On the Applications of Some Poisson Related Distributions

M. A. Umar[1]; W. B. Yahya[2]

Department of Statistics,
University of Ilorin, Ilorin, Nigeria.
E-mail: mumarad90@gmail.com[1]; dr.yah2009@gmail.com[2]

Abstract — **Poisson distribution plays an important role in count data analysis, but it cannot model some data with over-dispersion or under-dispersion because of its classical equi-dispersion property. Nevertheless, in an effort to handle such a situation, a number of works have been proposed. These have resulted in the development of Poisson mixture distributions such as the Negative Binomial, Poisson – Exponential-Gamma, Poisson – Exponential and Poisson-Lindley distributions among others. In this Paper, these distributions were applied to real-life datasets from different fields of study. Their Goodness-of-fit has been discussed based on the values of; −2logLik, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). These were achieved by estimating the parameters of the distributions using the real-life datasets considered. The distributions fit the datasets satisfactorily but Poisson – Exponential-Gamma distribution is the best model. On the basis of which it is concluded that the distribution can serve as an important alternative to real-life count data modeling.**

Keywords- Poisson, Poisson – Exponential-Gamma, Poisson-Exponential, Poisson-Lindley, Negative Binomial, Goodness-of-fit.

I. INTRODUCTION

Poisson distribution (PD) is a discrete probability function that expresses the probability of a given number of events occurring in a fixed interval of time or space, especially if these events occur with a known constant rate and independent of time (Anderson et al, 2012; Donelly, 2012; Jaggia & Kelly, 2012). It can also arise as an approximation to the binomial distribution when the proportion, p, is small and the size, n, is large (Triola, 2007).

It is often known as the distribution of rare events, thereby dealing with a process where discrete events occur in a continuous but finite interval of time or space with the conditions; for a small interval, the probability of event occurring is proportional to the size of the interval, the probability of more than one occurrence in small interval is negligible (that is, they are rare events and must not occur simultaneously), each occurrence must be independent of others and must be at random (Anderson et al., 2012; Donelly, 2012; Jaggia & Kelly, 2012). Thus, the distribution (PD) plays an important role in count data analysis. However, it cannot model some data with over-dispersion or under-dispersion because of its equi-dispersion property.

Similarly, for fitting Negative Binomial distribution (NBD) to count datasets, the datasets have to be over-dispersed, that is the mean is less than the variance (Shanker & Hagos, 2015). In biological and medical sciences, these conditions are not fully satisfied. Nevertheless, a number of works have proposed methods for modeling count data that violate this classical property. These have resulted to the development of Poisson mixture distributions apart from the Negative Binomial distribution (Cook, 2009; Kongrod et al, 2014), the Poisson – Exponential-Gamma (Umar, 2019), Poisson – Exponential (Umar, 2019) and Poisson-Lindley distributions (Sankaran, 1970) among others.

Analysis and modeling of lifetime data are crucial in applied sciences and other fields of knowledge. Thus, a number of models were in-exhaustively constructed to facilitate better modeling and significant progress (Asad et al, 2018). This has attracted the attention and interest of researchers all over the world. These models have been shown to perform better than one another in the various fields tested. However, this work is carried out to apply some of these models to count data especially from community health and other fields of knowledge.

II. RESEARCH METHODOLOGY

This The Poisson distribution is defined by its probability mass function (R Core Team, 2010) as follows;

$$P(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}; x = 0, 1, 2, \dots, \lambda > 0 \quad (1)$$

A Poisson mixture was used and obtained the Negative Binomial distribution (Greenwood & Yule, 1920), Poisson – Exponential-Gamma (Umar, 2019), Poisson – Exponential (Umar, 2019) and Poisson-Lindley distributions (Sankaran, 1970) among others. The Negative Binomial distribution was originally derived as limiting case of the Gamma-Poisson distribution, where the mixing distribution of the Poisson rate is Gamma distribution, i.e. the λ itself is a random variable distributed as a Gamma distribution with shape parameter $\alpha = r$ and a scale parameter $\theta = p(1-p)^{-1}$ (Greenwood & Yule, 1920).
The Poisson mixture distribution can be obtained by putting (1) and the equation of the distribution to be mixed in the expression below (Greenwood & Yule, 1920);

$$P(x;\alpha,\theta) = \int_0^\infty P_{Poi(\lambda)}(x)\cdot f(\lambda)d\lambda \qquad (2)$$

