Pakistan Journal of Statistics and Operation Research

The Negative Binomial-New Generalized Lindley Distribution for Count Data: Properties and Application

Sirinapa Aryuyuen[1*]

* Corresponding Author

1. Department of Mathematics and Computer Science, Faculty of Science and Technology, Rajamangala University of Technology Thanyaburi, Pathum Thani, 12110, Thailand, sirinapa_a@rmutt.ac.th

Abstract

In this paper, a new mixture distribution for count data, namely the negative binomial-new generalized Lindley (NB-NGL) distribution is proposed. The NB-NGL distribution has four parameters, and is a flexible alternative for analyzing count data, especially when there is over-dispersion in the data. The proposed distribution has sub-models such as the negative binomial-Lindley (NB-L), negative binomial-gamma (NB-G), and negative binomial-exponential (NB-E) distributions as the special cases. Some properties of the proposed distribution are derived, i.e., the moments and order statistics density function. The unknown parameters of the NB-NGL distribution are estimated by using the maximum likelihood estimation. The results of the simulation study show that the maximum likelihood estimators give the parameter estimates close to the parameter when the sample is large. Application of NB-NGL distribution is carry out on three samples of medical data, industry data, and insurance data. Based on the results, it is shown that the proposed distribution provides a better fit compared to the Poisson, negative binomial, and its sub-model for count data.

Key Words: Count data; Over-dispersion; Mixture distribution; Negative binomial-new generalized Lindley distribution; Maximum likelihood estimation.

Mathematical Subject Classification: 60E05, 62E10, 62E20

1. Introduction

A Poisson distribution is typically used to fit count data when the number of phenomena is randomly distributed over the time and/or space in which the counts of the phenomenon occur. Equality of the mean and variance is characteristic of the Poisson distribution. Let X be a Poisson random variable with the parameter μ. Then, the probability mass function (pmf) of X is given by

$$f(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!};\ x = 0,1,2,\ldots \text{ and } \mu > 0. \tag{1}$$

However, in practice, the observed count data often display features like over-dispersion or under-dispersion, which is common in applied data analysis (Rainer, 2000). Greenwood and Yule (1920) suggested a model in which the mean of the Poisson distribution has a gamma distribution namely the negative binomial (NB) distribution. The NB distribution has become increasingly popular as a more flexible alternative to the Poisson distribution (Johnson et al., 2005). Let X be the number of r successes that occur for a given number of x failures. Note that p still refers to the probability of success. We have that X be a random variable distributed as the NB distribution with the parameters r and p, denoted by $X \sim NB(r,p)$, and its pmf is

$$f(x;r,p)=\binom{r+x-1}{x}p^{r}(1-p)^{x};\ x=0,1,2,...,r>0\ \text{ and }\ 0<p<1. \tag{2}$$

The NB distribution is better for overdispersed count data that are not necessarily heavy-tailed. The extreme heavy tail implies overdispersion, but the converse does not hold (Wang, 2011). The traditional statistical distributions or models, such as the Poisson and NB distributions, cannot be used effectively for count data with a heavy tail. The Poisson distribution tends to underestimate the number of zeros given the mean of the data while the NB distributions may overestimate zeros and underestimate observations to be a count data (Lord and Geedipally, 2011). Many researchers proposed the mixture distribution, which is one of the most important ways to obtain new probability distributions in applied probability and operational research (Gómez-Déniz et al., 2008). In this study, we are considering the mixture NB distribution as a more flexible alternative to analyze count data, especially, count data with over-dispersion. It is a mix between the NB distribution and a lifetime distribution.

Elbatal et al. (2013) proposed a new generalized Lindley (NGL) distribution as an alternative for modeling lifetime data in many areas. Let λ be a random variable distributed as the NGL distribution with parameters α, β and θ, i.e., $\lambda\sim \text{NGL}(\alpha,\beta,\theta)$. Then the probability density function (pdf) of λ is given by

$$g(\lambda;\alpha,\beta,\theta)=\frac{1}{1+\theta}\left(\frac{\theta^{\alpha+1}\lambda^{\alpha-1}}{\Gamma(\alpha)}+\frac{\theta^{\beta}\lambda^{\beta-1}}{\Gamma(\beta)}\right)e^{-\theta\lambda};\ \ \lambda>0. \tag{3}$$

The corresponding moment generating function (mgf) of λ is

$$M_{\lambda}(t;\alpha,\beta,\theta)=\int_{0}^{\infty}e^{t\lambda}g(\lambda;\alpha,\beta,\theta)\,d\lambda=\frac{1}{1+\theta}\left(\frac{\theta^{\alpha+1}}{(\theta-t)^{\alpha}}+\frac{\theta^{\beta}}{(\theta-t)^{\beta}}\right);\ t>0. \tag{4}$$

Three sub-models of the NGL distribution as follows; (i) if $\alpha=1$ and $\beta=2$, we get the Lindley distribution with a parameter θ, which is proposed by Lindley in 1958 (see Ghitany et al., 2008), (ii) if $\alpha=\beta$, we get the gamma distribution with parameters α and θ (Jambunathan, 1954), and (iii) if $\alpha=\beta=1$, we get the exponential distribution with a parameter θ (Gupta and Kundu, 1999).

