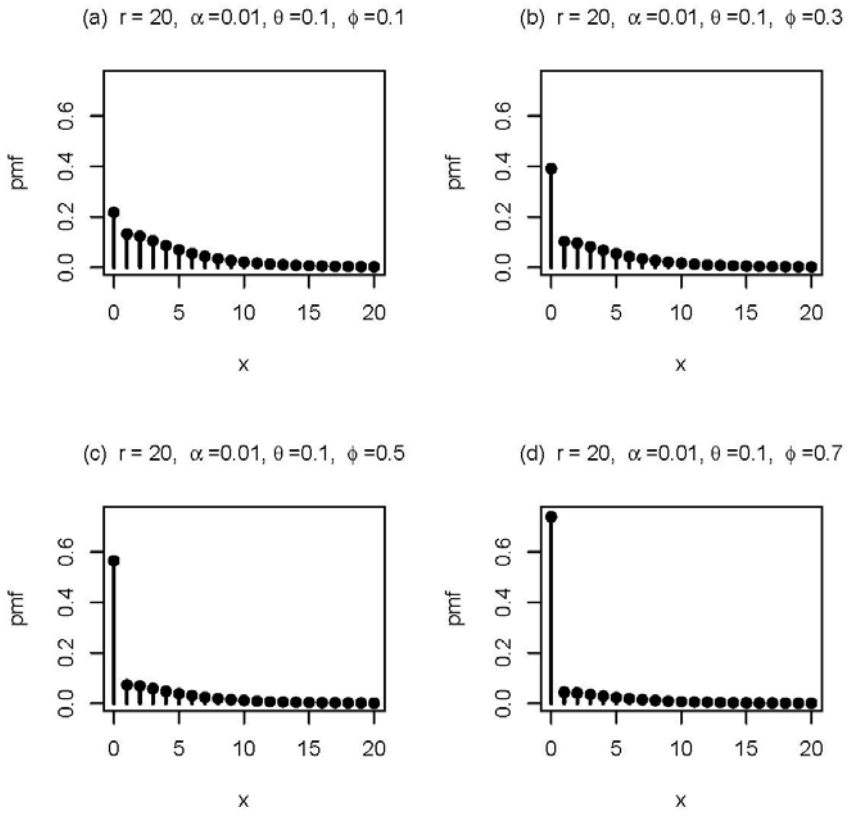

Figure 1 Some plots of pmf of the ZINB-S random variable with specified parameter values in (a)-(d)

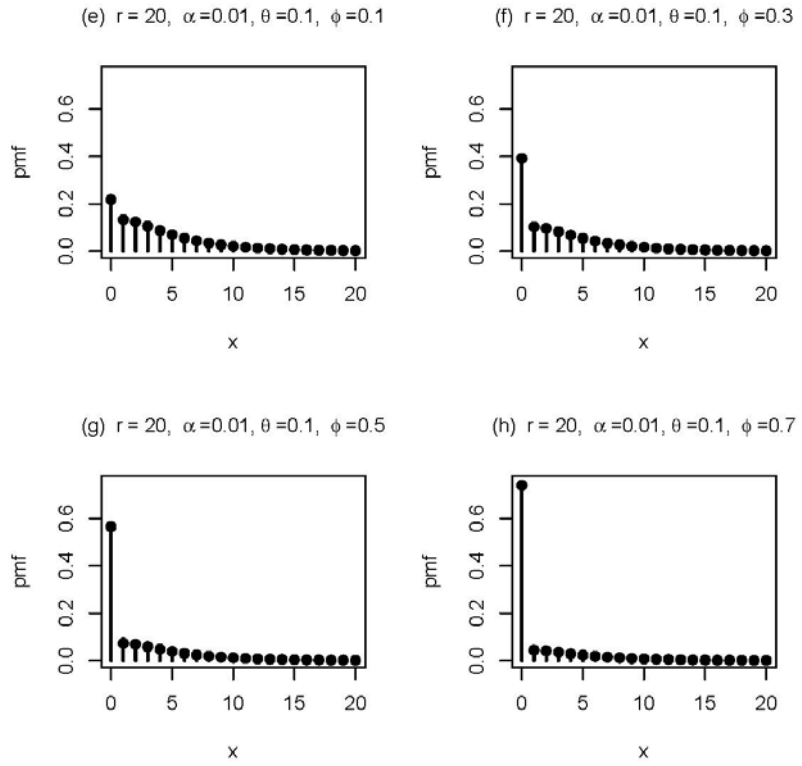

Figure 1 (Continued) Some plots of pmf of the ZINB-S random variable with specified parameter values in (e)-(h)