Through this expression in (2), the Negative Binomial distribution (Cook, 2009), Poisson – Exponential-Gamma (Umar, 2019), Poisson – Exponential (Umar, 2019) and Poisson-Lindley distributions (Sankaran, 1970) among others were obtained. The mathematical expressions of these distributions together with their graphical representations, important properties and parameter(s) estimates with their goodness-of-fit (in comparison with other related distributions) were discussed accordingly.
The p.m.f. of the Poisson – Exponential-Gamma distribution is defined (Umar, 2019) as follows:

$$P(x;\alpha,\theta) = \frac{\theta}{(\theta+\Gamma(\alpha))x!}\left[\frac{\theta(\theta+1)^{\alpha}x!+\theta^{\alpha-1}(\theta+1)\Gamma(\alpha+x)}{(\theta+1)^{x+\alpha+1}}\right];\ x = 0,1,2,\dots,\theta>0,\alpha>0 \qquad (3)$$

It can be easily verified that when $\alpha = 1$, the Poisson-Exponential-Gamma distribution in (3) reduces to a Poisson-Exponential distribution (Umar, 2019), and a Poisson-Lindley distribution when $\alpha = 2$ (Sankaran, 1970). That is;

$$P(x;1,\theta) = \frac{\theta}{(\theta+1)^{x+1}};\ x = 0,1,2,3,\dots,\theta>0 \qquad (4)$$

which is the Poisson-Exponential distribution, and

$$P(x;2,\theta) = \frac{\theta^2(x+\theta+2)}{(\theta+1)^{x+3}};\ x = 0,1,2,3,\dots,\theta>0 \qquad (5)$$

which is the Poisson-Lindley distribution (Sankaran, 1970).

III. Applications

In this section, the goodness-of-fit of the distributions is discussed with an application to real-life datasets. The parameters of the distributions were solved using the MLE method while the goodness-of-fit was evaluated using the Akaike Information Criterion (AIC, Akaike, 1974), Bayesian Information Criterion (BIC, Schwarz, 1978) and –2logLik with their respective statistics given below.

$$AIC = -2\ln L + 2k \qquad (6)$$
$$BIC = -2\ln L + k\ln n \qquad (7)$$

where k is the number of parameters and n is the sample size. The distribution that has a lower value of these criteria is judged to be the best among others.

Data Description

Dataset 1: This consists of the number of yeast cell counts per square reported by Shanker & Hagos (2015).
Dataset 2: This is the number of times a member's name appeared in an Edited Conference Proceedings (PSSN Journal of 2018) according to membership registration.
Dataset 3: This is the number of European red mites on Apple leaves (Shanker & Hagos, 2015).

Tables 1 – 3 present the observed and expected frequencies of the datasets. The expected frequencies according to the Poisson (PD), Poisson-Exponential (PED), Poisson-Lindley (PLD) and Poisson – Exponential-Gamma (PEGD) distributions were given and compared. It can be seen that the distributions gave satisfactory fits to the datasets. This can also be confirmed by the values of AIC, AICC, BIC and the graphs in Figures 1 – 3.

Table 1: Observed and Expected frequencies of yeast cell counts per square

The number of Cells per square	Observed Frequency	Expected Frequency			
		PD	**PED**	**PLD**	**PEGD**
0	128	118.1	128.1	127.4	128.1
1	37	54.3	40.4	41.1	40.9
2	18	12.3	12.7	12.9	12.6
3	3	1.9	4.0	3.9	3.8
4	1	0.2	1.3	1.2	1.1
5	0	0.0	0.4	0.4	0.3
TOTAL	187	187	187	187	187
ML Estimates		$\hat{\theta}$=0.46	$\hat{\theta}$=2.17	$\hat{\theta}$=2.75	$\hat{\alpha}$=1.51 $\hat{\theta}$=2.50
-2logLik		195.30	176.16	660.68	149.64
AIC		197.30	178.16	662.68	153.64
AICC		197.37	178.23	662.75	153.71
BIC		200.53	181.39	665.91	154.88

