A negative binomial Erlang-Lindley distribution with applications

Somporn Thepchim[1*], Kajita Matchima[2], Thanakorn Suthison[3], Yaovaruk Thongphum[4]

Received: 03 October 2023; Revised: 22 April 2024; Accepted: 07 May 2024

Abstract

In this paper, A negative binomial Erlang-Lindley distribution obtained by mixing the negative binomial distribution and the two-parameter Erlang-Lindley distribution is introduced for modeling count data. Its probability mass function, factorial moment, mean, variance and index of dispersion have been obtained and discussed. Estimation of the parameters is illustrated using the maximum likelihood method. In conclusion, real datasets were employed to demonstrate the discussed theory. We anticipate that this distribution can offer an alternative approach for modeling count data.

Keywords: negative binomial distribution, Erlang-Lindley distribution, Count data, Maximum likelihood estimation, overdispersion

[*]*somporn.t@ubru.ac.th*
[1,2,3,4]*Program of Mathematics, Faculty of Science, Ubon Ratchathani Rajabhat University*

1. Introduction

Count data consists of non-negative integer values that represent the occurrences of an event. Examples include the number of reported COVID-19 cases, the instances of monkeypox cases, claims under an insurance policy, cases of dengue hemorrhagic fever per week, and occurrences of traffic violations for drivers. These instances exemplify various applications of count data in different contexts.

In statistical analysis, count data often follows a discrete probability distribution, such as the Poisson distribution or the negative binomial distribution. The Poisson distribution is appropriate for modeling count data when the events occur at a constant rate over time or space and are independent of each other. It is defined for non-negative integer values and has a single parameter, usually denoted as λ, representing the mean and variance of the distribution. These distributions are commonly used to model and analyze count data due to their ability to handle non-negative integer values and their overdispersion (variance exceeding the mean), which is often observed in count data. Additionally, the Poisson distribution may not be the best fit. The negative binomial distribution is a generalization of the Poisson distribution that can handle overdispersed count data. It is characterized by two parameters: r (the number of successes until the experiment is stopped) and p (the probability of success on each trial). The variance of the negative binomial distribution is greater than the mean, allowing it to account for overdispersion commonly seen in count data. The mixed negative binomial distributions can be considered a compelling alternative for analyzing count data characterized by overdispersion.

Some articles also used other mixed distributions, a mixed negative binomial, which provided a better fit compared to Poisson and negative binomial distributions, such as the negative binomial-new weighted Lindley distribution (Samutwachirawong, 2017), generalized negative binomial (GNB) distributions (Korolev & Zeifman, 2019), negative binomial-quasi Lindley (NB-QL) distribution (Aryuyuen & Tonggumnead, 2022)

Abd EI-Monsef et al. (2017) proposed a two-parameter Erlang-Lindley (ErL) distribution is a mixture of Lindley and Erlang distributions. Its probability density function (pdf) is

$$f(x)=\frac{\theta^3}{(\theta+1)}\left(\frac{x+1}{\theta+1}+\frac{\theta^{k-3}x^{k-1}}{(k-1)!}\right)e^{-\theta x}, \tag{1.1}$$

where $x\geq 0,\ \theta>0,\ k=1,2,3,...$

The cumulative distribution function is

$$F(x)=1-\frac{\theta e^{-x\theta}\Gamma(k)(1+\theta+x\theta)+(1+\theta)\Gamma(k,x\theta)}{\Gamma(k)(1+\theta)^2}, \qquad x\geq 0,\ \theta>0, k=1,2,3,... \tag{1.2}$$

The moment generating $M_x(t)$ is

$$M_x(t)=\frac{\theta^k(1+\theta)(\theta-t)^{2-k}+\theta^3(1+\theta-t)}{(1+\theta)^2(\theta-t)^2}. \tag{1.3}$$

The parameter k is the shape parameter and θ is the rate parameter. The following summary of the key characteristics of the density function of ErL distribution is decreasing when $k=1,\ \forall\theta$ and $k=2$, $k=3$ at $\theta>1$, ErL distribution is unimodal when $k=2,\ k=3$, $k=3$ at $0<\theta\leq 1$, ErL distribution is bimodal when $k>3$, at $0<\theta<1$ and decreasing-increasing-decreasing, It indicates a fluctuation or variation in the trend of the sequence when $k>3$, at $\theta\geq 1$. The application of the ErL distribution demonstrates that, for some information criteria, it yields a better fit compared to the Weibull, inverse Gaussian, Lindley, and log-logistic distributions, as evidenced by its smaller values.

