Thailand Statistician
January 2017; 15(1): 69-77
http://statassoc.or.th
Contributed paper

The Negative Binomial-Sushila Distribution with Application in Count Data Analysis

Darika Yamrubboon [a], Winai Bodhisuwan*[a], Chookait Pudprommarat [b] and Luckana Saothayanun [c]

[a] Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand.

[b] Department of Science, Faculty of Science and Technology, Suan Sunandha Rajabath University, Bangkok 10300, Thailand.

[c] Department of Statistics, Faculty of Science and Technology, University of the Thai Chamber of Commerce, Bangkok 10325, Thailand.

*Corresponding author; e-mail: fsciwnb@ku.ac.th

Received: 8 September 2015
Accepted: 23 February 2016

Abstract

In this paper, we introduce a negative binomial-Sushila distribution which is a new mixed negative binomial distribution. The probability mass function (pmf) has been expressed as mixtures of the negative binomial and the Sushila distribution. The factorial moments, the first four moments, variance and skewness have been derived. Moreover, we found that the negative binomial-Lindley distribution is its special case. We also discuss maximum likelihood estimation of the model parameters. For application to real data set, it shows that the new distribution can provide a better fit the data than the Poisson and negative binomial distributions. We hope that this distribution may be an alternative model to over-dispersed count data analysis.

Keywords: Mixed negative binomial, Sushila distribution, negative binomial-Lindley distribution, over-dispersed data.

1. Introduction

Count data are often encountered in real world applications. There is a very large collection of literature on how to analyze and model count data. Many discrete distributions have been developed for modeling count data (Li et al. 2011). Model for count data have been prominent in many disciplines such as health economics, management and industrial organization (Greene 2008). The Poisson distribution is usually employed to fit count data in practice. However theoretical prediction may not match empirical observations for moment of higher order due to the only one parameter, which does not allow the variance to be adjusted independently of mean (Wang 2011). For this reason, the negative binomial (NB) distribution has become increasingly popular as a more flexible alternative to the Poisson distribution. Especially it is doubtful whether the strict requirements, particularly independence, for the Poisson distribution will be satisfied (Johnson et al. 2005).

Many attempts were implemented in expanding the classes of mixed and compound distributions, especially in the distribution of exponential family, resulting in a better fit on count data. In some case, it is proven that mixed distribution, in particular mixed Poisson and mixed negative binomial, provided better fit compared to other distributions (Zamani and Ismail 2010). Mixed Distributions define one of the most important ways to obtain new probability distributions in applied probability and operational research (Gómez-Déniz et al. 2008). One mixed distribution has been proposed in application to count data, especially a mixed negative binomial distribution. The mixed negative binomial distribution has been discussed by many authors such as NB-inverse Gaussian distribution, NB-beta exponential distribution, NB-generalized exponential distribution and NB-Erlang distribution, etc. (Gómez-Déniz et al. 2008; Pudprommarat et al. 2012; Aryuyuen and Bodhisuwan 2013; Kongrod et al. 2014). These have been proposed in fitting count data under the overdispersion. As it can be seen from above, those distributions that are lifetime distributions were used as mixing models to produce the mixed negative binomial distributions.

The Sushila distribution which is a new lifetime distribution was introduced by Shanker et al. (2013). This distribution, being a modified Lindley distribution, is mixture of the exponential and gamma distributions. The Lindley distribution has been discussed by Ghitany et al. (2008). It was applied as a mixing model for the negative binomial parameters to generate the mixed negative binomial distribution known as the NB-Lindley distribution (Zamani and Ismail 2010). The Sushila distribution was discussed that its failure rate function and mean residual life function show flexibility over the Lindley and exponential distributions (Shankar et al. 2013). Moreover, we found that the moment generating function of Sushila distribution can be expresses in closed form. Therefore, we have been interested in using the Suhila distribution for creating a new alternative mixed model.

In this study, we introduce a new mixed negative binomial distribution by mixing the NB distribution and the Sushila distribution. This new mixed distribution is called the negative binomial-Sushila (NB-S) distribution. Additionally, we present some properties of the NB-S distribution are the factorial moment, the first four moments, variance and skewness. The parameter estimations of the NB-S distribution are estimated by the maximum likelihood method, and we also present comparison analysis between the Poisson, NB and NB-S distributions based on real data set. The NB-S distribution may be an alternative model to count data analysis.