The contents of the article are structured as follows. In Section 2, a new mixed negative binomial distribution by mixing the NB and NGL distributions to create the negative binomial-new generalized Lindley distribution, is proposed. In Section 3, we present some characteristic properties of the proposed distribution. In Section 4, the method to estimate unknown parameters of the proposed distribution is introduced. Next, we illustrate simulation study and application study of the proposed distribution with three real data sets in Section 5. Finally, the conclusion is provided in Section 6.

2. A new mixture distribution for count data

In this section, we provide the definition and theorem of the new mixture negative binomial distribution. Next, its sub-model is provided.

Definition 1. Let X be a random variable distributed as the NB distribution with parameters r and $p=e^{-\lambda}$ where λ be a random variable distributed as the NGL distribution with parameters α, β and θ, i.e., $X\,|\,\lambda\sim \text{NB}(r,p=e^{-\lambda})$ and $\lambda\sim \text{NGL}(\alpha,\beta,\theta)$. Thus, X be a random variable distributed as a negative binomial-new generalized Lindley (NB-NGL) distribution with parameters r, α, β and θ, denoted by $X\sim \text{NB-NGL}(r,\alpha,\beta,\theta)$.

Theorem 1. Let $X\sim \text{NB-NGL}(r,\alpha,\beta,\theta)$. Then, the pmf of X is

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

where $x=0,1,2,...$ and parameters r, α, β and θ.

Proof. If $X \mid \lambda \sim NB(r, p = e^{-\lambda})$ with the pmf in (2) and $\lambda \sim NGL(\alpha, \beta, \theta)$ with the pdf in (3), then the pmf of X can be obtained by

$$f(x; r, \alpha, \beta, \theta) = \int_0^\infty f(x \mid \lambda) g(\lambda; \alpha, \beta, \theta)\, d\lambda, \tag{6}$$

where

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

By substituting (7) into (6), we obtain

$$\begin{aligned} f(x; r, \alpha, \beta, \theta) &= \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \int_0^\infty e^{-\lambda(r+j)} g(\lambda; \alpha, \beta, \theta)\, d\lambda \\ &= \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j M_\lambda\left(-(r+j); \alpha, \beta, \theta\right). \end{aligned} \tag{8}$$

By replacing the mgf of NGL distribution in (4) with $t = -(r+j)$ into (8), then the pmf of X is

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

Some pmf plots of the NB-NGL distribution with some fixed values of parameters r, α, β and θ are shown in Figure 1.

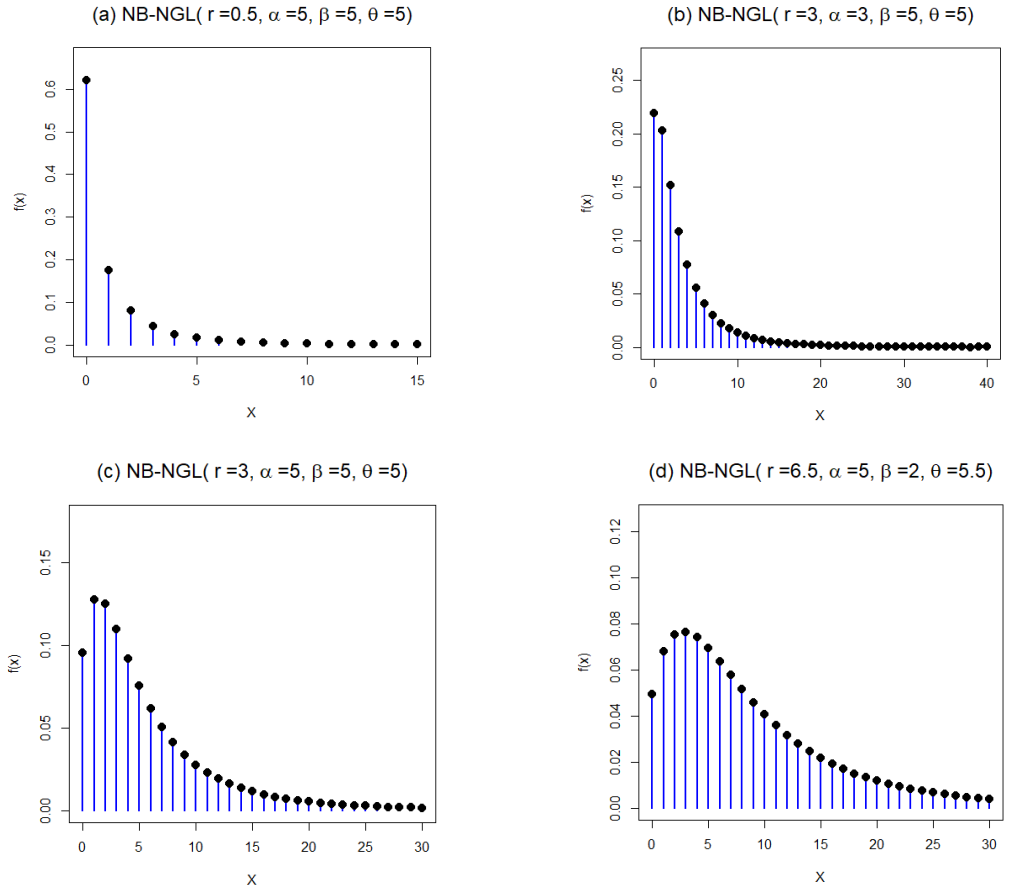


Figure 1: The pmf plots of the NB-NGL distribution for some fixed value of parameters

The NB-NGL distribution has three sub-models as follows.