The motivation for proposing A negative binomial Erlang-Lindley distribution for count data analysis is the usefulness of the ErL distribution illustrated in an analysis of lifetime models. Its probability mass function, factorial moment, mean, variance, and index of dispersion have been obtained and discussed. The parameters of the negative binomial-Erlang-Lindley (NB-ErL) distribution are estimated by the maximum likelihood estimation. The NB-ErL distribution is a probability distribution that is often used for comparing the goodness-of-fit of count data to the Poisson and negative binomial distributions. It considers the observed values and expected values to determine the fit of each distribution, through a comparison between the NB-ErL distribution and the observed data. The distribution is applied when the count data with overdispersion.

2. A new three-parameter negative binomial distribution discrete distribution

We introduce a new discrete three-parameter distribution known as the negative binomial-Erlang-Lindley (NB-ErL) distribution. This distribution is derived through a mixing of the negative binomial distribution and the Erlang-Lindley distribution as described in (1.1).

Theorem 2.1 The probability mass function (pmf) of a Negative Binomial-Erlang-Lindley distribution is given by

$$f(x;r,\theta,k)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}\frac{\theta^{k}(1+\theta)(\theta+r+j)^{2-k}+\theta^{3}(1+\theta+r+j)}{(1+\theta)^{2}(\theta+r+j)^{2}}, \tag{2.1}$$

where $x=0,1,2,...,\ ,r>0,\theta>0$ and $k=1,2,3,...$

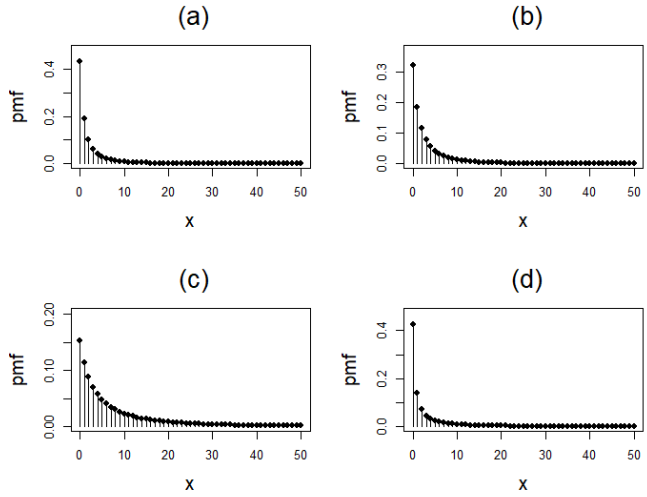

Figure 1 The pmf of the NB-ErL distribution of some specified values of *r*, θ , and *k*.

Proof. We assume a that is distributed as the ErL distribution with parameter k is the shape parameter and θ is the rate parameter.

If $X|a\sim \mathrm{NB}(r,p=\exp(-a))$ and $a\sim \mathrm{ErL}\ (\theta,k))$, then the pmf of X can be obtained by

$$f(x)=\int_{0}^{\infty}f_{1}(x|a)g(\lambda;\theta,k)da,$$

where $f_1(x|a)$ is express as $f_{1}(x|a)=\binom{r+x-1}{x}\exp(-ar)(1-\exp(-a))^{x},$

$$=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}\exp(-a(r+j)).$$

By substituting $f_1(x|a)$ into $f(x)=\int_{0}^{\infty}f_{1}(x|a)g(a;\theta,k)da,$ we obtain

$$f(x)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}\left(\int_{0}^{\infty}\exp(-a(r+j))g(a;\theta,k)da\right),\rightarrow$$

$$=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^{j}M_{a}(-(r+j)),$$

where $M_a(-(r+j))$ is the moment generating function (mgf) of the ErL distribution in (1.3). Replacing the mgf of the ErL distribution in the equation above, we have the pmf of the NB-ErL distribution is given as

$$f(x;r,\theta,k)=\binom{r+x-1}{x}\sum_{j=0}^{x}\binom{x}{j}(-1)^j\frac{\theta^k(1+\theta)(\theta+r+j)^{2-k}+\theta^3(1+\theta+r+j)}{(1+\theta)^2(\theta+r+j)^2}.$$