Corollary 1 For $\alpha = 1$ we obtain the Zero Inflated Negative Binomial-Lindley distribution with pmf

$$h(x) = \begin{cases} \phi + (1-\phi)\dfrac{\theta^2(\theta+r+1)}{(\theta+1)(\theta+r)^2}, & x = 0 \\ (1-\phi)\dfrac{\theta^2}{\theta+1}\binom{r+x-1}{x}\displaystyle\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta+(r+j)+1}{[\theta+(r+j)]^2}, & x = 1,2,\cdots \end{cases}$$

where $r > 0$, $\theta > 0$ and $0 < \phi < 1$.

2.2 Some Mathematical Properties of the Zero Inflated Negative Binomial-Sushila Distribution

In this section, we present some mathematical properties of the ZINB-S distribution.

Theorem 3 If $X \sim \text{ZINB-S}(r,\alpha,\theta,\phi)$, then the *k*th moment of X about origin is

$$E(X^k) = (1-\phi)\frac{\theta^2}{\theta+1}\sum_{x=1}^{\infty}x^k\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{[\theta+\alpha(r+j)]^2},$$

$k = 1,2,\cdots$. Consequently, we obtain the first four moments of X as

$$E(X) = (1-\phi)\frac{r}{\delta_1}(\delta_2-\delta_1), \quad E(X^2) = (1-\phi)\frac{r^2+r}{\delta_1}(\delta_3-2\delta_2+\delta_1),$$

$$E(X^3)=(1-\phi)\frac{r^3+3r^2+2r}{\delta_1}(\delta_4-3\delta_3+3\delta_2-\delta_1),$$

and

$$E(X^4)=(1-\phi)\frac{r^4+6r^3+11r^2+6r}{\delta_1}(\delta_5-4\delta_4+6\delta_3-4\delta_2+\delta_1).$$

where

$$\delta_1=\frac{\theta+1}{\theta^2},\quad \delta_2=\frac{-\alpha+\theta+1}{(\theta-\alpha)^2},\quad \delta_3=\frac{-2\alpha+\theta+1}{(\theta-2\alpha)^2},\quad \delta_4=\frac{-3\alpha+\theta+1}{(\theta-3\alpha)^2},\ \text{and}\ \ \delta_5=\frac{-4\alpha+\theta+1}{(\theta-4\alpha)^2}.$$

and $\theta\neq k\alpha,\ k=1,2,3,4.$ In particular, if $X\sim \text{ZINB-S}(r,\alpha,\theta,\phi)$, then the variance, the skewness, and the kurtosis of X according to its first four moments respectively are

$$Var(X)=(1-\phi)\frac{r^2+r}{\delta_1}(\delta_3-2\delta_2+\delta_1)-\left((1-\phi)\frac{r}{\delta_1}(\delta_2-\delta_1)\right)^2,$$

$$Skewness(X)=\frac{1}{\sigma^3}\left\{\frac{1}{\delta_1}(1-\phi)(r^3+3r^2+2r)(\delta_4-3\delta_3+3\delta_2-\delta_1)\right.$$

$$\left.-\frac{3}{\delta_1^2}(1-\phi)^2(r^3+r^2)(\delta_3-2\delta_2+\delta_1)(\delta_2-\delta_1)+\frac{2}{\delta_1^3}r^3(1-\phi)^3(\delta_2-\delta_1)^3\right\},$$

and

$$Kurtosis(X)=\frac{1}{\sigma^4}\left\{\frac{1}{\delta_1}(1-\phi)(r^4+6r^3+11r^2+6r)(\delta_5-4\delta_4+6\delta_3-4\delta_2+\delta_1)\right.$$

$$-\frac{4}{\delta_1^2}(1-\phi)^2(r^4+3r^3+2r^2)(\delta_4-3\delta_3+3\delta_2-\delta_1)(\delta_2-\delta_1)$$

$$\left.+\frac{6}{\delta_1^3}(1-\phi)^3(r^4+r^3)(\delta_2-\delta_1)^2(\delta_3-2\delta_2+\delta_1)-\frac{3}{\delta_1^4}r^4(1-\phi)^4(\delta_2-\delta_1)^4\right\}$$

where $\sigma=\sqrt{Var(X)}$.

