MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

On Poisson-Sujatha Distribution and its Applications to Model Count Data from Biological Sciences

Abstract

In this paper a simple method for finding moments of Poisson-Sujatha distribution (PSD) introduced by Shanker [1] has been suggested and hence the first four moments about origin and the variance has been given. The PSD has been fitted to the same data-sets relating to ecology and genetics to which earlier Shanker & Hagos [2] has fitted Poisson-Lindley distribution (PLD) introduced by Sankaran [3] and Poisson-distribution (PD) and the goodness of fit of PSD shows satisfactory fit in majority of data-sets.

Keywords: Sujatha distribution; Poisson-Sujatha distribution; Lindley distribution; Poisson-Lindley distribution; Moments; Compounding; Estimation of parameter; Goodness of fit

Research Article

Volume 3 Issue 4 - 2016

Rama Shanker* and Hagos Fesshaye
Department of Statistics, Eritrea Institute of Technology, Eritrea

***Corresponding author:** Rama Shanker, Department of Statistics, Eritrea Institute of Technology, Asmara, Eritrea, Email: shankerrama2009@gmail.com

Received: January 29, 2016 | **Published:** March 10, 2016

Introduction

The Poisson-Sujatha distribution (PSD) having probability mass function

$$P(X=x)=\frac{\theta^3}{\theta^2+\theta+2}\cdot\frac{x^2+(\theta+4)x+(\theta^2+3\theta+4)}{(\theta+1)^{x+3}};x=0,1,2,\ldots,\theta>0 \qquad (1.1)$$

has been introduced by Shanker [1] for modeling count data-sets. The PSD arises from Poisson distribution when its parameter λ follows Sujatha distribution introduced by Shanker [4] having probability density function

$$f(\lambda;\theta)=\frac{\theta^3}{\theta^2+\theta+2}(1+\lambda+\lambda^2)e^{-\theta\lambda};\lambda>0,\ \theta>0 \qquad (1.2)$$

We have

$$P(X=x)=\int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\cdot\frac{\theta^3}{\theta^2+\theta+2}(1+\lambda+\lambda^2)e^{-\theta\lambda}d\lambda \qquad (1.3)$$

$$=\frac{\theta^3}{(\theta^2+\theta+2)x!}\int_0^\infty \lambda^x(1+\lambda+\lambda^2)e^{-(\theta+1)\lambda}d\lambda$$

$$=\frac{\theta^3}{\theta^2+\theta+2}\cdot\frac{x^2+(\theta+4)x+(\theta^2+3\theta+4)}{(\theta+1)^{x+3}};x=0,1,2,\ldots,\theta>0 \qquad (1.4)$$

Which is the Poisson-Sujatha distribution (PSD).

Shanker [4] has shown that the Sujatha distribution (1.2) is a three component mixture of an exponential (θ) distribution, a gamma ($2,\theta$) distribution, and a gamma ($3,\theta$) distributionwith their mixing proportions $\frac{\theta^2}{\theta^2+\theta+2}$, $\frac{\theta}{\theta^2+\theta+2}$ and $\frac{2}{\theta^2+\theta+2}$ respectively. Shanker [4] has discussed its various mathematical and statistical properties including its shape, moment generating function, moments, skewness, kurtosis, hazard rate function, mean residual life function, stochastic orderings, mean deviations, distribution of order statistics, Bonferroni and Lorenz curves, Renyi entropy measure, stress-strength reliability , amongst others along with the estimation of the parameter and applications for modeling lifetime data.

Shanker [1] has detailed study about various mathematical and statistical properties of PSD including moment generating function, coefficient of variation, skewness, kurtosis, over-dispersion, hazard rate and unimodality along with the estimation of the parameter and applications. Shanker & Hagos [5,6] have obtained size-biased Poisson-Sujatha distribution (SBPSD) and zero-truncated Poisson-Sujatha distribution(ZTPSD) and discussed their statistical properties, estimation of the parameter and applications. Further, Shanker & Hagos [7] have detailed study about zero-truncation of Poisson, Poisson-Lindley and Poisson-Sujatha distributions and their applications.