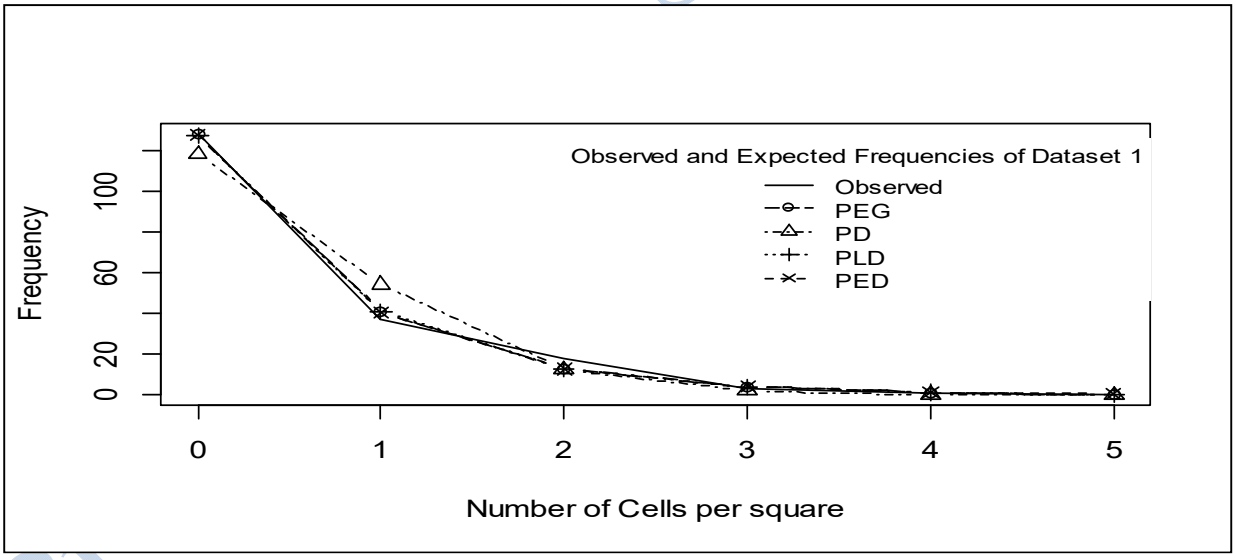

Figure 1: Graph of the observed and expected frequency of yeast cell counts per square

Table 2: Observed and Expected frequencies of Authors' name in PSSN Journal, Volume 2

The number of times a name appeared	Observed Frequency	Expected Frequency			
		PD	**PED**	**PLD**	**PEGD**
0	311	300.3	305.3	236.0	310.2
1	40	57.8	49.2	84.3	36.5
2	10	5.6	7.9	29.1	11.3
3	2	0.4	1.3	9.8	3.9
4	1	0.0	0.2	3.2	1.4
5	0	0.0	0.0	1.1	0.5
TOTAL	**364**	**364.1**	**363.9**	**363.5**	**363.8**
ML Estimates		$\hat{\theta} = 0.19$	$\hat{\theta} = 5.20$	$\hat{\theta} = 2.43$	$\hat{\alpha} = 0.16$ $\hat{\theta} = 1.62$
-2logLik		189.46	182.78	1346.20	110.28
AIC		191.46	184.78	1350.20	114.28
BIC		195.36	184.68	1357.99	122.07