3. Characteristics of the NB-ErL distribution

Definition 3.1 If $X:$ NB-ErL (r,θ,k), then the factorial moment of order m is given by

$$\mu_{[m]}(X)=\frac{\Gamma(r+m)}{\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^j\frac{\theta^k(1+\theta)(\theta-m+j)^{2-k}+\theta^3(1+\theta-m+j)}{(1+\theta)^2(\theta-m+j)^2}, \qquad (3.1)$$

with $m=1,2,3,\ldots$

Proof. Gomez et al. (2010) demonstrated the expression for the factorial moment of order m of the mixed negative binomial distribution as follows:

$$\mu_{[m]}(X)=E_a\left(\frac{\Gamma(r+m)}{\Gamma(r)}\frac{(1-\exp(-a))^m}{\exp(-am)}\right),$$

$$=\frac{\Gamma(r+m)}{\Gamma(r)}E_a(\exp(a)-1)^m,\ \text{using a binomial expansion of } (\exp(a)-1)^m, \text{ then shows}$$

that $\mu_{[m]}(X)$ can be written as

$$\mu_{[m]}(X)=\frac{\Gamma(r+m)}{\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^jE_a(\exp(a(m-j))),$$

$$=\frac{\Gamma(r+m)}{\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^jM_a(m-j).$$

Substituting the mgf of the ErL distribution in (1.3), the $\mu_{[m]}(X)$ is finally given as

$$\mu_{[m]}(X)=\frac{\Gamma(r+m)}{\Gamma(r)}\sum_{j=0}^{m}\binom{m}{j}(-1)^j\frac{\theta^k(1+\theta)(\theta-m+j)^{2-k}+\theta^3(1+\theta-m+j)}{(1+\theta)^2(\theta-m+j)^2},$$

with $m=1,2,3,\ldots$

Definition 3.2 If $X:$ NB-ErL (r,θ,k), then the mean; $E(X)=\mu_{[m]}(X)$, The second moment about zero of X is $E(X^2)$ and variance; $Var(X)=E(X^2)-\left[E(X)\right]^2$ are given by:

$$E(X) = r\left[\frac{\frac{2\theta^2-1}{(\theta-1)^2}+(\theta+1)\theta^k(\theta-1)^{-k}}{(\theta+1)^2}\right],$$

$$E(X^2) = (r^2+r)\left[\frac{\frac{\theta^3(\theta^2-\theta-1)}{(\theta^2-3\theta+2)^2}-(\theta+1)((\theta-2)^k-(\theta-1)^k)(\theta-2)^{-k}(\theta-1)^{-k}\theta^k}{(\theta+1)^2}\right],$$

and

$$Var(X) = (r^2+r)\left[\frac{\frac{\theta^3(\theta^2-\theta-1)}{(\theta^2-3\theta+2)^2}-(\theta+1)((\theta-2)^k-(\theta-1)^k)(\theta-2)^{-k}(\theta-1)^{-k}\theta^k}{(\theta+1)^2}\right]-r^2\left[\frac{\frac{2\theta^2-1}{(\theta-1)^2}+(\theta+1)\theta^k(\theta-1)^{-k}}{(\theta+1)^2}\right]^2.$$

The index of dispersion (ID) (Cox & Lewis, 1966) for count data is a normalized measure of the dispersion of a probability function. It is defined as the ratio of the variance to the mean; $ID = \frac{Var(X)}{E(X)}$. The ID value greater than 1 suggests that the data is overdispersion than expected under a random distribution, while the ID less than 1 indicates that the data is under dispersed than expected. The ID equal to 1 suggests that the data is dispersed as expected under a random distribution (i.e., Poisson distribution).