The probability mass function of Poisson-Lindley distribution (PLD) given by

$$P(X=x)=\frac{\theta^2(x+\theta+2)}{(\theta+1)^{x+3}};\ x=0,1,2,\ldots,\ \theta>0. \qquad (1.5)$$

has been introduced by Sankaran [3] to model count data. The distribution arises from the Poisson distribution when its parameter λ follows Lindley [8] distribution with its probability density function

$$f(\lambda,\theta)=\frac{\theta^2}{\theta+1}(1+\lambda)\,e^{-\theta\lambda}\;;\quad x>0,\theta>0 \qquad (1.6)$$

In this paper a simple method for finding moments of Poisson-Sujatha distribution (PSD) introduced by Shanker [1] has been suggested and hence the first four moments about origin and the variance has been presented. It seems that not much work has been done on the applications of PSD so far. The PSD has been fitted to the same data-sets relating to ecology and genetics to which Shanker & Hagos [2] has fitted Poisson-Lindley distribution (PLD) introduced by Sankaran [3] and Poisson-distribution (PD) and the goodness of fit of PSD shows satisfactory fit in majority of data-sets.

Moments of Poisson-Sujatha Distribution

Using (1.3) the r th moment about origin of PSD (1.1) can be obtained as

$$\mu_r'=E\left[E\left(X^r|\lambda\right)\right]=\frac{\theta^3}{\theta^2+\theta+2}\int_0^\infty\left[\sum_{x=0}^\infty x^r\frac{e^{-\lambda}\lambda^x}{x!}\right]\left(1+\lambda+\lambda^2\right)e^{-\theta\lambda}d\lambda \qquad (2.1)$$

Clearly the expression under the bracket in (2.1) is the r th moment about origin of the Poisson distribution. Taking $r=1$ in (2.1) and using the first moment about origin of the Poisson distribution, the first moment about origin of the PSD (1.1) can be obtained as

$$\mu_1'=\frac{\theta^3}{\theta^2+\theta+2}\int_0^\infty\lambda\left(1+\lambda+\lambda^2\right)e^{-\theta\lambda}d\lambda=\frac{\theta^2+2\theta+6}{\theta\left(\theta^2+\theta+2\right)} \qquad (2.2)$$

Again taking $r=2$ in (2.1) and using the second moment about origin of the Poisson distribution, the second moment about origin of the PSD (1.1) is obtained as

$$\mu_2'=\frac{\theta^3}{\theta^2+\theta+2}\int_0^\infty\left(\lambda^2+\lambda\right)\left(1+\lambda+\lambda^2\right)e^{-\theta\lambda}d\lambda=\frac{\theta^3+4\theta^2+12\theta+24}{\theta^2\left(\theta^2+\theta+2\right)} \qquad (2.3)$$

Similarly, taking $r=3$ and 4 in (2.1) and using the third and the fourth moment about origin of the Poisson distribution, the third and the fourth moment about origin of the PSD (1.1) are obtained as

$$\mu_3'=\frac{\theta^4+8\theta^3+30\theta^2+96\theta+120}{\theta^3\left(\theta^2+\theta+2\right)} \qquad (2.4)$$

$$\mu_4'=\frac{\theta^5+16\theta^4+84\theta^3+336\theta^2+840\theta+720}{\theta^4\left(\theta^2+\theta+2\right)} \qquad (2.5)$$

Thus the variance of the PSD (1.1) can be obtained as

$$\mu_2=\frac{\theta^5+4\theta^4+14\theta^3+28\theta^2+24\theta+12}{\theta^2\left(\theta^2+\theta+2\right)^2} \qquad (2.6)$$

Shanker [1] has shown that the PSD is always over-dispersed, has increasing hazard rate and unimodal. Further, Shanker [1] has also shown that the graphs of coefficient of variation, skewness, and kurtosis of PSD are increasing for increasing values of the parameter.