Corollary 1. Let $X \sim$ NB-NGL(r,α,β,θ). If $\alpha=1$ and $\beta=2$, we get the negative binomial-Lindley (NB-L) distribution with positive parameters r and θ. Its pmf is

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

where the NB-L distribution was proposed by Zamani and Ismail (2010).

Proof. If $X \sim$ NB-NGL(r,α,β,θ). and substituting $\alpha=1$ and $\beta=2$ in (5) then pmf of X is given by

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

which is the pmf of NB-L distribution. In the same way, we get the pmf of the NB-L distribution as follow; If λ has the Lindley distribution (see Zamani and Ismail, 2010) with the pdf and mgf as

$$g(\lambda;\theta)=\frac{\theta^2}{1+\theta}(1+\lambda)e^{-\theta\lambda};\ \lambda>0,\theta>0, \tag{10}$$

$$M_\lambda(t;\theta)=\int_0^\infty e^{t\lambda}g(\lambda;\theta)\,d\lambda=\frac{\theta^2}{1+\theta}\frac{\theta-t+1}{(\theta-t)^2};\ t>0. \tag{11}$$

If $X\mid\lambda \sim$ NB$(r,p=e^{-\lambda})$ with the pmf in (2) and $\lambda \sim$ Lindley(θ) with the pdf and mgf as (3) and (4) respectively. The pmf of X is obtained by

$$f(x;r,\theta)=\int_0^\infty f(x\mid\lambda)g(\lambda;\theta)\,d\lambda=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^jM_\lambda\left(-(r+j);\theta\right). \tag{12}$$

By replacing the mgf of the Lindley distribution in (11) with $t=-(r+j)$ into (12), then the pmf of X is

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

Corollary 2. Let $X \sim$ NB-NGL(r,α,β,θ). If $\alpha=\beta$, we get the negative binomial-gamma (NB-G) distribution with positive parameters r, α and θ. The pmf of the NB-G distribution is

$$f(x;r,\alpha,\theta)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\left(\frac{\theta}{\theta+r+j}\right)^\alpha;\ x=0,1,2,.... \tag{13}$$

Proof. If $X \sim$ NB-NGL(r,α,β,θ), and we substitute $\alpha=\beta$ in (5), then the pmf of X is

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

which is the pmf of the NB-G distribution (Gençtürk and Yiğiter, 2016)

Corollary 3. Let $X \sim$ NB-NGL(r,α,β,θ). If $\alpha=\beta=1$ we get the negative binomial-exponential (NB-E) distribution. The pmf of the NB-E distribution is

$$f(x;r,\theta)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta}{\theta+r+j};\ x=0,1,2,.... \tag{14}$$

Proof. If $X \sim$ NB-NGL(r,α,β,θ), and we substitute $\alpha=\beta=1$ in (5) then the pmf of X is

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

which is the pmf of the NB-E distribution (Panjer and Willmot, 1981)

3. Mathematical properties

Some properties of the proposed distribution including the moments and order statistics density function, are introduced in section.

3.1 Moments

Theorem 2. If $X \sim \text{NB-NGL}(r,\alpha,\beta,\theta)$. then the m*th* factorial moment of X is

$$\mu_{[m]}(x;r,\alpha,\beta,\theta)=\frac{\Gamma(r+m)}{(1+\theta)\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^{j}\left(\frac{\theta^{\alpha+1}}{(\theta-(m-j))^{\alpha}}+\frac{\theta^{\beta}}{(\theta-(m-j))^{\beta}}\right), \tag{15}$$

where $m=1,2,3,...,$ and $r,\alpha,\beta,\theta>0.$

Proof. The *m*th factorial moment of X, i.e., $\mu_{[m]}(x;r,p)=E[X(X-1)\cdots(X-m+1)]$, is

$$\mu_{[m]}(x;r,p)=\frac{\Gamma(r+m)}{\Gamma(r)}\frac{(1-p)^{m}}{p^{m}};\ m=1,2,3,..., \tag{16}$$

where $\Gamma(\cdot)$ is the complete gamma function, i.e., $\Gamma(t)=\int_0^{\infty}x^{t-1}e^{-x}dx,\ t>0.$ From (16), where $p=e^{-\lambda}$ we can write it as follows (Gómez-Déniz et al. 2008)

$$\mu_{[m]}(x;r,e^{-\lambda})=E_{\lambda}\left[\frac{\Gamma(r+m)}{\Gamma(r)}\frac{(1-e^{-\lambda})^{m}}{e^{-\lambda m}}\right]=\frac{\Gamma(r+m)}{\Gamma(r)}E_{\lambda}\left(e^{\lambda}-1\right)^{m}. \tag{17}$$

Using a binomial expansion in the term $(e^{\lambda}-1)^{m}$ we can write (17) as

$$\mu_{[m]}(x;r,e^{-\lambda})=\frac{\Gamma(r+m)}{\Gamma(r)}\sum_{j=0}^{m}(-1)^{j}E_{\lambda}\left(e^{\lambda(m-j)}\right)=\frac{\Gamma(r+m)}{\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^{j}M_{\lambda}\left(m-j\right).$$

