

A New Generalized Poisson Mixed Distribution and Its Application

Samuel Adewale Aderoju

Department of Statistics and Mathematical Sciences
Kwara State University, Malete
P.M.B. 1530, Ilorin, Kwara State, Nigeria

This article is distributed under the Creative Commons by-nc-nd Attribution License.
Copyright © 2020 Hikari Ltd.

Abstract

A new generalized Poisson mixed distribution is proposed in this study called New Generalized Poisson-Sujatha distribution (NGPSD). The properties and application of the distribution are studied. The two parameter distribution is obtained by compounding Poisson distribution with a two parameter generalized Sujatha distribution. The distribution has a tendency to account for over-dispersion in count data. The first four moments, variance and coefficient of variation of the distribution are also obtained. The estimators of its parameters are obtained via maximum likelihood method using R-software. The goodness-of-fit of the distribution is compared with other distributions such as Poisson distribution (*PO*), negative binomial (*NB*), Generalized Poisson-Lindley (*GPL*) and a New Generalized Poisson-Lindley (*NGPL*) Distributions. It can be seen that the test statistic, AIC and BIC for the *NGPSD* are lower than those of competing distributions implying that the proposed distribution satisfactorily fits better to the data set.

Keywords: maximum likelihood estimation, goodness-of-fit, moments, compound distribution, over-dispersion, moments

1. Introduction

In modeling count data, Poisson distribution is one common model in literatures. However, the distribution has a unique properties that makes it unsuitable for most count data, especially when there are issues of over- or under-dispersion. The mean and variance of Poisson distribution are equal (equi-dispersion), this assumption is violated by most count data. Hence, different authors had proposed Poisson mixed

model or extend Poisson model to accommodate the dispersion. Shanker and Hagos [1] proposed Poisson-Sujatha distribution and they studied its properties and applications to real life datasets related to ecology and genetics. They observed that their proposed distribution satisfactorily fit majority of data-sets used. Bhati et al., [2] proposed a new generalized Poisson-Lindley distribution by compounding Poisson distribution with a two parameter generalised Lindley distribution. They studied mathematical properties of the distribution and compared its application with some other existing distributions. They found the new generalized Poisson-Lindley distribution performed better in term of goodness-of-fit than other competing models. Aderoju et al., [3] proposed zero-truncated generalized Poisson-Lindley distribution. They studied the properties and applications of the distribution. They compared the goodness-of-fit with zero-truncated Poisson and Zero-truncated Poisson-Lindley distributions and discovered that the proposed distribution yielded satisfactory results.

Tesfay and Shanker [4] proposed a distribution they called “another two-parameter Sujatha distribution” (ATPSD). The distribution includes exponential distribution and Sujatha distribution as special cases. They study the properties and applications of ATPSD with two real lifetime datasets. The probability density function (pdf) of ATPSD was given as:

$$P(Y = y) = g(y) = \frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha} (1 + \alpha y + \alpha y^2) e^{-\beta y};$$

$$y > 0, \beta > 0, \alpha \geq 0, \quad (1.1)$$

Where β is a scale parameter and α is a shape parameter. Obviously, for $\alpha = 0$ and $\alpha = 1$ the ATPSD reduces to exponential distribution and Sujatha distribution respectively.

2. Proposed model

Definition 1: A random variable Y is said to be a new generalized Poisson-Sujatha distribution (NGPSD) if it follows

$$Y|\lambda \sim Po(\lambda)$$

$$\text{while } \lambda|\beta, \alpha \sim ATPSD(\beta, \alpha)$$

for $\lambda > 0, \beta > 0$ and $\alpha \geq 0$. Hence, we denote unconditional distribution of NGPSD by $NGPS(\beta, \alpha)$.