Estimation of the Parameter

Maximum likelihood estimate (MLE) of the parameter

Let $(x_1,x_2,\ldots,x_n)$ be a random sample of size n from the PSD (1.1) and let f_x be the observed frequency in the sample corresponding to $X=x$ $(x=1,2,3,\ldots,k)$ such that $\sum_{x=1}^{k}f_x=n$, where k is the largest observed value having non-zero frequency. The likelihood function L of the PSD (1.1) is given by

$$L=\left(\frac{\theta^3}{\theta^2+\theta+2}\right)^n\frac{1}{(\theta+1)^{\sum_{x=1}^{k}f_x(x+3)}}\prod_{x=1}^{k}\left[x^2+(\theta+4)x+\left(\theta^2+3\theta+4\right)\right]^{f_x}$$

The log likelihood function is thus obtained as

$$\log L=n\log\left(\frac{\theta^3}{\theta^2+\theta+2}\right)-\sum_{x=1}^{k}f_x(x+3)\log(\theta+1)+\sum_{x=1}^{k}f_x\log\left[x^2+(\theta+4)x+\left(\theta^2+3\theta+4\right)\right]$$

The first derivative of the log likelihood function is given by

$$\frac{d\log L}{d\theta}=\frac{n\left(\theta^2+2\theta+6\right)}{\theta\left(\theta^2+\theta+2\right)}-\frac{n(\bar{x}+3)}{\theta+1}+\sum_{x=1}^{k}\frac{\left[x+(2\theta+3)\right]f_x}{\left[x^2+(\theta+4)x+\left(\theta^2+3\theta+4\right)\right]}$$

where $\bar{x}$ is the sample mean.

The maximum likelihood estimate (MLE), $\hat{\theta}$ of θ of PSD (1.1) is the solution of the equation $\frac{d\log L}{d\theta}=0$ and is given by the solution of the following non-linear equation

$$\frac{n\left(\theta^2+2\theta+6\right)}{\theta\left(\theta^2+\theta+2\right)}-\frac{n(\bar{x}+3)}{\theta+1}+\sum_{x=1}^{k}\frac{\left[x+(2\theta+3)\right]f_x}{\left[x^2+(\theta+4)x+\left(\theta^2+3\theta+4\right)\right]}=0$$

This non-linear equation can be solved by any numerical iteration methods such as Newton- Raphson method, Bisection method, Regula–Falsi method etc.

Method of moment estimate (MOME) of the parameter

Let $(x_1,x_2,\ldots,x_n)$ be a random sample of size n from the PSD (1.1). Equating the population mean to the corresponding sample mean, the MOME $\tilde{\theta}$ of θ of PSD (1.1) is the solution of the following cubic equation

$$\bar{x}\theta^3+(\bar{x}-1)\theta^2+2(\bar{x}-1)\theta-6=0$$

Where $\bar{x}$ is the sample mean.

Applications of Poisson-Sujatha Distribution

The Poisson distribution is a suitable statistical model for the situations where events seem to occur at random including the number of customers arriving at a service point, the number of telephone calls arriving at an exchange, the number of fatal traffic accidents per week in a given state, the number of radioactive particle emissions per unit of time, the number of meteorites that collide with a test satellite during a single orbit, the number of

organisms per unit volume of some fluid, the number of defects per unit of some materials, the number of flaws per unit length of some wire, are some amongst others. Since the condition for the applications for Poisson distribution is the independence of events and the equality of mean and variance, this condition is rarely satisfied completely in biological and medical science due to the fact that the occurrences of successive events are dependent. Further, the negative binomial distribution is a possible alternative to the Poisson distribution when successive events are possibly dependent (see Johnson et al. [9]), but for fitting negative binomial distribution (NBD) to the count data, mean should be less than the variance. In biological and medical sciences, these conditions are also not fully satisfied. Generally, the count data in biological science and medical science are either over-dispersed or under-dispersed. The main reason for selecting PLD and PSD to fit biological science data is that these two distributions are always over-dispersed and PSD has some flexibility over PLD.