If $X\,|\,\lambda\sim\text{NB}(r,p=e^{-\lambda})$ and $\lambda\sim\text{NGL}(\alpha,\beta,\theta)$ when substituting the mgf of λ as in (4) with $t=(m-j)$ into $\mu_{[m]}(x;r,e^{-\lambda})$. The m*th* factorial moment of X is

$$\mu_{[m]}(x;r,\alpha,\beta,\theta)=\frac{\Gamma(r+m)}{(1+\theta)\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^{j}\left(\frac{\theta^{\alpha+1}}{(\theta-(m-j))^{\alpha}}+\frac{\theta^{\beta}}{(\theta-(m-j))^{\beta}}\right).$$

From the factorial moment of the NB-NGL distribution, we have the first four moments about zero, the variance and skewness, respectively

$$E(X)=r(\pi_1-1),$$

$$E(X^2)=(r^2+r)\pi_2-(2r^2+r)\pi_1+r^2,$$

$$E(X^3)=(r^3+3r^2+2r)\pi_3-(3r^3+6r^2+3r)\pi_2+(3r^3+3r^2+r)\pi_1-r^3,$$

$$E(X^4)=(r^4+6r^3+11r^2+6r)\pi_4-(4r^4+18r^3+26r^2+12r)\pi_3+(6r^4+18r^3+19r^2+7r)\pi_2-(4r^4+6r^3+4r^2+r)\pi_1+r^4,$$

$$V(X)=(r^2+r)\pi_2-r\pi_1(1+r\pi_1),$$

$$\text{Skewness}=\left\{E(X^3)-3E(X^2)E(X)+2E(X^3)\right\}\Big/\sigma_X^3=\left\{(r^3+3r^2+2r)\pi_3-(3r^3+6r^2+3r)\pi_2+(3r^3+3r^2+r)\pi_1-r^3\right.$$
$$\left.-r\left(\pi_1-1\right)\left[3(r^2+r)\pi_2-\left(2r^2+3r+2r^2\pi_1\right)\pi_1+r^2\right]\right\}\Big/\sigma_X^3,$$

$$\text{Kurtosis}=\left\{E(X^4)-4E(X^3)E(X)+6E(X^2)[E(X)]^2-3[E(X)]^4\right\}\Big/\sigma_X^4=\left\{(r^4+6r^3+11r^2+6r)\pi_4\right.$$
$$-(4r^4+18r^3+26r^2+12r)\pi_3+(6r^4+18r^3+19r^2+7r)\pi_2-(4r^4+6r^3+4r^2+r)\pi_1+r^4$$
$$-4r\left(\pi_1-1\right)\left[(r^3+3r^2+2r)\pi_3-(3r^3+6r^2+3r)\pi_2+(3r^3+3r^2+r)\pi_1-r^3\right]$$
$$\left.+3r^2\left(\pi_1-1\right)^2\left[2(r^2+r)\pi_2-\left(2(r^2+r)+r^2\pi_1\right)\pi_1+r^2\right]\right\}\Big/\sigma_X^4,$$

where $\pi_c = \frac{1}{1+\theta}\left(\frac{\theta^{\alpha+1}}{(\theta-c)^{\alpha}} + \frac{\theta^{\beta}}{(\theta-c)^{\beta}}\right)$ and $\sigma_X = \sqrt{V(X)}$.

3.2 Order statistic density function

Let $X_1, X_2, ..., X_n$ be n independent and identically distributed (iid) random variables defined on Ω with the cumulative density function (cdf) $F_X(x)$ and the pmf $f_X(x)$. Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote these random variables rearranged in non-descending order of magnitude. Thus, $X_{(k)}$ is the k*th* smallest number in the sample, $k = 1, 2, ..., n$. Because order statistics are random variables, it is possible to compute probability values associated with values in their support. The k*th* order statistics density function of $X_{(k)}$ is (e.g., Casella and Berger, 2002) given by

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} f_X(x)\left[F_X(x)\right]^{k-1}\left[1-F_X(x)\right]^{n-k};\ x \in \Omega. \tag{18}$$

If $X_1, X_2, ..., X_n$ be n iid variables with the pmf $f_X(x)$ as in (5) and cdf $F_X(x)$ as follows

$$F_X(x) = \sum_{s\le x} f_X(x) = \sum_{s\le x}\binom{r+s-1}{s}\sum_{j=0}^{s}\binom{s}{j}(-1)^j \frac{1}{\theta+1}\left(\frac{\theta^{\alpha+1}}{(\theta+r+j)^{\alpha}} + \frac{\theta^{\beta}}{(\theta+r+j)^{\beta}}\right)$$

Definition 2. Let $X \sim$ NB-NGL$(r, \alpha, \beta, \theta)$. Then, the order statistic density function of X is

$$\begin{aligned} f_{X_{(k)}}(x) = {} & \frac{n!}{(k-1)!(n-k)!(}\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j \frac{1}{\theta+1}\left(\frac{\theta^{\alpha+1}}{(\theta+r+j)^{\alpha}} + \frac{\theta^{\beta}}{(\theta+r+j)^{\beta}}\right) \\ & \times\left\{\sum_{s\le x}\binom{r+s-1}{s}\sum_{j=0}^{s}\binom{s}{j}(-1)^j \frac{1}{\theta+1}\left(\frac{\theta^{\alpha+1}}{(\theta+r+j)^{\alpha}} + \frac{\theta^{\beta}}{(\theta+r+j)^{\beta}}\right)\right\}^{k-1} \\ & \times\left\{1-\sum_{s\le x}\binom{r+s-1}{s}\sum_{j=0}^{s}\binom{s}{j}(-1)^j \frac{1}{\theta+1}\left(\frac{\theta^{\alpha+1}}{(\theta+r+j)^{\alpha}} + \frac{\theta^{\beta}}{(\theta+r+j)^{\beta}}\right)\right\}^{n-k}, \end{aligned} \tag{19}$$

where $k = 1, 2, ..., n,\ s, x = 0, 1, 2, ...$ for $s \le x$ and $r, \alpha, \beta, \theta > 0$.