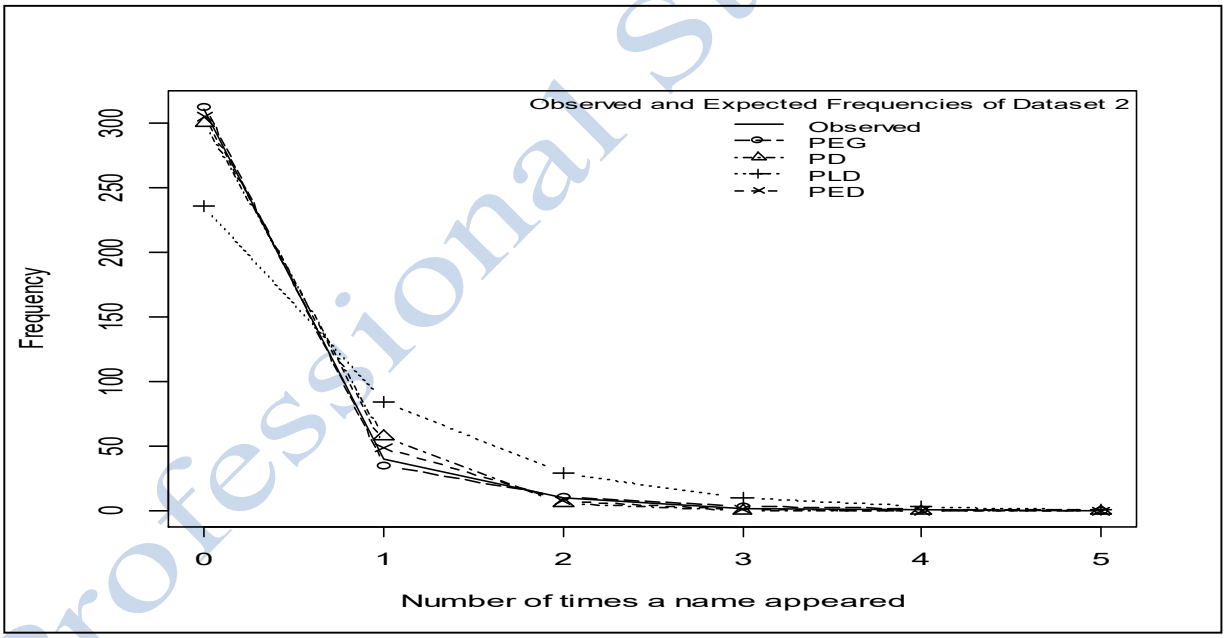

Figure 2: Graph of the observed and expected frequency of the number of times a name appeared

Table 3: Observed and Expected frequencies of European Red mites on Apple leaves

The number mites per leaf	Observed Frequency	Expected Frequency			
		PD	**PED**	**PLD**	**PEGD**
0	38	25.3	37.2	35.8	39.9
1	17	29.1	19.9	20.7	20.8
2	10	16.8	10.6	11.4	10.2
3	9	6.4	5.7	6.0	4.9
4	3	1.8	3.0	3.1	2.3
5	2	0.4	1.6	1.6	1.1
6	1	0.1	0.9	0.8	0.5
7	0	0.0	0.5	0.6	0.2
TOTAL	80	79.9	79.4	80	79.9
ML Estimates		$\hat{\theta} = 1.15$	$\hat{\theta} = 0.87$	$\hat{\theta} = 1.26$	$\hat{\alpha} = 1.62$ $\hat{\theta} = 1.31$
-2logLik		176.17	142.78	136.85	126.53
AIC		178.17	144.78	138.85	130.53
BIC		180.56	147.36	141.24	135.29