2. Methodology

In this section, we provide the definition of a new mixed NB distribution which is a NB-S distribution obtained by mixing the NB distribution with the Sushila distribution. Furthermore, we present some properties and parameter estimations of the NB-S distribution.

2.1. The negative binomial-Sushila distribution

In this part, the NB-S distribution is a mixture of the NB and Sushila distributions. First we present the NB distribution and some of its properties.

A classical negative binomial distribution is specified by the probability mass function (pmf)

$$P(X=x)=\binom{r+x-1}{x}p^{r}(1-p)^{x}, \tag{1}$$

where $x=0,1,2,\ldots,$ for $r>0$ and $0<p<1$.

If X denotes a random variable under the negative binomial distribution with parameters r and p, the first two moments about zero and the factorial moment of X (Gómez-Déniz et al. 2008) are given respectively by

$$E(X) = \frac{r(1-p)}{p},$$

$$E(X^2) = \frac{r(1-p)(1+r(1-p))}{p^2},$$

and

$$\mu_{[k]}(X) = E\left[X(X-1)\ldots(X-k+1)\right] = \frac{\Gamma(r+k)}{\Gamma(r)}\frac{(1-p)^k}{p^k}, \tag{2}$$

where $k = 1,2,\ldots,$ and $\Gamma(\cdot)$ is the complete gamma function denoted by

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx,\ t>0.$$

The Sushila distribution is specified by the probability density function (pdf)

$$f(x;\alpha,\theta) = \frac{\theta^2}{\alpha(\theta+1)}(1+\frac{x}{\alpha})e^{(-\frac{\theta}{\alpha}x)}, \tag{3}$$

where $x>0, \theta>0$ and $\alpha > 0$.

The Sushila distribution, which is a two parameter continuous distribution, was proposed by Shanker et al. (2013). It can easily be seen that the Lindley distribution is a particular case of (3) when $\alpha = 1$. The Sushila distribution can be shown as a mixture of the exponential $\left(\frac{\theta}{\alpha}\right)$ and gamma$\left(2,\frac{\theta}{\alpha}\right)$ distributions as follows

$$f(x;\alpha,\theta) = pf_1(x) + (1-p)f_2(x),$$

where $p = \frac{\theta}{\theta+1}$, $f_1(x) = \frac{\theta}{\alpha}e^{-\frac{\theta}{\alpha}x}$ and $f_2(x) = \frac{\theta^2}{\alpha^2}xe^{-\frac{\theta}{\alpha}x}$.

The moment generating function of the Sushila distribution is given by

$$M_X(t) = \frac{\theta^2(\theta-\alpha t+1)}{(\theta+1)(\theta-\alpha t)^2}. \tag{4}$$

Now, we provide a general definition of the NB-S distribution.

Definition Let X be a random variable of the NB-S(r,α,θ) distribution where X has the NB distribution with parameter $r>0$ and $p=\exp(-\lambda)$ when λ has the Sushila distribution with positive parameters α and θ, i.e., $X\,|\,\lambda \sim \text{NB}\big(r, p=\exp(-\lambda)\big)$ and $\lambda \sim$ Sushila(α,θ) for $r,\alpha,\theta > 0$.

Theorem 1 *Let* $X \sim$ NB-S(r,α,θ). *The pmf of* X *is given by*

$$f(x;r,\alpha,\theta) = \frac{\theta^2}{\theta+1}\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{(\theta+\alpha(r+j))^2}, \tag{5}$$

where $x = 0,1,2,\ldots,$ *, for* r, α *and* $\theta > 0$.