4. Maximum likelihood estimation

For this study, the unknown parameters of the proposed distribution are estimated via the method of maximum likelihood estimation (MLE). Let $\tilde{x} = (x_1, x_2, ..., x_n)$ be a random sample of size n from the NB-NGL distribution with parameters of $\tilde{\omega} = (r, \alpha, \beta, \theta)$. From the pmf of the NB-NGL distribution in (5), we have the likelihood function as follows

$$L(\tilde{x}; \tilde{\omega}) = \frac{1}{(1+\theta)^n}\prod_{i=1}^{n}\frac{\Gamma(r+x_i)}{\Gamma(r)\Gamma(x_i+1)}\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\left(\frac{\theta^{\alpha+1}}{(\theta+r+j)^{\alpha}} + \frac{\theta^{\beta}}{(\theta+r+j)^{\beta}}\right),$$

with the corresponding log-likelihood function, i.e., $\ell(\tilde{x}; \tilde{\omega}) = \log L(\tilde{x}; \tilde{\omega})$,

$$\begin{aligned} \ell(\tilde{x}; \tilde{\omega}) = {} & -n\log(1+\theta) + \sum_{i=1}^{n}\log\Gamma(r+x_i) - \sum_{i=1}^{n}\log\Gamma(r) - \sum_{i=1}^{n}\log\Gamma(x_i+1) \\ & + \sum_{i=1}^{n}\log\sum_{j=0}^{x_i}\binom{x_i}{j}(-1)^j\left(\frac{\theta^{\alpha+1}}{(\theta+r+j)^{\alpha}} + \frac{\theta^{\beta}}{(\theta+r+j)^{\beta}}\right). \end{aligned}$$

The partial derivatives $\ell(\tilde{x}; \tilde{\omega})$ with respect to r, α, β and θ are, respectively

$$\frac{\partial\ell(\tilde{x}; \tilde{\omega})}{\partial r} = 0, \frac{\partial\ell(\tilde{x}; \tilde{\omega})}{\partial\alpha} = 0, \frac{\partial\ell(\tilde{x}; \tilde{\omega})}{\partial\beta} = 0, \frac{\partial\ell(\tilde{x}; \tilde{\omega})}{\partial\theta} = 0. \tag{20}$$

The maximum likelihood estimators, $\hat{r},\hat{\alpha},\hat{\beta}$ and $\hat{\theta}$, are obtained by solving the expression (20). In this study, $\hat{r},\hat{\alpha},\hat{\beta}$ and $\hat{\theta}$, are obtained by using the numerical optimization with the **nlm** function in the ***stats*** package in R (R Core Team, 2018).

5. Simulation and application studies

5.1 Simulation

The simulation study of parameter estimation is illustrated for verification of the MLE performance before application to real data is illustrated. We conducted Monte Carlo simulation studies to assess on the finite sample behavior of the maximum likelihood estimators of r,α,β,θ. All results were obtained from 1000 replications ($T=1000$) and the simulations were carried out using the statistical software package R. In each replication a random sample of size n (i.e., $n=20$, 50, 100, 150 and 200) is drawn from the NB-NGL(r,α,β,θ) with two cases, i.e., (i) $r=0.5$, $\alpha=5$, $\beta=5$, $\theta=5$ and (ii) $r=3$, $\alpha=3$, $\beta=5$, $\theta=5$. Random variate generation of the NB-NGL distribution has been generated from follows.

(i) Generate λ from the NGL distribution with positive parameters $\alpha,\beta,$ and θ.

(ii) Generate X from the NB distribution with the parameters r and $p=e^{-\lambda}$.

The results of simulation study present the mean maximum likelihood estimates of four parameters, i.e.,

$$\hat{r}=\frac{1}{T}\sum_{t=1}^{T}\hat{r}_t,\ \hat{\alpha}=\frac{1}{T}\sum_{t=1}^{T}\hat{\alpha}_t,\ \hat{\beta}=\frac{1}{T}\sum_{t=1}^{T}\hat{\beta}_t,\ \text{and}\ \hat{\theta}=\frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t,$$

and the Root Mean Squared Errors (RMSE) of estimators, i.e.,

$$\text{RMSE}(\hat{r})=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{r}_t-r)^2},\ \text{RMSE}(\hat{\alpha})=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\alpha}_t-\alpha)^2},$$

$$\text{RMSE}(\hat{\beta})=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\beta}_t-\beta)^2},\quad \text{and}\ \text{RMSE}(\hat{\theta})=\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_t-\theta)^2}.$$