Theorem 1: If $Y \sim NGPS(\beta, \alpha)$, then probability mass function (pmf) of Y is

$$f(y|\beta, \alpha) = \frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha} \left(\frac{\alpha(y^2 + \beta + 3) + (\beta + 4)\alpha y + (\beta^2 + 2\beta + 1)}{(\beta + 1)^{y+3}} \right);$$

$$y = 0, 1, \dots \quad (2.1)$$

$$\text{for } \beta > 0 \text{ and } \alpha \geq 0.$$

Proof: Suppose $Y|\lambda \sim Po(\lambda)$ and $\lambda|\beta, \alpha \sim ATPSD(\beta, \alpha)$, then the pmf of unconditional random variable Y is given as:

$$f(y) = \int_0^{\infty} P(Y = y|\lambda) f(\lambda|\beta, \alpha) d\lambda$$

Here $P(Y = y|\lambda)$ is distribution and $f(\lambda|\beta, \alpha)$ is ATPSD. Therefore,

$$\begin{aligned} f(y) &= \int_0^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} \frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha} (1 + \alpha\lambda + \alpha\lambda^2) e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^3}{y! (\beta^2 + \alpha\beta + 2\alpha)} \int_0^{\infty} \lambda^y e^{-\lambda(\beta+1)} (1 + \alpha\lambda + \alpha\lambda^2) d\lambda \\ &= \frac{\beta^3}{y! (\beta^2 + \alpha\beta + 2\alpha)} \left[\int_0^{\infty} \lambda^y e^{-\lambda(\beta+1)} d\lambda + \alpha \int_0^{\infty} \lambda^{y+1} e^{-\lambda(\beta+1)} d\lambda \right. \\ &\quad \left. + \alpha \int_0^{\infty} \lambda^{y+2} e^{-\lambda(\beta+1)} d\lambda \right] \\ &= \frac{\beta^3}{y! (\beta^2 + \alpha\beta + 2\alpha)} \left[\frac{y!}{(\beta+1)^{y+1}} + \frac{\alpha(y+1)!}{(\beta+1)^{y+2}} + \frac{\alpha(y+2)!}{(\beta+1)^{y+3}} \right] \\ &= \frac{\beta^3}{(\beta^2 + \alpha\beta + 2\alpha)} \left[\frac{1}{(\beta+1)^{y+1}} + \frac{\alpha(y+1)}{(\beta+1)^{y+2}} + \frac{\alpha(y+2)(y+1)}{(\beta+1)^{y+3}} \right] \\ &= \frac{\beta^3}{(\beta^2 + \alpha\beta + 2\alpha)} \left[\frac{\alpha(y^2 + \beta + 3) + (\beta+4)\alpha y + (\beta^2 + 2\beta + 1)}{(\beta+1)^{y+3}} \right] \\ \therefore f(y) &= \frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha} \left(\frac{\alpha(y^2 + \beta + 3) + (\beta+4)\alpha y + (\beta^2 + 2\beta + 1)}{(\beta+1)^{y+3}} \right); \\ &\quad y = 0, 1, \dots \end{aligned}$$

where $\beta > 0$ and $\alpha \geq 0$

3. Properties of NGPSD

The first four moments were obtained as:

$$\mu_r = \sum_{y=0}^{\infty} Y^r f(y) \quad (2.2)$$

$$\mu'_1 = \frac{\beta^2 + 2\alpha(\beta + 3)}{\beta(\beta^2 + \alpha(\beta + 2))} \quad (2.3)$$

$$\mu'_2 = \frac{\beta^2(\beta + 2) + 2\alpha(12 + 6\beta + \beta^2))}{\beta^2(\beta^2 + \alpha(\beta + 2))} \quad (2.4)$$

$$\begin{aligned} \mu'_3 &= \frac{\beta^2(6 + 6\beta + \beta^2) + 2\alpha(60 + 48\beta + 12\beta^2 + \beta^3))}{\beta^3(\beta^2 + \alpha(\beta + 2))} \end{aligned} \quad (2.5)$$

$$\mu'_1 = \frac{\beta^2(24 + 36\beta + 14\beta^2 + \beta^3) + 2\alpha(360 + 420\beta + 156\beta^2 + 24\beta^3 + \beta^4)}{\beta^4(\beta^2 + \alpha(\beta + 2))} \quad (2.6)$$