4. Applications study of NB-ErL distribution

We illustrated the NB-ErL, negative binomial; NB, negative binomial-Lindley; NB-L and Poisson distributions by applying to real data sets. Three count datasets characterized by many zeros are considered to assess the performance of the NB-ErL distribution. The first dataset consists of the number of accidents of each policy recorded (Klugman et al., 2008). The second dataset is the number of injured persons from the accident on main roads in Bangkok in 2007 (Pudprommarat et al., 2011). The number of hospital stays of United States residents aged 66 and over is contained in the last dataset (Flynn & Francis, 2009). Use R programming to analyze the data. Fit the data to the distribution of interest using appropriate *p*-value of chi-square statistics.

With the least *p*-value of chi-square statistics, the proposed distribution (NB-ErL distribution) provides the best fit to the first dataset (**Table 1**) while its Poisson distribution is the worst. Such information is overdispersed with a mean of 0.21, variance of 0.29 and ID is 1.34. It is also observed that the NB-L performs better than its Poisson and NB distributions.

For the second dataset (**Table 2**), the mean is 0.34, the variance is 0.54, and the ID is 1.59, indicating overdispersion. The NB-ErL distribution also gives the best fit with the lowest *p*-value of chi-square statistics. Like the first data, the NB-ErL distribution also performs better than the classical NB distribution, while the NB-L performs better than the Poisson and NB distributions.

Table 3 reveals that the third dataset is best fitted by the NB-ErL distribution. This dataset has a mean of 0.30, a variance of 0.56, and an ID of 1.88. The NB-ErL distribution can be chosen as the best model. Also, based on the *p*-values of chi-square statistics, the NB-ErL is appropriate to fit the data compared to the Poisson, NB, and NB-L distributions.

Table 1 Estimated parameters, *p*-value of chi-square statistics for Dataset I

The number of accidents	The number of cases	Expected value by fitting distribution			
		Poisson	NB	NB-L	NB-ErL
0	7840	7638.3	7843.3	7853.6	7824.2
1	1317	1634.6	1290.2	247.6	1324.5
2	239	174.9	257.7	54.2	244.1
3	42	12.5	54.5	13.2	47.3
4	14	0.7	11.8	3.5	12.3
5	4	0 = 13.2	2.6 = 15.3	1.0 = 5	4.2
6	4	0	0.6	0.3	3.8 = 8.89
7	1	0	0.3	0.2	0.9
Total	**9461**				
Parameter estimates		$\hat{\lambda} = 0.214$	$\hat{r} = 0.700$	$\hat{r} = 4.63$	$\hat{r} = 4.52$
			$\hat{p} = 0.765$	$\hat{\theta} = 23.55$	$\hat{\theta} = 22.04$
					$\hat{k} = 1.00$
Chi-square		207.89	6.49	5.65	3.06
d.f.		3	4	4	5
p-value		<0.001	0.17	0.23	0.69

Table 2 Estimated parameters, *p*-value of chi-square statistics for Dataset II

The number of injured	The number of cases	Expected value by fitting distribution			
		Poisson	NB	NB-L	NB-ErL
0	1273	1187.6	1278.2	1275.3	1275.3
1	300	410.5	278.4	285.3	295.3
2	71	70.9	81.9	83.9	72.2
3	18	8.2	26.2	16.8	21.8
4	9	0.7	8.7	6.7	8.4
5	4	0.1 = 9.0	3.0 = 12.7	3.4 = 10.1	3.4 = 5.03
6	3	0.0	1.0	0	1.7
Total	**1678**				
Parameter estimates		$\hat{\lambda} = 0.35$	$\hat{r} = 0.59$	$\hat{r} = 4.08$	$\hat{r} = 16.89$
			$\hat{p} = 0.63$	$\hat{\theta} = 13.90$	$\hat{\theta} = 58.05$
					$\hat{k} = 8$
Chi-square		106.24	6.30	6.99	1.04
d.f.		3	4	4	5
p-value		<0.001	0.07	0.32	0.99