Applications in ecology

Ecology is the branch of biology dealing with the relations and interactions between organisms and their environment, including other organisms. The organisms and their environment in the nature are complex, dynamic, interdependent, mutually reactive and interrelated. Ecology deals with the various principles which govern such relationship between organisms and their environment. It was Fisher et al. [10] who have firstly discussed the applications of Logarithmic series distribution (LSD) to model count data in the science of ecology. Later, Kempton [11] who fitted the generalized form of Fisher's Logarithmic series distribution (LSD) to model insect data and concluded that it gives a superior fit as compared to ordinary Logarithmic series distribution (LSD). He also concluded that it gives better explanation for the data having exceptionally long tail. Tripathi & Gupta [12] proposed another generalization of the Logarithmic series distribution (LSD) which is flexible to describe short-tailed as well as long-tailed data and fitted it to insect data and found that it gives better fit as compared to ordinary Logarithmic series distribution. Mishra & Shanker [13] have discussed applications of generalized logarithmic series distributions (GLSD) to models data in ecology. Shanker & Hagos [2] have tried to fit PLD for data relating to ecology and observed that PLD gives satisfactory fit.

In this section we have tried to fit Poisson distribution (PD), Poisson-Lindley distribution (PLD) and Poisson-Sujatha distribution (PSD) to many count data from biological sciences using maximum likelihood estimates. The data were on haemocytometer yeast cell counts per square, on European red mites on apple leaves and European corn borers per plant (Table 1-3).

It is obvious from above Tables that both PSD and PLD give much closer fit than Poisson distribution. Further, in some data-sets PSD gives much closer fit than PLD while in some data-sets PLD gives much closer fit than PSD and thus both PSD and PLD can be considered as important tools for modeling data in ecology.

Table 1: Observed and expected number of Haemocytometer yeast cell counts per square observed by Gosset [14].

Number Of Yeast Cells per Square	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	213	202.1	234.0	233.2
1	128	138.0	99.4	99.6
2	37	47.1	40.5	41.0
3	18	10.7	16.0	16.3
4	3	1.8	6.2	6.7
5	1	0.2	2.4	2.3
6	0	0.1	1.5	0.9
Total		400.0	400.0	400.0
Estimate of Parameter		$\hat{\theta} = 0.6825$	$\hat{\theta} = 1.950236$	$\hat{\theta} = 2.373052$
χ^2		10.08	11.04	10.86
d.f.		2	2	2
p-value		0.0065	0.0040	0.0044

Table 2: Observed and expected number of red mites on Apple leaves.

Number of red mites per Leaf	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	38	25.3	35.8	35.3
1	17	29.1	20.7	20.9
2	10	16.7	11.4	11.6
3	9	6.4	6	6.1
4	3	1.8	3.1	3.1
5	2	0.4	1.6	1.5
6	1	0.2	0.8	0.7
7+	0	0.1	0.6	0.8
Total	80	80	80	80
Estimate of Parameter		$\hat{\theta}$ =1.15	$\hat{\theta}$ =1.255891	$\hat{\theta}$ =1.64683
χ^2		18.27	2.47	2.52
d.f.		2	3	3
p-value		0.0001	0.4807	0.4719

Table 3: Observed and expected number of European corn- borer of Mc Guire et al. [15].

Number of Corn-borer per plant	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	188	169.4	194.0	193.6
1	83	109.8	79.5	79.6
2	36	35.6	31.3	31.6
3	14	7.8	12.0	12.1
4	2	1.2	4.5	4.5
5	1	0.2	2.7	2.6
Total	324	324.0	324.0	324.0
Estimate of parameter		$\hat{\theta}$ = 0.648148	$\hat{\theta}$ = 2.043252	$\hat{\theta}$ = 2.471701
χ^2		15.19	1.29	1.16
d.f.		2	2	2
p-value		0.0005	0.5247	0.5599

It is obvious from above tables that in table 1, PD gives better fit than PLD and PSD; in table 2 PLD gives better fit than PD and PSD while in table 3, PSD gives better fit than PD and PLD.