Recall that μ'_1 is the mean while variance is obtain as $Var(Y) = \sigma^2 = \mu'_2 - (\mu'_1)^2$

$$\begin{aligned} \sigma^2 &= \frac{\beta^2(\beta + 2) + 2\alpha(12 + 6\beta + \beta^2)}{\beta^2(\beta^2 + \alpha(\beta + 2))} - \left(\frac{\beta^2 + 2\alpha(\beta + 3)}{\beta(\beta^2 + \alpha(\beta + 2))} \right)^2 \\ \sigma^2 &= \frac{\beta^4(\beta + 1) + \alpha\beta^2(16 + 12\beta + 3\beta^2) + 2\alpha^2(6 + 12\beta + 6\beta^2 + \beta^3)}{\beta^2(\beta^2 + \alpha(2 + \beta))^2} \end{aligned} \quad (2.7)$$

Coefficient of Variation (CV)

$$\begin{aligned} CV &= \frac{\sqrt{\sigma^2}}{\mu'_1} \\ CV &= \frac{\sqrt{\frac{\beta^4(\beta + 1) + \alpha\beta^2(16 + 12\beta + 3\beta^2) + 2\alpha^2(6 + 12\beta + 6\beta^2 + \beta^3)}{\beta^2(\beta^2 + \alpha(2 + \beta))^2}}}{\frac{\beta^2 + 2\alpha(\beta + 3)}{\beta(\beta^2 + \alpha(\beta + 2))}} \\ CV &= \frac{\sqrt{(\beta^4(\beta + 1) + \alpha\beta^2(16 + 12\beta + 3\beta^2) + 2\alpha^2(6 + 12\beta + 6\beta^2 + \beta^3))}}{\beta(\beta^2 + \alpha(\beta + 2))(\beta^2 + 2\alpha(\beta + 3))} \end{aligned} \quad (2.8)$$

4. Estimation of the parameters of NGPSD

The likelihood function of the NGPSD is given as:

$$\begin{aligned} L(\beta, \alpha | y_i) &= \prod_{i=1}^n \frac{\beta^3}{\beta^2 + \alpha\beta + 2\alpha} \left(\frac{\alpha(y_i^2 + \beta + 3) + (\beta + 4)\alpha y_i + (\beta^2 + 2\beta + 1)}{(\beta + 1)^{y_i+3}} \right) \end{aligned} \quad (4.1)$$

The log-likelihood function of the

$$\begin{aligned} \ell &= 3n\log(\beta) - n\log(\beta^2 + \alpha\beta + 2\alpha) + \sum_{i=1}^n \log(\alpha(y_i^2 + \beta + 3) + (\beta + 4)\alpha y_i + (\beta^2 + 2\beta + 1)) \\ &\quad - \sum_{i=1}^n (y_i + 3) \log(\beta + 1) \end{aligned} \quad (4.2)$$

We are to find the first and second partial derivative of (4.2) with respect to each parameter and equate them to zero as:

$$\frac{\partial \ell}{\partial \beta} = 0, \quad \frac{\partial^2 \ell}{\partial \beta^2} = 0 \quad \text{and} \quad \frac{\partial \ell}{\partial \alpha} = 0, \quad \frac{\partial^2 \ell}{\partial \alpha^2} = 0$$

Nevertheless, (4.2) does not have closed form. Therefore, the maximum likelihood estimates (MLEs) of NGPSD cannot be solved analytically, an iterative methods such as Fisher Score Algorithm, Bisection method Regula-Falsi method or *Newton-Raphson* (NR) iterative method can be used.

Hence, we obtained the MLEs of the parameters by direct maximization of the log-likelihood function using “*optim*” routine of R software [5] with “*L-BFGS-B*” method. This can as well be done by using **PROC NLMIXED** in SAS.

5. Application

To examine the goodness-of-fit of the NGPSD, we use the epileptic seizure data used by Bhati et al., [2] and compared its performance with the Poisson (PO), Negative Binomial (NB), Generalized Poisson-Lindley (GPL) and New Generalized Poisson-Lindley (NGPL) distributions considered in the same paper.