These results are shown in Table 1, and we notice that the RMSE values of the maximum likelihood estimators of r,α,β and θ decay toward zero as the sample size increases, as expected. The simulation study of parameter estimation is illustrated for verification of the MLE performance to estimate the proposed distribution. The results show that the maximum likelihood estimators give the parameter estimates close to the parameter when the sample is large ($n=200$)

Table 1: Statistic values of the NB-NGL parameter estimation by using the MLE

		Case 1				Case 2			
n		$r=0.5$	$\alpha=5$	$\beta=5$	$\theta=5$	$r=3$	$\alpha=5$	$\beta=5$	$\theta=5$
20	Estimate	2.0641	7.5268	9.2883	6.8231	3.4151	9.0536	4.7923	8.8767
	RMSE	3.4654	6.2030	6.6143	8.3313	5.2272	7.1718	3.3585	9.9589
50	Estimate	1.5316	8.4751	9.0835	6.3211	3.2951	6.4069	4.3781	6.1393
	RMSE	2.3571	7.3469	6.4617	7.3423	3.4035	2.8893	4.9896	3.9645
100	Estimate	0.8963	8.9364	8.8037	6.3764	3.1462	5.4977	4.4436	4.7436
	RMSE	1.2020	5.7132	5.2789	5.3984	3.3141	1.8272	3.4852	3.4501
150	Estimate	0.6787	7.2673	7.9638	5.7561	2.9322	5.3598	4.6076	4.6034
	RMSE	0.5357	3.4432	4.1797	5.2218	3.1803	1.5775	3.2877	2.7369
200	Estimate	0.5429	5.6641	5.8436	5.4971	2.9073	5.1217	3.8657	4.3329
	RMSE	0.1536	1.6727	1.6945	3.3271	2.9264	1.3691	2.1588	2.5161

5.2 Application to real data sets

We provide the application of the NB-NGL distribution to show its importance for count data analysis by considering three real data sets. These data sets are shown to fit by distributions, such as the Poisson, NB, NB-E, NB-G, NB-L and NB-NGL distributions. The goodness-of-fit of the Kolmogorov-Smirnov (K-S) is used to decide if a sample comes from a population with a specific distribution. For a discrete distribution, we used ks.test function in the dgof package in R to find the value of the K-S statistics (Marsaglia et al., 2003; R Core Team, 2018). The best distribution corresponds to the smaller statistic value of KS test. For two real data sets, we compare the fits of the proposed distribution with the Poisson, NB, and its sub-models of NB-NGL distribution, i.e., NB-L, NB-G, and NB-E distributions. The estimated parameters in each distribution are obtained by using the MLE with the **nlm** function in R. To compare the parameter estimation of each distribution, we consider the $\hat{\ell}$, which is the maximized value of the log-likelihood function under the considered distributions.

Data set I: The data are the number of hospital stays of United States residents age 66 and over from the National Medical Expenditure survey in 1987 and 1988 (see Flynn, 2009). This data set has a sample size of 4406. The mean and variance of the number of hospital stays are 0.296 and 0.557 respectively. The result of the expected values, and statistics of each distribution for this data set is provided in Table 1. Based on the KS test in Table 2 indicates that the NB-NGL distribution is a strong competitor to others considered for fitting data set (see Figure 2 (a)).

Data set II: The data set corresponds to an uncensored data set from Sankaran (1970) on the number of the mistakes in copying groups of random digits; it was used for illustrating the distribution of the Poisson, NB, NB-E, NB-G, NB-L and NB-NGL distributions. The statistic value of the KS test in Table 3 indicates that the NB-NGL distribution is a strong competitor to the others considered for fitting data set (see Figure 2 (b)).

Data set III: The data which was obtained from Klugman et al. (2008), provides information on 9,461 automobile insurance policies whereby the number of accidents of each policy is recorded. The distributions of the Poisson, NB, NB-E, NB-G, NB-L and NB-NGL are fitted to the data. Based on the log likelihood and KS test in Table 4, the NB-NGL provides a better fit for the numbers of hospital stays than the NB, NB-L, NB-G, NB-E, and Poisson distributions respectively. The plot comparing the expected values of the considered distributions with the observed value of count data are shown in Figure 2 (c).

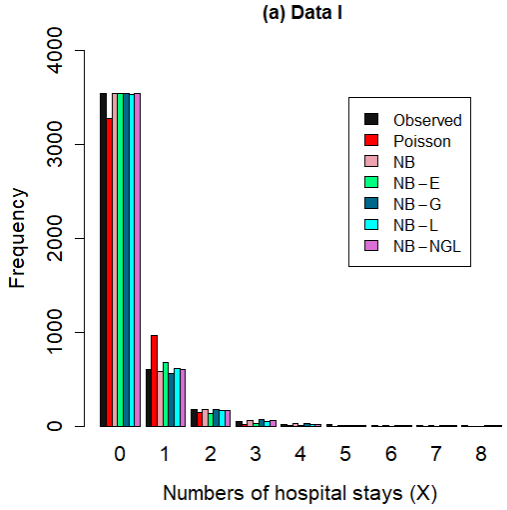
Table 2: Observed values, expected values, and statistics of each distribution for data set I