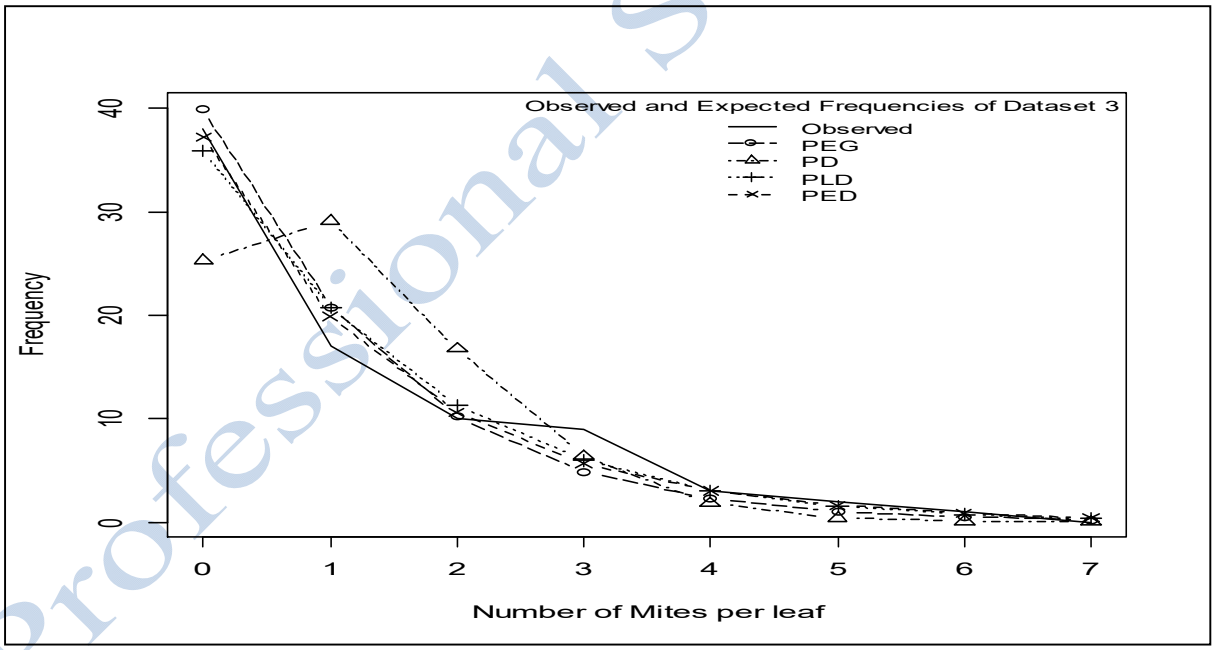

Figure 3: Graph of the observed and expected frequency of the number of mites per leaf

IV Concluding Remarks

The distributions considered in this paper were estimated and compared using real-life datasets. It is obvious that the expected frequencies given by the distribution were satisfactory. But the values given by the PEGD were closer to the observed frequencies of all the datasets considered in the paper than the competing distributions.

It has the minimum values of AIC and BIC and the fits were shown graphically. These distributions can, therefore, be considered important alternatives to modeling real-life count datasets.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, AC-* ***19***, 716-723.

Anderson, D. R., Sweeney, D. J. and Williams, T. A. (2012). *Essentials of Modern Business Statistics with Microsoft® Excel*. Mason, OH: South-Western, Cengage Learning.

Asad, A., Qaisar, R., Muhammad, Z. and Muhammad, T. J. (2018). A Quasi Lindley Pareto distribution. *Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Science.* ***55***(2): 32 – 40.

Cook, J. D. (2009, October, 28). Notes on Negative Binomial Distribution. Retrieved February, 4th 2019, from John D. Cook. http://www.johndcook.com/negative_binomial.pdf

Donnelly, Jr. R. A. (2012). *Business Statistics*. Upper Saddle River, NJ: Pearson Education, Inc.

Greenwood, M. and 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. http://doi:10.2307/2341080

Jaggia, S. and Kelly, A. (2012). *Business Statistics - Communicating with Numbers*. New York, NY: McGraw-Hill Irvin.

Kongrod, S., Bodhisuwan, W. & Payakkapong, P. (2014). The Negative Binomial – Erlang Distribution with Applications. *International Journal of Pure and Applied Mathematics,* ***92***(3); 389 – 401

R – forge distributions, Core Team University Year 2009 – 2010. Handbook on Probability Distributions. Retrieved from http://r-forge.r-project.org, available in http://tinyurl.com/yc7tem9z

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Sankaran, M. (1970). The discrete Poisson-Lindley distribution. *Biometrics,* ***26***: 145-149.

Schwarz, G. (1978). Noop. *The Annals of Statistics* ***6***, 461–464

Shanker, R. and Hagos, F. (2015). On Poisson-Lindley Distribution and its Applications to Biological Sciences. *International Journal of Biometrics and Biolstatistics,* ***2***(4): 00036. http://DOI:10.15406/bbij.2015.02.0036

Triola, M. F. (2007). *Elementary Statistics Using Excel®*. Boston, MA: Addison Wesley, Pearson Education, Inc.

Umar, M. A. (2019). A Zero-truncated Poisson – Exponential-Gamma Distribution and its Applications. *An M.Sc. dissertation Submitted to the Department of Statistics, University of Ilorin, Ilorin, Nigeria*. Unpublished.