Illustration of data: The dataset was obtained from [2] representing epileptic seizure counts (see Chakraborty [6]) are used.

Table 1: Distribution of epileptic seizure counts

x	Observation	Expected Frequency				
	Frequency	PO	NB	GPL	NGPL	NGPSD
0	126	74.9	91.1	121.5	122.0	122.3
1	80	115.7	86.6	92.0	91.0	89.6
2	59	89.3	63.4	59.0	58.8	58.8
3	42	46.0	42.6	35.1	35.2	35.8
4	24	17.8	27.6	20.1	20.5	20.6
5	8	5.5	17.6	11.2	11.3	11.4
6	5	1.4	10.5	6.1	6.4	6.1
7	4	0.3	6.6	3.3	3.3	3.2
8	3	0.1	5.0	2.7	2.5	3.2
Total	351	351	351	351	351	351
MLE		$\hat{\lambda} = 1.54$	$\hat{r} = 1.757$ $\hat{p} = 0.463$	$\hat{\alpha} = 1.292$ $\hat{\beta} = 1.139$	$\hat{\alpha} = 12.906$ $\hat{\beta} = 1.116$	$\hat{\alpha} = 1.3155$ $\hat{\beta} = 1.3716$
LogLik		-636.05	-595.22	-594.61	-594.48	-594.0483
χ^2		80.76	22.53	5.94	5.75	4.0447
df		4	5	4	4	4
G^2		70.15	25.55	5.176	4.958	4.1275
Pvalue		< 0.001	< 0.001	0.27	0.30	0.40
AIC		1276.1	1194.44	1193.22	1192.96	1192.097
BIC		1276.26	1194.599	1193.379	1193.119	1192.256

The underlying distribution for this data is Poisson. However, when the Poisson model is applied to the data, it fits very poorly, with $\hat{\lambda} = 1.54$ and $\hat{\sigma}^2 = 1.54$. The observed mean, $\lambda = 0.6825$ and $\sigma^2 = 0.8137$. Obviously, variance is greater than

expected value in the observed data. Therefore, we would need models that will account for this dispersion in the data, so we considered other models considered by Bhati et al., [2]. The proposed NGPSD was as well compared with these models and observe that the NGPSD fits the data better than other models, using X^2 , Deviance (G^2) and p-values. Alternative measures of fit provided by the AIC, Log-likelihood (LogLik) and BIC also imply that the NGPSD is a better fit than the NB, GPL and NGPL distributions.

Conflicts of Interest. There is no conflict of interest.

References

- [1] Shanker R., Hagos F., On Poisson-Sujatha Distribution and its Applications to Model Count Data from Biological Sciences, *Biometrics and Biostatistics International Journal*, **2** (4) (2016), 1-5.
<https://doi.org/10.15406/bbij.2016.03.00069>
- [2] Bhati, D., Sastry, D.V.S. and Qadri, P.Z. Maha, A New Generalized Poisson-Lindley Distribution: Applications and Properties, *Austrian Journal of Statistics*, **44** (2015), 35–51. <https://doi.org/10.17713/ajs.v44i4.54>
- [3] Aderoju, S.A., Jolayemi, E.T. and Ibrahim A.O., A Comparative Study on Zero-truncated Generalized Poisson-Lindley and Zero-truncated Poisson-Lindley Distributions, *International Journal of Mathematical Archive*, **8** (3) (2017), 92-98.
- [4] Tesfay, M. and Shanker R., Another Two-Parameter Sujatha Distribution with Properties and Applications, *Journal of Mathematical Sciences and Modelling*, **2** (1) (2019), 1-13. <https://doi.org/10.33187/jmsm.416628>
- [5] R Core Team, R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria, 2018*. <https://www.R-project.org/>
- [6] Chakraborty S., On Some Distributional Properties of the Family of Weighted Generalized Poisson distribution, *Communications in Statistics - Theory and Methods*, **39** (15) (2010), 2767–2788. <https://doi.org/10.1080/03610920903129141>

Received: March 23, 2020; Published: April 7, 2020