3. Parameter Estimation

The maximum likelihood method for parameter estimations of ZINB-S(r,α,θ,ϕ) distribution is implemented in this section. Let $X_1,X_2,\cdots,X_n$ be a set of independent and identically distributed random variables with ZINB-S distribution (sample) and $x_1,x_2,\cdots,x_n$ be its corresponding set of sample values. Let $\Theta=(r,\alpha,\theta,\phi)^T$ be the vector of the ZINB-S parameters. The likelihood function of ZINB-S distribution is

$$L(\Theta)=\prod_{i=1}^{n}\left\{I_{(x_i=0)}\left(\phi+(1-\phi)\frac{\theta^2(\theta+\alpha r+1)}{(\theta+1)(\theta+\alpha r)^2}\right)\right.$$

$$+I_{(x_i>0)}\left((1-\phi)\frac{\theta^2}{\theta+1}\binom{r+x_i-1}{x_i}\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{[\theta+\alpha(r+j)]^2}\right)\Bigg\}.$$

The log-likelihood function of the ZINB-S(r,α,θ,ϕ), $\ell(r,\alpha,\theta,\phi)$, is written as

$$\ell(\Theta)=\sum_{i=1}^{n}\Bigg\{I_{(x_i=0)}\log\left(\phi+(1-\phi)\frac{\theta^2(\theta+\alpha r+1)}{(\theta+1)(\theta+\alpha r)^2}\right)+I_{(x_i>0)}\Big[\log(1-\phi)+2\log\theta-\log(\theta+1)$$

$$+\log\Gamma(r+x_i)-\log\Gamma(x_i+1)-\log\Gamma(r)+\log\left(\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{[\theta+\alpha(r+j)]^2}\right)\Big]\Bigg\}$$

and the first partial derivatives of $\ell(r,\alpha,\theta,\phi)$ with respect to each parameter, called the score functions, are derived as follows:

$$\frac{\partial\ell(\Theta)}{\partial r}=\sum_{i=1}^{n}\Bigg\{I_{(x_i=0)}\left(\frac{2\phi(\theta+1)(\theta\alpha+\alpha^2 r)}{\zeta}+\frac{(1-\phi)\theta^2\alpha}{\zeta}-\frac{2\alpha}{\theta+\alpha r}\right)$$

$$+I_{(x_i>0)}\cdot\frac{1}{\omega}\left[\psi(r+x_i)-\psi(r)+\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\left(\frac{\alpha}{Z_j^2}-\frac{2\alpha(Z_j+1)}{Z_j^3}\right)\right]\Bigg\},$$

$$\frac{\partial\ell(\Theta)}{\partial\alpha}=\sum_{i=1}^{n}\Bigg\{I_{(x_i=0)}\left(\frac{2\phi r(\theta+1)(\theta+\alpha r)}{\zeta}+\frac{(1-\phi)\theta^2 r}{\zeta}-\frac{2r}{\theta+\alpha r}\right)$$

$$+I_{(x_i>0)}\cdot\frac{1}{\omega}\left[\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\left(\frac{r+j}{Z_j^2}-\frac{2(r+j)(Z_j+1)}{Z_j^2}\right)\right]\Bigg\},$$

$$\frac{\partial\ell(\Theta)}{\partial\theta}=\sum_{i=1}^{n}\Bigg\{I_{(x_i=0)}\left(\frac{2\phi(\theta+1)(\theta+\alpha r)}{\zeta}+\frac{\phi(\theta+\alpha r)^2}{\zeta}+\frac{(1-\phi)[\theta^2+2\theta(\theta+\alpha r+1)}{\zeta}-\frac{3\theta+\alpha r+2}{(\theta+1)(\theta+\alpha r)}\right)$$

$$+I_{(x_i>0)}\cdot\frac{1}{\omega}\left[\frac{\theta+2}{\theta^2+\theta}+\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\left(\frac{1}{Z_j^2}-\frac{2(Z_j+1)}{Z_j^3}\right)\right]\Bigg\},$$

and

$$\frac{\partial\ell(\Theta)}{\partial\phi}=\sum_{i=1}^{n}\Bigg\{I_{(x_i=0)}\left(\frac{(\theta+1)(\theta+\alpha r)^2}{\zeta}+\frac{\theta^2(\theta+\alpha r+1)}{\zeta}\right)+I_{(x_i>0)}\left(-\frac{1}{1-\phi}\right)\Bigg\}$$

where $\psi(\cdot)$ is the digamma function and

$$Z_j=\theta+\alpha(r+j),\quad \zeta=\phi(\theta+1)(\theta+\alpha r)^2+(1-\phi)\theta^2(\theta+\alpha r+1),$$

and

$$\omega = \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \frac{\theta + \alpha(r+j) + 1}{[\theta + \alpha(r+j)]^2}.$$