Proof: If $X \mid \lambda \sim \mathrm{NB}(r, p = \exp(-\lambda))$ in (1) and $\lambda \sim$ Sushila (α, θ) in (3), then the pmf of X can be obtained by

$$f(x; r, \alpha, \theta) = \int_0^\infty P(X = x \mid \lambda) f(\lambda; \alpha, \theta) d\lambda, \qquad (6)$$

where $P(X = x \mid \lambda)$ is defined by

$$P(X = x \mid \lambda) = \binom{r+x-1}{x} e^{-\lambda r}(1-e^{-\lambda})^x = \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j e^{-\lambda(r+j)}. \qquad (7)$$

By replaced (7) in (6) we have

$$P(X = x \mid \lambda) = \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \int_0^\infty e^{-\lambda(r+j)} f(\lambda; \alpha, \theta) d\lambda = \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j M_\lambda(-(r+j)). \qquad (8)$$

Then, the pmf of the $\mathrm{NB\text{-}S}(r, \alpha, \theta)$ in (5) is obtained by subsituting the moment generating function of the Sushila distribution (4) with $t = -(r+j)$ into (8). Finally, it can be written as

$$f(x; r, \alpha, \theta) = \frac{\theta^2}{\theta+1} \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \frac{\theta + \alpha(r+j) + 1}{(\theta + \alpha(r+j))^2}.$$

Next, we will represent the NB- Lindley distribution as a special case of the NB-S distribution. Furthermore, we will illustrate the pmf of NB-S with different values of r, α and θ in Figure 1.

Corollary 1 *If* $\alpha = 1$ *then the NB-S distribution reduces to the NB-Lindley distribution with pmf given by*

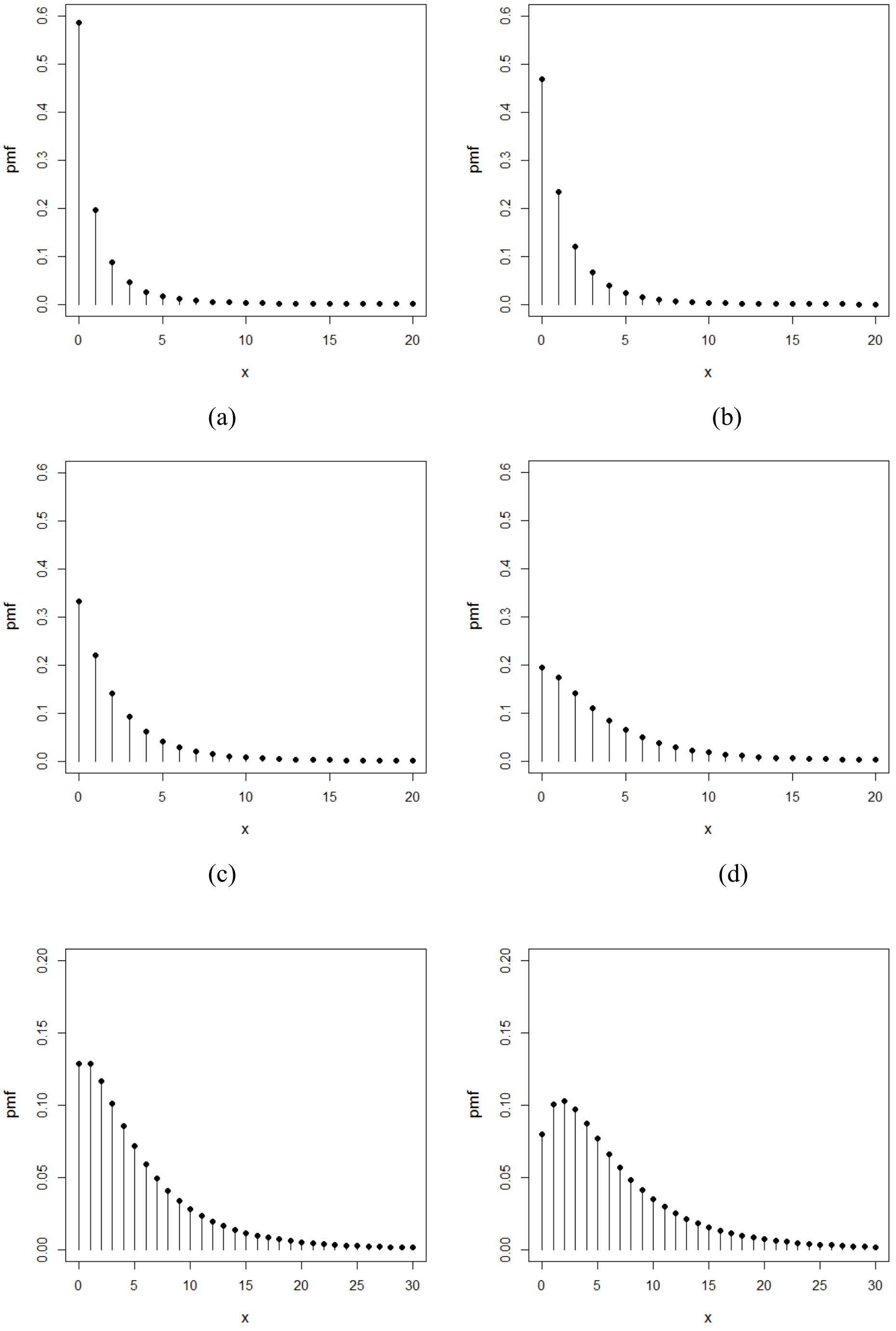
$$f(x; r, \theta) = \frac{\theta^2}{\theta+1} \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \frac{\theta + r + j + 1}{(\theta + r + j)^2}, \qquad (9)$$