Application in genetics

Genetics is the branch of biological science which deals with heredity and variation. Heredity includes those traits or characteristics which are transmitted from generation to generation, and is therefore fixed for a particular individual. Variation, on the other hand, is mainly of two types, namely hereditary and environmental. Hereditary variation refers to differences in inherited traits whereas environmental variations are those which are mainly due to environment. The segregation of chromosomes has been studied using statistical tool, mainly chi-square (χ^2). In the analysis of data observed on chemically induced chromosome aberrations in cultures of human leukocytes, Loeschke & Kohler [16] suggested the negative binomial distribution while Janardan & Schaeffer [17] suggested modified Poisson distribution. Mishra and Shanker [13] have discussed applications of generalized Logarithmic series distributions (GLSD) to model data in mortality, ecology and genetics. Shanker & Hagos [2] have detailed study on the applications of PLD to model data from genetics. Much quantitative works seem to be done in genetics but so far no works has been done on fitting of PSD to data relating to genetics. In this section an attempt has been made to fit to data relating to genetics using PSD, PLD and PD using maximum likelihood estimate. Also an attempt has been made to fit PSD, PLD, and PD to the data of Catcheside et al. [18,19] in Table 4-7.

It is obvious from the fitting of PSD, PLD, and PD that both PSD and PLD gives much satisfactory fit than PD while in some data-sets PSD gives much closer fit than PLD whereas PLD gives much closer fit than PSD in some data-sets. Thus both PSD and PLD can be considered as important tools for modeling data in genetics

Table 4: Distribution of number of Chromatid aberrations (0.2 g chinon 1, 24 hours).

Number of Aberrations	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	268	231.3	257	257.6
1	87	126.7	93.4	93
2	26	34.7	32.8	32.7
3	9	6.3	11.2	11.2
4	4	0.8	3.8	3.7
5	2	0.1	1.2	1.2
6	1	0.1	0.4	0.4
7+	3	0.1	0.2	0.2
Total	400	400	400	400
Estimate of Parameter		$\hat{\theta}$ =0.5475	$\hat{\theta}$ =2.380442	$\hat{\theta}$ =2.829241
χ^2		38.21	6.21	6.28
d.f.		2	3	3
p-value		0	0.1018	0.0987

Table 5: Mammalian cytogenetic dosimetry lesions in rabbit lymphoblast induced by streptonigrin (NSC-45383), Exposure -60 $\mu g \mid kg$.

Class/Exposure $\mu g \mid kg$	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	413	374	405.7	406.1
1	124	177.4	133.6	132.9
2	42	42.1	42.6	42.7
3	15	6.6	13.3	13.4
4	5	0.8	4.1	4.1
5	0	0.1	1.2	1.2
6	2	0.0	0.5	0.6
Total	601	601	601	601
Estimate of parameter		$\hat{\theta}$ =0.47421	$\hat{\theta}$ =2.685373	$\hat{\theta}$ =3.125788
χ^2		48.17	1.34	1.1
d.f.		2	3	3
p-value		0	0.7196	0.7771

Table 6: Mammalian cytogenetic dosimetry lesions in rabbit lymphoblast induced by streptonigrin (NSC-45383), Exposure -70 $\mu g \mid kg$.

Class/Exposure $\mu g \mid kg$	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	200	172.5	191.8	192
1	57	95.4	70.3	70.1
2	30	26.4	24.9	24.9
3	7	4.9	8.6	8.6
4	4	0.7	2.9	2.9
5	0	0.1	1.0	0.9
6	2	0.0	0.5	0.6
Total	300	300	300	300
Estimate of parameter		$\hat{\theta}$ =0.55333	$\hat{\theta}$ =2.353339	$\hat{\theta}$ =2.795745
χ^2		29.68	3.91	3.81
d.f.		2	2	2
p-value		0	0.1415	0.1488

Table 7: Mammalian cytogenetic dosimetry lesions in rabbit lymphoblast induced by streptonigrin (NSC-45383), Exposure -90 $\mu g \mid kg$.