Table 3 Estimated parameters, *p*-value of chi-square statistics for Dataset III

The number of hospital stays	The number of cases	Expected value by fitting distribution			
		Poisson	NB	NB-L	NB-ErL
0	3541	3277.1	3555.6	3538.0	3537.1
1	599	970.0	577.2	612.4	611.1
2	176	143.6	176.2	163.0	161.7
3	48	14.2	61.7	52.9	54.6
4	20	1.0	22.0	22.0	21.6
5	12	0.1	8.8	8.8	9.7
6	5	0.0	4.4	4.4	8.8
7	1	0.0	0.0	4.4	0.4
8	4	0.0	0.0	0.0	0.9
		= 15.3	= 13.2	= 8.8	= 10.2
Total	**4406**				
Parameter estimates		$\hat{\lambda} = 0.296$	$\hat{r} = 0.37$	$\hat{r} = 1.38$	$\hat{r} = 1.42$
			$\hat{p} = 0.56$	$\hat{\theta} = 6.40$	$\hat{\theta} = 7.68$
					$\hat{k} = 3$
Chi-square		615.03	9.94	3.28	2.98
d.f.		3	5	6	6
p-value		<0.001	0.08	0.77	0.81

5. Conclusion

The performances of the NB-ErL distribution were evaluated using three count observations, which had different percentages of zero counts. Comparisons are made with negative binomial; NB, negative binomial-Lindley; NB-L and Poisson distributions. The maximum likelihood estimation using different algorithms that come with R-language is used to provide estimates for the parameters of the distributions. The *p*-value of chi-square statistics are used for model selection. Results show that the NB-ErL distribution outperforms its overdispersion count data. The finding shows that the new three-parameter negative binomial distribution obtained by mixing the negative binomial distribution and the two-parameter Erlang-Lindley distribution is naturally suitable to model observation with the data is overdispersion than expected under a random distribution.

References

Abd El-Monsef, M. M. E., Hassanein, W. A., & Kilany, N. M. (2017). Erlang-Lindley distribution. **Communications in Statistics - Theory and Methods**, 46(19), 9494-9506.

Aryuyuen, S., & Tonggumnead, U. (2022). Bayesian inference for the negative binomial-Quasi Lindley model for time series count data on the COVID-19 pandemic. **Trends in Sciences**, 19(21), 1-16.

Aryuyuen, S., & Bodhisuwan, W. (2013). The negative binomial - generalized exponential distribution. **Applied Mathematical Sciences**, 7 (22), 1093–1105.

Cox D.R., & Lewis, P.A.W. (1966). The statistical analysis of series of events. **Ann. Math. Statist**, 37(6), 1852-1853

Flynn, M., & Francis, L. A. (2009). **More flexible GLMs zero - inflated models and hybrid models**. In Casualty Actuarial Society E-Forum, Winter 2009, 148-224. Las Vegas. URL: https://www.casact.org/sites/default/files/database/forum_09wforum_completew09.pdf

Gomez-Deniz, E., Sarabia, J. M., & Calderi-n-Ojeda, E. (2008). Univariate and multivariate versions of the negative binomial - inverse Gaussian distributions with applications. **Insurance Mathematics and Economics**, 42 (1), 39–49.

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2008). **Loss models: from data to decisions**. (3rd ed.). New Jersey: John Wiley & Sons.

Korolev V. Y., & Zeifman, A. I. (2019). Generalized negative binomial distributions as mixed geometric laws and related limit theorems. **Lithuanian Mathematical Journal**, 59, 366–388.

Samutwachirawong, S. (2017). **A negative binomial-new weighted Lindley distribution for count data and its application to hospitalized patients with diabetes at Ratchaburi hospital, Thailand**. In the 2nd International Conference of Multidisciplinary Approaches on UN Sustainable Development Goals (UNSDGs), SCI 25–32. Bangkok. URL: http://dept.npru.ac.th/unsdgs2017/data/files/vol%20%206.1.pdf

Zamani, H., Ismail, N., & Faroughi, P. (2014). Poisson-weighted exponential univariate version and regression model with applications. **Journal of Mathematics and Statistics**, 10(2), 148–154.