The estimated parameters of the ZINB-S distribution ($\hat{r}$, $\hat{\alpha}$, $\hat{\theta}$, and $\hat{\phi}$) using the maximum likelihood method can be achieved by setting the score functions equal to zero and solving this system of equations. In this case, the score functions are non-linear and do not have analytical solutions. Hence, the maximum likelihood estimates can be gained by using the SANN method in optim function of R language [15]. The function *optim* in stats package provides basic optimization capabilities and is the most widely used functions in R language [16, 17].

4. Applications Study

In this section, three data sets with excess zero counts are considered to fit with the three zero inflated distributions: ZIP, ZINB and ZINB-S distributions. The results of fitting by the different distributions are compared by using minus log-likelihood (-LL), Akaike's information criteria (AIC), and Anderson-Darling (AD) test statistic for discrete distributions [18]. The discrete AD test statistic can be obtained by using *dgof* package [19] in R language. Tables 1, 2, and 3 present the results of fitting the different distributions to these data sets.

4.1 Hospital Stays Data

The first data set is number of hospital stays for individuals age 66 and over (4406 observations), see [20]. This data set was obtained from the National Medical Expenditure Survey in 1987 and 1988, see [21]. The percentage of zero in the number of hospital stays is 80.37 and the index of dispersion is 1.882, indicated that there is a high percentage of zeros and over-dispersion with mean 0.296 and variance 0.557. The fitted distributions for number of hospital stays are presented in Table 1. It shows the fitting results of distribution based on ZINB-S, which is compared with the ZIP and ZINB distributions. The result shows that ZINB-S distribution gives the smallest -LL and AIC values. The *p*-value based on the discrete AD test statistic presents that the ZINB-S distribution is closely fitted to the observed values.

4.2 Insurance Policies Data

We consider the second data set which reports the number of claims for 9,461 automobile insurance policies, see [22]. The percentage of zeros in insurance policies data is 81.32. Likewise, this data indicates over-dispersion problem with mean 0.214 and variance 0.289 (index of dispersion 1.348).

The fitted distributions for the number of claims are shown in Table 2. It illustrates that the best fit is the ZINB-S distribution, followed by the ZINB and finally the ZIP distributions. Based on *p*-value, it indicates that the number of claims following the ZINB-S distribution with the highest *p*-value can model these data well.

Table 1 Estimated Parameters for Number of Hospital Stays

Number of hospital stays	Observed values	Expected values		
		ZIP	ZINB	ZINB-S
0	3541	3539.110	3539.046	3540.039
1	599	353.905	585.677	599.371
2	176	238.613	180.597	168.547
3	48	70.415	62.947	57.065
4	20	11.558	23.205	22.102
5	12	2.759	8.834	9.487
6	5	0.407	3.434	4.418
7	1	0.051	1.355	2.199
8	4	0.005	0.540	1.157
Parameter estimates		$\hat{\phi}$ = 0.665 $\hat{\lambda}$ = 0.885	$\hat{\phi}$ = 0.095 $\hat{r}$ = 0.438 $\hat{p}$ = 0.571	$\hat{\phi}$ = 0.164 $\hat{r}$ = 2.207 $\hat{\alpha}$ = 1.575 $\hat{\theta}$ = 12.184
-LL		3059.421	3010.143	3007.494
AIC		6122.841	6026.285	6022.987
AD-statistic		1.731	0.139	0.020
p-value		0.071	0.752	0.986