Class/Exposure $\mu g \mid kg$	Observed Frequency	Expected Frequency		
		PD	PLD	PSD
0	155	127.8	158.3	157.5
1	83	109	77.2	77.5
2	33	46.5	35.9	36.4
3	14	13.2	16.1	16.4
4	11	2.8	7.1	7.1
5	3	0.5	3.1	3.0
6	1	0.2	2.3	2.1
Total	300	300	300	300
Estimate of parameter		$\hat{\theta}$ =0.853333	$\hat{\theta}$ =1.617611	$\hat{\theta}$ =2.034077
χ^2		24.97	1.51	1.74
d.f.		2	3	3
p-value		0	0.6799	0.6281

Acknowledgement

None.

Conflict of Interest

None.

References

1. Shanker R (2016) The discrete Poisson-Sujatha distribution. International Journal of Probability and Statistics 5(1): 1- 9.
2. Shanker R, Hagos F (2015) On Poisson-Lindley distribution and Its applications to Biological Sciences. Biometrics and Biostatistics International Journal 2(4): 1-5.
3. Sankaran M (1970) The discrete Poisson-Lindley distribution. Biometrics 26(1): 145-149.
4. Shanker R (2015) Sujatha distribution and Its Applications, Accepted for publication in "Statistics in Transition new Series".
5. Shanker R, Hagos F (2016 a) Size-biased Poisson-Sujatha distribution with Applications, Communicated.
6. Shanker R, Hagos F (2016 b) Zero-truncated Poisson-Sujatha distribution with Applications, Communicated.
7. Shanker R, Hagos F (2016 c) On zero-truncation of Poisson, Poisson-Lindley, and Poisson-Sujatha distribution and their Applications, Communicated.
8. Lindley DV (1958) Fiducial distributions and Bayes theorem. Journal of the Royal Statistical Society 20(1): 102-107.
9. Johnson NL, Kotz S, Kemp AW (1992) Univariate Discrete Distributions. (2nd edn), John Wiley & sons Inc, USA.
10. Fisher RA, Corpet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12(1): 42-58.
11. Kempton RA (1975) A generalized form of Fisher's logarithmic series. Biometrika 62(1): 29-38.
12. Tripathi RC, Gupta RC (1985) A generalization of the log-series distribution. Communications in Statistics (Theory and Methods) 14(8): 1779-1799.
13. Mishra A, Shanker R (2002) Generalized logarithmic series distribution-Its nature and applications, Proceedings of the V International Symposium on Optimization and Statistics. 28-30, 155-168.
14. Gosset WS (1908) The Probable error of a mean. Biometrika 6(1): 1-25.
15. Mc Guire JU, Brindley TA, Bancroft TA (1957) The distribution of European corn-borer larvae pyrausta in field corn. Biometrics 13(1): 65-78.
16. Loeschke V, Kohler W (1976) Deterministic and Stochastic models of the negative binomial distribution and the analysis of chromosomal aberrations in human leukocytes. Biometrische Zeitschrift 18(6): 427-451.
17. Janardan KG, Schaeffer DJ (1977) Models for the analysis of chromosomal aberrations in human leukocytes. Biometrical Journal 19(8): 599-612.

18. Catcheside DG, Lea DE, Thoday JM (1946) Types of chromosome structural change induced by the irradiation on Tradescantia microspores. J Genet 47: 113-136.

19. Catcheside DG, Lea DE, Thoday JM (1946) The production of chromosome structural changes in Tradescannia microspores in relation to dosage, intensity and temperature. J Genet 47: 137–149.