Numbers of hospital stays	Observed values	Observed values of fitting with distributions					
		Poisson	NB	NB-E	NB-G	NB-L	NB-NGL
0	3,541	3,277.26	3,544.26	3,547.43	3,542.93	3,537.55	3,540.89
1	599	969.94	583.56	677.09	556.89	612.07	601.72
2	176	143.53	177.53	131.29	176.10	161.25	166.46
3	48	14.16	62.27	25.98	65.88	54.07	56.57
4	20	1.05	23.29	5.24	26.78	21.30	21.95
5	12	0.06	9.04	1.08	11.47	9.43	9.39
6	5	0.00	3.59	0.23	5.11	4.57	4.33
7	1	0.00	1.45	0.05	2.34	2.37	2.13
8	4	0.00	0.59	0.01	1.10	1.31	1.10
Estimated value of parameters		$\hat{\lambda}$ =0.2960	$\hat{p}$ =0.5562 $\hat{r}$ =0.3710	$\hat{r}$ =38.4510 $\hat{\theta}$ =163.0027	$\hat{r}$ =31.4690 $\hat{\alpha}$ =0.2203 $\hat{\theta}$ =20.3122	$\hat{r}$ =1.3783 $\hat{\theta}$ =6.3925	$\hat{r}$ =0.6173 $\hat{\alpha}$ =3.8042 $\hat{\beta}$ =1.0138 $\hat{\theta}$ =9.6358
$-\hat{\ell}$		3,304.51	3,009.63	3,098.10	3,061.01	3,007.89	3,007.50
KS test		0.0599	0.0028	0.0192	0.0091	0.0022	0.0016

(*p*-value)	(<0.0001)	(0.9999)	(0.0781)	(0.8572)	(0.9999)	(0.9999)

Table 3: Observed values, expected values, and statistics of each distribution for data set II

Mistakes in copying groups	Observed values	Observed values of fitting with distributions					
		Poisson	NB	NB-E	NB-G	NB-L	NB-NGL
0	35	27.41	33.95	30.37	31.81	36.61	34.48
1	11	21.47	14.49	14.93	17.01	14.23	13.66
2	8	8.41	6.39	7.34	6.98	5.56	6.17
3	4	2.20	2.85	3.63	2.60	2.18	2.91
4	2	0.43	1.28	1.80	0.92	0.86	1.41
Estimated value of parameters		$\hat{\lambda}$ =0.7833	$\hat{p}$ =0.5449 $\hat{r}$ =0.9380	$\hat{r}$ =115.580 $\hat{\theta}$ =119.504	$\hat{r}$ =79.4892 $\hat{\alpha}$ =1.8817 $\hat{\theta}$ =200.659	$\hat{r}$ =150.531 $\hat{\theta}$ =236.678	$\hat{r}$ =0.7893 $\hat{\alpha}$ =134.264 $\hat{\beta}$ =41.1337 $\hat{\theta}$ =190.068
$-\hat{\ell}$		77.55	73.57	74.18	74.21	73.91	73.48
KS test		0.1264	0.0407	0.0771	0.0532	0.0808	0.0356
(*p*-value)		(0.2928)	(0.9999)	(0.8675)	(0.9957)	(0.8282)	(0.9999)



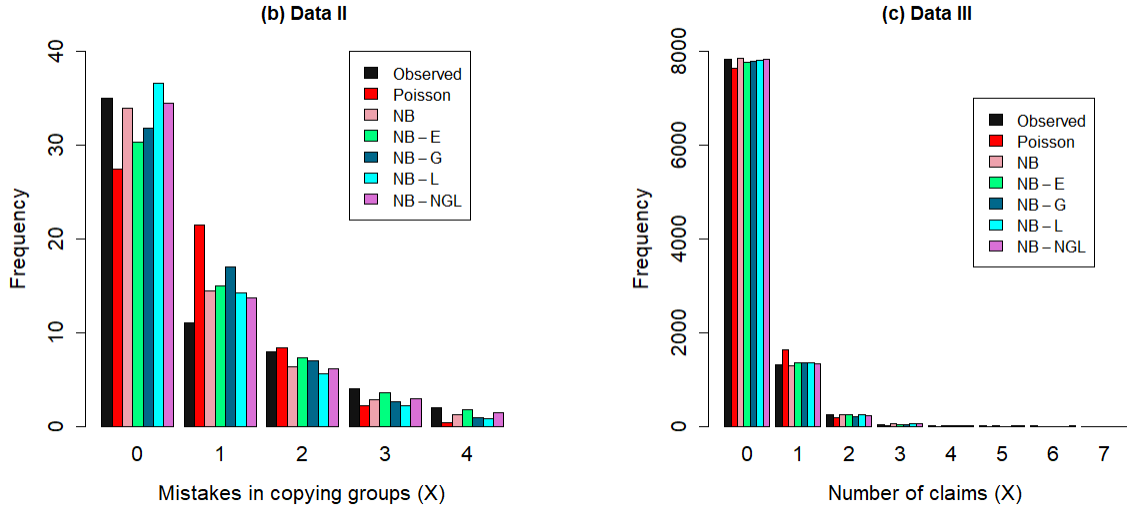


Figure 2: Plots of the observed and expected values of fitting distributions for three data sets

Table 4: Observed values, expected values, and statistics of each distribution for data set III