Table 2 Estimated Parameters for Number of Claims

Number of claims	Observed values	Expected values		
		ZIP	ZINB	ZINB-S
0	7840	7835.279	7867.870	7845.504
1	1317	1275.796	1276.604	1294.457
2	239	297.675	262.926	249.020
3	42	46.303	45.345	54.136
4	14	5.402	7.061	13.040
5	4	0.504	1.029	3.429
6	4	0.039	0.143	0.973
7	1	0.003	0.019	0.295
8+	0	0.0001	0.003	0.094
Parameter estimates		$\hat{\phi}$ = 0.467 $\hat{\lambda}$ = 0.539	$\hat{\phi}$ = 0.442 $\hat{r}$ = 2.905 $\hat{p}$ = 0.895	$\hat{\phi}$ = 0.003 $\hat{r}$ = 4.946 $\hat{\alpha}$ = 0.734 $\hat{\theta}$ = 18.469
-LL		5375.622	5359.021	5344.785
AIC		10755.244	10724.042	10697.570
AD-statistic		0.777	0.469	0.109
p-value		0.244	0.390	0.793

4.3 Crash Data

The third dataset contains of crash data in Texas which the number of crashes fall between 2003 and 2008. This dataset includes single-vehicle roadway departure fatal crashes that occurred on 32672 rural two-lane horizontal curves, see [23]. Based on mean 0.138, variance 0.204 and index of dispersion 1.485, it can be illustrated that this dataset presents an over-dispersed data. Moreover, the dataset has a high percentage of zeros which is 89.03. The fitting results are shown in Table 3. The result shows that ZINB-S distribution also gives the smallest -LL and AIC values. The *p*-value based on the discrete AD test statistic verifies that

the number of crashes follows the ZINB-S distribution with the largest p-value and also implies that the ZINB-S distribution is the best fit for this data set. The plots of the observed and expected values of the ZINB-S distribution for the number of hospital stays, number of claims and number of crashes are illustrated in Figures 2, 3 and 4, respectively. The plots show that the ZINB-S distribution is closely fitted to these data sets. Therefore, the ZINB-S distribution can be considered as an alternative model for count data with excess zeros.

Table 3 Estimated Parameters for Number of Crashes

Number of crashes	Observed values	Expected values		
		ZIP	ZINB	ZINB-S
0	29087	29076.930	29091.770	29092.030
1	2952	2817.803	2831.857	2911.578
2	464	660.664	622.406	500.948
3	108	103.266	107.469	116.406
4	40	12.106	16.024	33.093
5	9	1.135	2.163	10.903
6	5	0.089	0.272	4.026
7	2	0.006	0.032	1.628
8	3	0.0003	0.004	0.710
9	1	0	0.0004	0.329
10+	1	0	0	0.161
Parameter estimates		$\hat{\phi}$ = 0.469 $\hat{\lambda}$ = 0.706	$\hat{\phi}$ = 0.650 $\hat{r}$ = 4.605 $\hat{p}$ = 0.922	$\hat{\phi}$ = 0.008 $\hat{r}$ = 1.168 $\hat{\alpha}$ = 1.190 $\hat{\theta}$ = 12.060
-LL		13660.96	13614.72	13528.99
AIC		27325.91	27235.43	27065.98
AD-statistic		2.164	1.381	0.110
p-value		0.040	0.104	0.745