where $x = 0,1,2,\ldots,$ *for* r *and* $\theta > 0$.

Proof: If $X \mid \lambda \sim \mathrm{NB}(r, p = \exp(-\lambda))$ and $\lambda \sim$ Sushila $(\alpha = 1, \theta)$, then the pmf of X is

$$f(x; r, \theta) \; = \; \frac{\theta^2}{\theta+1} \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \frac{\theta + r + j + 1}{(\theta + r + j)^2}.$$

From Corollary 1, we can find the NB-Lindley distribution represented in (9), that was introduced by Zamani and Ismai (2010).



(a) (b)

(c) (d)

(e) (f)

Figure 1 The pmf of a NB-S random variable (X) of some values of parameters (a) $r = 1$, $\alpha = 0.05$, $\theta = 0.15$ (b) $r = 3$, $\alpha = 0.10$, $\theta = 0.50$ (c) $r = 5$, $\alpha = 0.10$, $\theta = 0.50$ (d) $r = 10$, $\alpha = 0.05$, $\theta = 0.30$ (e) $r = 15$, $\alpha = 0.05$, $\theta = 0.30$ and (f) $r = 20$, $\alpha = 0.01$, $\theta = 0.07$

2.2. Some properties of the negative binomial-Sushila distribution

In this part, we introduce some basic properties of the NB-S distribution. We begin with the factorial moment of this distribution. The factorial moment is perhaps one of the most important for characteristics of distributions.

Theorem 2 *If* $X \sim$ NB-S(r,α,θ)*, then the factorial moment of order* k *of* X *is given by*

$$\mu_{[k]}(X)=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\frac{\theta^2(\theta-\alpha(k-j)+1)}{(\theta+1)(\theta-\alpha(k-j))^2}, \tag{10}$$

where $k=1,2,\ldots,$ *for* r,α *and* $\theta>0$.

Proof: From the factorial moment of order k of a mixed negative binomial distribution is shown (Gómez-Déniz et al. 2008), where $p=\exp(-\lambda)$ in (2) can be obtained by

$$\mu_{[k]}(X)=E_\lambda\left[\frac{\Gamma(r+k)}{\Gamma(r)}\frac{(1-e^{-\lambda})^k}{e^{-\lambda k}}\right]=\frac{\Gamma(r+k)}{\Gamma(r)}E_\lambda(e^{\lambda}-1)^k.$$

Using of a binomial expansion for the term $(e^{\lambda}-1)^k$ we can write $\mu_{[k]}(X)$ as

$$\mu_{[k]}(X)=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j E_\lambda(e^{\lambda(k-j)})=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j M_\lambda(k-j). \tag{11}$$

From the moment generating function of the Sushila distribution in (4) with $t=k-j,$ we insert (4) in (11). Finally, $\mu_{[k]}(X)$ is obtained

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

Corollary 2 *If* $\alpha=1$ *then the factorial moment of order* k *of the NB-S reduces to*

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

which is the same as the factorial moment of order k *of the NB-Lindley distribution.*