Number of claims	Observed values	Observed values of fitting with distributions					
		Poisson	NB	NB-E	NB-G	NB-L	NB-NGL
0	7,840	7,635.27	7,846.64	7,763.54	7,796.23	7,811.55	7,837.40
1	1,317	1,637.00	1,288.58	1,356.39	1,361.03	1,346.35	1,326.20
2	239	175.49	256.64	241.98	205.80	244.61	226.34
3	42	12.54	54.10	44.29	30.25	46.67	48.74
4	14	0.67	11.72	8.31	4.45	9.32	14.03
5	4	0.03	2.58	1.60	0.66	1.94	4.97
6	4	0.00	0.57	0.31	0.10	0.42	1.95
7	1	0.00	0.13	0.06	0.02	0.10	0.79
Estimated value of parameters		$\hat{\lambda}$ =0.2144	$\hat{p}$ =0.7659 $\hat{r}$ =0.7015	$\hat{r}$ =30.4775 $\hat{\theta}$ =143.966	$\hat{r}$ =21.7876 $\hat{\alpha}$ =1.4737 $\hat{\theta}$ =162.372	$\hat{r}$ =15.0765 $\hat{\theta}$ =72.3754	$\hat{r}$ =4.8261 $\hat{\alpha}$ =1.8355 $\hat{\beta}$ =12.1718 $\hat{\theta}$ =48.1251
$-\hat{\ell}$		5,490.78	5,396.29	5,396.29	5,431.77	5,348.97	5,341.93
KS test		0.0216	0.0023	0.0081	0.0066	0.0031	0.0007
(p-value)		(0.0003)	(0.9999)	(0.5669)	(0.8039)	(0.9999)	(0.9999)

6. Conclusions

In this paper, we proposed a new four-parameter distribution called the negative binomial-new generalized Lindley (NB-NGL) distribution that mixes the negative binomial distribution (Greenwood and Yule, 1920) and new generalized Lindley distribution (Elbatal et al., 2013). Its moments and order statistics density function are introduced. The negative binomial-Lindley (Zamani and Ismail, 2010), negative binomial-gamma (Gençtürk and Yiğiter, 2016), and negative binomial-exponential (Panger and Willmot, 1981) distributions are special cases of the NB-NGL distribution. The maximum likelihood estimation is used to estimate the unknown parameters of the proposed distribution. The results of the simulation study show that the maximum likelihood estimators give the parameter estimates close to the parameter when the sample is large. Finally, these real data sets are used to illustrate the fitting distribution by using the proposed distribution. The statistic value of the KS test indicates that the NB-NGL distribution is a strong competitor to the others considered for fitting data sets. I expect that the NB-NGL distribution will be a flexible alternative for count data analysis.

References

1. Casella, G. & R. Berger. (2002). Statistical Inference, 2nd ed. Duxbury Press, Pacific Grove, CA.
2. Elbatal, I., Merovci, F. & Elgarhy, M. (2013). A new generalized Lindley distribution. Mathematical Theory and Modeling, 3(13), 30-47.
3. Flynn, M. & Francis, L. A. (2009). More flexible GLMs zero-inflated models and hybrid models. Casualty Actuarial Soc, 148-224.
4. Gençtürk, Y. & Yiğiter, A. (2016). Modelling claim number using a new mixture model: negative binomial gamma distribution. Journal of Statistical Computation and Simulation, 86(10), 1829-1839.
5. Ghitany, M. E., Atieh, B. & Nadarajah, S. (2008). Lindley distribution and its application. Mathematics and Computers in Simulation, 78(4), 493-506.
6. Gómez-Déniz, E., Sarabia, J. M. & Calderín-Ojeda, E. (2008). Univariate and multivariate versions of the negative binomial-inverse Gaussian distributions with applications. Insurance: Mathematics and Economics, 42(1), 39-49.
7. Greenwood, M. & Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. Journal of the Royal Statistical Society, 83(2), 255-279.
8. Gupta, R. D. & Kundu, D. (1999). Theory & methods: Generalized exponential distributions. Australian & New Zealand Journal of Statistics, 41(2), 173-188.
9. Jambunathan M. B. (1954). Some properties of beta and gamma distributions. The annals of mathematical statistics. 25(2), 401-405.
10. Johnson, N. L., Kemp, A. W. & Kotz, S. (2005). Univariate discrete distributions, 444, John Wiley & Sons.
11. Klugman, S.A., Panjer H.H. & Willmot G.E. (2008). Loss Models: From Data to Decision. 3rd Edn., John Wiley and Sons, USA : 101-159.
12. Lord, D. & 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.
13. Marsaglia, G., Tsang, W. W. & Wang, J. (2003), Evaluating Kolmogorov's distribution. Journal of Statistical Software, 8(18), 1-4.
14. Panjer, H. H. & Willmot, G. E. (1981). Finite sum evaluation of the negative binomial-exponential model. ASTIN Bulletin: The Journal of the IAA, 12(2), 133-137.
15. R Core Team. (2018). R: A language and environment for statistical computing.
16. Rainer, W. (2000). Econometric Analysis of Count Data, 3rd edn, Springer Verlag, Berlin, Germany.
17. Sankaran, M. (1970). The discrete Poisson-Lindley distribution, Biometrics, 26(1), 145-149.
18. Wang, Z. (2011). One mixed negative binomial distribution with application. Journal of Statistical Planning and Inference, 141(3), 1153-1160.
19. Zamani, H. & Ismail, N. (2010). Negative binomial-Lindley distribution and its application. Journal of Mathematics and Statistics, 6(1), 4-9.