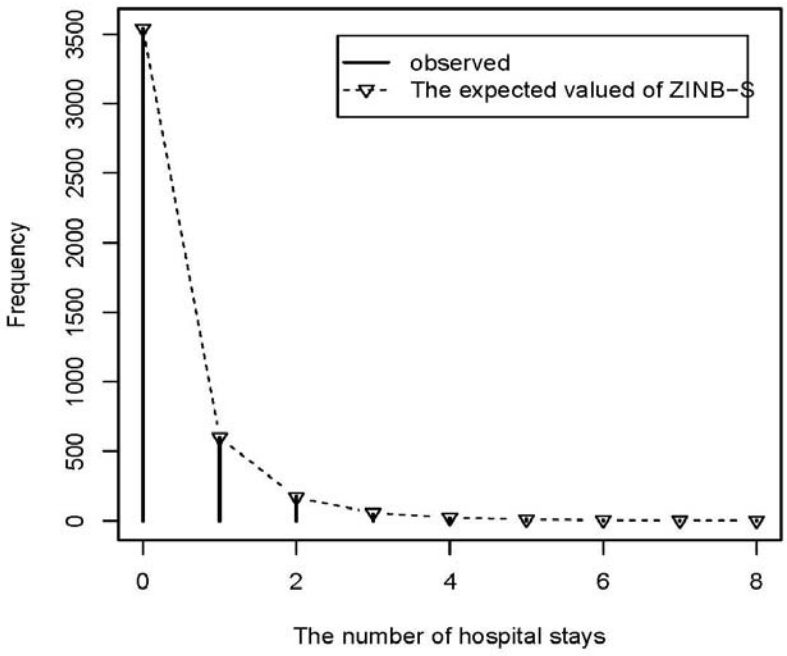

Figure 2 Results of fitting distribution to number of hospital stays

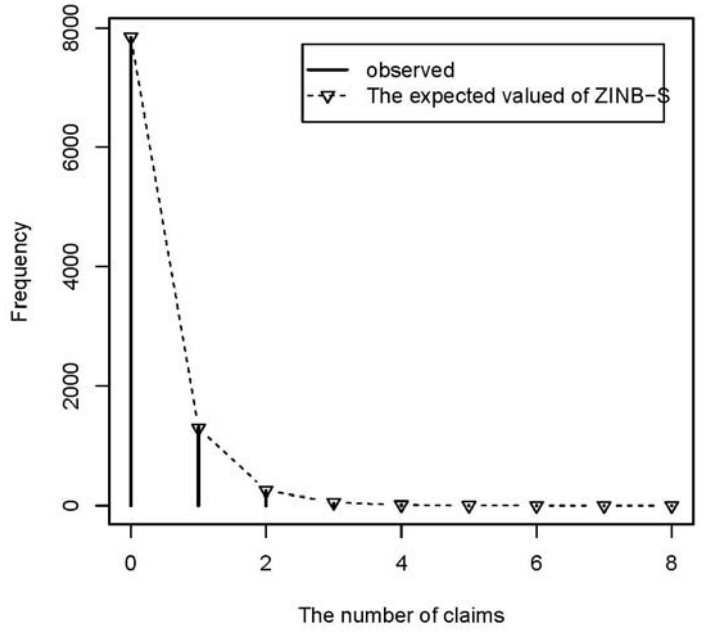

Figure 3 Results of fitting distribution to number of claims

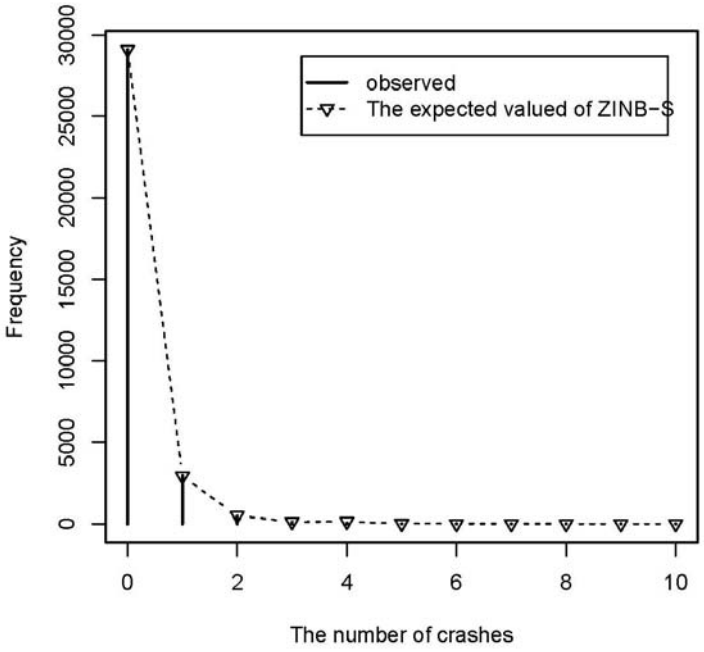

Figure 4 Results of fitting distribution to number of crashes

5. Conclusion

In this work, the Zero Inflated Negative Binomial-Sushila (ZINB-S) distribution is proposed. We show that its special case is the ZINB-L distribution. In particular, we derive some mathematical properties of the ZINB-S variable. The estimation procedure for parameters is also implemented by the maximum likelihood method. The proposed distribution is applied to

three real datasets and it is compared with ZIP and ZINB distributions. The comparison results of the minus log-likelihood and AIC values for distributions show that the best fit model is the ZINB-S distribution, followed by the ZINB and finally the ZIP distributions. In addition, The proposed distribution gives the largest *p*-value based on discrete AD test statistic and implies that the ZINB-S distribution outperforms other distributions. In conclusion, the ZINB-S distribution is a flexible model that can be an alternative way to model count data with too many zeros. For future study, we will study simulation of the ZINB-S distribution to analyze its performance and will be considered the utilization of generalized linear model to develop a ZINB-S linear model.