Proof: Substituting $\alpha=1$ into (10), we get

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

From the factorial moments of the NB-S distribution, it is straightforward to deduce the first four moments, variance and skewness respectively are

$$E(X)=\frac{r}{\delta_1}(\delta_2-\delta_1),$$

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

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

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

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

$$Skewness\,(X)=\left[E(X^3)-3E(X^2)E(X)+2(E(X))^3\right]/\sigma^3,$$

$$=\left[\frac{6r\delta_2-6r^2\delta_3+6r\delta_3+r^3\delta_4+3r^2\delta_4+2r\delta_4}{\delta_1}+\frac{6r^2\delta_2{}^2-3r^3\delta_2\delta_3-3r^2\delta_2\delta_3}{\delta_1^2}+\frac{2r^3\delta_2{}^3}{\delta_1^3}-2r\right]/\sigma^3,$$

where $\delta_1=\frac{\theta+1}{\theta^2}$, $\delta_2=\frac{-\alpha+\theta+1}{(\theta-\alpha)^2}$, $\delta_3=\frac{-2\alpha+\theta+1}{(\theta-2\alpha)^2}$, $\delta_4=\frac{-3\alpha+\theta+1}{(\theta-3\alpha)^2}$, $\delta_5=\frac{-4\alpha+\theta+1}{(\theta-4\alpha)^2}$, $\sqrt{Var(X)}=\sigma$, and $\theta\neq\alpha$.

2.3. Parameter estimations

In this section, the estimation of parameters for the NB-S (r,α,θ) via the maximum likelihood procedure is provided.

The likelihood function of the NB-S (r,α,θ) is given by

$$L(r,\alpha,\theta)=\prod_{i=1}^{n}\frac{\theta^2}{\theta+1}\binom{r+x_i-1}{x_i}\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{(\theta+\alpha(r+j))^2},$$

with the corresponding log-likelihood function:

$$\log L(r,\alpha,\theta)=\sum_{i=1}^{n}\left[\log\Gamma(r+x_i)-\log\Gamma(x_i+1)-\log\Gamma(r)\right]+2n\log(\theta)-n\log(\theta+1)$$
$$+\sum_{i=1}^{n}\log\left[\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\frac{\theta+\alpha(r+j)+1}{(\theta+\alpha(r+j))^2}\right]. \quad (12)$$

The log-likelihood function in (12) leads to the following partial derivatives with respect to r,α and θ which the optimal values of the parameters can be obtained. The score equations are derived as follows

$$\frac{\partial}{\partial r}\log L(r,\alpha,\theta)=\sum_{i=1}^{n}\psi(r+x_i)-n\psi(r)+\sum_{i=1}^{n}\left\{\left[\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\times\left(\frac{\alpha}{Z_j^2}-\frac{2\alpha(Z_j+1)}{Z_j^3}\right)\right]/\omega\right\},$$

$$\frac{\partial}{\partial\alpha}\log L(r,\alpha,\theta)=\sum_{i=1}^{n}\left\{\left[\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\left(\frac{r+j}{Z_j^2}-\frac{2(r+j)(Z_j+1)}{Z_j^3}\right)\right]/\omega\right\},$$

$$\frac{\partial}{\partial\theta}\log L(r,\alpha,\theta)=\left(\frac{2n}{\theta}-\frac{n}{\theta+1}\right)+\sum_{i=1}^{n}\left\{\left[\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\times\left(\frac{1}{Z_j^2}-\frac{2(Z_j+1)}{Z_j^3}\right)\right]/\omega\right\},$$

where $\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function, $Z_j=\theta+\alpha(r+j)$ 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 score equations can be solved numerically by using Newton-Raphson method. In this paper, we obtain the MLE estimates of $\hat{r}, \hat{\alpha}$ and $\hat{\theta}$ by using nlm function in stats package of R language (R Core Team 2015). R language provides a powerful and flexible system for statistical computations. A new-like method for unconstrained ploblems with at least first derivatives such as nlm often performs well (Nash 2014).