Acknowledgments

The authors would like to thank to Department of Statistics, Faculty of Science, Kasetsart University. Also we thank to the Faculty of Science Kasetsart University Post-graduate Studentship (ScKUPGS) for awarding a scholarship to the first listed author.

References

[1] Mouatassim, Y. and Ezzahid, E. H. (2012). Poisson regression and zero-inflated Poisson regression: application to private health insurance data, *European Actuarial Journal*, **2**(2), 187-204.

[2] Ozmen, I. and Famoye, F. (2007). Count regression models with an application to zoological data containing structural zeros, *Journal of Data Science*, **5**(4), 491-502.

[3] Zamri, N. S. N. and Zamzuri, Z. H. (2017). A review on models for count data with extra zeros, *in AIP Conference Proceedings 2017*, vol. 1830, no. 1, pp. 80010-1-080010-6.

[4] Xie, F.-C., Lin, J.-G., and Wei, B.-C. (2014) Bayesian zero-inflated generalized Poisson regression model: estimation and case influence diagnostics, *Journal of Applied Statistics*, **41**(6), 1383-1392.

[5] Garay, A. M., Lachos,V. H., and Bolfarine, H. (2015) Bayesian estimation and case influence diagnostics for the zero-inflated negative binomial regression model, *Journal of Applied Statistics*, **42**(6), 1148-1165.

[6] Ridout, M., Demétrio, C. G. B., and Hinde, J. (1998). Models for count data with many zeros, in 1998 *Proceedings of the XIXth International Biometric Conference*, **19**, 179-192.

[7] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing, *Technometrics*, **34**(1), 1-14.

[8] Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). Univariate Discrete Distributions, John Wiley & Sons.

[9] Garay, A. M., Hashimoto, E. M., Ortega, E. M. M., and Lachos, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models, *Computational Statistics and Data Analysis.*, **55**(3), 1304-1318.

[10] Aryuyuen, S., Bodhisuwan, W., and Supapakorn, T. (2014). Zero inflated negative binomial-generalized exponential distribution and its applications, *Songklanakarin Journal of Science & Technology*, **36**(4), 483-491.

[11] Saengthong, P., Bodhisuwan, W., and Thongteeraparp, A. (2015). The zero inflated negative binomial-Crack distribution: some properties and parameter estimation, *Songklanakarin Journal of Science & Technology*, **37**(6), 701-711.

[12] Yamrubboon, D., Bodhisuwan, W., Pudprommarat, C., and Saothayanun, L. (2017). The negative binomial-Sushila distribution with application in count data analysis, *Thailand Statistician*, **15**(1), 69-77.

[13] Shanker, R., Sharma, S., Shanker, U., and Shanker, R. (2013). Sushila distribution and its application to waiting times data, *International Journal of Business Management*, **3**(2), 1-11.

[14] Cameron, A. C. and Trivedi, P. K. (2013). Regression Analysis of Count Data, Cambridge University Press, 2013.

[15] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria, 2017.

[16] Nash, J. C. (2014). On best practice optimization methods in R, *Journal of Statistical Software*, **60**(2), 1-14.

[17] Varadhan, R. (2014). Numerical optimization in R: Beyond optim, *Journal of Statistical Software*, **60**(1), 1-3.

[18] Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994). Cramér-von Mises statistics for discrete distributions, *Canadian Journal of Statistics*, **22**(1), 125-137.

[19] Arnold, T. B. and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions, *R Journal*, **3**(2), 34-39.

[20] Flynn, M. and Francis, L. A. (2009). More flexible GLMs zero-inflated models and hybrid models, *Casualty Actuarial Society E-Forum*, Winter 2009, 148-224.

[21] Deb, P. and P. K. Trivedi, P. K. (1997). Demand for medical care by the elderly: a finite mixture approach, *Journal of Applied Econometrics*, **12**(3), 313-336.

[22] Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). Loss Models: from Data to Decisions, John Wiley & Sons.

[23] Lord, D. and Geedipally, S. R. (2011). The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros, *Accident Analysis & Prevention*, **43**(5), 1738-1742.