Table 1 Estimated parameters for number of claims under the policy

Number of claims	Observed	Expected by Poisson		Expected by NB		Expected by NB-S	
0	7840	7635.63		7847.02		7846.30	
1	1317	1636.72		1288.35		1299.21	
2	239	175.42		256.54		244.41	
3	42	12.53		54.07		52.97	
4	14	0.67	13.23	11.71		12.99	
5	4	0.03		2.58	15.02	3.54	
6	4	0		0.57		1.05	5.04
7	1	0		0.13		0.34	
8+	0	0		0.03		0.11	
Estimated parameters		$\hat{\lambda} = 0.2143$		$\hat{r} = 0.7015$		$\hat{r} = 2.0180$	
				$\hat{p} = 0.7659$		$\hat{\alpha} = 0.0100$	
						$\hat{\theta} = 0.1881$	
Chi-squared (Degree of freedom)		293.4276 (2)		8.7776 (2)		5.8079 (2)	
KS test statistic		0.0216		0.0023		0.0012	
-Log likelihood		5490.78		5348.04		5344.00	
AIC		10983.56		10700.08		10694.02	

3. Results and Discussion

The NB-S distribution is applied on count data. We fit the NB-S distribution with a real data set. The data set is the number of claims under the policy for the 9,461 automobile insurance policies. This data obtained from Klungman et al. (2012) is recorded in Table 1. The data set is an over-dispersed data with mean = 0.2144, variance = 0.2889 and index of dispersion = 1.347. It is fitted by the Poisson, the NB and the NB-S distributions. The maximum likelihood method provides parameter estimations. The performances of the model fittings in Table 1 are compared by using Chi-squared statistic, Kolmogorov - Smirnov (KS) test statistic for discrete distributions (Arnold and Emerson 2011), negative Log likelihood and Akaike Information Criteria (AIC). Those values of the NB-S distribution are the smallest of competitive distributions. As a result, the NB-S distribution provides a better fit for this data compared to the Poisson and NB distributions.

4. Conclusions

In this study, we introduce the NB-S distribution which is obtained by mixing the NB distribution with the Sushila distribution. We present that the NB-Lindley distribution is a special case of this distribution. In particular, the factorial moments, the first four moments variance and skewness of the NB-S distribution are derived. In addition, the parameter estimations are shown via the maximum likelihood method. For application to real data set, it shows that the NB-S distribution fits the data better than the Poisson and NB distributions. We hope that the NB-S distribution may attract wider applications in over-dispersed count data analysis. The NB-S distribution can be extened to include covairates in the model. In future study, we will consider regression models based on the NB-S distribution for count data.

Acknowledgements

The authors would like to thank Department of Statistics, Faculty of Science, Kasetsart University for supporting the scholarship and Graduate school Kasetsart University.

References

Arnold TB, Emerson, JW. Nonparametric goodness-of-fit tests for discrete null distributions. R J. 2011; 3: 34-39.

Aryuyuen S, Bodhisuwan W. The negative binomial-generalized exponential (NB-GE) Distribution. Appl. Math. Sci. 2013; 7: 1093-1105.

Ghitany ME, Atieh B, Nadarajah S. Lindley distribution and its application. Math. Comput. Simulat. 2008; 78: 493-506.

Gómez-Déniz E, Sarabia JM, Calder´ın-Ojeda E. Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications. Insur. Math. Econ. 2008; 42: 39-49.

Greene W. Functional forms for the negative binomial model for count data. Econ. Lett. 2008; 99: 585-590.

Johnson LN, Kemp WA, Kotz S. Discrete Univariate Distributions. 3rd ed. Wiley; 2005.

Klungman AS, Harry P H, Willmot EG. Loss models: From data to decisions. 4th ed. Willey; 2012.

Kongrod S, Bodhisuwan W, Payakkapong P. The negative binomial-Erlang (NB-EL) distribution with applications. Int. J. Pure Appl. Math. 2014; 92: 389-401.

Li S, Yang F, Famoye F, Lee C, Black D. Quasi-negative binomial distribution: Properties and applications. Comput. Stat. Data Anal. 2011; 55: 2363-2371.

Nash JC. On Best Practice Optimization methods in R. J. Stat. Software. 2014; 60: 1-14.

Pudprommarat C, Bodhisuwan W, Zeephongsekul P. A new mixed negative binomial distribution. J. Appl. Sci. 2012; 12: 1853-1858.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2015. Available from: http://www.R-project.org/.

Shanker R, Sharma S, Shanker U, Shanker R. Sushila distribution and its application to waiting times data. Opinion. 2013; 3: 1-11.

Wang Z. One mixed negative binomial distribution with application. J. Stat. Plann. Infer. 2011; 141: 1153-1160.

Zamani H, Ismail N. Negative binomial-Lindley distribution and its application. J. Math. Stat. 2010; 6: 4-9.